Stochastic Simulation

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester

Notation

- Let X be a random variable (for example continuous), with $X \sim f_X(x)$, $X \in \mathcal{X}$ ($f_X(x)$ pdf).
  Then
  $$\mathbb{E}[X] = \int_{\mathcal{X}} x \; f_X(x) dx$$
  and
  $$\mathbb{V}[X] = \int_{\mathcal{X}} (x - \mathbb{E}[X])^2 \; f_X(x) dx$$
- Let $Y = h(X)$, $Y \in \mathcal{Y}$. If $Y \sim f_Y(y)$. Then
  $$\mathbb{E}[Y] = \int_{\mathcal{Y}} y \; f_Y(x) dx.$$
  We can prove that
  $$\mathbb{E}[Y] = \mathbb{E}_X[h(X)] = \int_{\mathcal{X}} h(x) f_X(x) dx$$
- Thus
  $$\mathbb{V}[X] = \mathbb{E}_X[(X - \mathbb{E}[X])^2]$$
- We also can prove that
  $$\mathbb{V}[X] = \mathbb{E}_X[X^2] - \mathbb{E}[X]^2$$
- When needed we will use the notation $\mathbb{E}_X$, or $\mathbb{E}_f$ equivalently, where f is the pdf of X, to explicitly state that the mean is w.r.t. the pdf f of X.

Introduction

- Whatever you would like to know about a distribution can be achieved simply by simulating random values from it.
- E.g. $f(x)$ pdf, $\quad X \sim f(x)$, $X \in \mathcal{X}$.

$$\mu = \mathbb{E}[X] = ?, \quad \sigma = \sqrt{\mathbb{V}[X]} = ?, \quad p = \mathbb{P}(X > 2) = ?, \quad \text{Graph?}$$

Let $X_1, \ldots X_n$ be i.i.d. r.v.'s from $f(x)$. Then,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > 2\}},$$

Graph $\rightarrow$ Kernel

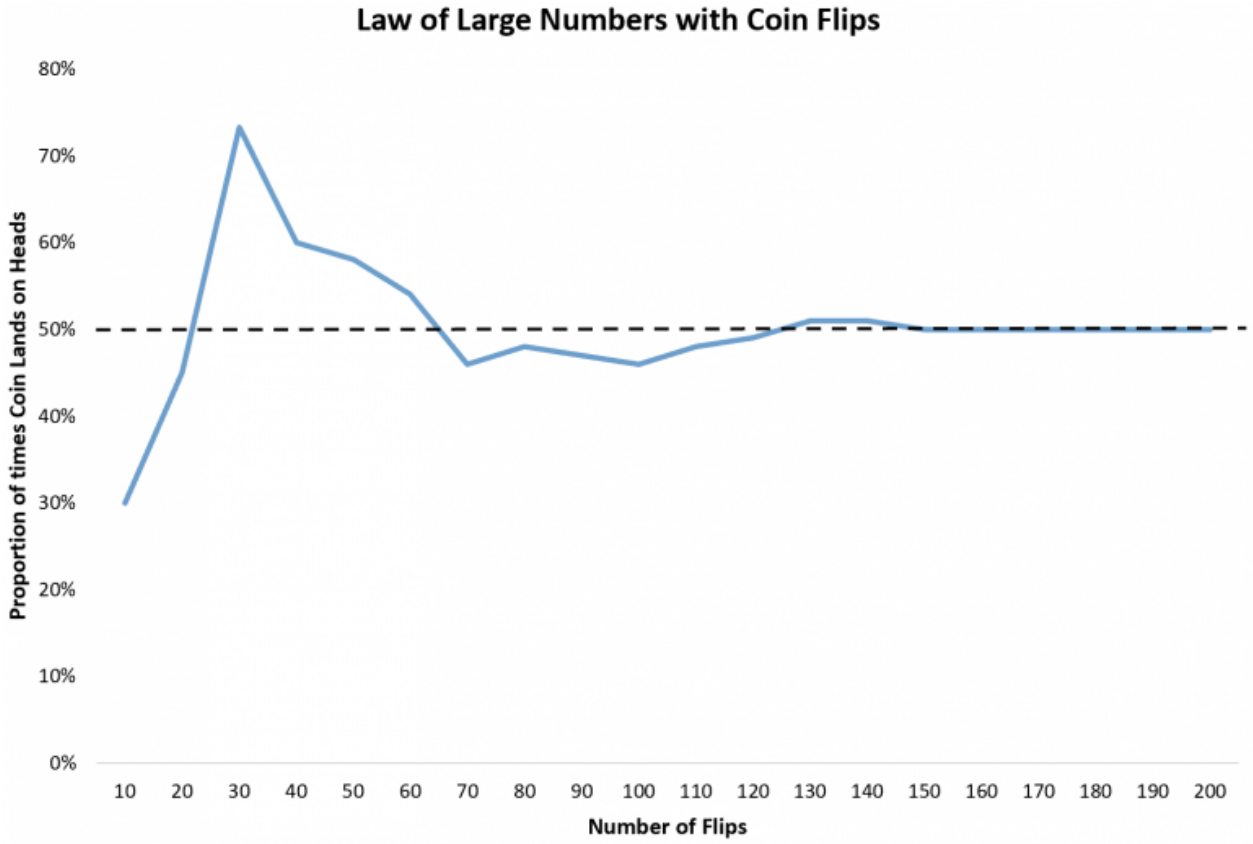
- In general, $\mathbb{E}_X[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx \;\rightarrow\; \frac{1}{n}\sum_{i=1}^{n} h(X_i)$
  $\rightarrow$ Monte Carlo Integration

  The above idea is used often in Bayesian statistics for computing integrals of the posterior (MCMC)

Theoretical Justification for Monte Carlo Integration

Theorem (Strong Law of Large Numbers)

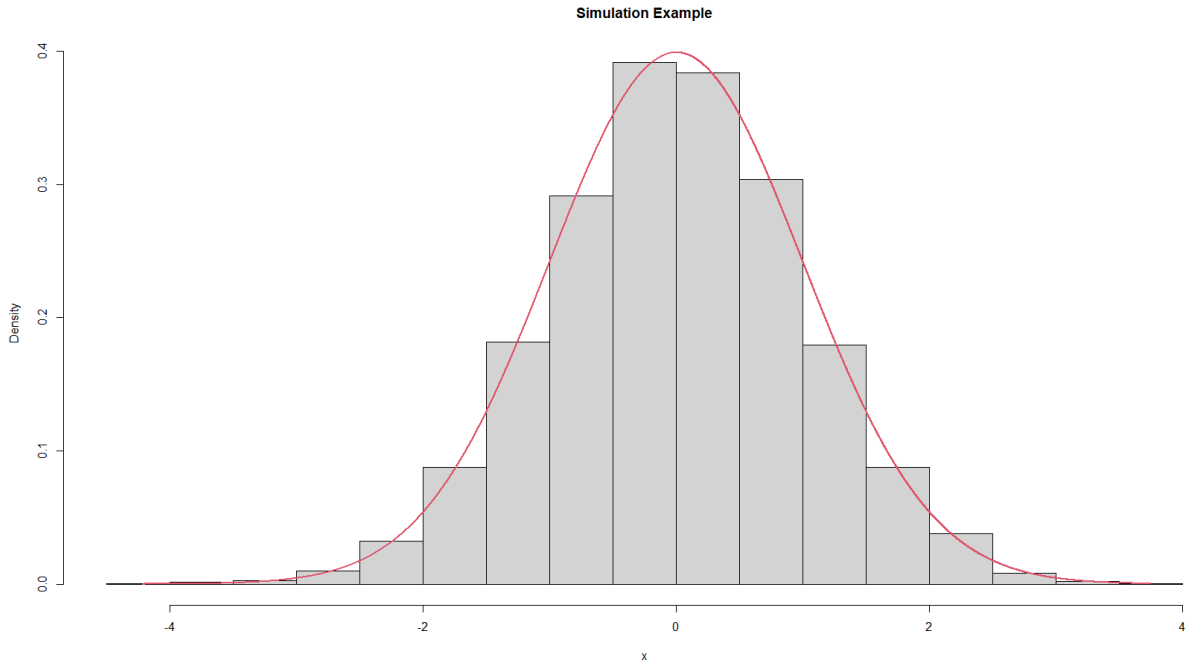
If $X_1, \ldots X_n$ are i.i.d. r.v.'s from $f(x)$ with $\mathbb{E}_X[|h(X)|] < +\infty$ then $\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \mathbb{E}_X[h(X)]$ a.s.



Introduction (cont'd)

Q: HOW DO WE SIMULATE R.V.'S from $f(x)$?

- Samples should be generated in proportion to the height (density) of the pdf f.



Introduction (cont'd)

- Statistical packages generate from known distributions and especially from the uniform $(0, 1)$.
- All algorithms we are going to see begin with $U_1, U_2 \ldots U_n \sim U[0, 1]$.

$$x_{n+1} = (ax_n + b) \mathrm{mod} m, \quad n = 0, 1, 2, \ldots$$

where x_0 (:seed), m (:modulus), b (:increment), a (:multiplier) non-negative integers with $x_0, b, a \in \{0, 1, \ldots m - 1\}$. In this way, we create a class of "generators" of numbers $x_0, x_1, \ldots \in \{0, 1, \ldots m - 1\}$. Then,

$$u_n = \frac{x_n}{m} \in [0, 1)$$

For large m and "appropriate" $x_0, b, a \rightarrow u_i \sim U[0, 1]$ (pseudo-random).

- If $U \sim U(0, 1)$ then $(\beta - \alpha)U + \alpha \sim U(\alpha, \beta)$ (inversion method)

Inversion Method

- We would like to generate $X \sim F(x)$ (cdf) and assume that the inverse function $F^{-1}(u)$ is well defined for $0 \leq u \leq 1$. If $U \sim U[0,1]$ then $X = F^{-1}(U)$ has the desired distribution. Indeed:

$$\mathbb{P}[X \leq x] = \mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[U \leq F(x)] = F(x)$$

(since $U \sim U[0,1] \rightarrow F_U(u) = u$)

- If X is discrete the above methodology needs to be modified, defining

$$F^{-1}(u) = \min\{x : \ F(x) \geq u\}$$

e.g. Consider we would like to simulate $X = \begin{cases} 1 & \text{prob} \quad p \\ 0 & \text{prob} \quad 1-p \end{cases}$.
We generate $U \sim U[0,1]$ and if $U \geq 1-p$, we set $X = 1$ else $X = 0$. (more about discrete r.v.'s later)

Inversion Method: Examples

1. $X \sim \text{Exp}(\lambda)$, $f(x) = \lambda e^{-\lambda x}$, $F(x) = 1 - e^{-\lambda x}$, $(x, \lambda > 0)$

   We set $X = F^{-1}(U)$, i.e. $U = 1 - e^{-\lambda X} \Rightarrow X = -\lambda^{-1} \log(1 - U)$

   But $(1 - U) \equiv U \sim U[0, 1]$. Thus,
   i. $U \sim U[0, 1]$
   ii. $X = -\lambda^{-1} \log(U)$
   iii. $X \sim \text{Exp}(\lambda)$
2. $X \sim$ Cauchy, $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{\pi(1+s^2)} \mathrm{d}s = 1/2 + \pi^{-1} \arctan x = u \\ &\Rightarrow \quad F^{-1}(u) = \tan(\pi(u - 1/2)) \end{aligned}$$

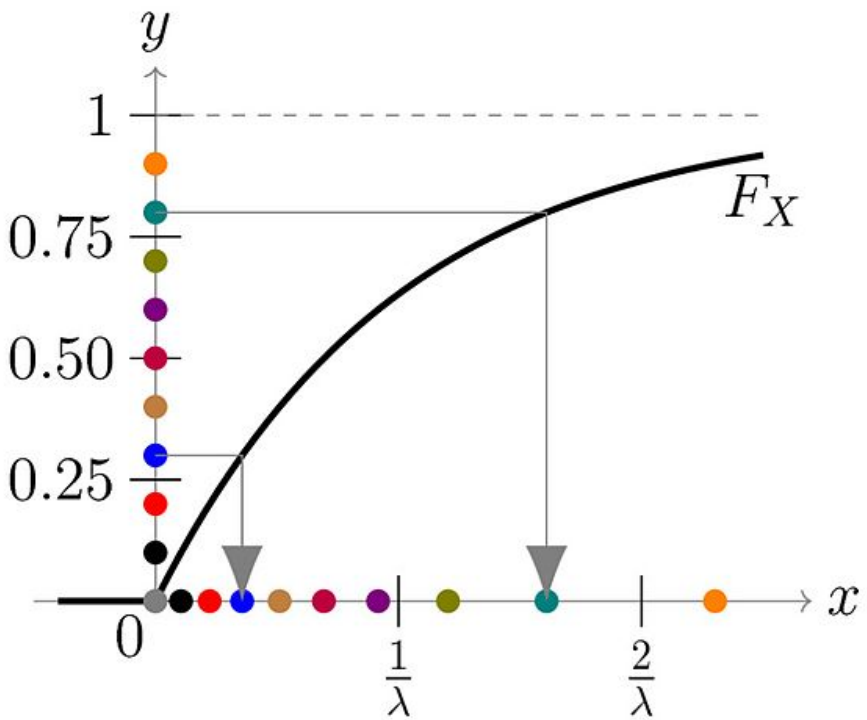
   Thus,
   i. $U \sim U[0, 1]$
   ii. $X = \tan(\pi(U - 1/2))$
   iii. $X \sim$ Cauchy
3. $X \sim \mathcal{N}(0, 1)$ with cdf $\Phi(x)$.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-2s^2} \mathrm{d}s = u \;\Rightarrow\; \Phi^{-1}(u) = ?$$

The requirement of the existence of the inverse is an Important Limitation of the method.

Inversion Method: Illustration

We generate random numbers u_i from $U(0,1)$ (y-axis). Each of the points is mapped according to $x = F^{-1}(u)$ (gray arrows). We use the exponential distribution with pdf $f(x) = \lambda e^{-\lambda x}$ and cdf $F(x) = 1 - e^{-\lambda x}$. The mean is equal to $\mu = 1/\lambda$, the standard deviation is $\sigma = 1/\lambda$ and thus $\mu + \sigma = 2/\lambda$.
We can see that using this method, many points end up close to 0 and only few points end up having high x-values - just as it is expected for an exponential distribution.



Inversion Method: Truncated Distributions

Let $X \sim f(x)$ (f is a pdf) and F be the corresponding cdf, with inverse (quantile function) F^{-1}. Let t denote the pdf of the truncated f distribution on $[a, b]$ and T denote the corresponding cdf.
Then

$$t(x) = \begin{cases} \frac{f(x)}{F(b)-F(a)} & x \in [a, b] \\ 0 & \text{else} \end{cases}$$

and

$$T(x) = \begin{cases} 1 & x > b \\ \frac{F(x)-F(a)}{F(b)-F(a)} & x \in [a, b] \\ 0 & x < a \end{cases}$$

Thus $T^{-1}(u) = F^{-1}\left[F(a) + u(F(b) - F(a))\right]$

Rejection Sampling

Imagine we would like to generate $X \sim f(x)$ (f is a pdf) but this is hard while at the same time we could easily get $Y \sim g(y)$ (proposal distribution) for which we could find an $M \geq 1$:

$$f(x) \leq Mg(x) \equiv G(x) \ \ (\textbf{envelope} - \text{also denoted by } e(x)),$$

for all x for which $f(x) > 0$.

Algorithm

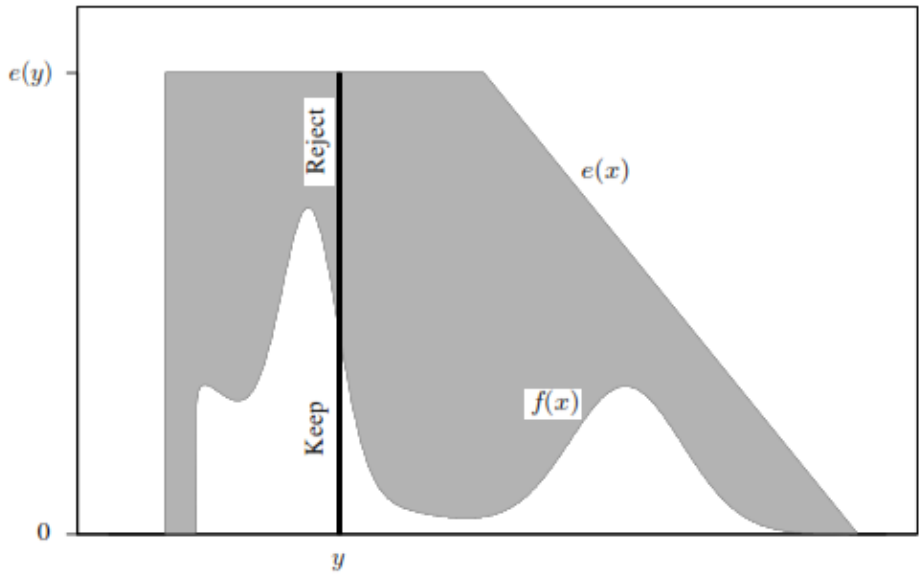
1. Generate $Y \sim g(y)$ and set $y = Y$
2. Generate $U \sim U(0, 1)$ and set $u = U$
3. If $u \leq \frac{f(y)}{Mg(y)}$ then $X = y$ else go back to Step 1.

This algorithm gives a way to generate randomly within the area under the curve $G = Mg$. We only accept those points that fall under f. This can be seen from the condition $u \leq \frac{f(y)}{Mg(y)} \Leftrightarrow Mg(y)u \leq f(y)$, where $Mg(y)u$ is a random point under Mg.

Rejection Sampling: Illustration

- Let $e(x) \equiv G(x) = Mg(x)$ denote the envelope
- $y \sim g$, g proposal
- $u \sim U(0,1)$ and keep y if $u \leq \frac{f(y)}{Mg(y)} \Rightarrow$
  $\Rightarrow u' = (u|y) \equiv Mg(y)u \sim U(0, e(y))$ and keep if $u' < f(y)$.

Suppose y falls at the point indicated by the figure. Then imagine sampling u' uniformly along the vertical bar. The rejection rule eliminates the y value with probability proportional to the length of the bar above $f(y)$ relative to the overall bar length.



Rejection sampling: Proof

Why would this give values from f?

It suffices to show that the cdf of X that we accept with the algorithm is $F(y) = \int_{-\infty}^{y} f(z)\mathrm{d}z$.

Indeed,

$$\begin{aligned}
\mathbb{P}[X \leq y] &= \mathbb{P}\left[Y \leq y | U \leq \frac{f(Y)}{G(Y)}\right] = \frac{\mathbb{P}\left[Y \leq y, U \leq \frac{f(Y)}{G(Y)}\right]}{\mathbb{P}\left[U \leq \frac{f(Y)}{G(Y)}\right]} \\
&= \frac{\int_{-\infty}^{y} \int_{0}^{f(z)/G(z)} \mathrm{d}u g(z)\mathrm{d}z}{\int_{-\infty}^{\infty} \int_{0}^{f(z)/G(z)} \mathrm{d}u g(z)\mathrm{d}z} = \frac{\int_{-\infty}^{y} f(z)/G(z) g(z)\mathrm{d}z}{\int_{-\infty}^{\infty} f(z)/G(z) g(z)\mathrm{d}z} \\
&= \frac{\frac{1}{M}\int_{-\infty}^{y} f(z)\mathrm{d}z}{\frac{1}{M}\int_{-\infty}^{\infty} f(z)\mathrm{d}z} = \int_{-\infty}^{y} f(z)\mathrm{d}z = F(y)\,.
\end{aligned}$$

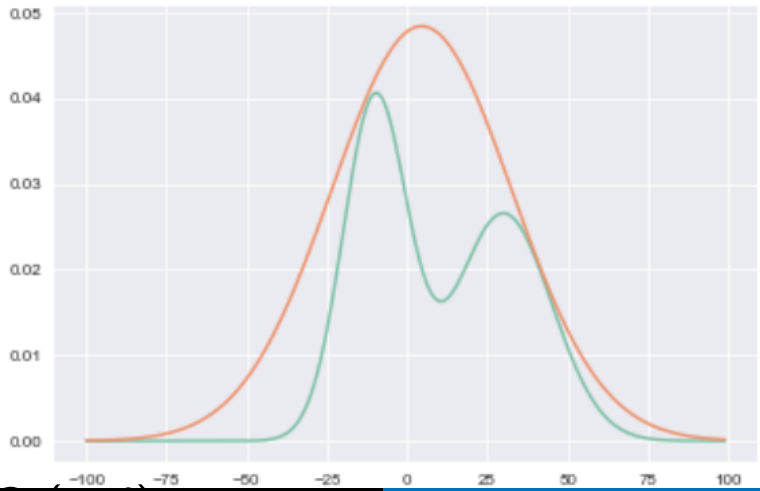
Rejection sampling (cont'd)

How many trials (on average) do we need till we accept the value $X = y$ at Step 3?

$$\mathbb{P}[\text{accept the value } X = y] = \int_{-\infty}^{\infty} \underbrace{\frac{f(y)}{G(y)}}_{\downarrow} g(y)\mathrm{d}y = 1/M\,.$$

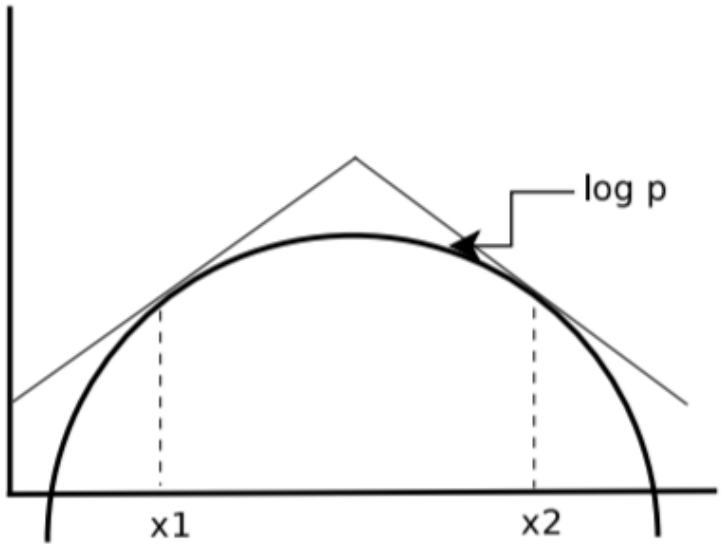
from every $y \sim g$ we only accept those with prob $f(y)/G(y)$

Thus, in order to achieve high acceptance probability, M should be as small as possible (closer to 1). From the graph on p. 12, we also see that the closer G is to f, the higher the acceptance probability will be. Thus, the smaller the M, the closer G to f. In general, the better g mimicks the shape of f, the smaller the value of M required to find G covering f.



Adaptive Rejection Sampling

- If $f \equiv p$ is log-concave (i.e. its logarithm is a concave function ($d^2 \log f(x)/dx^2 < 0$) – most distributions are) then one way of choosing G is to draw tangents at each side of the maximum.



Adaptive rejection sampling: choose x_1, x_2 yielding small rejection probability or draw more tangents.

Adaptive Rejection Sampling (cont'd)

- These piecewise linear proposals on the log scale result in a set of piecewise exponential proposal distributions on the original scale.
- We can easily simulate values from exponential distributions, using the inversion method (and a stratified sampling method):
  1. Sample from multinomial distribution to determine the "piece".
  2. Sample from the truncated exponential distribution.

Remarks on Rejection Sampling

- If we do not want to do adaptive rejection sampling then choosing g is art. We usually choose g with the same support as f being easy to simulate values from and mimicking f as much as possible. In general, almost always we can find a pdf g and $M > 1$: $f \leq Mg$ unless f is not bounded or its tails are very heavy (e.g. f : Cauchy, g : Normal $\rightarrow \nexists M : f(x)/g(x) \leq M \ \ \forall x$)
- If we know f up to a normalizing constant c, that is we have $f^*(x) = f(x)/c$ (Bayesian Statistics) then we can work with $f^*(x)$ without a problem. The envelope then is $G(y) \geq f^*(y)$ and $u \leq f^*(y)/G(y)$. Finally, the probability to accept X at step 3 becomes $1/cM$. $\left(\text{e.g. } f(x) \sim \text{Beta}(a, b),\ f^*(x) = (1-x)^{b-1}x^{a-1}\right)$
- As we saw before, M should be the smallest possible provided that G covers f, for all x: $f(x) > 0$:

$$f \leq Mg \Rightarrow M \geq f/g \Rightarrow M = \max_y \frac{f(y)}{g(y)}$$

Rejection Sampling: Example

Assume we want $X \sim \text{Beta}(a, b)$

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(1-x)^{b-1}x^{a-1} \propto x^{a-1}(1-x)^{b-1} \equiv f^*(x), x \in (0,1)$$

Let $a, b > 1$ (these values ensure concavity). It is easier to work with f^* in the place of f. Then,

i) Choice of $g \equiv U(0,1)$ (g has the same support with f).

ii) Find M so that $M = \max \frac{f^*}{g}$. But,

$$\frac{f^*(x)}{g(x)} = x^{a-1}(1-x)^{b-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{f^*(x)}{g(x)}\right) = 0 \Rightarrow x = \frac{a-1}{a+b-2}$$

Thus,

$$\frac{f^*}{g} \leq \frac{(a-1)^{a-1}(b-1)^{b-1}}{(a+b-2)^{a+b-2}} = M$$

Rejection Sampling: Example II

Suppose we want $X \sim \mathcal{N}(0,1)$, i.e. $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

It is easy to generate samples from a double exponential distribution with density $g(x) = \frac{1}{2} e^{-|x|}$.

Indeed, if $E \sim$exp(1) and $S \sim$ Uniform$\{-1, +1\}$, then $Y = SE$ has density g.

It is easy to check that

$$\frac{f_X(x)}{g(x)} = \sqrt{\frac{2e}{\pi}} e^{-\frac{1}{2}(|x|-1)^2} \leq \sqrt{\frac{2e}{\pi}} = M \ (\simeq 1.3155)$$

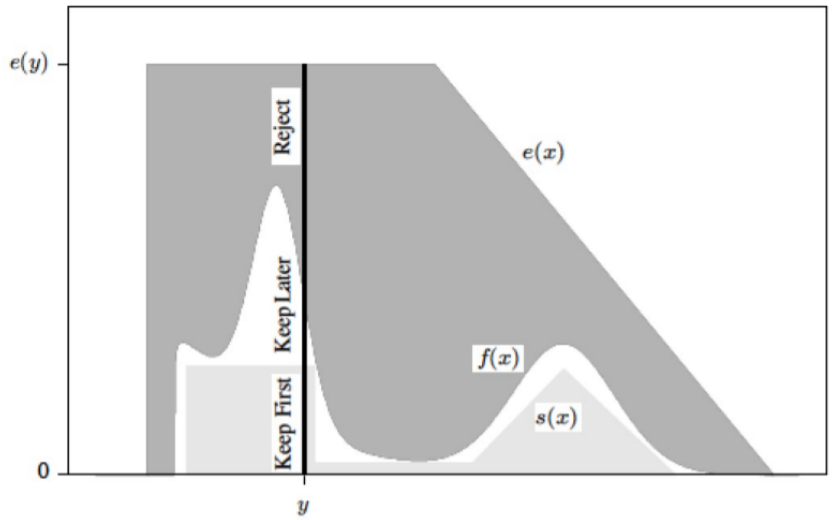
The algorithm for generating samples of X from samples of Y is as follows

REPEAT
    draw a sample y from density g.
    draw a sample u form Uniform(0,1)
UNTIL $u \leq e^{-\frac{1}{2}(|y|-1)^2}$
RETURN y

Squeezed Rejection Sampling

If computing f is time-consuming then the method becomes slow since at every step we need to compute $f(y)$. Let $e(x) \equiv G(x) = Mg(x)$ denote the envelope and s a (squeezing) function: $f(y) \geq s(y) \geq 0 \; \forall y$, which does not take time to compute. Then,

1. Generate $Y \sim g(y)$ and set $y = Y$
2. Generate $U \sim U(0, 1)$ and set $u = U$
3. If $u \leq \frac{s(y)}{G(y)}$ then $X = y$
   else if $u \leq \frac{f(y)}{G(y)}$ then $X = y$
   else go back to Step 1.



For a given $Y = y$, the total acceptance probability is $\frac{f(y)}{G(y)}$ like in the rejection sampling method, while $\frac{1}{M}$ remains the probability to accept some $X = y$. Finally, the proportion of repetitions where we avoid computing f is $\frac{\int_{-\infty}^{\infty} s(x)\mathrm{d}x}{\int_{-\infty}^{\infty} G(x)\mathrm{d}x}$.

Methods of generation from Standard Normal Distribution

A. Box-Müller

Let 2 standard normal $\mathcal{N}_1(0,1), \mathcal{N}_2(0,1)$ that generate two independent values N_1, N_2 respectively. In this way, a point is defined in the two-dimensional space with Cartesian coordinates and if we would like to transform to polar coordinates then $N_1 = R\cos(\Theta), N_2 = R\sin(\Theta)$. It can be shown (see Appendix 1) that R and Θ are independent r.v.'s with $\Theta \sim U(0, 2\pi)$ and $R^2 = N_1^2 + N_2^2 \sim X_2^2 \equiv \text{Exp}(1/2)$. To generate Θ, we simply generate $U_2 \sim U(0,1)$ and set $\Theta = 2\pi U_2$, while to generate R we set $R = (-2\ln U_1)^{1/2}$ ($U_1 \sim U(0,1)$ - inversion method). Consequently, the method generates polar coordinates and afterwards transforms in Cartesian coordinates N_1, N_2. For each pair of values of the normal distribution, two independent values $U_1, U_2 \sim U[0,1]$ are required as well as the computation of trigonometric functions. If a large sample is necessary, then the computation of trigonometric functions has a negative effect on the efficiency of the algorithm, so we use the Polar-Marsaglia method.

Note: If $Z \sim \mathcal{N}(0,1)$ then $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$

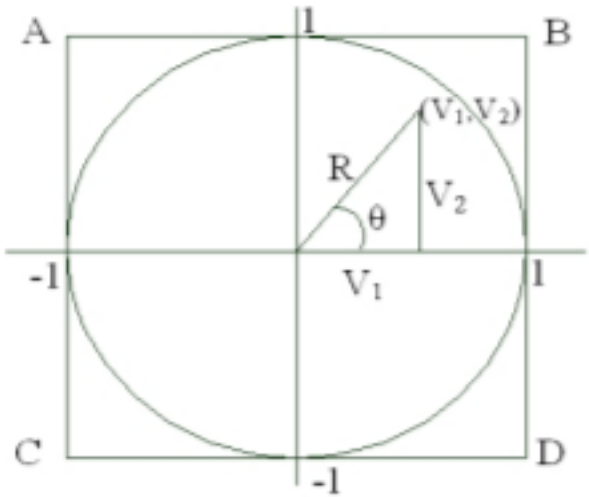
Methods of generation from Standard Normal Distribution (cont'd)

B. Polar-Marsaglia

The idea is based on the construction of sines and cosines of uniformly distributed angles without having to simulate the angles! This is achieved with the rejection sampling method as follows:

$$U_1, U_2 \sim U(0,1) \quad \text{and} \quad V_i = 2U_i - 1 \Rightarrow V_i \sim U(-1,1)\ (i = 1,2)\,.$$

Thus, (V_1, V_2) corresponds to a random pair inside a square of side 2 and center $(0,0)$. We continue to generate such pairs $(V_1, V_2), (V_3, V_4), \ldots$ till $V_i^2 + V_{i+1}^2 < 1$, i.e. till the point belongs to the unit circle with center $(0,0)$ (rejection sampling method: the pairs are rejected with probability $1 - \pi/4$).



Methods of generation from Standard Normal Distribution (cont'd)

Then, we set

$$R^2 = V_1^2 + V_2^2, \quad \tan\Theta = V_2/V_1.$$

It can be shown that $R^2 \sim U(0,1)$ and $\Theta \sim U(0, 2\pi)$ and they are independent r.v.'s. Thus, the pair (R, Θ) is what is required for the Box-Müller method and we can write:

$$\sin\Theta = V_2(V_1^2 + V_2^2)^{-1/2}, \quad \cos\Theta = V_1(V_1^2 + V_2^2)^{-1/2}$$

Thus,

$$\begin{aligned} N_1 &= (-2\ln(R^2))^{1/2} V_1(V_1^2 + V_2^2)^{-1/2} \\ N_2 &= (-2\ln(R^2))^{1/2} V_2(V_1^2 + V_2^2)^{-1/2} \end{aligned}$$

or equivalently

$$\begin{aligned} N_1 &= (-2\ln(V_1^2 + V_2^2))^{1/2} V_1(V_1^2 + V_2^2)^{-1/2} \\ N_2 &= (-2\ln(V_1^2 + V_2^2))^{1/2} V_2(V_1^2 + V_2^2)^{-1/2} \end{aligned}$$

yielding

$$N_1 = V_1\left(\frac{-2\ln W}{W}\right)^{1/2}, \quad N_2 = V_2\left(\frac{-2\ln W}{W}\right)^{1/2}$$

where $W = V_1^2 + V_2^2$.

Discrete Random Variables

Let X be a discrete r.v. with values $i = 1, 2, \ldots n$.

If $f(i) = \mathbb{P}[X = i] = p_i$ is the pmf and $F(i) = \mathbb{P}[X \leq i] = \sum_{j \leq i} p_j = P_i$ the cdf of the r.v. X then

$$F^{-1}(u) = \min(x | F(x) \geq u) = i, \text{ if } P_{i-1} < u \leq P_i$$

where $P_0 = 0$ and $P_n = 1$.

Thus,

i) We simulate $U \sim U(0, 1)$ and set u=U

ii) We set $i = 1$

iii) If $P_i \leq u$ then $i \to i + 1$ and repeat iii), else $X = i$

Example (Inversion method on discrete)

Flip a fair coin 3 times and X : $\#heads + 1$

X discrete $(i = 1, 2, 3, 4)$

$$f(x) = \begin{cases} 1/8 & x = 1 \rightarrow p_1 \\ 3/8 & x = 2 \rightarrow p_2 \\ 3/8 & x = 3 \rightarrow p_3 \\ 1/8 & x = 4 \rightarrow p_4 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \quad \rightarrow P_0 = 0 \\ 1/8 & 0 \leq x < 1 \rightarrow P_1 = 1/8 \\ 3/8 & 1 \leq x < 2 \rightarrow P_2 = 4/8 \\ 3/8 & 2 \leq x < 3 \rightarrow P_3 = 7/8 \\ 1/8 & 3 \leq x \rightarrow P_4 = 1 \end{cases}$$

$u \sim U(0, 1)$

ex1. 0.43, $i = 1$

$P_1 \leq u \rightarrow i = 2$

$P_2 \not\leq u \rightarrow$ X=2

ex2. 0.77, $i = 1$

$P_1 \leq u \rightarrow i = 2$

$P_2 \leq u \rightarrow i = 3$

$P_3 \not\leq u \rightarrow$ X=3

Mixture Representations

- Sometimes probability distributions can be naturally represented as a *mixture distribution*, i.e.

$$f_X(x) = \int_{\mathcal{Y}} f_{X|Y}(x|y) f_Y(y) dy$$

or

$$f_X(x) = \sum_{i \in \mathcal{Y}} p_i f_i(x),$$

depending on whether the auxiliary space $\mathcal{Y}$ is continuous or discrete. In the above expressions $f_{X|Y}(x|y), f_Y(y)$ are pdfs, $f_i(x),\ i \in \mathcal{Y}$ are pdfs or pmfs (of the same or of different types) and $p_i,\ i \in \mathcal{Y}$ are probabilities such that $\sum_{i \in \mathcal{Y}} p_i = 1$ (therefore we have a discrete r.v. $\gamma \in \mathcal{Y}$ with pmf p, such that $\mathbb{P}(\gamma = i) = p_i,\ i \in \mathcal{Y}$).

- To generate a r.v. X using such a representation, we can first generate a variable Y from the mixing distribution and then generate X from the selected conditional distribution (given the generated y value).

Mixture Representations (cont'd)

- Therefore:
  - If $Y \sim f_y(y)$ and $X|Y \sim f_{X|Y}(x|y)$, then $X \sim f_X(x)$ (continuous case).
  - If $\gamma \sim p$ and $X \sim f_\gamma(x)$, then $X \sim f_X(x)$ (discrete case).
- **Example (discrete case) - Mixtures of Normals**:

$$f_X(x) = \frac{1}{3}N(0,1) + \frac{2}{3}N(1,1)$$

Simulate $\gamma \sim Bernoulli(1/3)$. If $\gamma = 1$ then $X \sim N(0,1)$, if $\gamma = 0$ then $X \sim N(1,1)$.

- **Example (continuous case)**: The Student's density T_ν, with ν degrees of freedom, can be written as a mixture:

$$X|y \sim N(0, \nu/y) \text{ and } Y \sim X^2_\nu$$

Therefore to generate a r.v. from T_ν, we generate $Y \sim X^2_\nu$ (call the value y) and then $X \sim N(0, \nu/y)$.

Special Cases

i) Cauchy:
If N_1 and N_2 independent r.v.'s following N(0,1), then
$X = N_1/N_2 \sim$ Cauchy.

ii) $\Gamma(n, \theta)$, $n \in \mathbb{N}$, $\theta > 0$:
If $X_1, \ldots, X_n$ iid sample from $Exp(\theta)$ then $X = \sum_{i=1}^{n} X_i \sim \Gamma(n, \theta)$.

iii) X_n^2:
If $Z \sim N(0, 1)$, then $Z^2 \sim X_1^2$. If $X_1, \ldots, X_n$ iid sample from X_1^2 then $X = \sum_{i=1}^{n} X_i \sim X_n^2$.

iv) $Beta(p, q)$:
If $Y \sim \Gamma(p, \theta)$ and $Z \sim \Gamma(q, \theta)$, then $\frac{Y}{Y+Z} \sim Beta(p, q)$

Special Cases (cont'd)

v) Poisson(μ), $\mu \leq 30$:
- Set $P = 1, N = 0, c = e^{-\mu}$
- Repeat: simulate $U_i \sim U(0,1), P = PU_i, N = N + 1$ till $P < c$
- $X = N - 1 \sim \mathrm{P}(\mu)$

vi) Geometric:
Let $Y \sim \mathrm{Exp}(\lambda), X = int[Y]$ (i.e. the largest integer $\leq Y$). Then,

$$\mathbb{P}[X = r] = \mathbb{P}[r \leq Y \leq r + 1] = e^{-\lambda r} - e^{-\lambda(r+1)} = e^{-\lambda r}(1 - e^{-\lambda})$$

i.e. $X \sim \mathrm{Ge}(p = 1 - e^{-\lambda})$. Thus, for $\lambda = \log(1/(1-p))$ we simulate from $\mathrm{Ge}(p)$.

vii) Negative Binomial:
$X|y \sim Poisson(y)$ and $Y \sim Gamma(n, \beta) \Rightarrow X \sim NB(n, p)$, where $\beta = (1-p)/p$

vii) t-distribution: $T_\nu = \frac{\mathcal{N}(0,1)}{\sqrt{X^2_\nu/\nu}}$ (equivalent to the mixture representation we have seen before)

viii) F-distribution: $F(m, n) = \frac{X^2_m/m}{X^2_n/n}$

R functions

- runif(n,par1,par2)
- rnorm(n,par1,par2)
  ↓
  sd
- rgamma(n,shape,rate) $\quad (f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax})$
  ↓ ↓
  p a
- rbeta(n,shape1,shape2)
  ↓ ↓
  a b
- rcauchy(n,par1,par2)
  ↓ ↓
  0 1
- rf(n,par1,par2)
- rt(n,par1,par2)

Variance Reduction Techniques

Let $X \sim f(x)$, $X \in \mathcal{X}$ (f: pdf) and assume that the value $\theta = \mathbb{E}_X[\phi(X)] = \int_{\mathcal{X}} \phi(x) f(x) \mathrm{d}x$ is unknown. Let $X_1, \ldots X_n$ be a random sample from f. According to Monte Carlo integration we have:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i) \, .$$

Then,

$$\mathbb{E}[\hat{\theta}] \quad = \quad n^{-1} \sum_{i=1}^{n} \mathbb{E}_{X_i}[\phi(X_i)] = n^{-1} n \mathbb{E}_X[\phi(X)] = \int_{\mathcal{X}} \phi(x) f(x) \mathrm{d}x = \theta$$

and

$$\mathbb{V}[\hat{\theta}] = n^{-2} \sum_{i=1}^{n} \mathbb{V}_X[\phi(X)] = n^{-1} \int_{\mathcal{X}} [\phi(x) - \theta]^2 f(x) \mathrm{d}x = \frac{c}{n}$$

where c: constant. Thus, $\hat{\theta}$ is an unbiased estimator of θ with s.e. proportional to $1/\sqrt{n}$, thus $\hat{\theta}$ is a consistent estimator of θ. Let us call the current method ϕ-f. (Monte-Carlo integration or parametric Bootstrap).

Variance Reduction Techniques (cont'd)

It is clear that different choices of f and ϕ yield different constants c.

e.g. let $\theta = \int_2^\infty \frac{1}{\pi}\frac{1}{1+x^2}\mathrm{d}x$ which is equal to $\mathbb{P}[X > 2]$ when $X \sim$ Cauchy (symmetric around zero)

i)

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad \phi(x) = \mathbb{1}_{\{x>2\}} = \begin{cases} 1, & x > 2 \\ 0, & x \leq 2 \end{cases}$$

ii)

$$\theta = \frac{1}{2}\mathbb{P}[|X| > 2], \quad f(x) = \frac{1}{\pi(1+x^2)}, \quad \phi(x) = \mathbb{1}_{\{|x|>2\}}$$

iii)

$$1 - 2\theta = \int_{-2}^{2} \frac{1}{\pi(1+x^2)}\mathrm{d}x = 2\int_0^2 \frac{1}{\pi(1+x^2)}\mathrm{d}x$$

$$f(x) = \frac{1}{2}, \; x \in [0,2] \; (i.e.\; U(0,2)), \quad \phi(x) = \frac{2}{\pi(1+x^2)}$$

Method of Control Variates

Let $X \sim f(x)$, $X \in \mathcal{X}$ (f: pdf) and assume that the value $\theta = \mathbb{E}_X[\phi(X)] = \int_{\mathcal{X}} \phi(x) f(x) \mathrm{d}x$ is unknown. Let $X_1, \ldots X_n$ be a random sample from f. According to Monte Carlo integration we have that

$$\delta_1 = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$$

is an unbiased estimator of θ.

In some settings, there exist functions ϕ_0 whose mean under f is known. For instance, if f is symmetric around its mean μ, the mean of $\phi_0(X) = \mathbb{1}_{\{X \geq \mu\}}$ is $1/2$. Let's suppose we simulate i.i.d. sample $\phi_0(X_1), \ldots \phi_0(X_n)$ and

$$\delta_3 = \frac{1}{n} \sum_{i=1}^{n} \phi_0(X_i)$$

is the Monte Carlo unbiased estimator of $\mathbb{E}_X[\phi_0(X)]$.

Define the new unbiased estimator of θ:

$$\delta_2 = \frac{1}{n} \sum_{i=1}^{n} [\phi(X_i) - b\, \phi_0(X_i)] + b\, \mathbb{E}_X[\phi_0(X)].$$

Method of Control Variates (cont'd)

Goal: Choose b to make the variance of δ_2 smaller.

$$\mathbb{V}_X\left[\phi(X) - b\,\phi_0(X) + b\,\mathbb{E}_X[\phi_0(X)]\right] = \mathbb{V}_X\left[\phi(X) - b\,\phi_0(X)\right] =$$
$$\mathbb{V}_X[\phi(X)] + b^2\,\mathbb{V}_X[\phi_0(X)] - 2b\,\mathrm{Cov}\left[\phi(X), \phi_0(X)\right]$$

This is minimised when $b = \frac{\mathrm{Cov}[\phi(X),\phi_0(X)]}{\mathbb{V}_X[\phi_0(X)]}$ and the minimum variance achieved is $\mathbb{V}_X[\phi(X)]\left(1 - \rho^2_{\phi(X),\phi_0(X)}\right)$. Thus

$$\mathbb{V}_X[\delta_3] = \mathbb{V}_X[\delta_1]\left(1 - \rho^2_{\phi(X),\phi_0(X)}\right).$$

Caveat 1: The reduction in variance comes with increase in simulation time, since we also need to simulate $\phi_0(X)$. Suppose we need time τ to sample from X and time $\lambda\tau$ to sample from $\phi_0(X)$. In the time needed to draw n samples from X, we can only draw $n/(1+\lambda)$ samples of the pair $(X, \phi_0(X))$. Hence, a fair comparison would be between

$$\frac{\mathbb{V}_X[\phi(X)]}{n} \quad \text{and} \quad \mathbb{V}_X[\phi(X)]\left(1 - \rho^2_{\phi(X),\phi_0(X)}\right) \times \frac{1+\lambda}{n}$$

The more correlated $\phi(X)$ and $\phi_0(X)$ are, the more advantageous it is to use $\phi_0(X)$ as a control variate.

Method of Control Variates (cont'd)

Caveat 2: The optimal $b = \frac{\text{Cov}[\phi(X),\phi_0(X)]}{\mathbb{V}_X[\phi_0(X)]}$ requires knowledge of $\text{Cov}\left[\phi(X), \phi_0(X)\right]$, which is not very realistic, considering we do not even know $\mathbb{E}_X[\phi(X)]$.

Remedy: Use an estimator of the optimal b, i.e.,

$$\hat{b}_n = \frac{\sum_{i=1}^n (\phi(X_i) - \delta_1)(\phi_0(X_i) - \delta_3)}{\sum_{i=1}^n (\phi_0(X_i) - \delta_3)^2}.$$

In this case, the new Monte Carlo estimator becomes

$$\delta_2 = \sum_{i=1}^n \Big(\frac{1}{n} + \underbrace{\frac{(\delta_3 - \phi_0(X_i))\,(\delta_3 - \mathbb{E}_X[\phi_0(X)])}{\sum_{i=1}^n (\phi_0(X_i) - \delta_3)^2}}_{\text{correction term}} \Big) \phi(X_i).$$

Example on Control Variates

We wish to find the value of the following integral (true value is $\log 2 = 0.69314718$):

$$\theta = \int_0^1 \frac{1}{1+x} \mathrm{d}x$$

This integral is the expected value of $\phi(X) = \frac{1}{1+X}$, with $X \sim U(0,1)$.

- Standard Monte-Carlo Method

  Take a sample of let's say $n = 1500$ $U(0,1)$ variates $X_1, \ldots X_{1500}$ and take $\delta_1 = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.

- Control Variates

  Let $\phi_0(X) = 1 + X \sim U(1,2)$. Then

  $$\mathbb{E}_X[\phi_0(X)] = \int_0^1 (1+x)\mathrm{d}x = \frac{3}{2}.$$

  Let

  $$\delta_2 = \frac{1}{n}\sum_{i=1}^{n} [\phi(X_i) - b\,\phi_0(X_i)] + b\,\frac{3}{2},$$

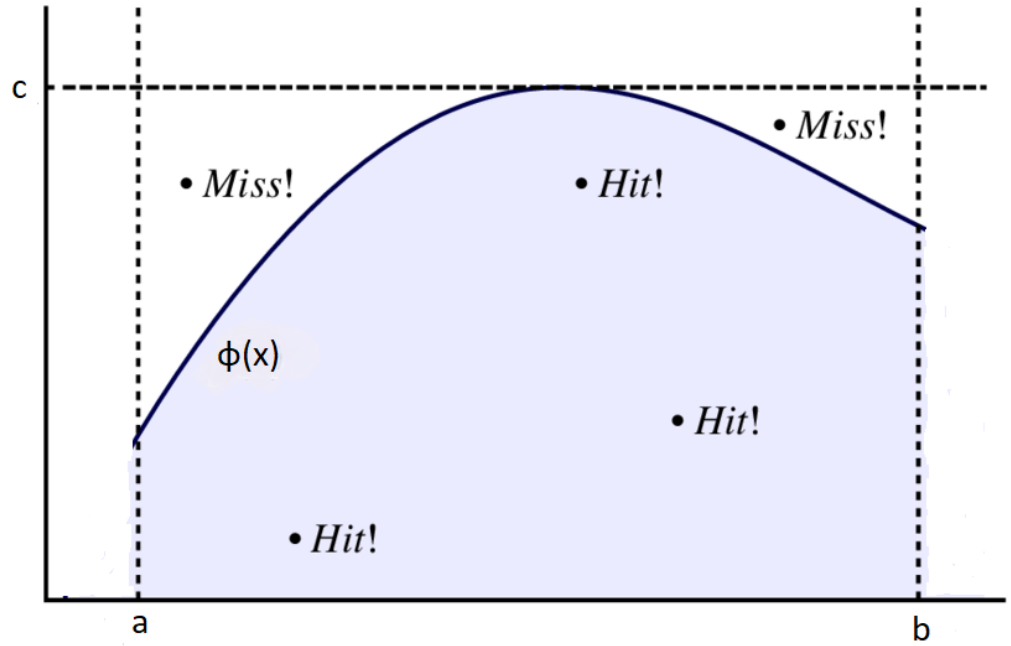
  The estimated optimal $\hat{b}_n = -0.47$. Then $\delta_1 = 0.69475$ with variance $= 0.01947$ & $\delta_2 = 0.69295$ with variance $= 0.00060$.

Method "hit and miss"

Let $\phi(x)$, $x \in [a, b]$ be a bounded function $0 \leq \phi(x) \leq c$ and we want to compute the area under ϕ, i.e.

$$Y = \int_a^b \phi(x)\mathrm{d}x = (b-a)\int_a^b \phi(x)f(x)\mathrm{d}x, \quad f(x) = \frac{1}{b-a}, x \in [a, b]$$

Thus $f(x)$ denotes the pdf of $U(a, b)$. The method generates points randomly inside the rectangle defined by the points $(a, 0), (b, 0), (a, c)$, (b, c) and estimates Y by the relative frequency of the points that fall under $\phi(x)$.



Method "hit and miss" (cont'd)

More specifically, let $U_i \sim U(a,b)$ and $V_i \sim U(0,c)$, $i = 1, 2 \ldots n$ (independent r.v.'s). We thus create a random sample of size n in the rectangle enclosing ϕ. Then we estimate Y by

$$\widetilde{Y} \underset{\text{Appendix 2}}{\overset{\text{see}}{=}} c(b-a)\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{V_i \leq \phi(U_i)\}}$$

rectangle area relative frequency of points under $\phi(x)$

$$\mathbb{E}[\widetilde{Y}] \underset{\text{call them } V \text{ \& } U}{\overset{V_i \text{ \& } U_i \text{ i.i.d.}}{=}} c(b-a)\frac{1}{n} n \mathbb{E}\left[\mathbb{1}_{\{V \leq \phi(U)\}}\right] = c(b-a)\mathbb{P}\left[V \leq \phi(U)\right],$$

since $\mathbb{1}_{\{V \leq \phi(U)\}} \sim$ Bernoulli$(\mathbb{P}[V \leq \phi(U)])$. Furthermore, $V \sim U(0,c)$ and $U \sim U(a,b)$. Thus:

$$\mathbb{E}[\widetilde{Y}] \underset{\text{Appendix 3}}{\overset{\text{see}}{=}} c(b-a)\frac{\int_a^b \phi(x)\mathrm{d}x}{c(b-a)} = \int_a^b \phi(x)\mathrm{d}x = Y$$

Method "hit and miss" (cont'd)

Additionally

$$\begin{aligned}
\mathbb{V}[\widetilde{Y}] &= c^2(b-a)^2\frac{1}{n^2}n\mathbb{V}\left[\mathbb{1}_{\{V\leq\phi(U)\}}\right] \\
&= c^2(b-a)^2\frac{1}{n}\left\{\mathbb{P}\left[V\leq\phi(U)\right]\left(1-\mathbb{P}\left[V\leq\phi(U)\right]\right)\right\} \\
&= \frac{c^2(b-a)^2}{n}\left[\frac{Y}{c(b-a)}\left(1-\frac{Y}{c(b-a)}\right)\right] = \frac{1}{n}Y[c(b-a)-Y]\,.
\end{aligned}$$

Method "hit and miss" (cont'd)

With the "ϕ-f" method, we have respectively:
$\hat{Y} = (b-a)\frac{1}{n}\sum_{i=1}^{n}\phi(X_i),\ X_i \sim U(a,b)$
$\mathbb{E}[\hat{Y}] = Y$

$$
\begin{aligned}
\mathbb{V}[\hat{Y}] &= (b-a)^2\frac{1}{n^2}n\mathbb{V}_X[\phi(X)] = \frac{(b-a)^2}{n}\left\{\mathbb{E}_X\left[(\phi^2(X))\right] - \mathbb{E}_X^2[\phi(X)]\right\} \\
&= \frac{(b-a)^2}{n}\left\{\int_a^b \frac{\phi^2(x)}{b-a}\mathrm{d}x - \left[\int_a^b \frac{\phi(x)}{b-a}\mathrm{d}x\right]^2\right\} \\
&= \frac{(b-a)^2}{n}\left\{\int_a^b \frac{\phi^2(x)}{b-a}\mathrm{d}x - \frac{Y^2}{(b-a)^2}\right\} \\
&\underset{c\phi(x)\geq\phi^2(x)}{\overset{c\geq\phi(x)}{\leq}} \frac{(b-a)^2}{n}\left\{c\int_a^b \frac{\phi(x)}{b-a}\mathrm{d}x - \frac{Y^2}{(b-a)^2}\right\} \\
&= \frac{1}{n}[c(b-a)Y - Y^2] = \mathbb{V}[\widetilde{Y}]
\end{aligned}
$$

Thus, "hit and miss" is worse than "ϕ-f" (has larger variance).

Antithetic R.V.'s

Assume that $\hat{\theta}_1, \hat{\theta}_2$ are 2 unbiased estimators of θ with variances $\mathbb{V}[\hat{\theta}_1]$ and $\mathbb{V}[\hat{\theta}_2]$ respectively. Let $\hat{\theta} = \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)$. Then, $\mathbb{E}[\hat{\theta}] = \theta$ and

$$\mathbb{V}\left[\frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)\right] = \frac{1}{4}\mathbb{V}[\hat{\theta}_1] + \frac{1}{4}\mathbb{V}[\hat{\theta}_2] + \frac{1}{2}Cov(\hat{\theta}_1, \hat{\theta}_2)$$

If further $\mathbb{V}[\hat{\theta}_1] = \mathbb{V}[\hat{\theta}_2]$ then

$$\begin{aligned}\mathbb{V}\left[\frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)\right] &= \frac{1}{2}\mathbb{V}[\hat{\theta}_1] + \frac{1}{2}Cov(\hat{\theta}_1, \hat{\theta}_2) \\ &= \frac{1}{2}\mathbb{V}[\hat{\theta}_1](1 + Corr(\hat{\theta}_1, \hat{\theta}_2))\end{aligned}$$

Consequently, if $Corr(\hat{\theta}_1, \hat{\theta}_2)$ is negative (thus $\hat{\theta}_1, \hat{\theta}_2$ are **antithetic** r.v.'s (same distribution but negatively correlated)) then $\mathbb{V}[\hat{\theta}] \leq \frac{1}{2}\mathbb{V}[\hat{\theta}_1]$.

Assuming we roughly need double the time now to generate both $\hat{\theta}_1$ and $\hat{\theta}_2$, this is an overall reduction of variance in the same computational time.

Antithetic R.V.'s (cont'd)

In the standard Monte Carlo setting, when $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$, where $\{X_i\}_{i\in\{1,\ldots,n\}}$ are i.i.d. r.v.'s. antithetic random variables can be used when the distribution of X has some symmetry: $R(X) \sim X$, for some transformation R.

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} f\big(R(X_i)\big)$$

If $\mathrm{Cov}\big(X, R(X)\big) < 0$, using $R(X)$ as an antithetic variable reduces variance to a certain extent.

A typical example is to take $R(X) = -X$, when the distribution of X is symmetric around zero, as the next example shows.

Example on Antithetic R.V.'s

$$\theta = \int_{-\infty}^{\infty} \underbrace{\frac{x}{2^x - 1}}_{\phi} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}_{f} \mathrm{d}x$$

- Standard Monte-Carlo Method

  Take a sample of let's say $n = 1000$ $\mathcal{N}(0,1)$ variates $X_1, \ldots X_{1000}$ and take $\hat{\theta}_{\mathrm{MC}} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.

- Antithetic sample

  $\hat{\theta}_{\mathrm{AS}} = \frac{1}{n}\sum_{i=1}^{500} [\phi(x_i) + \phi(-x_i)]$

  $\widehat{Corr}\{\phi(x_i), \phi(-x_i)\} = -0.95, \quad i = 1, \ldots 500$

  $\hat{\theta}_{\mathrm{MC}} = 1.4993 \rightarrow s.e. = 0.016$

  $\hat{\theta}_{\mathrm{AS}} = 1.4992 \rightarrow s.e. = 0.0035$

Importance Sampling

Let $X \sim f(x)$, (f: pdf) and $X \in \mathcal{X}$. Like before we would like to estimate

$$\theta = \int_{\mathcal{X}} \phi(x) f(x) \mathrm{d}x = \int_{\mathcal{X}} h(x) \mathrm{d}x = \mathbb{E}_f[\phi(X)]$$

Let g be another pdf that is strictly positive when $h(x) = \phi(x)f(x)$ is different from zero, for $x \in \mathcal{X}$. Let $\psi(x) = \frac{\phi(x)f(x)}{g(x)}$, $x \in \mathcal{X}$. Then,

$$\theta = \int_{\mathcal{X}} \frac{\phi(x)f(x)}{g(x)} g(x) \mathrm{d}x = \mathbb{E}_g[\psi(X)]$$

Therefore we either generate a sample from f and estimate θ from $\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) = \hat{\theta}_f$ (ϕ-f method) or we generate a sample from g and estimate θ from $\frac{1}{n}\sum_{i=1}^{n} \psi(X_i) = \hat{\theta}_g$.

$$\mathbb{E}[\hat{\theta}_g] = \frac{1}{n}\mathbb{E}_g\left[\sum_{i=1}^{n} \psi(X_i)\right] = \mathbb{E}_g[\psi(X)] = \theta, \text{ i.e. unbiased}$$

Importance Sampling (cont'd)

$$\begin{aligned}\mathbb{V}[\hat{\theta}_g] &= \frac{1}{n}\int_{\mathcal{X}}[\psi(x)-\theta]^2 g(x)\mathrm{d}x = \frac{1}{n}\int_{\mathcal{X}}[\phi(x)f(x)/g(x)-\theta]^2 g(x)\mathrm{d}x \\ &= \frac{1}{n}\int_{\mathcal{X}}[h(x)/g(x)-\theta]^2 g(x)\mathrm{d}x = \frac{1}{n}\mathbb{V}_g\left[\frac{h(X)}{g(X)}\right].\end{aligned}$$

Thus, the variance of $\hat{\theta}_g$ becomes zero when $\frac{h(x)}{g(x)} = \theta,\ x \in \mathcal{X}$, but since this requires apriori knowledge of θ, we choose g to make $\phi(x)f(x)/g(x)$ nearly constant. We can show that $\mathbb{V}[\hat{\theta}_g]$ is minimized by $g(x) \propto |\phi(x)f(x)| \equiv |h(x)|$. For this reason, the method is good when $g(x) \propto |\phi(x)f(x)|,\ x \in \mathcal{X}$.

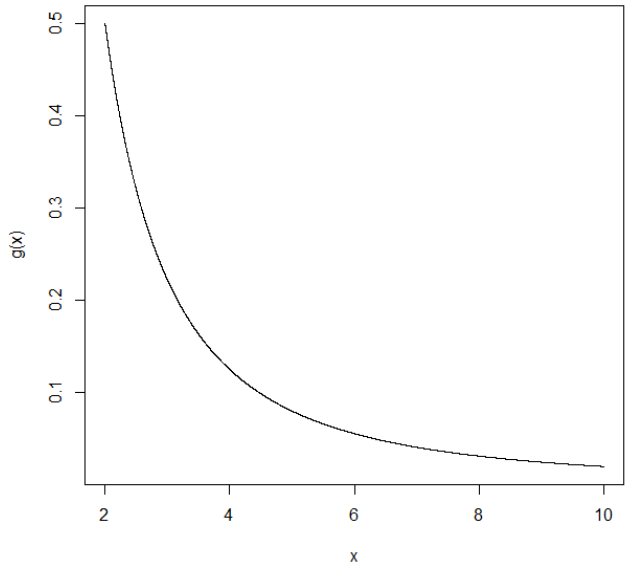
Importance Sampling: Example

Let X be a random variable in $\mathbb{R}$ following the Cauchy distribution. We would like to estimate

$$\theta = \mathbb{P}[X > 2] = \int_2^\infty \frac{1}{\pi} \frac{1}{1+x^2} \mathrm{d}x.$$

We choose $g(x) = 2/x^2,\ x \in (2, \infty) \ \rightarrow \ g > 0$. The function g is mimicking the Cauchy shape in $(2, \infty)$ and further

$$\int_2^\infty \frac{2}{x^2} \mathrm{d}x = \frac{-2}{x}\Big|_2^\infty = 1 \ \rightarrow \ g \text{ is a pdf in } (2, \infty)$$



Importance Sampling: Example (cont'd)

We can simulate values from g with the inversion method:

$$G(x) = \int_2^x \frac{2}{t^2}\mathrm{d}t = 1 - \frac{2}{x} = 1 - u \Rightarrow x = \frac{2}{u}$$

Thus $U \sim U(0,1) \Rightarrow 1 - U \sim U(0,1) \Rightarrow X = 2/U \sim g$. Note also that $X^{-1} \sim U(0, 1/2)$. So:

$$\begin{aligned}
\theta &= \int_2^\infty \frac{1}{\pi}\frac{1}{1+x^2}\mathrm{d}x = \int_2^\infty \frac{x^2}{2\pi(1+x^2)}\frac{2}{x^2}\mathrm{d}x \\
&= \mathbb{E}_g\left[\frac{X^2}{2\pi(1+X^2)}\right] = \int_2^\infty \frac{1}{\pi}\frac{1}{1+x^2}\mathrm{d}x \underset{\text{inversion method}}{\overset{x^2=y^{-2}}{=}} \\
&\int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})}\mathrm{d}y = \int_0^{1/2} \frac{1}{\pi(1+y^2)}\mathrm{d}y = \mathbb{E}_h\left[\frac{1}{2\pi(1+X^2)}\right],
\end{aligned}$$

where $h(x) = 2,\ x \in (0, 1/2)$, i.e. the pdf of the $U(0, 1/2)$.

Importance Sampling: Exercise

Let X be a random variable in $\mathbb{R}$ following the Cauchy distribution (symmetric around zero). We would like to estimate

$$\theta = \mathbb{P}[X > 2] = \int_2^\infty \frac{1}{\pi} \frac{1}{1+x^2} \mathrm{d}x.$$

We can find θ using the cdf of the Cauchy distribution. We have that $\theta = 1 - F(2) = 1/2 - \pi^{-1} \arctan 2 = 0.1476$. We will now use simulation methods to estimate θ and compare the variances of our estimators.

1.

$$f(x) = \frac{1}{\pi(1+x^2)} \text{ (Cauchy)}, \quad \phi(x) = \mathbb{1}_{\{x>2\}}, \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$$

We thus generate n Cauchy values and $\hat{\theta}$: proportion of values > 2. Then,

$$\mathbb{V}[\hat{\theta}] = \frac{\theta(1-\theta)}{n} = \frac{0.126}{n} \quad \left(n\hat{\theta} = \sum_{i=1}^{n} \phi(X_i) \sim \text{Bin}(n, \theta) \right)$$

Importance Sampling: Exercise (cont'd)

2.

$$\theta = \frac{1}{2}\mathbb{P}[|X| > 2], \quad f(x) = \frac{1}{\pi(1+x^2)}, \quad \phi(x) = \mathbb{1}_{\{|x|>2\}}$$

A moment's reflection reveals that this trick is essentially using the method of antithetic r.v.'s. Since the Cauchy distribution is symmetric around zero, using $-X$ as an anithetic variable we have

$$\frac{1}{2}\Big(\mathbb{1}\{X > 2\} + \mathbb{1}\{-X > 2\}\Big) = \frac{1}{2}\mathbb{1}\{|X| > 2\}$$

We thus generate n Cauchy values and $\hat{\theta}$: $1/2$ the proportion of values > 2 in absolute value.

$$2n\hat{\theta} \sim \text{Bin}(n, 2\theta) \Rightarrow \mathbb{V}[2n\hat{\theta}] = n2\theta(1-2\theta) \Rightarrow 4n^2\mathbb{V}[\hat{\theta}] = n2\theta(1-2\theta)$$

$$\Rightarrow \mathbb{V}[\hat{\theta}] = \frac{2\theta(1-2\theta)}{4n} \approx \frac{0.052}{n} \text{ (2.4 times reduction)}$$

Exercise (cont'd)

3.

$$1-2\theta=\int_{-2}^{2}\frac{1}{\pi(1+x^2)}\mathrm{d}x=2\int_{0}^{2}\frac{1}{\pi(1+x^2)}\mathrm{d}x$$

$$\Rightarrow\theta=\frac{1}{2}-\int_{0}^{2}\frac{1}{\pi(1+x^2)}\mathrm{d}x$$

To estimate the integral we take

$$f(x)=\frac{1}{2},\ x\in(0,2),\ \text{i.e. the pdf of } U(0,2),\quad \phi(x)=\frac{2}{\pi(1+x^2)}$$

We generate values $X_1,\dots X_n\sim U(0,2)$ and $\hat{\theta}=\frac{1}{2}-\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$

$$\begin{aligned}\mathbb{V}[\hat{\theta}] &= \frac{1}{n}\mathbb{V}_f[\phi(X)]=\frac{1}{n}\int_0^2[\phi(x)-(\tfrac{1}{2}-\theta)]^2 f(x)\mathrm{d}x\\ &= \frac{1}{2n}\int_0^2\left[\frac{2}{\pi(1+x^2)}-0.3524\right]^2\mathrm{d}x=\frac{0.028}{n}\end{aligned}$$

(1.85 times further reduction)

Exercise (cont'd)

4. Let $y = x^{-1}$. Then,

$$\theta = \int_2^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} \mathrm{d}x = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})} \mathrm{d}y = \int_0^{1/2} \frac{1}{\pi(1+y^2)} \mathrm{d}y$$

$f(x) = 2,\ x \in (0, 1/2)$, i.e. the pdf of $U(0, 1/2)$, $\quad \phi(x) = \dfrac{1}{2[\pi(1+x^2)]}$

We generate values $X_1, \ldots X_n \sim U(0, 1/2)$ and $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$

$$\mathbb{V}[\hat{\theta}] = \frac{2}{n} \int_0^{1/2} \left[\frac{1}{2\pi(1+x^2)} - 0.1476 \right]^2 \mathrm{d}x = \frac{0.0000955}{n}$$

(importance sampling $\rightarrow$ 293 times further reduction)

Importance Sampling for Rare Events

One should **always** consider importance sampling techniques when simulating rare events. For instance, let's suppose we wish to estimate the integral

$$I := \int_5^\infty e^{x-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = \mathbb{E}_Z\left[e^Z \mathbb{1}\{Z > 5\}\right], \quad Z \sim N(0,1).$$

Because $p = \mathbb{P}[Z > 5] \simeq 2.87 \times 10^{-7}$, it would take roughly $1/p$ samples to find one that contributes to the integral! Here, $I \simeq 5.22 \times 10^{-5}$ and $\mathbb{V}(\hat{\theta}) \simeq \frac{1}{100n}$. We need roughly 1.6×10^9 samples for 10% accuracy.

With importance sampling, we try to make use of a variable that typically takes values > 5, such as $Y \sim \mathcal{N}(5,1)$ for instance. Then:

$$\int_5^\infty e^{x-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = e^{25/2} \int_5^\infty e^{-4x} e^{-\frac{(x-5)^2}{2}} \frac{dx}{\sqrt{2\pi}} = e^{25/2}\mathbb{E}_Y\left[e^{-4Y}\mathbb{1}\{Y > 5\}\right].$$

Now $\mathbb{V}(\hat{\theta}_{\text{IS}}) \simeq \frac{1.23\times 10^{-8}}{n}$. We need roughly 2000 samples for the same accuracy, or even < 1000 if we use $W = 5 - Y$ as an antithetic r.v.

Dependent R.V.'s

- In several applications, instead of working with an independent sample $X_1, \dots X_n$ we use the dependent ordered sample $X_{(1)}, X_{(2)} \dots X_{(n)}$. The simplest approach is to generate $X_1, \dots X_n$ and subsequently sort them with the use of a sorting algorithm. (For large n though, the computational cost is large.)
- By using the inversion method $X_{(i)} = F_X^{-1}\left(U_{(i)}\right), U_{(1)} < U_{(2)} \dots < U_{(n)}$, thus the problem is transferred to the generation of ordered values from $U(0,1)$.
- Method of sequence

  $U_1, \dots U_n$ independent sample $U(0,1)$. We set

$$\begin{aligned} U_{(n)} &= U_n^{1/n} \\ U_{(n-1)} &= U_{(n)} \times U_{n-1}^{1/(n-1)} \\ &\vdots \quad\quad \vdots \\ U_{(k)} &= U_{(k+1)} \times U_k^{1/k} \\ &\vdots \quad\quad \vdots \\ U_{(1)} &= U_{(2)} \times U_1^{1/1} \end{aligned}$$

  Then, $U_{(1)}, \dots U_{(n)}$ ordered sample from $U(0,1)$.

Multivariate Normal Distribution

$\mathbf{X} = (X_1, \dots X_n) \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V})$

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu})\},$$

$$\boldsymbol{\mu} = (\mu_1, \dots \mu_n)^T, \quad \mathbf{V}_{n \times n} = (V_{ij})$$

$$\mathbb{E}[X_i] = \mu_i, \quad V_{ii} = \mathbb{V}[X_i], \quad V_{ij} = Cov(X_i, X_j)$$

Simulation of $\mathbf{X}$ is achieved by transforming it into a standard multivariate normal distribution. We set

$$\mathbf{Y} = \mathbf{L}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \text{, where } \mathbf{L} \text{ (lower triangular)} : \mathbf{L}\mathbf{L}^T = \mathbf{V}$$

Then $\mathbf{Y} \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{n \times n}) \Rightarrow Y_i \sim \mathcal{N}(0, 1), \quad i = 1, \dots n$

$\Leftarrow$

if independent

Thus, generating $\mathbf{Y}$ is simple. If we then set $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Y}$, we get values from $\text{MVN}(\boldsymbol{\mu}, \mathbf{V})$. To compute $\mathbf{L} \rightarrow$ Cholesky (beware: the R function chol() returns the upper triangular L^T matrix)

Travelling Salesman Problem

n cities and $d(i,j)$: cost travelling from city i to city j. One salesman will visit each city once. Which is the cheapest possible route?

There are $n!$ possible routes and denote by $c(\mathbf{x})$ the cost of the $\mathbf{x}$ route, $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_{n!}\}$

We would like to minimize $c(\mathbf{x})$ with respect to $\mathbf{x}$, and the complete search is not possible (e.g. $n = 100$) (stochastic (or combinatorial) optimization).

Trick: Define $p_\lambda(\mathbf{x}) = \dfrac{e^{-\lambda c(\mathbf{x})}}{\sum_{\mathbf{x}} e^{-\lambda c(\mathbf{x})}} = K e^{-\lambda c(\mathbf{x})}$

Then, $p_\lambda(\mathbf{x})$ is a joint pmf in $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_{n!}\}$.

Note that for large λ, $\mathbf{x}$ with large $c(\mathbf{x})$ give negligible $p_\lambda(\mathbf{x})$. Thus, the $\mathbf{x}$ that are generated from $p_\lambda(\mathbf{x})$ for large λ are those $\mathbf{x}$ that minimize $c(\mathbf{x})$ (work with $e^{-\lambda c(\mathbf{x})} \propto p_\lambda(\mathbf{x})$).

Rao-Blackwellization

- What if we only care about $\mathbb{E}_X[h(X)]$ when our sampling method produces (X, Y)? Naive method is to throw out Y and estimate the expectation by $\delta = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$.
- e.g. Y are samples from g in rejection sampling and X are samples that pass the acceptance test (X depends on Y and on some other r.v.'s that are integrated out).
- Rao-Blackwellization is a method to produce **lower-variance** using the following formula (*Law of Total Variance*):

$$\begin{aligned} \mathbb{V}_X[\delta] &= \mathbb{E}_Y\left[\mathbb{V}_{X|Y}[\delta|Y]\right] + \mathbb{V}_Y\left[\mathbb{E}_{X|Y}[\delta|Y]\right] \Rightarrow \\ \mathbb{V}_X[\delta] &\geq \mathbb{V}_Y\left[\mathbb{E}_{X|Y}[\delta|Y]\right] \end{aligned}$$

- Thus if $\mathbb{E}_X[h(X)]$ is the quantity we wish to estimate, then we can use $\mathbb{E}_{X|Y}[\delta|Y]$ instead of δ to produce better estimator (with lower variance).
- The two estimators have the same bias (*Law of Total Expectation*):

$$\mathbb{E}_X[\delta] = \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[\delta|Y]\right]$$

Rao-Blackwellization (cont'd)

- Let X be a random variable with pdf $f_X(x)$, $x \in \mathcal{X}$ and $\mathbf{X} = (X_1, X_2 \ldots X_n)$ be a random sample from $f_X(x)$. Let Y be another random variable (on the same probability space as X) with pdf $f_Y(y)$, $y \in \mathcal{Y}$ and $\mathbf{Y} = (Y_1, Y_2 \ldots Y_n)$ be a random sample from $f_Y(y)$.
- Rao-Blackwellization means taking advantage of the fact that, if $T(\mathbf{X})$ is an estimator of $\theta = \mathbb{E}_X[h(X)]$ and if $\mathbf{X}$ can be simulated from the joint distribution $f_{X,Y}(x, y)$ satisfying

$$\int_{\mathcal{Y}} f_{X,Y}(x, y)dy = f_X(x),$$

then the new estimator $T^*(\mathbf{Y}) = \mathbb{E}_{X|Y}[T(\mathbf{X})|\mathbf{Y}]$ dominates $T(\mathbf{X})$ in terms of variance, while the bias is the same. Obviously, this result is only useful in settings where $T^*(\mathbf{Y})$ can be explicitly computed.

Rao-Blackwellization: Example

- Consider computing the expectation of $h(x) = \exp(-x^2)$ when $X \sim St(\nu, \mu, \sigma^2)$ (scaled Student distribution). If $X \sim St(\nu, \mu, \sigma^2)$, then:

$$X = \mu + \sigma \frac{\epsilon}{\sqrt{\frac{\xi}{\nu}}}, \text{ with } \epsilon \sim N(0,1) \text{ and } \xi \sim X^2_\nu.$$

- Even though the Student distribution can be simulated directly using the function `rt()` in R, it allows for the marginal representation above in terms of the joint distribution on (x, ξ), or, equivalently on (x, y), where $y = \xi/\nu$ $(\sim X^2_\nu/\nu \equiv Gamma(\nu/2, \nu/2))$.

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\sqrt{y}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2 y}{2\sigma^2}\right) \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} y^{\nu/2-1} \exp(-y\nu/2) \\ &= f_{X|Y}(x|y) \times f_Y(y) \\ &= N(x|\mu, \sigma^2/y) \times Gamma(y|\nu/2, \nu/2) \end{aligned}$$

- In R we produce a sample of (X_i, Y_i):

```
> y = sqrt(rchisq(Nsim, df=nu)/nu)
> x = rnorm(Nsim, mu, sigma/y)
```

Rao-Blackwellization: Example (cont'd)

- In the above we use in the normal distribution the standard deviation and not the variance (thus we take sample of $\sqrt{Y}$).
- Therefore the usual estimate

$$\delta_n = \frac{1}{n}\sum_{i=1}^{n} \exp(-X_i^2)$$

can be improved using the Rao-Blackwellized version

$$\begin{aligned} \delta_n^* &= \mathbb{E}_{X|Y}\left[n^{-1}\sum_{i=1}^{n}\exp(-X_i^2)|Y_i\right] \\ &\overset{\substack{X_i \text{ i.i.d.} \\ \text{call them } X}}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X|Y}[\exp(-X^2)|Y_i] \end{aligned}$$

- We can show that

$$\int_{-\infty}^{+\infty} N(x|\mu_1,\sigma_1^2)N(x|\mu_2,\sigma_2^2)dx = N(\mu_1|\mu_2,\sigma_1^2+\sigma_2^2) = N(\mu_2|\mu_1,\sigma_1^2+\sigma_2^2)$$

Rao-Blackwellization: Example (cont'd)

- Therefore

$$\begin{aligned}\mathbb{E}_{X|Y}[\exp(-X^2)|y] &= \sqrt{\pi}\int_{-\infty}^{+\infty} N(x|0,1/2)N(x|\mu,\tfrac{\sigma^2}{y})dx \\ &= \frac{1}{\sqrt{2\sigma^2/y+1}}\exp\left(-\frac{\mu^2}{2\sigma^2/y+1}\right)\end{aligned}$$

- Thus

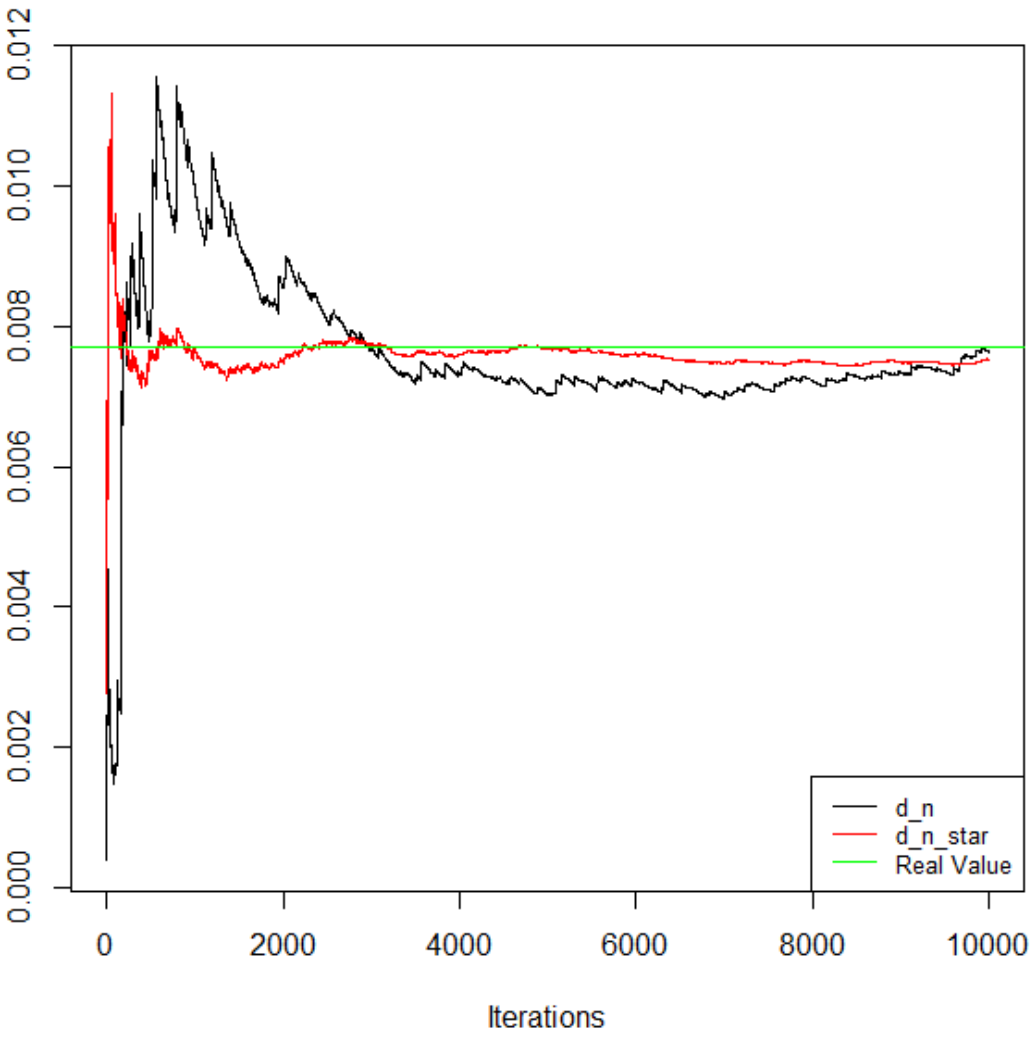
$$\delta_n^* = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\sigma^2/Y_i+1}}\exp\left(-\frac{\mu^2}{2\sigma^2/Y_i+1}\right)$$

- In R we type (we square $\sqrt{Y}$):

```
> d_n=cumsum(exp(-x^2))/(1:Nsim)
> d_n_star=cumsum(exp(-mu^2/(1+2*(sigma/y)^2))/
                  sqrt(1+2*(sigma/y)^2))/(1:Nsim)
```

Rao-Blackwellization: Example (cont'd)

- For `Nsim`=10000 and $(\nu, \mu, \sigma) = (5, 3, 0.5)$ we get



Appendix 1

Let $X, Y \sim N(0,1)$, independent. Then the joint pdf of (X, Y) is

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, \quad x, y \in \mathbb{R}.$$

We define the following r.v.'s

$$R = \sqrt{X^2 + Y^2}, \quad \Theta = \arctan Y/X, \quad R \in \mathbb{R}^+, \ \Theta \in [0, 2\pi]$$

Then

$$x = r\cos\theta, \quad y = r\sin\theta,$$

and

$$J(r,\theta) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r$$

Thus

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta)|r| = re^{-\frac{r^2}{2}} \frac{1}{2\pi}$$

Therefore R, Θ independent and

$R \sim$ Rayleigh(1) and $\Theta \sim U(0, 2\pi) \Rightarrow R^2 \sim \text{Exp}(1/2) \equiv X_2^2$ and $\Theta \sim U(0, 2\pi)$