http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo17/

Class notes: Monte Carlo methods
Week 2, Simulation and variance reduction
Jonathan Goodman
January 30, 2017

1 Simulation

I use the term *simulation* to mean producing a random object with a given description. This is different from *Monte Carlo*, which I take to mean computing quantities that themselves are not random. For example, simulation might mean making sample paths of a Markov chain. Evaluating the expected value of a function of the sample path is Monte Carlo. The distinction was emphasized to me by Malvin Kalos. In this terminology, the phrase *Monte Carlo simulation* is an oxymoron. This section describes some tricks for simulating discrete time and continuous time Markov chains with a discrete state space, first in discrete time then in continuous time.

A stationary Markov chain with states $\{1, \ldots, n\}$ is described by its *transition matrix* P. At each time $t = 0, 1, 2 \ldots$, the state of the chain is a random variable X_t which is one of the n states. The entry P_{ij} is the $i \to j$ transition probality

$$P_{ij} = P(X_{t+1} = j \mid X_t = i) \; .$$

The definition of a stationary (or time homogenous) Markov chain is that you get path probabilities from these by multiplication. If $X_0 = x_0$ is known and not random, the probability of the sequence $(X_1, \ldots, X_T)$ is

$$P(X_1, X_2, \ldots, X_T) = \prod_{t=0}^{T-1} P_{X_{t-1}, X_t} \; .$$

This is equivalent to the *Markov property*

$$\begin{aligned} &P(X_{t+1} = j \mid X_t = i) \\ &\quad = P(X_{t+1} = j \mid X_t = i \text{ , and } X_{t-1} = i_1 \; , \cdots \; , \; X_1 = i_{t-1} \;) \; . \end{aligned}$$

You can simulate a discrete time finite state space Markov chain using the discrete selection algorithm from Week 1. If you know $X_t = i$, let $N = j$ with probability P_{ij} and take $X_{t+1} = N$. This is not the best simulation method for most Markov chains you meet in practice. Either n is too large for this to be practical, or the chain has special structure that makes the generic sampling method unnecessary. In particular, MCMC (Markov chain Monte Carlo) sampling rarely uses an explicit transition matrix. Instead, the trick of *detailed balance* is used to avoid having to know the numbers P_{ij} exactly.

2 Continuous time processes

For a continuous time Markov chain, $X_t \in \{1, \ldots, n\}$ is defined for real numbers t. There are *transition rates*

$$R_{ij}dt = P(X_{t+dt} = j \mid X_t = i) \ .$$

To be a valid transition rate matrix, the row sums must be zero:

$$\sum_{j=1}^{n} R_{ij} = 0 \ ,$$

and the off diagonal entries must be non-negative:

$$R_{ij} \geq 0 \ , \quad \text{if } i \neq j.$$

One approximate simulation algorithm is to choose a small Δt and simulate a discrete time Markov chain with transition probabilities

$$P_{ij} = \delta_{ij} + \Delta t R_{ij} \ . \tag{1}$$

Here, δ_{ij} is the Kronecker delta. The number δ_{ij} is the (i,j) entry of the identity matrix, which is $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. The exact transition probabilities for time Δt are

$$P(\Delta t) = e^{\Delta t R} = I + \Delta t R + \frac{\Delta t^2}{2} R^2 + \cdots \ .$$

Therefore, the approximation (1) is first order accurate in the sense of numerical analysis. To get an accurate solution, you would like to take a small Δt. But if you do that, you have to to take many steps to advance to a given time T. Moreover, most of those steps are "wasted" in the sense that nothing happens. For small Δt, it is most likely that the state does not change:

$$P(X_{t+\Delta t} \neq X_t) = O(\Delta t) \ .$$

The simulation strategy is not very accurate and spends most of its time doing nothing.

There is an exact and method for simulating a continuous time Markov chain that is usually much faster. It is called the *embedded Markov chain* or *kinetic Monte Carlo*. in chemistry, it is sometimes called *Gillespie's stochastic simulation algorithm*, or *SSA*. It simulates only the transitions and the transition times. A sample path of a continuous time Markov chain may be described by the sequence of occupied states and the occupation times. The occupied states are $Y_k \in \{1, \ldots, n\}$. The transition times are $0 < T_1 < T_2 < \cdots$. We start at time T_0. These satisfy the relation

$$X_t = Y_k \ , \quad \text{if } T_k \leq t < T_{k+1} \ .$$

A non-trivial transition occurs at time T_k if $Y_{k-1} \neq Y_k$.

To describe the algorithm, suppose $X_t = i$ and ask what might happen in time dt. The probability of a transition is

$$P(X_{t+dt} \neq X_t) = \lambda_i dt \; ,$$

where

$$\lambda_i dt = \sum_{j \neq i} P(X_{t+dt} = j \mid X_t = i) = \sum_{j \neq i} R_{ij} dt \; .$$

Since X_t is a Markov process, the transition time is an exponential with rate λ_i. Conditional on having a transition in time dt, the probabilities for the new state are (using Bayes' rule)

$$\begin{aligned} Q_{ij} &= P(X_{t+dt} = j \mid X_{t+dt} \neq i) \\ &= \frac{P(X_{t+dt} = j \text{ and } X_{t+dt} \neq i)}{P(X_{t+dt} \neq i)} \\ &= \frac{R_{ij} dt}{\lambda_i dt} \\ Q_{ij} &= \frac{R_{ij}}{\lambda_i} \end{aligned} \tag{2}$$

To simulate a continuous time Markov chain up to time T, you go from transition to transition T_{k+1} until $T_{k+1} > T$. If $Y_k = i$, you generate the next transition time adding an exponential random variable to T_k:

$$T_{k+1} = T_k - \frac{1}{\lambda_{Y_k}} \log(U_k) \; .$$

Then you choose Y_{k+1} using the transition probabilities (2). Of course, U_k is uniformly distributed in $[0, 1]$.

People often talk about continuous time discrete state processes in terms of a *bell.* They will say "a bell rings and ... ". Here, suppose there is a transition at time T_k. We stay in state Y_k until a bell rings at time T_{k+1}. The waiting time, which is $T_{k+1} - T_k$ is an exponential with rate λ_{Y_k}. When the bell rings, we make a transition to a new state, chosen by probabilities (2).

3 Variance reduction

Suppose $\widehat{A}$ is a Monte Carlo estimate of A. The error is the sum of *bias*, which is $E[\widehat{A}] - A$, and *statistical error*, which is $\widehat{A} - E[\widehat{A}]$. Statistical error is usually larger than bias. Improving accuracy normally means reducing statistical error, which is called *variance reduction.*

There are many variance reduction methods. The nature of the problem determines which methods or methods might help. As a rule of thumb, the simpler the problem, the more difference variance reduction can make. "Simple" means things like low dimension, direct samplers, smooth functions, etc. You might think you will not encounter a problem like this, but “simple” problems are very common. They occur as sub-problems in complicated problems. They also occur by themselves. For example, financial institutions value some options by Monte Carlo. One single option valuation Monte Carlo is simple by Monte Carlo standards/ But they need to re-evaluate options very quickly (much less than a second) and in large numbers (thousands of options).

4 Control variates

Suppose $X \sim f$ is a random variable with known probability density, $V(x)$ is a known function, and we want to estimate

$$A = \mathrm{E}[\, V(X)] \ .$$

A *control variate* is a different function $W(X)$ whose expected value

$$B = E[W(X)]$$

is known. The more closely $W(x)$ resembles $V(x)$, the more variance can be removed. If W resembles V, then $U(X) = V(X) - \alpha W(X)$ may have less variance than $V(X)$. We make a Monte Carlo estimate of

$$A - \alpha B = \mathrm{E}[\, U(X)] \ ,$$

then we add back the known αB to get an estimate of A.

Suppose we have N independent samples $X_k \sim f$. The direct estimate of A is

$$\widehat{A}_d = \frac{1}{N} \sum_{k=1}^{N} V(X_k) \ . \tag{3}$$

The variance of this estimator is

$$\mathrm{var}\Big(\widehat{A}_d\Big) = \frac{\sigma_V^2}{N} \ .$$

The control variate estimator is

$$\widehat{A}_{cv} = \frac{1}{N} \sum_{k=1}^{N} \left(V(X_k) - \alpha W(X_k) \right) + \alpha B \ . \tag{4}$$

The variance of this is

$$\mathrm{var}\Big(\widehat{A}_{cv}\Big) = \frac{\sigma_U^2}{N} \ ,$$

where

$$\begin{aligned} \sigma_U^2 &= \mathrm{var}(U(X)) \\ &= \mathrm{var}(V(X) - \alpha W(X)) \\ &= \mathrm{var}(V(X)) - 2\alpha \,\mathrm{cov}(V(X), W(X)) + \alpha^2 \,\mathrm{var}(W(X)) \\ \sigma_U^2 &= \sigma_V^2 - 2\alpha \, \sigma_{V,W} + \alpha^2 \, \sigma_W^2 \ . \end{aligned} \tag{5}$$

The optimal α minimizes (5), which leads to

$$\alpha = \frac{\sigma_{V,W}}{\sigma_W^2} \; . \tag{6}$$

We substitute back into (5) to get the greatest variance reduction

$$\begin{aligned}
\sigma_U^2 &= \sigma_V^2 - \frac{\sigma_{V,W}^2}{\sigma_W^2} \\
&= \sigma_V^2 \left(1 - \frac{\sigma_{V,W}^2}{\sigma_V^2 \sigma_W^2} \right) \\
\sigma_U^2 &= \sigma_V^2 \left(1 - \rho_{V,W}^2 \right) \; , \qquad (7)
\end{aligned}$$

which involves the *correlation coefficient*

$$\rho_{V,W} \;\; = \operatorname{corr}(V(X), W(X)) = \frac{\operatorname{cov}(V(X), W(X))}{\sqrt{\operatorname{var}(V(X)) \operatorname{var}(W(X)}} \; . \tag{8}$$

These formulas show that the variance reduction depends on the correlation coefficient between the target V and the control variate W. The correlation coefficient is a dimensionless measure of the relationship between V and W. It is between -1 and 1. You drive the variance of the control variate estimator to zero by driving the correlation coefficient to ± 1.

Control variates in specific problems often come from simple approximate solutions.

It is unlikely that you will be able to evaluate the optimal α (6) analytically. It is not common to know $\operatorname{cov}(V, W)$ but not $E[V]$. One approach is to guess a good α to use in (4). Another approach is to estimate the optimal α from Monte Carlo data. This would replace α in (4) by some $\widehat{\alpha}$ that is a function of the numbers $V(X_k)$ and $W(X_k)$. This makes the estimator $\widehat{A}_{cv}$ a nonlinear function of the data, which might make us worry whether the central limit theorem applies.

An informal analysis shows that using $\widehat{\alpha}$ is almost the same as using the exact α, if your goal is to estimate A. Here (in notation statisticians often use) ε_* will represent the estimation error for quantity $*$. For example, if $\widehat{\alpha}$ is the estimate of α, then the estimation error is $\varepsilon_\alpha = \widehat{\alpha} - \alpha$. We also use an overbar to represent sample means, for example in

$$\overline{V} = \frac{1}{N} \sum_{k=1}^{n} V_k = \frac{1}{n} \sum_{k=1}^{N} V(X_k) \; .$$

The relevant estimation errors for this discussion are defined by

$$\begin{aligned}
\overline{V} &= A + \varepsilon_V \\
\overline{W} &= B + \varepsilon_W \\
\widehat{\sigma^2_W} &= \overline{\left(W - \overline{W}\right)^2} = \sigma^2_W + \varepsilon_{WW} \\
\widehat{\sigma_{V,W}} &= \overline{\left(V - \overline{V}\right)\left(W - \overline{W}\right)} = \sigma_{V,W} + \varepsilon_{VW} \\
\widehat{\alpha} &= \frac{\widehat{\sigma_{V,W}}}{\widehat{\sigma^2_W}} = \frac{\sigma_{V,W} + \varepsilon_{VW}}{\sigma^2_W + \varepsilon_{WW}} = \alpha + \varepsilon_\alpha
\end{aligned} \tag{9}$$

With $\widehat{\alpha}$ in (4), the overall estimation error is found by calculating

$$\begin{aligned}
\widehat{A} &= \overline{V} - \widehat{\alpha}\overline{W} + \widehat{\alpha}B \\
&= A + \varepsilon_V - (\alpha + \varepsilon_\alpha)(B + \varepsilon_W) + (\alpha + \varepsilon_\alpha)B \\
&= A + \varepsilon_V - \alpha\varepsilon_W - \varepsilon_\alpha\varepsilon_W \ .
\end{aligned}$$

If we had used the exact but unknown α, the result would have been

$$\begin{aligned}
\widehat{A} &= \overline{V} - \alpha\overline{W} + \alpha B \\
&= A + \varepsilon_V - \alpha(B + \varepsilon_W) + \alpha B \\
&= A + \varepsilon_V - \alpha\varepsilon_W \ .
\end{aligned}$$

The statistical errors differ by $\varepsilon_\alpha\varepsilon_W$. If you have a lot of data, the difference between the statistical errors should be much smaller than the errors themselves (next paragraph). This shows that estimating α, at least when there is a lot of data, is nearly as good as using the exact α.

The central limit theorem suggests that for large N, the statistical errors are on the order of $N^{-1/2}$. This is clear, for example, for ε_V, given that

$$\mathrm{E}\left[\varepsilon^2_V\right] = \frac{1}{N}\sigma^2_V \ .$$

For ε_{WW}, we calculate

$$\begin{aligned}
\overline{\left(W - \overline{W}\right)^2} &= \overline{W^2} - \left(\overline{W}\right)^2 \\
&= \mathrm{E}\left[\varepsilon^2_W\right] + \varepsilon_{W^2} - \left(B + \varepsilon_W\right)^2 \\
&= \mathrm{E}\left[\varepsilon^2_W\right] - B^2 + \varepsilon_{W^2} - 2B\varepsilon_W + \left(\varepsilon_W\right)^2 \\
&= \sigma^2_W + \varepsilon_{W^2} - 2B\varepsilon_W + \left(\varepsilon_W\right)^2 \ .
\end{aligned}$$

From this, we see that

$$\varepsilon_{WW} = \varepsilon_{W^2} - 2B\varepsilon_W + \left(\varepsilon_W\right)^2 \ .$$

The central limit theorem says the first two terms on the right are order $N^{-1/2}$. The last term is a product of two, so it is order N^{-1}, which is smaller. The point of this ε method is to derive results like this. We see when you put statistical approximations into nonlinear formulas. In this case the nonlinearity was just a square.

The nonlinearity is more complicated for ε_α. It is estimated from (9), assuming $\varepsilon \ll \sigma_{WW}$, using Taylor series:

$$\begin{aligned}\widehat{\alpha} &= (\sigma_{VW} + \varepsilon_{VW})\left(\frac{1}{\sigma_W^2} - \frac{\varepsilon_{WW}}{\sigma_W^4}\right) + O(\varepsilon_{WW}^2) \\ &= \frac{\sigma_{VW}}{\sigma_W^2} + \frac{1}{\sigma_W^2}\varepsilon_{VW} - \frac{\sigma_{VW}}{\sigma_W^4}\varepsilon_{WW} + O(\varepsilon_{WW}^2)\ .\end{aligned}$$

The first term on the right is α. The rest is

$$\varepsilon_\alpha = \frac{1}{\sigma_W^2}\varepsilon_{VW} - \frac{\sigma_{VW}}{\sigma_W^4}\varepsilon_{WW} + O(\varepsilon_{WW}^2)\ .$$

This shows, at least for large N, that ε_α also is of order $N^{-1/2}$.

It is possible, clearly, to use more than one control variate. If the control variates are $W_1, \ldots, W_m$, then the optimal control variate estimator is

$$\widehat{A}_{cv} = \overline{V} - \sum_{j=1}^{m} \alpha_j \overline{W}_j + \sum_{j=1}^{m} \alpha_j B_j\ .$$

The optimal coefficients α_j are found by a multi-variate version of (6), which statisticians will recognize as linear regression coefficients. The necessary covariances may be estimated from data, though the more control variates there are, the larger N is needed to estimate all of them accurately enough.

5 Rao Blackwellization

This is named after statisticians C. R. Rao and David Blackwell. It is the principle that partial averaging reduces variance. Suppose (X, Y) is a pair of random variables with some joint distribution $f(x, y)$. Suppose X by itself has marginal density $g(x)$. Suppose the average of $V(X, Y)$ over Y is $W(x)$, then the variance of W is less than the variance of V. In formulas, suppose x has n components and Y has m components. the marginal density of X is and that

$$g(x) = \int f(x, y)\, dy\ .$$

The conditional density of Y given x is

$$f(y|x) = \frac{f(x, y)}{g(x)}\ .$$

The conditional expectation of $V(X, Y)$, conditioned on $X = x$ is

$$W(x) = \int V(x, y)\, \frac{f(x, y)}{g(x)}\, dy\ . \qquad (10)$$

The *Rao Blackwellization* principle is, except in trivial cases where they are equal,

$$\mathrm{var}_g(W(X)) < \mathrm{var}_f(V(X,Y)) \ .$$

If you can replace a random variable by a partial average, the variance goes down. This is a consequence of the orthogonality relation

$$\mathrm{E}\Big[(V(X,Y)-A)^2\Big] = \mathrm{E}\Big[(W(X)-A)^2\Big] + \mathrm{E}\Big[(V(X,Y)-W(X))^2\Big] \ . \qquad (11)$$

Here $A = \mathrm{E}[V(X,Y)] = \mathrm{E}[W(X)]$. A statistician might remember this formula as the total sum of squares being equal to the explained plus the unexplained sums of squares.

A general abstract version of the same thing involves the abstract definition of conditional expectation. Suppose ω (the abstract version of (X,Y)) has probability measure P and sigma algebra $\mathcal{F}$. Suppose $\mathcal{G} \subset \mathcal{F}$ is a sub-algebra. If $V(\omega)$ is a function of the random variable and

$$W = \mathrm{E}[V \mid \mathcal{G}] \ ,$$

then $\mathrm{var}(W) < \mathrm{var}(V)$, except in trivial cases of equality. The analogue of (11) is

$$\mathrm{E}\Big[(V-A)^2\Big] = \mathrm{E}\Big[(W-A)^2\Big] + \mathrm{E}\Big[(V-W)^2\Big] \ .$$

The abstract Rao Blackwellization principle is that whenever you can substitute an the expected value for a random sample, you reduce the variance. This can be a very large variance reduction, or a reduction so small that it is not worth the trouble.

Some variance reduction tricks can be understood as instances or the Rao Blackwellization principle.

5.1 Antithetic variates

Many probability distributions, particularly centered Gaussians, are symmetric with respect to $x \leftrightarrow -x$. If $f(-x) = f(x)$ and X_k $(k = 1, \ldots, N)$ is a sample of f, then a consistent estimator of $E_f[V(X)]$ is

$$\widehat{A}_{av} = \frac{1}{2N} \sum_{k=1}^{N} \left[V(X_k) + V(-X_k) \right] \ . \qquad (12)$$

The variance of $\widehat{A}_{av}$ is less than the variance of the direct estimator (3).

The antithetic variates estimator may be seen as an instance of the Rao Blackwell principle. It is possible to choose a sample $Y \sim f$ as follows. First choose $Y \sim f$, then choose $s = \pm 1$ with equal probabilities. Finally, choose $Y = sX$. The expectation over Y is the same as the expectation over X, by the symmetry of f. The expectation over Y is the same as the average over the pair (X, s) The antithetic variates estimator (12) is the same as averaging over s. This justifies the claim that antithetic variates reduces variance. Of course, the estimator (12) uses twice the number of evaluations of V. If it is more expensive to evaluate V than to sample f, then we might prefer to use $2N$ independent samples of f.

Antithetic variates give large variance reductions in case $V(x)$ is a smooth function with $\nabla V(x_0) \neq 0$ and $f(x)$ centered about x_0 and concentrated close to x_0. In that case, $V(x) \approx V(x_0) + \nabla V(x_0)(x - x_0)$. If f is symmetric about x_0, then the linear term $\nabla V(x_0)(x - x_0)$ makes zero contribution to the expectation. The antithetic variates estimator (12) also gives zero contribution from the linear term.

5.2 Stratified sampling

Systematic sampling means choosing some aspect of the samples in a deterministic systematic way, while possibly allowing other aspects of the samples to be random. *Stratified sampling* is a simple instance of this idea. Suppose the sample space, Ω, is broken into L disjoint pieces of equal probability:

$$\Omega = \cup_{j=1}^{L} \Omega_j \ ,$$

with

$$\mathrm{P}(\Omega_j) = \int_{\Omega_j} f(x)\, dx = \frac{1}{L} \ .$$

The subsets Ω_j are *strata*. Suppose the overall number of samples, N, is chosen so that about $M = N/L$ samples should be in each stratum. Stratified sampling means choosing exactly M samples in each stratum.

The probability density of stratum Ω_j is $f_j(x) = Lf(x)$ if $x \in \Omega_j$ and $f_j(x) = 0$ otherwise. Stratified sampling chooses a subsample of size M from each stratum:

$$X_{j,k} \sim f_j(x) \ , \quad \text{for } k = 1, \ldots, M \ .$$

The stratified Monte Carlo estimate is

$$\begin{aligned} \widehat{A}_{ss} &= \frac{1}{L} \sum_{j=1}^{L} \left(\frac{1}{M} \sum_{k=1}^{M} V(X_{j,k}) \right) \\ &= \frac{1}{N} \sum_{j=1}^{L} \sum_{k=1}^{M} V(X_{j,k}) \ . \end{aligned}$$

On the top line, the quantity in parentheses is the estimate of the average of $V(x)$ in the stratum Ω_j, and the outer sum gives the average over strata.

The stratified sampling estimator has less variance than the direct estimator because it has less randomness. In the direct estimator, the number of samples landing in a particular stratum is random. The stratified sampler fixes these numbers all to be equal to M. We can see this directly by computing the variance

$$\mathrm{var}\left[\frac{1}{M} \sum_{k=1}^{M} V(X_{j,k}) \right] = \frac{1}{M} \mathrm{var}_{f_j}\left[V(X) \right] \ .$$

The variance of $V(X)$ over the stratum Ω_j is the expectation of $\left[V(X) - \overline{V}_j\right]^2$, where $\overline{V}_j$ is the average of $V(X)$ over Ω_j. The variance of the direct estimator depends on the overall variance of $V(X)$, which is higher.

It is not necessary that the strata all have the same probability. If

$$p_j = \int_{\Omega_j} f(x)\,dx \ ,$$

then $M_j = p_j N$ is the number of samples to take in Ω_j. This helps make the connection to partial quadrature (below).

5.3 Partial quadrature

Partial quadrature means doing some of the Monte Carlo integral by deterministic quadrature and the rest by Monte Carlo. Suppose the random variable may be written (X, Y) with PDF $f(x, y)$. Suppose that the marginal density for X is $g(x)$ and is known. Consider doing the integral over x and y using a deterministic quadrature over x and Monte Carlo in y. A weighted quadrature rule over x is a collection of points x_k and weights w_k so that

$$\int u(x)g(x)\,dx \approx \sum_k w_k u(x_k) \ .$$

Now suppose it is possible to sample the conditional density $Y_k \sim f(y|x_k)$. Then it is possible to form the partial quadrature estimator

$$\widehat{A}_{pq} = \sum_k w_k V(x_k, Y_k) \ . \tag{13}$$

This estimator is slightly biased. Suppose $u(x)$ is the conditional expectation

$$u(x) = \int V(x, y) f(y|x)\,dy \ .$$

The partial quadrature estimator (13) does not require you to know u. The expected value is

$$\mathrm{E}\left[\widehat{A}_{pq}\right] = \sum_k w_k \mathrm{E}_{x_k}[V(x_k, Y] = \sum_k w_k u(x_k) \ .$$

Therefore the bias is the error in the quadrature formula

$$\int u(x)g(x)\,dx - \sum_k w_k u(x_k) \ .$$

The variance of the partial quadrature estimator depends on the fixed-x variance

$$\sigma^2(x) = \mathrm{E}\left[(V(x, Y) - u(x))^2\right] = \int (V(x, y) - u(x))^2 f(y|x)\,dy \ .$$

The terms on the right of (13) are independent, so

$$\mathrm{var}\left(\widehat{A}_{pq}\right) = \sum_k w_k^2 \sigma^2(x_k) \; .$$

Stratified sampling and partial quadrature are related ideas. For example, suppose x is one dimensional. Suppose the x axis is divided into n pieces of size Δx each. A deterministic integration rule might take x_k to be the center of the $k - th$ piece. Stratified sampling would take X_k to be random in this interval. Taking the x_k to be deterministic is generally more accurate, because the bias that comes from the quadrature rule is probably much less than the statistical error that comes from a random choice.

5.4 Latin hypercube sampling

Suppose there are independent variables $X \sim f$ and $Y \sim g$ and we want to know $A = \mathrm{E}[V(X,Y)]$. Suppose that each variable can be stratified. Stratifying X means that there are disjoint sets in the $X-$space, B_j, so that $\mathrm{P}(X \in B_j) = \frac{1}{L}$. Suppose the sets C_j stratify Y in the same sense. A stratified sample of X would be a collection of samples $X_j \in B_j$, for $j = 1, \ldots, L$. These could be in a different order, which could be written j_k, for $k = 1, \ldots, L$. so that if $j_{k_1} = j_{k_2}$ then $k_1 = k_2$. This means that the indices $j_1, \ldots, j_L$ are a permutation of $1, \ldots, L$.

A *latin hypercube* sample $((X_1,Y_1), \ldots, (X_L,Y_L))$ comes with an associated permutation j_k. It has $Y_k \in C_k$ and $X_k \in B_{j_k}$, for $k = 1, \ldots, L$. Consider an $L \times L$ square of boxes, with the rows corresponding to the $Y-$strata and the columns corresponding $X-$strata. A latin hypercube sample has exactly one sample in each row and one sample in each column. Latin hypercube sampling means choosing a permutation "at random" (each permutation being equally likely), then choosing the Y_k and X_k from the appropriate strata. It is easy to choose a random permutation (choose the first index "at random", then choose the next index at random from the remaining ones, and so on).

Another stratified sampling strategy would be to create L^2 strata in (X,Y) space of the form $B_j \times C_k$. One stratified sample would be a sequence of L^2 (X,Y) pairs. A latin hypercube sample is a sequence of L pairs, which is a lot shorter, say, if $L = 100$.

It may not be immediately obvious that latin hypercube sampling is "correct" in the sense of being unbiased:

$$\mathrm{E}\left[\frac{1}{L} \sum_{k=1}^{L} V(X_k, Y_k) \right] = A \; .$$

This is true because each box $B_j \times C_k$ is equally likely to be chosen, which means that if you choose k at random, then

$$\mathrm{P}((X_k, Y_k) \in B_j \times C_k) = \frac{1}{L^2} \; .$$

Therefore, if you choose a term at random from the sum, which means choosing k at random, then

$$\mathrm{E}[V(X_k, Y_k)] = A \ .$$

Latin hypercube sampling, like much of Monte Carlo, is a clever idea.

The method just described, with strata B_j and C_k could be called "latin squares sampling". If there were three independent variables (X, Y, Z) that can be stratified, a *latin cube* sample would be a sequence (X_k, Y_k, Z_k) with one "hit" in each $X-$stratum, one in each $Y-$stratum, and one in each $Z-$stratum. For more than three independent stratified the method is *hypercube* sampling.

5.5 Low discrepancy sequences, quasi Monte Carlo

This is an approach to integration in d dimensions. The curse-of-dimensionality calculation shows that product form quadrature rules are impractical for large d, or even moderate d. If there are m points in each direction, there are $N = m^d$ points in all. This grows exponentially as a function of d. *Low discrepancy sequences* are quadrature strategies for d dimensional integration that are better than product quadrature.

A low discrepancy sequence is $X_k \in C_d$, where C_d is the unit cube in d dimensions. The sequence is uniformly distributed if, for any continuous function $V(x)$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} V(X_k) = \int_{C_d} V(x)\, dx \ . \tag{14}$$

A random i.i.d. sequence (uniformly distributed) does this, with an error (the difference between the sum and the limit) that goes to zero like $N^{-1/2}$. The deterministic sequence is a *low discrepancy sequence* if the error goes to zero faster than this.

It takes some time to describe good low discrepancy sequences. One popular one is the *Sobol sequence*. This one has errors that decay something like

$$\frac{C \log(N)^d}{N} \ .$$

For any fixed d, this is eventually better than $N^{-1/2}$, but for large d you need very large N for this. For this reason, low discrepancy sequences are most helpful when d is not very large. They are more useful as the deterministic part of a partial quadrature strategy. In that case, it is important to find variables that contain as much of the overall variance as possible. In problems involving Brownian motion, the Brownian bridge construction (below) is helpful for this purpose.

6 Brownian motion and the Brownian bridge construction

A Brownian motion is a random function function of t, written X_t, defined for t in the range $0 \le t \le T$. By convention we take $X_0 = 0$ unless we say otherwise. The *increment* of Brownian motion over the interval $[t_1, t_2]$ is just the difference $Y_{t_1,t_2} = X_{t_2} - X_{t_1}$. For a Brownian motion, the increment is a mean zero Gaussian with variance equal to the length of the interval:

$$\begin{aligned} \mathrm{E}[X_{t_2} - X_{t_1}] &= 0 \\ \mathrm{E}\Big[\left(X_{t_2} - X_{t_1}\right)^2\Big] &= t_2 - t_1 \ . \end{aligned}$$

Moreover, increments from disjoint intervals are independent. If $t_1 < t_2 \le t_3 < t_4 \le t_5 \cdots$, then the increments Y_{t_1,t_2}, Y_{t_3,t_4}, etc., are independent. The Brownian motion path also is a continuous function of t.

Consider times $t_0 = 0$, and $t_{k+1} > t_k$. We abuse notation by writing $X_k = X_{t_k}$ for the positions of the Brownian motion path at times t_k. Let $X = (X_1, X_2, \ldots, X_n)$. We write the PDF $f_n(x_1, \ldots, x_n)$, using the independent increments property. First, the first increment is $Y_{t_0,t_1} = X_1 - X_0 = X_1$ is normal with mean zero and variance t_1. Therefore its PDF is

$$f_1(x_1) = \frac{1}{\sqrt{2\pi t_1}} e^{-x_1^2/2t_1} \ .$$

Since $t_0 = 0$ and $X_0 = X_{t_0} = 0$, it is harmless to write this in the seemingly more complicated form

$$f_1(x_1) = \frac{1}{\sqrt{2\pi(t_1 - t_0)}} e^{-x_1 - x_0)^2/2(t_1 - t_0)} \ .$$

Independent of X_1, the next increment is Gaussian with mean zero and variance $t_2 - t_1$. This means that if X_1 is known then $X_2 = X_1 + Y_{t_1,t_2}$ is normal with mean X_1 and variance $t_2 - t_1$. Therefore, the joint density of (X_1, X_2) is the product of the marginal density of X_1 and the conditional density of X_2 conditioned on X_1. That is

$$\begin{aligned} f_2(x_1, x_2) &= \frac{1}{\sqrt{2\pi(t_1 - t_0)}} e^{-(x_1 - x_0)^2/2(t_1 - t_0)} \frac{1}{\sqrt{2\pi(t_2 - t_1)}} e^{-(x_1 - x_0)^2/2(t_2 - t_1)} \\ &= \frac{1}{Z(t_1, t_2)} e^{\frac{-1}{2}\left[\frac{(x_1 - x_0)^2}{(t_1 - t_0)} + \frac{(x_2 - x_1)^2}{(t_2 - t_1)}\right]} \ . \end{aligned}$$

The second line shows how convenient it can be not to write the normalization constant explicitly, because it is complicated and often not important. Continuing in this way leads to

$$f_n(x_1, \ldots, x_n) = \frac{1}{Z(t_1, \ldots, t_n)} \exp\left[\frac{-1}{2} \sum_{k=1}^{n} \frac{(x_k - x_{k-1})^2}{t_k - t_{k-1}}\right] \ . \qquad (15)$$

The normalization constant is

$$Z(t_1, \ldots, t_n) = (2\pi)^{n/2} \left(\prod_{k=1}^{n} (t_k - t_{k-1}) \right)^{1/2} .$$

These formulas show that the Brownian motion values at any sequence of times are distributed as a multi-variate Gaussian. I am aware of the conflict of notation, that Z represents a standard normal and the normalization constant in (15). I hope it doesn't cause too much confusion.

A Brownian motion path is a function $X(t)$ for t in some range. The path cannot exist in the computer because it takes an infinite number of numbers to describe it. A computer approximation is the values of X_t at a sequence of times $t_k = k\Delta t$. Generating a path means finding a set of numbers $X_k X_{t_k}$ (another conflict of notation) whose PDF is (15), with $t_k - t_{k-1} = \Delta t$. The independent increments property leads to a simple direct sampler for this. First generate $X_1 \sim \mathcal{N}(0, \Delta t)$ as $\sqrt{\Delta t}\, Z_1$, then generate $X_2 \sim X_1 + \mathcal{N}(0, \Delta t)$as $X_2 = X_1 + \sqrt{\Delta t}\, Z_2$, and so on. A sequence of i.i.d. standard normals $Z_1, \ldots, Z_N$ turns into the values of a Brownian motion path $X_k = X_{t_k}$ using $X_0 = 0$ and

$$X_k = X_{k-1} + \sqrt{\Delta t}\, Z_k .$$

This is not the only way to do it, but it is the simplest.

6.1 The Brownian bridge construction

Many variance reduction methods are ways to exploit your understanding of what the most important components of X are. It may be that $V(x)$ is more sensitive to some components of V than to other components. If so, we can look for ways to relate X to a lower dimensional model with just the most important components, or to use the most important component or components in partial quadrature. Mathematically, this may involve a change of coordinates.

The *Brownian bridge* construction is a coordinate change that identifies the most important components of Brownian motion in many cases. The Brownian bridge parametrization of Brownian motion allows for remarkable variance reduction in many applications. You can find examples from finance in the work of Russ Caflisch and Bill Morokoff. This section "recalls" enough of the basics about Brownian motion to describe it.

The Brownian bridge construction consists of first choosing the end point, then the midpoint, then the two quarter points, then the 4 eighth points, and so on. That is we first choose X_T, then $X_{T/2}$, conditional on the value of X_T, then the two values $X_{T/4}$ and $X_{3T/4}$, etc. To start, X_T is Gaussian mean zero and variance T, which we can do as (all Z random variables are independent standard normals):

$$X_T = \sqrt{T} Z_1 .$$

We then need the conditional distribution of $X_{T/2}$, given X_T and $X_0 = 0$. This conditional density, as a function of the unknown $x_{T/2}$, is the same (up to a normalization constant) as the joint density. Look to (15) for the joint density, and you get the conditional density as

$$f_2(x_{T/2}|x_T) = \frac{1}{Z}\exp\left[\frac{-1}{2}\left(\frac{x_{T/2}^2}{T/2} + \frac{(X_T - x_{T/2})^2}{T/2}\right)\right] \; .$$

The random variable is $x_{T/2}$ and the exponent depends quadratically on this quantity. A quadratic is characterized by its minimum (in this case) and the coefficient of $x_{T/2}^2$. We find the minimizer by differentiating with respect to $x_{T/2}$ and setting the derivative to zero.

$$\begin{aligned} 0 &= \partial_{x_{T/2}}\left(\frac{x_{T/2}^2}{T/2} + \frac{(X_T - x_{T/2})^2}{T/2}\right) \\ &= \frac{2}{T/2}\left(x_{T/2} - (X_t - x_{T/2})\right) \\ &= \frac{2}{T/2}\left(2x_{T/2} - X_T\right) \; . \end{aligned}$$

The most likely value of $X_{T/2}$, which is the same as mean for a Gaussian, is $\frac{1}{2}X_T$. The coefficient of $x_{T/2}^2$ is what you see when you let $x_{T/2}$ go to infinity, which is $\frac{2}{T/2}$. Therefore, the conditional variance of $X_{T/2}$ is the reciprocal of this, which is $\frac{T}{4}$. This shows that the conditional distribution of $X_{T/2}$ is Gaussian with mean $\frac{1}{2}X_T$ and variance $T/4$. If we did not know X_T, the unconditional variance of $X_{T/2}$ would be $T/2$. Knowing X_T adds information and removes uncertainty – hence the lower variance. This conditional $X_{T/2}$ distribution may be sampled using

$$X_{T/2} = \frac{1}{2}X_T + \sqrt{\frac{T}{4}}\, Z_2 \; .$$

Note that the conditional mean, which is $\frac{1}{2}X_T$, is on the line connecting $X_0 = 0$ to X_T. Continuing, we sample the conditional distribution of the quarter points using

$$\begin{aligned} X_{T/4} &= \frac{1}{2}X_{T/2} + \sqrt{\frac{T}{8}}\, Z_3 \\ X_{3T/4} &= \frac{1}{2}\left(X_{T/2} + X_T\right) + \sqrt{\frac{T}{8}}\, Z_4 \; . \end{aligned}$$

At the next level, normals Z_5, Z_6, Z_7, and Z_8 would be used to create values for $X_{T/8}$, $X_{3T/8}$, $X_{5T/8}$, and $X_{7T/8}$. For example, $X_{5T/8} = \frac{1}{2}(X_{T/2} + X_{3T/4}) + \sqrt{\frac{T}{16}}\, Z_7$.

The Brownian bridge construction is useful in Monte Carlo for systematic sampling. In many applications of Brownian motion, the first variables Z_1, Z_2, etc., have more impact on the answer than the later ones. Suppose there are L levels and $N = 2^L$ times, $t_i = i\Delta t$, with $\Delta t = T/N$. Then the Brownian motion path at those times is determined by $Z_1, \ldots, Z_N$. Let $V(X)$ be some function of Brownian motion. Then this V is ultimately a function of the Brownian bridge variables, which we write as $V(Z_1, Z_2, \ldots)$. In many applications it happens that the variance of $V(z_1, Z_2, \ldots)$ (which is V with the first Z fixed) is significantly smaller than the overall V. Systematic sampling over Z_1 may reduce the overall variance in such cases.

7 Importance sampling and rare event simulation

Suppose R is some event and we want

$$A = \mathrm{P}(X \in R) = \int_{x \in R} f(x)\, dx \; . \tag{16}$$

Rare event simulation is the problem of estimating A when A is very small. Direct simulation is a poor approach in the sense that you need many samples to estimate A accurately. More precisely, the *relative error* in $\widehat{A}$, an estimate of A, is

$$\frac{\widehat{A} - A}{A} \; .$$

If $A = 10^{-5}$ and $\widehat{A} = 3 \cdot 10^{-5}$, the error is "only" $\widehat{A} - A = 2 \cdot 10^{-5}$, but the relative error is 2. The estimate is off by a factor of 3. Direct estimation suffers from this if A is small.

Suppose $X_1, \ldots, X_n$ are independent samples of f, and $Y_i = 1$ if $X_i \in R$ and $Y_i = 0$ otherwise. The number of *hits* is

$$N = \sum_{k=1}^{n} Y_i \; .$$

Each Y is a Bernoulli random variable with mean A and variance $A(1 - A)$. The standard deviation of N is

$$\sigma_N = \frac{1}{\sqrt{n}} \sqrt{A(1-A)} \approx \sqrt{\frac{A}{n}} \; .$$

The second approximation is valid when A is small, so 1 approximates $1 - A$ to a high relative accuracy. Since $\widehat{a} - A$ is on the order of σ_N, the relative error is of the order

$$\frac{\sqrt{\frac{A}{n}}}{A} = \frac{1}{\sqrt{nA}} \; .$$

This shows that the relative accuracy depends on the expected number of hits (which is nA), not the number of samples (which is n). It takes a reasonable number of hits to get even a rough idea of A. If A is very small, most of the samples will not be hits, and will therefore be wasted.

There are other Monte Carlo computations that depend on rare events. The quantity $A = \mathrm{E}[V(X)]$ may be determined by rare events in the sense that (just think of ϵ and δ as small numbers)

$$\mathrm{P}(\, V(X) > \varepsilon A) \leq \delta \; . \tag{17}$$

For example, consider the moments of a standard normal

$$M_{2n} = \mathrm{E}_{\mathcal{N}(0,1)}\big[X^{2n}\big] = (2n-1)(2n-3)\cdot \cdots \cdot 3 \cdot 1 \; .$$

For $2n = 12$ (the twelth moment), we have

$$A = M_{12} = 11 \cdot 9 \cdot 7 \cdot 5 \cdot 3 = 135135 \; .$$

On the other hand,

$$\mathrm{P}\big(X^{2n} \geq 2^{2n}\big) = \mathrm{P}(|X| \geq 2) \approx .052 = \delta \; .$$

Take the ε in (17) corresponding to $|X| = 2$ and you get $\varepsilon = 2^{2n}/M_{12} = .03$. The probability that $V(X)$ is more than 3% of the answer is about 5%. This can be put in a form more relevant for Monte Carlo error bars. The difficulty of estimating A is proportional to the dimensionless measure

$$D = \frac{\sigma_{V(X)}}{A} \; .$$

The relative error after n independent samples is of the order of $D/\sqrt{n}$. In this example, with $n = 6$,

$$\begin{aligned} D &= \frac{\sqrt{\mathrm{E}[X^{12}] - \mathrm{E}[X^6]^2}}{\mathrm{E}[X^6]} \\ &= \frac{\sqrt{11 \cdot 9 \cdot 7 \cdot 5 \cdot 3 - (5 \cdot 3)^2}}{5 \cdot 3} \\ &= \sqrt{11 \cdot 9 \cdot 7/(5 \cdot 3) - 1} \\ &= 6.7 \; . \end{aligned}$$

For $n = 12$ the corresponding number is $D = 54$ (check this!). This suggests that if you use $n = 100^2 = 10,000$ samples, the relative accuracy is on the order of

$$\frac{54}{\sqrt{10,000}} = \frac{54}{100} = .54\% \; .$$

Rare event sampling is a way to get better than 54% error after ten thousand samples.

There are several approaches to rare event sampling and variance reduction. Many of these call for MCMC methods, so we talk about them later. *Importance sampling* is an approach based on understanding the mechanism of the rare event. The understanding is used to create a different probability density,

$g(x)$, that puts more weight on the parts of probability space relevant for the rare event. We will see that if this is not done carefully, the variance can be increased. A supposed variance *reduction* method will become just a variance *altering* method.

For any probability g, the target quantity may be written

$$\begin{aligned} A &= \mathrm{E}_f[V(X)] \\ &= \int V(x) f(x)\, dx \\ &= \int V(x) \frac{f(x)}{g(x)}\, g(x)\, dx \\ A &= \mathrm{E}_g[L(X)V(X)] \quad , \text{ where } \quad L(x) = \frac{f(x)}{g(x)}\ . \end{aligned} \tag{18}$$

Here, L is the *likelihood ratio*, a term borrowed from statistics. The original rare event problem (16) takes V to be the *indicator* function $V(x) = \mathbf{1}_R(x)$, which takes the values $\mathbf{1}_R(x) = 1$ if $x \in R$ and $\mathbf{1}_R(x) = 0$ if $x \notin R$. The variance that is relevant for the direct estimator is

$$\sigma_d^2 = \int \left(V(x) - A\right)^2 f(x)\, dx\ .$$

The variance for the importance sampling estimator is

$$\sigma_{is}^2 = \int \left(L(x)V(x) - A\right)^2 g(x)\, dx\ .$$

The goal of this section is to study f and V and learn how to identify a g so that σ_{is} is as small as possible.

To find an importance function, we try to identify the mechanism by which the rare event $X \in R$ happens. A "mechanism", when there is one, is a small region of R where that contains most of the probability that is in R. To say this differently, rare events are not predictable, but they happen in predictable ways. We cannot say *when* a rare event will happen, but we can say *how* it will happen.

A direct application of "how rare events happen" philosophy is to find the most likely point in the rare event set:

$$x_0 = \arg \max_{x \in R} f(x)\ .$$

One might hope that $f(x_0)$ is an estimate of the probability (16), and that x_0 suggests the mechanism. But this cannot be exactly right because $f(x_0)$ is a probability *density*, not a probability. To get a probability, you need to know something about the the size of the set in R where $f(x)$ is not too different from $f(x_0)$. Subsection **??** gives an example where this direct approach works. Subsection 7.2 gives an example where you have to work harder. There the philosophy works only in some generalized sense.

7.1 Asymptotic methods of integration

Mathematical analysis of rare events proceeds by introducing a small parameter, ε. The probability density f or the event R are made to depend on ε in a way that $\mathrm{P}_\varepsilon[X \in R_\varepsilon] \to 0$ as $\varepsilon \to 0$. The event R is rare if R does not come close to the "center" of f where the probability is large. To make this happen as $\varepsilon \to 0$, we can make R_ε move away from the center, or we can fix R and make f_ε increasingly concentrated about a point $x_* \notin R$. Following tradition, we take the f_ε approach here, though the R_ε approach is equivalent.

We have seen that probability densities often are expressed as exponentials. Here is a density that becomes concentrated about a point

$$f_\varepsilon(x) = \frac{1}{Z_\varepsilon} e^{-\phi(x)/\varepsilon} \ . \tag{19}$$

The astute reader will see that this is equivalent to the Gibbs Boltzmann distribution from Week 1, with ε playing the role of temperature, and $\beta = 1/\varepsilon$. This subsection discusses methods for estimate integrals involving f_ε in the limit $\varepsilon \to 0$. The *Laplace method* of asymptotic integration is for integrals involving f_ε that include a neighborhood of x_*, the minimizer of ϕ and the maximizer of f. The other method, which is simpler and doesn't seem to have a name, is for integrals that exclude a neighborhood of x_*. The Laplace approximation is the basis of the *BIC* (*B*ayesian *I*nformation *C*riterion) for Bayesian model selection.

Suppose $\phi(x)$ is a smooth function of $x \in \mathbb{R}^n$ so that $\phi(x) > \phi(x_*)$ if $x \neq x_*$. Let $H = \mathrm{D}^2\phi(x_*)$ be the Hessian matrix of ϕ at x_*. We know H is positive semi-definite, because x_* is a local minimizer of ϕ. The minimum is *non-degenerate* if H is positive definite. The Laplace approximation for this case is

$$\int e^{-\phi(x)/\varepsilon}\, dx = (2\pi)^{n/2} \frac{\varepsilon^{n/2}}{\sqrt{\det(H)}} e^{-\phi(x_*)/\varepsilon} \left(1 + O(\varepsilon)\right) \ . \tag{20}$$

This formula and $O(\varepsilon)$ error are true provided that ϕ has partial derivatives up to order 4, and x_* is a non-degenerate global minimum, and if $\phi(x) \to \infty$ as $|x| \to \infty$ fast enough. It suffices, for example, that $\phi(x) \geq C_1 + C_2 |x|^p$ for any $p > 0$.

The most significant content of the Laplace approximation formula (20) is the exponential $e^{-\phi(x_*)/\varepsilon}$. If we write $A(\varepsilon)$ for the integral on the left, the simplest Laplace approximation is

$$A(\varepsilon) = \int e^{-\phi(x)/\varepsilon}\, dx \sim e^{-\phi(x_*)/\varepsilon} \ .$$

This tells us that $A(\varepsilon) \to 0$, or $A(\varepsilon) \to \infty$, exponentially as $\varepsilon \to 0$, depending on the sign of $\phi(x_*)$. Multiplying the exponential term is the *prefactor*

$$(2\pi)^{n/2} \frac{\varepsilon^{n/2}}{\sqrt{\det(H)}} \ .$$

The prefactor behaves like a power of ε as $\varepsilon \to 0$. The prefactor is usually harder to identify than the the exponential factor. The exponential factor just depends on $\phi(x_*)$, not on the dimension, or the behavior of ϕ near x_*. The prefactor contains a dimension dependent constant, $(2\pi)^{n/2}$, a dimension dependent power of ε, and derivative information on ϕ. The prefactor is particularly hard to identify in rare event problems involving paths, which are infinite dimensional. Theorems about rare events are often called *large deviation* theorems, because large deviations are rare. If you can only identify the exponential and not the prefactor, the statement would be

$$\lim_{\varepsilon \to 0} -\varepsilon \log(A(\varepsilon)) = \phi(x_*) \ .$$

We do not give a complete proof of the Laplace formula (20), the main idea was present in Assignment 1. Most of the integral is determined by a small neighborhood of x_*. Near x_*, the exponent may be approximated by the lowest order Taylor approximation that leads to a finite integral, which is, as we saw last week,

$$\phi(x) \approx \phi(x_*) + \tfrac{1}{2}\left(x - x_*\right)^t H \left(x - x_*\right) \ .$$

If we replace $\phi(x)$ by this Taylor approximation, the integral becomes exactly the Laplace approximation, without the $O(\varepsilon)$ correction. The correction comes from corrections to the Taylor expansion. The cubic correction term is

$$C(x - x_*) = \tfrac{1}{6} \sum_{ijk} \partial_{x_i} \partial_{x_j} \partial_{x_k} \phi(x_*) \left(x_i - x_{*i}\right) \left(x_j - x_{*j}\right) \left(x_k - x_{*k}\right) \ .$$

The quartic correction term is

$$Q(x - x_*) = \tfrac{1}{24} \sum_{ijkl} \partial_{x_i} \partial_{x_j} \partial_{x_k} \partial_{x_l} \phi(x_*) \left(x_i - x_{*i}\right) \left(x_j - x_{*j}\right) \left(x_k - x_{*k}\right) \left(x_l - x_{*l}\right) \ .$$

The Taylor approximation that uses these terms is

$$\phi(x) \approx \phi(x_*) + \tfrac{1}{2}\left(x - x_*\right)^t H \left(x - x_*\right) + C(x - x_*) + Q(x - x_*) + O\left(\left|x - x^*\right|^5\right) \ .$$

The integral, using the approximate ϕ up to order four, is

$$A(\varepsilon) \approx e^{\phi(x_*)} \int e^{-\frac{1}{2\varepsilon}(x - x_*)^t H (x - x_*)} e^{-\frac{1}{\varepsilon}\left(C(x - x_*) + Q(x - x_*)\right)} \, dx \ .$$

A re-scaling change of variables lets us understand the relative sizes of the terms in the exponent. Let $y = \frac{1}{\sqrt{\varepsilon}}(x - x_*)$. The quadratic term transforms as

$$\tfrac{1}{2\varepsilon}\left(x - x_*\right)^t H \left(x - x_*\right) = \tfrac{1}{2} y^t H y \ .$$

The cubic and quartic terms transform as

$$\tfrac{1}{\varepsilon} C(x - x_*) = \sqrt{\varepsilon} C(y) \ , \quad \tfrac{1}{\varepsilon} Q(x - x_*) = \varepsilon Q(y) \ .$$

The differential transforms as $dx = \varepsilon^{n/2}dy$. The integral takes the form

$$A(\varepsilon) \approx \varepsilon^{n/2}\, e^{\phi(x_*)} \int e^{-\frac{1}{2}y^tHy} e^{-\left(\sqrt{\varepsilon}C(y)+\varepsilon Q(y)\right)}\, dy \;.$$

We can expand the small terms in the second exponential to order ε

$$e^{-\left(\sqrt{\varepsilon}C(y)+\varepsilon Q(y)\right)} = 1 - \sqrt{\varepsilon}C(y) + \varepsilon(\tfrac{1}{2}C(y)^2 - Q(y)) \;.$$

Then

$$\begin{aligned} A(\varepsilon) \approx \varepsilon^{n/2}\, e^{\phi(x_*)} \Bigg(& \int e^{-\frac{1}{2}y^tHy}\, dy \\ & - \sqrt{\varepsilon} \int e^{-\frac{1}{2}y^tHy} C(y)\, dy \\ & + \varepsilon \int e^{-\frac{1}{2}y^tHy} (\tfrac{1}{2}C(y)^2 - Q(y))\, dy \Bigg) \;. \end{aligned}$$

The top line gives the main term of the Laplace approximation (20). The second line, the $O(\sqrt{\varepsilon})$ term, integrates to zero because $C(y)$ is odd and the exponent is even. The third line gives the $O(\varepsilon)$ correction in (20).

These notes will not give a complete mathematical proof of the Laplace approximation formula. But it is not hard, should a reader have the interest and background, to give a proof along the lines of the problem from Assignment 1. You can split the integration domain into a part within $O(\varepsilon^{5/12})$ of x_*, where the above calculations are valid, and the rest, which integrates to something exponentially smaller. The power $p = \frac{5}{12}$ is chosen so that $p < \frac{1}{2}$, which makes the outside integral exponentially smaller, and so that $p > \frac{1}{3}$, so that $\frac{1}{\varepsilon}\left|x - x_*\right|^3 \ll 1$ in the inside integral, making the Taylor expansion of the exponential valid there. The reader who wants to learn something, and have a more elegant proof, should look up the *Morse lemma*.

The other integral approximation is for one dimensional integrals over a range that does not include x_*. Let

$$A(\varepsilon) = \int_{x_0}^{\infty} e^{-\frac{1}{\varepsilon}\phi(x)}\, dx \;.$$

Suppose $\phi'(x) > C > 0$ for all $x \geq x_0$. Then most of the "mass" of this integral is at the left endpoint, and the exponential part of $A(\varepsilon)$ should be $e^{-\frac{1}{\varepsilon}\phi(x_0)}$ We find the prefactor and correction terms using Taylor series, as for the Laplace approximation. This is, near x_0,

$$\phi(x) \approx \phi(x_0) + \phi'(x_0)(x - x_0) + \tfrac{1}{2}\phi''(x_0)(x - x_0)^2 \;.$$

This approximation in the $A(\varepsilon)$ integral gives (using Laplace approximation calculations)

$$
\begin{aligned}
A(\varepsilon) &\approx e^{-\frac{1}{\varepsilon}\phi(x_0)} \int_{x_0}^{\infty} e^{-\frac{1}{\varepsilon}\phi'(x_0)(x-x_0)} e^{-\frac{1}{2\varepsilon}\phi''(x_0)(x-x_0)^2}\, dx \\
&= \varepsilon e^{-\frac{1}{\varepsilon}\phi(x_0)} \int_0^{\infty} e^{-\phi'(x_0)y} e^{-\frac{\varepsilon}{2}\phi''(x_0)y^2}\, dy \\
&\approx \varepsilon e^{-\frac{1}{\varepsilon}\phi(x_0)} \int_0^{\infty} e^{-\phi'(x_0)y} \left(1 - \frac{\varepsilon}{2}\phi''(x_0)y^2\right) dy \\
A(\varepsilon) &\approx \frac{\varepsilon}{\phi'(x_0)} e^{-\frac{1}{\varepsilon}\phi(x_0)} \left(1 - \varepsilon\frac{\phi''(x_0)}{\phi'(x_0)^2}\right) \; . \qquad (21)
\end{aligned}
$$

The leading order prefactor is $\frac{\varepsilon}{\phi'(x_0)}$. It would be the exact answer if ϕ were linear.

In applications of these approximations, we do not always bother to put the integral into the precise forms given. For example, suppose $X \sim \mathcal{N}(0,1)$ is a standard normal, and we want an approximate expression for $\text{P}(X > x) = 1 - N(x)$ when x is a large positive number. The integral for this is

$$
\text{P}(X > x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-y^2/2}\, dy \; .
$$

We see that $\phi(y) = y^2/2$, and $\phi'(x) = x$. Therefore

$$
\begin{aligned}
\text{P}(X > x) &\approx \frac{1}{\sqrt{2\pi}} \frac{1}{\phi'(x)} e^{-\phi(x)} \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2} \; .
\end{aligned}
$$

This is reasonably accurate already when $x = 2$. The approximation gives

$$
\text{P}(X > 2) \approx \frac{1}{\sqrt{2\pi}} \frac{1}{2} e^{-2} = .026695 \; .
$$

The exact answer is .02275. The relative error is $(.026695 - .02275)/.02275 =$ 18%. The correction term in (21) suggests that the relative error is approximately $\phi''(x)/\phi'(x)^2 = 1/x^2$, which suggests that it decays by a factor of 4 if you increase x from 2 to 4. In fact, the relative error at $x = 4$ is 5.4%, which is a little larger than $18\%/4 = 4.5\%$.

7.2 Cramer's theorem

Suppose Y_k are independent with common density $g(y)$ and mean (the purpose of the subscript will be clear soon)

$$
x_0 = \text{E}_0[Y] = \int y g(y)\, dy \; .
$$

Cramer's theorem gets at the probability that the empirical mean of n samples is very different from x_0. If $x > x_0$, we want to estimate

$$\mathrm{P}(X \geq x) \ ,$$

where X is the empirical mean

$$X = \frac{1}{n}\sum_{k=1}^{n} Y_k \ .$$

Cramer's theorem estimates this probability when n is large, provided the probability density g has *exponential tails*, which means that the *exponential moments* are finite:

$$Z(\lambda) = \mathrm{E}\big[e^{\lambda Y}\big] = \int e^{\lambda y} g(x)\, dy < \infty \ , \tag{22}$$

at least for a range of λ values. The result is that the probability is exponentially small:

$$\mathrm{P}(X \geq x) \sim e^{-nC(x)} \ .$$

The derivation also gives the prefactor. It suggests a way to do rare event sampling in this case.

We start by writing the probability density, $f(x)$, for X.

$$f(x) = \int\int\cdots\int g(y_1)g(y_2)\cdots g(y_n)\,\delta\!\left(x - \frac{y_1 + \cdots + y_n}{n}\right) dy_1\cdots dy_n \ .$$

This is because $X = x$ is equivalent to $x = \frac{1}{n}\left(Y_1 + \cdots + Y_n\right)$. (Reasoning with delta functions may be tricky. For example, $X = x$ is also equivalent to $nx = Y_1 + \cdots + Y_n$. You can check that our f is a probability by integrating with respect to x and getting 1. You can derive an equivalent expression without delta functions but with convolutions. Our formula is more convenient for calculations.) The trick of Cramer's theorem has several natural motivations, none of which are presented here. It is that if $nx = y_1 + \cdots + y_n$, then

$$e^{n\lambda x} = e^{\lambda y_1}\cdots e^{\lambda x_n} \ .$$

Therefore

$$e^{n\lambda x} f(x) = \int\cdots\int e^{\lambda y_1} g(y_1)\cdots e^{\lambda y_n} g(y_n)\,\delta\!\left(x - \frac{y_1 + \cdots + y_n}{n}\right) dy_1\cdots dy_n \ .$$

Now, $e^{\lambda y} g(y)$ is not a probability density, but it can be normalized to be a probability density

$$g(y,\lambda) = \frac{1}{Z(\lambda)} e^{\lambda y} g(y) \ . \tag{23}$$

Therefore

$$\begin{aligned} &Z(\lambda)^{-n} e^{n\lambda x} f(x) \\ &\quad = \int\cdots\int g(y_1,\lambda)\cdots g(y_n,\lambda)\,\delta\!\left(x - \frac{y_1 + \cdots + y_n}{n}\right) dy_1\cdots dy_n \ . \end{aligned} \tag{24}$$

The integral on the right is the probability density of X if $Y \sim g(y, \lambda)$. The density $g(\cdot, \lambda)$ given by (23) is the original density modified by an *exponential twist*. For positive λ, this twist "pulls" the expected value to the right. The twisted expectation is

$$x_\lambda = \mathrm{E}_\lambda[Y] = \int y g(y, \lambda)\, dy = \frac{1}{Z(\lambda)} \int y e^{\lambda y} g(y)\, dy \; .$$

If $\lambda > 0$, then $x_\lambda > x_0$. In fact, x_λ is a strictly increasing function of λ. Therefore the inverse relation is well defined. For "any" x, there is a $\lambda(x)$ so that $x_{\lambda(x)} = x$.

Here's the important observation (motivation exists but isn't given here): If we use $\lambda(x)$ so that $\mathrm{E}_\lambda[Y] = x$, then the central limit theorem gives an estimate of the right side of (24). If $\mathrm{E}_\lambda[Y] = x$, then X has the approximate density $\mathcal{N}(x, \frac{\sigma^2}{n})$, where the appropriate variance is

$$\sigma^2 = \mathrm{var}_{\lambda(x)}(Y) = \mathrm{E}_\lambda\Big[(Y-x)^2\Big] = \frac{1}{Z(\lambda)} \int (y-x)^2 e^{\lambda y} g(y)\, dy \; .$$

If $X \sim \mathcal{N}(x, \frac{\sigma^2}{n})$, then the right side of (24) is

$$f_{\lambda(x)}(x) = \frac{1}{\sqrt{2\pi \frac{\sigma^2}{n}}} \; .$$

Therefore

$$f(x) \approx \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} \exp\big\{-n\left[x\lambda(x) - \log(Z(\lambda(x))\right]\big\} \; . \tag{25}$$

8 Exercises and examples

1. Control variates often help estimate the difference between the true answer and an analytic approximation. In an earlier lecture we discussed using a Gaussian to approximate the probability density

$$f(x) = \frac{1}{Z(\beta)} e^{-\beta\phi(x)}$$

when β is large (low temperature). We suppose the minimum of ϕ is at $x = x_0$. The leading order Taylor series approximation of ϕ about x_0 is given by the Hessian matrix $H = \phi''(x_0)$. It is

$$\phi(x) \approx \phi(x_0) + \tfrac{1}{2}(x-x_0)^t H (x-x_0) \; .$$

The problem will be to estimate the *specific heat*, for which we need

$$\mathrm{E}_\beta[\phi(X)] = \frac{\int \phi(x) e^{-\beta\phi(x)}\, dx}{\int e^{-\beta\phi(x)}\, dx} \; .$$

We apply the weighted direct sampling idea from last week. This leads us to the problem of evaluation

$$A(\beta) = \int \phi(x) e^{-\beta\phi(x)}\, dx \; .$$

We write this as an expectation with respect to the Gaussian probability distribution of the low temperature approximation

$$g(x) = \frac{\det(H)^{1/2}}{(2\pi)^{d/2}} e^{-\beta(x-x_0)^t H(x-x_0)/2}$$

After some algebra, we find

$$A(\beta) = \frac{(2\pi)^{d/2}\, e^{-\beta\phi(x_0)}}{\det(H)^{1/2}} \, \mathrm{E}_g\Big[\phi(X)\, e^{-\beta(\phi(X)-\phi(x_0))}\Big] \; .$$

Now (finally coming to the point), there is a formula for Gaussian expectation of $(x-x_0)^t H(x-x_0)$. Independent of H, depending only on the dimension,

$$\mathrm{E}_g\big[(x-x_0)^t H(x-x_0)\big] = d \; .$$

Therefore, we are in the situation of Section 4, with

$$V(x) = \phi(x)\, e^{-\beta(\phi(x)-\phi(x_0))} \;\; , \quad W(x) = (x-x_0)^t H(x-x_0) \; .$$