AM207 - Lecture 04 Variance reduction methods

Short recap from last lecture:

① MC for integration

② Sampling methods

③ Variance reduction methods

①: - Mean value theorem for integrals:

$$\int_a^b f(x)\,dx = (b-a)\cdot\langle f\rangle$$

- MC estimate for $\langle f\rangle$: $\langle f\rangle \approx \frac{1}{N}\sum_{x_i} f(x_i)$

- The error of our estimate is

$$\sigma_{MC} = \frac{\sigma_f}{\sqrt{N}}$$

②: - inverse transform

- rejection sampling

- rejection sampling with an envelope function.

③: - Importance Sampling

and today we will add:

- control variates
- antithetic variates
- stratification

Note: Altough it is called importance "sampling" it is not a sampling method!

Example for importance sampling:

$$\int_0^{\pi} \sin(x) \cdot x \, dx$$

Reminder for importance sampling:

We need $q(x) > 0$ whenever $f(x) \cdot p(x) \neq 0$

$q(x)$ and $p(x)$ need to be normalized.

$$\int_a^b f(x)\, dx \approx \frac{(b-a)}{N} \sum_{x_i \sim U(a,b)} f(x_i) = \frac{1}{N} \sum_{x_i \sim q(x_i)} \frac{f(x)}{q(x)}$$

Control variates

Intuition: Let's say we want to open a snow removal business in Cambridge and want to estimate the average amount of snow that is falling per day:

$E[f(x)]$.

We have access to snow data from Cambridge and Rhode Island: $\{f(x_i), g(x_i)\}$

A close friend has been living in Rhode Island all his life and from his experience can tell us the perfect (or at least very precise estimate of)

$E[g(x)]$

We know that it tends to snow at the same time in Cambridge and Rhode Island

$\Rightarrow$ How can we exploit our knowledge of $E[g(x)]$ to better estimate $E[f(x)]$?

$\Rightarrow$ we estimate $E[g(x)] \approx \frac{1}{N} \sum_i g(x_i) = \langle \hat{g} \rangle$ and compare it to the true $E[g(x)]$.

if $\langle \hat{g} \rangle > E[g(x)]$ then we also expect

$\langle \hat{f} \rangle > E[f(x)]$

Control variates

Basic idea: transform $f(x)$ to something else such that the new function has the same estimate of the mean:

$$f(x) \to f^*(x) \quad \text{with} \quad E[f(x)] = E[f^*(x)]$$

This means instead of using $f(x)$ for our estimate, we can as well use $f^*(x)$.

If $var(f^*) < var(f)$ we get a better estimate of our expectation value.

How can we find this mysterious function f^*?

Let's take another function $g(x)$ such that we know the expectation value:

$$g(x) \text{ s.t. } E[g(x)] = \tau \quad \text{with } \tau \text{ known}$$

| $\Rightarrow$ put snow removal business motivation here |

Now we do some magic:

$$f^*(x) = f(x) + b\,[g(x) - \tau]$$

$$E[f^*] = E[f] + b\,(E[g] - \tau)$$

$\hookrightarrow$ τ is just a number so $E[\tau] = \tau$

We said we know $E[g] = \tau$, that means

$$E[g] - \tau = \tau - \tau = 0$$

$$\Rightarrow E[f^*] = E[f]$$

So we reached our goal, we found a function f^* that has the same expectation value as f.

But we also need to reduce the variance to get better estimates:

$$\mathrm{var}(f^*) = \mathrm{var}(f) + b^2 \mathrm{var}(g) + 2b \,\mathrm{cov}(f, g)$$

Why? Because $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y)$

with $\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$

We want $\mathrm{var}(f^*)$ to be small and $\mathrm{var}(f^*) < \mathrm{var}(f)$

How do we minimize a function?

$\Rightarrow$ Take the derivative and set to zero

$$\frac{\partial}{\partial b} \mathrm{var}(f^*) \stackrel{!}{=} 0$$ so let's do this!

$$\frac{\partial}{\partial b} \mathrm{var}(f^*) = 2b \,\mathrm{var}(g) + 2\,\mathrm{cov}(f, g) \stackrel{!}{=} 0$$

and we solve for b:

$$b = -\frac{2\,\mathrm{cov}(f, g)}{2\,\mathrm{var}(g)}$$

this is the minimum, but is it smaller than $\mathrm{var}(f)$?

Let's look at our variance of f^* with the optimal value for b:

$$\mathrm{var}(f^*) = \mathrm{var}(f) + \frac{\mathrm{cov}^2(f,g)}{\mathrm{var}^{\cancel{2}}(g)} \cdot \cancel{\mathrm{var}(g)} - 2\,\frac{\mathrm{cov}^2(f,g)}{\mathrm{var}(g)} \cdot \cancel{\mathrm{cov}(f,g)}$$

$$= \mathrm{var}(f) - \frac{\mathrm{cov}^2(f,g)}{\mathrm{var}(g)}$$

$$= \mathrm{var}(f)\left[1 - \frac{\mathrm{cov}^2(f,g)}{\mathrm{var}(g)\cdot\mathrm{var}(f)}\right]$$

correlation coefficient: $\rho(f,g) = \dfrac{\mathrm{cov}(f,g)}{\sqrt{\mathrm{var}(g)\cdot\mathrm{var}(f)}}$

$\Rightarrow \mathrm{var}(f^*) = \mathrm{var}(f)\cdot(1-\rho^2(f,g))$

$\Rightarrow$ If ρ gets close to one, $\mathrm{var}(f^*)$ gets close to zero

$\Rightarrow$ If our function g has some correlation with f, then we reduce the variance of f^*.

So how do we use this for integration?

$$I = \int_V f(x)\,dx \approx \frac{V}{N}\sum f^*(x) = \frac{V}{N}\sum_i f(x_i) + b_c\,(g(x_i) - \tau)$$

we just showed this

solution from before

$\Rightarrow$ draw random x_i, plug them in, we've done!

There is one little problem: we need to calculate

$b_c = -\frac{cov(f,g)}{var(g)}$ based on our samples.

If we have enough samples, this works perfectly fine

Summary: Need to find $g(x)$ such that we know $\tau = E[g(x)]$ and such that $g(x)$ is correlated with $f(x)$.

1. Draw $x_i \sim U[x_{min}, x_{max}]$
2. Calculate $f(x) = Y$

   $g(x) = Z$

   τ

   $b_c = -\frac{cov(Y,Z)}{Var(Z)}$

3. Compute $\frac{V}{N} \sum Y + b(Z - \tau)$

$\boxed{\text{Show example}}$

Multiple control variates:

$$f^*(x) = f(x) + b_1[g_1(x) - T_1] + b_2[g_2(x) - T_2] + \dots + b_n[g_n(x) - T_n]$$

why do we need this? You might not be able to find directly a function that correlates with $f(x)$, but a linear combination of functions where we know the Ts might do it.

$$\mathrm{var}(f^*) = \mathrm{var}(f) + 2\sum_i b_i \,\mathrm{cov}(f, g_i) + \sum_{i,j} b_i b_j \,\mathrm{cov}(g_i, g_j)$$

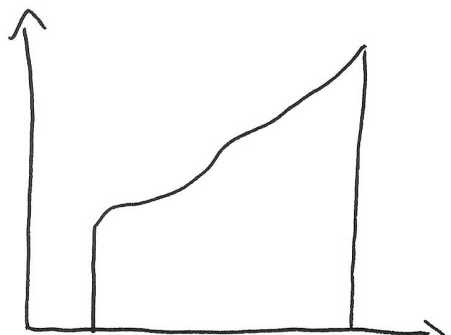
The last term is there because every time we have the variance of a sum, we pick up the covariance terms for all possible combinations.

To find the optimal b_i we again just take the derivative and set it to zero

$\Rightarrow$ n linear equations with n unknowns so we can solve this.

Antithetic variates:

Intuition:



If we have a monotonic function, it helps to balance the samples $\Rightarrow$ if u_i is high, also take a low u_j into the sample.

$\Rightarrow$ for example set $u_1, u_2, u_3, \dots u_n$ also create $1-u_1, 1-u_2, 1-u_3, \dots$

Let's say we have two variates Y_1 and Y_2 (two sets of random numbers) from the same source

Now we construct a new set $Z = \frac{Y_1 + Y_2}{2}$

$$E[Z] = E\left[\frac{Y_1+Y_2}{2}\right] = \frac{1}{2}\left(E[Y_1] + E[Y_2]\right) = E[Y_1] = E[Y_2]$$

So this part works.

Now lets look at the variance:

$$\mathrm{var}(Z) = \frac{1}{4}\left(\mathrm{var}(Y_1) + \mathrm{var}(Y_2) + 2\,\mathrm{Cov}(Y_1, Y_2)\right)$$

↳ from the denominator of Z

again we have $\mathrm{var}(Y_1) = \mathrm{var}(Y_2)$ so we can write

$$\mathrm{var}(Z) = \frac{1}{2}\mathrm{var}(Y_1) + \frac{1}{2}\mathrm{cov}(Y_1, Y_2)$$

⇒ If the covariance of Y_1 and Y_2 is negative then we reduce our variance.

⇒ We are not talking about functions here, just two sets of random samples.

So how can we create covariance?

$cov(Y_1, Y_1) = \sigma^2$ ⇒ The covariance of a variable to itself is just its variance

$corr(Y_1, Y_1) = 1$ ⇒ perfect correlation

For a completely independent variable $\varepsilon \perp Y_1$ we have:

$$Cov(Y_1, \varepsilon) = Corr(Y_1, \varepsilon) = 0$$

We can build Y_2 as a linear combination of Y_1 and ε to control the amount of correlation

$$Y_2 = aY_1 + (1-a)\varepsilon$$

If ~~fa~~ we want a to directly correspond to ϱ
~~For normally distributed Y_1 and ε we have~~
we have:

$$Y_2 = \varrho Y_1 + \sqrt{1-\varrho^2}\,\varepsilon \quad \text{with } Corr(Y_1, Y_2) = \varrho$$

Let's compare again the variance of our estimate.
To be fair we need to take into account that we are using two samples: u, $1-u$

we define $Y_1 = f(u)$, $Y_2 = f(1-u)$

$$Z = \frac{Y_1 + Y_2}{2}$$

$$I = \int f(x)\,dx$$

$$\mathrm{var}(\hat{I}_{2n}) \approx \frac{\mathrm{var}(Y)}{2n}$$

$\Leftarrow$ this is our vanilla MC error, but we are omitting the factor for the coverage, because it is the same for both methods

Compare this to:

$$\mathrm{var}(\hat{I}_{n+n}) = \frac{\mathrm{var}(Y_1+Y_2)}{4n} = \frac{\mathrm{var}(Y_1) + \mathrm{var}(Y_2) + 2\,\mathrm{cov}(Y_1 + Y_2)}{4n}$$

$$= \frac{\mathrm{var}(Y_1)}{2n} + \frac{\mathrm{cov}(Y_1, Y_2)}{2n}$$

$\Rightarrow \mathrm{var}(\hat{I}_{2n}) > \mathrm{var}(\hat{I}_{n+n})$ if $\mathrm{cov}(Y_1, Y_2) < 0$

$\Rightarrow$ If $\mathrm{cov}(Y_1, Y_2)$ is negative, then the ~~estim~~ variance of my estimation is reduced.

This looks like a free lunch, but it isn't.
The catch is that drawing from u and $1-u$ does not guarantee that $Y_1 = f(u_i)$ and $Y_2 = f(u_j)$ are correlated.

Theorem: If f is a monotonic function, then

$$\mathrm{Cov}(f(x), f(1-x)) < 0$$

Note: If not iff: condition is sufficient, but not necessary.

In real life it is often sufficient if $f(x)$ is approximately monotonic.