Monte Carlo Integration

In these notes we first review basic numerical integration methods (using Riemann approximation and the trapezoidal rule) and their limitations for evaluating multidimensional integrals. Next we introduce stochastic integration methods based on Monte Carlo and importance sampling. We conclude with a section on computationally efficient generation of random numbers, when the sampling density is known up to a normalizing constant.

An excellent reference for this material is the book by Robert and Casella [1]. These stochastic methods have found numerous applications in engineering; see for instance the papers in the 2002 special issue of the IEEE Transactions on Signal Processing [2].

1 Riemann Integration

Consider the problem of evaluating an integral $I = \int_a^b \phi(x)\,dx$. The Riemann approximation to I is given by

$$\hat{I}_n = \sum_{i=1}^{n} (x_i - x_{i-1})\phi(x_i) \tag{1}$$

where $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. This may be viewed as approximating $\phi(x)$ with a piecewise-constant function $\hat{\phi}_n(x)$ which is equal to $\phi(x_i)$ for all $x \in [x_i, x_{i-1}]$ and $1 \le i \le n$. Indeed $\hat{I}_n = \int \hat{\phi}_n$. Assuming that the derivative $\phi'(x)$ is bounded, and that $x_i = a + (b-a)\frac{i}{n}$, the maximum absolute error due to this approximation is upper bounded as

$$|\phi(x) - \hat{\phi}_n(x)| \le \frac{1}{n}(b-a)\|\phi'\|_\infty$$

with equality if $\phi(x)$ is an affine function. Hence the error incurred by approximating the integral with a Riemann sum is at most $|\hat{I}_n - I| \le \frac{C}{n}$ for some constant $C = (b-a)^2\|\phi'\|_\infty$ independent of n.

2 Trapezoidal Rule

The approximation formula (1) can be improved by replacing $\phi(x_i)$ with $\frac{1}{2}[\phi(x_i) + \phi(x_{i-1})]$:

$$\hat{I}_n = \sum_{i=1}^{n} (x_i - x_{i-1})\,\frac{1}{2}[\phi(x_i) + \phi(x_{i-1})]. \tag{2}$$

This is the so-called trapezoidal rule, which is extensively used for numerical integration. For instance, if $\phi(x)$ is an affine function, the approximation is *exact*. For general functions $\phi(x)$, the approximation error is due to the curvature of ϕ. If the second derivative $\phi''(x)$ exists and is bounded, it may be shown (by application of Taylor's theorem again) that $|\hat{I}_n - I| \le \frac{C}{n^2}$ for some constant C.

3 Multidimensional Integration

For d-dimensional integrals, $\mathcal{X}$ is a subset of $\mathbb{R}^d$. An integral can be approximated by a Riemann sum, similarly to Sec. 1, or using a trapezoidal rule as in Sec. 2. If a n-point approximation is used, the trapezoidal rule yields an approximation error $|\hat{I}_n - I| \leq \frac{C}{n^{2/d}}$ for some constant C. This is the same formula as in 1D, except that n is replaced with $n^{1/d}$ (the number of points per coordinate in case $\mathcal{X}$ is discretized using a cubic lattice). Hence n needs to increase *exponentially* with d to achieve a target approximation error. This phenomenon is known as the *curse of dimensionality.*

The stochastic methods for numerical integration avoid the curse of dimensionality, as the resulting integrals may be approximated with an accuracy of the order of $1/\sqrt{n}$, where n is the number of samples $X_1, \cdots, X_n$ taken from $\mathcal{X}$. Hence *the stochastic methods outperform the deterministic ones for dimensions* $d > 4$ and are worse for $d < 4$.

4 Classical Monte Carlo Integration

The basic problem considered in this section and the following one is as follows. Given a pdf $f(x)$, $x \in \mathcal{X}$ and a function $h(x)$, $x \in \mathcal{X}$, evaluate the integral

$$\mu = \mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x) f(x)\, dx.$$

Note these methods can be used to evaluate any integral $I = \int_{\mathcal{X}} \phi(x)\, dx$ by expressing ϕ as the product of a pdf f and another function h.

Given $X_1, X_2, \cdots, X_n$ drawn iid from the pdf f, estimate μ by the empirical average

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

By the strong law of large numbers, we have $\hat{\mu}_n \overset{a.s}{\to} \mu$ as $n \to \infty$. The variance of $\hat{\mu}_n$ is

$$\mathrm{Var}(\hat{\mu}_n) = \frac{1}{n} \mathrm{Var}[h(X)] = \frac{1}{n} \int_{\mathcal{X}} (h(x) - \mu)^2 f(x)\, dx.$$

We will henceforth assume that $\mathbb{E}_f[h^2(X)] < \infty$.

Example. Let f be the Cauchy distribution, $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$, and $h(x)$ the indicator function for the interval $[0, 2]$. We have

$$\mu = \int_0^2 \frac{1}{\pi(1+x^2)}\, dx \approx 0.35.$$

The estimator of μ is given by

$$\hat{\mu}_n = \frac{1}{n} 1_{\{0 \le X_i \le 2\}}.$$

Its variance is

$$\mathrm{Var}(\hat{\mu}_n) = \frac{\mu(1-\mu)}{n} \approx \frac{0.21}{n}.$$

This method is intuitively inefficient because only 35% of the samples contribute to the sum giving $\hat{\mu}_n$. Can we do better?

5 Importance Sampling

The idea here is to drawn samples not from f, but from an auxiliary pdf g (often called *instrumental density*). Specifically, given $X_1, X_2, \cdots, X_n$ drawn iid from the pdf g, estimate μ by the empirical average

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

Clearly this method reduces to standard Monte-Carlo if $g = f$. It is required that $\mathrm{supp}\{f\} \subseteq \mathrm{supp}\{g\}$, i.e., $f(x) > 0 \Rightarrow g(x) > 0$.

By the strong law of large numbers, we have

$$\hat{\mu}_n \overset{a.s}{\to} \mathbb{E}_g \left[\frac{f(X)}{g(X)} h(X) \right] = \int_{\mathcal{X}} f(x) h(x)\, dx = \mu$$

as $n \to \infty$. Hence the estimator remains unbiased. Its variance is

$$\begin{aligned} \mathrm{Var}_g(\hat{\mu}_n) = \frac{1}{n} \mathrm{Var}_g \left[\frac{f(X)}{g(X)} h(X) \right] &= \frac{1}{n} \left\{ \mathbb{E}_g \left(\frac{f(X)}{g(X)} h(X) \right)^2 - \mu^2 \right\} \\ &= \frac{1}{n} \left\{ \int_{\mathcal{X}} \frac{f^2(x)}{g(x)} h^2(x)\, dx - \mu^2 \right\} \end{aligned}$$

which generally differs from $\mathrm{Var}_f(\hat{\mu}_n)$. The idea of importance sampling is to find a good g such that

$$\mathrm{Var}_g(\hat{\mu}_n) < \mathrm{Var}_f(\hat{\mu}_n).$$

For the Cauchy example above, consider the uniform pdf over $[0, 2]$:

$$g(x) = \frac{1}{2} 1_{\{0 \le x \le 2\}} = \frac{1}{2} h(x).$$

Then we have

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{2}{\pi(1 + X_i^2)}.$$

The variance of this estimator is

$$\mathrm{Var}_g(\hat{\mu}_n) = \frac{1}{n}\left\{\int_0^2 2f^2(x)\,dx - \mu^2\right\} \approx \frac{0.009}{n}$$

i.e., about 20 times smaller than $\mathrm{Var}_f(\hat{\mu}_n)$!

In principle one may seek g that minimizes $\mathrm{Var}_g(\hat{\mu}_n)$ over all possible pdf's. The solution is otained using the method of Lagrange multipliers: minimize the Lagrangian

$$\begin{aligned} L(g,\lambda) &= \mathrm{Var}_g(\hat{\mu}_n) + \lambda\int_{\mathcal{X}} g(x)\,dx \\ &= \int_{\mathcal{X}} \frac{v(x)}{g(x)}\,dx + \lambda\int_{\mathcal{X}} g(x)\,dx \end{aligned}$$

where λ is the Lagrange multiplier, and we have used the shorthand $v(x) = f^2(x)h^2(x)$. Taking the Fréchet derivative of $L(g,\lambda)$ with respect to $g(x)$, we obtain

$$0 = \frac{\partial L(g,\lambda)}{\partial g(x)} = -\frac{v(x)}{g^2(x)} + \lambda, \quad x \in \mathcal{X}$$

whence

$$g(x) = \sqrt{v(x)/\lambda} = \frac{f(x)|h(x)|}{\int_{\mathcal{X}} f(x)|h(x)|\,dx}$$

where the value of λ was selected to ensure that $\int g = 1$. The expression above is elegant, however evaluating $g(x)$ requires computation of the integral in the numerator, which is as hard as the original problem! In practice thus one is content to find a "good" g that assigns high probability to regions where $f(x)|h(x)|$ is large. Ideally the ratio $\frac{f(x)\,|h(x)|}{g(x)}$ would be roughly constant over $\mathcal{X}$.

6 Random Number Generation

A classical method for generating a real random variable X from an arbitrary cdf $F(x)$ is to generate a random variable U uniformly distributed over $[0,1]$ and then apply the inverse cdf to U, resulting in $X = F^{-1}(U)$ with the desired distribution. Indeed

$$Pr[X \le x] = Pr[U \le F(x)] = F(x).$$

Now suppose the pdf $f(x)$ is known up to a normalization constant which is difficult or expensive to compute. An example is when samples have to be generated from a posterior distribution $f(x|y) = \frac{p(y|x)\,p(x)}{\int p(y|x)\,p(x)\,dx}$, where the integral in the denominator is the normalization constant.

A good method in this case is the so-called *Accept-reject* method [1, Ch. 2.3]. We are given an auxiliary pdf $g(x)$ which is easy to sample, and a constant M such that $\frac{f(x)}{Mg(x)} \le 1$ holds and is easy to evaluate for all $x \in \mathrm{supp}(f)$. The Accept-reject method works as follows:

(1) Generate independent random variables $X \sim g$ and $U \sim \text{Uniform}\,[0,1]$.

(2) Accept $Y = X$ if $U \leq \frac{f(X)}{Mg(X)} \leq 1$.
Return to (1) otherwise.

Claim: $Y \sim f$.

Proof: The cdf for Y is

$$\begin{aligned} Pr[Y \leq y] &= Pr\left[X \leq y \;\middle|\; U \leq \frac{f(X)}{Mg(X)}\right] \\ &= \frac{Pr\left[X \leq y,\; U \leq \frac{f(X)}{Mg(X)}\right]}{Pr\left[X \leq \infty,\; U \leq \frac{f(X)}{Mg(X)}\right]} = \frac{N(y)}{N(\infty)}. \end{aligned} \tag{3}$$

The numerator of (3) takes the form

$$\begin{aligned} N(y) &= \int_{-\infty}^{y} dx\, g(x) \int_{0}^{\frac{f(x)}{Mg(x)}} du \\ &= \frac{1}{M} \int_{-\infty}^{y} f(x)\, dx \end{aligned}$$

hence $N(\infty) = \frac{1}{M}$. Substituting back into (3), we obtain $Pr[Y \leq y] = \int_{-\infty}^{y} f(x)\, dx$, which proves the claim. □

As a final observation, in Step 2 of the Accept-reject algorithm, the probability of acceptance is equal to $N(\infty) = \frac{1}{M}$. If $\mathcal{X} = \mathbb{R}$, this forces the tails of g to be heavier than those of f, otherwise the ratio f/g would be unbounded, and so would M.

References

[1] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, 1999.

[2] *IEEE Transactions on Signal Processing* special issue on Monte Carlo methods for statistical signal processing, Vol. 50, No. 2, Feb. 2002.