

# Chapter III

## Simulating random variables

- ▶ Inference techniques used so far have been based on **simulation**
- ▶ We now consider how to simulate  $X$  from  $f_X(x)$ .
- ▶ In semester 1 used MCMC - but simpler methods needed in order to do MCMC.
- ▶ Starting point: generate  $U$  from  $U[0, 1]$  distribution
- ▶ Then consider transformation  $g(U)$  to obtain a random draw from  $f_X(x)$ .

How could we generate  $U[0, 1]$  r.v.s with coin tosses?

### Sampling from $U(0, 1)$

Need to simulate independent random variables uniformly distributed on  $[0, 1]$ .

**Definition:** A sequence of pseudo-random numbers  $\{u_i\}$  is a deterministic sequence of numbers in  $[0, 1]$  having the same statistical properties as a similar sequence of random numbers. Ripley 1987.

The sequence  $\{u_i\}$  is reproducible provided  $u_1$  is known.

A good sequence would be “unpredictable to the uninitiated”.

### Congruential generators (D.H. Lehmer, 1949)

The general form of a congruential generator is

$$N_i = (aN_{i-1} + c) \bmod M,$$

$$U_i = N_i/M, \text{ where integers } a, c \in [0, M - 1]$$

If  $c = 0$ , it is called a *multiplicative congruential generator* (otherwise, *mixed*).

These numbers are restricted to the  $M$  possible values

$$0, \frac{1}{M}, \frac{2}{M}, \dots, \frac{M-1}{M}.$$

Clearly, they are *rational* numbers, but if  $M$  is large they will practically cover the reals in  $[0, 1]$ .

$N_1$ : the **seed**. Can be re-set so you can reproduce same set of uniform random numbers. In R, use `set.seed(i)`, where  $i$  an integer.

As soon as some  $N_i$  repeats, say,  $N_i = N_{i+T}$ , then the whole subsequence repeats, i.e.  $N_{i+t} = N_{i+T+t}$ ,  $t = 1, 2, \dots$

The least such  $T$  is called the *period*.

A good generator will have a long period.

The period cannot be longer than  $M$  and also depends on  $a$  and  $c$ .

Several useful Theorems exist concerning periods of congruential generators. For example, for  $c > 0$ ,  $T = M$  if and only if

1.  $c$  and  $M$  have no common factors (except 1),
2.  $1 = a \pmod{p}$  for every prime number that divides  $M$ ,
3.  $1 = a \pmod{4}$  if 4 divides  $M$ .

## Lattice structure

Notice that for a congruential generator

$$N_i - aN_{i-1} = c - bM,$$

where  $b > 0$  is an integer. Therefore,

$$U_i - aU_{i-1} = \frac{c}{M} - b.$$

The LHS lies in  $(-a, 1)$  since  $U_i \in [0, 1)$ .

Therefore,  $b$  can take at most  $a + 1$  distinct values.

If we plot points  $(U_{i-1}, U_i)$ , all the points will lie on at most  $a + 1$  parallel lines.

Usually  $M$  is chosen to make the modulus operation efficient, and then  $a$  and  $c$  are chosen to make the period as long as possible. Ripley suggests  $c = 0$  or  $c = 1$  is usually a good choice.

The NAG Fortran Library G05CAF

$$M = 2^{59} \quad a = 13^{13} \quad c = 0$$

Another recommended one is

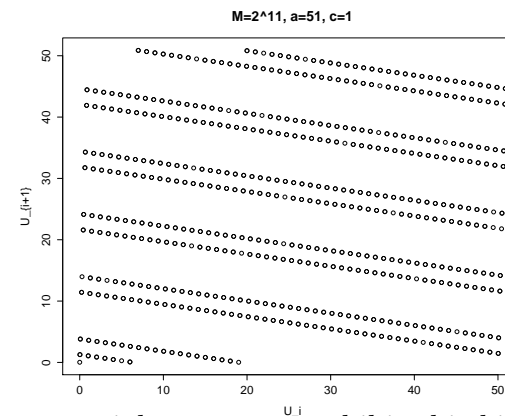
$$M = 2^{32} \quad a = 69069 \quad c = 1.$$

so that

$$N_i = (69069N_{i-1} + 1) \bmod 2^{32}$$

and

$$U_i = 2^{-32}N_i$$



All linear congruential generators exhibit this kind of lattice structure, not just for pairs  $(U_{i-1}, U_i)$ , but also for triples  $(U_{i-2}, U_{i-1}, U_i)$ , and in higher dimensions.

A good generator is expected to have *fine lattice structure*, that is, points  $(U_{i-k+1}, \dots, U_{i-1}, U_i) \in [0, 1)^k$  must lie on many hyperplanes in  $\mathbb{R}^k$  for all small  $k$  ( $k \ll M$ ).

Let  $U_i = N_i/m$  then for this generator

$$U_{i+2} - 6U_{i+1} + 9U_i = k \text{ an integer.}$$

Since  $0 \leq U_i < 1$

$$-6 < U_{i+2} - 6U_{i+1} + 9U_i < 10.$$

Therefore  $k = -5, -4, \dots, -1, 0, +1, \dots, 9$ .

Hence  $k$  can take on 15 integer values only, and subsequently  $(U_{i-2}, U_{i-1}, U_i)$  must lie on at most 15 parallel planes.

This is an example of *coarse lattice structure*, unsatisfactory coverage of  $[0, 1)^3$ .

## Generation from non- $U(0, 1)$

We have a sequence  $U_1, U_2, U_3, \dots$  of independent uniform random numbers in  $[0, 1]$ .

We want  $X_1, X_2, \dots$  distributed independently and identically from some specified distribution.

The answer is to transform the  $U_1, U_2, \dots$  sequence into  $X_1, X_2, \dots$  sequence.

The idea is to find a function  $g(U_1, U_2, U_3, \dots)$  that has the required distribution.

There are always many ways of doing this. A good algorithm should be quick because millions of random numbers may be required.

## 3.2 The inversion method

Let  $X$  be any continuous random variable and define  $Y = F_X(X)$ , where  $F_X$  is the distribution function of  $X$ :  $F_X(x) = P(X \leq x)$ .

**Claim:**  $Y \sim U[0, 1]$ .

**Proof**  $Y \in [0, 1]$  and the distribution function of  $Y$  is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y \end{aligned}$$

which is the distribution function of a uniform random variable on  $[0, 1]$ .

So whatever the distribution of  $X$ ,  $Y = F_X(X)$  is uniformly distributed on  $[0, 1]$ . The inversion method turns this backwards. Let  $U = F_X(X)$ , then  $X = F_X^{-1}(U)$ .

- So to generate  $X \sim F_X$  take a single uniform variable  $U$ , and set  $X = F_X^{-1}(U)$ .

## Example: exponential distribution

Let  $X \sim \text{Exp}(1/\lambda)$  (mean  $\lambda$ ), i.e.

$$f(x) = \lambda^{-1} e^{-x/\lambda} \quad (x \geq 0)$$

$$F(x) = \int_0^x \lambda^{-1} e^{-z/\lambda} dz = [-e^{-z/\lambda}]_0^x = 1 - e^{-x/\lambda}.$$

Set  $U = 1 - e^{-X/\lambda}$  and solve for  $X$

$$X = -\lambda \ln(1 - U).$$

Note that  $1 - U$  is uniformly distributed on  $[0, 1]$ , so we might as well use

$$X = -\lambda \ln U.$$

**Question:** What are the limitations of the inversion method?

## Discrete distributions - example

Let  $X \sim \text{Bin}(4, 0.3)$ . The probabilities are

$$P(X = 0) = .2401, \quad P(X = 1) = .4116, \quad P(X = 2) = .2646$$

$$P(X = 3) = .0756, \quad P(X = 4) = .0081.$$

The algorithm says  $X = 0$  if  $0 \leq U \leq .2401$ ,  
 $X = 1$  if  $.2401 < U \leq .6517$ ,  
 $X = 2$  if  $.6517 < U \leq .9163$ ,  
 $X = 3$  if  $.9163 < U \leq .9919$ ,  
 $X = 4$  if  $.9919 < U \leq 1$ .

Carrying out the binomial algorithm means the following. Let  $U \sim U(0, 1)$ .

1. Test  $U \leq .2401$ . If true, return  $X = 0$ .
2. If false, test  $U \leq .6517$ . If true, return  $X = 1$ .
3. If false, test  $U \leq .9163$ . If true, return  $X = 2$ .
4. If false, test  $U \leq .9919$ . If true, return  $X = 3$ .
5. If false, return  $X = 4$ .

## Discrete distributions

The inversion method works for discrete random variables in the following sense.

Let  $X$  be discretely distributed with possible values  $x_i$  having probabilities  $p_i$ . So

$$P(X = x_i) = p_i, \quad \sum_{i=1}^k p_i = 1.$$

Then  $F_X(x) = \sum_{x_i \leq x} p_i$  is a step function.

Inversion gives  $X = x_i$  if  $\sum_{x_j < x_i} p_j < U \leq \sum_{x_j \leq x_i} p_j$  which clearly gives the right probability values.

- Think of this as splitting  $[0, 1]$  into intervals of length  $p_i$ . The interval in which  $U$  falls is the value of  $X$ .

**Question:** What problems might we face using this method?  
Eg Consider a Poisson(100) distribution.

## Discrete distributions - example

Consider the speed of this. The expected number of steps (which roughly equates to speed) is

$$\begin{aligned} 1 \times .2401 + 2 \times .4116 + 3 \times .2646 + 4 \times .0756 + 4 \times .0081 \\ = 1 + E(X) - 0.0081 = 2.1919 \end{aligned}$$

To speed things up we can rearrange the order so that the later steps are less likely.

1. Test  $U \leq .4116$ . If true return  $X = 1$ .
2. If false, test  $U \leq .6762$ . If true return  $X = 2$ .
3. If false, test  $U \leq .9163$ . If true return  $X = 0$ .
4. and 5. as before.

Expected number of steps:

$$1 \times .4116 + 2 \times .2646 + 3 \times .2401 + 4 \times (0.0956 + 0.0081) = 1.9959.$$

Approximate 10% speed increase.

### 3.3 Transformations

- (a) If  $U \sim U(0, 1)$  set  $V = (b - a)U + a$  then  $V \sim U(a, b)$  where  $a < b$ .
- (b) If  $Y_i$  are iid exponential with parameter  $\lambda$  then

$$X = \sum_{i=1}^n Y_i = -\frac{1}{\lambda} \sum_{i=1}^n \log U_i = -\frac{1}{\lambda} \log \left( \prod_{i=1}^n U_i \right)$$

has a  $Ga(n, \lambda)$  distribution.

- (c) If  $X_1 \sim Ga(p, 1)$ ,  $X_2 \sim Ga(q, 1)$ ,  $X_1$  and  $X_2$  independent then  $Y = X_1 / (X_1 + X_2) \sim Be(p, q)$ .
- (d) Composition: if

$$f = \sum_{i=1}^r p_i f_i$$

where  $\sum p_i = 1$  and each  $f_i$  is a density, then we can sample from  $f$  by first sampling  $I$  from the discrete distribution  $p = \{p_1, \dots, p_r\}$  and then taking a sample from  $f_I$ .

### 3.4 Rejection Algorithm

#### Fundamental Theorem of Simulation:

Simulating

$$X \sim f(x)$$

is equivalent to simulating

$$(X, U) \sim U\{(x, u) : 0 < u < f(x)\}.$$

Note that  $f(x, u) = \mathbb{I}_{0 < u < f(x)}$  so that

$$\int f(x, u) du = \int_0^{f(x)} du = f(x)$$

as required.

Hence,  $f$  is the marginal density of the joint distribution  $(X, U) \sim U\{(x, u) : 0 < u < f(x)\}$ .

### The Box-Müller algorithm for the normal distribution

We cannot generate a normal random variable by inversion, because  $F_X$  is not known in closed form (nor its inverse).

**The Box-Müller method (1958).** Let  $U_1, U_2 \sim U[0, 1]$ . Calculate

$$X_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2),$$

$$X_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2).$$

Then  $X_1$  and  $X_2$  are independent  $N(0, 1)$  variables.

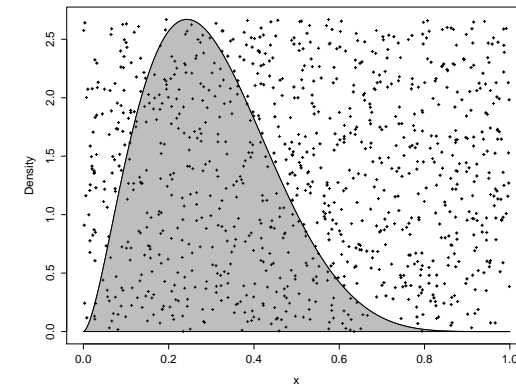
The method is not particularly fast, but is easy to program and quite memorable.

### Rejection Algorithm Explained

The problem with this result is that simulating uniformly from the set

$$\{(x, u) : 0 < u < f(x)\}$$

may not be possible. A solution is to simulate the pair  $(X, U)$  in a bigger set, where simulation is easier, and then take the pair if the constraint is satisfied.



## Rejection: Uniform bounding box

Suppose that  $f(x)$  is zero outside the interval  $[a, b]$  (so that  $\int_a^b f(x)dx = 1$ ) and that  $f$  is bounded above by  $m$ .

- ▶ Simulate the pair  $(Y, U) \sim U[a, b] \times [0, m]$  ( $Y \sim U[a, b]$ ,  $U \sim U[0, m]$  independently).
- ▶ Accept the pair if the constraint  $0 < U < f(Y)$  is satisfied.

This results in the correct distribution for the accepted  $Y$  value, call it  $X$ .

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(Y \leq x | U < f(Y)) \\ &= \frac{\int_a^x \int_0^{f(y)} du dy}{\int_a^b \int_0^{f(y)} du dy} \\ &= \int_a^x f(y) dy.\end{aligned}$$

Note: we can use the rejection algorithm even if we only know  $f$  upto a normalising constant (as is often the case in Bayesian statistics - see chapter 4).

## Generalising the Rejection Idea

If the support of  $f$  is not finite, then bounding it within a rectangle will not work. Instead of using a box to bound the density  $f(x)$  (ie requiring  $f(x) < m$  for some constant  $m$ ) we can use a function  $m(x)$  such that  $f(x) \leq m(x)$  for all  $x$ .

Suppose the larger bounding set is

$$\mathcal{L} = \{(y, u) : 0 < u < m(y)\}$$

then all we require is that simulation of a uniform from  $\mathcal{L}$  is feasible. Note

- ▶ The closer  $m$  is to  $f$  the more efficient our algorithm.
- ▶ Because  $m(x) \geq f(x)$ ,  $m$  cannot be a probability density. We write

$$m(x) = Mg(x) \text{ where } \int m(x)dx = \int Mg(x)dx = M$$

for some density  $g$ .

## Example: Sampling from a beta distribution

Consider sampling from  $X \sim \text{Beta}(\alpha, \beta)$  for  $\alpha, \beta > 1$  which has pdf

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 < x < 1.$$

We note

$$f(x) \propto f_1(x) = x^{\alpha-1} (1-x)^{\beta-1} \quad 0 < x < 1$$

and that  $M = \sup_{0 < x < 1} x^{\alpha-1} (1-x)^{\beta-1}$  occurs at  $x = \frac{\alpha-1}{\alpha+\beta-2}$  (mode) and hence

$$M = \frac{(\alpha-1)^{\alpha-1} (\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}.$$

The rejection algorithm is

1. Generate  $Y \sim U(0, 1)$  and  $U \sim U(0, M)$ .
2. If  $U \leq f_1(Y) = Y^{\alpha-1} (1-Y)^{\beta-1}$  then let  $X = Y$  (accept) else go to 1 (reject).

## Generalising the Rejection Idea II

This suggests a more general implementation of the fundamental theorem:

**Corollary:** Let  $X \sim f(x)$  and let  $g(x)$  be a density function that satisfies  $f(x) \leq Mg(x)$  for some constant  $M \geq 1$ . Then, to simulate  $X \sim f$ , it is sufficient to generate

$$Y \sim g \quad \text{and} \quad U | Y = y \sim U(0, Mg(y))$$

and set  $X = Y$  if  $U \leq f(Y)$ .

**Proof:**

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(Y \in A | U \leq f(Y)) \\ &= \frac{\int_A \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} \\ &= \int_A f(y) dy\end{aligned}$$

## The Rejection Algorithm

The rejection algorithm is usually stated in a slightly modified form:

### Rejection Algorithm

If  $g$  is such that  $f/g$  is bounded, so there exists  $M$  such that  $Mg(x) \geq f(x)$  for all  $x$  then

1. Generate  $Y$  from density  $g$ , and  $U$  from  $U(0, 1)$ .
2. If  $U \leq f(Y)/Mg(Y)$  set  $X = Y$ . Otherwise, return to step 1.  
produces simulations from  $f$

We keep sampling new  $Y$  and  $U$  until the condition is satisfied.

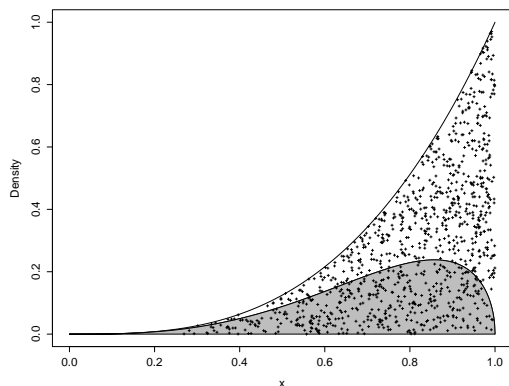
Exercise: Convince yourself that these two descriptions of the rejection algorithm are the same.

How to simulate  $Y$  with pdf  $g(y) = \alpha y^{\alpha-1}$ ?

- We note that the cdf of  $Y$  is  $G(y) = y^\alpha$ ,  $0 < y < 1$ .
- Therefore we can use inversion. Let  $Z \sim U(0, 1)$  then solve  $Z = G(Y) = Y^\alpha$  and so  $Y = Z^{\frac{1}{\alpha}}$ .

**Full algorithm is:**

1. Generate  $U \sim U(0, 1)$  and  $Z \sim U(0, 1)$ . Let  $Y = Z^{\frac{1}{\alpha}}$ .
2. If  $U \leq (1 - Y)^{\beta-1}$  then set  $X = Y$  else go to 1.



## Example: Sampling from a beta distribution revisited

Use rejection to sample from  $X \sim \text{Beta}(\alpha, \beta)$ . Let  $g(y) = \alpha y^{\alpha-1}$ ,  $0 < y < 1$ , then

$$\frac{f_1(x)}{g(x)} = \frac{(1-x)^{\beta-1}}{\alpha} \quad \text{is bounded if and only if } \beta \geq 1$$

Then  $M = \sup_x \left\{ \frac{f_1(x)}{g(x)} \right\} = \frac{1}{\alpha}$  occurs at  $x = 0$ .

1. Simulate  $Y$  with pdf  $g(y) = \alpha y^{\alpha-1}$ ,  $0 < y < 1$  and  $U \sim U(0, 1)$ .
2. If  $U \leq \frac{f_1(Y)}{Mg(Y)} = \frac{(1-Y)^{\beta-1}}{(\frac{1}{\alpha})\alpha} = (1-Y)^{\beta-1}$  then set  $X = Y$  else go to 1.

## Efficiency of the rejection method

Each time we generate a  $(Y, U)$  pair,

$$\text{Prob(Reject)} = P(U \geq f(Y)/Mg(Y)) = 1 - \frac{1}{M}, \quad \text{Prob(Accept)} = \frac{1}{M}.$$

The number of tries until we accept  $Y$  is a geometric random variable with expectation  $M$ .

Note that  $M$  here must be calculated with the normalised density  $f$ , i.e.,  $M = \sup \frac{f(x)}{g(x)}$ .

If we used an unnormalised density  $f_1(x)$ , where  $\int f_1(x)dx = c$ , so that  $f(x) = \frac{1}{c}f_1(x)$ , then if we used

$$M = \sup \frac{f_1(x)}{g(x)}$$

the acceptance rate is

$$\mathbb{P}(\text{Accept}) = \frac{c}{M}$$

## Rejection Example III

Let  $\theta$  have von Mises distribution with pdf

$$f(\theta) = \frac{\exp(k \cos \theta)}{2\pi I(k)} \quad 0 < \theta < 2\pi \quad (k \geq 0)$$

where  $I(k)$  is the normalising constant.

Let  $f_1(\theta) = \frac{1}{2\pi} \exp(k \cos \theta)$ ,  $0 < \theta < 2\pi$ .

$Y \sim U(0, 2\pi)$  so that  $g(y) = \frac{1}{2\pi}$ ,  $0 < y < 2\pi$ .

Then

$$M = \sup_{\theta} \left\{ \frac{f_1(\theta)}{g(\theta)} \right\} = \sup_{\theta} \{ \exp(k \cos \theta) \} = \exp k.$$

Let  $U \sim U(0, 1)$ .

If

$$U \leq \frac{f_1(Y)}{Mg(Y)} = \frac{\exp(k \cos Y)}{2\pi \cdot \frac{1}{2\pi} \cdot \exp k} = \exp(k(\cos Y - 1))$$

we accept  $\theta = Y$  otherwise reject.

## Truncated distributions

Suppose we wish to sample  $X$  from the following distribution:

$$f_X(x) \propto \begin{cases} g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

where  $g_X(x)$  is a known density that we can sample from, e.g.  $g_X(x)$  is the  $N(0, 1)$  density, and  $A = [0, \infty)$ .

$$f_X(x) = \begin{cases} k g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

where  $k$  is a normalising constant, given by

$$k^{-1} = \int_A g_X(x) dx$$

$$f_X(x) \propto \begin{cases} g_X(x) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Consider using rejection method to sample  $X$  from  $f_X(x)$ . We sample  $Y$  from the full (non-truncated) density  $g_X(x)$ .

$$\frac{f_X(x)}{g_X(x)} = \begin{cases} k & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$



So  $M = \sup_x \frac{f_X(x)}{g_X(x)} = k$ .

Rejection algorithm: sample  $u$  from  $U[0, 1]$  and  $y$  from  $g_Y(y)$ , and accept  $X = y$  if  $u \leq \frac{f_X(y)}{M g_Y(y)}$ .

But since

$$\frac{f_X(x)}{M g_X(x)} = \begin{cases} \frac{f_X(x)}{k g_X(x)} = 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

we will always have  $u \leq \frac{f_X(y)}{M g_Y(y)}$  if  $y \in A$ , and  $u \geq \frac{f_X(y)}{M g_Y(y)}$  if  $y \notin A$ .

So we don't need to sample  $u$ . Can just do

1. generate  $y$  from  $g_Y(y)$
2. if  $y \in A$ , accept  $X = y$
3. otherwise, return to step 1.

As usual, acceptance probability will be high if  $M$  is small, i.e.  $\int_A g_Y(y) dy$  is near 1. So if the truncated region is large, rejection sampling will be inefficient.

## Sequential methods

We can obviously write

$$f(\mathbf{x}) = f_1(x_1)f_2(x_2|x_1)f_3(x_3|x_1, x_2) \dots$$

So we can first generate  $X_1$  from  $f_1$ . Then for that given value of  $X_1$ , generate  $X_2$  from  $f_2$ , and so on.

## 3.5 Multivariate generators

Now suppose we want to generate a random vector

$\mathbf{X} = (X_1, \dots, X_p)$  from density  $f(\mathbf{x})$ . We can note the following simple points.

1. If the elements of  $\mathbf{X}$  are to be independent, i.e.

$$f(\mathbf{x}) = f_1(x_1)f_2(x_2) \dots f_p(x_p),$$

then we can separately generate  $X_1$  from  $f_1$ ,  $X_2$  from  $f_2, \dots, X_p$  from  $f_p$  using different uniforms.

2. Inversion no longer works as the theorem can't be generalised.
3. Rejection *does* work. If we can generate from  $g(\mathbf{x})$  (and  $g$  may be a product of independent components) and find  $M \geq \sup_{\mathbf{x}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$  and otherwise reject.

## Example

REPLACE THIS EXAMPLE WITH Normal inverse gamma instead.

We wish to sample  $\{\theta, \phi\}$  from the density function

$$f(\theta, \phi) \propto \phi^3 \exp\{-1(1 + 5\phi)\theta^2 + 40\phi\theta - 81\phi\}.$$

Firstly, it can be shown that the marginal density of  $\phi$  is

$$f(\phi) \propto \phi^3(1 + 5\phi)^{-1/3} \exp\left\{-\phi\left(1 + \frac{80}{1 + 5\phi}\right)\right\}.$$

Additionally, the conditional distribution of  $\theta|\phi$  is normal:

$$\theta|\phi \sim N\left(\frac{20\phi}{1+5\phi}, \frac{1}{2(1+5\phi)}\right).$$

To sample from  $f(\theta, \phi)$  we could first sample  $\phi$  from  $f(\phi)$ , using rejection sampling, then generate  $Z$  from  $N(0, 1)$ , and finally set

$$\theta = \frac{20\phi}{1+5\phi} + \frac{1}{\sqrt{2(1+5\phi)}}Z.$$

### Multivariate normal distributions

Consider generating  $\mathbf{X}$  from  $N(\mathbf{m}, V)$ , for some non-diagonal matrix  $V$ , given a sample of independent standard normal random variables  $Z_1, Z_2, \dots$

One technique involves the use of the **Cholesky square root** of the matrix  $V$ . For any (symmetric, square) positive definite matrix  $V$ , we can find a square root  $U$ , such that  $U^T U = V$ .

Multiple solutions for  $U$ , but upper triangular matrix  $U$  known as the Cholesky square root. To find the Cholesky square root of a matrix  $V$  in R, type `chol(V)`.

Define  $\mathbf{Z} \sim N(\mathbf{0}, I)$  (equal dimension to  $\mathbf{X}$ ), consider

$\mathbf{Y} = \mathbf{m} + U^T \mathbf{Z}$ . Then have  $\mathbf{Y}$  normally distributed, (each element of  $\mathbf{Y}$  is linear combination of the elements of  $\mathbf{Z}$ ), and

$$\begin{aligned} E(\mathbf{m} + U^T \mathbf{Z}) &= \mathbf{m}, \\ Var(\mathbf{m} + U^T \mathbf{Z}) &= U^T I U = V, \end{aligned}$$

(with  $I$  the identity matrix, the variance of  $\mathbf{Z}$ ).

Hence to generate  $\mathbf{X}$ , we generate independent standard normal random variables  $\mathbf{Z}$ , and then transform them by  $\mathbf{m} + U^T \mathbf{Z}$  to obtain  $\mathbf{X}$ .

## 3.6 Importance sampling

In order to estimate an integral of the form  $\int h(x)f(x)dx$  we find that it is sometimes better to generate values not from the distribution  $f(x)$ , but instead from some other distribution  $g(x)$  and to then account for this by using a weighting. This is the idea behind importance sampling. To introduce the idea we consider a simple example.

### Example of Monte Carlo/Importance Sampling

Let  $X$  be Cauchy  $f(x) = \frac{1}{\pi(1+x^2)}$ ,  $-\infty < x < \infty$ .

Let  $\theta = P(X > 2) = I = \int_2^\infty \frac{1}{\pi(1+x^2)} dx$  ( $= 0.1476$ ).

Use Monte Carlo Methods to estimate  $\theta$ .

(i) Generate  $n$  Cauchy variates,  $X_1, \dots, X_n$ .

Let  $Y_1$  be the number that are greater than 2,

$Y_1 = \sum \mathbb{I}_{X_i > 2}$ . Then  $Y_1 \sim B(n, \theta)$  so that

$$E(Y_1) = n\theta, \quad V(Y_1) = n\theta(1-\theta)$$

$$\hat{\theta}_1 = \frac{Y_1}{n}$$

$$E(\hat{\theta}_1) = \frac{E(Y_1)}{n} = \frac{n\theta}{n} = \theta$$

and

$$V(\hat{\theta}_1) = \frac{V(Y_1)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} = \frac{0.126}{n}.$$

## Example of Monte Carlo/Importance Sampling - II

- (ii) Note that  $\theta = \frac{1}{2}P(|X| > 2)$  - we want to use this to reduce the variance of our estimator  $\hat{\theta}$ .

Generate  $n$  Cauchy variates.

Let  $Y_2$  be the number that are greater than 2 in modulus then  $Y_2 \sim B(n, 2\theta)$

$$\text{and } \hat{\theta}_2 = \frac{1}{2} \frac{Y_2}{n}$$

$$\implies E(\hat{\theta}_2) = \frac{1}{2} \frac{E(Y_2)}{n} = \frac{1}{2} \cdot \frac{n2\theta}{n} = \theta$$

and

$$V(\hat{\theta}_2) = \frac{V(Y_2)}{2^2 n^2} = \frac{n2\theta(1-2\theta)}{2^2 n^2} = \frac{\theta(1-2\theta)}{2n} = \frac{0.052}{n}.$$

## Example of Monte Carlo/Importance Sampling - IV

We can see that

$$\mathbb{E}(\hat{\theta}_3) = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n \int_0^2 \frac{2}{\pi(1+x^2)} dx = \frac{1}{2} - \mathbb{P}(0 < X < 2)$$

where  $X \sim \text{Cauchy}$ , so that it too is an unbiased estimator.

The variance of  $\hat{\theta}_3$  is  $\text{Var}(h(U))/n$  and we can see that

$$\begin{aligned} \mathbb{E}h(U) &= \int_0^2 h(x) \frac{1}{2} dx = 0.5 - 0.1475 = 0.3525 \\ \mathbb{E}h(U)^2 &= \int_0^2 h(x)^2 \frac{1}{2} dx = \int_0^2 \frac{2}{\pi^2(1+x^2)^2} dx \\ &= \frac{1}{\pi^2} \left[ \frac{x}{x^2+1} + \tan^{-1}(x) \right]_0^2 = 0.1527 \end{aligned}$$

Hence  $\text{Var}(h(x)) = 0.1527 - 0.3525^2 = 0.02851$  and thus

$$\text{Var}(\hat{\theta}_3) = \frac{0.02851}{n}$$

## Example of Monte Carlo/Importance Sampling - III

- (iii) The relative inefficiency of these methods is due to generation of values outside the domain of interest  $[2, \infty)$ . Alternatively note we can write

$$\theta = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx.$$

This integral can be considered the expectation of  $h(X) = \frac{2}{\pi(1+x^2)}$  where  $X \sim U[0, 2]$  as the density of  $U[0, 2]$  is  $g(x) = 1/2$ .

An alternative method of evaluation of  $\theta$  is therefore

$$\hat{\theta}_3 = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n h(U_i)$$

where  $U_i \sim U[0, 2]$ .

## Example of Monte Carlo/Importance Sampling - V

- (iv) Finally, note that another possibility is to note that if

$$y = \frac{1}{x}$$

$$\theta = \int_{+2}^{\infty} \frac{1}{\pi(1+x^2)} dx = \int_0^{\frac{1}{2}} \frac{y^{-2} dy}{\pi(1+y^{-2})} = \int_0^{\frac{1}{2}} h(y) dy.$$

This can be seen as the expectation of  $h(X) = \frac{X^{-2}}{2\pi(1+X^{-2})}$  where  $X \sim U[0, \frac{1}{2}]$ . We can estimate this as

$$\hat{\theta}_4 = \frac{1}{n} \sum_{i=1}^n h(U_i)$$

where  $U_1, \dots, U_n \sim U[0, 1/2]$ .

Again, we have  $\mathbb{E}\hat{\theta}_4 = \theta$  and now

$$\mathbb{E}h(U)^2 = \int_0^{1/2} h(x)^2 \cdot 2 dx = \frac{1}{4\pi^2} \left[ \frac{x}{x^2+1} + \tan^{-1}(x) \right]_0^{1/2} = 0.02188$$

$$\text{Hence } \text{Var}(\hat{\theta}_4) = \frac{0.02188 - 0.1476^2}{n} = \frac{0.0000955}{n}$$

## Summary of Example

We found 4 unbiased estimators of  $\theta$ , each with a different variance.

$$\begin{aligned}\text{Var}(\hat{\theta}_1) &= \frac{0.126}{n} & \text{Var}(\hat{\theta}_2) &= \frac{0.052}{n} \\ \text{Var}(\hat{\theta}_3) &= \frac{0.02851}{n} & \text{Var}(\hat{\theta}_4) &= \frac{0.0000955}{n}\end{aligned}$$

The best estimator is the one with the smallest variance, namely  $\hat{\theta}_4$ .

Compared with  $\hat{\theta}_1$ , the evaluation of  $\hat{\theta}_4$  requires  $\sqrt{(0.126/0.0000955)} \approx 36$  times fewer simulations to achieve the same precision.

By carefully considering our simulation method we can hope to get more accurate estimates.

Estimate  $\hat{\theta}_2$  and  $\hat{\theta}_4$  are both types of importance sampling.

Some comments:

- ▶  $g(\mathbf{x})$  is called the importance function, and  $w(\mathbf{X}_i)$  are called the importance weights.
- ▶ The sum (1) will converge for the same reasons the Monte Carlo sum does.
- ▶ Notice that this sum is valid for any choice of the distribution  $g$ , as long as  $\text{supp}(f) \subseteq \text{supp}(g)$ .
- ▶ This is a very general representation that expresses the fact that a given integral is not intrinsically associated with a given distribution.
- ▶ Because very little restriction is put on the choice  $g$ , we can choose a distribution which is easy to sample from, and one which gives nice properties for the sum.

## Importance Sampling

Consider calculating the integral

$$I = \mathbb{E}_f h(X) = \int h(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x}.$$

### Importance sampling

Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be independently and identically distributed random variables with common density  $g(\mathbf{x})$ .

Define  $w(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$ , so that

$$\mathbb{E}_g \{h(\mathbf{X}_i)w(\mathbf{X}_i)\} = \int h(\mathbf{x})w(\mathbf{x})g(\mathbf{x}) \, d\mathbf{x} = \int h(\mathbf{x})f(\mathbf{x}) \, d\mathbf{x} = I.$$

Therefore

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n w(\mathbf{X}_i)h(\mathbf{X}_i) \quad (1)$$

is an unbiased estimator of  $I$ .

## Cauchy example revisited

We can now understand the estimator  $\hat{\theta}_4$  in the Cauchy example. Recall that we want to estimate

$$\mathbb{E}\mathbb{I}_{X>2} = \int h(x)f(x)dx$$

where  $h(x) = \mathbb{I}_{x>2}$  and  $f(x) = \frac{1}{\pi(1+x^2)}$ .

Noticing that for large  $x$ ,  $f(x)$  is similar to the density

$$g(x) = 2/x^2 \text{ for } x > 2.$$

suggests  $g()$  might be a good importance density. We can sample from  $g$  by letting  $X_i = 1/U_i$  where  $U_i \sim U[0, \frac{1}{2}]$  (inversion method). Thus our estimator is

$$\begin{aligned}\hat{\theta} &= \frac{1}{n} \sum_{i=1}^n h(x_i) \frac{f(x_i)}{g(x_i)} = \frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{2\pi(1+x_i^2)} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{u_i^{-2}}{2\pi(1+u_i^{-2})} = \hat{\theta}_4\end{aligned}$$

## The variance of the estimator

Since the  $\mathbf{X}_i$ s are iid,  $\text{Var}(\hat{I}) = \frac{\sigma^2}{n}$ , where

$$\begin{aligned}\sigma^2 &= \text{Var}_g\{h(\mathbf{X})w(\mathbf{X})\} = \mathbb{E}\{h(\mathbf{X})^2w(\mathbf{X})^2\} - \mathbb{E}\{h(\mathbf{X})w(\mathbf{X})\}^2 \\ &= \int h(\mathbf{x})^2w(\mathbf{x})^2g(\mathbf{x}) \, d\mathbf{x} - I^2 \\ &= \int \frac{h(\mathbf{x})^2f(\mathbf{x})^2}{g(\mathbf{x})} \, d\mathbf{x} - I^2 \quad \text{since} \quad g(\mathbf{x}) = \frac{f(\mathbf{x})}{w(\mathbf{x})}.\end{aligned}$$

We do not of course know  $\sigma^2$  in practice, but we can see that  $\hat{I}$  will be a better estimator if we can make  $w(\mathbf{X})$  less variable. Our objective, therefore, is to find a distribution  $g(\mathbf{x})$  that we know how to obtain independent samples from, and which mimics  $h(\mathbf{x})f(\mathbf{x})$  as closely as possible.

## Unnormalised densities

Suppose we only know  $f$  upto a normalising constant, i.e., we know

$$f(x) = \frac{f_1(x)}{c} \quad \text{where } c = \int f_1(x)dx$$

We can still use importance sampling

### Importance sampling with unnormalised densities

Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be independently and identically distributed random variables with common density  $g(\mathbf{x})$ .

Define  $\tilde{w}(\mathbf{x}) = f_1(\mathbf{x})/g(\mathbf{x})$ . Estimate  $I$  by

$$\hat{I} = \frac{\sum_{i=1}^n \tilde{w}(\mathbf{X}_i)h(\mathbf{X}_i)}{\sum_{i=1}^n \tilde{w}(\mathbf{X}_i)}$$

Alternatively, we can write this as

$$\hat{I} = \sum_{i=1}^n w_i h(\mathbf{X}_i) \quad \text{where} \quad w_i = \frac{\tilde{w}(\mathbf{X}_i)}{\sum \tilde{w}(\mathbf{X}_i)}$$

## Optimal choice of $g$

**Theorem** The choice of  $g = g^* = \frac{|h(x)|f(x)}{\int |h(z)|f(z)dz}$  minimises the variance of the estimator (1).

**Proof** We've seen that it is sufficient to minimise

$$\int \frac{h^2(\mathbf{x})f^2(\mathbf{x})}{g(\mathbf{x})} \, d\mathbf{x} = \mathbb{E}_g \left( \frac{h^2(\mathbf{X})f^2(\mathbf{X})}{g^2(\mathbf{X})} \right)$$

and using Jensen's inequality we can see that

$$\begin{aligned}\mathbb{E}_g \left( \frac{h^2(\mathbf{X})f^2(\mathbf{X})}{g^2(\mathbf{X})} \right) &\geq \left( \mathbb{E}_g \left[ \frac{|h(\mathbf{X})|f(\mathbf{X})}{g(\mathbf{X})} \right] \right)^2 \\ &= \left( \int |h(x)|f(x)dx \right)^2\end{aligned}$$

and that this lower bound is achieved by choosing  $g = g^*$ .

NB: We won't be able to calculate  $g^*$ ! But the theorem suggests that choosing  $g$  to look like  $hf$  will be a good choice.

$\frac{1}{n} \sum \tilde{w}(\mathbf{X}_i)$  is an unbiased estimator of  $c$  as

$$\mathbb{E}_g \tilde{w}(\mathbf{X}) = \int \frac{f_1(x)}{g(x)} g(x) dx = \int f_1(x) dx = c.$$

When we use unnormalised densities,  $\hat{I}$  is a biased estimator of  $I$ , however it is possible to prove that we still have  $\hat{I} \rightarrow I$  almost surely as  $n \rightarrow \infty$ .

This will be important when we use importance sampling to estimate Bayesian quantities.

## Effective sample size

How variable the weights are tells us how efficient our choice of  $g$  is. s

In the best case, where  $g = f$ , then  $\tilde{w}(X) = 1$  so that  $w_i = \frac{1}{n}$ , which is the case in plain Monte Carlo. In this case  $\text{Var}(w(X)) = 0$ .

If  $f$  and  $g$  are very different, then the weights will be very variable, and we can find that one or two particles ( $X_i$ ) dominate the sum.

We often calculate the **effective sample size**

$$ESS = \frac{1}{\sum w_i^2}$$

- ▶ In the best case,  $w_i = \frac{1}{n}$  and  $ESS = n$  - so we have an effective sample size equal to the true sample size.
- ▶ The worst case is when one of the  $w_i = 1$  and all the others are equal to zero. Then  $ESS = 1$ , i.e., we effectively have only a single sample.

We want to choose  $g$  so that the ESS is large.

## Antithetic variables - II

We need to find two estimators which are negatively correlated.

This can be done as follows:

- ▶ If  $U \sim U[0, 1]$  then  $1 - U \sim U[0, 1]$  also.
- ▶ If  $F$  is the distribution function of  $X$  then  $X_1 = F^{-1}(U)$  and  $X_2 = F^{-1}(1 - U)$  are both distributed according to  $F$
- ▶ and  $\text{Cov}(X_1, X_2) < 0$ .

**Proof (non-examinable):**

Let  $h(u) = F^{-1}(u)$ . Then  $h(u)$  is a non-decreasing function.

We need to show

$$\mathbb{E}h(U)h(1 - U) \leq (\mathbb{E}h(U))^2$$

Let  $Q = \mathbb{E}h(U)$ . The since  $h$  is non-decreasing on  $[0, 1]$

$$h(0) \leq Q \leq h(1)$$

## 3.7 Variance reduction techniques

### Antithetic variables

The method of antithetic variables uses two correlated estimators and combines them to get an estimator with a lower variance (i.e. a better estimator).

Suppose we have two different estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  of  $\theta$ ,

- ▶ with the same mean and variance
- ▶ but which are negatively correlated

Define  $\hat{\theta}_3 = \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)$ . Then

$$\begin{aligned}\text{Var}(\hat{\theta}_3) &= \frac{1}{4}(\text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2) + 2\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)) \\ &= \frac{1}{2}(\text{Var}(\hat{\theta}_1) + \text{Cov}(\hat{\theta}_1, \hat{\theta}_2)) \\ &< \frac{1}{2}\text{Var}(\hat{\theta}_1)\end{aligned}$$

This is twice the cost of computing  $\hat{\theta}_1$  but the variance is more than halved!

Let  $f(y) = \int_0^y h(1 - x)dx - Qy$  on  $[0, 1]$

Then  $f(0) = f(1) = 0$  and

$$f'(y) = h(1 - y) - Q$$

is also a non-increasing function.

Since  $f'(0) = h(1) - Q \geq 0$  and  $f'(1) = h(0) - Q \leq 0$  we must have

$$f(u) \geq 0 \text{ on } [0, 1]$$

Therefore

$$\begin{aligned}0 &\leq \int_0^1 f(y)h'(y)dy = [fh]_0^1 - \int_0^1 f'h(y)dy \\ &= - \int_0^1 f'(y)h(y)dy\end{aligned}$$

Therefore

$$\int_0^1 f'(y)h(y)dy = \int_0^1 h(y)(h(1 - y) - Q)dy = \int_0^1 h(y)h(1 - y)dy - Q^2 \leq 0$$

Hence  $\int_0^1 h(y)h(1 - y)dy \leq Q^2$  as required.

## Cauchy Example Revisited

Above we used

$$\hat{\theta}_3 = \frac{1}{2} - \frac{2}{n} \sum_{i=1}^n \left[ \frac{1}{\pi(1 + u_i^2)} \right]$$

as an estimator of  $\mathbb{P}(X > 2)$  where  $X \sim \text{Cauchy}$ .

An estimator with a smaller variance can be found using antithetic variables

$$\frac{1}{2} \left( \frac{1}{2} - \frac{2}{n} \sum_{i=1}^n \left[ \frac{1}{\pi(1 + u_i^2)} \right] + \frac{1}{2} - \frac{2}{n} \sum_{i=1}^n \left[ \frac{1}{\pi(1 + (2 - u_i)^2)} \right] \right)$$

which gives

$$\hat{\theta}_{\text{antithetic}} = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n \left[ \frac{1}{\pi(1 + u_i^2)} + \frac{1}{\pi(1 + (2 - u_i)^2)} \right]$$

The for  $n = 10$  we find the variance of  $\hat{\theta}_3$  is  $2.7 \times 10^{-4}$  whereas the variance of  $\hat{\theta}_{\text{antithetic}}$  is  $5.5 \times 10^{-6}$  - a substantial improvement.