Brief Introduction to Probability and Simulation: Part 2 — Monte Carlo and Importance Sampling

Elaine Spiller

Marquette University

Data Assimilation Research Programme: Monsoon Summer School

July 7, 2011

Suppose $X_1, X_2, \ldots$ are independent rvs taken from a distribution $f(x)$. Given a function $g(x)$, define

$$G = \frac{1}{N}\sum_{n=1}^{N} g(X_n)$$

$$E[G] = E[\frac{1}{N}\sum_{n=1}^{N} g(X_n)] = \frac{1}{N}\sum_{n=1}^{N} E[g(X)] = E[g(X)]$$

And

$$var\{G\} = \sum_{n=1}^{N} \frac{1}{N^2} var\{g(X)\} = \frac{1}{N} var\{g(X)\}$$

So, as $N \to \infty$, $var\{G\} \to 0$.

Recall, $$E_f[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

And, we just saw $E[G] = E[g]$, so we have the basis for Monte Carlo integration

$$\int_{-\infty}^{\infty} g(x)f(x)dx \approx \frac{1}{N}\sum_{n=1}^{N} g(X_n) \qquad \text{where } X_n \sim f.$$

Another common case: $g(x) = 1$

$$P[X > a] = \int_{a}^{\infty} f(x)dx \approx \frac{1}{N}\sum_{n=1}^{N} I(X_n > a)$$

where I is called and *indicator function*, that is

$$I(X > a) = \begin{cases} 1 & \text{for } X > a \\ 0 & \text{for } X < a. \end{cases}$$

How to justify this

Law of Large Numbers:

$$\text{If} \quad \bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n, \quad \text{then} \quad P\{\lim_{N\to\infty} \bar{X}_N = E[X]\} = 1$$

$$\text{So,} \quad G \xrightarrow[n\to\infty]{} \int_{-\infty}^{\infty} g(x)f(x)dx$$

If we take enough samples, Monte Carlo (MC) integration will **always** converge.

Question: How much is *enough*?

$$P\Big\{|G - E[G]| \geq \Big[\frac{var\{G\}}{\delta}\Big]^{1/2}\Big\} \leq \delta$$

This could be called the Fundamental Theorem of Monte Carlo methods because it estimates the probably of a large deviation of an MC calculation.

Rewriting, we have

$$P\Big\{\Big(\frac{1}{N}\sum_{n=1}^{N} g(X_n) - \int_{-\infty}^{\infty} g(x)f(x)dx\Big)^2 \geq \frac{var\{g(X)\}}{\delta N}\Big\} \leq \delta.$$

In words, this says that the probability that a sample calculation & the exact solution differ by $\sqrt{\frac{1}{\delta N} var\{g(x)\}}$ is no more than δ.

Some consequences of Chebyshev's Inequality

For a specified "certainty", δ, there are only two ways to the magnitude of the error

$$\text{error}^2 = \frac{var\{g(X)\}}{\delta N}$$

- increase the number of MC samples, N
- decrease $var\{g(X)\}$

If you think about MC as a method to calculate an integral

$$\int_a^b h(x)dx,$$

you get to choose how to "break up" h into $h(x) = g(x)f(x)$.

Monte Carlo example

How many samples, N, must we take to have a 0.99 probability of calculating the integral

$$\int_0^2 x^3 dx$$

withing an error of 0.1?

Note, probability (certainty) of 0.99 means $\delta = 0.01$.

To do this, we need to write $x^3 = g(x)f(x)$ where $\int_0^2 f(x)dx = 1$.

Let's do two cases

1. $f(x)$ Uniform, so $f(x) = \frac{1}{2} \quad 0 < x < 2$ and $g(x) = 2x^3$
2. $f(x) = \frac{3}{8}x^2$ and $g(x) = \frac{8}{3}x$

MC example continued

Really what we are asking for is

$$\frac{1}{\delta N} var\{g(x)\} < \text{error}^2 < (\frac{1}{10})^2.$$

So, we need to calculate $var\{g(x)\}$ and solve for N.

For our two cases, we'll see

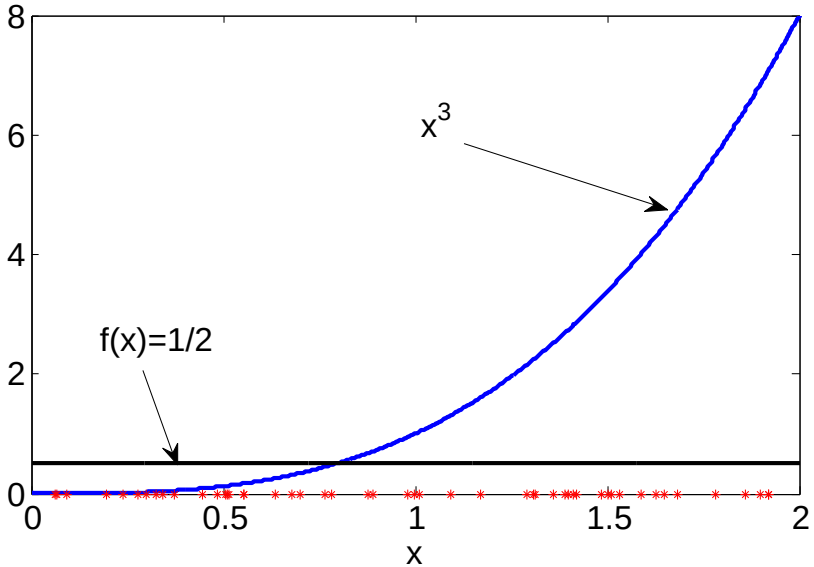
1. $N \approx 20x10^4$
2. $N \approx 1x10^4$

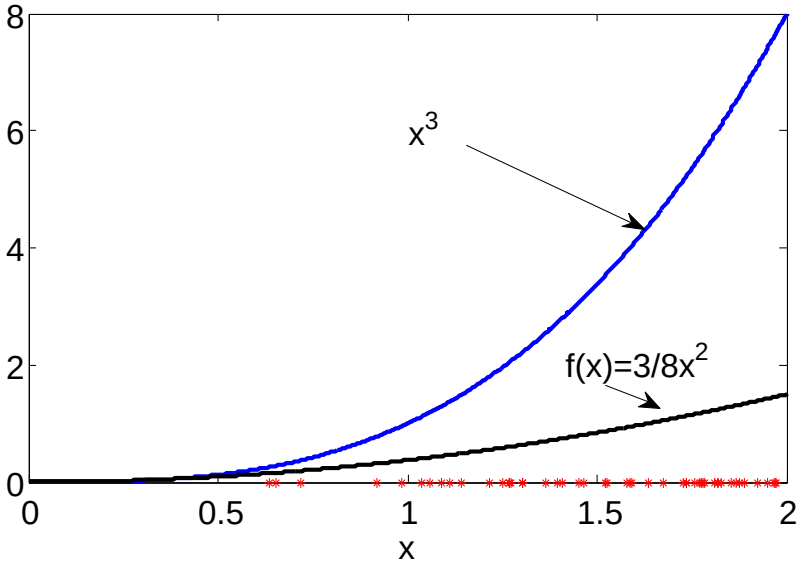
Questions:

Why is this happening?

What if we had a tolerance of 0.01?

What's going on in MC example





Moral of the story:

We want to sample the domain where the integrand is big.

The magnitude of the variance of $g(X)$ in some sense tells us "how well" we're accomplishing that goal.

Pros and cons of MC

$$\text{error}^2 = \frac{var\{g(X)\}}{\delta N}$$

Cons:

- quite ineficient compared to grid-based methods, $O(1/\sqrt{N}) = O(\sqrt{\Delta x})$
- $f(x)$ can be hard to sample

Pros:

- error is *independent* of dimension
- we have some freedom to reduce $var\{g(X)\}$
- natural framework for inherently stochastic problems

Importance sampling: why?

Recall, the key to efficient Monte Carlo algorithms is a reduction in the variance of the estimator,

$$G = \frac{1}{N} \sum_{n=1}^{N} g(X_n)$$

The greater the variance, the more sample points, N, needed to (accurately) esitmate the quantity of interest.

Suppose we wish to integrate

$$E_f[g] = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

we could estimate this with samples from *any* probability distribution we like.

Importance sampling: idea

Let's consider $\tilde{f}(x) > 0$ and rewrite our integral as

$$E_f[g] = \int_{-\infty}^{\infty} \frac{g(x)f(x)}{\tilde{f}(x)} \tilde{f}(x)dx.$$

At this point, we could sample from $f(x)$ or $\tilde{f}(x)$, but if we sample $\tilde{f}(x)$, then

$$\tilde{g}(x) = \frac{g(x)f(x)}{\tilde{f}(x)}$$

and the MC error is determined by

$$var\{\tilde{g}\} = \int_{-\infty}^{\infty} \Big[\frac{g^2(x)f^2(x)}{\tilde{f}^2(x)}\Big] \tilde{f}(x)dx - E^2_{\tilde{f}}[\tilde{g}].$$

Note, $E_{\tilde{f}}[\tilde{g}] = E_f[g] =$ some (unkown) constant. And recall, we'd like to minimize $var\{\tilde{g}\}$.

Importance sampling: how to choose $\tilde{f}$?

To minimize $var\{\tilde{g}\}$, we want an $\tilde{f}$ such that

$$\int_{-\infty}^{\infty} \left[\frac{g^2(x)f^2(x)}{\tilde{f}^2(x)}\right]\tilde{f}(x)dx \quad \text{is minimized, subject to} \quad \int_{-\infty}^{\infty} \tilde{f}(x)dx = 1.$$

This can be solved via Lagrange multipliers, resulting in an optimal $\tilde{f}$ of

$$\tilde{f}(x) \propto g(x)f(x)$$

- Does this result seem reasonable?
- Is it useful???

Optimal biasing distribution, $\tilde{f}$

Is $\tilde{f} \propto g(x)f(x)$ reasonable ?

- Yes, choosing this $\tilde{f}$, we get $var\{\tilde{g}\} = 0$. This is the best we can do!
- No, to sample from $\tilde{f}$, we need to normalize it by

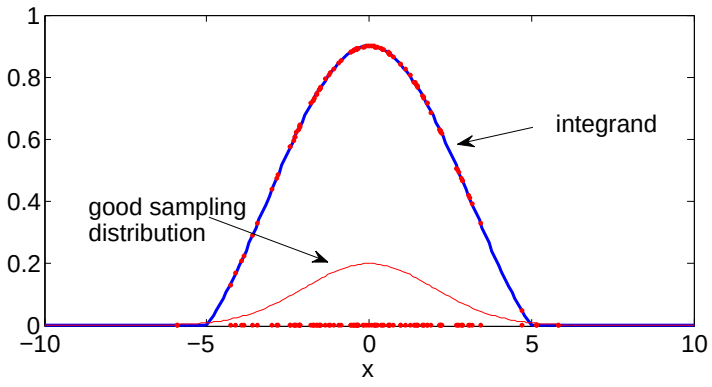
$$\tilde{f}(x) = \frac{f(x)g(x)}{\int_{-\infty}^{\infty} f(x)g(x)dx}.$$

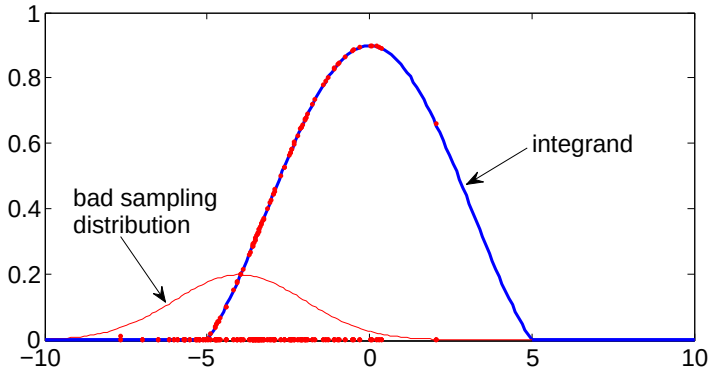
  But that normalization factor is exactly the integral we'd like to approximate!

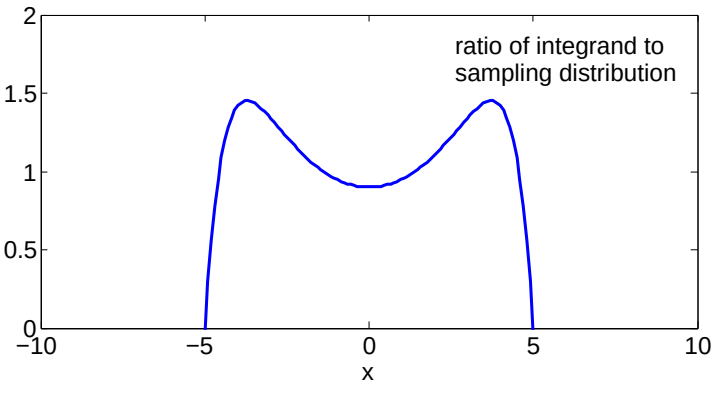
Is $\tilde{f} \propto g(x)f(x)$ useful?

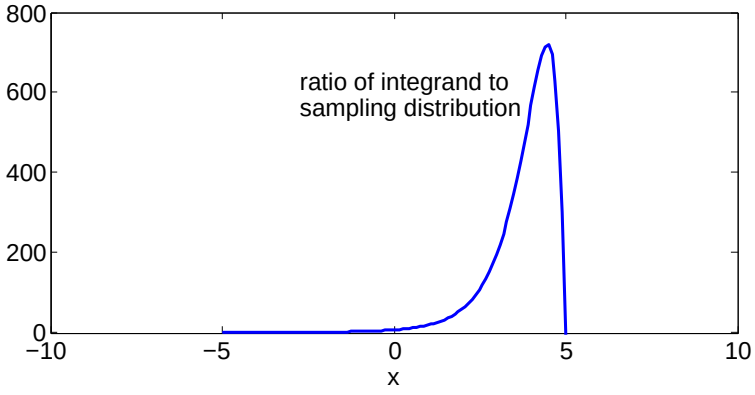
- Yes, the general goal is to make the importance distribuiton $\tilde{f}$ "look like" the integrand, $f(x)g(x)$

Good vs.bad importance distribution









Imporance sampling example

Consider the integral

$$E[g(x)] = \int_0^1 \cos(\frac{\pi}{2}x)dx$$

If we pick $\tilde{f}$ to be $U(0,1)$, the we get a variance of

$$\int_0^1 \cos^2(\frac{\pi}{2}x)dx - E^2[g] \approx 0.09472$$

Recall, we want $\tilde{f}(x) \approx \cos(\frac{\pi}{2}x)$, we could choose a $\tilde{f}$ by expanding

$$\cos(\frac{\pi}{2}x) = 1 - \frac{\pi^2}{8}x^2 + \frac{\pi^4}{2^4 4!}x^4 + \dots$$

This suggests
$\tilde{f}(x) = \alpha(1 - \frac{\pi}{8}x^2)$ with α chosen so $\int_0^1 (1 - \frac{\pi}{8}x^2)dx = 1/\alpha$

IS example continued

Problem, $\tilde{f}(x) = \alpha(1 - \frac{\pi}{8}x^2) < 0$ on $0 < x < 1$.

So let's try

$$\tilde{f}(x) = \alpha(1 - x^2) \qquad \text{with } \alpha = 3/2.$$

Computing the variance for the resulting $\tilde{g}$, we get

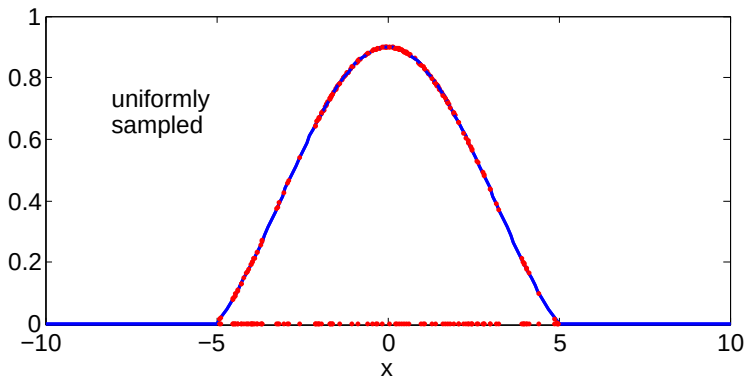
$$\int_0^1 \frac{\cos^2(\frac{\pi}{2})x}{\frac{3}{2}(1 - x^2)} dx - \frac{4}{\pi^2} \approx 0.000990$$

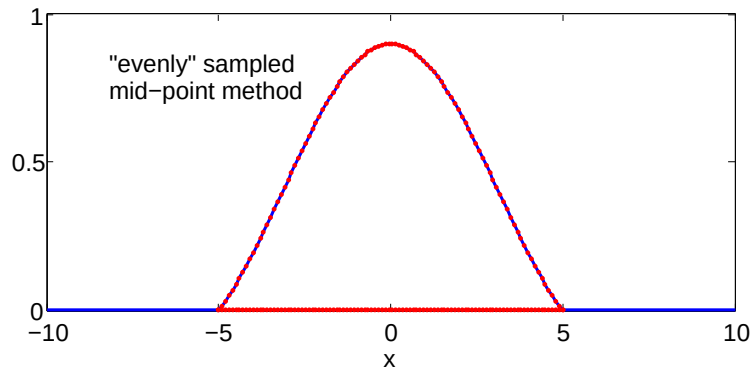
So by choosing a better $\tilde{f}$, we have succeeded in reduce that variance of our estimator by a factor of 100.

Quadrature methods: MC vs. Trapezoid

Both cases looking to approximation

$$I = \int_a^b f(x)g(x)dx$$





Averaged over random points, e.g. $X_n \sim U(a, b)$

$$I \approx \frac{b-a}{N} \sum_{n=1}^{N} g(X_n) f(X_n)$$

Averaged over points sampled on a grid , e.g., $x_n = a + n\Delta x$

$$I \approx \frac{b-a}{N} \sum_{n=1}^{N} g(x_n) f(x_n)$$

Recall, we are looking to estimate

$$E_f[g] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

In practice, we sample X_n from $\tilde{f}$ and we have an estimate to the mean of g,

$$E_f[g] = E_{\tilde{f}}[\tilde{g}] \approx \frac{1}{N}\sum_{n=1}^{N} g(X_n)\frac{f(X_n)}{\tilde{f}(X_n)}.$$

$\frac{f(X_n)}{\tilde{f}(X_n)}$ is called the likelihood ratio.

In some sense it tells us how likely each realization of X_n would have been if it had come from f instead of $\tilde{f}$.

Note on IS in multiple dimensions

Suppose g is a function of many random variables, say $\mathbf{Y} = (X_1, \ldots, X_k)$ and the $X_k's$ are independent. In this case

$$f(\mathbf{Y}) = \prod_{i=1}^{k} f_i(X_i) \quad \text{and similarly for } \tilde{f}(\mathbf{Y}), \quad \text{so}$$

$$\frac{f(\mathbf{Y})}{\tilde{f}(\mathbf{Y})} = \prod_{i=1}^{k} \frac{f_i(X_i)}{\tilde{f}_i(X_i)}$$

So in words, the likelihood ratio is the product of individual likehood ratios.

(Note, in practice products of "small things" are often numerically unstable.)

Another motivation for importance sampling: Rare Events

Recall MC for our second "type" of problem

$$P(X > a) = \int_a^\infty f(x)dx \approx \frac{1}{N}\sum_{n=1}^{N} I(X_n > a) \quad X_n \sim f$$

- Guaranteed to coverge by the Law of Large Numbers.
- In practice if $a >> \mu$, it will **not** converge.
- A *rare event* is loosy defined to have $P \leq 10^{-6}$

Using importance sampling, we have

$$P(X > a) = \int_a^\infty f(x)dx \approx \frac{1}{N}\sum_{n=1}^{N} I(X_n > a)\frac{f(X_n)}{\tilde{f}(X_n)}$$

with $X_n \sim \tilde{f}$.

Simplest example: 100 coin flips

question: What is P(70 or more heads) ?

answer: $2.4x10^{-13}$

importance sampling: Use weighted coin

$$\begin{aligned} \tilde{p} &= 0.7 \quad \text{for heads} \\ \tilde{p} &= 0.3 \quad \text{for tails} \end{aligned}$$

likelihood ratios for flipping weighted coin:

$$\frac{p}{\tilde{p}} = \begin{cases} 0.5/0.7 & \text{for heads} \\ 0.5/0.3 & \text{for tails} \end{cases}$$

key: correcting with likelihood ratio gives statistics for fair coin

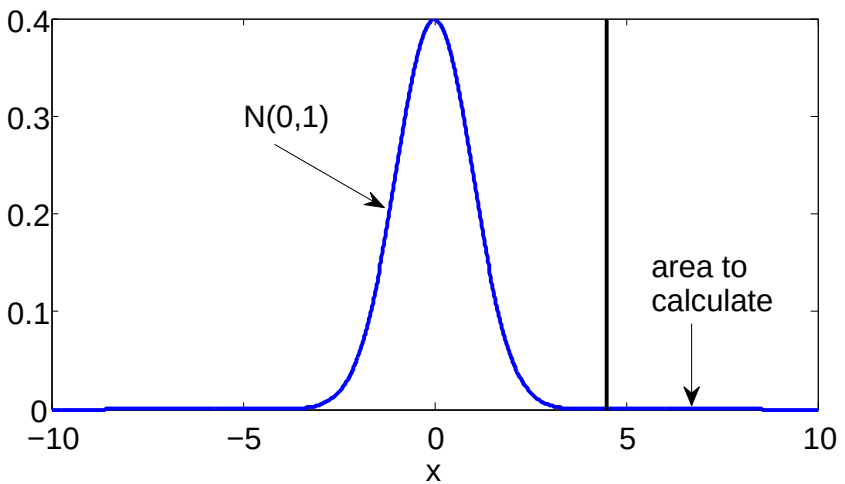
efficiency: over 10 orders of magnitude speed-up

Another rare event example

Say $Z \sim N(0,1)$ and we are interested in $P(Z > 4.5)$. Approximating this with MC gives us

$$P(Z > 4.5) \approx \frac{1}{N} \sum_{n=1}^{N} I(Z_n > 4.5). \qquad Z_n \sim \tilde{f} = N(0,1)$$

Typically $N = 10,000$ samples produces *all zeros* of the indicator function.

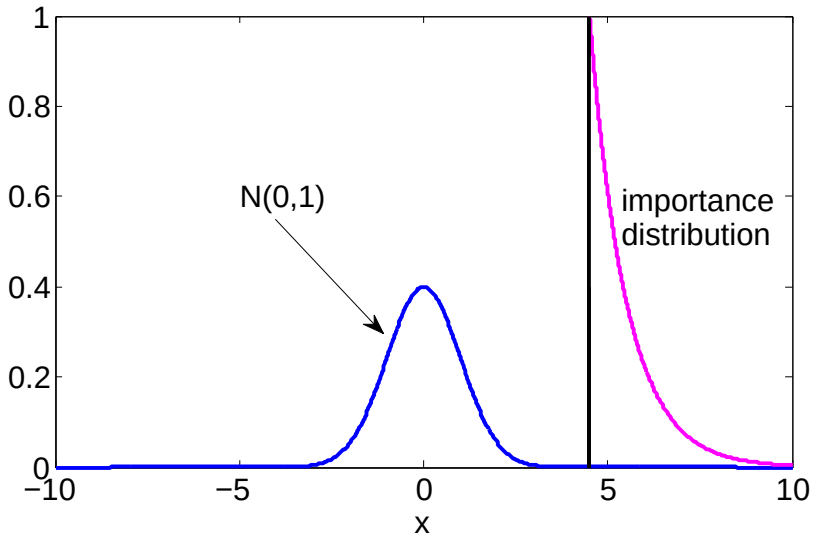


Instead use a shifted Exponential distribution.

Example continued

A good shifted exponential would be

$$\tilde{f}(x) = \frac{e^{-(x-4.5)}}{\int_{4.5}^{\infty} e^{-(x-4.5)} dx} \qquad \text{for } x > 4.5$$



Now if $X \sim \tilde{f}$

$$P(Z > 4.5) \approx \frac{1}{N} \sum_{n=1}^{N} \frac{f(X_n)}{\tilde{f}(X_n)} = 0.000003377$$

Reference Books

- Better for introduction

  Simulation, Ross (third ed. is quite similar to fourth)

- For people with background in prob/stats

  Monte Carlo Statistical Methods, Robert and Casella