Lecture Notes on Monte Carlo Methods
Fall Semester, 2007 (half done)
Courant Institute of Mathematical Sciences, NYU
Jonathan Goodman, goodman@cims.nyu.edu

Chapter 3: Variance Reduction.

1 Introduction

Variance reduction is the search for alternative and more accurate estimators of a given quantity. The possibility of variance reduction is what separates Monte Carlo from direct simulation. Simple variance reduction methods often are remarkably effective and easy to implement. It is good to think about them as you wait for a long Monte Carlo computation to finish. In some applications, such as rare event simulation and quantum chemistry, they make practicle what would be impossible otherwise. Most advanced Monte Carlo is some kind of variance reduction.

Among the many variance reduction techniques, which may be used in combination, are *control variates*, *partial integration*, *systematic sampling*, *re-weighting*, and *importance sampling*. The method of control variates is useful when a crude version of the problem can be solved explicitly. This is often the case in simple problems (possibly the definition of "simple") such pricing problems in quantitative finance where the crude solvable version could be Black Scholes. Partial integration, also called *Rao Blackwellization* lowers variance by replacing integrals over come variables or over parts of space by their averages. Systematic sampling methods range from the simplest, *antithetic variates*, to the slightly more sophisticated *stratified sampling*, to *quasi Monte Carlo* integration. Re-weighting means giving different weight to different samples. They serve many functions in Monte Carlo, one being to choose weights so that certain known functions give the exact answer. Importance sampling has appeared already as sampling with a weight function. It also is the basis of reweighting and score function strategies for sensitivity analysis. Methods for *rare event sampling* mostly use importance functions, often suggested by the mathematical theory of *large deviations*.

2 Control variates

Suppose X is a random variable and that we want to evaluate

$$A = E[V(X)] \; .$$

We may estimate A by generating L independent samples of X and taking

$$\widehat{A} = \frac{1}{L} \sum_{k=1}^{L} V(X_k) \; . \tag{1}$$

The error is of the order of

$$\widehat{A} - A \sim \frac{\sigma_V}{\sqrt{L}} \ , \quad \sigma_V^2 = E\left[(V(X) - A)^2\right] \ .$$

Thus, the number of samples and the run time needed to achieve a given accuracy is inversely proportional to the variance.

A *control variate* is an easily evaluated random variable, $W(X)$, so that $B = E[W(X)]$ is known. If $W(X)$ is correlated with $V(X)$ with covariance

$$C_{VW} = \mathrm{cov}(V, W) = E\left[(V - A)(W - B)\right] \ ,$$

then the random variable,

$$Z = V(X) - \alpha\left(W(X) - B\right) \ , \tag{2}$$

can have less variance than $V(X)$. This will make the control variate estimator

$$\widehat{A} = \frac{1}{L}\sum_{k=1}^{L}\left(V(X_k) - \alpha W(X_k)\right) + \alpha B \tag{3}$$

more accurate than the simple one (1), often dramatically so.

We choose α to minimize the variance of Z in (2). The variance is

$$\sigma_Z^2 = \sigma_V^2 - 2\alpha C_{VW} + \alpha^2\sigma_W^2 \ .$$

The optimal α is

$$\alpha^* = \frac{C_{VW}}{\sigma_W^2} \ , \tag{4}$$

and the corresponding optimal variance is

$$\sigma_Z^2 = \sigma_V^2 - \frac{C_{CW}^2}{\sigma_W^2} = \sigma_V^2\left(1 - \rho_{VW}^2\right) \ , \tag{5}$$

in terms of the correlation coefficient

$$\rho_{VW} = \mathrm{corr}(V, W) = \frac{C_{VW}}{\sigma_V\sigma_W} \ .$$

Thus, the quality of W as a control variate depends on the correlation between V and W.

In practice it is not likely that one would know the optimal α (4) in advance, but it can be estimated from Monte Carlo data. From the samples X_k we can evaluate $V_k = V(X_k)$ and $W_k = W(X_k)$, then

$$\begin{aligned}
\widehat{\sigma_W^2} &= \frac{1}{L}\sum_{k=1}^{L}\left(W_k - B\right)^2 \ , \\
\widehat{A}^{(1)} &= \frac{1}{L}\sum_{k=1}^{L} V_k \qquad \text{(simple estimator of } A\text{)},
\end{aligned}$$

$$
\begin{aligned}
\widehat{C_{VW}} &= \frac{1}{L}\sum_{k=1}^{L}\left(V_k - \widehat{A}^{(1)}\right)(W_k - B) \ , \\
\widehat{\alpha^*} &= \frac{\widehat{C_{VW}}}{\widehat{\sigma_W^2}} \ , \\
\widehat{A} &= \widehat{A}^{(1)} - \widehat{\alpha^*}\frac{1}{L}\sum_{k=1}^{L}(W_k - B) \ .
\end{aligned} \tag{6}
$$

The estimate (6) may not be a very accurate estimate of (4), but the perfomance does not depend strongly on α when α is close α^*, where the derivative is zero.

One can use more than one control variate. Given $W_1(X)$, ..., $W_n(X)$ with $B_l = E[W_l(X)]$ known, we can form

$$
Z = V(X) - \sum_{l=1}^{n}\alpha_l\left(W_l(X) - B_l\right) \ . \tag{7}
$$

The optimal coefficients, the α_l that minimize $\mathrm{var}(Z)$, are found by solving the system of linear equations

$$
\mathrm{cov}(V, W_l) = \sum_{l=1}^{n}\mathrm{cov}(W_l, W_m)\alpha_m \ . \tag{8}
$$

Should the coefficients in (8) be unknown, we can estimate them from Monte Carlo data as above.

Let $V_{\mathcal{S}}$ denote the control variate sum on the right of (7) so that $V = Z + V_{\mathcal{S}}$. The optimality conditions for the coefficients α_l imply that $V_{\mathcal{S}}$ is uncorrelated with Z. If this were not so, we could use $W = V_{\mathcal{S}}$ as an additional control variate and further reduce the variance. Because they are uncorrelated, $\mathrm{var}(V) = \mathrm{var}(Z) + \mathrm{var}(V_{\mathcal{S}})$. In statisticians' terminology, the total variance of V is the sum of the *explained* part, $\mathrm{var}(V_{\mathcal{S}})$, and the *unexplained* part, $\mathrm{var}(Z)$.

Linear algebra has a geometrical way to express this. Given a random variable, X, there is a vector space consisting of mean zero functions of X with finite variance. If $V(X) - A$ and $W(X) - B$ are two such, their *inner product* is $\langle V - A, W - B\rangle = \mathrm{cov}(V, W)$. The corresponding length is $\|V\|^2 = \langle V - A, V - A\rangle = \mathrm{var}(V)$. In the vector space is the subspace, $\mathcal{S}$, spanned by the vectors $W_l(X) - B_l$. Minimizing $\mathrm{var}(Z)$ in (7) is the same as finding the $V_{\mathcal{S}} \in \mathcal{S}$ that minimizes $\|V - V_{\mathcal{S}}\|^2$. This is the element of $V_{\mathcal{S}}$ closest to $V - A$. In this way we write $V = Z + V_{\mathcal{S}}$ with $V_{\mathcal{S}}$ perpendicular to Z.

Example: From the introduction. Let $B \subset R^3$ be the *unit ball* of points with $|x| \le 1$. Suppose X and Y are independent and uniformly distibuted in B and try to evaluate

$$
E\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] \ .
$$

Since the functional $V(X,Y) = \frac{e^{-\lambda|X-Y|}}{|X-Y|}$ depends on $|X-Y|$, we seek control variates that have this dependence, the difficulty being finding functionals whose

expected value is known. One possibility is $W_1(X,Y) = |X-Y|^2$ with

$$E\left[W_1\right] = E\left[|X|^2\right] + 2E\left[\langle X, Y\rangle\right] + E\left[|Y|^2\right] \ .$$

The middle term on the right vanishes because X and Y are independent. The other two each are equal to $\frac{3}{5}$, so $E[W_1] = \frac{6}{5}$. With $\lambda = .2$, the improvement takes us from $\mathrm{var}(V) \approx .99$ to $\mathrm{var}(Z) \approx .72$, about 26% lower. Another possibility is $W_2 = |X-Y|^4$ with $E\left[W_2\right] = \frac{6}{7} + \frac{6}{5}$. Using these two control variates together gives $\mathrm{var}(Z) \approx .58$, an almost 50% reduction. The Matlab program that does this, CV1.m, is posted.

This example shows a relatively modest variance reduction from two not very insightful control variates. Variance reduction methods that seem impressive in one dimensional examples may become less effective in higher dimensional problems, as this relatively modest six dimensional problem illustrates.

3 Partial averaging

Partial averaging[1], or Rao-Blackwellization, reduces variance by averaging over some of the variables or over part of the integration domain. for example, suppose (X,Y) is a random variable wtih probability density $f(x,y)$. Let $V(X,Y)$ be a random variable and

$$\widetilde{V}(x) = E\left[V(X,Y) \mid x\right] = \frac{\int V(x,y) f(x,y) dy}{\int f(x,y) dy} \tag{9}$$

A simple inequality shows that except in the trivial case where V already was independent of y,

$$\mathrm{var}(\widetilde{V}) < \mathrm{var}(V) \ . \tag{10}$$

In fact, the reader can check that

$$\mathrm{var}(V) = \mathrm{var}(\widetilde{V}) + E\left[\left(V - \widetilde{V}\right)^2\right] \ . \tag{11}$$

The conclusion is that if a problem can be solved partially, if some of the integrals (9) can be computed explicitly, the remaining problem is easier.

A more abstract and general version of of the partial averaging method is that if $\mathcal{G}$ is a sub $\sigma-$algebra and

$$\widetilde{V} = E\left[V \mid \mathcal{G}\right] \ ,$$

then we again have (11) and the variance reduction property (10). Of course, the method still depends on being able to evaluate $\widetilde{V}$ efficiently.

[1]This has nothing to do with integration by parts. Here, it means integrating over some but not all of the variables in a multi-dimensional integral.

Subset averaging is another concrete realization of the partial averaging principle. Suppose B is a subset (i.e., an event) and that $E[V \mid B]$ is known. If

$$\widetilde{V}(x) = \begin{cases} E[V \mid B] & \text{if } x \in B, \\ V(x) & \text{if } x \notin B, \end{cases}$$

then again $\mathrm{var}(V) < \mathrm{var}(\widetilde{V})$ except in trivial situations. For example, we might take B to be the largest set for which $E[V \mid B]$ can be evaluated by symmetry.
Example. Consider just the Y integration in the previous example.:

$$E_Y\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] = \frac{3}{4\pi}\int_{|y|\le 1}\frac{e^{-\lambda|x-y|}}{|x-y|}dy\ .$$

For each x, define $B_x = \{y \mid |x-y| \le 1-|x|\}$. This is the largest round ball about x contained in the integration domain $|y| \le 1$. The conditional expectation

$$E_Y[V(x,Y) \mid B_x] = \frac{\frac{3}{4\pi}\int_{y\in B_x}\frac{e^{-\lambda|x-y|}}{|x-y|}dy}{P(Y \in B_x)}$$

may be evaluated using radial symmetry. The numerator is

$$\begin{aligned}\frac{3}{4\pi}\int_{r=0}^{1-|x|}\frac{e^{-\lambda r}}{r}4\pi r^2 dr &= 3\int_{r=0}^{1-|x|}e^{-\lambda r}r\,dr \\ &= \frac{3}{\lambda^2}\left(1-e^{-\lambda(1-|x|)}\left(1+\lambda(1-|x|)\right)\right)\ ,\end{aligned}$$

And $P(Y \in B_x) = (1-|x|))^3$, so that

$$\begin{aligned}&E_Y[V(x,Y) \mid B_x] \\ &= u(x) = (1-|x|))^{-3}\frac{3}{\lambda^2}\left(1-e^{-\lambda(1-|x|)}\left(1+\lambda(1-|x|)\right)\right)\ . \qquad (12)\end{aligned}$$

Therefore

$$A = E_{(X,Y)}\left[\frac{e^{-\lambda|X-Y|}}{|X-Y|}\right] = E_{(X,Y)}\left[\widetilde{V}(X,Y)\right]\ ,$$

where

$$\widetilde{V}(X,Y) = \begin{cases} \dfrac{e^{-\lambda|X-Y|}}{|X-Y|} & \text{if } |X-Y| \ge 1-|X|\ , \\ u(X) & \text{if } |X-Y| < 1-|X|\ . \end{cases}$$

Computational experiments (Matlab script CV3.m posted) with $\lambda = .2$ show that $\mathrm{var}(\widetilde{V}) \approx .61$. We may further reduce the variance using the earlier control variates $W_1 = |X-Y|^2$ and $W_2 = |X-Y|^4$. Using only W_1 gives $\mathrm{var}(Z) \approx .35$. Using W_1 and W_2 together gives $\mathrm{var}(Z) \approx .24$. Thus, the combined effects of not very sophisticated partial averaging and two simple control variates reduces the variance, and the work needed to achieve a given accuracy, by a factor of 4 (from .99 to .24).

4 Importance sampling

Suppose $f(x)$ is the probability density for the random variable X and that $g(x)$ is another probability with the property that $f(x) = 0$ for all x with $g(x) = 0$. If is true, we say f is *absolutely continuous* with respect to g. The *likelihood ratio* between f and g is $L(x) = f(x)/g(x)$. It is well defined if f is absolutely continuous with respect to g. Importance sampling is based on the simple identity

$$E_f\left[V(X)\right] = \int V(x)f(x)dx = \int V(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[V(X)L(X)\right] \ . \quad (13)$$

The variance is reduced if

$$\mathrm{var}_g\left[V(X)L(X)\right] < \mathrm{var}_f\left[V(X)\right] \ .$$

Importance sampling is helpful is cases where $V(x)$ is largest for values of x that are unlikely in the f probability distribution. Then we seek a g distribution that puts more weight on the most important (for the expected value of V) x values. Of course, it can take some ingenuity to identify the most important x values and even more to find a simple g that puts it's probability mass in those places.

Importance sampling is one approach to the general problem of *rare event simulation*. This means estimating the probability of an unlikely event using Monte Carlo. Applications call for estimating probabilities as large as 5% or as small as one part in 10^9. For example, in nuclear reactor shielding, we want fewer than one neutron out of 10^9 to succeed in traveling from the reactor core to where people are. As we will see near (15) below, the accuracy of direct rare event simulation depends on the number of *hits* rather than on the number of trials. A hit is a simulation in which the rare event happens. We hope for a method that requires fewer than 10^9 simulations to estimate the rare event probability.

For a very simple example, let f correspond to the standard normal distribution, let r be a large number, and ask $P(X > r)$. This is the same as $E[V(X)]$ where $V(x)$ is the characteristic function $\mathbf{1}_{x>r}$. Clearly, f puts most weight near $x = 0$, while $V = 0$ there. The probability distribution $g = \mathcal{N}(r, 1)$ puts more mass where $V \neq 0$. The likelihood ratio is

$$L(x) = f(x)/g(x) = \frac{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-(x-r)^2/2}} = e^{-rx+r^2/2} \ .$$

Therefore,

$$A = P_{\mathcal{N}(0,1)}\left[X > r\right] = e^{r^2/2}E_{\mathcal{N}(r,1)}\left[\mathbf{1}_{x>r}(X)\ e^{-rX}\right] \ . \quad (14)$$

Let us compare the Monte Carlo algorithms corresponding to the two sides of (14). The right hand side asks us to generate L standard normals, $X_k \sim \mathcal{N}(0, 1)$,

and count the number of *hits*, $N = \#\{X_k > r\}$. The estimate of A is N/L. The right hand side asks us to generate L independent normals with mean r and variance one, $\widetilde{X}_k \sim \mathcal{N}(r,1)$. Approximately half of these will be hits in the sense that $\widetilde{X}_k > r$. Each hit is re-weighted with a factor $w_k = e^{-r\widetilde{X}_k + r^2/2}$. The w_k are small for $\widetilde{X}_k$ in the hit region. Indeed, when $\widetilde{X}_k > r$, we have $-r\widetilde{X}_k - r^2/2 < -r^2/2 << 0$. Therefore, the right hand side estimator,

$$\frac{e^{r^2/2}}{L} \sum_{\widetilde{X}_k > r} e^{-r\widetilde{X}_k} \; ,$$

estimates A using a large number of very small contributions, rather than a very small number of order one contributions.

We can make this more quantitative using the idea of *relative accuracy*. The *absolute error* of an estimator is $\widehat{A} - A$, while the relative error is $\left(\widehat{A} - A\right)/A$. This is particularly important when A is very small. If $A = 10^{-6}$, then the estimate $\widehat{A} = 2 \cdot 10^{-6}$ is off by 100%, although that is only 10^{-6} in absolute terms. In the present example, random variables $Y_1 = \mathbf{1}_{x>r}(X)$, $X \sim \mathcal{N}(0,1)$, and $Y_2 = e^{r^2/2}\mathbf{1}_{x>r}(X)e^{-rX}$, $X \sim \mathcal{N}(r,1)$. Corresponding to these are the two estimators

$$\widehat{A}_j = \frac{1}{L}\sum_{k=1}^{L} Y_{j,k} \; .$$

We will find the natural condition that implies that the naive estimator $\widehat{A}_1$ has some relative accuracy and see how much better the importance sampling $\widehat{A}_2$ is.

The random variable Y_1 is Bernoulli, and has (using the approximate formula from the next section)

$$p \;=\; P(Y_1 = 1) \;=\; A \;\approx\; \frac{1}{\sqrt{2\pi}r} e^{-r^2/2} \; .$$

The variance of a Bernoulli is $p(1-p) \approx p$ (the last for small p). The relative accuracy measured by the standard deviation of the estimator normalized by the exact answer

$$\frac{\sigma\left(\widehat{A}_1\right)}{A} \;=\; \frac{\sigma\left(Y_1\right)}{\sqrt{L}A} \;\approx\; \frac{1}{\sqrt{Lp}} \; . \tag{15}$$

This result simply says that the relative accuracy of the naive estimator is determined by the Lp, which is the expected number of hits in L trials. The relative accuracy depends on the (expected) number of hits, not the number of samples.

By contrast, we calculate (using approximate integration as below) that

$$E\left[Y_2^2\right] \;=\; e^{r^2}\frac{1}{\sqrt{2\pi}}\int_r^\infty e^{-2rx}e^{-(x-r)^2/2} \;\approx\; e^{-r^2}\,\frac{1}{\sqrt{2\pi}}\,\frac{1}{2r} \; .$$

Since $E[Y_2] = A$, this leads to (getting A^2 from (17))

$$\mathrm{var}[Y_2] \approx e^{-r^2} \left(\frac{1}{\sqrt{2\pi}} \frac{1}{2r} - \frac{1}{2\pi} \frac{1}{r^2} \right) \approx e^{-r^2} \frac{1}{\sqrt{2\pi}} \frac{1}{2r} .$$

Thus, we have relative accuracy

$$\frac{\sigma\left(\widehat{A}_2\right)}{\sqrt{L}A} \approx (2\pi)^{1/4} \sqrt{\frac{r}{L}} . \tag{16}$$

Although both the naive and the importance sampling methods have relative error that grows with r, the naive estimator (15) has a factor $1/\sqrt{p} \sim e^{r^2/4}$, which grows exponentially with r, while importance sampling (16) has relative error that grows only like $\sqrt{r}$. A more sophisticated method can remove even the $\sqrt{r}$ growth.

4.1 Approximate integration I

If an integral depends on a parameter, it may be possible to find the approximate value of the integral when the parameter is large or small. One simple way this happens is that most of the mass of the integral becomes concentrated in small sets. The integral approximation them comes from approximations to the integrand that are valid where the mass is.

A simple example is

$$A(r) = P_{\mathcal{N}(0,1)}(X > r) = \frac{1}{\sqrt{2\pi}} \int_{x=r}^{\infty} e^{-x^2/2} dx ,$$

when r is large. Of course, all the mass in this integral is from $x > r$, but for large r, most of the mass is very close to $x = r$. This is because $x^2/2$ is rapidly increasing at $x = r$ if r is large, which in turn makes $e^{-x^2/2}$ rapidly decreasing. Informally, there are two ranges of $x > r$, the near field and far field. In the near field the Taylor approximation, $x^2/2 \approx r^2/2 + (x - r)r$, is valid. In the far field, $x^2/2$ is so much larger than $r^2/2$, that is makes a negligible contribution to the integral. In other words, the total integral is almost the same as the near field integral, which, with the Taylor approximation, gives

$$A(r) \approx \frac{1}{\sqrt{2\pi}} \int_{x=r}^{\infty} e^{-r^2/2 - r(x-r)} dx = \frac{1}{\sqrt{2\pi} r} e^{-r^2/2} .$$

Extending the near field integral out to $x = \infty$ has a negligible effect on the answer.

Let us do this example more carefully. With the change of variables $x = r + y$, we have $x^2/2 = r^2/2 + ry + y^2/2$, and

$$A(r) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \int_{y=0}^{\infty} e^{-ry} e^{-y^2/2} dy .$$

There is a dichotomy in the integrand on the right side. Either y is small or ry is so large that the integrand is negligible. In either case (but for different reasons), we may use the Taylor approximation $e^{-y^2/2} \approx 1 - y^2/2 + y^4/8 - \cdots$ to get (after integrating term by term)

$$A(r) \approx \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \left(\frac{1}{r} - \frac{1}{r^3} + \frac{3}{r^5} + \cdots \right) . \tag{17}$$

An error bound for this case follows from

$$\left| e^{-u} - \left(1 - u + u^2/2\right) \right| \leq Cu^3 ,$$

for $u > 0$. Putting this in the $A(r)$ integral gives

$$\left| A(r) - \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \left(\frac{1}{r} - \frac{1}{r^3} + \frac{3}{r^5} + \cdots \right) \right| \leq \frac{Ce^{-r^2/2}}{r^7} .$$

This shows that for large enough r, the error in (17) is smaller than any of the terms in the approximation.

Note what the error bound did not say. It did not say that the approximation always improves when you add another term, or that the approximations converge as the number of terms goes to infinity. Approximations like (17) are *asymptotic* approximations[2] rather than convergent series. The approximation with any fixed number of terms improves as $r \to \infty$, but for fixed r there is an optimal number of terms. Adding more makes the approximation worse.

A general case of this involves a *phase* function $\phi(x)$ and an *amplitude* function $f(x)$ in the integral

$$A(r) = \int_{x=0}^{\infty} f(x) e^{-r\phi(x)} dx .$$

If we suppose that ϕ and f are smooth functions of x with $\phi' \geq C > 0$ for all $x > 0$, then for large r the mass of the integral is concentrated about $x = 0$. Using the Taylor expansions $f(x) \approx f(0) + xf'(0) + \cdots$ and $\phi(x) = \phi(0) + x\phi'(0) + \frac{1}{2}x^2\phi''(0) + \cdots$ gives

$$A(r) = e^{-r\phi(0)} \left(\frac{1}{r} \frac{f(0)}{\phi'(0)} + \frac{1}{r^2} \frac{f'(0) + f(0)\phi''(0)}{\phi'(0)^2} + O\left(\frac{1}{r^3}\right) \right) . \tag{18}$$

The reader should check that this is consistent with the special case (17).

4.2 Cramer's theorem

Cramer's theorem is the first example of what now is called *large deviation* theory, which is a general theory that tries to explain how rare events happen.

[2]This distinction and terminology are due to Poincare. It is discussed at length in many good books on methods of applied mathematics such as the one by Bender and Orszag.

The analysis also leads to highly efficient importance sampling methods for this kind of rare event.

Suppose X is a random variable with density $f(x)$ and expected value μ_0. Let $X_1, \ldots, X_n$ be n independent samples of f and let $S_n = (X_1 + \cdots + X_n)/n$ be mean of these samples. For large n, we expect $S_n \approx \mu_0$. Cramer's theorem asks about the possibility that $S_n \approx \mu$ with $\mu \neq \mu_0$. Under suitable hypotheses on f, this is exponentially unlikely, as we (following Cramer) now show.

One derivation of the Cramer result starts from the view that the central limit theorem is an asymptotic integration method. Suppose $g(x)$ is a probability density the X_k are an i.i.d. sequence of samples of g, and $h_n(s)$ is the probability density for the mean $S_n = (X_1 + \cdots + X_n)/n$. We can express h_n in terms of the delta function as

$$h_n(s) = \int_{x_1} \cdots \int_{x_n} \delta\left(s - (x_1 + \cdots + x_n)/n\right)\ g(x_1) \cdots g(x_n)\ dx_1 \cdots dx_n\ . \quad (19)$$

For most values of s this integral is hard to evaluate. However, the central limit theorem states that for $s \approx \mu = E_g[X]$, h_n corresponds to a Gaussian with mean μ and variance $\sigma_g^2(X)/n$. That is, $h_n(s) \approx \frac{1}{\sqrt{2\pi\sigma_g^2(X)n}} e^{-(s-\mu)^2/2\sigma_g^2(X)n}$. In particular, setting $s = \mu$ gives (writing μ_g for emphasis)

$$h_n(\mu_g) \approx \frac{1}{\sqrt{2\pi\sigma_g^2 n}}\ . \quad (20)$$

The rest of the argument is a clever trick that reduces general integrals (19) to this case using the *exponential twist*. There will be some motivation for this below, but for now, the twist is a factor $e^{\lambda x}$. If $f(x)$ is a probability density, the exponentially twisted density is

$$f_\lambda(x) = \frac{1}{Z(\lambda)} e^{\lambda x} f(x)\ , \quad (21)$$

where the normalization factor is chosen so that f_λ has total mass one:

$$Z(\lambda) = \int e^{\lambda x} f(x) dx = E\left[e^{\lambda X}\right]\ . \quad (22)$$

The Cramer theory depends on the *moment generating function*, or *exponential moments*, (22) being finite, at least for a range of λ. Otherwise, the results are very different. For example, if $Y \sim \mathcal{N}(0,1)$ and X is the *lognormal* $X = e^Y$, then (22) is (for $\lambda > 0$)

$$E\left[e^{\lambda e^Y}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\lambda e^y - y^2/2\right) dy = \infty\ ,$$

because $\lambda e^y - y^2/2 \to \infty$ as $y \to \infty$. The Cramer theory does not apply to the lognormal because lognormal tails are too fat. You can see how stringent the

requirement of finite exponential moments is by noting that the lognormal has finite (power) moments of all orders: $E[X^n] < \infty$ for all n.

The exponential twist in Cramer theory is used to move the mean of X from the true value μ_0 to the twisted value

$$\mu(\lambda) = E_\lambda[X] = \int x f_\lambda(x) dx = \frac{1}{Z(\lambda)} E_0\left[Xe^{\lambda X}\right] = \frac{1}{Z(\lambda)} \int x e^{\lambda x} f(x) dx \; . \tag{23}$$

We write $\lambda_*(\mu)$ for the value of λ that achieves that μ in (23) The trick is that because of the delta function in the integral (19), we have the identity

$$Z^n(\lambda)\, e^{-n\lambda s}\, \frac{e^{\lambda x_1}}{Z(\lambda)} \cdots \frac{e^{\lambda x_n}}{Z(\lambda)} \;=\; 1 \; .$$

Therefore, for any λ we may rewrite (19) as

$$h_n(s) = Z^n(\lambda)\, e^{-n\lambda s} \int \cdots \int \delta(\cdots) f_\lambda(x_1) \cdots f_\lambda(x_n) dx_1 \cdots dx_n \; .$$

If we choose $\lambda_*(s)$ then we can apply (20). The result is

$$h_n(s) \;\approx\; \frac{1}{\sqrt{2\pi n \sigma^2_{\lambda_*(s)}}} Z^n(\lambda_*(s))\, e^{-ns\lambda_*(s)} \; . \tag{24}$$

Much of the interpretation of (24) uses the *free energy*,[3] $F(\lambda) = \ln(Z(\lambda))$. The derivatives of F are (using (22) and (23))

$$F'(\lambda) = \frac{\partial_\lambda Z(\lambda)}{Z(\lambda)} = \frac{E\left[Xe^{\lambda X}\right]}{Z(\lambda)} = E_\lambda[X] \; , \tag{25}$$

and

$$F''(\lambda) = \frac{\partial^2_\lambda Z(\lambda)}{Z(\lambda)} - \left(\frac{\partial_\lambda Z(\lambda)}{Z(\lambda)}\right)^2 = E_\lambda\left[X^2\right] - \left(E_\lambda[X]\right)^2 = \mathrm{var}_\lambda(X) \; . \tag{26}$$

From (26) we learn that $F(\lambda)$ is strictly convex (unless X is trivial), which implies that there is a unique λ with $s = F'(\lambda) = E_\lambda[X]$. Any convex function, $F(\lambda)$ has a *convex conjugate* function (or *Legendre transform*)

$$F^*(s) = \min_\lambda F(\lambda) - s\lambda \; . \tag{27}$$

The following facts are easy to verify. If F is strictly convex then the minimum in (27), if there is one, is taken at a unique point, $\lambda_*(s)$. This satisfies $\partial_s \lambda_*(s) = 1/F''(\lambda_*(s)) > 0$. The conjugate function satisfies $\partial_s F^*(s) = \lambda_*(s)$

[3]The term *free energy* comes from statistical mechanics, where it refers to the part of the total energy that can be used for doing mechanical work in a certain way. See a good book on statistical physics for details.

and $\partial_s^2 F^*(s) = \partial_s \lambda_*(s) > 0$, so F^* also is strictly convex wherever it is defined. Finally, the relation $\partial_s F^* = \lambda_*$ implies that $F^{**}(\lambda) = \min_s F^*(s) - s\lambda = F(\lambda)$.

Returning to (24), which may be written

$$h_n(s) \;\approx\; \frac{1}{\sqrt{2\pi n \sigma^2_{\lambda_*(s)}}} e^{-n\{F(\lambda_*(s)) - s\lambda_*(s)\}} \quad .$$

From what we know about F and F^*, we recognize this as

$$h_n(s) \;\approx\; \frac{1}{\sqrt{2\pi n \sigma^2_{\lambda_*(s)}}} e^{-nF^*(s)} \quad . \tag{28}$$

It is a useful exercise for the reader to check that the minimum of $F^*(s)$ is attained at $s = \mu_0 = E_0[X]$. The minimum value is $F^*(\mu_0) = 0$, and $\partial_s F^*(\mu_0) = 0$, and, finally, $\partial_s^2 F^*(\mu_0) = 1/\mathrm{var}(X)$. Using this in (28) gives, for $s \approx \mu_0$,

$$h_n(s) \;\approx\; \frac{1}{\sqrt{2\pi n \sigma^2}} e^{-n(s-\mu)^2/2\sigma^2} \;,$$

which is the central limit theorem.

In discussions of the theory of large deviations, one often sees a weaker version of (28), namely

$$\lim_{n\to\infty} \frac{-1}{n} \ln\left[h_n(s)\right] \;=\; F^*(s) \;.$$

This gives information only about the exponent, not about the algebraic *pre-factor*. To be fair, even this information is hard to come by in harder cases. Still, the pre-factor has a factor of $\sqrt{n}$, so the approximate formula based only on the exponent, $h_n \sim e^{-nF^*}$, is not accurate in the usual sense. Even though large deviation "approximations" are not very accurate in themselves, the derivations often identify the most likely way for the rare events to happen, the exponential twist in this case. This leads to effective importance sampling strategies.

Let us compare the direct simulation algorithm to an importance sampling strategy using the exponential twist. We wish to evaluate $A_n(r) = P(S_n > s)$ for $s > \mu_0$. Integrating the approximation (28) using the approximate integration method above gives (using $\partial_s F^*(s) = \lambda_*(s)$)

$$\begin{aligned} A_n(s) \;&\approx\; \frac{1}{\sqrt{2\pi n \sigma^2_{\lambda_*(s)}}} \int_s^\infty e^{-nF^*(s)} ds \\ &\approx\; n^{-3/2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\lambda_*(s)} \lambda_*(s)} e^{-nF^*(s)} \end{aligned} \tag{29}$$

Of course, this shows that the event is exponentially unlikely and direct simulation will produce few hits.