Simulation

1. Simulating random variables
2. Variance reduction techniques
3. Simulating stochastic processes
4. Markov chain Monte Carlo methods

Index

The problem

Given a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with density function $f(x_1, \ldots, x_n)$, we want to compute

$$\theta = \mathbb{E}[g(\mathbf{X})] = \int\int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n,$$

for some n-dimensional function g.

Often it is not analytic possible to compute it exactly or to numerically approximate it. However, we can approximate it using simulation.

Monte Carlo simulation.

- Simulate r independent random vectors $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$, $i = 1, \ldots, r$, having density $f(x_1, \ldots, x_n)$.
- Compute $Y_i = g(\mathbf{X}^{(i)})$.
- SLLN: $\lim_{r\to\infty} \frac{\sum_{i=1}^{r} Y_i}{r} = \mathbb{E}[Y_i] = \mathbb{E}[g(\mathbf{X})]$ a.s..

How to simulate random vectors having a specified joint distribution?

Simulating a uniform random variable

- Number ten balls from 0 to 9, put them in a bag, and draw n balls, with replacement. The sequence of digits obtained represents the fractional part of a $U(0,1)$ r.v. rounded off to the nearest $\left(\frac{1}{10}\right)^n$.
- Digital computers use **pseudo random numbers**. Start with an initial value X_0 and, for suitable choices of a, c, m, recursively compute the values

$$X_{n+1} = (aX_n + c) \text{ modulo } m, \qquad n \geq 0.$$

  Each $\frac{X_n}{m}$ approximates a $U(0,1)$ and the sequence $\left(\frac{X_n}{m}\right)_n$ seems generated from independent $U(0,1)$ r.v.'s.

We assume we can simulate **random numbers**, i.e., independent $U(0,1)$, and we present three methods for simulating continuous r.v.'s.

The inverse transformation method

- **The inverse transformation method.** Let $U \sim U(0,1)$. For any continuous distribution function F, the r.v.
$$X = F^{-1}(U) = \inf\{x \,|\, F(x) \geq U\}$$
has distribution function F.

Proof. Since F is monotone,

$$F_X(a) = \mathbb{P}(X \leq a) = \mathbb{P}(F^{-1}(U) \leq a) = \mathbb{P}(U \leq F(a)) = F(a). \qquad \square$$

Note that the definition in the book ($F^{-1}(U) = x$ s.t. $F(x) = U$) is not proper: if the density is zero on an interval, then the value of x is not unique.

Hence, when F^{-1} is computable, we can simulate X from F by simulating $U \sim U(0,1)$ and then setting $X = F^{-1}(U)$.

Example

Example 11.3: Simulating an exponential r.v..

Exponential r.v.'s have distribution $F(x) = 1 - e^{-\lambda x}$. Hence, if $U \sim U(0,1)$, then $\frac{-\log(U)}{\lambda} \sim \mathrm{Exp}(\lambda)$ and $\frac{-c\log(U)}{\lambda} \sim \mathrm{Exp}(\frac{\lambda}{c})$.

Note that the log function is not the cheapest function to work with in mathematical programs.

The rejection method

- **The rejection method.** Suppose that we can simulate a r.v. with density $g(x)$. If $\frac{f(x)}{g(x)} \leq c$ for all x, then we can simulate a continuous r.v. X with density $f(x)$.
  1. Simulate Y with density g and $U \sim U(0,1)$.
  2. If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$, otherwise return to step 1.

Proof. For $K = \mathbb{P}\big(U \leq \frac{f(Y)}{cg(Y)}\big)$,

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(Y \leq x \,\middle|\, U \leq \frac{f(Y)}{cg(Y)}\right) = \frac{\mathbb{P}\left(Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right)}{K}$$

$$= \frac{\int_{-\infty}^{x} \mathbb{P}\big(U \leq \frac{f(y)}{cg(y)}\big) g(y)\, dy}{K} = \frac{\int_{-\infty}^{x} \frac{f(y)}{c}\, dy}{K},$$

and letting $x \to \infty$ shows that $K = \frac{1}{c}$. □

- Each iteration will result in an accepted value with probability

$$K = \mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \frac{1}{c},$$

hence number of iterations is geometric with mean c.

- It is not necessary to simulate a new $U(0,1)$ after rejection, but we can suitably modify the previous one, at the cost of some computation. Indeed, if Y is rejected, we can use

$$\frac{U - \frac{f(Y)}{cg(Y)}}{1 - \frac{f(Y)}{cg(Y)}} = \frac{cUg(Y) - f(Y)}{cg(Y) - f(Y)} \sim U(0,1).$$

Cost of simulation vs. cost of computation.

Examples

Example 11.4: simulating a beta random variable.

Example 11.5: simulating a normal random variable.

The hazard rate function

Consider a continuous positive r.v. X with distribution F and density f. The **hazard rate function** $\lambda(t)$ is defined by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

It represents the conditional probability density that a t-year-old item with lifetime X will fail. Indeed,

$$\begin{aligned}\mathbb{P}(X \in (t, t + dt) \,|\, X > t) &= \frac{\mathbb{P}(X \in (t, t + dt), X > t)}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(X \in (t, t + dt))}{\mathbb{P}(X > t)} \approx \frac{f(t)dt}{1 - F(t)} = \lambda(t)dt.\end{aligned}$$

The hazard rate method

- **The hazard rate method.** Given a bounded function $\lambda(t)$ s.t. $\int_0^\infty \lambda(t)\,dt = \infty$, we can simulate a r.v. S having $\lambda(t)$ as its hazard rate function.
  1. Simulate a Poisson process with rate λ s.t. $\lambda(t) \leq \lambda$ for all $t \geq 0$.
  2. Accept an event that occurs at time t with probability $\frac{\lambda(t)}{\lambda}$.
  3. Set S to be the time of the first accepted event.

Simulate pairs of r.v.'s $U_i \sim U(0,1), X_i \sim \mathrm{Exp}(\lambda), i \geq 1$. Stop at

$$N = \min\left\{ n : U_n \leq \frac{\lambda(\sum_{i=1}^n X_i)}{\lambda} \right\}$$

and set $S = \sum_{i=1}^N X_i$.

From Wald's equation, $\mathbb{E}[S] = \mathbb{E}\big[\sum_{i=1}^N X_i\big] = \mathbb{E}[X_i]\mathbb{E}[N] = \frac{\mathbb{E}[N]}{\lambda}$, hence the expected number of iterations is $\mathbb{E}[N] = \lambda\mathbb{E}[S]$.

Proof.

$$
\begin{aligned}
&\mathbb{P}(S \in (t, t+dt) \,|\, S > t) \\
&= \mathbb{P}(\text{first accepted event in } (t, t+dt) \,|\, \text{no accepted events prior to } t) \\
&= \mathbb{P}(\text{accepted Poisson event in } (t, t+dt) \,|\, \text{no accepted events prior to } t) \\
&= \mathbb{P}(\text{accepted Poisson event in } (t, t+dt)) \\
&= (\lambda dt + o(dt)) \frac{\lambda(t)}{\lambda} \\
&= \lambda(t) dt + o(dt).
\end{aligned}
$$

□

Special techniques for simulating continuous r.v.'s

Section 11.3: normal, gamma, chi-square, beta and exponential distributions.

Simulating from discrete distributions

The general methods for simulating from continuous distributions have analogues in the discrete case.

- **Analogue of the inverse transformation method.** In order to simulate a r.v. X having probability mass function

$$\mathbb{P}(X = x_j) = P_j,\ j = 1, 2, \dots, \qquad \sum_j P_j = 1,$$

let $U \in U(0,1)$ and set

$$X = \begin{cases} x_1, & \text{if } U < P_1, \\ x_2, & \text{if } P_1 < U < P_1 + P_2, \\ \vdots & \\ x_j, & \text{if } \sum_{i=1}^{j-1} P_i < U < \sum_{i=1}^{j} P_i, \\ \vdots & \end{cases}$$

Note that $\mathbb{P}(X = x_j) = \mathbb{P}(\sum_{i=1}^{j-1} P_i < U < \sum_{i=1}^{j} P_i) = P_j$.

Example

Example 11.9: simulating a Poisson r.v..

Index

Motivation for variance reduction

Given a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with density $f(x_1, \ldots, x_n)$ and some n-dimensional function g, we want to compute

$$\theta = \mathbb{E}[g(\mathbf{X})] = \int \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Monte Carlo simulation.

- Simulate r independent random vectors $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$, $i = 1, \ldots, r$, having density $f(x_1, \ldots, x_n)$.
- Compute $Y_i = g(\mathbf{X}^{(i)})$.
- SLLN: $\lim_{r\to\infty} \frac{\sum_{i=1}^r Y_i}{r} = \mathbb{E}[Y_i] = \mathbb{E}[g(\mathbf{X})]$ a.s..

Let $\bar{Y} = \frac{\sum_{i=1}^r Y_i}{r}$. To know how fast the convergence is, we need **control on the variance**

$$\mathrm{Var}(\bar{Y}) = \mathbb{E}[(\bar{Y} - \mathbb{E}[g(\mathbf{X})])^2],$$

and we will see three techniques for reducing it.

Use of antithetic variables

Example: Suppose we have generated Y_1, Y_2, identically distributed. If they are independent, then $\mathrm{Var}\left(\frac{Y_1+Y_2}{2}\right) = \frac{\mathrm{Var}(Y_1)}{2}$. However, if they are dependent and negatively correlated, i.e., $\mathrm{Cov}(Y_1, Y_2) \leq 0$, then the variance is reduced. Indeed,

$$\begin{aligned}\mathrm{Var}\left(\frac{Y_1+Y_2}{2}\right) &= \frac{\mathrm{Var}(Y_1)+\mathrm{Var}(Y_2)+2\mathrm{Cov}(Y_1,Y_2)}{4} \\ &= \frac{\mathrm{Var}(Y_1)}{2} + \frac{\mathrm{Cov}(Y_1,Y_2)}{2} \leq \frac{\mathrm{Var}(Y_1)}{2}.\end{aligned}$$

When simulating via the inverse transformation method ($X_i = F_i^{-1}(U_i)$ with $U_i \sim U(0,1)$, for $i = 1, \ldots, n$), we can use the following technique.

- **Use of antithetic variables.** If $U \sim U(0,1)$, then $1 - U \sim U(0,1)$ and they are negatively correlated. Hence, rather than generating r sets of n variables $U(0,1)$, we should generate $r/2$ sets and use each set twice.

Theorem

If $X_1, \ldots, X_n$ are independent, then, for any increasing functions f and g of n variables,

$$\mathbb{E}[f(\mathbf{X})g(\mathbf{X})] \geq \mathbb{E}[f(\mathbf{X})]\mathbb{E}[g(\mathbf{X})].$$

Proof. Proof by induction on n. For $n = 1$, for any i.i.d. r.v.'s X and Y, we have that $(f(X) - f(Y))(g(X) - g(Y)) \geq 0$ and

$$\begin{aligned} 0 &\leq \mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \\ &= \mathbb{E}[f(X)g(X) + f(Y)g(Y) - f(X)g(Y) - f(Y)g(X)] \\ &= 2\mathbb{E}[f(X)g(X)] - 2\mathbb{E}[f(X)]\mathbb{E}[g(X)]. \end{aligned}$$

For larger n, see the book. □

Corollary

If $U_1, \dots, U_n$ are independent, and h is either an increasing or decreasing function, then

$$\mathrm{Cov}\big(h(U_1, \dots, U_n), h(1 - U_1, \dots, 1 - U_n)\big) \leq 0.$$

Proof. If h is increasing, let $g(x_1, \dots, x_n) = -h(1 - x_1, \dots, 1 - x_n)$, and if h is decreasing, replace it with its negative. □

When simulating via the inverse transformation method ($X_i = F_i^{-1}(U_i)$), since $F_i^{-1}(U_i)$ is increasing in U_i, we have that $g(F_1^{-1}(U_1), \dots, F_n^{-1}(U_n)$ is monotone whenever g is monotone. Hence, the **antithetic variable approach of twice using each set** of $U_1, \dots, U_n$ by computing

$$g\big(F_1^{-1}(U_1), \cdots, F_n^{-1}(U_n)\big) \text{ and } g\big(F_1^{-1}(1 - U_1), \cdots, F_n^{-1}(1 - U_n)\big)$$

(which are identically distributed and negatively correlated) will reduce the variance of the estimate of $\theta = \mathbb{E}[g(X_1, \dots, X_n)]$.

Variance reduction by conditioning

Recall the conditional variance formula for r.v.'s Y and Z

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbb{E}[Y \mid Z]) \geq \mathrm{Var}(\mathbb{E}[Y \mid Z]).$$

- **Variance reduction by conditioning.** If we can compute $\mathbb{E}[Y \mid Z]$ for some cleverly chosen r.v. Z, then $\mathbb{E}[Y \mid Z]$ is a better estimator of $\mathbb{E}(Y)$ than is Y.

  Moreover, for any $\lambda_i \geq 0$ s.t. $\sum_i \lambda_i = 1$, and for a sequence of r.v.'s $Z_i, i \geq 1$, we have that

  $$\mathbb{E}\left[\sum_i \lambda_i \mathbb{E}[Y \mid Z_i]\right] = \mathbb{E}[Y]$$

  and

  $$\mathrm{Var}\left(\sum_i \lambda_i \mathbb{E}[Y \mid Z_i]\right) \leq \mathrm{Var}(Y).$$

Examples

Example 11.16: queueing system with capacity.

Example 11.18: estimating the renewal function.

Importance sampling

Suppose we want to estimate $\theta = \mathbb{E}[h(\mathbf{X})] = \int h(\mathbf{x})f(\mathbf{x})\,d\mathbf{x}$, but simulating $\mathbf{X}$ with density f is difficult or $\mathrm{Var}(h(\mathbf{X}))$ is large.

- **Importance sampling.** Let g be another density s.t. $f(\mathbf{x}) = 0$ if $g(\mathbf{x}) = 0$, and $\mathrm{Var}\left(\frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}\right)$ is small. Simulate $\mathbf{X}$ from g and let

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

Intuition: Since $\mathbf{X}$ has density $g(\mathbf{X})$, the ratio $\frac{f(\mathbf{X})}{g(\mathbf{X})}$ is usually small in comparison to 1. However, since $\mathbb{E}\left[\frac{f(\mathbf{X})}{g(\mathbf{X})}\right] = 1$, $\frac{f(\mathbf{X})}{g(\mathbf{X})}$ is occasionally large and $\mathrm{Var}\left(\frac{f(\mathbf{X})}{g(\mathbf{X})}\right)$ will tend to be large. We should choose g s.t. this ratio is large exactly when h is very small, so that $\frac{h(\mathbf{X})f(\mathbf{X})}{g(\mathbf{X})}$ is always small.

Tilted densities

Let X be a r.v. with density f, and $M(t) = \mathbb{E}[e^{tX}] = \int e^{tx} f(x)\, dx$ be its moment generating function. The **tilted density** of X is defined as

$$f_t(x) = \frac{e^{tx} f(x)}{M(t)}.$$

Example 11.22:

- If $X \sim \mathrm{Exp}(\lambda)$, then, for $t \leq \lambda$, $f_t(x)$ is an exponential density with rate $\lambda - t$.
- If $X \sim \mathrm{Ber}(p)$, then $f_t(x)$ is the probability mass function of a Bernoulli r.v. with parameter $p_t = \frac{pe^t}{pe^t + 1 - p}$.

For the importance sampling estimator, we can use $g = f_t$ for an appropriate choice of t.

Sum of independent random variables

If $\mathbf{X} = (X_1, \cdots X_n)$ is a vector of independent random variables with densities f_i, for $i = 1, \ldots, n$, then the joint density function is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

It is useful to simulate the X_i's according to their $f_{i,t}$ with a common t.

Example 11.23: **sum of independent r.v.'s**. For $S = \sum_{i=1}^{n} X_i$ and $a > \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, we want to approximate $\theta = \mathbb{P}(S \geq a) = \mathbb{E}[\mathbb{1}_{\{S \geq a\}}]$. At each iteration, estimate

$$\begin{aligned}\hat{\theta} &= \mathbb{1}_{\{S \geq a\}} \prod_{i=1}^{n} \frac{f_i(X_i)}{f_{i,t}(X_i)} = \mathbb{1}_{\{S \geq a\}} \prod_{i=1}^{n} M_i(t) e^{-tX_i} \\ &= \mathbb{1}_{\{S \geq a\}} M(t) e^{-tS} \leq M(t) e^{-ta},\end{aligned}$$

and choose t that minimizes $M(t)e^{-ta}$. It can be shown that the optimal $t = t^*$ is such that $\mathbb{E}[S] = a$ when the X_i's are simulated from f_{i,t^*}.

Index

1. Simulating random variables
2. Variance reduction techniques
3. Simulating stochastic processes
4. Markov chain Monte Carlo methods

Simulating stochastic processes

So far, we have seen how to simulate r.v.'s and random vectors. We can easily **simulate a stochastic process** by simulating a sequence of r.v.'s (not creative but often effective).

Example: simulating a renewal process.

1. Given an interarrival distribution F, simulate i.i.d. r.v.'s $X_1, X_2, \ldots$ with distribution F.
2. Stop at $N = \min\{n : \sum_{i=1}^{n} X_i > t\}$
3. The X_i's represents the interarrival times and the simulation yields $N - 1$ events by time t.

Simulating Poisson processes

Suppose that we want to **simulate a Poisson process** with rate λ until time t.

- Simulate the sequence of exponentially distributed arrival times.
- Another approach:
  1. Simulate $N(t) \sim \mathrm{Po}(\lambda t)$, the number of events by time t.
  2. If $N(t) = n$, simulate n $U(0,1)$ r.v.'s.
  3. To order them, rather than ordering a single list, create n random lists and put U in list i if $\frac{i-1}{n} \leq U < \frac{i}{n}$. Then order each list (quick) and obtain $U_1 < \cdots < U_n$.
  4. The values $\{tU_1, \ldots, tU_n\}$ represent the ordered times at which the events occur.

Nonhomogeneous Poisson processes (where $\lambda = \lambda(t)$) are usually not mathematically tractable, hence are strong candidates for simulations. We will present three methods.

Sampling a Poisson process

Sampling a Poisson process. By simulating a Poisson process with rate $\lambda \geq \lambda(t)$ for all $t \leq T$, and then randomly counting its events with probability $\frac{\lambda(t)}{\lambda}$ (thinning), we can simulate a nonhomogeneous Poisson process with intensity function $\lambda(t)$ up to time T.

Thinning algorithm:

1. Simulate independent r.v.'s $\{X_i \sim \mathrm{Exp}(\lambda)\}_i$ and $\{U_i \sim U(0,1)\}_i$.
2. Stop at $N = \min\{n : \sum_{i=1}^{n} X_i > T\}$.
3. For $j = 1, \ldots, N-1$, let $I_j = \begin{cases} 1, & \text{if } U_j \leq \frac{\lambda(\sum_{i=1}^{j} X_i)}{\lambda} \\ 0, & \text{otherwise,} \end{cases}$ and set $J = \{j : I_j = 1\}$.
4. The counting process having events at the set of times $\{\sum_{i=1}^{j} X_i : j \in J\}$ is a nonhomogeneous Poisson process on $[0, T]$ with intensity function $\lambda(t)$.

Most efficient if $\lambda(t)$ is close to λ throughout the interval, since we would have the fewest number of rejected events.

Improve the thinning method by breaking up the inteval $[0, T]$ into k subintervals $\{I_i = [t_{i-1}, t_i), i = 1, \dots, k\}$, with $t_0 = 0, t_k = T$, on which we sample Poisson processes using $\lambda_1, \dots, \lambda_k$ s.t. $\lambda(t) < \lambda_i$ for $t \in I_i$.

In the algorithm, t is the present time and I is the present interval.

1. Start with $t = 0$ and $I = 1$.
2. Simulate $X \sim \mathrm{Exp}(\lambda_I)$.
3. If $t + X < t_I$, set $t \to t + X$, simulate $U \sim U(0,1)$ and accept the event time t if $U \leq \frac{\lambda(t)}{\lambda_I}$. Return to step 2.
4. If $t + X \geq t_I$, stop if $I = k$, or set $X \to \frac{(X-(t_I-t))\lambda_I}{\lambda_{I+1}} \sim \mathrm{Exp}(\lambda_{I+1})$, $t \to t_I$, $I \to I + 1$, and go to step 3.

If on the subinterval I_i we have that $\underline{\lambda}_i = \min\{\lambda(s) : s \in I_i\} > 0$, then it is better to first simulate a Poisson process with rate λ_i, then simulate a nonhomogeneous Poisson process with intensity function $\lambda(s) - \lambda_i$, and merge the two processes.

Conditional distribution of the arrival times

For a nonhomogeneous Poisson process on $[0, T]$, given $N(T)$, the event times are i.i.d. with **conditional distribution**

$$F(t) = \frac{\int_0^t \lambda(s)}{m(T)} = \frac{\int_0^t \lambda(s)}{\int_0^T \lambda(s)\, ds}, \quad t \in (0, T).$$

Since $N(T) \sim \mathrm{Po}(m(T))$, we can simulate the nonhomogeneous Poisson process by first simulating $N(T)$ and then simulating $N(T)$ r.v.'s from their common density function $f(t) = \frac{\lambda(t)}{m(T)}$.

Example 11.12: $\lambda(t) = ct$.

Simulating the event times

The most basic approach is to **simulate the event times** in the order in which they occur.

If an event occurs at time x, then, independently of what has occurred prior to x, the time until the next event has distribution F_x s.t.

$$\begin{aligned} 1 - F_x(t) &= \mathbb{P}(\text{no events in } (x, x+t) \,|\, \text{event at } x) \\ &= \mathbb{P}(\text{no events in } (x, x+t)) \\ &= e^{-\int_x^t \lambda(s)\, ds}, \end{aligned}$$

and density

$$f_x(t) = \lambda(x+t) e^{-\int_0^t \lambda(x+s)\, ds}.$$

Simulate X_1 from F_0. If $X_1 = x_1$, simulate X_2 by adding x_1 to a value simulated from F_{x_1}. If $X_2 = x_2$, simulate X_3 by adding x_2 to a value simulated from F_{x_2}, and so on.

Example 11.13: $\lambda(t) = \frac{1}{t+a}$.

Index

Markov chain Monte Carlo methods

Let $\mathbf{X}$ be a discrete random vector taking values $\mathbf{x}_i$, $i \geq 1$, and with probability mass function $\mathbb{P}(\mathbf{X} = \mathbf{x}_i)$, for $i \geq 1$. For a given h, we want to compute

$$\theta = \mathbb{E}[h(\mathbf{X})] = \sum_{i=1}^{\infty} h(\mathbf{x}_i)\mathbb{P}(\mathbf{X} = \mathbf{x}_i).$$

- **Monte Carlo simulation.** Use $U(0,1)$ r.v.'s to simulate i.i.d. $\mathbf{X}_1, \ldots, \mathbf{X}_r$ with mass function $\mathbb{P}(\mathbf{X} = \mathbf{x}_i)$ for $i \geq 1$. From the SLLN, $\theta = \lim_{r\to\infty} \frac{\sum_{i=1}^{r} h(\mathbf{X}_i)}{r}$ a.s..
  Difficult to simulate the $\mathbf{X}_i$'s, especially if they are vectors of dependent r.v.'s. Moreover, often $\mathbb{P}(\mathbf{X} = \mathbf{x}_i) = Cb_i$, $i \geq 1$, with only the b_i's specified, and it is computationally hard to compute C.
- **Markov chain Monte Carlo (MCMC) method.** Simulate a sequence of the successive states of a (vector-valued) Markov chain $\mathbf{X}_1, \mathbf{X}_2, \ldots$ whose stationary distribution is π with $\pi_i = \mathbb{P}(\mathbf{X} = \mathbf{x}_i)$ for $i \geq 1$. Then $\theta = \lim_{r\to\infty} \frac{\sum_{i=1}^{r} h(\mathbf{X}_i)}{r}$.

Metropolis-Hastings algorithm

For $b_i > 0$ for $i \geq 1$ and $B = \sum_i b_i < \infty$, we want to generate a Markov chain with stationary probabilities $\pi_i = \frac{b_i}{B}$ for $i \geq 1$. In particular, we want to allow arbitrary stationary distributions that may only be specified up to a multiplicative constant.

Metropolis-Hastings algorithm to define a Markov chain with state space $\{X_n, n \geq 0\}$.

1. Let Q be any irreducible transition matrix with entries $q(i,j)$.
2. When $X_n = i$, simulate a r.v. Y s.t. $\mathbb{P}(Y = j) = q(i,j), j \geq 1$. If $Y = j$, then set $X_{n+1} = \begin{cases} j, & \text{w.p. } \alpha(i,j) \\ i, & \text{w.p. } 1 - \alpha(i,j). \end{cases}$.

- The Markov chain has transition probabilities $P_{i,j}$ given by

$$\begin{aligned} P_{i,j} =& q(i,j)\alpha(i,j), \qquad \text{if } j \neq i, \\ P_{i,i} =& q(i,i) + \sum_{k\neq i} q(i,k)(1-\alpha(i,k)). \end{aligned}$$

- The Markov chain has stationary probabilities π_i if it satisfies the balance equations for $j \neq i$

$$\begin{aligned} \pi_i P_{i,j} &= \pi_j P_{j,i} \\ \pi_i q(i,j)\alpha(i,j) &= \pi_j q(j,i)\alpha(j,i), \end{aligned}$$

  which are solved by taking $\pi_i = \frac{b_i}{B}$ and $\alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right)$.

- Since $\alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b_j q(j,i)}{b_i q(i,j)}, 1\right)$, the value of B is not needed to define the Markov chain.

- Almost always the stationary probabilities π_i's are also limiting probabilities (a sufficient condition is $P_{i,i} > 0$ for some i).

Gibbs sampling

We want to simulate a discrete random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with probability mass function $p(\mathbf{x}) = Cg(\mathbf{x})$, where g is known and C is not.

Gibbs sampling to define a vector-valued Markov chain.

1. When in state $\mathbf{x} = (x_1, \ldots, x_n)$, choose u.a.r. one coordinate, say the i-th coordinate.
2. Simulate a r.v. X with mass $\mathbb{P}(X = x) = \mathbb{P}(X_i = x \,|\, X_j = x_j, j \neq i\}$ (assume we can). If $X = x$, then consider as the candidate next state $\mathbf{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$.
3. Use the Metropolis-Hastings algorithm with
$$q(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\,\mathbb{P}(X_i = x \,|\, X_j = x_j, j \neq i\} = \frac{p(\mathbf{y})}{n\,\mathbb{P}(X_j = x_j, j \neq i)}.$$
4. The candidate state $\mathbf{y}$ is accepted with probability
$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{p(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{p(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1\right) = \min\left(\frac{p(\mathbf{y})p(\mathbf{x})}{p(\mathbf{x})p(\mathbf{y})}, 1\right) = 1,$$
hence it is always accepted.

The Ising model

- The **Ising model** is the simplest model of ferromagnetism, which arises when atomic spins align s.t. their magnetic moments all point in the same direction, yielding a macroscopic net magnetic moment.
- Consider discrete r.v.'s with spins $+1$ or -1 and a state space $\{-1,1\}^V$, where V is a large part of a lattice.
- The spins interact with their neighbors: spins that agree have a lower energy than spins that disagree. The energy of a state σ is given by the Hamiltonian

  $$H(\sigma) = \sum_{v \sim w} \mathbb{1}_{\{\sigma(v) \neq \sigma(w)\}}, \qquad \sigma \in \{-1,1\}^V, v, w \in V,$$

  and its probability by $\pi_\sigma = C_\beta e^{-\beta H(\sigma)}$, where $\beta > 0$ is a constant (inverse temperature) and C_β is a normalizing constant.
- The system tends to the lowest energy, but heat can disturb this tendency and create the possibility of different structural phases (phase transitions).

Gibbs sampling applied to the Ising model

Since C_β is hard to compute, direct sampling from the distribution π is hard. We can then **apply Gibbs sampling** to define a Markov chain on

$$\left\{X^{(k)} \in \{-1,1\}^V, k \geq 0\right\}.$$

1. Start in $X^{(0)}$ with all -1 or all 1. When in state $X^{(k)}$, choose u.a.r. one element, say the i-th element $X_i^{(k)}$.
2. The next state $X^{(k+1)}$ is s.t. $X_j^{(k+1)} = X_j^{(k)}$ for all $j \neq i$ and $X_i^{(k+1)}$ is a simulated r.v. with mass $\mathbb{P}\left(X_i^{(k+1)} = x \,|\, X_j^{(k)} = x_j, j \neq i\right)$.
   In particular, for the 1-dim Ising model,

$$\mathbb{P}\left(X_i^{(k+1)} = 1 \,|\, X_{i-1}^{(k)} + X_{i+1}^{(k)} = 0\right) = \frac{1}{2},$$

$$\mathbb{P}\left(X_i^{(k+1)} = 1 \,|\, X_{i-1}^{(k)} = X_{i+1}^{(k)} = 1\right) = \frac{1}{1+e^{-2\beta}},$$

$$\mathbb{P}\left(X_i^{(k+1)} = 1 \,|\, X_{i-1}^{(k)} = X_{i+1}^{(k)} = -1\right) = \frac{e^{-2\beta}}{1+e^{-2\beta}}.$$

Exercises

Session 8. Chapter 11: 1, 5, 7, 8, 13, 30-33. For 8 use Stirling's formula.

Session 9. Chapter 11: 17, 23, 24. For 23 assume that the intensity is strictly positive everywhere.