Monte Carlo methods

Stat 580

Monte Carlo integration

Monte Carlo integration

- a numerical approximation for expectation
- often useful for multidimensional problems that require the estimation of $\mu = E\{h(\boldsymbol{X})\}$, where $\boldsymbol{X}$ is a random vector and h is a function
- simplest approach: approximate μ by $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{X}_i)$ where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid copies of $\boldsymbol{X}$
- properties:
    - $\hat{\mu}_m$ is consistent: by the Strong Law of Large Numbers, with probability 1, $\hat{\mu}_m \to \mu$ as $m \to \infty$
    - $\hat{\mu}_m$ is unbiased: $E(\hat{\mu}_m) = \mu = E\{h(\boldsymbol{X})\}$.
    - $\mathrm{Var}(\hat{\mu}_m) = \mathrm{Var}\{h(\boldsymbol{X})\}/m$ and can be estimated by

$$\widehat{\mathrm{Var}}(\hat{\mu}_m) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\{h(\boldsymbol{X}_i) - \hat{\mu}_m\}^2.$$

Monte Carlo integration

- we will see methods:

  1. that are applicable when $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m$ cannot be easily sampled

  2. to reduce $\mathrm{Var}(\hat{\mu}_m)$

- MC integration can be used to evaluate a "usual" integral $I = \int_{\mathcal{X}} H(x)dx$

  1. the idea is to "factorize" $H(x) = f(x)h(x)$ with $f(x)$ as a pdf with support $\supseteq \mathcal{X}$ (we take $h(x) = 0$ if $x \notin \mathcal{X}$.)

  2. approximate I by $\frac{1}{m}\sum_{i=1}^{m} h(X_i)$, where $X_1, \dots, X_m$ are iid with pdf $f(x)$

Example

We want to compute

$$\int_{-\infty}^{\infty} \log |x| e^{-\frac{(x+1)^2}{8}}\, dx.$$

Set

$$f(x) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(x+1)^2}{8}}$$

and

$$h(x) = \sqrt{8\pi} \log |x|.$$

Note that $f(x)$ is the pdf for $\mathcal{N}(-1, 4)$, so the integral can be approximated by $\frac{\sqrt{8\pi}}{n} \sum_{i=1}^{m} \log |X_i|$ with $\{X_i\}$ iid $\sim \mathcal{N}(-1, 4)$.

Example

In a particular missing data problem, suppose $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})$ but only $\boldsymbol{X} = M(\boldsymbol{Y})$ is observed, where M is a (known and nonrandom) many-to-fewer mapping.

Frequentist approach:

To estimate $\boldsymbol{\theta}$ by the MLE, you need the likelihood

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{M(\boldsymbol{y})=\boldsymbol{x}} p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}.$$

- This integral is often hard to compute analytically
- Using MC integration
  - one designs a suitable function h and a distribution P
  - then sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from P and get

$$L(\boldsymbol{\theta}|\boldsymbol{x}) \approx \frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{Y}_i).$$

Example

Bayesian approach:

- Suppose $\boldsymbol{\theta}$ has a prior $\pi(\boldsymbol{\theta})$.
- We are interested in the posterior distribution of $\boldsymbol{\theta}$.
  - If the full data $\boldsymbol{Y} = \boldsymbol{y}$ is observed, by Bayes formula, the posterior is

$$p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\mathbf{y}) \propto p_{\boldsymbol{Y}}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

  - If only $\boldsymbol{X} = \boldsymbol{x}$ is observed, the posterior is

$$p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\mathbf{x}) = \int p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\mathbf{y})p_{\boldsymbol{Y}|\boldsymbol{M}(\boldsymbol{Y})}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}.$$

    - If $p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y})$ has a closed form, a MC integration method may sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \sim p_{\boldsymbol{Y}|\boldsymbol{M}(\boldsymbol{Y})}(\boldsymbol{y}|\boldsymbol{x})$ and approximate

$$p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\mathbf{x}) \approx \frac{1}{m}\sum_{i=1}^{m} p_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\mathbf{Y}_i).$$

Importance sampling

Importance sampling

- Same setup: want to estimate

$$\mu = \int h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}, \quad f(\boldsymbol{x}) \text{ is a pdf.}$$

  (the integral is taken over the region where the integrand is positive)

  - but it is difficult to sample from f

- Rewrite

$$\mu = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x},$$

  where g is a pdf such that $g(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x}) > 0$.

- Let $\boldsymbol{X}$ have density $g(\boldsymbol{x})$.
- Then $\mu = E\{h(\boldsymbol{X}) w^*(\boldsymbol{X})\}$ with $w^*(\boldsymbol{X}) = \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}$.

Importance sampling

- Consider the following steps:

  1. Generate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ iid $\sim g(\boldsymbol{x})$.

  2. Estimate μ by $\hat{\mu}_g = \frac{1}{m} \sum_{i=1}^m h(\boldsymbol{X}_i) w^*(\boldsymbol{X}_i)$.

- $\hat{\mu}_g$ is the importance sampling (IS) estimator of μ associated with g.

- $w^*(\boldsymbol{X}_i)$'s are referred to as importance ratios

- Note that $\hat{\mu}_g$ is a weighted sum of $h(\boldsymbol{X}_i)$.

- If $f = g$, then $\hat{\mu}_g$ is the ordinary MC estimator.

Example

We want to compute $\mu = E(U^5)$ where $U \sim \text{Unif}(0,1)$.

- the straightforward MC estimator: $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} U_i^5$
  - oversample data U_i^5 near the origin and undersample the data near 1
  - $\text{Var}(\hat{\mu}) = 0.0631/m$
- Use IS to put more weights near 1:
  - use $g(x) = 5x^4$ for $0 < x < 1$
  - the IS estimator:$\hat{\mu}_g = \frac{1}{n}\sum_{i=1}^{m} X_i^5 w^*(X_i)$ where $w^*(X_i) = 1/(5X_i^4)$
  - $\text{Var}(\hat{\mu}_g) = 0.00794/m$ (verify!)
  - resulting a variance reduction of 98.74%!
- the IS can be used as a variance reduction technique!

Properties

1. $\hat{\mu}_g$ is unbiased for μ.

2. $\mathrm{Var}(\hat{\mu}_g) = \frac{1}{m}\mathrm{Var}\{h(\boldsymbol{X})w^*(\boldsymbol{X})\}$.

- To reduce the variance of $\hat{\mu}_g$, $g(\boldsymbol{x})$ should be in proportion to $h(\boldsymbol{x})f(\boldsymbol{x})$ as much as possible.

Properties

To show this, we need to reduce

$$E\left[\left\{h(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right\}^2\right] = \mathrm{Var}(\hat{\mu}_g) + \mu^2.$$

We can use

$$E\left[\left\{h(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right\}^2\right] \geq \left[E\left\{h(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right\}\right]^2.$$

The equality holds if $h(\boldsymbol{x})\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}$ is a constant. That is, when $g(\boldsymbol{x}) \propto h(\boldsymbol{x})f(\boldsymbol{x})$. (Why is it called importance sampling?)

When f is only known up to a constant

- That means, $f(\boldsymbol{x}) = cq(\boldsymbol{x})$ with $c > 0$ unknown, then

$$\mu = \frac{E\{h(\boldsymbol{X})w^*(\boldsymbol{X})\}}{E\{w^*(\boldsymbol{X})\}}$$

  with $w^*(\boldsymbol{X}) = \frac{q(\boldsymbol{X})}{g(\boldsymbol{X})}$.

- In this case, standardized weights have to be used in IS:

  1. Generate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ iid from $g(\boldsymbol{x})$.
  2. Estimate μ by $\hat{\mu}_g = \frac{1}{m} \sum_{i=1}^m h(\boldsymbol{X}_i)w(\boldsymbol{X}_i)$ where $w(\boldsymbol{X}_i) = \frac{w^*(\boldsymbol{X}_i)}{\sum_{i=1}^m w^*(\boldsymbol{X}_i)}$.

Processes

- Let $\{S_1, S_2, \dots\}$ be a discrete time stochastic process.
- For example, S_i = asset price $S(t_i)$ at time t_i.
- In general, to estimate $\mu = E\{h(S_1, \dots, S_n)\}$, one treats $(S_1, \dots, S_n)$ as a random vector $\boldsymbol{S}$ with density f.
- Then IS estimates μ by $\hat{\mu}_g = \frac{1}{m} \sum_{i=1}^m h(\boldsymbol{X}_i) \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}$, where $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m$ are iid with density $g(\boldsymbol{x})$, and $g(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x}) > 0$.
- It is useful to think of $\boldsymbol{X}_i = (X_{i,1}, \dots, X_{i,n})$ as a sample path of a process (up to the n-th index).
- From this point, IS uses a process different from $\boldsymbol{S}$ to estimate μ.
- An important issue is how to design the auxiliary process to take advantage of any special dependence structure of $S_1, S_2, \dots$.

Markov processes

- $S_1, \ldots, S_n$ form a Markov process; i.e., we can decompose their joint density as

$$f(x_1, \ldots, x_n) = f_1(x_1) \prod_{j=2}^{n} f_j(x_j | x_{j-1}),$$

where $f_1(x_1)$ is the density of the initial distribution and $f_j(x_j | x_{j-1})$ is the transition density at step j.

- To utilize the Markov property in IS, the auxiliary density $g(x_1, \ldots, x_n)$ may also have a similar structure as f:

$$g(x_1, \ldots, x_n) = g_1(x_1) \prod_{j=2}^{n} g_j(x_j | x_{j-1})$$

- Now in IS, the weight for $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$w(\boldsymbol{X}) = \frac{f_1(X_1)}{g_1(X_1)} \prod_{j=2}^{n} \frac{f_j(X_j | X_{j-1})}{g_j(X_j | X_{j-1})}.$$

Markov processes

- The IS in this case is, for $i = 1, \dots, m$:

  1. Draw $X_1 \sim g_1$.
  2. For $j = 2, \dots, n$, draw $X_j \sim g_j(x_j | x_{j-1})$.
  3. Set $h_i = h(X_1, \dots, X_n)$ and

  $$w_i = \frac{f_1(X_1)}{g_1(X_1)} \prod_{j=2}^{n} \frac{f_j(X_j | X_{j-1})}{g_j(X_j | X_{j-1})}.$$

- The IS estimator is

$$\hat{\mu}_g = \frac{1}{m} \sum_{i=1}^{m} h_i w_i.$$

Control variates

Control variates

- We still want to compute $\mu = E\{h(\boldsymbol{X})\}$.
- Suppose we know the exact value of $\theta = E\{c(\boldsymbol{Y})\}$, where c is a function of another random variable $\boldsymbol{Y}$.
- The simple MC estimators for μ and θ are, respectively,

$$\hat{\mu}_{\text{MC}} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i), \qquad \hat{\theta}_{\text{MC}} = \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{Y}_i).$$

- Of course, for θ, $\hat{\theta}_{\text{MC}}$ is unnecessary. However, it can be helpful for the estimation of μ.
- How? Suppose $h(\boldsymbol{X})$ and $c(\boldsymbol{Y})$ are positively correlated. (If not, this method is not applicable.)

Control variates

- If we see $\hat{\theta}_{\text{MC}} > \theta$, then due to the positive correlation, $\hat{\mu}_{\text{MC}}$ is more likely to be $> \mu$.
- Then we can decrease the value of $\hat{\mu}_{\text{MC}}$ to obtain a better estimate.
- To be specific, suppose we can sample $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ iid.
- The control variate estimator for μ is

$$\hat{\mu}_{\text{CV}} = \hat{\mu}_{\text{MC}} - b(\hat{\theta}_{\text{MC}} - \theta),$$

  where b is a constant.
- $\hat{\mu}_{\text{CV}}$ is unbiased and consistent (as MC estimators are unbiased and consistent).
- If $b = 0$, then $\hat{\mu}_{\text{CV}} = \hat{\mu}_{\text{MC}}$, so the optimal control variate estimator is at least as good as $\hat{\mu}_{\text{MC}}$.
- Given a control variate $c(\boldsymbol{Y})$, need to choose b (choosing $c(\boldsymbol{Y})$ is harder)

Choice of b

- How should we choose b?
  - for any given b,

$$\mathrm{Var}(\hat{\mu}_{\mathrm{CV}}) = \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) + b^2\mathrm{Var}(\hat{\theta}_{\mathrm{MC}}) - 2b\mathrm{Cov}(\hat{\mu}_{\mathrm{MC}}, \hat{\theta}_{\mathrm{MC}}).$$

  - the minimum of $\mathrm{Var}(\hat{\mu}_{\mathrm{CV}})$ happens when $b = b^*$, where

$$b^* = \frac{\mathrm{Cov}(\hat{\mu}_{\mathrm{MC}}, \hat{\theta}_{\mathrm{MC}})}{\mathrm{Var}(\hat{\theta}_{\mathrm{MC}})} = \frac{\mathrm{Cov}\{h(\boldsymbol{X}), c(\boldsymbol{Y})\}}{\mathrm{Var}\{c(\boldsymbol{Y})\}}.$$

  - in practice b^* is unknown, but we can estimate it.
  - plug b^* into $\mathrm{Var}(\hat{\mu}_{\mathrm{CV}})$ to get $\mathrm{Var}(\hat{\mu}_{\mathrm{CV}}^{\mathrm{opt}})$, and we can show the variance reduction factor is

$$\frac{\mathrm{Var}(\hat{\mu}_{\mathrm{CV}}^{\mathrm{opt}})}{\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})} = 1 - \rho^2,$$

    where ρ is the correlation coefficient between $h(\boldsymbol{X})$ and $c(\boldsymbol{Y})$.

Estimation of b^*

- the optimal b^* can be estimated by

$$\hat{b}_n = \frac{\widehat{\mathrm{Cov}}\,(\hat{\mu}_{\mathrm{MC}}, \hat{\theta}_{\mathrm{MC}})}{\widehat{\mathrm{Var}}\,(\hat{\theta}_{\mathrm{MC}})},$$

  where

$$\widehat{\mathrm{Var}}\,(\hat{\theta}_{\mathrm{MC}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\{ c(\boldsymbol{Y}_i) - \hat{\theta}_{\mathrm{MC}} \right\}^2$$

  and

$$\widehat{\mathrm{Cov}}\,(\hat{\mu}_{\mathrm{MC}}, \hat{\theta}_{\mathrm{MC}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \{h(\boldsymbol{X}_i) - \hat{\mu}_{\mathrm{MC}}\}\{c(\boldsymbol{Y}_i) - \hat{\theta}_{\mathrm{MC}}\}.$$

- $\hat{b}_n$ is the slope of the least-squares regression line for $(h(\boldsymbol{X}_i), c(\boldsymbol{Y}_i))$, $i = 1, \ldots, n$

General idea

- Overall, the general idea for the control variate method is to search $c(\boldsymbol{Y})$ such that

    1. $E\{c(\boldsymbol{Y})\}$ is known.

    2. the scatterplot of $(h(\boldsymbol{X}_i), c(\boldsymbol{Y}_i))$ shows strong correlation.

- In practice, $\hat{\mu}_{\mathrm{MC}}$ and $\hat{\theta}_{\mathrm{MC}}$ often depend on the same random variable; i.e., $\boldsymbol{Y}_i = \boldsymbol{X}_i$.

- It is possible to use more than one control variate; i.e.,

$$\hat{\mu}_{\mathrm{CV}} = \hat{\mu}_{\mathrm{MC}} - b_1(\hat{\theta}_{1,\mathrm{MC}} - \theta_1) - b_2(\hat{\theta}_{2,\mathrm{MC}} - \theta_2).$$

Example

Let $\mu = E(e^U)$ where $U \sim \text{Unif}(0, 1)$. Theoretical study of CV estimator with b^* (when $n = 1$):

- Use U as the control variate
- $E(U) = 1/2, \text{Cov}(e^U, U) = 1 - (e - 1)/2 = 0.14086$ and $\text{Var}(U) = 1/12$
- the CV estimator: $\hat{\mu}_{CV} = e^U - b^*(U - 1/2)$, where $b^* = 12(0.14086)$
- $\text{Var}(\hat{\mu}_{CV}) = 0.0039$ (verify!)
- resulting a variance reduction of 98.4% when compared to $Var(e^U) = 0.242$

Monte Carlo Testing

Monte Carlo testing

- We use Monte Carlo testing when
  - we cannot calculate the null distribution of the test statistic
  - but we can simulate from the null hypothesis.
- We "simulate" the null distribution of the test statistic.

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Suppose we want to test $H_0 : \mu = 3, \sigma^2 = 4$ against the alternative $H_A : \mu \neq 3, \sigma^2 = 4$.

- This is a simple z-test. The test statistic is

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- To use Monte Carlo testing,
  - we generate $X_1^*, \ldots, X_n^*$ from $\mathcal{N}(3, 4)$, the distribution in the null hypothesis.
  - We then calculate $\bar{X}^* = \frac{\sum X_i^*}{n}$ and repeat this procedure m times (e.g., set $m = 999$) to get m $\bar{X}^*$ values.
  - This approximates the distribution of the test statistic using m simulated values of $\bar{X}$.
  - If $\bar{X}$ is amongst the smallest 2.5% or the largest 2.5% of these $\bar{X}^*$ values, we reject H_0.