Monte Carlo "Swindles"

Charles J. Geyer
School of Statistics
University of Minnesota

Stat 8054 Lecture Notes

GOFMC

Good old-fashioned Monte Carlo (GOFMC), also called independent, identically distributed Monte Carlo (IIDMC), also called ordinary Monte Carlo (OMC) is the practice of using independent and identically distributed (IID) simulations to calculate (estimate, approximate, whatever) integrals that one cannot do by hand or using a computer algebra system.

Theory of GOFMC

Theory of GOFMC is very simple — just elementary statistics!

You want to calculate

$$\mu = E\{g(X)\}$$

but you can't exactly.

Use

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

as your *Monte Carlo approximation*, where X_1, X_2, ... are IID simulations having the same distribution as X.

Theory of GOFMC (cont.)

Think of μ as a *parameter*.

Think of $\widehat{\mu}_n$ as a *statistical estimator* of μ.

Think of n as the *sample size*.

Apply elementary statistical theory!

Terminology

Original Problem Statement

The original problem need not be statistical!

If the original problem is statistical, then μ need not be a parameter in it, and n is certainly not the sample size in it.

Monte Carlo Calculation

We need to be careful. Emphasize that n is the *Monte Carlo sample size*.

Theory of GOFMC (cont.)

The central limit theorem says

$$\widehat{\mu}_n \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

where

$$\sigma^2 = \text{var}\, g(X)$$

can be estimated by

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \Big(g(X_i) - \widehat{\mu}_n\Big)^2$$

That's all the statistical theory we need!

Terminology (cont.)

Original Problem Statement

If the original problem is statistical, then $\sigma/\sqrt{n}$ need not be anything we would call sampling error in it, because n is certainly not the sample size in it.

Monte Carlo Calculation

We need to be careful. Emphasize that $\widehat{\sigma}_n/\sqrt{n}$ is the *Monte Carlo standard error*.

It gives the size (on average) of the difference between the quantity we want to calculate (μ) and our Monte Carlo approximation to it ($\widehat{\mu}_n$).

Variance Reduction

There are often multiple ways to do Monte Carlo approximation of the same quantity.

Markov chain Monte Carlo (MCMC) gives many ways. Here we work with GOFMC, which also has multiple ways.

The boring name for this subject is "variance reduction" because the idea is to use the way with the smallest *Monte Carlo standard error*.

The cute name is "swindles".

Variance Reduction (cont.)

If you have taken a course like this you are *assumed to know* that the five types of swindles are

- Control Variates
- Antithetic Variates
- Common Random Numbers
- Rao-Blackwellization
- Importance Sampling

Control Variates

If

$$\eta = E\{h(X)\}$$

is an expectation we *can* calculate by hand, then

$$\tilde{\mu}_n = \hat{\mu}_n - \frac{1}{n}\sum_{i=1}^{n} \beta\Big(h(X_i) - \eta\Big)$$

is *another* estimator of μ.

Any β works — second term estimates zero.

Elementary Theory Reminder

$$\mathsf{var}(X - Y) = \mathsf{var}(X) - 2\,\mathsf{cov}(X, Y) + \mathsf{var}(Y)$$

Control Variates (cont.)

Monte Carlo variance (MCV) is

$$\mathrm{var}(\tilde{\mu}_n) = \mathrm{var}(\hat{\mu}_n) - \frac{2\beta}{n}\,\mathrm{cov}\{g(X), h(X)\} + \frac{\beta^2}{n}\,\mathrm{var}\{h(X)\}$$

Minimize by choosing β as

$$\beta_{\mathrm{opt}} = \frac{\mathrm{cov}\{g(X), h(X)\}}{\mathrm{var}\{h(X)\}}$$

the slope in the regression of $g(X)$ on $h(X)$.

Control Variates (cont.)

Also works with multiple covariates. If know

$$\eta_j = E\{h_j(X)\}, \qquad j = 1, \ldots, k$$

then

$$\tilde{\mu}_n = \hat{\mu}_n - \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \beta_j \Big(h_j(X_i) - \eta_j \Big)$$

is *another* estimator of μ.

Optimal β_j are the partial regression coefficients in the multiple regression of $g(X)$ on $h_1(X)$, ..., $h_k(X)$.

Antithetic Variates

Simulate IID exchangeable pairs $(X_i, \widetilde{X}_i)$, $i = 1, 2, \ldots$, with X_i and $\widetilde{X}_i$ having the same distribution as X.

Then

$$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i) + g(\widetilde{X}_i)}{2}$$

is *another* estimator of μ.

X_i and $\widetilde{X}_i$ may be correlated — negatively correlated best.

Like matched pairs experiment.

Antithetic Variates (cont.)

How to simulate antithetic variates when using inversion method: U_1, U_2, ... are IID Uniform$(0, 1)$.

$$X_i = G(U_i)$$
$$\widetilde{X}_i = G(1 - U_i)$$

where G is the quantile function of the distribtion we want to simulate.

How to simulate antithetic variates when not using inversion method: No known methods. Good luck!

Common Random Numbers

Again like a matched pairs experiment.

When comparing two things by Monte Carlo, use the same random numbers in both experiments.

If you want to estimate $E\{g_1(X)\} - E\{g_2(X)\}$ use

$$\frac{1}{n}\sum_{i=1}^{n} g_1(X_i) - \frac{1}{n}\sum_{i=1}^{n} g_2(X_i)$$

(same random numbers in both terms).

If $g_1(X)$ and $g_2(X)$ are nearly perfectly positively correlated and have nearly equal variance, then the variance above is nearly zero.

Elementary Theory Reminder

$$E(X) = E\{E(X \mid Y)\}$$
$$\text{var}(X) = \text{var}\{E(X \mid Y)\} + E\{\text{var}(X \mid Y)\}$$

Hence X and $E(X \mid Y)$ have same expectation, and conditioning reduces variance

$$\text{var}(X) \geq \text{var}\{E(X \mid Y)\}$$

Rao-Blackwellization

For any function $h(X)$

$$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} E\{g(X_i) \mid h(X_i)\}$$

is *another* estimator of μ — better than naive one.

$\tilde{\mu}_n$ converges to μ by the iterated expectation theorem, $\tilde{\mu}_n$ has smaller Monte Carlo error than $\widehat{\mu}_n$ by the iterated variance theorem.

Importance Sampling

Suppose we want to generate samples from a distribution having density $f(x)$ but don't know how. Suppose we can generate samples X_1, X_2, ... having density $h(x)$.

Then

$$\mu = E_f\{g(X)\} = E_h\left\{g(x)\frac{f(x)}{h(x)}\right\}$$

so long as there is no divide by zero. Then

$$\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{f(X_i)}{h(X_i)}$$

is *another* estimator of μ.

Importance Sampling (cont.)

Monte Carlo variance will be empirical variance of

$$Z_i = g(X_i) f(X_i) / h(X_i)$$

divided by Monte Carlo sample size.

Minimized when

$$\text{var}_h \left\{ g(X) \frac{f(X)}{h(X)} \right\}$$

is minimized.

Optimal choice of $h(x)$ is density proportional to $|g(x)| f(x)$ by theorem in Casella and Robert, *Monte Carlo Statistical Methods*.

Conclusions

Many ways to do GOFMC.

Naive way not necessarily the best.

"There's more than one way to do it" (TMTOWTDI, pronounced "Tim Toady") is the Perl slogan.

Could also be the GOFMC slogan.

Anti-Importance Sampling

any sample can come from any distribution

Trotter and Tukey (1956)

Importance sampling is most important when it is *not* used for variance reduction.

Replace $f(x)$ by $f_\theta(x)$. Then

$$\tilde{\mu}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \frac{f_\theta(X_i)}{h(X_i)}$$

is sensible estimator of

$$\mu(\theta) = E_\theta\{g(X)\}.$$

Anti-Importance Sampling (cont.)

With one sample from one distribution $h(x)$ we learn about $\mu(\theta)$ for all θ.

Using both

- Common Random Numbers
- Importance Sampling

Anti-Importance Sampling (cont.)

$$\nabla \tilde{\mu}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \frac{\nabla f_\theta(X_i)}{h(X_i)}$$

is sensible estimator of

$$\nabla \mu(\theta) = E_\theta\{\nabla g(X)\}.$$

Principle of common random numbers is crucial. Naive estimator

$$\frac{1}{\epsilon} \left(\frac{1}{n} \sum_{i=1}^{n} g(X_i) - \frac{1}{n} \sum_{i=1}^{n} g(X_i^*) \right)$$

where X_i are samples from f_θ and X_i^* are independent samples from $f_{\theta+\epsilon}$ would be terrible — useless.

Unknown Normalizing Constant Models

Let $h_\theta(x)$ be a parametric family of functions that are

- nonnegative and
- integrable.

When divided by what the integrate to, they become probability densities

$$f_\theta(x) = \frac{1}{c(\theta)} f_\theta(x)$$

where

$$c(\theta) = \int h_\theta(x) \mu(dx)$$

Unknown Normalizing Constant Models (cont.)

When the integral defining the "normalizing constant" $c(\theta)$ cannot be done analytically, it can be done by Monte Carlo.

Suppose X_1, X_2, ... are samples from f_ψ, then

$$\frac{c(\theta)}{c(\psi)} = \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x) \mu(dx) = E_\psi \left\{ \frac{h_\theta(x)}{h_\psi(x)} \right\}$$

so

$$c_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\psi(X_i)}$$

is a sensible estimator of $c(\theta)/c(\psi)$.

Unknown Normalizing Constant Models (cont.)

Hence

$$l_n(\theta) = \log h_\theta(x) - \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\psi(X_i)} \right)$$

is a sensible estimator of the log likelihood.