Lecture 1.

Chp 1.1, 1.2. Liu's Book.

Note Title 4/2/2006

Goal: Teach Monte Carlo methods with some related material on optimization.

Text Book: "Monte Carlo Strategies in Scientific Computing". J.S. Liu.

Grading: 4 Homework Assignments + Final.

Motivation: Many problems in Statistics can be formulated as probabilistic inference, or as optimization.

Find $\hat{\underline{x}} = \underset{\underline{x}}{\text{ARG-MIN}}\, E(\underline{x})$ $\qquad \min_{\underline{x}} E(\underline{x}) = E(\hat{\underline{x}})$

$\hat{\underline{x}} = \underset{\underline{x}}{\text{ARG-MAX}}\, P(\underline{x} | \underline{d})$

Or as evaluating an integral

$$I = \int_D g(\underline{x})\, d\underline{x}$$

Monte Carlo gives a way to perform this

History: Developed after the 2nd world war. The Manhattan project. Computers available.

Page 2 Count de Buffon (1707-1788) use sampling to estimate π.

Early Sampling Experiment

Needle length $\ell < D$

Drop needle at random.

spacing D between lines

probability that needle will intersect a line is $\frac{2\ell}{\pi D}$. (check $\frac{2\ell}{\pi D} < 1$, because $\ell < D$).

Let p_N be the proportion of "intersects" in N samples (i.e. drop the needle N times, count no. times it intersects line - say M, set $p_N = M/N$).

$$\lim_{N \to \infty} p_N = \frac{2\ell}{\pi D}$$

Hence $$\pi = \lim_{N \to \infty} \frac{2\ell}{p_N D}$$ //

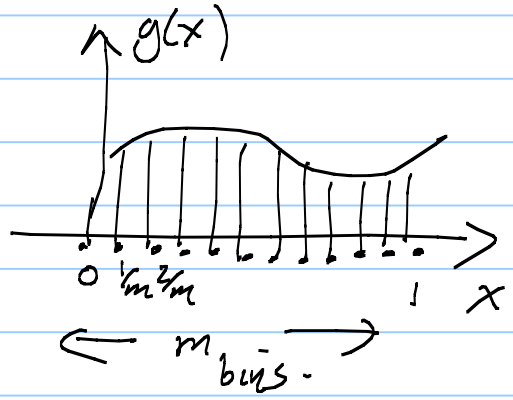
Note: Sampling has been around for 1,000's of years.

For example, Banks / Kings count number of coins in a pile by selecting a few coins at random, weighing them; then weigh the pile to estimate total number of coins.

Integration

$$I = \int_0^1 g(x)\,dx$$

$D = [0,1]$



Riemannian Approximation

Approximate I by $\tilde{I}_m = \frac{1}{m}\left\{g(1/m)+g(2/m)+\cdots+g(1)\right\}$

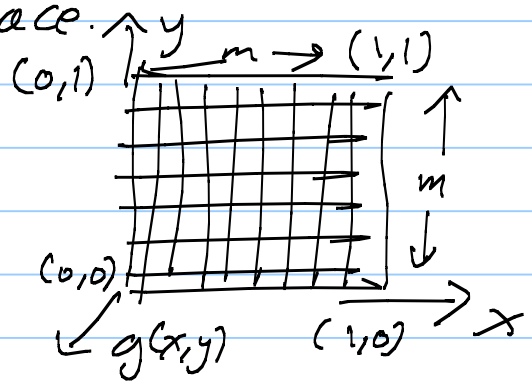
The typical error is $O(m^{-1})$. Not so bad.

As $m \to \infty$, $\tilde{I}_m \to I$.

But no. of bins increases exponentially with the dimensionality of the space.

In 2-dimensions

$D = [0,1]^2$



$$\tilde{I}_m = \frac{1}{m^2}\sum_{\substack{i=1\\ j=1}}^{m} g\left(\frac{i}{m},\frac{j}{m}\right)$$

$\tilde{I}_m \to I$, as $m \to \infty$

But requires evaluating m^2 points to get error $O(m^{-1})$

In n-dimensions, to get $O(m^{-1})$ error requires evaluating m^n points — too many!

Monte Carlo approximation

$I = \int_D \pi(x) g(x) dx$, x in n-dimensional space

$\pi(x)$ a probability distribution.

Note: If we let $\pi(x)$ be the uniform distribution

$\pi(x) = 1/|D|$, $x \in D$ so that $\int_D \pi(x) dx = 1$

then $I = \frac{1}{|D|} \int_D g(x) dx$. (Eg. if $D = [0,1]^2$, $\pi(x,y) = 1$ for $0 \le x \le 1$, $0 \le y \le 1$.)

Use Monte Carlo (MC) to draw m independent & identically distributed (i.i.d) samples from $\pi(x)$:

$$x^{(1)}, \ldots, x^{(m)}$$

Approximate I by $\hat{I}_m = \frac{1}{m} \{ g(x^{(1)}) + \ldots + g(x^{(m)}) \}$

By the law of large numbers

$$\lim_{m \to \infty} \hat{I}_m = I, \text{ with probability 1.}$$

$$(\forall \epsilon > 0, \lim_{m \to \infty} P(|\hat{I}_m - I| > \epsilon) = 0)$$

By central limit theorem: $\hat{I}_m = I + \epsilon/\sqrt{m}$, $\epsilon \sim N(0, \bar{\sigma}^2)$

$\sqrt{m}(\hat{I}_m - I) \to N(0, \bar{\sigma}^2)$ - error is $O(m)^{-1/2}$

where $N(0, \bar{\sigma}^2)$ is a zero-mean Gaussian with variance $\bar{\sigma}^2$:

$\bar{\sigma}^2 = \int_D dx \, \pi(x) (g(x) - \bar{g})^2$, with $\bar{g} = \int_D dx \, \pi(x) g(x)$

Note: error is $O(m^{-1/2})$ with only m computations and independent of the dimensions of the space.

The Monte Carlo approach gives better error rates with far fewer samples! (bonus)

But

(1.) We need to be able to draw samples from $\pi(\underline{x})$ — this is not easy (the main purpose of the course is to show how to sample from probability distributions.)

(2.) The variance σ^2 may be very large. Recall Error $\sim \sigma/\sqrt{m}$.

There are ways to reduce the variance E.G. Importance Sampling. Suppose $\pi(\underline{x})$ is uniform. Then get i.i.d samples $\underline{x}^{(1)}, \ldots, \underline{x}^{(m)}$ from distribution $\pi_1(\underline{x})$ that puts more probability on important parts of $\mathcal{D}$.

Estimate $\hat{I}_m = \frac{1}{m} \sum_{j=1}^{m} \frac{g(\underline{x}^j)}{\pi_1(\underline{x}^j)}$ | $E_{\pi_1}\left(\frac{g(\underline{x})}{\pi_1(\underline{x})}\right) = \int_{\mathcal{D}} g(\underline{x})\, d\underline{x}$

As before $\sqrt{m}\,(\hat{I}_m - I) \to N(0, \sigma_{\pi_1}^2)$

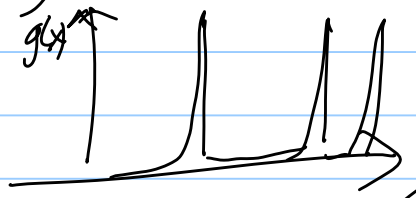
with $\sigma_{\pi_1}^2 = \mathrm{var}_{\pi_1}\left(\frac{g(\underline{x})}{\pi_1(\underline{x})}\right)$. If $\pi_1(\underline{x}) = K g(\underline{x})$ (K constant) then $\sigma_{\pi_1}^2 = 0$, ideal!

(6) Best distribution to sample from is $g(x)$ (if $\pi(x)$ is uniform) but this may not be possible.

|| Some distributions are easy to sample from but others are very difficult. ||

Notice that if $\pi_1(x)$ and $g(x)$ are very different, then the variance $\sigma^2_{\pi_1} = \text{var}_{\pi_1}\left(\frac{g(x)}{\pi_1(x)}\right)$ may be huge — so sampling might be even worse than using the Riemmann approximation.

For example, in 1-D suppose $g(x)$ is very 'spikey' and you sample from a uniform distribution $\pi_1(x) = 1/|D|$



There is very low probability that the samples from $\pi_1(x)$ will lie on the spikes — so your estimate of $\int g(x)dx$ will often be bad.

Moral → you need to pick a sampling distribution $\pi_1(x)$ which is close to $g(x)$.

So the "miracle" of MC — the ability to estimate an integral with error independent of dimension — occurs only if you know a lot about the function $g(x)$ that you want to integrate. (described in more detail in other lectures)