Introduction to Monte Carlo Integration

Lecture #6: Thursday, 16 October 1997
Lecturer: Eric Veach
Scribe: Lucas Pereira
Reviewer: Li-Yi Wei

1 The Development of Monte Carlo Methods

Monte Carlo integration methods were first used on a computer, the ENIAC, just after World War II. The ENIAC, completed in 1946, was the first electronic (as opposed to electromechanical) computer in the United States, following the Colossus completed in England in 1943. It was huge (24 meters long, 18,000 vacuum tubes), and could run at a (then) phenomenal rate of 5000 operations per second.

At that time, scientists at the Los Alamos National Laboratory were working on building an H-bomb, and were considering how the ENIAC could help them. Stanislaw Ulam suggested that random sampling could be used to simulate the flight paths of neutrons. John Von Neumann expanded on the idea, and came up with a detailed proposal in 1947. He estimated that, following 100 neutrons, they could compute one round of collisions in 3 minutes. This was fast enough so that a simulation of 100 collisions could be completed in under 5 hours. Nick Metropolis named the new method "Monte Carlo", after the city in Monaco famous for its casinos.

In 1949, Metropolis and Ulam published a paper on Monte Carlo methods, which sparked a lot of work on the methods in the 50's. Unfortunately, the computers at that time were only capable of handling trivial examples of many applications. Also, some researchers tried to apply Monte Carlo methods to every problem in sight, even those that were already well solved using previous techniques. These mis-applications gave Monte Carlo methods a bad reputation for a while. However, things gradually improved over the 60's as people discovered more efficient Monte Carlo techniques, and they learned which problems were and were not appropriate for the methods. Monte Carlo methods are now an important tool for solving many numerical problems.

2 A Brief History of Random Sampling

In isolated instances, random sampling had been used much earlier to solve numerical problems. For example, in 1777 the Comte de Buffon computed that, given parallel

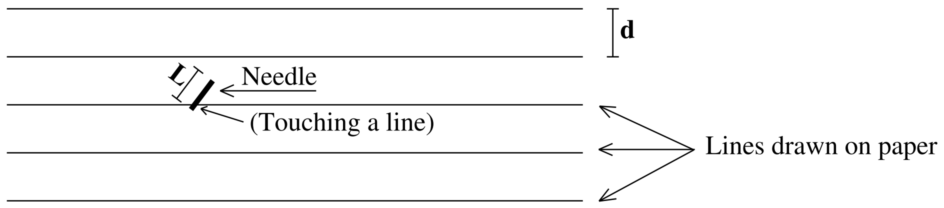

Figure 1: Using a needle to estimate π.

ruled lines separated by a distance d, and a needle of length L (where $L < d$) dropped randomly, the probability of the needle touching a line would be:

$$P_{touch} = \frac{2L}{\pi d} .$$

He verified this experimentally. Laplace, 100 years later, noticed that you could actually use this method to estimate π. In the latter part of the 19th century, this became quite a popular experiment (or party game).

Similarly, in 1900 Lord Kelvin was researching the kinetic theory of gasses, and used random sampling to help with his calculations. His random number generator consisted of writing numbers on slips of paper, putting them in a jar, and pulling them out one at a time. He was quite concerned that they weren't getting mixed well enough, due to such effects as static electricity.

As a final example, Student (an alias for W.S. Gosset) used random sampling in 1908 as an aid to deriving his famous t-distribution. He did not know the exact distribution, so he did some experiments to help him guess the analytic form.

3 Numerical Quadrature

Suppose we want to calculate some integral,

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

(We will do this example in 1-D, to keep things simple.)

The best way to calculate this integral would be to solve it analytically, and get:

$$I = \int_a^b f(x)\,dx = F(b) - F(a) .$$

However, there are many functions we would like to integrate that we cannot do analytically. So people have developed methods for approximating the integral through *quadrature rules* of the form:

$$\hat{I} = \sum_{i=1}^{n} w_i f(x_i) ,$$

which is essentially a weighted sum of samples of the function at various points. There are many different quadrature rules, each with their own sampling patterns and weights. A few common ones include:

3.1 Midpoint Rule:

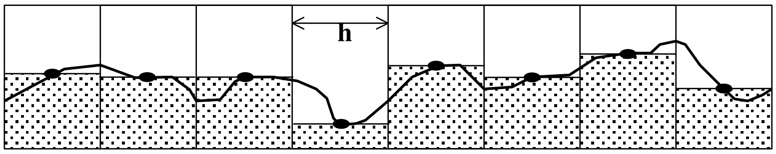

Figure 2: The midpoint rule.

We divide up the interval into some fixed number n of intervals, each of size $h = (b-a)/n$. We then choose one sample point at the midpoint of each interval:

$$\begin{aligned}\hat{I} &= h\sum_{i=1}^{n} f(a + (i - \frac{1}{2})h) \\ &= h\left[f(a + \frac{h}{2}) + f(a + \frac{3h}{2}) + \cdots + f(b - \frac{h}{2})\right] .\end{aligned}$$

3.2 Trapezoid Rule:

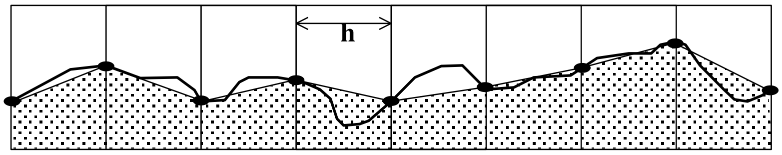

Figure 3: The trapezoid rule.

This is similar to the midpoint rule, except we sample the function at the ends of each interval, and compute the area of a trapezoid for each interval.

$$\begin{aligned}\hat{I} &= \sum_{i=1}^{n} \frac{h}{2}\left[f(a + (i-1)h) + f(a + ih)\right] \\ &= h\left[\frac{1}{2}f(a) + f(a+h) + f(a+2h) + \cdots + f(b-h) + \frac{1}{2}f(b)\right] .\end{aligned}$$

3.3 Simpson's Rule:

This is similar to the trapezoid rule, except we compute the area under a quadratic polynomial approximation (instead of a linear approximation for the trapezoid). The equation is:

$$\hat{I} = h\left[\frac{1}{3}f(a) + \frac{4}{3}f(a+h) + \frac{2}{3}f(a+2h) + \frac{4}{3}f(a+3h) + \frac{2}{3}f(a+4h) + \cdots + \frac{4}{3}f(b-h) + \frac{1}{3}f(b)\right] .$$

3.4 Convergence:

The Midpoint Rule is exact for constant or linear functions ($f \sim 1, x$). Its error is given by

$$\hat{I} - I = -\frac{(b-a)^3}{24n^2} f''(\xi) = O(n^{-2})$$

where $a \leq \xi \leq b$, provided that f has at least two continuous derivatives on $[a, b]$. For the trapezoid rule, the error is

$$\hat{I} - I = \frac{(b-a)^3}{12n^2} f''(\xi^*) = O(n^{-2}) .$$

So, the midpoint and trapezoid rules have the same convergence rate, and their errors have opposite signs if $f'' > 0$ (i.e. the true answer is bounded between them).

Simpson's Rule is exact for polynomial functions up to cubics ($f \sim 1, x, x^2, x^3$). The error can be bounded by the fourth derivative:

$$|\hat{I} - I| = \frac{(b-a)^5}{180(2n)^4} f^{(4)}(\xi) = O(n^{-4}) .$$

This converges very quickly, assuming that f has a continuous fourth derivative. There are higher-order rules called *Newton-Cotes* rules that can achieve even faster convergence, but require the function to be even smoother. Another popular family of integration rules are the *Gauss-Legendre rules*, which optimize the sample locations as well as the weights to get better convergence.

3.5 Multi-Dimensional Integration:

Multi-dimensional integration is more complex. A common way to extend a one-dimensional quadrature rule is to use a *tensor product rule*. These have the form

$$\hat{I} = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_s=1}^{n} w_{i_1} w_{i_2} \cdots w_{i_s} f(x_{i_1}, x_{i_2}, \ldots, x_{i_s}) ,$$

where s is the dimension, and the w_i and x_i are the weights and sample locations for a given one-dimensional quadrature rule. For example, if we needed to evaluate the 5-D integral

$$I = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3, x_4, x_5)\, dx_1\, dx_2\, dx_3\, dx_4\, dx_5 ,$$

we would compute a sum of the form

$$I = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} \sum_{i_5=1}^{n} w_{i_1} w_{i_2} w_{i_3} w_{i_4} w_{i_5} f(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}, x_{i_5}) \; .$$

The 2-D version of Simpson's rule looks like this:

$$h^2 \cdot \frac{1}{9} \cdot \begin{bmatrix} 1 & 4 & 2 & 4 & 2 & \cdots & 4 & 1 \\ 4 & 16 & 8 & 16 & 8 & \cdots & 16 & 4 \\ 2 & 8 & 4 & 8 & 4 & \cdots & 8 & 2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 4 & 16 & 8 & 16 & 8 & \cdots & 16 & 4 \\ 1 & 4 & 2 & 4 & 2 & \cdots & 4 & 1 \end{bmatrix}$$

However, this does not work very well in high dimensions. If we start with an n-point quadrature rule in 1-D, we need $N = n^s$ sample points for an s-dimensional integral. Thus we can see that when the dimensionality is large (say $s = 20$), the number of required samples grows astronomically:

$$\begin{aligned} 1^{20} &= 1 \\ 2^{20} &= 1048576 \\ 3^{20} &= 3486784401 \\ 4^{20} &= 1099511627776 \\ \cdots \end{aligned}$$

Furthermore, suppose that the 1-D rule has a convergence rate of $O(n^{-r})$. The s-dimensional rule doesn't work any better than the one-dimensional rule along each axis, so it converges at $O(n^{-r})$ as well. However, the total number of sample points used is much larger ($N = n^s$), so that in terms of the total number of samples the convergence is only $O(N^{-r/s})$. For example, the rate of convergence for Simpson's method is:

$$\begin{aligned} s = 1 :&\quad O(n^4) \\ s = 2 :&\quad O(n^2) \\ s = 4 :&\quad O(n^1) \\ s = 8 :&\quad O(n^{1/2}) \\ &\cdots \end{aligned}$$

So even with Simpson's assumption of 4 continuous derivatives, the convergence is already looking bad for high dimensions.

If we throw in a discontinuity in f, things get even worse. Consider the function

$$f(x) = \begin{cases} 1 & \text{if } x < X^* \\ 0 & \text{if } x > X^* \end{cases} \; .$$

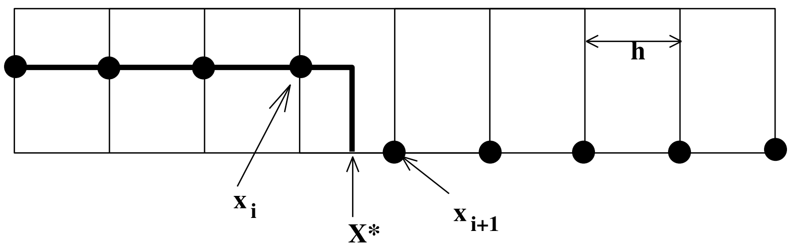

Figure 4: A discontinuous f.

and suppose that the quadrature points x_i are evenly spaced.

We see that the value computed the quadrature rule does not change when X^* is anywhere between x_i and x_{i+1}. Thus the error of *any* fixed quadrature rule is directly proportional to $h = 1/n$. So if f has a discontinuity, the convergence rate is $O(n^{-1})$ even in one dimension. Larger dimensions only compound this problem, leading to a convergence rate of $O(N^{-1/s})$.

3.6 Bakhvalov's Theorem:

There is an important result which limits the convergence rate for any deterministic quadrature rule, called *Bakhvalov's theorem.* Essentially, it says that given any s-dimensional quadrature rule, there is function f with r continuous, bounded derivatives, for which the convergence rate of the chosen quadrature rule is only $O(N^{-r/s})$.

Specifically, let C^r_M denote the set of function on $[0,1]^s$ with r continuous, bounded derivatives. That is, we require

$$\left| \frac{\partial^r}{\partial x_1^{a_1} \cdots \partial x_s^{a_s}} f \right| \leq M$$

for all $a_1, \ldots, a_s$ such that $\sum a_i = r$.

Now consider any N-point quadrature rule, and let

$$\hat{I}(f) = \sum_{i=1}^{N} w_i f(x_i)$$

be the approximation to the true integral

$$I(f) = \int_{[0,1]^s} f(x_1, \ldots, x_s)\, dx_1 \cdots dx_s \; .$$

Then there exists a function $f \in C^r_M$ such that the error is

$$\left| \hat{I}(f) - I(f) \right| > k \cdot N^{-r/s} \; ,$$

where $k > 0$ depends only on M and r.

4 Monte Carlo Integration

The basic Monte Carlo method (in 1-dimension, for simplicity) is:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{N}\sum_{i=1}^{N} f(X_i) \tag{1}$$

where the points X_i are chosen independently and uniformly at random on the interval $[a,b]$. As we will see, this method has a convergence rate of $O(N^{-1/2})$ in any dimension, *regardless* of the smoothness of the function f. This is particularly useful in graphics, where we often need to calculate multi-dimensional integrals of discontinuous functions, where Simpson's rule and Gaussian quadrature don't do well.

4.1 A Bit of Probability Theory

First we review some terms from that probability course you took a long time ago.

A *random variable* X is simply a quantity that is chosen by some random process. (This is good enough for our purposes; to define it more precisely involves a lot of math.)

Given a random variable X, its **cumulative distribution function** $P(x)$ is defined by

$$P(x) = Pr\{X_i \leq x\}$$

i.e. $P(x)$ is the probability that the random variable is not greater than a given value x.

So, for example, the cumulative distribution function $U[0,1]$ is:

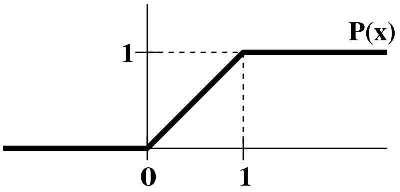

Figure 5: The cumulative distribution of $U[0,1]$.

Next, we define the *probability density function* $p(x)$:

$$p(x) = \frac{dP(x)}{dx}\ ,$$

which is often just called a *density function* or *pdf.* From these definitions, we get the important relationship

$$Pr\{\alpha \leq X \leq \beta\} = \int_\alpha^\beta p(x)\,dx = P(\beta) - P(\alpha)\ .$$

Let $Y = f(X)$ (note that Y is also a random variable). The **expectation** of Y is defined as

$$E[Y] = \int f(x)p(x)\,dx\ ,$$

where $p(x)$ is the density function for X. The **variance** of Y is defined as

$$V[Y] = E[(Y - E[Y])^2] \, ,$$

which is to say, the variance of Y is the expected squared value of the difference between Y and its mean.

From these definitions, it is easy to see that

$$E[aY] = a\, E[Y] \qquad \text{and} \qquad V[aY] = a^2\, V[Y]$$

for any constant a. Another important rule that always holds is

$$E\left[\sum_i Y_i\right] = \sum_i E[Y_i] \, .$$

For example, we can use these rules to derive a simpler expression for the variance:

$$\begin{aligned} V[Y] &= E[(Y - E[Y])^2] \\ &= E[Y^2 - 2Y E[Y] + E[Y]^2] \\ &= E[Y^2] - E[Y]^2 \end{aligned}$$

noting that the inner $E[Y]$'s are just constants.

The following rule, however, holds only when the Y_i are independent:

$$V\left[\sum_i Y_i\right] = \sum_i V[Y_i]$$

Two variables Y_1 and Y_2 are defined to be **independent** if:

$$Pr\{Y_1 \leq x_1 \text{ and } Y_2 \leq x_2\} = Pr\{Y_1 \leq x_1\} \cdot Pr\{Y_2 \leq x_2\}$$

for all values of x_1 and x_2.

4.1.1 Basic Monte Carlo

Let's return now to the basic Monte Carlo estimate (1), namely

$$F_N = \frac{b-a}{N} \sum_{i=1}^{N} f(X_i) \, . \tag{2}$$

Our first task is to show that this gives the correct answer on average, i.e.

$$E[F_N] = \int_a^b f(x)\, dx = I \, .$$

We have used the notation F_N (rather than $\hat{I}$) to emphasize that the result is a random variable, and that its properties depend on how many sample points are chosen.

Recall that th X_i are independent samples from $U[a,b]$ (the uniform distribution on $[a,b]$). The density function for each X_i is

$$p(x) = \begin{cases} 1/(b-a) & \text{when } a \le x \le b \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

We can now compute the expected value of F_N:

$$\begin{aligned} E[F_N] &= E\left[\frac{b-a}{N}\sum_{i=1}^{N} f(X_i)\right] \\ &= \frac{b-a}{N}\sum_{i=1}^{N} E[f(X_i)] \\ &= \frac{b-a}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} f(x)p(x)\,dx \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{a}^{b} f(x)\,dx \\ &= I \ , \end{aligned}$$

which is what we wanted.

It is actually not necessary to choose the samples uniformly to make this work. Suppose that rather than choosing the X_i uniformly, we choose them according to some arbitrary density function $p(x)$ on the interval $[a,b]$. (We will discuss how to do this in the next lecture.) Now consider the estimate

$$F'_N = \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)}{p(X_i)} \ , \tag{3}$$

i.e. we compute f/p at each sample point and average all of these values together. The expected value of F'_N is

$$\begin{aligned} E[F'_N] &= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\frac{f(x)}{p(x)}\,p(x)\,dx \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{a}^{b} f(x)\,dx \\ &= I \ . \end{aligned}$$

This is an important generalization that is often used in practice. The only condition we require on the density $p(x)$ is that $p(x) > 0$ whenever $f(x) \neq 0$. The easiest way to achieve this is to have $p(x) > 0$ on the whole interval $[a,b]$. (If $p(x) = 0$ for some interval where $f(x) \neq 0$, then we will miss sampling some of the integral.)

4.2 Convergence of Monte Carlo Methods:

From the previous discussion, we know that F_N gives the right answer on average. However, we would also like to know how large an error we can expect, and what the rate of convergence is.

To simplify the notation, let $Y_i = f(X_i)/p(X_i)$, so that the estimate F_N becomes

$$F_N = \frac{1}{N} \sum_{i=1}^{N} Y_i \ .$$

We also let Y be a synonym for Y_1. The value that we are trying to approximate is $E[Y] = I$.

First, we have the **strong law of large numbers**, which says that given enough samples, the mean will converge to the expected value with probability 1:

$$Pr\left\{ \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Y_i = E[Y] \right\} = 1$$

assuming that the Y_i are i.i.d. (independent and identically distributed). This says that if we take enough samples, F_N is guaranteed to converge to the correct answer.

To determine the rate of convergence, we need **Chebychev's inequality**:

$$Pr\left\{ |F - E[F]| \geq \left(\frac{V[F]}{\delta}\right)^{1/2} \right\} \leq \delta \ ,$$

where F is any random variable such that $V[F] < \infty$. We will apply this inequality to F_N to get a bound on the error.

First, we compute the variance of F_N is terms of the variance of Y (a single sample):

$$\begin{aligned} V[F_N] &= V\left[\frac{1}{N}\sum_{i=1}^{N} Y_i\right] \\ &= \frac{1}{N^2} V\left[\sum_{i=1}^{N} Y_i\right] \\ &= \frac{1}{N^2} \sum_{i=1}^{N} V[Y_i] \\ &= \frac{1}{N} V[Y] \end{aligned}$$

where we have used $V[aY] = a^2 V[Y]$ and the fact that the Y_i are independent samples. Thus, the variance decreases linearly with N.

Plugging this into Chebychev's inequality, we get

$$Pr\left\{ |F_N - I| \geq N^{-1/2} \left(\frac{V[Y]}{\delta}\right)^{1/2} \right\} \leq \delta \ .$$

Thus for any fixed threshold δ, we see that the error decreases at the rate $N^{-1/2}$.

It is possible to get much tighter bounds on the error using the **Central Limit Theorem**, which states that F_N converges to a normal distribution as $N \to \infty$. It is most conveniently stated in terms of the *standard deviation* σ_{F_N} of F_N, which is simply the square root of the variance:

$$\sigma_{F_N} = (V[F_N])^{1/2} = N^{-1/2}\sigma_Y \ .$$

The standard deviation itself is a common measurement of error (it is sometimes called RMS error), and notice that it also converges at the $O(N^{-1/2})$ rate.

The central limit theorem then states that

$$\lim_{N\to\infty} Pr\left\{\frac{1}{N}\sum_{i=1}^{N} Y_i - E[Y] \le t\,\frac{\sigma_Y}{\sqrt{N}}\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\,dx \ ,$$

where the expression on the right is the *normal distribution* (the integral of the familiar "bell curve" or Gaussian).

This equation can be rearranged to give

$$Pr\left\{|F_N - I| \ge t\sigma_{F_N}\right\} = \sqrt{\frac{2}{\pi}}\int_{t}^{\infty} e^{-x^2/2}\,dx \ .$$

The integral on the right decreases very quickly with t; for example when $t = 3$ the right-hand side is approximately 0.003. Thus, there is only about a 0.3% chance that F_N will differ from its mean by more than three standard deviations, provided that N is large enough for the central limit theorem to hold. Recall that the standard deviation is $\sigma_{F_N} = O(N^{-1/2})$, so this also verifies the $O(N^{-1/2})$ convergence.

4.3 Properties of Estimators:

The purpose of a Monte Carlo estimator is to approximate the value of some *quantity of interest* Q. Normally Q will be the value of an integral, although more general situations are possible (e.g. Q could be the ratio of two integrals).

An *estimator* is now a function of the form

$$F_N \;=\; F_N(X_1, \ldots, X_N) \ , \tag{4}$$

where the X_i are random variables. A particular numerical value of F_N is called an *estimate*. So far, we have only considered estimators where the X_i are independent and identically distributed (i.i.d.). However, in general the X_i can depend on each other, and they can have different distributions.

An estimator F_N is called *unbiased* if

$$E[F_N] = Q \qquad \text{for all values of } N \ .$$

The estimators (2) and (3) we have already discussed are unbiased.

Otherwise, the *bias* of the estimator is defined as

$$\beta[F_N] = E[F_N] - Q \ .$$

If the bias goes to zero as N increases, then the estimator is called *consistent*:

$$\lim_{N\to\infty} \beta[F_N] = 0 \ ,$$

or equivalently

$$\lim_{N\to\infty} E[F_N] = Q \ .$$

Such an estimator will ultimately converge to the right answer, but it may take many samples.

4.3.1 Example: a biased, consistent estimator

Suppose we are attempting to do antialiased sampling on a (1-D) pixel. To determine the pixel value, we need to evaluate an integral

$$I = \int_0^1 w(x) f(x)\, dx \tag{5}$$

where f is the image function, and w is the filter function, which is assumed to integrate to one:

$$\int_0^1 w(x)\, dx = 1 \ .$$

For example, w could be a tent filter:

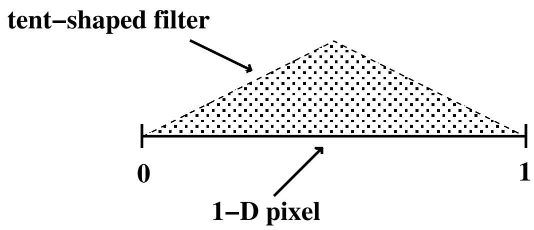

Figure 6: Antialiasing samples in a 1-D pixel with a tent filter.

In graphics, a common way to evaluate this integral is

$$F_N = \frac{\sum_{i=1}^N w(X_i) f(X_i)}{\sum_{i=1}^N w(X_i)}$$

where the X_i are uniformly distributed on $[0,1]$. This is an example of a biased estimator, since when $N = 1$ we get

$$E[F_1] = E\left[\frac{w(X_1) f(X_1)}{w(X_1)}\right] = E[f(X_1)] = \int_0^1 f(x)\, dx \ .$$

This is not the same as the desired answer I given by (5), so the estimator is biased. Similarly, when $N = 2$ we get

$$E[F_2] = \int_0^1 \int_0^1 \frac{w(x_1)f(x_1) + w(x_2)f(x_2)}{w(x_1) + w(x_2)}\, dx_1\, dx_2 \ ,$$

which is difficult to evaluate but is certainly not equal to I. In general, this estimator is biased for every value of N.

However, as $N \to \infty$ the bias goes to zero. To see this, we rewrite F_N as

$$F_N = \frac{(1/N)\sum_{i=1}^N w(X_i)f(X_i)}{(1/N)\sum_{i=1}^N w(X_i)}$$

and then evaluate

$$\begin{aligned} \lim_{N\to\infty} E[F_N] &= \frac{\lim_{N\to\infty}(1/N)\sum_{i=1}^N w(X_i)f(X_i)}{\lim_{N\to\infty}(1/N)\sum_{i=1}^N w(X_i)} \\ &= \frac{\int_0^1 w(x)f(x)\,dx}{\int_0^1 w(x)\,dx} \\ &= \int_0^1 w(x)f(x)\,dx \\ &= I \ . \end{aligned}$$

Next time, we will look at some other important properties of estimators, including the *mean squared error* and *efficiency* of an estimator.