MONTE CARLO AND MARKOV CHAIN MONTE CARLO METHODS

History: Monte Carlo (MC) and Markov Chain Monte Carlo (MCMC) have been around for a long time. Some (very) early uses of MC ideas:

- Conte de Buffon (1777) dropped a needle of length L onto a grid of parallel lines spaced $d > L$ apart to estimate $P[\text{needle intersects a line}]$.
- Laplace (1786) used Buffon's needle to evaluate π.
- Gosset (Student, 1908) used random sampling to determine the distribution of the sample correlation coefficient.
- von Neumann, Fermi, Ulam, Metropolis (1940s) used games of chance (hence MC) to study models of atomic collisions at Los Alamos during WW II.

We are concerned with the use of MC and, in particular, MCMC methods to solve estimation and detection problems.

As discussed in the introduction to MC methods (handout # 4), many estimation and detection problems require evaluation of integrals.

Monte Carlo Integration

MC Integration is essentially numerical integration and thus may be thought of as estimation — we will discuss MC estimators that estimate integrals.

Although MC integration is most useful when dealing with high-dimensional problems, the basic ideas are easiest to grasp by looking at 1-D problems first.

Suppose we wish to evaluate

$$G = \int_{\Omega} g(x)\, p(x)\, dx$$

where $p(\cdot)$ is a density, i.e. $p(x) \geq 0$ for $x \in \Omega$ and $\int_{\Omega} p(x)\, dx = 1$. Any integral can be written in this form if we can transform their limits of integration to $\Omega = (0, 1)$ — then choose $p(\cdot)$ to be uniform$(0, 1)$.

The basic MC estimate of G is obtained as follows:

1. Draw i.i.d. samples $x_1, x_2, \ldots, x_N$ from $p(\cdot)$ and

2. Estimate G as
$$\widehat{G}_N = \frac{1}{N} \sum_{i=1}^{N} g(x_i).$$

Clearly, $\widehat{G}_N$ will be such that

$$\mathrm{E}\,[\widehat{G}_N] = G \quad \text{and} \quad \widehat{G}_N \xrightarrow{\mathrm{p}} G$$

which follows by applying the law of large numbers (LLN).

Comments:

- When we get to MCMC, we will do essentially the same thing except that independence will not hold for the samples $x_1, x_2, \ldots, x_N$. We then need to rely on ergodic theorems rather than LLN.

- $\widehat{G}_N$ has rate of convergence $N^{-1/2}$ which is

  (i) slow (quadrature may have $\sim N^{-4}$ convergence)
  (ii) more efficient than quadrature or finite-difference methods when the dimensionality of integrals is large ($>$ 6–8)

Define

$$\sigma^2 = \underbrace{\int g^2(x)\,p(x)\,dx - \Big[\int g(x)\,p(x)\,dx\Big]^2}_{\mathrm{var}(G(X_i))}.$$

The overall efficiency of MC calculation is proportional to $t\,\sigma^2$ where t is the time required to sample x from $p(x)$.

Bias and Variance in MC Estimators

Assume

$$\widehat{G}_N = \frac{1}{N} \sum_{i=1}^{N} g(x_i).$$

for

$$G = \int_\Omega g(x)\, p(x)\, dx$$

where $x_i \sim$ i.i.d. with pdf $p(x)$. As before

$$\mathrm{E}\,[\widehat{G}_N] = G \quad \text{and} \quad \widehat{G}_N \xrightarrow{\mathrm{p}} G.$$

Also

$$\mathrm{E}\,[(\widehat{G}_N - G)^2] = \mathrm{E}\,[(\widehat{G}_N - \mathrm{E}\,[\widehat{G}_N])^2] + (\mathrm{E}\,[\widehat{G}_N] - G)^2$$

which is just MSE $=$ variance $+$ bias^2 (recall handout # 1). As we know from the estimation theory, it might be possible to find an estimator with smaller MSE than $\widehat{G}_N$ at the expense of being biased.

Example. Estimate the mean of uniform$(0, 1)$ using MC:

$$G = \int_0^1 x\, dx, \quad \widehat{G}_N^{(1)} = \frac{1}{N} \sum_i x_i$$

for x_i sampled from uniform$(0,1)$. Consider

$$\widehat{G}_N^{(2)} = \tfrac{1}{2} \max\{x_1, x_2, \ldots, x_N\}$$

for x_i sampled from uniform$(0,1)$.

$$\mathrm{E}\,[\widehat{G}_N^{(1)}] = G \quad \text{whereas} \quad \mathrm{E}\,[\widehat{G}_N^{(2)}] = \frac{N}{N+1} G.$$

But

$$\begin{aligned} \mathrm{MSE}(\widehat{G}_N^{(1)}) &= \mathrm{var}(\widehat{G}_N^{(1)}) = \frac{G}{6N} \quad \text{(order } N^{-1}\text{)} \\ \mathrm{MSE}(\widehat{G}_N^{(2)}) &= \frac{2G^2}{(N+1)(N+2)} \quad \text{(order } N^{-2}\text{).} \end{aligned}$$

Comment:

- Of course, we may be able to get a better (and more complicated) unbiased estimator of G than $\widehat{G}_N^{(1)}$. Yet, the biased estimator $\widehat{G}_N^{(2)}$ that performs well is really simple.

We often arrange for MC estimators to be unbiased, but this is not a necessary consequence of MC methodology.

Example. Formulate an MC estimator which samples from uniform$(0, 1)$ for the following integral:

$$G = \frac{\int_0^1 g_1(x)\, dx}{\int_0^1 g_2(x)\, dx}.$$

An estimate might be

$$\widehat{G}_N = \frac{\sum_{i=1}^N g_1(x_i)}{\sum_{i=1}^N g_2(x_i)}$$

where $x_1, x_2, \ldots, x_N \sim$ i.i.d. uniform$(0, 1)$. Here, $\widehat{G}_N$ will be biased for G although it will be consistent with $\widehat{G}_N \xrightarrow{\mathrm{p}} G$.

If $\widehat{G}_N$ is *unbiased*, we wish to reduce its variance (which is equal to the MSE in this case).

To implement MC methods we

1. must be able to generate from $p(x)$ (or from a suitable alternative),
2. must be able to do so in reasonable time (i.e. small t),
3. need small σ^2.

Monte Carlo addresses problem 3.
Markov Chain addresses problems 1 and 2.

MC Variance Reduction: Importance Sampling

We have an integral

$$G = \int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

where $\boldsymbol{x}$ is p-dimensional. $p(\boldsymbol{x})$ may not be the best distribution to sample from. Certainly, $p(\boldsymbol{x})$ is not the only distribution we could sample from: For any $\widetilde{p}(\boldsymbol{x}) > 0$ *over the same support* as $p(\boldsymbol{x})$, satisfying $\int \widetilde{p}(\boldsymbol{x})\, d\boldsymbol{x} = 1$:

$$G = \int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = \int \frac{g(\boldsymbol{x})\, p(\boldsymbol{x})}{\widetilde{p}(\boldsymbol{x})}\, \widetilde{p}(\boldsymbol{x})\, d\boldsymbol{x}$$

as long as

$$\frac{g(\boldsymbol{x})\, p(\boldsymbol{x})}{\widetilde{p}(\boldsymbol{x})} < \infty$$

up to a countable set and the original integral exists. Let $\widehat{G}_{\widetilde{p},N}$ be the MC estimator of G based on sampling from $\widetilde{p}$:

$$\widehat{G}_{\widetilde{p},N} = \frac{1}{N} \sum_i \frac{g(\boldsymbol{x}_i) p(\boldsymbol{x}_i)}{\widetilde{p}(\boldsymbol{x}_i)} \qquad (1)$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N \sim$ i.i.d. $\widetilde{p}(\boldsymbol{x})$. Clearly,

$$\mathrm{E}\,[\widehat{G}_{\widetilde{p},N}] = \mathrm{E}\,[\widehat{G}_N] = G$$

for any legitimate $\widetilde{p}$ and

$$\mathrm{var}(\widehat{G}_{\widetilde{p},N}) = \frac{1}{N}\int \Big[\frac{g(\boldsymbol{x})p(\boldsymbol{x})}{\widetilde{p}(\boldsymbol{x})}\Big]^2 \widetilde{p}(\boldsymbol{x})\, d\boldsymbol{x} - \frac{G^2}{N}.$$

The idea is to pick $\widetilde{p}$ to minimize this variance, subject to the constraint that $\widetilde{p}$ is a density.

Suppose that we pick

$$\widetilde{p}(\boldsymbol{x}) \propto |g(\boldsymbol{x})|\, p(\boldsymbol{x})$$

i.e.

$$\widetilde{p}(\boldsymbol{x}) = \frac{1}{C}\,|g(\boldsymbol{x})|\, f(\boldsymbol{x}) \quad \text{where} \quad C = \int |g(\boldsymbol{x})|\, p(\boldsymbol{x})\, d\boldsymbol{x}.$$

This would give

$$\mathrm{var}(\widehat{G}_{\widetilde{p},N}) = \frac{1}{N}\int C\,|g(\boldsymbol{x})|\, f(\boldsymbol{x})\, d\boldsymbol{x} - \frac{G^2}{N} = \frac{1}{N}\,(C^2 - G^2).$$

Note that
if $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$ in the support of p, then

$$C = \int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = G$$

and if $g(\boldsymbol{x}) < 0$ for all $\boldsymbol{x}$ in the support of p, then

$$C = -\int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = -G.$$

In both cases $\mathrm{var}(\widehat{G}_{\widetilde{p},N}) = 0$!

However, if we could compute $\int |g(\boldsymbol{x})|\, p(\boldsymbol{x})\, d\boldsymbol{x}$, we would already be done and would not use MC. Nevertheless, we can draw some conclusions from the above exercise:

1. There may exist an importance pdf $\widetilde{p}(\boldsymbol{x})$ that gives smaller variance than $p(\boldsymbol{x})$ (of the corresponding estimate of G in (1)).

2. Good $\widetilde{p}(\boldsymbol{x})$ are those that match the behavior of $g(\boldsymbol{x})\, p(\boldsymbol{x})$. This is particularly true near the maximum of the integrand $g(\boldsymbol{x})\, p(\boldsymbol{x})$.

The distribution $\widetilde{p}(\boldsymbol{x})$ we actually sample from is usually called the *importance function* or importance density.

How can we *actually* find useful importance densities?

Example 1: Find an MC estimator of

$$G = \int_0^1 (1 - x^2)^{1/2}\, dx.$$

Obviously, we could draw i.i.d. samples $x_1, x_2, \ldots, x_N$ from uniform$(0, 1)$ and use

$$\widehat{G}_N = \frac{1}{N}\sum_{i=1}^{N}(1 - x_i^2)^{1/2}.$$

It turns out that in this case

$$\mathrm{var}(\widehat{G}_N) = \frac{0.050}{N}.$$

Now, consider finding a good importance density. Observe that $(1 - x^2)^{1/2}$ has a maximum at $x = 0$ on $[0, 1]$. To find a function that looks like $(1-x^2)^{1/2}$ near $x = 0$, we could expand $(1 - x^2)^{1/2}$ in a Taylor series around zero (i.e. the maximum).

In this example, let $g(x) = (1 - x^2)^{1/2}$ and $p(x) = 1$
$\Longrightarrow g(x)p(x) = g(x)$ and

$$g(0) = 1, \quad g'(0) = 0, \quad g''(0) = -1$$

implying that

$$g(x) \approx 1 - \tfrac{1}{2}\,x^2, \quad \int_0^1 (1 - \tfrac{1}{2}\,x^2)\,dx = 1 - 1/6 = 5/6.$$

So, we could form

$$\widetilde{p}(x) = \frac{6}{5} \cdot (1 - \tfrac{1}{2}x^2), \quad 0 < x < 1.$$

If we sample $x_1, x_2, \ldots, x_N$ i.i.d. from this $\widetilde{p}(\cdot)$ and use this MC estimator:

$$\widehat{G}_{\widetilde{p},N} = \frac{1}{N} \sum_{i=1}^{N} \frac{5}{6} \cdot \frac{(1 - x_i^2)^{1/2}}{1 - \frac{1}{2} x_i^2}$$

it turns out that $\mathrm{var}(\widehat{G}_{\widetilde{p},N}) = 0.011/N$, about $1/5$ of $\mathrm{var}(\widehat{G}_N)$.

In general, this is not a great enough reduction to make finding $\widetilde{p}(\cdot)$ worth the effort. But, suppose now that we generalize $1 - \frac{1}{2}x^2$ to $1 - \beta x^2$ so that

$$\int_0^1 (1 - \beta x^2)\, dx = 1 - \tfrac{1}{3}\beta \quad \Longrightarrow \quad \widetilde{p}(x) = \frac{1 - \beta x^2}{1 - \beta/3}, \; 0 < x < 1.$$

We now look for β that minimizes the variance of

$$\frac{(1 - x^2)^{1/2}(1 - \beta/3)}{1 - \beta x^2}$$

where x is drawn from $\widetilde{p}(x) = (1 - \beta x^2)/(1 - \beta/3)$, $0 < x < 1$. It turns out that $\beta = 0.74$ minimizes this variance and the end result is

$$\mathrm{var}(\widehat{G}_{\widetilde{p},N}) = \frac{0.0029}{N}$$

which is a significant (by an order of magnitude) improvement compared with $\mathrm{var}(\widehat{G}_N)$.

Example 2: Estimate $P[2 < X]$ for $X \sim$ Cauchy:

$$G = \int_2^\infty [\pi(1+x^2)]^{-1} dx.$$

Note that, for $2 < x$, $1/(1+x^2)$ looks much like $1/x^2$ and

$$\int_2^\infty \frac{1}{x^2} dx = \tfrac{1}{2}.$$

Then, we might try

$$\widetilde{p}(x) = \frac{2}{x^2}, \quad 2 < x$$

and estimate G using

$$\widehat{G}_{\widetilde{p},N} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i^2}{2\pi(1+x_i^2)}$$

where $x_1, x_2, \ldots, x_N$ are i.i.d. $\widetilde{p}(\cdot)$. Note that for X following $\widetilde{p}(x) = \frac{2}{x^2}, \quad 2 < x$, we have

$$Y = \frac{2}{X} \sim \text{uniform}(0,1)$$

$\Longrightarrow$ we could just sample $y_1, y_2, \ldots, y_N$ from uniform$(0,1)$ and take

$$x_i = \frac{2}{y_i}, \quad i = 1, 2, \ldots, N.$$

In this example, variance reduction was not our major problem — we focused on finding a density that is easy to sample from.

Example 3: Find an MC estimator of

$$G = \int_0^1 x^{-1/2}(1-x)^{-1/2}\,dx.$$

Note that the integrand is the kernel of a beta density with parameters 1/2 and 1/2, see the table of distributions handed out in class. The integrand has singularities at the endpoints and a "standard" MC estimator based on sampling from uniform$(0, 1)$ will have infinite variance!

Define

$$g(x) = \frac{1}{x^{1/2}\,(1-x)^{1/2}}$$

and observe that it has singularities at $x = 0$ and $x = 1$. Note that

$$g(x) \approx \begin{cases} 1/x^{1/2} & \text{for } x \text{ near } 0 \\ 1/(1-x)^{1/2} & \text{for } x \text{ near } 1 \end{cases}$$

and choose

$$\widetilde{p}(x) \propto \frac{1}{x^{1/2}} + \frac{1}{(1-x)^{1/2}}$$

which is the sum of the approximating functions near the singularities. Now

$$\int_0^1 \Big[\frac{1}{x^{1/2}} + \frac{1}{(1-x)^{1/2}}\Big]\,dx = 4 \Longrightarrow \widetilde{p}(x) = \frac{1}{4x^{1/2}} + \frac{1}{4(1-x)^{1/2}}$$

yielding

$$\widehat{G}_{\widetilde{p},N} = \frac{1}{N} \sum_{i=1}^{N} \frac{4}{x_i^{1/2} + (1 - x_i)^{1/2}}.$$

All terms in the above expression are bounded by 4 so $\mathrm{var}(\widehat{G}_{\widetilde{p},N})$ exists!

The above three examples illustrate importance sampling for

- variance reduction (Example 1),
- finding a density to sample from (Example 2), and
- producing bounded $g(x)p(x)/\widetilde{p}(x)$ even if $g(x)p(x)$ is not (Example 3).

Summary of Importance Sampling

We wish to estimate

$$G = \int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}.$$

Suppose that we have an "importance" density $\widetilde{p}(\boldsymbol{x})$ such that it approximates $|g(\boldsymbol{x})|\, p(\boldsymbol{x})$ well, the support of $\widetilde{p}(\boldsymbol{x})$ covers that of $p(\boldsymbol{x})$, and it is easy to sample from $\widetilde{p}(\boldsymbol{x})$. Then, the observation that

$$G = \int g(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = \int \frac{g(\boldsymbol{x})\, p(\boldsymbol{x})}{\widetilde{p}(\boldsymbol{x})} \widetilde{p}(\boldsymbol{x})\, d\boldsymbol{x}$$

suggests the following algorithm:

1) Draw

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N \overset{\text{i.i.d.}}{\sim} \widetilde{p}(\boldsymbol{x})$$

where $\widetilde{p}(\boldsymbol{x}) \equiv$ the importance distribution.

2) Compute the importance weights

$$\phi_n = \frac{p(\boldsymbol{x}_n)}{\widetilde{p}(\boldsymbol{x}_n)}, \quad n = 1, 2, \ldots, N.$$

3) Approximate G by either of the following estimators:

$$\widehat{G}_{\mathrm{WE}} = \frac{\sum_{n=1}^{N} g(\boldsymbol{x}_n)\phi_n}{\sum_{n=1}^{N} \phi_n} \quad \text{(weighted average)}$$

or

$$\widehat{G} = \sum_{n=1}^{N} g(\boldsymbol{x}_n)\phi_n \quad \text{(simple average, discussed earlier).}$$

Note that $\widehat{G}$ is an unbiased estimator of G whereas $\widehat{G}_{\mathrm{WE}}$ is asympotitically consistent (not necessarily unbiased).

Why is $\widehat{G}_{\mathrm{WE}}$ asympotitically consistent? Because

$$\frac{1}{N}\sum_{n=1}^{N} g(\boldsymbol{x}_n)\,\phi_n \xrightarrow{p} \mathrm{E}_{\widetilde{p}}[g(\boldsymbol{X})\phi(\boldsymbol{X})] = G$$

and

$$\frac{1}{N}\sum_{n=1}^{N} \phi_n \xrightarrow{p} \mathrm{E}_{\widetilde{p}}[\phi(\boldsymbol{X})] = 1.$$

Although $\widehat{G}_{\mathrm{WE}}$ is biased, it often has smaller mean-square error than $\widehat{G}$.

Since both numerator and denominator of $\widehat{G}_{\mathrm{WE}}$ involve ϕ_n, we only need to know

$$\phi(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\widetilde{p}(\boldsymbol{x})}$$

up to a proportionality constant! Therefore, $\widehat{G}_{\mathrm{WE}}$ is great for Bayesian computations where $p(\boldsymbol{x})$ is often known only up to a proportionality constant.

Sampling–Importance Resampling

Here, we use the importance sampling idea in a slightly different context.

Suppose we have N draws $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ from a proposal distribution $\widetilde{p}(\boldsymbol{x})$. Can we convert these samples to samples from a desired distribution $p(\boldsymbol{x})$?

A *sampling–importance resampling method* for this conversion:

- For each $\boldsymbol{x}_n$, compute

$$\begin{aligned} \phi_n &= \frac{p(\boldsymbol{x}_n)}{\widetilde{p}(\boldsymbol{x})} \\ w_n &= \frac{\phi_n}{\sum_{k=1}^N \phi_k}. \end{aligned}$$

- Draw $\boldsymbol{x}_\star$ from the discrete distribution over $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ with weight w_n on θ_n.

If we need multiple samples $\boldsymbol{x}_\star$, it is suggested to draw them without replacement (which, of course, makes sense only if the number of samples to be drawn is a few times smaller than N). If the number of samples to be drawn is N, then sample with replacement. The resampled $\boldsymbol{x}_\star$ are drawn approximately from $p(\boldsymbol{x})$.

Proof. For simplicity, we focus on univariate x. Then

$$\begin{aligned}
P\{x_\star \le a\} &= \overbrace{\sum_{n=1}^{N} w_n \, i_{-\infty,a}(x_n)}^{\text{empirical cdf}} \\
&= \frac{N^{-1}\sum_{n=1}^{N} \phi_n \, i_{-\infty,a}(x_n)}{N^{-1}\sum_{n=1}^{N} \phi_n} \\
&\longrightarrow \frac{\mathrm{E}_{\widetilde{p}}[\frac{p(X)}{\widetilde{p}(X)} \, i_{-\infty,a}(X)]}{\mathrm{E}_{\widetilde{p}}[\frac{p(X)}{\widetilde{p}(X)}]} \\
&= \frac{\int_{-\infty}^{a} p(x)\,dx}{\int_{-\infty}^{\infty} p(x)\,dx} = \int_{-\infty}^{a} p(x)\,dx.
\end{aligned}$$

□

MC Variance Reduction: Rao-Blackwellization

Suppose that we have drawn i.i.d. samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ from $p(\boldsymbol{X})$ and wish to estimate

$$I = \mathrm{E}_{p(\boldsymbol{x})}[g(\boldsymbol{X})].$$

Recall the straightforward estimator that we mentioned before:

$$\widehat{G}_N = \frac{1}{N} \cdot [g(\boldsymbol{x}_1) + g(\boldsymbol{x}_2) + \cdots g(\boldsymbol{x}_N)]$$

known as the *simple average estimator* or *histogram estimator*.

Suppose that the random vector $\boldsymbol{X}$ can be divided into two blocks:

$$\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$$

where we can compute

$$\mathrm{E}_{p(\boldsymbol{x} \,|\, \boldsymbol{x}^{(2)})}[g(\boldsymbol{X}) \,|\, \boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]$$

analytically. Since the law of iterated expectations states

$$\mathrm{E}_{p(\boldsymbol{x}^{(2)})}[\mathrm{E}_{p(\boldsymbol{x} \,|\, \boldsymbol{x}^{(2)})}[g(\boldsymbol{X}) \,|\, \boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]] = \mathrm{E}_{p(\boldsymbol{x})}[g(\boldsymbol{X})]$$

or, more informally, after removing the annoying subscripts:

$$\mathrm{E}\,\{\mathrm{E}\,[g(\boldsymbol{X}) \,|\, \boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]\} = \mathrm{E}\,[g(\boldsymbol{X})]$$

the following *Rao-Blackwellized estimator of G*

$$\begin{aligned}\widetilde{G}_N &= \frac{1}{N} \cdot \Big\{ \mathrm{E}_{\,p(\boldsymbol{x} \,|\, \boldsymbol{x}^{(2)})}[g(\boldsymbol{x}_1)] + \mathrm{E}_{\,p(\boldsymbol{x} \,|\, \boldsymbol{x}^{(2)})}[g(\boldsymbol{x}_2)] \\ &\quad + \cdots \mathrm{E}_{\,p(\boldsymbol{x} \,|\, \boldsymbol{x}^{(2)})}[g(\boldsymbol{x}_N)] \Big\}\end{aligned}$$

is unbiased. The above estimator is also known as the *mixture estimator*.

Which one of the above two MC estimators is better? Since

$$\begin{aligned}\mathrm{var}[g(\boldsymbol{X})] &= \mathrm{E}\,\{\mathrm{var}[g(\boldsymbol{X})|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]\} \\ &\quad +\mathrm{var}\{\mathrm{E}\,[g(\boldsymbol{X})|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]\}\end{aligned}$$

we have

$$\mathrm{var}[\widehat{G}_N] = \frac{\mathrm{var}[g(\boldsymbol{X})]}{N} \geq \frac{\mathrm{var}\{\mathrm{E}\,[g(\boldsymbol{X}) \,|\, \boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}]\}}{N} = \mathrm{var}[\widetilde{G}_N].$$

Of course, the computational effort for obtaining the two estimates should also be taken into account when deciding which one is "better."

A basic rule in MC computation: One should carry out analytical computations as much as possible!

Comments:

- We can use the mixture estimator for density estimation as well: in particular, if we can compute $p_{\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}^{(2)})$ analytically, then

$$\begin{aligned} p_{\boldsymbol{x}^{(1)}}(\boldsymbol{x}^{(1)}) \; &\approx \frac{1}{N}\cdot\Big[p_{\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}_1^{(2)}) + p_{\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}_2^{(2)}) \\ &\quad +p_{\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)}\,|\,\boldsymbol{x}_N^{(2)})\Big]. \end{aligned}$$

- If the samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, then it is not so straightforward to claim superiority of Rao-Blackwellization, but it has been shown for Gibbs sampler.

ELEMENTARY TECHNIQUES FOR RANDOM SAMPLING

Our concern now is “sampling” observations from a given distribution using observations from another distribution [which is often uniform$(0, 1)$].

Inversion

Probability Integral Transform: If x is a continuous random variable with cdf F, then

$$F(x) \sim \text{uniform}(0, 1).$$

To use this idea for sampling from both continuous and discrete distributions, we extend the probability integral transform as follows.

Theorem 1. *Assume that we wish to sample a random variable x with cdf $F(x)$. Define*

$$x(u) = \min\{\widetilde{x} \,:\, F(\widetilde{x}) \geq u\}$$

where $u \sim \text{uniform}(0, 1)$. Then

$$x(u) \sim F.$$

Proof. Ripley, p. 59. □

For continuous x and known F^{-1}:

1. Sample $u_i \sim$ i.i.d. uniform$(0, 1)$,
2. $x_i = F^{-1}(u_i)$.

Then $x_1, x_2, \ldots, x_N \sim$ i.i.d. F.

Example: Sample Poisson(4):

$$p(x) = \frac{1}{x!} 4^x e^{-4}$$

So, if we generate $u_1 = 0.324$ from uniform$(0, 1)$, then $x_1 = 3$.

x	$p(x)$	cdf $F(x) = P[X \leq x]$
0	0.018316	0.018316
1	0.073262	0.091578
2	0.146525	0.238103
3	0.195367	0.433470
$\cdots$	$\cdots$	$\cdots$

Some Useful Inversion Formulas

pdf: $\mathbf{p}(x)$	cdf: $F(x)$	$X = F^{-1}(U)$	simplified form
Exponential(λ)			
$\lambda e^{\lambda x},\ x \geq 0$	$1 - e^{-\lambda x}$	$-\frac{1}{\lambda}\log(1-U)$	$-\frac{1}{\lambda}\log(U)$
Laplace(λ)			
$\frac{\lambda}{2}e^{\lambda x},\ x \in \mathbb{R}^1$	$\frac{1}{2}e^{\lambda x}$, if $x \leq 0$ $1 - \frac{1}{2}e^{-\lambda x}$, if $x > 0$	$\frac{1}{\lambda}\log(2U)$, if $U \leq \frac{1}{2}$ $-\frac{1}{\lambda}\log(2(1-U))$, if $U > \frac{1}{2}$	$\frac{1}{\lambda}\log(2U)$ $-\frac{1}{\lambda}\log(2U)$
Cauchy(θ)			
$\frac{\theta}{\pi(x^2+\theta^2)},\ x \in \mathbb{R}^1$	$\frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x}{\theta}\right)$	$\theta\tan\left(\pi\left(U - \frac{1}{2}\right)\right)$	$\theta\tan(\pi U)$
Rayleigh(σ)			
$\frac{x}{\sigma}e^{-\frac{x^2}{2\sigma^2}},\ x \geq 0$	$1 - e^{-\frac{x^2}{2\sigma^2}}$	$\sigma\sqrt{-\log(1-U)}$	$\sigma\sqrt{-\log(U)}$
Pareto$(\mathbf{a}, \mathbf{b})$			
$\frac{ab^a}{x^{a+1}},\ x \geq b > 0$	$1 - \left(\frac{b}{x}\right)^a$	$\frac{b}{(1-U)^{1/a}}$	$\frac{b}{U^{1/a}}$

Composition of Random Variables

The basic idea: be clever in manipulating functions that describe distributions. We now present one technique to illustrate the flavor of this approach. Consider sampling from a distribution that has the following form:

$$p(x) = \sum_{i=1}^{K} \alpha_i \, g_i(x), \quad \alpha_i > 0, \quad g_i(x) \geq 0.$$

Note that we do not require $\int g_i(x)\, dx = 1$.

In our original problem, $g_i(x)$ may not be densities. However, we can write

$$p(x) = \sum_{i=1}^{K} \underbrace{\alpha_i \int g_i(x)\, dx}_{\beta_i} \cdot \underbrace{\frac{g_i(x)}{\int g_i(x)\, dx}}_{h_i(x)}.$$

Note that we need $\sum \beta_i = 1$, but this should hold if $p(x)$ is a valid distribution.

We can find functions $h_i(x)$ and coefficients β_i such that $h_i(x) \geq 0, \beta_i > 0, \sum \beta_i = 1$, and

$$\int h_i(x)\, dx = 1 \quad \text{(i.e. densities).}$$

Sampling Scheme:

- We first sample a value m from the set $\{1, 2, \ldots, K\}$ with probabilities $\beta_1, \beta_2, \ldots, \beta_K$;
- Then, we take one observation from $h_m(x)$.

Repeating the above scheme N times, we end up with $x_1, x_2, \ldots, x_N$ from

$$h(x) = \sum_i \beta_i \, h_i(x) \qquad \text{(a finite mixture).}$$

Example: We wish to sample from

$$p(x) = \tfrac{3}{5} + \tfrac{3}{5}\, x + \tfrac{3}{10}\, x^2, \quad 0 < x < 1.$$

Clearly, the above $p(x)$ can be written as $\sum_i \alpha_i \, g_i(x)$. Over the interval $(0, 1)$, we know that $h_1(x) = 1, h_2(x) = 2x$ and $h_3(x) = 3x^2$ are densities; hence

$$p(x) = (\underbrace{\tfrac{3}{5}}_{\beta_1} \cdot \underbrace{1}_{h_1(x)}) + (\underbrace{\tfrac{3}{10}}_{\beta_2} \cdot \underbrace{2x}_{h_2(x)}) + (\underbrace{\tfrac{1}{10}}_{\beta_3} \cdot \underbrace{3x^2}_{h_3(x)}) \ \text{ form } \textstyle\sum_i \beta_i \, h_i(x).$$

Now, choose $k = 1$ with probability $\beta_1 = \frac{3}{5}$, $k = 2$ with probability $\beta_2 = \frac{3}{10}$, and $k = 3$ with probability $\beta_3 = \frac{1}{10}$ or

$$p(k) = \begin{cases} 6/10, & i = 1 \\ 3/10, & i = 2 \\ 1/10, & i = 3 \end{cases}.$$

We can use simple inversion to simulate samples from the above pmf.

Entire algorithm:

1. Generate $u_i \sim$ uniform$(0, 1)$.
2. - If $u_i < 6/10$ set $k = 1$;
   - If $6/10 \leq u_i < 9/10$ set $k = 2$;
   - If $9/10 \leq u_i$ set $k = 3$.
3. - If $k = 1$, set $x_i = w_1$ where $w_1 \sim$uniform$(0, 1)$;
   - If $k = 2$, set $x_i = \max\{w_1, w_2\}$
     where $w_1, w_2 \sim$ i.i.d. uniform$(0, 1)$;
   - If $k = 3$, set $x_i = \max\{w_1, w_2, w_3\}$
     where $w_1, w_2, w_3 \sim$ i.i.d. uniform$(0, 1)$.

Note:

$$\begin{aligned} P[\max\{w_1, w_2\} < x] &= x^2, \quad \text{for } x \geq 0 \\ P[\max\{w_1, w_2, w_3\} < x] &= x^3, \quad \text{for } x \geq 0. \end{aligned}$$

Grid Approach

Suppose that we wish to sample from $p(x)$. Here is how we can (approximately) accomplish that:

- Make a grid of values of x spanning the support of $p(x)$:

  $$x_1, x_2, \ldots, x_m.$$

  For convenience, define also $x_0 = -\infty$ and $x_{m+1} = +\infty$.

- Evaluate

  $$p(x_1), p(x_2), \ldots, p(x_m).$$

- Estimate the cdf of this distribution as follows:

  $$\underbrace{0}_{\text{at } x_0}, \underbrace{0 + p(x_1)}_{\text{at } x_1}, \underbrace{0 + p(x_1) + p(x_2)}_{\text{at } x_2}, \ldots, \underbrace{\sum_{i=1}^{m} p(x_i)}_{\text{at } x_m}, \underbrace{1}_{\text{at } x_{m+1}} .$$

- Generate a uniform$(0, 1)$ random variable u.

- If $u \in [\sum_{i=1}^{l-1} p(x_i), \sum_{i=1}^{l} p(x_i)]$, draw x_l.

Comments:

- **Nice properties:**
  - It is easy to implement and understand (no tricks).
- **Disadvantages:**
  - Draws are from an approximation to the true distribution.
  - It is not clear how many grid points m shoyld be chosen. If m is too small this will be a poor approximation. But how much is "too small"? Clearly, small is "good" in terms of computational complexity.
  - Most values of a continuous random variable can not be generated by this scheme.

This approach has a larger educational (perhaps debugging?) than practical value.

Basic Rejection Sampling

A powerful technique that allows us to sample from a distribution known only up to a constant. It is due to von Neumann, see

J. von Neumann, "Various techniques used in connection with random digits," in *John von Neumann, Collected Works,* vol. V, A.H. Taub (Ed.), New York: Pergamon, 1961, pp. 768–770.

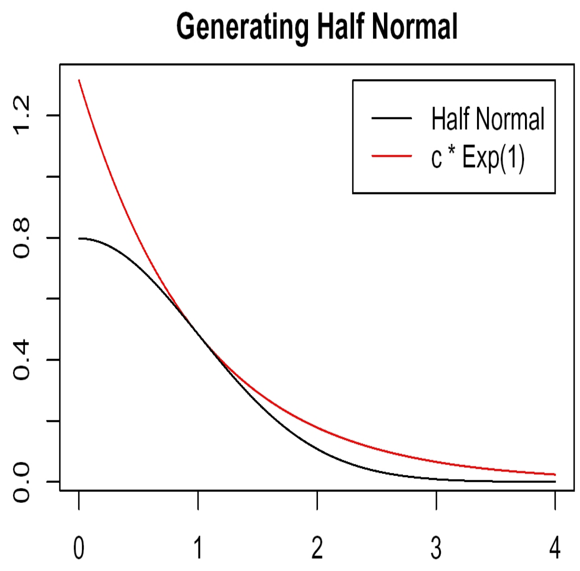
Also known as acceptance-rejection algorithm.

Say we wish to simulate from a distribution with density $p(x)$. In fact, rejection sampling "works" fine with discrete random variables and with random vectors (at least in principle — computational efficiency is important). Here, we focus on one-dimensional continuos random variables.

To implement this method, we need to find a dominating or "majorizing" pdf $\widetilde{p}(x)$ where

- $\widetilde{p}(\cdot)$ is easy to sample from and
- $$p(x) \leq m\,\widetilde{p}(x) = h(x)$$

  for all x and some constant $m > 1$.



Rejection Sampling Scheme:

1. Draw a proposal x from $\widetilde{p}(x)$ and compute the acceptance ratio:

$$r(x) = \frac{p(x)}{m\,\widetilde{p}(x)} = \frac{p(x)}{h(x)} \leq 1.$$

2. Sample $u \sim \text{uniform}(0,1)$.

   - If $u \leq r(x)$, accept the draw and return x;
   - If $u > r(x)$, reject the draw and go back to 1 (and continue the loop).

   Note that this step is equivalent to flipping a *biased coin* with success probability $r(x)$.

Then, the sample obtained using the above procedure is a draw from $p(x)$.

Proof. Let I be the indicator of whether a sample x is accepted. Then

$$P\{\underbrace{I=1}_{\text{sample } x \text{ accepted}}\} = \int P\{I=1|X=x\}\,\widetilde{p}(x)\,dx$$

$$= \int r(x)\,\widetilde{p}(x)\,dx$$

$$= \int \frac{p(x)}{m\,\widetilde{p}(x)}\,\widetilde{p}(x)\,dx = \frac{1}{m}.$$

Next, we prove the desired result:

$$p(x|I=1) = \frac{p(x)}{m\,\widetilde{p}(x)}\cdot\widetilde{p}(x)\Big/P\{I=1\} = \frac{p(x)}{m}\cdot m = p(x). \quad (2)$$

□

The acceptance probability is $1/m$ and, clearly, $m \geq 1$. The number of trials until accepting a draw is a geometric random variable, geometric$(1/m)$; hence, the average number of trials until acceptance is the mean of this geometric random variable, which is

$$\frac{1}{1/m} = m.$$

One consequence of this result is that m should be made as small as possible to minimize the number of rejections. The optimal m is given by

$$m = \sup \frac{p(x)}{\widetilde{p}(x)}.$$

Note that we do not need to find the best m, just one that satisfies

$$p(x) \leq m\, \widetilde{p}(x) = h(x)$$

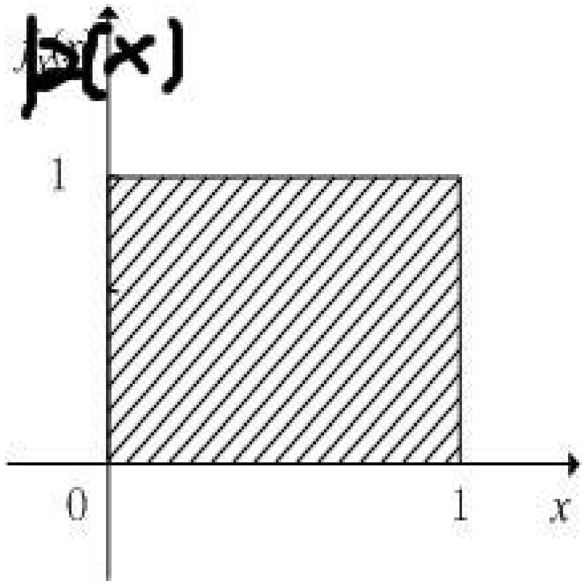
for all x.

Comments:

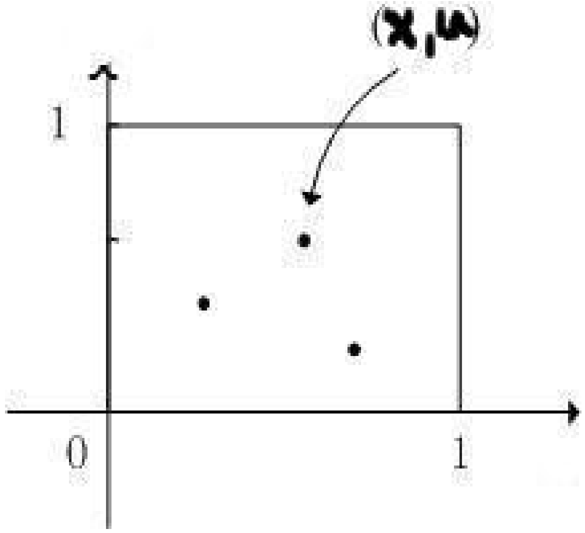
- The key to rejection sampling is finding $\widetilde{p}(\cdot)$ with *correct support* such that $\widetilde{p}(\cdot)$ is *easy to sample from* and $P[\text{accept } y]$ is *high*.
  - Intuitively, $\widetilde{p}(\cdot)$ needs to have *thicker tails* than $p(\cdot)$ for $p(x)/\widetilde{p}(x)$ to remain bounded for all y. [For example, we cannot use a Gaussian pdf $\widetilde{p}(\cdot)$ to generate samples from a Cauchy pdf $p(\cdot)$. We can do the opposite, however.]
- Rejection sampling is self-monitoring — if the method is not working efficiently, very few draws will be accepted.

Example 1 (trivial). $x \sim \text{uniform}(0, 1)$:

$$p(x) = \begin{cases} 1, & 0 < x < 1, \\ 0 & \text{otherwise} \end{cases}.$$



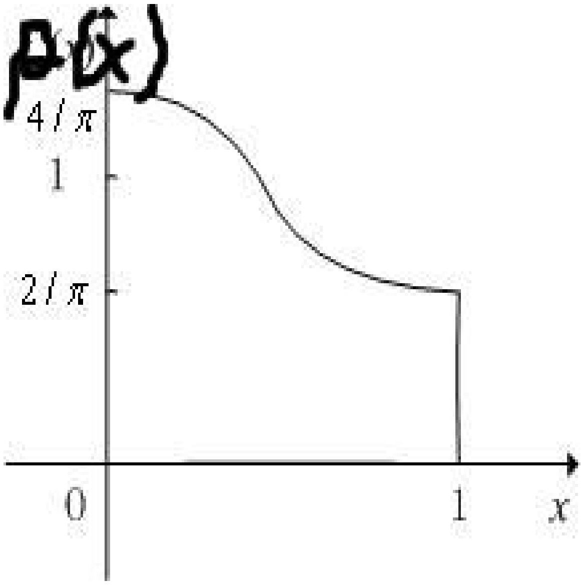
We could sample by generating two uniform$(0,1)$ variables x and u, which we can visualize as



Accept x whenever $u \leq 1$. So, the candidate value is x, $m = 1$, $p(x) = \text{uniform}(0,1)$, $\widetilde{p}(x) = 1$:

$$r(x) = \frac{p(x)}{m\,\widetilde{p}(x)} = 1.$$

Example:



$$p(x) = \frac{4}{\pi\,(1 + x^2)}, \quad 0 < x < 1.$$

Here, we can still use $\widetilde{p}(x) = 1$, $0 < x < 1$ because it has the correct support.

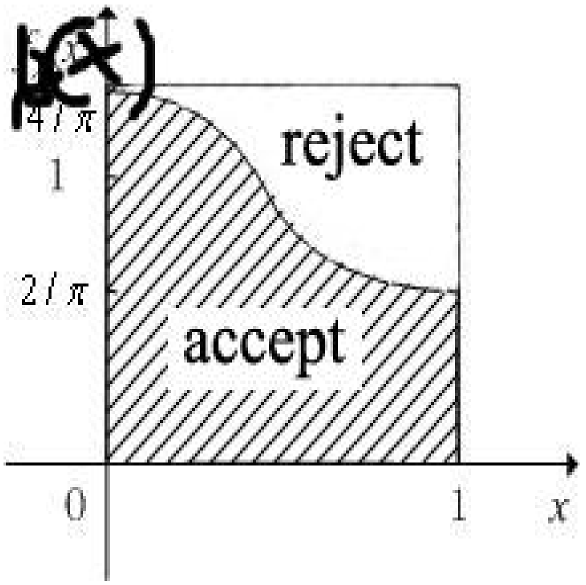
Note: $m = 1$ (say) will not give the correct rejection region. Recall that we need

$$p(x) \leq m\,\widetilde{p}(x)$$

which, in this case, will be satisfied if

$$m = \frac{4}{\pi}.$$

We sample uniformly over the region $[0, 1] \times [0, 4/\pi]$ and reject values in the upper right corner:

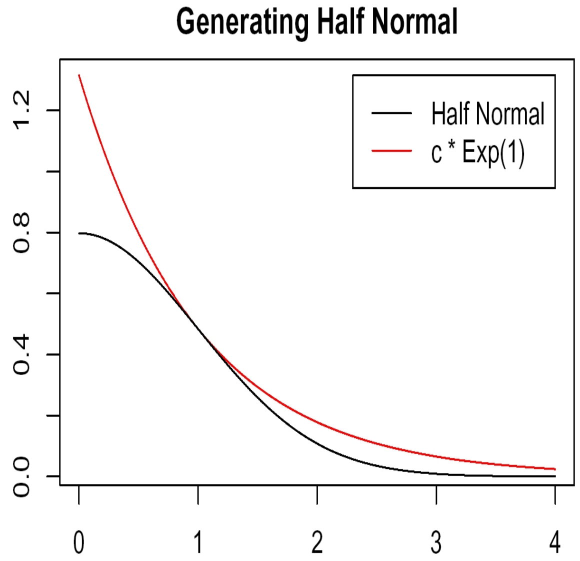


For this example, we can write the general algorithm as follows:

1. Draw x from $\underbrace{\text{uniform}\,(0,1)}_{\widetilde{p}(y)}$.

2. Draw u from uniform$(0,1)$.

3. If
$$u \leq r(x) = \frac{p(x)}{m\,\widetilde{p}(x)} = \frac{1}{1+x^2}$$
then return x. Otherwise go back to 1 and continue the loop.



Example:

We wish to sample from a half-normal pdf with mean and variance parameters 0 and 1:

$$p(x) = \sqrt{2/\pi} \cdot \exp(-x^2/2) \cdot i_{[0,\infty)}(x).$$

One possibility: draw $x \sim \mathcal{N}(0,1)$ and reject all $x < 0 \Longrightarrow$ inefficient since 50% of the draws end up being rejected.

Consider using the exponential(1) proposal distribution:

$$\widetilde{p}(x) = \exp(-x)\, i_{[0,\infty)}(x)$$

We need to find m such that

$$p(x) \leq m\,\widetilde{p}(x) \geq \quad \Longleftrightarrow \quad \sqrt{2/\pi} \cdot \exp(-x^2/2) \leq \exp(-x).$$

Differentiate the log of the above expression and find the smallest m such that the above inequality is satisfied. As a result, we obtain $x = 1$ and the bound

$$m = \sqrt{\frac{2}{\pi}} \cdot \exp(\tfrac{1}{2}) \approx 1.315$$

which is the optimal m that we can choose in this case, leading to the acceptance rate of 76%.

To summarize, here is the rejection sampler for this case:

1. Draw x from the exponential(1) pdf and compute the acceptance ratio:

$$r(x) = \frac{p(x)}{m\,\widetilde{p}(x)} = \exp[-0.5(x-1)^2].$$

2. Sample $u \sim \mathrm{uniform}(0, 1)$.
   - If $u \leq r(x)$, accept the draw and return x;
   - If $u > r(x)$, reject the draw and go back to 1 (and continue the loop).

Note that this example is somewhat artificial since we do not need rejection to sample from a half normal distribution. Half normal distribution is the distribution of the absolute value of a standard normal random variable.

In the previous discussion, it was assumed that $p(x)$ was a density function. In fact, $p(x)$ only needs to be known up to a *multiplicative constant*:

$$l(x) = b\,p(x).$$

Here, b is the multiplicative constant that may be unknown. This case is particularly common in Bayesian inference where the posterior density is usually known up to a proportionality factor:

$$p(\theta|\mathrm{data}) \propto p(\mathrm{data}|\theta)\,\pi(\theta)$$

and the normalizing constant is difficult to calculate exactly.

Rejection sampling approach *does not* require knowing the constant b! The original procedure can be modified as follows:

Find $\widetilde{m}$ such that

$$l(x) \le \widetilde{m}\,\widetilde{p}(x) = h(x)$$

for all x and some positive constant $\widetilde{m}$. Then, the rejection sampling scheme is:

1. Draw a proposal x from $\widetilde{p}(x)$ and compute the acceptance ratio:

$$r(x) = \frac{l(x)}{\widetilde{m}\,\widetilde{p}(x)} = \frac{l(x)}{h(x)} \le 1.$$

2. Sample $u \sim \mathrm{uniform}(0, 1)$.

   - If $u \le r(x)$, accept the proposal and return x;
   - If $u > r(x)$, reject the proposal and go back to 1 (and continue the loop).

Everything is the same as before, except the unnormalized density $l(x)$ is used instead of the normalized density $p(x)$.

The acceptance probability for this scheme is $b/\widetilde{m}$.

A general comment about the choice of $\widetilde{p}(\cdot)$:

A good choice $\widetilde{p}(x)$ will normally be "close to" $p(x)$ — we wish to minimize the separation between the two densities. Often, a parametric family of candidates $\widetilde{p}(\cdot)$ is chosen and the member from the parametric family with the smallest m or $\widetilde{m}$ is determined and used.

A Trick for Finding m or $\widetilde{m}$ when Sampling from a Posterior Distribution

Say we want to sample $\boldsymbol{\theta}$s from

$$\overbrace{p(\boldsymbol{\theta}|\boldsymbol{x})}^{p(\cdot)} \propto \overbrace{\pi(\boldsymbol{\theta})\, p(\boldsymbol{x}|\boldsymbol{\theta})}^{l(\cdot)}.$$

by generating samples from the prior $\overbrace{\pi(\boldsymbol{\theta})}^{\widetilde{p}(\boldsymbol{\theta})}$ and rejecting some of them. Hence, our basic rejection algorithm is

1. Generate $\boldsymbol{\vartheta}$ from $\overbrace{\pi(\boldsymbol{\vartheta})}^{\text{proposal dist.}}$.

2. Generate u from uniform$(0, 1)$.

3. Repeat Steps 1 and 2 until

$$u \le r(\boldsymbol{\vartheta}) = \frac{p(\boldsymbol{x}|\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})}{\widetilde{m}\,\pi(\boldsymbol{\vartheta})} = \frac{p(\boldsymbol{x}|\boldsymbol{\vartheta})}{\widetilde{m}}.$$

4. Return $\boldsymbol{\theta} = \boldsymbol{\vartheta}$.

Here, we need to find $\widetilde{m}$ such that

$$p(\boldsymbol{x}|\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta}) \le \widetilde{m}\,\pi(\boldsymbol{\vartheta}) \quad \Longleftrightarrow \quad p(\boldsymbol{x}|\boldsymbol{\vartheta}) \le \widetilde{m}$$

for all $\boldsymbol{\vartheta}$. Choose $\widetilde{m} = \max_{\boldsymbol{\vartheta}} p(\boldsymbol{x} | \boldsymbol{\vartheta})$, which is the *maximized likelihood*!

This yields the following rejection-sampling scheme:

1. Generate $\boldsymbol{\vartheta}$ from $\pi(\boldsymbol{\vartheta})$.
2. Generate u from uniform$(0, 1)$.
3. Repeat Steps 1 and 2 until

$$u \leq \frac{p(\boldsymbol{x} | \boldsymbol{\vartheta})}{\max_{\boldsymbol{\vartheta}} p(\boldsymbol{x} | \boldsymbol{\vartheta})}.$$

4. Return $\boldsymbol{\theta} = \boldsymbol{\vartheta}$.

Hence, those $\boldsymbol{\vartheta}$ from the prior $\pi(\cdot)$ that are likely according to the likelihood are kept in the posterior sample! For example, a random draw equal to the ML estimate of $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{x} | \boldsymbol{\vartheta})$$

will *always* be accepted!

Background: (Univariate) Slice Sampler

Consider now sampling a random variable ϕ from a nonstandard $p(\phi) \propto h(\phi)$.

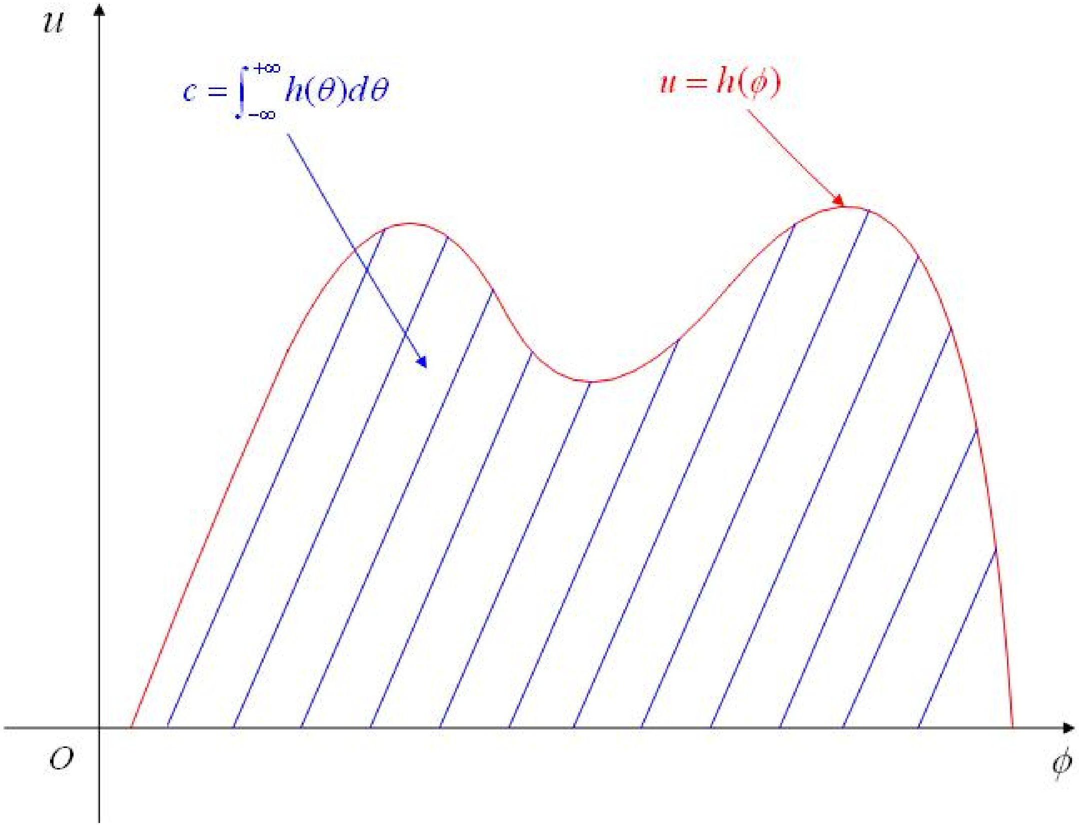
(Seemingly Counter-Intuitive!) Idea:

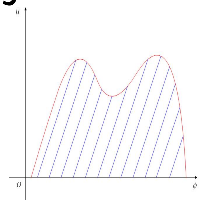
- Invent a convenient bivariate distribution for, say, ϕ and u, with marginal pdf for ϕ specified by $h(\phi)$.
- Then, use Gibbs sampling to make

$$(\phi^{(0)}, u^{(0)}), (\phi^{(1)}, u^{(1)}), (\phi^{(2)}, u^{(2)}), \ldots, (\phi^{(T)}, u^{(T)}).$$

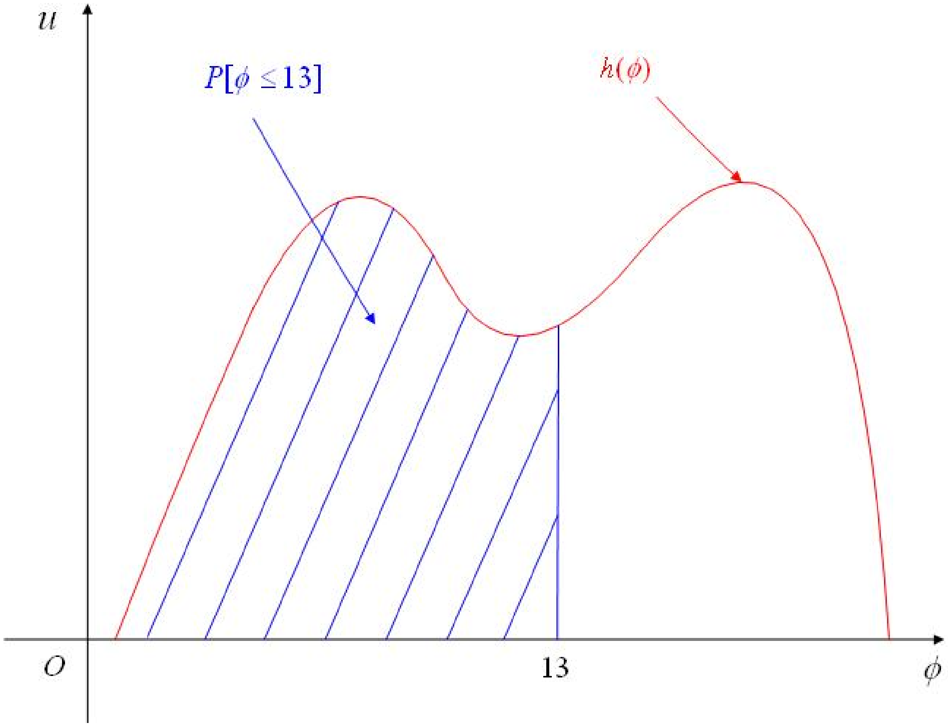
Create an auxiliary variable u just for convenience!

(Univariate) Slice Sampler



"Invent" a joint distribution for ϕ and u by declaring it to be uniform on  :

$$p(\phi, u) = \begin{cases} \frac{1}{c}, & 0 < u < h(\phi) \\ 0, & \text{otherwise} \end{cases} \propto \; i_{(0,h(\phi))}(u).$$



With this joint pdf, $P[\phi \leq 13] = \int_{-\infty}^{13} \frac{h(\phi)}{c}\, d\phi$.

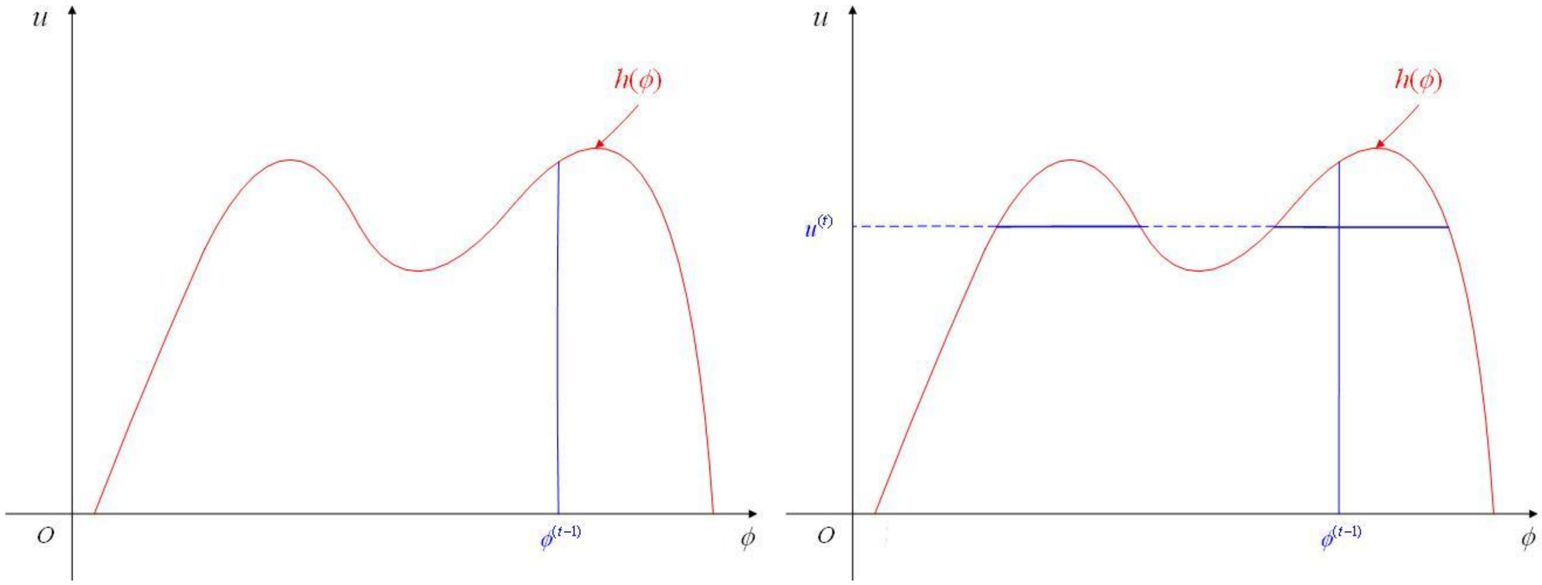
☺ The marginal pdf of ϕ is indeed specified by $h(\phi) \Longrightarrow$ if we figure out how to do Gibbs sampling, we know how to generate a ϕ from $h(\phi)$.

Gibbs Sampler is Easy in This Case!

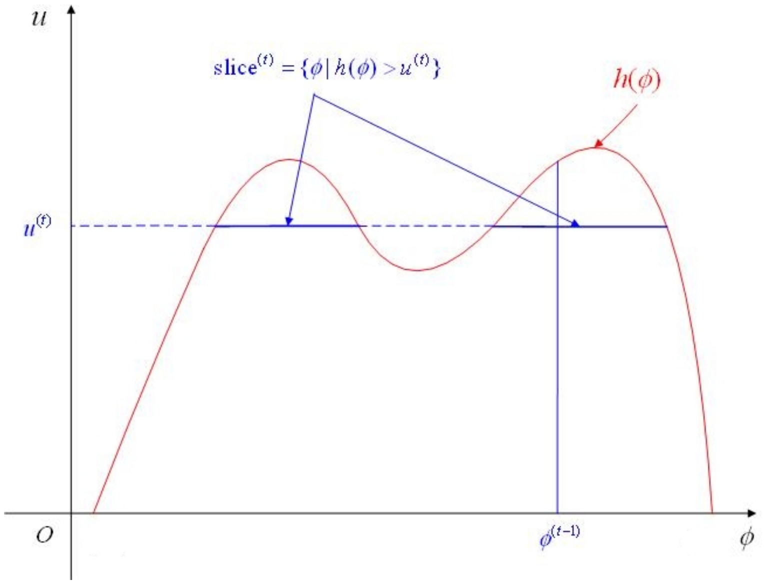
$$p(u \,|\, \phi) \;=\; \text{uniform}\Big(0, h(\phi)\Big)$$

$$p(\phi \,|\, u) \;=\; \text{uniform on } \underbrace{\{\phi \,|\, h(\phi) > u\}}_{\text{“slice”}}.$$

Step 1: Given $\phi^{(t-1)}$, sample $u^{(t)} \sim \text{uniform}\Big(0, h(\phi^{(t-1)})\Big)$



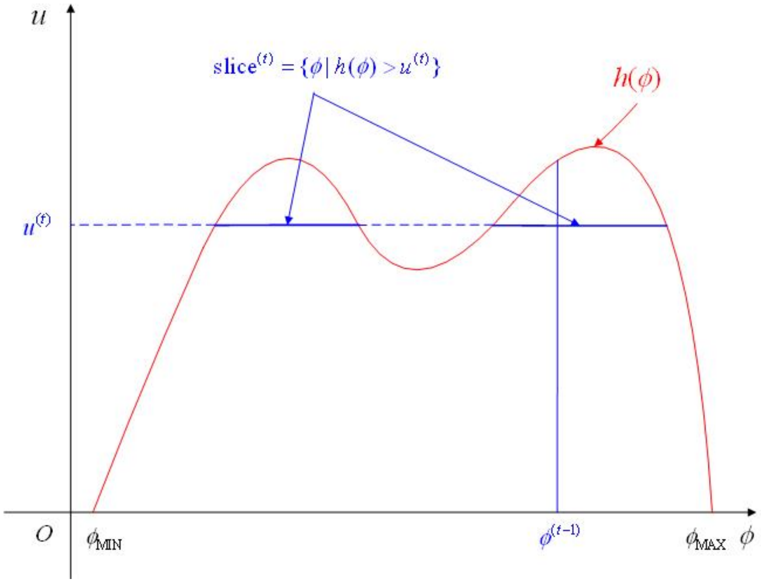
Step 2: Given $u^{(t)}$, sample $\phi^{(t)}$ Uniform from slice$^{(t)}$



If we can algebraically solve $h(\phi) = u^{(t)}$, our task is easy. What if not?

Step 2 implementation using the rejection method

When we have band bounds on ϕ, say $\phi_{\text{MIN}} \leq \phi \leq \phi_{\text{MAX}}$



generate i.i.d. values ϕ from uniform$(\phi_{\text{MIN}}, \phi_{\text{MAX}})$ until we produce a ϕ in the slice [i.e. $h(\phi) > u^{(t)}$], which we then accept as $\phi^{(t)}$.

Note: For multivariate extensions of the slice sampler (particularly the "shrinkage idea"), see

R.M. Neal, "Slice sampling," *Ann. Statist.*, vol. 31, pp. 705–741, June 2003.