Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 4
Variance reduction for MC methods
January 30, 2020

Plan of today's lecture

1. Introduction to variance reduction
2. Control variates
3. Antithetic sampling
4. Stratified Sampling

We are here $\longrightarrow$ ●

1. Introduction to variance reduction
2. Control variates
3. Antithetic sampling
4. Stratified Sampling

Confidence interval from a simulation viewpoint

Assume, as usual, that we estimate $\tau = \mathbb{E}(\phi(X))$ by means of MC, providing the level $(1-\alpha)$ confidence interval

$$\left(\tau_N - \lambda_{\alpha/2}\frac{\sigma(\phi)}{\sqrt{N}}, \tau_N + \lambda_{\alpha/2}\frac{\sigma(\phi)}{\sqrt{N}}\right)$$

for τ. Conversely, assume that we want to choose N large enough to assure that we estimate τ with an error less than a given $\epsilon > 0$ on the specified level. This means that

$$\lambda_{\alpha/2}\frac{\sigma(\phi)}{\sqrt{N}} < \epsilon \quad \Leftrightarrow \quad N > \lambda^2_{\alpha/2}\frac{\sigma^2(\phi)}{\epsilon^2}.$$

Thus, the required MC sample size N (and consequently the required work) increases linearly with the variance $\sigma^2(\phi)$.

Alternative representations of τ

Thus, in general, a strategy to gain computational efficiency would thus be to find an alternative representation (ϕ', f') of τ, in the sense that

$$\tau = \mathbb{E}_f(\phi(X)) = \int_{\mathsf{X}} \phi(x) f(x) \, \mathrm{d}x = \int_{\mathsf{X}} \phi'(x) f'(x) \, \mathrm{d}x = \mathbb{E}_{f'}(\phi'(X)),$$

for which $\sigma^2_{f'}(\phi') < \sigma^2_f(\phi)$.

Last time we saw that importance sampling (IS) was one way to do this ($f' \leftarrow g$ and $\phi' \leftarrow \phi\omega$).

Last time: Importance sampling

The basis of importance sampling was to take an instrumental density g on X such that $g(x) = 0 \Rightarrow f(x)\phi(x) = 0$ and rewrite the integral as

$$\begin{aligned}\tau = \mathbb{E}_f\left(\phi(X)\right) &= \int_{\mathsf{X}} \phi(x) f(x)\,\mathrm{d}x = \int_{f(x)|\phi(x)|>0} \phi(x) f(x)\,\mathrm{d}x \\ &= \int_{g(x)>0} \phi(x) \frac{f(x)}{g(x)} g(x)\,\mathrm{d}x = \mathbb{E}_g\left(\phi(X)\frac{f(X)}{g(X)}\right) = \mathbb{E}_g\left(\phi(X)\omega(X)\right),\end{aligned}$$

where

$$\omega : \{x \in \mathsf{X} : g(x) > 0\} \ni x \mapsto \frac{f(x)}{g(x)}$$

is the so-called importance weight function.

Last time: Importance sampling (cont.)

We may now estimate $\tau = \mathbb{E}_g(\phi(X)\omega(X))$ using standard MC:

```
for $i = 1 \to N$ do
    draw $X_i \sim g$
end for
set $\tau_N \leftarrow \sum_{i=1}^{N} \phi(X_i)\omega(X_i)/N$
return $\tau_N$
```

The CLT provides immediately, as $N \to \infty$,

$$\sqrt{N}(\tau_N - \tau) \xrightarrow{\text{d.}} \mathcal{N}(0, \sigma_g^2(\phi\omega)),$$

where $\sigma_g^2(\phi\omega) = \mathbb{V}_g(\phi(X)\omega(X))$ can be estimated using `var`.

Conclusion: Try to choose g so that the function $x \mapsto \phi(x)\omega(x)$ is close to constant in the support of g.

Last time: Self-normalized IS

Often $f(x)$ is known only up to a normalizing constant $c > 0$, i.e. $f(x) = z(x)/c$, where we can evaluate $z(x) = cf(x)$ but not $f(x)$. We could then however show that τ can be rewritten as

$$\tau = \mathbb{E}_f\left(\phi(X)\right) = \ldots = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))},$$

where

$$\omega : \{x \in \mathsf{X} : g(x) > 0\} \ni x \mapsto \frac{z(x)}{g(x)}$$

is known and can be evaluated.

Last time: Self-normalized IS (cont.)

Thus, having generated a sample $X_1, \ldots, X_N$ from g we may estimate the numerator $\mathbb{E}_g(\phi(X)\omega(X))$ as well as the denominator $\mathbb{E}_g(\omega(X))$ using standard MC:

$$\tau = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}$$

$$\approx \frac{\frac{1}{N}\sum_{i=1}^{N} \phi(X_i)\omega(X_i)}{\frac{1}{N}\sum_{\ell=1}^{N} \omega(X_\ell)} = \sum_{i=1}^{N} \underbrace{\frac{\omega(X_i)}{\sum_{\ell=1}^{N} \omega(X_\ell)}}_{\text{normalized weight}} \phi(X_i) = \tau_N.$$

We also concluded that the denominator yields an estimate of the normalizing constant c:

$$c = \mathbb{E}_g(\omega(X)) \approx \frac{1}{N}\sum_{\ell=1}^{N} \omega(X_\ell).$$

A CLT for self-normalized IS estimators

One may establish the following CLT.

Theorem

Assume that $\sigma_g^2(\omega\phi) = \mathbb{V}_g(\omega(X)\phi(X)) < \infty$. *Then*

$$\sqrt{N}(\tau_N - \tau) \xrightarrow{d.} \mathcal{N}(0, \sigma_g^2(\omega\{\phi - \tau\})/c^2),$$

where, as usual, $c = \mathbb{E}_g(\omega(X))$*.*

The asymptotic standard deviation can be estimated by
(1) letting, for each of the draws $X_i \sim g$,

$$Z_i = \omega(X_i)(\phi(X_i) - \tau_N),$$

(2) applying `std` to the vector containing all the Z_i's, and, finally, (3) dividing the result by the estimate of the normalizing constant c.

We are here ⟶ ●

1 Introduction to variance reduction

2 Control variates

3 Antithetic sampling

4 Stratified Sampling

What do we need to know?

OK, so what do we need to master for having practical use of the MC method?

We agreed on that, for instance, the following questions should be answered:

1: How do we generate the needed input random variables?

2: How many computer experiments should we do? What can be said about the error?

3: Can we exploit problem structure to speed up the computation?

Control variates

Assume that we have at hand another real-valued random variable Y, referred to as a control variate such that

1. $\mathbb{E}(Y) = m$ is **known** and
2. $\phi(X)$ and Y can be simulated at the same complexity as $\phi(X)$.

Then we may set, for some $\beta \in \mathbb{R}$,

$$Z = \phi(X) + \beta(Y - m),$$

so that

$$\mathbb{E}(Z) = \mathbb{E}(\phi(X) + \beta(Y - m)) = \underbrace{\mathbb{E}(\phi(X))}_{=\tau} + \beta \underbrace{(\mathbb{E}(Y) - m)}_{=0} = \tau.$$

Control variates (cont.)

In addition, if $\phi(X)$ and Y have covariance $\mathbb{C}(\phi(X), Y)$ it holds that

$$\begin{aligned}\mathbb{V}(Z) = \mathbb{V}(\phi(X) + \beta Y) &= \mathbb{C}(\phi(X) + \beta Y, \phi(X) + \beta Y) \\ &= \mathbb{V}(\phi(X)) + 2\beta\mathbb{C}(\phi(X), Y) + \beta^2\mathbb{V}(Y).\end{aligned}$$

Differentiating w.r.t. β and minimizing yields

$$0 = 2\mathbb{C}(\phi(X), Y) + 2\beta\mathbb{V}(Y) \quad \Leftrightarrow \quad \beta = \beta^* = -\frac{\mathbb{C}(\phi(X), Y)}{\mathbb{V}(Y)},$$

which provides the optimal coefficient β^* in terms of variance.

Control variates (cont.)

Plugging β^* into the formula for $\mathbb{V}(Z)$ gives

$$\mathbb{V}(Z) = \mathbb{V}(\phi(X)) + 2\beta^*\mathbb{C}(\phi(X), Y) + (\beta^*)^2\mathbb{V}(Y)$$
$$= \ldots = \mathbb{V}(\phi(X))\left(1 - \frac{\mathbb{C}(\phi(X), Y)^2}{\mathbb{V}(\phi(X))\mathbb{V}(Y)}\right) = \mathbb{V}(\phi(X))\{1-\rho(\phi(X), Y)^2\},$$

where

$$\rho(\phi(X), Y) \overset{\text{def}}{=} \frac{\mathbb{C}(\phi(X), Y)}{\sqrt{\mathbb{V}(\phi(X))}\sqrt{\mathbb{V}(Y)}}$$

is the correlation between $\phi(X)$ and Y.

Consequently, we can expect large variance reduction if $|\rho(\phi(X), Y)|$ is close to 1.

Example: another tricky integral

We estimate

$$\tau = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(x))\, \mathrm{d}x \overset{sym}{=} 2\int_0^{\pi/2} \exp(\cos^2(x))\, \mathrm{d}x$$

$$\int_0^{\pi/2} \underbrace{2\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)}\, \mathrm{d}x = \mathbb{E}_f(\phi(X))$$

using

$$Z = \phi(X) + \beta^*(Y - m),$$

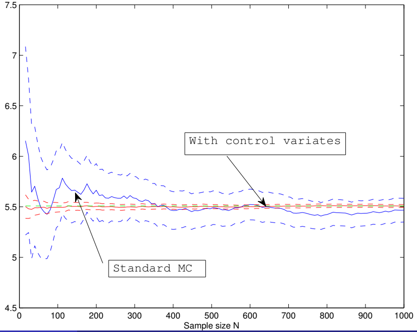
where $Y = \cos^2(X)$ is a control variate with

$$m = \mathbb{E}(Y) = \int_0^{\pi/2} \cos^2(x)\frac{2}{\pi}\, \mathrm{d}x = \{\text{use integration by parts}\} = \frac{1}{2}$$

and $\hat{\beta}^*$ is an estimate of the optimal coefficient.

Example: another tricky integral (cont.)

```
cos2 = @(x) cos(x).^2;
phi = @(x) 2*(pi/2)*exp(cos2(x));
X = (pi/2)*rand(1,N);
tau = mean(phi(X));
Y = cos2(X);
m = 1/2;
beta = - cov([phi(X)' Y'])./var(Y); % appr. optimal beta
Z = phi(X) + beta(1,2)*(Y - m);
tau_CV = mean(Z);
```



We are here $\longrightarrow \bullet$

1. Introduction to variance reduction
2. Control variates
3. Antithetic sampling
4. Stratified Sampling

Antithetic sampling

Again, assume that we wish to estimate $\tau = \mathbb{E}_f(\phi(X))$ by means of MC. For simplicity, let $V \stackrel{\text{def}}{=} \phi(X)$, so that $\tau = \mathbb{E}(V)$.

Now, assume we can generate another variable V' such that

1. $\mathbb{E}(V') = \tau$,
2. $\mathbb{V}(V') = \mathbb{V}(V)$ $(= \sigma^2(\phi))$,
3. V' can be simulated at the same complexity as V.

Then for

$$W \stackrel{\text{def}}{=} \frac{V + V'}{2}$$

it holds that $\mathbb{E}(W) = \tau$ and

$$\mathbb{V}(W) = \mathbb{V}\left(\frac{V + V'}{2}\right) = \frac{1}{4}\left(\mathbb{V}(V) + 2\mathbb{C}(V, V') + \mathbb{V}(V')\right)$$
$$= \frac{1}{2}\left(\mathbb{V}(V) + \mathbb{C}(V, V')\right).$$

Antithetic sampling (cont.)

Now, recall that the number N_V of V:s that we need to generate in order to estimate τ with an error less than a given $\epsilon > 0$ must satisfy

$$N_V > \lambda_{\alpha/2}^2 \frac{\mathbb{V}(V)}{\epsilon^2}.$$

Similarly,

$$N_W > \lambda_{\alpha/2}^2 \frac{\mathbb{V}(W)}{\epsilon^2}.$$

Consequently, it is better to use the W's if

$$\begin{aligned} 2\lambda_{\alpha/2}^2 \frac{\mathbb{V}(W)}{\epsilon^2} < \lambda_{\alpha/2}^2 \frac{\mathbb{V}(V)}{\epsilon^2} \quad &\Leftrightarrow \quad \mathbb{V}(V) + \mathbb{C}(V, V') < \mathbb{V}(V) \\ &\Leftrightarrow \quad \mathbb{C}(V, V') < 0. \end{aligned}$$

So, if we can find V' such that the antithetic variables V and V' are negatively correlated, then we will gain computational work.

Antithetic sampling (cont.)

For this purpose, the following theorem can be very useful.

Theorem

Let $V = \varphi(U)$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a monotone function. Moreover, assume that there exists a non-increasing transform $T : \mathbb{R} \to \mathbb{R}$ such that $U \overset{d.}{=} T(U)$. Then $V = \varphi(U)$ and $V' = \varphi(T(U))$ are identically distributed and

$$\mathbb{C}(V, V') = \mathbb{C}(\varphi(U), \varphi(T(U))) \leq 0.$$

Example: a tricky integral (reconsidered)

We estimate again

$$\tau = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(x))\,\mathrm{d}x \overset{sym}{=} \int_0^{\pi/2} \underbrace{2\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)} \,\mathrm{d}x$$
$$= \mathbb{E}(\phi(X)).$$

Now, set

$$\begin{cases} U = \cos^2(X), X \sim \mathcal{U}(0, \pi/2) \\ T(u) = 1 - u, \\ \varphi(u) = 2\frac{\pi}{2}\exp(u). \end{cases}$$

Then

$$T(U) = 1 - \cos^2(X) = \sin^2(X) = \cos^2\left(\frac{\pi}{2} - X\right) \overset{\mathrm{d.}}{=} \cos^2(X) = U.$$

Example: a tricky integral (reconsidered)

Since in addition $T(u) = 1 - u$ is non-increasing and $\varphi(u) = 2\frac{\pi}{2}\exp(u)$ monotone, the theorem above applies. Thus,

$$\mathbb{C}\left(\pi\exp(\cos^2(X)), \pi\exp(\underbrace{1-\cos^2(X)}_{=\sin^2(X)})\right) \leq 0,$$

and we may apply antithetic sampling with

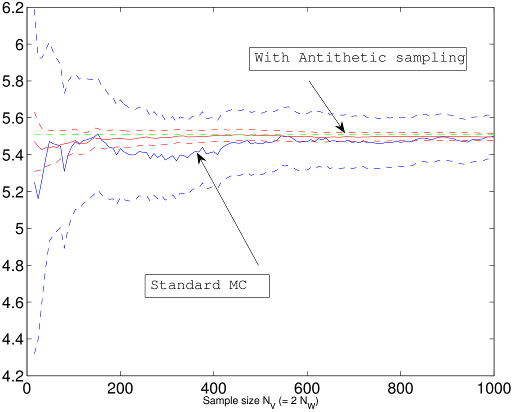
$$\begin{cases} V &= \pi\exp(\cos^2(X)), \\ V' &= \pi\exp(\sin^2(X)), \\ W &= \frac{V+V'}{2}. \end{cases}$$

Example: another tricky integral (cont.)

In Matlab:

```
cos2 = @(x) cos(x).^2;
phi = @(x) pi*exp(cos2(x));
X = (pi/2)*rand(1,N);
tau = mean(phi(X));
XX = (pi/2)*rand(1,N/2); % only half the sample size
V_1 = pi*exp(cos2(XX));
V_2 = pi*exp(1 - cos2(XX));
W = (V_1 + V_2)/2;
tau_AS = mean(W);
UB = tau + norminv(0.975)*std(phi(X))./sqrt(N);
LB = tau - norminv(0.975)*std(phi(X))./sqrt(N);
UB_AS = tau_AS + norminv(0.975)*std(W)./sqrt(N/2);
LB_AS = tau_AS - norminv(0.975)*std(W)./sqrt(N/2);
```

Example: another tricky integral (cont.)



We are here ⟶ •

1. Introduction to variance reduction

2. Control variates

3. Antithetic sampling

4. Stratified Sampling

Stratified Sampling

Let $A_1, \dots, A_J$ be disjoint sets such that

$$\cup_{i=1}^{J} A_i = \{\text{support of distribution for } X \text{ with density } f\}.$$

We then have that (Law of total probability)

$$f(x) = \sum_{i=1}^{J} f_{X|X\in A_i}(x)\mathbb{P}(X \in A_i) \overset{\text{def}}{=} \sum_{i=1}^{J} f_i(x)p_i.$$

So if we want to calculate $\tau = \mathbb{E}_f(\phi(X))$ we can use that

$$\mathbb{E}_f[\phi(X)] = \sum_{i=1}^{J} p_i \mathbb{E}_{f_i}[\phi(X)].$$

Stratified Sampling (Cont)

We can now simulate $N_1, N_2, \ldots, N_J$ independent random variables $\{X_{1,k}\}_{k=1}^{N_1}, \{X_{2,k}\}_{k=1}^{N_2}, \ldots, \{X_{J,k}\}_{k=1}^{N_J}$ from $f_1, f_2, \ldots, f_J$ respectively where $\sum_{i=1}^{J} N_i = N$. The simulation can be done using problem one on the home assignment if the A_i:s are intervals. We then form the Monte Carlo estimate

$$\tau_N^{\text{strat}} = \sum_{i=1}^{J} \frac{p_i}{N_i} \sum_{k=1}^{N_i} \phi(X_{i,k}).$$

$$\mathbb{E}[\tau_N^{\text{strat}}] = \sum_{i=1}^{J} p_i \mathbb{E}_{f_i}[\phi(X_{i,1})] \overset{\text{def}}{=} \sum_{i=1}^{J} p_i \mu_i = \mathbb{E}_f[\phi(X)]$$

and

$$\mathbb{V}[\tau_N^{\text{strat}}] = \sum_{i=1}^{J} \frac{p_i^2}{N_i} \mathbb{V}_{f_i}[\phi(X_{i,1})] \overset{\text{def}}{=} \sum_{i=1}^{J} \frac{p_i^2}{N_i} \sigma_i^2.$$

When is this better than standard Monte Carlo?

First we have that for τ_N being the standard MC estimate

$$\mathbb{V}[\tau_N] = \frac{1}{N}\mathbb{V}_f[\phi(X)] = \frac{1}{N}\sum_{i=1}^{J} p_i\sigma_i^2 + \frac{1}{N}\sum_{i=1}^{J} p_i(\mu_i - \tau)^2$$

Comparing this with

$$\mathbb{V}[\tau_N^{\text{strat}}] = \sum_{i=1}^{J} \frac{p_i^2}{N_i}\sigma_i^2$$

we see that choosing $N_i = p_i N$ (proportional allocation) will give

$$\mathbb{V}[\tau_N^{\text{strat,prop}}] = \frac{1}{N}\sum_{i=1}^{J} p_i\sigma_i^2 \leq \mathbb{V}[\tau_N].$$

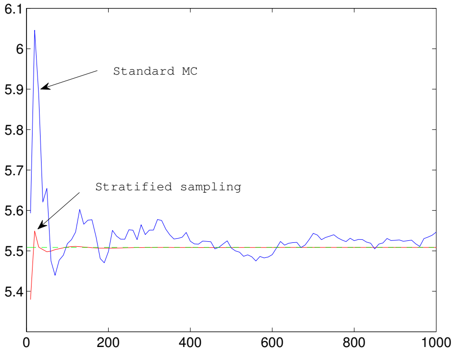
One extreme choice would be $J = N$,
$A_i = (F^{-1}((i-1)/N), F^{-1}(i/N)] \Rightarrow p_i = 1/N \Rightarrow N_i = 1.$

Example: Tricky integral again

We estimate

$$\tau = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(x))\,\mathrm{d}x$$

with (extreme) stratified sampling



Optimal allocation (Neyman allocation)

Minimizing

$$\sum_{i=1}^{J} \frac{p_i^2}{N_i} \sigma_i^2$$

under the restriction $N_1 + N_2 + \ldots + N_J = N$ give us

$$N_j = N \frac{p_j \sigma_j}{\sum_{i=1}^{J} p_i \sigma_i}, j = 1, 2, \ldots, J$$

which gives the variance

$$\mathbb{V}[\tau_N^{\text{strat,optimal}}] = \frac{1}{N} \left(\sum_{i=1}^{J} p_i \sigma_i \right)^2 \overset{\text{C-S}}{\leq} \frac{1}{N} \sum_{i=1}^{J} p_i \sigma_i^2 = \mathbb{V}[\tau_N^{\text{strat,prop}}] \leq \mathbb{V}[\tau_N].$$

This optimal choice of N_i is called Neyman allocation. The problem is however that we do not know the σ_i:s. We can use MC-methods with a smaller sample size to estimate them approximately (pilot sampling).