3. The convergence rate of Monte Carlo integration is independent of the dimensionality of the integral, m. The variance of the integration estimate, σ^2/N, does not contain m. In contrast, a typical quadrature method has a convergence rate of $\mathrm{O}(N^{-4/m})$. For this reason, Monte Carlo integration is an especially useful method for high-dimensional integrals.

4. The error $\sigma/\sqrt{N}$ of Monte Carlo integration can be estimated, thereby allowing the user to adjust the number of sample points until a desired error tolerance is matched.

In summary, Monte Carlo integration does best for integrals that are high dimensional and are not smooth.

However, the Monte Carlo method also has some disadvantages:

1. Monte Carlo integration is not the most accurate integration method. Its error is $\sigma/N^{-1/2}$, where σ is the standard deviation of the random variable to be sampled. So halving σ reduces the error as much as quadrupling the number of sample points. The inverse square root dependence on N means that to improve the accuracy by one decimal place (a factor of ten reduction in RMSE) requires 100 times more evaluations of the function $f(x)$. In contrast, for 1D integrals with continuous fourth derivatives, Simpson's rule yields $\mathrm{O}(N^{-4})$.

   To get a sense of the error, we can remove half the sample points and recompute the integrals.

0.7 Variance reduction

To improve the accuracy of Monte Carlo estimates, various methods of variance reduction have been proposed. Variance reduction methods attempt to reduce the variance of statistical estimators such as $\hat{\mu}_N$. Some variance reduction methods attempt to use prior information about the integrals. An example is the method of control variates. Other methods strategically choose the location of sample points. Within the latter category, two classes of variance reduction are importance sampling and stratified sampling (Lemieux 2009).

0.8 Antithetics

Here we follow Section 8.2, Owen (2013).

Monte Carlo integration can lead to inaccurate results if, by random chance, a disproportionate subset of sample points x_i that we draw all lead to a large value of $f(x_i)$. To avoid this problem and to improve our estimate of $\mathbb{E}[f(\mathbf{x})]$, we can try to pair each sample point with an antithetic sample point that, in a loose sense, produces the opposite value of f. I.e., if x_1 produces a large value of $f(x_1)$, then we next deliberately choose a sample point x_2 that produces a small value of $f(x_2)$.

For example, suppose a PDF is symmetric about 0, such as a Gaussian with zero mean. Then we pair each sample value x with a second sample value $-x$. Then our integral becomes

$$\hat{\mu}_{\text{anti}} = \frac{1}{N}\sum_{i=1}^{N/2}[f(X_i) + f(-X_i)]. \tag{142}$$

If f were a linear function of x, then the sample estimate would be exact. But, of course, we only need to resort to Monte Carlo integration for non-linear functions. In that case, the improvement from antithetic sampling is less spectacular, but it still can be useful. The variance of the estimator is given as follows. First, note that, for two random variables X and Y,

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Covar}(X,Y). \tag{143}$$

Generalizing, we find

$$\text{Var}\sum_i X_i = \sum_i \text{Var}(X_i) + 2\sum_{i<j}\text{Covar}(X_i, X_j). \tag{144}$$

This allows us to pull Var inside a sum, when samples are drawn iid and hence the covariance is zero. Then,

$$\begin{aligned}
\text{Var}(\hat{\mu}_{\text{anti}}) &= \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N/2}[f(X_i) + f(-X_i)]\right) \\
&= \frac{1}{N^2}\text{Var}\left(\sum_{i=1}^{N/2}[f(X_i) + f(-X_i)]\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N/2}\text{Var}\left(f(X_i) + f(-X_i)\right) \\
&= \frac{N/2}{N^2}\text{Var}\left(f(X_i) + f(-X_i)\right) \\
&= \frac{N/2}{N^2}\mathbb{E}\left[\left((f(X_i)-\mu) + (f(-X_i)-\mu)\right)^2\right] \\
&= \frac{1}{2N}\left(\text{Var}[f(X_i)] + \text{Var}[f(-X_i)] + 2\text{Covar}[f(X_i), f(-X_i)]\right) \\
&= \frac{\sigma^2}{N}\left(1 + \text{Corr}\right),
\end{aligned} \tag{145}$$

where $-1 \le \text{Corr} \le 1$. So if $f(X_i)$ and $f(-X_i)$ have negative covariance, antithetic sampling is beneficial. But if they have positive covariance, the antithetic estimate is worse! E.g., our samples will be clumped if we use antithetic sampling when f is a parabola centered on zero.

If f is monotonic in all its variates, then $\text{Corr} < 0$, and antithetic sampling will help. But it won't help much if Corr is only slightly negative. On the other hand, the converse is not necessarily true: a non-monotonic function can have a strong negative correlation. Sometimes we can deduce that f is monotone by computing its derivative.

Antithetic sampling is beneficial when f is an odd function. Any function can be written as a sum of even and odd parts:

$$f(x) = \underbrace{\frac{1}{2}\left(f(x) + f(-x)\right)}_{f_{\text{even}}} + \underbrace{\frac{1}{2}\left(f(x) - f(-x)\right)}_{f_{\text{odd}}} \tag{146}$$

An estimate of the mean is:

$$\hat{\mu}_{\text{anti}} = \frac{2}{N}\sum_{i=1}^{N/2} f_{\text{even}}(X_i) \tag{147}$$

The variance can be estimated by noting that

$$\text{Covar}[f(X_i), f(-X_i)] = \pm\text{Var}(f(X_i)) = \pm\text{Var}(f(-X_i)) \tag{148}$$

depending on whether the function f is even (+) or odd (-). Hence, using (145), we obtain:

$$\text{Var}\left(\hat{\mu}_{\text{anti}}\right) = \frac{2}{N}\sigma^2_{\text{even}} \tag{149}$$

Hence if f is purely odd, then antithetic sampling yields the exact integral. If f is purely even, then antithetic sampling is twice as bad as simple Monte Carlo.

0.9 Control variates

This section follows closely that of Section 8.9, Owen (2013).

Suppose that we wish to integrate a complicated function, $f(\mathbf{x})$, and $f(\mathbf{x})$ is related to a similar but simpler function $h(\mathbf{x})$ that we can analytically integrate. This extra knowledge is exploited by the method of control variates in order to reduce the error in sampling $f(\mathbf{x})$.

Our goal is to estimate the average

$$\mu = \mathbb{E}[f(\mathbf{x})]. \tag{150}$$

The basic Monte Carlo estimate of μ is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i). \tag{151}$$

However, suppose that by exact analytic integration, we can also compute the average of a similar function, $h(\mathbf{x})$:

$$\theta = \mathbb{E}[h(\mathbf{x})], \tag{152}$$

where $h(\mathbf{x}) \approx f(\mathbf{x})$. Then, instead of estimating the average of f by $\hat{\mu}$, we use the difference estimator

$$\begin{aligned} \hat{\mu}_{\text{diff}} &= \frac{1}{N}\sum_{i=1}^{N}[f(\mathbf{X}_i) - h(\mathbf{X}_i)] + \theta \\ &= \hat{\mu} - \hat{\theta} + \theta \end{aligned} \tag{153}$$

The estimator $\hat{\mu}_{\text{diff}}$ is unbiased. Its variance is

$$\text{Var}(\hat{\mu}_{\text{diff}}) = \frac{1}{N}\text{Var}[f(\mathbf{X}) - h(\mathbf{X})] \tag{154}$$

Therefore, if h is similar to f, the noise in the sample is reduced. If $h = f$, then the noise is eliminated and the error is zero. (Of course, if we could analytically integrate $h = f$, then integrating f would not require Monte Carlo methods at all.) If h is not perfectly correlated to f, then we can reduce the noise by reducing the magnitude of h by a factor β:

$$\begin{aligned} \hat{\mu}_{\beta} &= \frac{1}{N}\sum_{i=1}^{N}[f(\mathbf{X}_i) - \beta h(\mathbf{X}_i)] + \beta\theta \\ &= \hat{\mu} - \beta\left(\hat{\theta} - \theta\right) \end{aligned} \tag{155}$$

This is called the regression estimator. Again, it is unbiased. Setting $\beta = 1$ recovers the difference estimator, $\hat{\mu}_{\text{diff}}$. Setting $\beta = 0$ recovers the basic Monte Carlo estimator.

The variance of the regression estimator is

$$\text{Var}(\hat{\mu}_{\beta}) = \frac{1}{N}\left\{\text{Var}[f(\mathbf{X})] - 2\beta\text{Cov}[f(\mathbf{X}), h(\mathbf{X})] + \beta^2\text{Var}[h(\mathbf{X})]\right\} \tag{156}$$

We have flexibility in the choice of β. The minimum variance of $\hat{\mu}_{\beta}$ occurs when β is chosen according to:

$$\left.\frac{\partial\text{Var}(\hat{\mu}_{\beta})}{\partial\beta}\right|_{\beta=\beta_{\text{opt}}} = 0. \tag{157}$$

This condition holds when

$$\beta = \beta_{\text{opt}} = \frac{\text{Cov}[f(\mathbf{X}), h(\mathbf{X})]}{\text{Var}[h(\mathbf{X})]}. \tag{158}$$

If we use $\beta = \beta_{\text{opt}}$ in Eq. (156), then the variance of the resulting estimator is

$$\text{Var}(\hat{\mu}_{\beta_{\text{opt}}}) = \frac{\sigma_f^2}{N}\left(1 - \rho^2\right), \tag{159}$$

where

$$\rho = \text{Corr}[f(\mathbf{X}), h(\mathbf{X})] = \frac{\text{Cov}[f(\mathbf{X}), h(\mathbf{X})]}{\sqrt{\text{Var}[f(\mathbf{X})]\text{Var}[h(\mathbf{X})]}}. \tag{160}$$

To see this, substitute (158) into (156). How effective is the use of control variates? If we can find an h that is close enough to f to have a correlation $\rho = 0.9$, then the standard deviation of the sampling noise is reduced by more than a factor of 2. Because of the dependence on ρ^2, slight degradations in ρ lead to large degradations in the quality of the estimator. However, the control variate never increases the variance, even when the correlation ρ is negative!

We can estimate β_{opt} by

$$\hat{\beta}_{\text{opt}} = \frac{\sum_{i=1}^{n}[f(\mathbf{X}_i) - \hat{\mu}_f][h(\mathbf{X}_i) - \hat{\mu}_h]}{\sum_{i=1}^{n}[h(\mathbf{X}_i) - \hat{\mu}_h]^2}, \tag{161}$$

thereby creating the estimator $\hat{\mu}_{\hat{\beta}}$. It turns out that this estimator is biased slightly.

Once $\hat{\beta}_{\text{opt}}$ has been determined, it is possible to find both $\hat{\mu}_{\hat{\beta}}$ and $\text{Var}(\hat{\mu}_{\hat{\beta}})$ by

$$\text{Var}(\hat{\mu}_{\hat{\beta}}) = \frac{1}{N^2}\sum_{i=1}^{N}[f(\mathbf{x}_i) - \hat{\mu}_{\hat{\beta}} - \hat{\beta}(h(\mathbf{x}_i) - \overline{h})]^2. \tag{162}$$

We want to choose $\hat{\beta}$ and $\hat{\mu}_{\hat{\beta}}$ so as to minimize $\text{Var}(\hat{\mu}_{\hat{\beta}})$. Mathematically, this is the same problem as linear regression. We can use standard regression software to solve it. $\hat{\mu}_{\hat{\beta}}$ is the y-intercept.

0.10 Acceptance-Rejection and Variable transformation

Section 7.6, Press et al. (2007)

Making a transformation of variable can more equally distribute the sample points. This leads to better convergence of the integral. Press et al. (2007) gives an example involving the density of a torus. We describe that here.

We want to estimate the mass of an object

$$\int\int\int \rho dxdydz \tag{163}$$

and its moments

$$\int\int\int x\rho dxdydz \tag{164}$$

$$\int\int\int y\rho dxdydz \tag{165}$$

$$\int\int\int z\rho dxdydz \tag{166}$$

The object is in the shape of a part of a torus

$$z^2 + \left(\sqrt{x^2+y^2} - 3\right)^2 \leq 1 \tag{167}$$

inside of a box

$$x \geq 1 \qquad\qquad y \geq -3. \tag{168}$$

Because of the irregular shape, we can sample using acceptance-rejection. That is, first we sample uniformly in the box, and then we throw away points that fall outside the torus.

The torus piece has density

$$\rho(x,y,z) = e^{5z}. \tag{169}$$

So most of the mass is concentrated in the very top of the torus. If we sample uniformly throughout the torus, we'll waste a lot of sample points on low-density regions that don't contribute much to the integral.

To provide a better sample distribution, first transform the z variable:

$$ds = e^{5z}dz \qquad\qquad s = \frac{1}{5}e^{5z} \qquad\qquad z = \frac{1}{5}\ln(5s) \tag{170}$$

With this transformation, we find

$$\int\int\int z\rho dxdydz = \int\int\int ze^{5z}dxdydz = \int\int\int zdxdyds = \int\int\int_{s=\frac{1}{5}e^{-5}}^{s=\frac{1}{5}e^{5}} \frac{1}{5}\ln(5s)\, dsdxdy \tag{171}$$

The integrand in s varies much less than the original integral in z. Hence the Monte Carlo integration will be smoother.

The integral is still difficult because the shape of the boundary is irregular.

0.11 Importance sampling

Section 9.1, Owen (2013); Section 9.6, Ross (2012); Section 7.9.1, Press et al. (2007); Section 7.5.2, Gentle (2003).

Suppose that once again, we wish to integrate

$$\mu_f = \int_D f(x)P(x)\,dx, \tag{172}$$

but that we know that $f(x)P(x)$ is much larger for some values of x than others. A "rough" integrand like this leads to larger errors than smooth integrands. The regions with large-magnitude values of the integrand will contribute more to the integral than other regions and hence are more "important." This knowledge can be exploited in order to reduce the error in the integral. One method that does so is importance sampling.

In importance sampling, we find another PDF, $q(x)$, that is preferentially weighted toward large-magnitude values of the integrand, $f(x)P(x)$, so that more samples are drawn from that region:

$$\begin{aligned}
\mu_f &= \int_D f(x)P(x)\,dx && (173)\\
&= \int_D \frac{f(x)P(x)}{q(x)}q(x)\,dx && (174)\\
&\equiv \int_D h(x)q(x)\,dx. && (175)
\end{aligned}$$

where $h(x) \equiv f(x)P(x)/q(x)$. (Sketch $f(x)$, $P(x)$, and $q(x)$.) Stated differently,

$$\mu_f = \int_D \frac{f(x)P(x)}{q(x)}q(x)\,dx \equiv \mathbb{E}_q\left(\frac{f(X)P(X)}{q(X)}\right) \equiv \mathbb{E}_q[h(X)] \tag{176}$$

Therefore, $\overline{f}$ can be estimated by drawing a set of sample points X_i from the PDF $q(x)$ and then summing $h(X_i)$ over those sample points:

$$\hat{\mu}_{fq} = \frac{1}{N}\sum_{i=1}^{N} h(X_i) = \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)P(X_i)}{q(X_i)}, \qquad X_i \sim q \tag{177}$$

This method is more accurate than direct Monte Carlo integration (Eqns. 132 and 133) when $h(x)$ is smoother than $f(x)$.

Whereas the error of basic Monte Carlo integration (Eqn. 132) is $\mathrm{Var}(\hat{\mu}_f) = \sigma^2/N$, with

$$\begin{aligned}
\sigma^2 &= \int_D (f(x)-\mu)^2 P(x)\,dx\\
&= \left[\int_D f(x)^2 P(x)\,dx\right] - \mu^2,
\end{aligned} \tag{178}$$

the error associated with importance sampling is $\mathrm{Var}_q(\hat{\mu}_{fq}) = \sigma_q^2/N$, where

$$\begin{aligned}
\sigma_q^2 &= \int_D (f(x)P(x)/q(x)-\mu)^2 q(x)\,dx\\
&= \left[\int_D \frac{(f(x)P(x))^2}{q(x)}\,dx\right] - \mu^2.
\end{aligned} \tag{179}$$

This expression indicates the promise and peril of importance sampling. If $q(x)$ is chosen to be proportional to $f(x)P(x)$ insofar as feasible, then $h(x)$ will be more uniform than $f(x)$, and

importance sampling will reduce variance. For instance, if $f(x) > 0$ everywhere and $\mu > 0$, then $q(x) = f(x)P(x)/\mu$ yields an integral with exact accuracy. On the other hand, any discrepancy in the numerator will be magnified by a small value of $q(x)$ in the denominator. This means that a $q(x)$ PDF with light tails can lead to large errors. For instance, large errors can occur even if $|f|P$ is small, if q is even smaller. The essential problem is the age-old problem of dividing by zero. Therefore, if you want q to work with a variety of f, then it is safer to choose a $q(x)$ that has tails at least as heavy as those of $P(x)$. When performing importance sampling, one must be careful to ensure that there are no localized regions where $q(x)$ is small and $f(x)P(x)$ is large. Otherwise, $f(x)P(x)/q(x)$ will be large, and a spike will appear in the integrand that is difficult to sample accurately. At a minimum, we must have $q(x) > 0$ wherever $f(x)P(x) \neq 0$.

The PDF $q(x)$ that leads to the most accurate integration is proportional to

$$q(x) = |f(x)|\, P(x) \tag{180}$$

Of course, in most realistic applications, simple expressions of $q(x)$ cannot approximate $|f(x)|\, P(x)$ with perfect accuracy.

Note also that the most accurate integration requires that the PDF $q(x)$ be tailored to the function $f(x)$. The same $q(x)$ may not be well tailored to a new and different function $g(x)$.

0.11.1 Importance sampling diagnostics

Section 9.3, Owen (2013).

The likelihood ratio, or weighting function, is defined as

$$w(x) = \frac{P(x)}{q(x)} \tag{181}$$

Sample weights are defined as

$$w_i = \frac{P(X_i)}{q(X_i)} \tag{182}$$

for $i = 1, ..., N$, where N is the number of sample points. w_i is a sample value of a random variable, the likelihood ratio P/q, with mean

$$\mathbb{E}_q(w(X)) = \mathbb{E}_q(P(X)/q(X)) = 1. \tag{183}$$

(See (176).) Are a few weights much larger than the others? If so, only a few sample points will contribute significantly to the average, reducing the effective sample size, and possibly making the sample noisy and inaccurate.

A few weights with excessively large values might occur if $q(x)$ has tails that are too light, so that far from the important region, but where $P(x)$ is still non-zero, $P(x) >> q(x)$.

To quantify the effective sample size, consider an independent and identically distributed sample of size n_e drawn from a random variable Z_i with variance σ^2. The unweighted average of this sample

has a variance σ^2/n_e. Now consider a *weighted* average of a sample with size N drawn from Z_i and weighted by w_i:

$$S_w = \frac{\sum_{i=1}^N w_i Z_i}{\sum_{i=1}^N w_i}. \tag{184}$$

What is the variance of this weighted average? It turns out to be:

$$\mathrm{Var}(S_w) = \sigma^2 \frac{\sum_{i=1}^N w_i^2}{(\sum_{i=1}^N w_i)^2}. \tag{185}$$

If we define an effective sample size n_e by

$$\mathrm{Var}(S_w) \equiv \frac{\sigma^2}{n_e}, \tag{186}$$

then we can solve for n_e:

$$n_e = \frac{(\sum_{i=1}^N w_i)^2}{\sum_{i=1}^N w_i^2} = \frac{N\overline{w}^2}{\overline{w^2}}. \tag{187}$$

Here

$$\overline{w} = \frac{1}{N}\sum_{i=1}^N w_i \tag{188}$$

and

$$\overline{w^2} = \frac{1}{N}\sum_{i=1}^N w_i^2 \tag{189}$$

If $n_e << N$, then we might become concerned that our average is coming from just a few sample points because the weights are too uneven. However, if n_e is large, this does not guarantee that importance sampling worked. The sample still might have missed an important region. Furthermore, n_e depends only on the weights and not the function to be integrated, f. A large weight at x might not matter if $f(x) = 0$ at that point. Hence the weights are only indicators, and do not include complete information.

Although n_e is based on the ratio of weights, another diagnostic is $\overline{w}$ itself. Since the expected value of w_i is 1, if $\overline{w}$ is far from 1, then it indicates that q was poorly chosen.

0.11.2 Exponential tilting

Section 9.6, Owen (2013).

How can we choose a functional form for q? Here we discuss the method of exponential tilting.

Consider an example in which the PDF $P(x)$ in our integral is a normal distribution, $\mathcal{N}(\mu, \sigma)$. Then choose $q(x)$ from the same PDF family as $P(x)$, in this case a normal distribution family, but with different parameter values. That is, choose

$$q(x) = \mathcal{N}(\mu', \sigma'). \tag{190}$$

This method of exponential tilting is convenient because sampling of q can use the same method and code as is used for sampling of P. In addition, sometimes the calculation of P/q can be simplified. For instance, if the mean of a normal is perturbed, but not the variance, then the $\exp -x^2/\sigma^2$ term cancels from the numerator and the denominator.

When we choose a value of μ', recall that we wish to choose q such that

$$q(x) \approx f(x)P(x)/\mu. \tag{191}$$

Hence, if $f(x)$ peaks at values of x larger than μ, then we would want to set $\mu' > \mu$.

0.11.3 Defensive importance sampling

Section 9.11, Owen (2013).

We have seen that choosing a q function with tails that are too light can lead to a poor set of weights w_i and a poor sample. How can we ensure that the tails of q are heavy enough?

In defensive importance sampling, we choose an importance distribution, $q_\alpha(x)$, that is a mixture of two PDFs:

$$q_\alpha(x) = \alpha_1 P(x) + \alpha_2 q(x), \tag{192}$$

where α_1 and α_2 are positive weights that sum to 1: $\alpha_1 + \alpha_2 = 1$. The second term is designed to match $f(x)P(x)$ as closely as possible. The first term is proportional to P itself. It is designed to prevent the spikiness associated with light tails in q. The tails cannot be excessively light if $P(x)$ itself is included in q_α. On the other hand, if fP peaks at a very different value than P, the inclusion of P in q_α may degrade the sampling.

The estimator for defensive importance sampling is

$$\hat{\mu}_\alpha = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)P(X_i)}{\alpha_1 P(X_i) + \alpha_2 q(X_i)}. \tag{193}$$

0.11.4 What-if simulations

Section 9.14, Owen (2013).

In importance sampling, we integrate $f(x)P(x;\theta_0)$ by drawing points from a different distribution, q. Here θ_0 is a parameter on which P depends (e.g., its mean). The usual goal is to reduce sampling noise. The estimator is

$$\hat{\mu}_{fq} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)P(X_i;\theta_0)}{q(X_i)}, \quad X_i \sim q \tag{194}$$

This is accurate if $q(x) \approx f(x)P(x;\theta_0)/\mu$.

Sometimes we have a different problem. We wish to solve multiple integrals, each with the same f but with slightly different distributions, P. Suppose that we have already solved an integral over $P(x;\theta_0)$. To solve another integral with the same f but a slightly different PDF, $P(x;\theta)$, where $\theta \approx \theta_0$, we can re-use the same sample points X_i and the same values of $f(X_i)$, but we weight them by $P(X_i;\theta)$ rather than $P(X_i;\theta_0)$:

$$\hat{\mu}_{fq} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)P(X_i;\theta)}{q(X_i)}, \qquad X_i \sim q \tag{195}$$

This is likely to work well if $q(x) \approx f(x)P(x;\theta)/\mu$, which is likely to be true if $\theta \approx \theta_0$.

The method does not require importance sampling. In the latter equation, one could replace $q(x)$ with $P(x;\theta_0)$.

$$\hat{\mu}_{fP0} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)P(X_i;\theta)}{P(X_i;\theta_0)}, \qquad X_i \sim P(X_i;\theta_0) \tag{196}$$

0.11.5 Simultaneous use of control variates and importance sampling

Section 9.10, Owen (2013).

Importance sampling and control variates can be combined. Suppose we have an approximation $h(x)$ to $f(x)$ whose integral is known analytically:

$$\int_D h(x)P(x)dx = \theta. \tag{197}$$

Then we can form the estimate

$$\begin{aligned}\hat{\mu}_{\mathrm{q},\beta} &= \frac{1}{N}\sum_{i=1}^{N} \frac{[f(\mathbf{X}_i) - \beta h(\mathbf{X}_i)]P(\mathbf{X}_i)}{q(\mathbf{X}_i)} + \beta\theta \qquad \mathbf{X}_i \sim q \\ &= \hat{\mu} - \beta\left(\hat{\theta} - \theta\right)\end{aligned} \tag{198}$$

By taking the expectation of this estimator, we can show that the estimator is unbiased.

0.12 Stratified sampling

Section 8.4, Owen (2013); Section 7.9.2, Press et al. (2007); Section 9.6, Ross (2012)

Stratified sampling spreads out, or stratifies, the sample points. This reduces variance by avoiding clumping of sample points. The points are no longer distributed randomly but rather quasi-randomly.

Consider a simple example. Again we wish to estimate

$$\overline{f} = \int_D f(x)P(x)\,dx, \tag{199}$$

but suppose, for simplicity, that $P(x)$ is a uniform distribution with $P(x) = 1$. The variance of the exact integral $\overline{f}$ is

$$\mathrm{Var}(f) = \overline{f^2} - \overline{f}^2 \tag{200}$$

The basic Monte Carlo estimator of $\overline{f}$ is $\overline{f} \approx \hat{\mu}_f = \sum f(X_i)/N$. The variance of the estimator $\hat{\mu}_f$ is

$$\mathrm{Var}(\hat{\mu}_f) = \frac{\mathrm{Var}(f)}{N}, \tag{201}$$

which is similar to Eqn. (178).

Now suppose the domain D is divided into two equal subdomains, a and b. (Draw a $(0,1)$ domain divided into two parts labeled a and b.) Then

$$\overline{f} = \frac{1}{2}\left(\overline{f}^a + \overline{f}^b\right) \tag{202}$$

and

$$\overline{f^2} = \frac{1}{2}\left(\overline{f^2}^a + \overline{f^2}^b\right). \tag{203}$$

Also, we define

$$\mathrm{Var}_a(f) \equiv \overline{f^2}^a - \left(\overline{f}^a\right)^2 \tag{204}$$

and similarly for $\mathrm{Var}_b(f)$.

To estimate $\overline{f}$, rather than drawing sample points randomly from the entire domain D, instead let's draw $N/2$ sample points from each of the two subdomains. The corresponding estimator for $\overline{f}$ is

$$\overline{f} \approx \hat{\mu}_{f,(a+b)} \equiv \frac{1}{2}\left(\hat{\mu}_{f,a} + \hat{\mu}_{f,b}\right). \tag{205}$$

Here,

$$\hat{\mu}_{f,a} = \frac{1}{N/2}\sum_{i=1}^{N/2} f(X_i) \tag{206}$$

with the understanding that all points are drawn from domain a. A similar expression holds for $\hat{\mu}_{f,b}$.

The variance of this new estimator is

$$\begin{aligned}
\mathrm{Var}\left(\hat{\mu}_{f,(a+b)}\right) &= \frac{1}{4}\left[\mathrm{Var}\left(\hat{\mu}_{f,a}\right) + \mathrm{Var}\left(\hat{\mu}_{f,b}\right)\right] \\
&= \frac{1}{4}\left[\frac{\mathrm{Var}_a(f)}{N/2} + \frac{\mathrm{Var}_b(f)}{N/2}\right] \\
&= \frac{1}{2N}\left[\mathrm{Var}_a(f) + \mathrm{Var}_b(f)\right] \qquad (207)
\end{aligned}$$

Using (201), (200), (202), (203), (204) and other relations listed above, we can show that

$$\mathrm{Var}(\hat{\mu}_f) = \frac{\mathrm{Var}(f)}{N} = \frac{\overline{f^2} - \overline{f}^2}{N} = \frac{1}{2N}\left[\mathrm{Var}_a(f) + \mathrm{Var}_b(f)\right] + \frac{1}{4N}\left(\overline{f}^a - \overline{f}^b\right)^2 \qquad (208)$$

By comparing (207) and (208), we see that

$$\mathrm{Var}\left(\hat{\mu}_{f,(a+b)}\right) \leq \mathrm{Var}(\hat{\mu}_f) \qquad (209)$$

Stratifying the domain into two parts reduces the variance of the estimator when the means in each subdomain, $\overline{f}^a$ and $\overline{f}^b$, are different. The reason is that, when the subdomain means are different, putting too many sample points in one subdomain and too few in another will cause an under- or over-estimate. Stratifying sampling prevents this "clumping," thereby reducing errors. An extreme but illustrative example is given by a stair-step function f with half the domain at $f = 0$ and half at $f = 1$. Then $\overline{f} = 1/2$, but we will only obtain this result if half the sample points are in one subdomain and half are in the other.

Now let's write a more general formulation of stratified sampling. Divide the entire domain D into J disjoint categories ("strata") D_j, with $j = 1, ..., J$. The size of the strata, i.e. the probability of selecting from each stratum j, is given by $\omega_j = \mathbb{P}(X \in D_j)$. The conditional PDF of X given that $X \in D_j$ is given by $P_j(x) = (1/\omega_j)P(x)1_{x \in D_j}$. (Show that P_j is normalized to 1.) Suppose that n_j sample points are drawn from each category j. The stratified sampling estimate of μ is

$$\hat{\mu}_{\text{strata}} = \sum_{j=1}^{J} \frac{\omega_j}{n_j} \sum_{i=1}^{n_j} f(X_{ij}) \qquad (210)$$

This estimate is unbiased:

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{\text{strat}}) &= \sum_{j=1}^{J} \omega_j \mathbb{E}\left(\frac{1}{n_j}\sum_{i=1}^{n_j} f(X_{ij})\right) \\ &= \sum_{j=1}^{J} \omega_j \int_{D_j} f(x) P_j(x) dx \\ &= \sum_{j=1}^{J} \int_{D_j} f(x) P(x) dx \\ &= \int_{D} f(x) P(x) dx \\ &= \mu.\end{aligned} \tag{211}$$

The variance of f can be written as follows. The within-stratum mean is $\mu_j = \int_{D_j} f(x) P_j(x) dx$. The within-stratum variance is $\sigma_j^2 = \int_{D_j} (f(x) - \mu_j)^2 P_j(x) dx$. Then one can show

$$\sigma^2 = \sum_{j=1}^{J} \omega_j \sigma_j^2 + \sum_{j=1}^{J} \omega_j (\mu_j - \mu)^2. \tag{212}$$

This formula is a special case of the general relationship

$$\text{Var}[f(X)] = \mathbb{E}\left\{\text{Var}\left[f(X|Z)\right]\right\} + \text{Var}\left\{\mathbb{E}\left[f(X|Z)\right]\right\} \tag{213}$$

where $Z \in \{1, ..., J\}$ is the stratum containing the random point X. This formula suggests that to reduce the noise, we can choose the strata such that f is nearly constant within each stratum, and hence σ_j is small. Indeed, we find that

$$\text{Var}(\hat{\mu}_{\text{strat}}) = \sum_{j=1}^{J} \omega_j \frac{\sigma_j^2}{n_j}. \tag{214}$$

That is, the total variance is the weighted sum of the variance in each category j, σ_j^2/n_j.

0.12.1 Stratified sampling with mixture distributions

A standard way to draw sample points from a PDF that is a mixture of two other PDFs is to first choose a mixture component at random, and then choose a point from within that component. However, this procedure will destroy stratification if the two components are overlapped. Even if the points are stratified within each component, a point in one component may reside close to another point in the other component.

To avoid this problem, one could draw samples using a multivariate version of the inverse CDF method described before. For a three-dimensional multivariate PDF $P(X_1, X_2, X_3)$, with variates (X_1, X_2, X_3), the steps are:

1. Choose a sample point x_1 from the marginal $P(X_1)$.

2. Choose a sample point from the conditional marginal $P(X_2|X_1 = x_1)$. Here, X_3 is integrated out.

3. Choose a sample point from the conditional $P(X_3|X_2 = x_2, X_1 = x_1)$.

0.12.2 Proportional allocation and Neyman allocation

Section 8.4, Owen (2013).

One can generalize this strategy in several ways. For example, the number of subdomains can be increased. A simple but effective strategy is to create multiple subdomains and draw an equal number of sample points from each subdomain. Second, different numbers of points can be placed in the subdomains. It turns out to be optimal to place the sample points proportionally in the subdomains according to the variance of f in each.

In proportional allocation, $n_j = n\omega_j$. The number of samples is proportional to the size of the category.

$$\hat{\mu}_{\text{prop}} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} f(X_{ij}) \tag{215}$$

$$\text{Var}(\hat{\mu}_{\text{prop}}) = \sum_{j=1}^{J} \omega_j^2 \frac{\sigma_j^2}{N\omega_j} = \frac{1}{N} \sum_{j=1}^{J} \omega_j \sigma_j^2. \tag{216}$$

The optimal, or Neyman allocation, is

$$n_j = \frac{N\omega_j\sigma_j}{\sum_{k=1}^{J} \omega_k\sigma_k}. \tag{217}$$

Of course, typically we don't know n_j in advance.

0.12.3 Latin hypercube sampling

Section 10.3, Owen (2013). Section 7.8, Press et al. (2007).

Latin hypercube sampling (LHS) stratifies the points in all d dimensions of the domain. LHS divides the domain into N categories in each dimension d. Consider a 2D ($d = 2$) example. LHS may divide the cloud number concentration into low, mid, and high values, and also divide cloud water mixing ratio into $N = 3$ categories. Then LHS chooses 3 points so that they are ensured to fall into low, mid, and high categories of number concentration and mixing ratio. In this example, the points may be drawn as follows. Divide the 2D domain into a 3×3 grid of boxes. Choose a box, and place a point somewhere at random within it. Cross out the row and column containing the box. From the remaining, uncrossed-out boxes, choose another box. Repeat until all rows and columns have been crossed out. Hence, if we choose one of the d dimensions, and we survey the N points in our sample, exactly one point lies in the low category, one in the mid category, and one in the high category.

Mathematically, to implement LHS, first, we choose sample points from a uniform distribution:

$$X_{ij} = \frac{\Pi_{ij} + U_{ij}}{N}, \tag{218}$$

where $1 \leq i \leq N$ and $1 \leq j \leq d$. Here N is the number of sample points, and d is the dimensions of each sample point, where $d > 1$ for multivariate samples. Each column of Π is a random permutation of the integers $0, 1, ..., (N-1)$. All columns are independent. (Show how Π would be constructed for the $d = 2$, $N = 3$ cloud example mentioned above.)

Each element of U_{ij} is distributed as $U(0, 1)$, and all elements are independent. The U_{ij} is needed to obtain an unbiased distribution. For instance, consider $P(x)$ to be a 1D uniform distribution and consider the function $f(x) = \delta(x - \epsilon)$. The sampling without U_{ij} will never sample $f(x)$. With the U_{ij}/N term, a Latin Hypercube sample is uniformly distributed: $X_{ij} \sim U(0, 1)$.

The $N \times d$ array of sample points looks like:

$$\begin{bmatrix} X_{11} & \cdots & X_{1d} \\ \vdots & & \vdots \\ \vdots & X_{ij} & \vdots \\ \vdots & & \vdots \\ X_{N1} & \cdots & X_{Nd} \end{bmatrix} \tag{219}$$

Then we transform to the PDF of interest, e.g., normal or lognormal.

The estimate given by LHS is

$$\hat{\mu}_{\text{LHS}} = \frac{1}{N} \sum_{i=1}^{N} f(X_i), \tag{220}$$

where each sample point is d-dimensional. This estimate is unbiased:

$$\mathbb{E}(\hat{\mu}_{\text{LHS}}) = \mu, \tag{221}$$

Because LHS stratifies in each of the d dimensions, it reduces noise whenever the function $f(X_i)$ varies a lot in any of the d dimensions. In particular, LHS is helpful when $f(X_i)$ is an additive function:

$$f^{\text{add}}(x_1, ..., x_d) = \mu + \sum_{j=1}^{d} f_j(x_j) \tag{222}$$

LHS is not as helpful when the variates of $f(X_i)$ have strong interaction effects. For simplicity, consider the 2D case. Then

$$f^{\text{add}}(x_1, x_2) = \mu + f_1(x_1) + f_2(x_2) \tag{223}$$

So, for example,

$$f(x_1, x_2) = (x_1^2 - 1/3) + (x_2^2 - 1/3) \tag{224}$$

is additive, but

$$f(x_1, x_2) = x_1 x_2 \tag{225}$$

is not. It seems plausible that the former function will be well sampled by LHS, which stratifies the points in each dimension. However, the latter function will not. Consider a sample with all point along a diagonal, i.e., the line $x_1 = x_2$ or the line $x_1 = -x_2$.

This intuition can be formalized by writing our function to integrate as a sum of additive and non-additive parts:

$$f(\mathbf{x}) = f^{\text{add}}(\mathbf{x}) + e(\mathbf{x}) \tag{226}$$

Then, for a large sample,

$$Var(\hat{\mu}_{\text{LHS}}) = \frac{1}{N} \int e(\mathbf{x})^2 d\mathbf{x} + O(1/N) \tag{227}$$

The error comes from the non-additive part. Owen shows that LHS cannot be much worse than simple MC:

$$Var(\hat{\mu}_{\text{LHS}}) \leq \frac{1}{N-1} \sigma^2. \tag{228}$$

0.13 Mixed importance and stratified sampling

Importance sampling and stratified sampling are different strategies. Importance sampling places more sample points from regions where the integrand contributes a lot, i.e., where $|f(x)P(x)|$ is large. Therefore, importance sampling requires knowledge about the characteristics of $f(x)$. Stratified sampling spreads the points in order to avoid clumping of points and consequent overrepresentation of a particular part of the integrand. In its simpler form, in which equal numbers of points are placed in equal-sized subdomains, it does not require any knowledge of $f(x)$. However, if the sample points are partitioned proportionally among the subdomains according to the variance of $f(x)$, then

knowledge of $f(x)$ is required, and stratified sampling produces a similar distribution of sample points as importance sampling.

Strategies that combine importance and stratified sampling are also possible. For instance, importance sampling can be done on a crude grid, with stratified sampling within each grid cell.

Bibliography

Gentle, J. E., 2003: *Random Number Generation and Monte Carlo Methods*. 2nd edition, Springer, 381 pp.

Germano, M., 1992: Turbulence: The filtering approach. *J. Fluid Mech.*, **238**, 325–336.

Gibson, M. M. and B. E. Launder, 1978: Ground effects on pressure fluctuations in the atmospheric boundary layer. *Journal of Fluid Mechanics*, **86**, 491–511.

Johnson, M. E., 1987: *Multivariate Statistical Simulation*. John Wiley and Sons, 230 pp.

Kalos, M. H. and P. A. Whitlock, 2008: *Monte Carlo Methods*. 2nd edition, Wiley-Blackwell, 203 pp.

Lemieux, C., 2009: *Monte Carlo and quasi-Monte Carlo sampling*. Springer Science & Business Media, 373 pp.

Leonard, A., 1974: Energy cascade in large-eddy simulations of turbulent fluid flows. *Adv. Geophys.*, **18A**, 237–248.

Owen, A. B., 2013: Monte Carlo theory, methods and examples. Https://artowen.su.domains/mc/.

Pope, S. B., 2000: *Turbulent Flows*. Cambridge University Press, 771 pp.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 2007: *Numerical Recipes: The art of scientific computing*. 3rd edition, Cambridge University Press, 1235 pp.

Ross, S. M., 2012: *Simulation*. 5th edition, Academic Press, 328 pp.