Sec 6.3 Variance Reduction Teocniques.

Pg 162

Recall, the simple Monte Carlo estimator of $\mu = \int h(x) f(x)\, dx$ is $\hat{\mu}_{mc} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$

where $X_1, \ldots, X_n$ are randomly sampled from f. However, better m.c. estimators (lower variances) can be derived by using clever sampling strategies.

Sec 6.3.1 Importance Sampling.

Very (Overly) Simple motivating Example:

Suppose we wish to estimate the prob of a die roll will yield a 1.

we roll the die n times $\longrightarrow$ Expect $n/6$ ones $\quad (p \sim \frac{1}{6})$

and our point estimate would be the proportion of 1's in the sample.

The variance of this estimator is $5/36n$ if the die is fair. (Bernoulli).
So, to achieve an est. w/ coef of variation $= \frac{\sqrt{var(x)}}{E(x)}$. of 5%, we would expect to have to roll the die 2000 times.

To redue required # of rolls, let's replace the die w/ 1,1,1,4,5,6. so the prob of rolling a1 is 1/2.
Problem: We are no longer sampling from the target dist of a fair die.

Solution: Weight each roll of 1 by 1/3.
Let $Y_i = 1/3$ if a 1 and $Y_i = 0$ o.w.
Then the exp. of sample mean of Y_i is $1/6$
however the variance is $1/36n$.

since for population

$$\left(E[Y_i] = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \; ; \; Var(Y_i) = E[Y_i^2] - E[Y_i]^2 = \frac{1}{9} \cdot \frac{1}{2} - \left(\frac{1}{6}\right)^2 = \frac{1}{18} - \frac{1}{36} = \frac{1}{36} \right)$$

∴ to achieve a coef. of var of 5% we only expect to need 400 rolls.

This improved accuracy is caused by forcing the event of interest to occur more freq.

Our die rolling example is successful because we used an "importance sampling dist" to oversample a portion of the state space that receives lower prob under the target dist. We used an "importance weighting" to correct for this bias & provide ~~our~~ improved estimator.

more formally.

The imp. sampling approach is upon the principle that exp of $h(x)$ w.r.t. density f can be written as

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

or sim.

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

where g is the imp. samp. function & another density. that is hopefully easy to sample from.

This alternative form suggests that a m.c. approach to estimating $E[h(x)]$ is to draw $X_1, \dots, X_n$ iid from g & use

$$\hat{\mu}^*_{IS} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)\, w^*(X_i)$$

→ no need to compare

where $w^*(X_i) = f(X_i)/g(X_i)$ are the importance weights or ratios.

Comments: ① Clearly $E[w^*(X)] = E\left[f(X)/g(X)\right] = 1$

② $E[\hat{\mu}^*_{IS}] = \frac{1}{n} \sum_{i=1}^{n} E[h(X_i) w^*(X_i)] = \mu$

③ $\text{var}[\hat{\mu}^*_{IS}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(h(X_i) w^*(X_i)) = \frac{1}{n} \text{var}(h(X) w(X))$

and is the value that we hope to reduce by our choice of $w^*(g)$.

Discussion pg 164

- we want $f(x)/g(x)$ to be bounded
- good to have g w/ heavier tails than f.
- want to avoid a rare draw from g getting a huge weight.
- In practice we want g to be nearly prop. to $|h(x) f(x)|$ so that $|h(x) f(x)|/g(x)$ is nearly a constant.

④ w^* as defined above are unstandardized weights. We obtain standardized weights by letting

$$w(X_i) = \frac{w^*(X_i)}{\sum_{i=1}^{n} w^*(X_i)}$$

to obtain

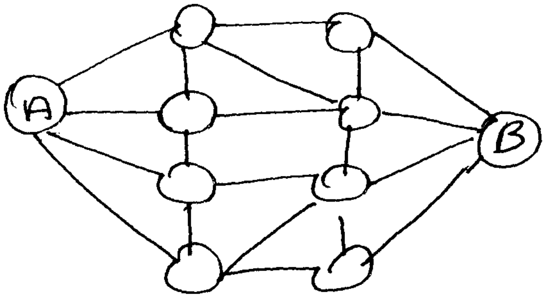
$$\hat{\mu}_{IS} = \sum_{i=1}^{n} h(x_i)\, w(x_i).$$

This approach can be used when f is known only up to a constant of prop. However (see discussion 165) a slight bias is introduced.

pg 166 Ex/ Network failure probability.

Many systems can be rep. by connected graphs - nodes & edges. (People, comp. communications & so on).

We are going to send a signal from Ⓐ to Ⓑ that can follow a path along any edges



We assume that with a small prob. p $(10^{-3} - 10^{-10})$ each edge may fail (independently).

The signal will only sucessfuly arrive at B from A if there is an unbroken path.

So we want to know the prob of a network failure.

Let X denote a network, summarizing random outcomes for each edge.

r.v. $X = (X_1, \ldots, X_{20})$ each X_i indicates broken intact

$b(X) = \#$ of broken edges in X.

$$h(X) = \begin{cases} 1 & \text{if network fails} \quad \cdot \text{ no A-B path} \\ 0 & \text{o.w.} \quad \cdot \text{ A-B paths exist.} \end{cases}$$

The propability of network failure

$$\mu = E[h(X)].$$

Computing μ directly for any realistically sized network can be a very difficult combinatorial problem, so we choose to use a M.C method.

Attempt ① Standard M.C.

Draw $X_1, \dots, X_n$ idep. & uniformly at at random from all possible network config whose edges fail w/ prob p. Then the estimator is

$$\hat{\mu}_{mc} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

(Bernoulli)

The variance of this estimator is $\mu(1-\mu)/n$. So For $n=100{,}000$ & $p=0.05$, simulation yields $\hat{\mu}_{mc} = 200 \times 10^{-5}$ w/ error 1.41×10^{-5} (same order of mag.). Only 2 networks failed.

The prob. is that when est. $\hat{\mu}_{mc}$, $h(x)$ is very rarely 1 and so a huge # of networks must be sampled to estimate μ with sufficient precision.

Attempt 2 Importance Sampling.

We will draw $X_1^*, \ldots, X_n^*$ by breaking edges w/ prob $p^* > p$ and then weighting.

Originally

$$\mu = \int h(x) f(x)\, dx$$

sim to Binomial

$$\text{where } f(x) = p^{b(x)} (1-p)^{20-b(x)}.$$

We want to use $g(x) = p^{*\,b(x)} (1-p^*)^{20-b(x)}$ so we need weights (unstandardized)

$$w^*(x_i^*) = f(x) / g(x) = \left(\frac{1-p}{1-p^*}\right)^{20} \left(\frac{p\,(1-p^*)}{p^*\,(1-p)}\right)^{b(x_i^*)}$$

And so our importance sampling estimator is

$$\hat{\mu}_{IS} = \frac{1}{n} \sum_{i=1}^{n} h(X_i^*)\, w^*(X_i^*)$$

What about the variance?

Let $\mathcal{C}$ be the set of all possible network configurations & $\mathcal{F}$ be the subset that fail.

$$\text{var}\{\hat{\mu}_{IS}\} = \frac{1}{n} \text{var}\{h(X_i^*)\, w^*(X_i^*)\}$$

$h(X_i^*)^2 = h(X_i^*)$ $= 1$ iff network fails

$$= \frac{1}{n}\left(E\{[h(X_i^*) w^*(X_i^*)]^2\} - [E\{h(X_i^*) w^*(X_i^*)\}]^2\right.$$

$$= \frac{1}{n}\left[\sum_{x \in \mathcal{F}} E[w^*(X_i)^2] - \mu^2\right)$$

$$= \frac{1}{n}\left[\sum_{x \in \mathcal{F}} \left[\left(\frac{1-p}{1-p^*}\right)^{20} \left(\frac{p(1-p^*)}{p^*(1-p)}\right)^{b(x)}\right]^2 \cdot p^{*\,b(x)} (1-p^*)^{20-b(x)} - \mu^2\right]$$

$$= \frac{1}{n}\left[\sum_{x\in\mathcal{F}_1} w^*(x)\, p^{b(x)}(1-p)^{20-b(x)} - \mu^2\right])$$

and noting that failure only occurs when $b \geq 4$

$$\forall x \in \mathcal{F}_1 \qquad w^*(x) \leq \left(\frac{1-p}{1-p^*}\right)^{20}\left(\frac{p(1-p^*)}{p^*(1-p)}\right)^4$$

So if

$p^* = .25$ & $p = .05$ $\quad w^*(x) \leq .07$ and

$$\mathrm{var}(\hat{\mu}_{IS}) \leq \frac{1}{n}\left(.07 \sum_{x\in F} p^{b(x)}(1-p)^{20-b(x)} - \mu^2\right)$$

$$= \frac{1}{n}\left(.07 \sum_{x\in \mathcal{E}} h(x)\, p^{b(x)}(1-p)^{20-b(x)} - \mu^2\right)$$

$$= \frac{1}{n}\left(.07\mu - \mu^2\right) < \mathrm{var}(\hat{\mu}_{mc})$$

$$= \frac{\mu}{n}(.07 - \mu)$$

In fact $\mathrm{var}\{\hat{\mu}_{mc}\}/\mathrm{var}\{\hat{\mu}_{IS}\} \approx 14$.

In our importance sampling 497 of the 100 000 networks failed, producing $\hat{\mu}_{IS} = 1.01\times10^{-5}$ and error 1.56×10^{-6}.

Sec 1.7
pg 14

Markov Chains

Consider a sequence of r.v. $\{X^{(t)}\}$, $t=0,1,2,\ldots$ where each value may equal 1 of an at most countably infinite set of possible values called states.

- $X^{(t)}=j$ indicates that the process is in state j at time t.
- The set $\mathcal{S}$ of possible values is called the state space.

Now suppose that

- $p_{ij}^{(t)}$ is the probability that the process changes from state i to state j at time $t+1$

If

$$p_{ij}^{(t)} = P\left[X^{(t+1)}=j \,\middle|\, X^{(0)}=x_0, X^{(1)}=x_1, \ldots, X^{(t)}=i\right]$$
$$= P\left[X^{(t+1)}=j \,\middle|\, X^{(t)}=i\right]$$

$\forall$ $t=0,1,\ldots$ & $X^{(0)}, x^{(1)}, \ldots x^{(t-1)}, i,j \in \mathcal{S}$

Then $\{X^{(t)}\}$, $t=0,1,\ldots$ is called a Markov Chain.
Basic idea: Given the "present" the "future" is independent of the "past", and so the process is Memoryless.

- [illegible] The $p_{ij}^{(t)}$ are called the single ~~the~~ ~~single~~ or one-step transition probabilities
- If they are independent of t, the chain is said to be homogeneous and $p_{ij}^{(t)}=p_{ij}$ and

o.w. inhomogeneous

$$P=[p_{ij}]=\begin{bmatrix} p_{00} & p_{01} & \cdots, & p_{0j} & \cdots \\ p_{10} & p_{11} & \cdots & p_{1j} & \cdots \\ \vdots & & \ddots & \vdots & \\ p_{i0} & p_{i1} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix}$$

is called the transition probability matrix - and P governs the behavior of the MC.

Note: ① $p_{ij} \geq 0 \quad \forall i,j$

row sums ② $\sum_j p_{ij} = 1 \quad \forall i$ } stochastic matrix.

③ The size of P is dep on size of S

Ex/ Consider a seq. of Bernoulli trials p - success & q-failure (prob).

Let X_n be the # of uninterupted successes that have been completed to this point.

For example

SFSSF gives $X_0=1$ $X_1=0$ $X_2=1$ $X_3=2$ $X_4=0$

then the state space for $\{X_n\}$ is $\{0,1,2,\dots\}$ and the transition prob matrix is

$$P = \begin{bmatrix} q & p & 0 & 0 & \cdots \\ q & 0 & p & 0 & \cdots \\ q & 0 & 0 & p & \cdots \\ q & 0 & 0 & 0 & p \\ q \\ \vdots \end{bmatrix}$$

The state 0 can be reached in 1 transition from any state, where as state i, $i>0$, can only be reached from state $i-1$.

This is a homogeneous M.C.

For homo M.C. we can define the m-step transition prob

$$p_{ij}^{(m)} = P\left[X^{(t+m)} = j \mid X^{(t)} = i\right]$$

the prob of going from i to j in m-steps.

and the corresponding mth transition matrix

$$P^{(m)} = [p_{ij}^{(m)}].$$

It can be shown that $P^{(m)} = P^m$.

ie. What happens to the process in the long run. Does it settle down to a steady state, distribution?

For the methods we will discuss, we wish to know the limiting behavior* & so we have the following def:

- A state to which the chain returns w/ probability 1 is called a <u>recurrent state</u>.
- If the expected time until a recurrence is finite it is called <u>nonnull</u>.
- If any state j can be reached from any state i in a finite number of steps, the chain is <u>irreducible</u>. (i.e, $\exists\ m>0 \ni P[X^{(m+n)}=j \mid X^{(n)}=i] > 0$)
- Let $d(i)$ be the G.C. Divisor of all integers n s.t. $p_{ii}^{(n)} > 0$.
- If $d(i)=1$, the i is said to be aperiodic o.w. periodic (& can only return to i after n steps where n is divisible by d). ie. the mc can only visit i at regularly spaced intervals.
- If $d(i)=1\ \forall i$, the M.C. is aperiodic

* If a mc is irreducible, aperiodic, and all states are nonnull & recurrent then the mc is said to be <u>ergodic</u>.

We like ergodic mc because they have nice limiting behaviors!!

marginal or unconditional prob.

- Let $\pi^{(t)}$ denote a vector of prob (sum to 1) with $\pi_i^{(t)} = \text{prob}(X^{(t)}=i)$.

- Then $\pi^{(t+1)} = [\pi^{(t)}] P$ is the marginal prob for $X^{(t+1)}$.

- If a (long run) limiting or stationary prob distribution exists for $\{X^{(t)}\}$ then $\pi^{(t+1)} = \pi^{(t)} = \pi$ true for all t and

$$\boxed{\pi = \pi P}$$ (Steady state)

Main Result: If a m.c. w/ trans matrix P is ergodic, then the stationary dist π ($\pi = \pi P$) is unique and limiting.

$$\lim_{n\to\infty} P[X^{(t+n)}=j \mid X^{(t)}=i]\ \pi_j$$

(ie the rows of $P^{(m)}$ go to π as $m \to \infty$).

Furthermore we compute this st. stated dist, by solving the system

$$\pi_j \geq 0,\ \sum_{i\in S} \pi_i = 1\ \text{\&}\ \pi_j = \sum_{i \in S} \pi_i p_{ij}\ \forall j \in S$$

$$\pi \geq 0 \quad \pi e = 0\ \text{\&}\ \pi = \pi P.$$

Furthermore if $\{X^t\}$ are real from eg. mc. w/ ss dist π. $\forall$ function h

$$\frac{1}{n} \sum_{i=1}^{n} h(X^{(t)}) \to E_\pi\{h(X)\}.$$

(generalization of S. Law of L.N).

pg 183. Chpt 7 MCMC.

Why? Suppose target density f can be eval. but not easily sampled. We use MCMC as a method for generating a sample from which exp of functs of $X \sim f(x)$ can be reliably estimated.

Basic idea: Create an ergodic M.C. whose stationary dist is f. then use the fact that

$$\frac{1}{n}\sum_{t=1}^{n} h(X^{(t)}) \rightarrow E_f\{h(X)\}.$$

Comments.

1. Methods often support Bayesian inference (can ignore constants of prop.).
2. We need t large enough to have reached stationarity.
3. Often times the initial behavior of the chain is not the limiting behavior → burn in period w/ realizations that are ignored
4. $X^{(0)}, X^{(1)}, \ldots$ are dependent.

How to choose a suitable Chain?

Sec 7.1
pg 184

Metropolis - Hastings Algorithm

- an acceptance rejection method.
- generate variates from density f by gen. variates from a M.C. w/ cond density $g(y|\cdot)$.

M-H Algorithm

Guess

0. For $t=0$ draw a random x_0, with $f(x_0) > 0$ & set $X^{(0)} = x_0$.

Given $X^{(t)} = x^{(t)}$ compute $X^{(t+1)}$ as follows:

1. Generate a value x^* from the proposal dist $g(\cdot \mid x^{(t)})$

2. Set r equal to the M-H ratio

$$r = R(x^{(t)}, x^*) = \frac{f(x^*)\, g(x^{(t)} \mid x^*)}{f(x^{(t)})\, g(x^* \mid x^{(t)})}$$

~~3. Generate u from U(0,1)~~

different from text

3. If $r \geq 1$, set $x^{(t+1)} = x^*$ (accept).
   o.w.
   Generate u from $U(0,1)$
   if $u < r$, set $x^{(t+1)} = x^*$ (accept)
   o.w set $x^{(t+1)} = x^{(t)}$ (reject).

4. increment t by 1 and return to step ①.

Comments:

① In step ③ we assign $X^{(t+1)}$ as follows

$$X^{(t+1)} = \begin{cases} x^* & \text{with prob } \min\{R, 1\}. \\ x^{(t)} & \text{o.w.} \end{cases}$$

② since f in $\frac{\rightarrow\infty}{\rightarrow 0}$ we only need to know up to a constant of prop.

③ If $g(x^{(t)} \mid x^*) \overset{?}{=} g(x^* \mid x^{(t)})$ called the Metropolis algorithm.
-just have density ratio.

(4) Clearly $\{X^{(t)}\}$ created by M-H is a MC since $X^{(t+1)}$ only depends on $X^{(t)}$.

(5) Whether the chain is ergodic depends on choice of g → Officially you should check. if so we know the chain has a unique limiting dist.

(6) The unique stationary dist is f for the MC.

Pf/ Suppose $X^{(t)} \sim f(x)$ and consider $x_1, x_2 \in S$ $x_1 \neq x_2$ for which $f(x_1) > 0$ & $f(x_2) > 0$.
w.l.o.g. assume $f(x_2)\, g(x_1|x_2) \geq f(x_1)\, g(x_2|x_1)$.

The unconditional joint density of $X^{(t)} = x_1$ & $X^{(t+1)} = x_2$ is $f(x_1)\, g(x_2|x_1)$.
Since we assume $X^{(t)} \sim f(x)$ and $X^{*} = x_2$ must have been the accepted guess. since $R \geq 1$.

Joint prob = Prob $X^{(t)}=x_1$ & prob $X^{*}=x_2$ and is accepted as $X^{(t+1)}$.

Also the unconditional joint density of $X^{(t)} = x_2$ and $X^{(t+1)} = x_1$ is

$$f(x_2)\, g(x_1|x_2)\, \frac{f(x_1)\, g(x_2|x_1)}{f(x_2)\, g(x_1|x_2)} = f(x_1)\, g(x_2|x_1)$$

because if we start w/ x_2 and propose $x^* = x_1$ then X^{t+1} is set = to x_2 with prob. $R(x_2, x_1)$.

∴ joint dist of $X^{(t)}$ & $X^{(t+1)}$ is symmetric.
∴ $X^{(t)}$ & $X^{(t+1)}$ have the same marginals.
∴ marginal of $X^{(t+1)}$ must be f.

7. Since lim dist of mc is f we have
$$E\{h(X)\} \approx \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)}).$$
with strong consistency. Keeping in mind
   - (a) Some people throw out burn in period
   - (b) there will be repeated points & you must keep them.

8. What makes a good proposal?
   - covers ~~s~~upport of f in reas. # of iter.
   - neither too many accept/rej.
   - seen Normal (x_t, σ^2)

     vs $X - X_t \sim$ ____

Ex/ Bayesian Inference: Binomial w/ nonstandard prior

- $Y = (Y_1, \dots, Y_n)^T$ & $Y_i \overset{iid}{\sim} \text{Bin}(1, \theta)$
- $S_n = \sum_{i=1} Y_i$
- prior $\pi(\theta) = 2\cos^2(4\pi\theta)$.

then

- posterior $\pi(\theta|Y) \sim f(Y|\theta)\pi(\theta)$
$$= \theta^{S_n}(1-\theta)^{n-S_n}\, 2\cos^2 4(\pi\theta).$$

use M-H.

proposal Normal mean $\theta^{old} = \theta$ $\theta' = \theta_{new}$
$$g(\theta'|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\tfrac{1}{2\sigma^2}(\theta'-\theta)^2\}.$$

M-H ratio ~~accept. prob~~
$$r = \frac{\pi(\theta'|Y)\, g(\theta|\theta')}{\pi(\theta|Y)\, g(\theta'|\theta)} = \frac{\theta'^{S_n}(1-\theta')^{n-S_n}\cos^2(4\pi\theta')}{\theta^{S_n}(1-\theta)^{n-S_n}\cos^2(4\pi\theta)}.$$

We can adjust proposal by adjusting σ^2.