Monte Carlo Methods for Statistical Inference: Variance Reduction Techniques

Hung Chen

hchen@math.ntu.edu.tw

Department of Mathematics

National Taiwan University

3rd March 2004

Meet at NS 104 On Wednesday from 9:10 to 12.

Outline

- **Numerical Integration**

  1. **Introduction**
  2. **Quadrature Integration**
  3. **Composite Rules**
  4. **Richardson's Improvement Formula**
  5. **Improper integrals**

- **Monte Carlo Methods**

  1. **Introduction**
  2. **Variance Reduction Techniques**
  3. **Importance Sampling**

- **References:**

- **Lange, K. (1999)** *Numerical Analysis for Statisticians*. **Springer-Verlag, New York**
- **Robert, C.P. and Casella, G. (1999).** *Monte Carlo Statistical Methods*. **Springer Verlag.**
- **Thisted, R.A. (1996).** *Elements of Statistical Computing: Numerical Computing* **Chapman & Hall.**
- *An Introduction to R* **by William N. Venables, David M. Smith (http://www.ats.ucla.edu/stat/books/ #DownloadableBooks)**

Monte-Carlo Integration

Integration is fundamental to statistical inference.

- Evaluation of probabilities, means, variances, and mean squared error can all be thought of as integrals.
- Very often it is not feasible to solve for the integral of a given function via analytical techniques and alternative methods are adapted.
- The approximate answers are presented in this lecture.

Suppose we wish to evaluate $I = \int f(x)dx$.

- Riemann integral: The definition starts with
  - Divide $[a, b]$ into n-disjoint intervals $\triangle x_i$, such that

$\cup_i \triangle x_i = [a,b]$ **and** $\{\triangle x_i\}$ **is called a partition** $\mathcal{P}$ **of** $[a,b]$**.**

- **The mesh of this partition is defined to be the largest size of sub-intervals,** $mesh(\mathcal{P}) = \max_{\rangle} \mid \triangle \S_{\rangle} \mid$**.**
- **Define a finite sum,**

$$S_n = \sum_{i=1}^{n} f(x_i)\triangle x_i,$$

  where $x_i \in \triangle x_i$ **is any point.**
- **If the quantity** $\lim_{mesh\mathcal{P}\downarrow 0} S_n$ **exists, then it is called the integral of** f **on** $[a,b]$ **and is denoted by** $\int_a^b f(x)dx$**.**

- **This construction demonstrates that any numerical approximation of** $\int_a^b f(x)dx$ **will have two features:**

(i) Selection of samples points which partition the interval

(ii) A finite number of function evaluations on these sample points

Now consider the problem of evaluating $\theta = \int \phi(x) f(x) dx$ **where** $f(x)$ **is a density function.**
Using the law of large numbers, we can evaluate θ **easily.**

- **Sample** $X_1, \ldots, X_n$ **independently from** f **and form**

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i),$$
$$Var(\hat{\theta}) = \frac{1}{n} \int [\phi(x) - \theta]^2 f(x) dx.$$

- **The precision of $\hat{\theta}$ is proportion to $1/\sqrt{n}$.** **In numerical integration, n points can achieve the precision of $O(1/n^4)$.**

- **Question: We can use Riemann integral to evaluate definite integrals. Then why do we need Monte Carlo integration?**
  - **As the number of dimensions d increases, the number of points n required to achieve a fair estimate of integral would increase dramatically, i.e., proportional to n^d.**
  - **Even when the value d is small, if the function to integrated is irregular, then it would inefficient to use the regular methods of integration.**

- *Curse of Dimensionality:*

- **It is known that the subefficiency of numerical methods compared with simulation algorithms for dimension d larger than 4 since the error is then of order $O(n^{-4/d})$.**
- **The intuitive reason behind this phenomenon is that a numerical approach like the Riemann sum method basically covers the whole space with a grid.**

  When the dimension of the space increases, the number of points on the grid necessary to obtain a given precision increases too.

Numerical Integration

Also named as "quadrature," it is related to the evaluation of the integral

$$I = \int_a^b f(x)dx.$$

It is equivalent to solving for the value $I = y(b)$ the differential equation

$$\frac{dy}{dx} = f(x)$$

with the boundary condition $y(a) = 0$.

- **When f is a simple function, I can be evaluated easily.**
- **The underlying idea is to approximate f by a simple**

function which can be easily integrated on $[a, b]$ and which agrees with f on the sampled points.

- **The technique of finding a smooth curve passing through a set of points is also called** *curve fitting*.
  - **To implement this idea is to sample $N + 1$ points and find an order-N polynomial passing through those points.**
  - **The integral of f over that region (containing $N+1$-points) can be approximated by the integral of the polynomial over the same region.**
  - **Given $N + 1$ sample points there is a unique polynomial passing through these points, though there are several methods to obtain it. We will use the Lagrange's method to find this polynomial, which**

we will call Lagrange's interpolating polynomial.

- **Fit a polynomial to the sample points over the whole interval $[a, b]$ we may end up with a high order polynomial which itself might be difficult to determine its coefficients.**

- **Focus on a smaller region in $[a, b]$, lets say $[x_k, x_{k+n}]$, containing the points $x_k, x_{k+1}, \ldots, x_{k+n}$.**

- **Let $P_{k+i} = (x_{k+i}, f(x_{k+i}))$ be the pairs of the sampled points and the function values; they are called $knots$.**

- **Let $p_{k,k+n}(x)$ denote the polynomial of degree less than or equal to n that interpolates $P_k, P_{k+1}, \ldots, P_{k+n}$.**

Now the question becomes
Given $P_k, \ldots, P_{k+n}$, find the polynomial $p_{k,k+n}(x)$ such

that

$$p_{k,k+n}(x_{k+i}) = f(x_{k+i}), \quad 0 \le i \le n.$$

To understand the construction of $p_{k,k+n}(x)$, we look at the case $n = 0$ first.

It is the so-called the extended midpoint rule of finding I.

1. **Pick N large.**
2. **Let $x_i = a + (i - 1/2)h$ for $i = 1, \ldots, N$ where $h = (b - a)/N$.**
3. **Let $f_i = f(x_i)$.**
4. **Then $I \approx h \sum_i f_i$.**

Sample code:

```
emr<- function(f,a,b,n=1000){
 h<- (b-a)/n
```

```
  h*sum(f(seq(a+h/2,b,by=h)))
 }
```

This is the simplest thing to do.

Extending these ideas to an arbitrary value of n, the polynomial takes the form

$$p_{k,k+n}(x) = \sum_{i=0}^{n} f(x_{k+i}) L_{k+i}(x),$$

where

$$L_{k+i}(x) = \prod_{j=0, j \neq i} \frac{x - x_{k+j}}{x_{k+i} - x_{k+j}}.$$

Note that

$$L_{k+i}(x) - \begin{cases} 1, & x = x_{k+i} \\ 0, & x = x_{k+j}, j \neq i \\ else, & \textbf{in between} \end{cases}$$

Therefore, $p_{k,k+n}(x)$ satisfies the requirement that it should pass through the $n+1$ knots and is an order-n polynomial.

This leads to

$$f(x) \sim p_n(x) = f(x_k)L_k(x) + f(x_{k+1})L_{k+1}(x) + \cdots + f(x_{k+n})L_{k+n}$$

Approximate $\int_{x_k}^{x_{k+n}} f(x)dx$ by the quantity $\int_{x_k}^{x_{k+n}} p_{k,k+n}(x)dx$.

$$\int_{x_k}^{x_{k+n}} f(x)dx \approx w_k f(x_k) + w_{k+1} f(x_{k+1}) + \cdots + w_{k+n} f(x_{k+n}),$$

where

$$w_{k+i} = \int_{x_k}^{x_{k+n}} L_{k+i}(x)dx.$$

calculate Calculation of these weights to derive a few well-known numerical integration techniques.

- Assume that the sample points $x_k, x_{k+1}, \ldots$, are uniformly spaced with the spacing $h > 0$.
- Any point $x \in [x_k, x_{k+n}]$ can now be represented by $x = x_k + sh$, where s takes values $1, 2, 3, \ldots, n$ at the sample points and other values in between.
- The weight is

$$L_{k+i}(x_k + sh) = \prod_{j=0, j\neq i} \frac{s-j}{i-j}$$

or

$$w_{k+i} = h \int_0^n L_{k+i}(x_k + hs)ds.$$

- **For** $n = 1$**, the weights are given by** $w_k = h \int_0^1 sds = h/2$ **and** $w_{k+1} = h \int_0^1 (1-s)ds = h/2$**. The integral value is given by**

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx \frac{h}{2}[f(x_k) + f(x_{k+1})].$$

  This approximation is called the Trapezoidal rule, because the integral is equal to the area of the trapezoid formed by the two knots.

– **For $n = 2$, the weights are given by**

$$\begin{aligned} w_k &= h\int_0^2 \frac{1}{2}(s^2 - 3s + 2)ds = \frac{h}{3}, \\ w_{k+1} &= h\int_0^2 \frac{1}{2}(s^2 - 2s)ds = \frac{4h}{3}, \\ w_{k+2} &= h\int_0^2 \frac{1}{2}(s^2 - s)ds = \frac{h}{3}. \end{aligned}$$

The integral value is given by

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx \frac{h}{3}[f(x_k) + 4f(x_{k+1}) + f(x_{k+2})].$$

This rule is called the Simpson's $1/3$ rule.

– **For $n = 3$, we obtain Simpson's-**$3/8$ **rule given by**

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx \frac{3h}{8}\left\{f(x_k) + 3[f(x_{k+1}) + f(x_{k+2})] + f(x_{k+}\right.$$

Composite Rules:
Considering the whole space, we will divide it into n sub-intervals of equal width.

- **Utilize a sliding window on $[a, b]$ by including only a small number of these sub-intervals at a time. That is,**

$$\int_a^b f(x)dx = \sum_{k=0}^{N/n} \int_{x_k}^{x_{k+n-1}} f(x)dx,$$

where each of the integrals on the right side can be approximated using the basic rules derived earlier.

- **The summation of basic rules over sub-intervals to obtain an approximation over $[a, b]$ gives rise to composite rules.**

- **Composite Trapezoidal Rule:**

$$I \approx h \left(\frac{f(a)+f(b)}{2} + \sum_{i=1}^{n-1} f(x_i) \right) .$$

The error is this approximation is given by $\frac{-1}{12} h^2 f^{(2)}(\xi)(b-a)$ **for** $\xi \in (a,b)$**.**

- **Composite Simpson's-1=3 Rule: The number of samples** $n+1$ **should be odd, or the number of intervals should be even. The integral approximation is given by**

$$I \approx \frac{h}{3} \left(\frac{f(a)+f(b)}{2} + 4 \sum_{i \textbf{ odd}} f(x_i) + 2 \sum_{i \textbf{ even})} f(x_i) \right) .$$

The error associated with this approximation is

$$\frac{-1}{90}h^4 f^{(4)}(\xi)(b-a)^4 \quad \textbf{for } \xi \in (a,b).$$

Richardson's Improvement Formula

Suppose that we use $F[h]$ as an approximation of I computed using h-spacing. Then,

$$I = F[h] + Ch^n + O(h^m),$$

where C is a constant and $m > n$. An improvement of $F[h]$ is possible if there is a separation between n and m.

- **Eliminates the error term Ch^n by evaluating I for two different values of h and mixing the results appropriately.**
- **Assume that I is evaluated for two values of h: h_1 and h_2.**
- **Let $h_2 > h_1$ and $h_2/h_1 = r$ where $r > 1$.**

- For the sample spacing given by h_2 or rh,

$$I = F[rh_1] + Cr^n h_1^n + O(h_1^m).$$

We have

$$r^n I - I = r^n F[h_1] - F[rh_1] + O(h_1^m).$$

Rearranging,

$$I = \frac{r^n F[h] - F[rh]}{r^n - 1} + O(h^m).$$

The first term on the right can now be used as an approximation for I with the error term given by $O(h^m)$.

- This removal to Ch^n from the error using two evaluations of $f[h]$ at two different values of h is called Richardson's Improvement Formula.

This result when applied to numerical integration is called Romberg's Integration.

Improper Integrals

How do we revise the above methods to handle the following cases:

- **The integrand goes to a finite limit at finite upper and lower limits, but cannot be calculated right on one of the limits (e.g., $\sin x/x$ at $x = 0$).**
- **The upper limit of integration is ∞ or the lower limit is $-\infty$.**
- **There is a singularity at each limit (e.g., $x^{-1/2}$ at $x = 0$.)**
- **Commonly used techniques:**

  1. **Change of variables**
     For example, if $a > 0$ and $f(t) \to 0$ faster than

$t^{-2} \to 0$ as $t \to \infty$, then we can use $u = 1/t$ as follows:

$$\int_a^\infty f(t)dt = \int_0^{1/a} \frac{1}{u^2} f\left(\frac{1}{u}\right) du.$$

This also works if $b < 0$ and the lower limit is $-\infty$. Refer to Lange for additional advice and examples.

2. **Break up the integral into pieces**

Monte-Carlo Method

The main goal in this technique is to estimate the quantity θ, where

$$\theta = \int_R g(x) f(x) dx = E[g(X)],$$

for a random variable X distributed according to the density function $f(x)$.
$g(x)$ is any function on R such that θ and $E[g^2(X)]$ are bounded.
Classical Monte Carlo approach:

- **Suppose that we have tools to simulate independent and identically distributed samples from $f(x)$, call them $X_1, X_2, \ldots, X_n$, then one can approximate θ by**

the quantity:

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

- **The variance of $\hat{\theta}_n$ is given by $n^{-1}Var(g(X))$.**

For this approach, the samples from $f(x)$ are generated in an i.i.d. fashion.

In order to get a good estimate, we need that the variance goes to zero and hence the number of samples goes to infinity.

For practical situations, the sample size never goes to infinity. This raises an interesting question on whether a better estimate can be obtained with a given amount computing constraint. Now we consider three widely used techniques to accomplish this task.

Using Antithetic Variables

This method depends on generating averages from the samples which have negative covariance between them, causing overall variance to go down.

- **Let Y_1 and Y_2 be two identically distributed random variables with mean θ. Then,**

  $$Var\left(\frac{Y_1 + Y_2}{2}\right) = \frac{Var(Y_1)}{2} + \frac{Cov(Y_1, Y_2)}{2}.$$

  - **If Y_1 and Y_2 are independent, then the last term is zero.**
  - **If Y_1 and Y_2 are positively correlated then the variance of $Var((Y_1 + Y_2)/2) > Var(Y_1/2)$.**
  - **If they are negatively correlated, the resulting variance reduces.**

- **Question: How to obtain random variables Y_1 and Y_2 with identical distribution but negative correlation?**
- **Illustration:**
  - **Let $X_1, X_2, \ldots, X_n$ be independent random variables with the distribution functions given by $F_1, F_2, \ldots,$**
  - **Let g be a monotonous function.**
  - **Using the inverse transform method, the X_i's can be generated according to $X_i = F_i^{-1}(U_i)$, for $U_i \sim UNIF[0,1]$.**
  - **Define**

$$Y_1 = g(F_1^{-1}(U_1), F_2^{-1}(U_2), \ldots, F_n^{-1}(U_n)).$$

Since U and $1 - U$ are identically distributed and negatively correlated random variables, if we de-

fine

$$Y_2 = g(F_1^{-1}(1-U_1), F_2^{-1}(1-U_2), \ldots, F_n^{-1}(1-U_n)).$$

- For monotonic function g, Y_1 and Y_2 are negatively correlated.
- Utilizing negatively correlated functions not only reduces the resulting variance of the sample average but also reduces the computation time as only half the samples need to be generated from $UNIF[0,1]$.

- Estimate θ by

$$\tilde{\theta} = \frac{1}{2n} \sum_{i=1}^{n} [g(U_i) + g(1-U_i)].$$

- Other possible implementation:

- **If f is symmetric around μ, take $Y_i = 2\mu - X_i$.**
- **See Geweke (1988) for the implementation of this idea.**

Variance Reduction by Conditioning:

Let Y and Z be two random variables. In general, we have

$$Var[E(Y \mid Z)] = Var(Y) - E[Var(Y \mid Z)] \leq Var(Y).$$

- **For the two random variables Y and $E(Y \mid Z)$, both have the same mean.**
  Therefore $E(Y \mid Z)$ is a better random variable to simulate and average to estimate θ.

- **How to find an appropriate Z such that $E(Y \mid Z)$ has significantly lower variance than Y?**

- **Example: Estimate π.**

– **We can do it by** $V_i = 2U_i - 1$**,** $i = 1, 2$**, and set**

$$I = \begin{cases} 1 \text{ **if** } V_1^2 + V_2^2 \le 1 \\ 0 \text{ **otherwise.**} \end{cases}$$

– **Improve the estimate** $E(I)$ **by using** $E(I \mid V_1)$**. Note that**

$$\begin{aligned} E[I \mid V_1 = v_1] &= P(V_1^2 + V_2^2 \le 1 \mid V_1 = v) \\ &= P(V_2^2 \le 1 - v^2) = (1 - v^2)^{1/2}. \end{aligned}$$

– **The conditional variance equals to**

$$Var[(1 - V_1^2)^{1/2}] \approx 0.0498,$$

which is smaller than $Var(I) \approx 0.1686$**.**

Variance Reduction using Control Variates:

Estimate θ which is the expected value of a function g of random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

- **Assume that we know the expected value of another function h of these random variables, call it μ.**
- **For any constant a, define a random variable W_a according to**

$$W_a = g(X) + a[h(X) - \mu].$$

- **We can utilize the sample averages of W_a to estimate θ since $E(W_a) = \theta$.**
- **Observe that**

$$Var(W_a) = Var(g(X)) + a^2 Var(h(X)) + 2aCov(g(X), h(X)).$$

It follows easily that the minimizer of $Var(W_a)$ as a function of a is

$$a = -\frac{Cov(g(X), h(X))}{Var(h(X))}.$$

- Estimate θ by averaging observations of

$$g(X) - \frac{Cov(g(X), h(X))}{Var(h(X))}[h(X) - \mu].$$

- The resulting variance of W is given by

$$Var(W) = Var(g(X)) - \frac{[Cov(f(X), g(X))]^2}{Var(f(X))}.$$

- Example: Use "sample mean" to reduce the variance of estimate of "sample median."
  - Find median of a Poisson random variable X with $\lambda = 11.5$ using a set of 100 observations.

– **Note that $\mu = 11.5$.**

– **Modify the usual estimate as**

$$\tilde{x} - \frac{corr(median, mean)}{s^2}(\bar{x} - 11.5),$$

where $\tilde{x}$ and $\bar{x}$ are the median and mean of sampling data, and s^2 is the sample variance.

Example on Control Variate Method

In general, suppose there are p control variates $W_1, \ldots, W_p$ and Z generally varies with each W_i, i.e.,

$$\hat{\theta} = Z - \sum_{i=1}^{p} \beta_i [W_i - E(W_i)].$$

- Multiple regression of Z on $W_1, \ldots, W_p$.
- How do we find the estimates of correlation coefficients between Z and W's?

Importance Sampling

Another technique commonly used for reducing variance in Monte Carlo methods is importance sampling. Importance sampling is different from a classical Monte Carlo method is that instead of sampling from $f(x)$ one samples from another density $h(x)$, and computes the estimate of θ using averages of $g(x)f(x)/h(x)$ instead of $g(x)$ evaluated on those samples.

- Rearrange the definition of θ as follows:

  $$\theta = \int g(x)f(x)dx = \int \frac{g(x)f(x)}{h(x)}h(x)dx.$$

  $h(x)$ can be any density function as long as the support of $h(x)$ contains the support of $f(x)$.

- Generate samples $X_1, \ldots, X_n$ from the density $h(x)$

and compute the estimate:

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)f(X_i)}{h(X_i)}.$$

It can be seen that the mean of $\hat{\theta}_n$ is θ and its variance is

$$Var(\hat{\theta}_n) = \frac{1}{n}\left(E_h \left[\frac{g(X)f(X)}{h(X)}\right]^2 - \theta^2 \right).$$

- **Recall that the variance associated with the classical Monte Carlo estimator differs in the first term.** **In that case, the first term is given by $E_f[g(X)^2]$.**
- **It is possible that a suitable choice of h can reduce the estimator variance below that of the classical Monte Carlo estimator.**

- By Jensen's inequality, we have a lower bound on the first term:

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

$$= \left(\int g(x)f(x)dx\right)^2.$$

  In practice, for importance sampling, we generally seek a probability density h that is nearly proportional to f.

- Example taken from Robert and Casella:
  - Let X be a Cauchy random variable with parameters $(0,1)$, i.e. X is distributed according to the

density function:

$$f(x) = \frac{1}{\pi(1+x^2)},$$

and $g(x) = 1(x > 2)$ **be an indicator function.**

– **Estimate**

$$\theta = Pr(X > 2) = \frac{1}{2} - \frac{\tan 2}{\pi} = 0.1476.$$

– **Method 1:**

∗ **Generate** $X_1, X_2, \ldots, X_n$ **as a random samples from** $f(x)$**.**

∗ $\hat{\theta}_n$ **is just the frequency of sampled values larger than** 2

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} 1(X_i > 2).$$

  - $*$ **Variance of this estimator is simply $\theta(1-\theta)/n$ or $0.126/n$.**
- **Method 2:**
  - $*$ **Utilize the fact that the density $f(x)$ is symmetric around 0, and θ is just half of the probability $Pr\{\mid X \mid > 2\}$.**
  - $*$ **Generating X_i's as i.i.d. Cauchy, one can estimate θ by**

$$\hat{\theta}_n = \frac{1}{2n}\sum_{i=1}^{n} 1(\mid X_i \mid > 2).$$

  - $*$ **Variance of this estimator is $\theta(1-2\theta)/n$ or $0.052/n$.**
- **Method 3:**

- $*$ **Write θ as the following integral:**

$$\theta = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx.$$

- $*$ **Generate $X_1, X_2, \ldots, X_n$ as a random samples from $UNIF(0,2)$.**
- $*$ **Define**

$$\hat{\theta}_n = \frac{1}{2} - \frac{1}{n} \sum_i \frac{1}{\pi(1+X_i^2)}.$$

- $*$ **Its variance is given by $0.0092/n$.**

$-$ **Let $y = 1/x$ and write θ as the integral**

$$\int_0^{1/2} \frac{x^{-2}}{\pi(1+x^{-2})} dx.$$

Using i.i.d. samples from $UNIF[0,1/2]$ and evaluating average of the function $g(x) = 1/[2\pi(1+x^2)]$

once can further reduce the estimator variance.

- **Importance sampling:**
  - Select h so that its support is $\{X > 2\}$.
  - For $x > 2$,

$$f(x) = \frac{1}{\pi(1+x^2)}$$

is closely matched by

$$h(x) = \frac{2}{x^2}.$$

  - Note that the cdf associated with h is $1 - 2/x$.
  - Sampling $X = 2/U$, $U \sim U(0,1)$, and let

$$\psi(x) = 1(X > 2) \cdot \frac{f(x)}{h(x)} = \frac{1}{2\pi(1+x^{-2})}.$$

- **By $\hat{\theta}_h = n^{-1} \sum_i \psi(x_i)$, this is equivalent to Method 3.**
- $Var(\hat{\theta}_h) \approx 9.3 \times 10^{-5}/n$**.**