Intro to Monte Carlo

Rebecca C. Steorts
Bayesian Methods and Modern Statistics: STA 360/601

Module 5

Studying for an Exam

1. The PhD notes
2. Homework
3. Lab (it's duplicated for a reason)
4. Class modules
5. Questions you ask me

How I write the exam

1. I sit down and I go through the slides
2. I think about what we talked about in class
3. I look at what I assigned for your for reading in the notes, homework, Hoff, and labs.
4. I think about the big concepts and I write problems to test your knowledge of this.

Where many of your went wrong

1. You derived the Normal-Normal (you wasted time).
2. You froze on the exam – it happens.
3. You didn't use your time wisely.
4. You wrote nothing down for some problems.
5. Advice: write down things that make sense. We know when you're writing things down that are wrong. (Like when I make a typo in class).

How to turn your semester around

1. Have perfect homework grades
2. Start preparing for midterm 2 now. (It's on March 10th).
3. You midterm grades were a worst case scenario.
4. There will be a curve and most of your grades will keep going up. (READ THIS AGAIN).
5. Keep working hard. (And yes, I know you're all working very hard).
6. Undergrads: stay after class for a few minutes. I want to talk to you apart from the grad students.

Where we are in Hoff, notes, class, labs

1. Homework 4: posted and due March 2, 11:55 PM.
2. Lab next week: importance and rejection sampling
3. Lab before midterm 2: Gibbs sampling
4. Class: importance sampling, rejection, and Gibbs sampling.

Goal: approximate

$$\int_X h(x)f(x)\, dx$$

that is intractable, where $f(x)$ is a probability density.

What's the problem? Typically $h(x)$ is messy!

Why not use numerical integration techniques?

In dimension $d = 3$ or higher, Monte carlo really improves upon numerical integration.

Numerical integration

- Suppose we have a d-dimensional integral.
- Numerical integration typically entails evaluating the integrand over some grid of points.
- However, if d is even moderately large, then any reasonably fine grid will contain an impractically large number of points.

Numerical integration

- Let $d = 6$. Then a grid with just ten points in each dimension will consist of 10^6 points.
- If $d = 50$, then even an absurdly coarse grid with just *two* points in each dimension will consist of 2^{50} points (note that $2^{50} > 10^{15}$).

What's happening here?

Numerical integration error rates (big Ohh concepts)

If $d = 1$ and we assume crude numerical integration based on a grid size n, then we typically get an error of order n^{-1}.

For most dimensions d, estimates based on numerical integrations required m^d evaluations to achieve an error of m^{-1}.

Said differently, with n evaluations, you get an error of order $n^{-1/d}$.

But, the Monte Carlo estimate retains an error rate of $n^{-1/2}$. (The constant in this error rate may be quite large).

Classical Monte Carlo Integration

The generic problem here is to evaluate

$$E_f[h(x)] = \int_X h(x)f(x)\,dx.$$

The classical way to solve this is generate a sample $(X_1, \ldots, X_n)$ from f.

Now propose as an approximation the empirical average:

$$\bar{h}_n = \frac{1}{n}\sum_{j=n}^{n} h(x_j).$$

Why? $\bar{h}_n$ converges a.s. (i.e. for almost every generated sequence) to $E_f[h(X)]$ by the Strong Law of Large Numbers.

Also, under certain assumptions[1], the asymptotic variance can be approximated and then can be estimated from the sample $(X_1, \ldots, X_n)$ by

$$v_n = 1/n \sum_{j=1}^{n} [h(x_j) - \bar{h}_n]^2.$$

Finally, by the CLT (for large n),

$$\frac{\bar{h}_n - E_f[h(X)]}{\sqrt{v_n}} \overset{\text{approx.}}{\sim} N(0,1).$$

(Technically, it converges in distribution).

[1]see Casella and Robert, page 65, for details

Importance Sampling

Recall that we have a difficult, problem child of a function $h(x)$!

- Generate samples from a distribution $g(x)$.
- We then "re-weight" the output.

Note: g is chosen to give greater mass to regions where h is large (the important part of the space).

This is called *importance sampling.*

Importance Sampling

Let g be an arbitrary density function and then we can write

$$I = E_f[h(x)] = \int_X h(x)\frac{f(x)}{g(x)}g(x)\,dx = E_g\left[\frac{h(x)f(x)}{g(x)}\right]. \qquad (1)$$

This is estimated by

$$\hat{I} = \frac{1}{n}\sum_{j=1}^{n}\frac{f(X_j)}{g(X_j)}h(X_j) \longrightarrow E_f[h(X)] \qquad (2)$$

based on a sample generated from g (not f). Since (1) can be written as an expectation under g, (2) converges to (1) for the same reason the Monte carlo estimator $\bar{h}_n$ converges.

The Variance

$$
\begin{aligned}
Var(\hat{I}) &= \frac{1}{n^2}\sum_i Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) &(3)\\
&= \frac{1}{n}Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) \implies &(4)\\
\widehat{Var}(\hat{I}) &= \frac{1}{n}\widehat{Var}\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right). &(5)
\end{aligned}
$$

Simple Example

Suppose we want to estimate $P(X > 5)$, where $X \sim N(0,1)$.

Naive method:

- Generate $X_1 \dots X_n \overset{iid}{\sim} N(0,1)$
- Take the proportion $\hat{p} = \bar{X} > 5$ as your estimate

Importance sampling method:

- Sample from a distribution that gives high probability to the "important region" (the set $(5, \infty)$).
- Do re-weighting.

Importance Sampling Solution

Let $f = \phi_o$ and $g = \phi_\theta$ be the densities of the $N(0,1)$ and $N(\theta,1)$ distributions (θ taken around 5 will work). Then

$$p = \int I(u > 5)\phi_o(u)\, du \tag{6}$$

$$= \int \left[I(u > 5)\frac{\phi_o(u)}{\phi_\theta(u)} \right] \phi_\theta(u)\, du. \tag{7}$$

In other words, if

$$h(u) = I(u > 5)\frac{\phi_o(u)}{\phi_\theta(u)}$$

then $p = E_{\phi_\theta}[h(X)]$.

If $X_1, \ldots, X_n \sim N(\theta, 1)$, then an unbiased estimate is $\hat{p} = \frac{1}{n}\sum_i h(X_i)$.

Simple Example Code

```
1 - pnorm(5)                       # gives 2.866516e-07
## Naive method
set.seed(1)
mySample <- 100000
x <- rnorm(n=mySample)
pHat <- sum(x>5)/length(x)
sdPHat <- sqrt(pHat*(1-pHat)/length(x)) # gives 0

## IS method

set.seed(1)
y <- rnorm(n=mySample, mean=5)
h <- dnorm(y, mean=0)/dnorm(y, mean=5) * I(y>5)
mean(h)                            # gives 2.865596e-07
sd(h)/sqrt(length(h))              # gives 2.157211e-09
```

Notice the difference between the naive method and IS method!

Harder example

Let $f(x)$ be the pdf of a $N(0,1)$. Assume we want to compute

$$a = \int_{-1}^{1} f(x)dx = \int_{-1}^{1} N(0,1)dx$$

Let $g(X)$ be an arbitrary pdf,

$$a(x) = \int_{-1}^{1} \frac{f(x)}{g(x)} g(x)\, dx.$$

We want to be able to draw $g(x) \sim Y$ easily. But how should we go about choosing $g(x)$?

Harder example

- Note that if $g \sim Y$, then $a = E[I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}]$.
- Some g's which are easy to simulate from are the pdf's of:
  - the Uniform$(-1, 1)$,
  - the Normal$(0, 1)$,
  - and a Cauchy with location parameter 0 (Student t with 1 degree of freedom).
- Below, there is code of how to get a sample from

  $$I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$$

  for the three choices of g.

Harder example

```
uniformIS <- function(sampleSize=10) {
  sapply(runif(sampleSize,-1,1),
    function(xx) dnorm(xx,0,1)/dunif(xx,-1,1)) }

cauchyIS <- function(sampleSize=10) {
  sapply(rt(sampleSize,1),
    function(xx)
    (xx <= 1)*(xx >= -1)*dnorm(xx,0,1)/dt(xx,2)) }

gaussianIS <- function(sampleSize=10) {
  sapply(rnorm(sampleSize,0,1),
    function(xx) (xx <= 1)*(xx >= -1)) }
```

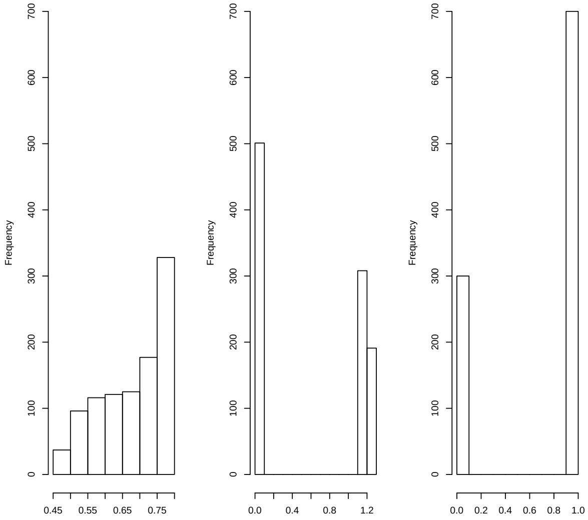

Figure 1: Histograms for samples from $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ when g is, respectivelly, a uniform, a Cauchy and a Normal pdf.

Importance Sampling with unknown normalizing constant

Often we have sample from μ, but know $\pi(x)$ except for a multiplicative $\mu(x)$ constant. Typical example is Bayesian situation:

- $\pi(x) = \nu_Y =$ posterior of $\theta \mid Y$ when prior density is ν.
- $\mu(x) = \lambda_Y =$ posterior of $\theta \mid Y$ when prior density is λ.[2]

Consider

$$\frac{\pi(x)}{\mu(x)} = \frac{c_\nu L(\theta)\nu(\theta)}{c_\lambda L(\theta)\lambda(\theta)} = c\frac{\nu(\theta)}{\lambda(\theta)} = c\,\ell(x),$$

where $\ell(x)$ is known and c is unknown.

This implies that

$$\pi(x) = c\,\ell(x)\mu(x).$$

[2] I'm motivating this in a Bayesian context. The way Hoff writes this is equivalent.

Then if we're estimating $h(x)$, we find

$$\begin{aligned}
\int h(x)\pi(x)\,dx &= \int h(x)\;c\;\ell(x)\mu(x)\;d(x) && (8)\\
&= \frac{\int h(x)\;c\;\ell(x)\mu(x)\;d(x)}{\int \pi(x)\;d(x)} && (9)\\
&= \frac{\int h(x)\;c\;\ell(x)\mu(x)\;d(x)}{\int c\;\ell(x)\mu(x)\;d(x)} && (10)\\
&= \frac{\int h(x)\;\ell(x)\mu(x)\;d(x)}{\int \ell(x)\mu(x)\;d(x)}. && (11)
\end{aligned}$$

Generate $X_1, \ldots, X_n \sim \mu$ and estimate via

$$\frac{\sum_i h(X_i)\;\ell(X_i)}{\sum_i \ell(X_i)} = \sum_i h(X_i)\left(\frac{\ell(X_i)}{\sum_j \ell(X_j)}\right) = \sum_i w_i h(X_i)$$

where $w_i = \dfrac{\ell(X_i)}{\sum_j \ell(X_j)} = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

Why the choice above for $\ell(X)$? Just taking a ratio of priors. The motivation is the following for example:

- Suppose our application is to Bayesian statistics where $\theta_1, \ldots, \theta_n \sim \lambda_Y$.
- Think of $\pi = \nu$ as a complicated prior.
- Think of $\mu = \lambda$ as a conjugate prior.
- Then the weights are $w_i = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

1. If μ and π i.e. ν and λ differ greatly most of the weight will be taken up by a few observations resulting in an unstable estimate.
2. We can get an estimate of the variance of

$$\sum_i \frac{h(X_i)\,\ell(X_i)}{\ell(X_i)}$$

   but we need to use theorems from advance probability theory (The Cramer-Wold device and the Multivariate Delta Method). These details are beyond the scope of the class.
3. In Bayesian statistics, the cancellation of a potentially very complicated likelihood can lead to a great simplification.
4. The original purpose of importance sampling was to sample more heavily from regions that are important. So, we may do importance sampling using a density μ because it's more convenient than using a density π. (These could also be measures if the densities don't exist for those taking measure theory).

Rejection Sampling

Rejection sampling is a method for drawing random samples from a distribution whose p.d.f. can be evaluated up to a constant of proportionality.

Difficulties? You must design a good proposal distribution (which can be difficult, especially in high-dimensional settings).

Uniform Sampler

Goal: Generate samples from Uniform(A), where A is complicated.

Example: $X \sim$ Uniform(Mandelbrot).

How? Consider $I_X(A)$.

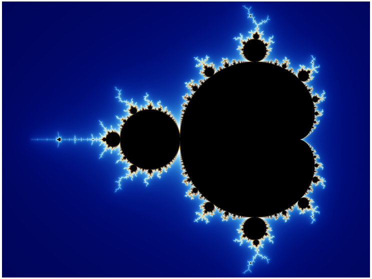

Figure 2: A complicated function A, called the Mandelbrot!

Proposition

- Suppose $A \subset B$.
- Let $Y_1, Y_2, \ldots \sim$ Uniform(B) iid and
- $X = Y_k$ where $k = \min\{k : Y_k \in A\}$,

Then it follows that

$$X \sim \text{Uniform}(A).$$

Proof: Exercise. Hint: Try the discrete case first and use a geometric series.

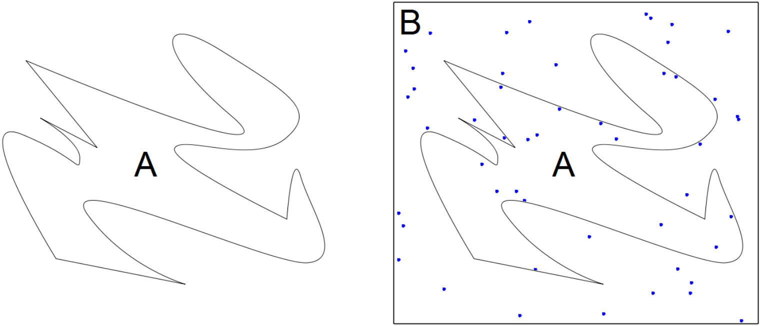

Figure 3: (Left) How to draw uniform samples from region A? (Right) Draw uniform samples from B and keep only those that are in A.

General Rejection Sampling Algorithm

Goal: Sample from a complicated pdf $f(x)$.

Suppose that

$$f(x) = \tilde{f}(x)/\alpha, \alpha > 0$$

.

Algorithm:

1. Choose a proposal distribution q such that $c > 0$ with

$$cq(x) \geq \tilde{f}(x).$$

2. Sample $X \sim q$, sample $Y \sim \text{Unif}(0, c\, q(X))$ (given X)
3. If $Y \leq \tilde{f}(X), Z = X$, we reject and return to step (2).

Output: $Z \sim f$
Proof: Exercise.

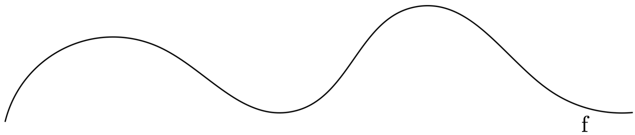

Figure 4: Visualizing just f.

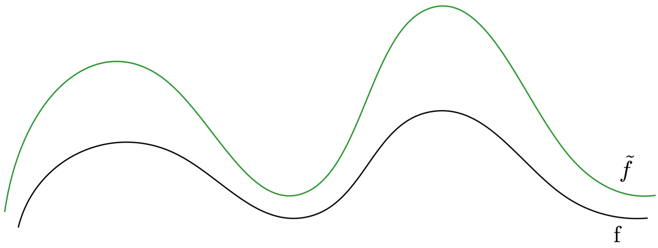

Figure 5: Visualizing just f and $\tilde{f}$.

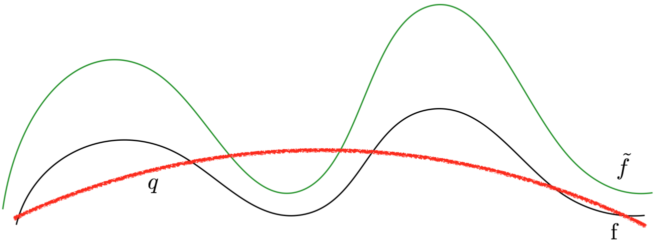

Figure 6: Visualizing f and $\tilde{f}$. Now we look at enveloping q over f.

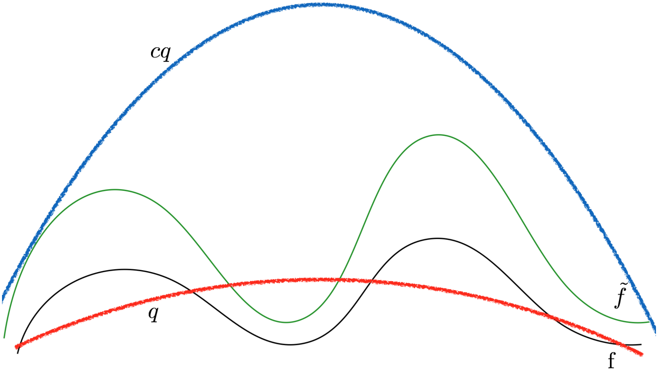

Figure 7: Visualizing f and $\tilde{f}$. Now we look at enveloping cq over $\tilde{f}$.

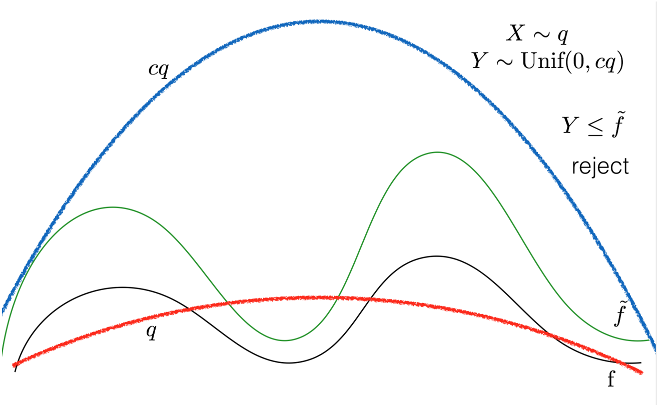

Figure 8: Recalling the sampling method and accept/reject step.

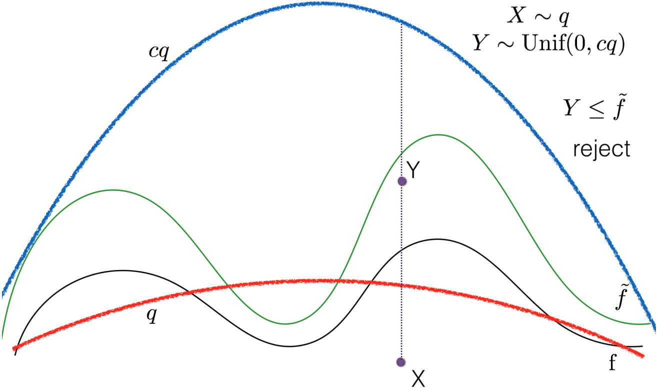

Figure 9: Entire picture and an example point X and Y.

- Suppose we want to generate random variables from the Beta(5.5,5.5) distribution.
- There are no direct methods for generating from Beta(a,b) if a,b are not integers.
- One possibility is to use a Uniform(0,1) as the trial distribution. A better idea is to use an approximating normal distribution.
- Do this as an exercise on your own.