

4. Importance Sampling

Applying simulation methodology is simply finding the right wrench to pound in the correct screw. *Anon.*

4.1 The Basic Problem of Rare Event Simulation

Large and/or nonlinear stochastic systems, due to analytic intractability, must often be simulated in order to obtain estimates of the key performance parameters. Typical situations of interest could be a buffer overload in a queueing network or an error event in a digital communication system. In many system designs or analyses a low probability event is a key parameter of the system's efficacy.

Since the test statistic's probability distribution is usually very difficult or impossible to compute in closed form, one is faced with the problem of computer simulation in order to find the probability of interest.

How does one go about simulating a rare event in order to find its probability? This is a more difficult problem than one may expect. Consider a sequence $\{X_j\}$ of i.i.d. Bernoulli random variables with

$$P(X_1 = 1) = \rho = 1 - P(X_1 = 0).$$

Suppose that we wish to estimate from the observed sequence the parameter ρ . We wish to have at most a 5% error on ρ with 95% confidence. This means that we must have

$$P(|\rho - \hat{\rho}| \leq .05\rho) = .95,$$

where $\hat{\rho}$ is the estimate of ρ . For example, the maximum likelihood estimate would be

$$\hat{\rho} = \frac{1}{k} \sum_{i=1}^k X_i.$$

The variance of the Bernoulli random variable X_1 is $\rho(1 - \rho)$. If ρ is very small (i.e., $\{X_1 = 1\}$ is a rare event), the variance of X_1 is approximately ρ . Hence the variance of $\hat{\rho}$ is

$$\frac{\rho(1-\rho)}{k} \approx \frac{\rho}{k}.$$

The mean value of $\hat{\rho}$ is ρ which (by definition) means that the estimate is unbiased. Hence using a Central Limit Theorem approximation, we have that

$$\begin{aligned} P(|\rho - \hat{\rho}| \leq .05\rho) &= P\left(\left|\frac{1}{\sqrt{k}} \sum_{i=1}^k \frac{X_i - \rho}{\sqrt{\rho}}\right| \leq .05\sqrt{\rho k}\right) \\ &\approx P(|Z| \leq .05\sqrt{\rho k}), \end{aligned}$$

where Z is a Gaussian, mean zero, variance 1 (i.e., standard Gaussian) random variable. Now for a standard Gaussian we have that $P(|Z| \leq z) = .95$ implies (from tables) that $z \approx 2$; that is, two standard deviations about the mean captures 95% of the probability of a Gaussian distribution. Thus we must have that $.05\sqrt{\rho k} = 2$ which in turn implies that $k = 1600/\rho$. Therefore, if ρ is somewhere on the order of 10^{-6} , we would need something like 1.6×10^9 number of samples to estimate it to the desired level of precision. This is a very large number of simulation runs and will impose severe demands on our random number generator. Unfortunately (for the simulation designer) error probabilities of this order are typical in digital communication systems and many other systems of engineering and scientific interest.

4.2 Importance Sampling

The principal method that we use to attack the rare event simulation problem is to utilize a variance reduction technique from simulation theory known as *importance sampling*. Suppose that we wish to estimate

$$\rho = \mathbb{E}[\eta(X)],$$

where X is a random variable (or vector) describing some observation on a random system. If η is the indicator function of some set (a typical case), then ρ would be the probability of that set. Suppose that the observation random variable X is controlled by a probability density function $p(\cdot)$. The direct simulation method would be to generate a sequence of i.i.d. random numbers $X^{(1)}, X^{(2)}, \dots, X^{(k)}$ from the density $p(\cdot)$ and form the estimate

$$\hat{\rho}_p = \frac{1}{k} \sum_{i=1}^k \eta(X^{(i)}).$$

Alternatively, we could generate a sequence of i.i.d. random numbers $Y^{(1)}, Y^{(2)}, \dots, Y^{(k)}$ distributed with density $q(\cdot)$. The density $q(\cdot)$ is called the *importance sampling biasing distribution*. We then form the *importance sampling estimator* or *estimate* as

$$\hat{\rho}_q = \frac{1}{k} \sum_{i=1}^k \eta(Y^{(i)}) \frac{p(Y^{(i)})}{q(Y^{(i)})}.$$

Immediately we see that we could have problems unless $q(x)$ is never zero for any value of x where $p(x)$ is positive. Mathematically, this means that the support of $q(\cdot)$ must include the support of $p(\cdot)$. However, we see that there can only be a problem if $\eta(x)$ is nonzero at x and the ratio of $p(\cdot)$ and $q(\cdot)$ blows up. This is equivalent to saying that support of $\eta(x)p(x)$ is included in the support of $\eta(x)q(x)$. Thus our requirement may be stated as

$$\text{support}(p \cdot \eta) \subset \text{support}(q \cdot \eta).$$

The expected value of $\hat{\rho}_q$ under the density $q(\cdot)$ is just

$$\begin{aligned} \mathbb{E}_q[\hat{\rho}_q] &= \frac{1}{k} \sum_{i=1}^k \int \eta(y^{(i)}) \frac{p(y^{(i)})}{q(y^{(i)})} q(y^{(i)}) dy^{(i)} \\ &= \int \eta(y) p(y) dy \\ &= \mathbb{E}_p[\eta(X)] \\ &= \rho. \end{aligned}$$

Therefore, the estimate $\hat{\rho}_q$ is unbiased and, as $k \rightarrow \infty$, we expect it to be converging by the law of large numbers to its mean value ρ .

Example 4.2.1. Suppose we are interested in

$$\rho = P\left(\frac{1}{n} \sum_{j=1}^n Z_j > T\right),$$

where $\{Z_j\}$ are i.i.d. \mathcal{R} -valued random variables with mean value zero, density function $p^*(\cdot)$, and T is a positive constant. We simulate with some other random variables $\{R_j\}$ with density function $q^*(\cdot)$ and form the importance sampling estimator,

$$\hat{\rho}_q = \frac{1}{k} \sum_{i=1}^k 1_{\{\frac{1}{n} \sum_{j=1}^n R_j^{(i)} > T\}} \frac{\prod_{j=1}^n p^*(R_j^{(i)})}{\prod_{j=1}^n q^*(R_j^{(i)})}.$$

To fit in with the theoretical framework given in the introduction, note that we have the correspondences

$$\begin{aligned}
(Z_1, Z_2, \dots, Z_n) &\Rightarrow X \\
1_{\{\frac{1}{n} \sum_{i=1}^n Z_i > T\}} &\Rightarrow \eta(X) \\
(R_1, R_2, \dots, R_n) &\Rightarrow Y \\
\prod_{j=1}^n p^*(Z_j^{(i)}) &\Rightarrow p(X^{(i)}) \\
\prod_{j=1}^n q^*(R_j^{(i)}) &\Rightarrow q(Y^{(i)}).
\end{aligned}$$

In this example we can see some of the possible utility in using an importance sampling estimator. Consider again the quantity ρ . It involves the sum of n random variables. Even if we know $p(\cdot)$, the distribution of the sum would involve the n -fold convolution of $p(\cdot)$. For large n , this could be a very difficult task even numerically. After that, then one would be forced to try to integrate over the tail of the resulting distribution, another task that could be very difficult analytically or numerically. With importance sampling, we see that knowledge of the one-dimensional densities is sufficient to come up with an unbiased estimator of ρ .

One key question of importance sampling is: Are there better choices for $q(\cdot)$ than just $p(\cdot)$ (the direct Monte Carlo choice)? Let us consider the variance of $\hat{\rho}_q$. Since this estimator is the average of k i.i.d. terms, the variance will be $1/k$ times the variance of one of the terms. Thus

$$\begin{aligned}
k \text{Var}(\hat{\rho}_q) &= \int [\eta(x) \frac{p(x)}{q(x)} - \rho]^2 q(x) dx \\
&= \int [\frac{\eta(x)^2 p(x)^2}{q(x)} - 2\rho p(x) \eta(x) + \rho^2 q(x)] dx \\
&= \int \frac{\eta(x)^2 p(x)^2}{q(x)} dx - \rho^2 \\
&= F_q - \rho^2.
\end{aligned} \tag{4.1}$$

In the above expression, we have emphasized that $k \text{Var}(\hat{\rho}_q)$ may be written as a difference of a first term, F_q , and a second term ρ^2 .

We now try to choose $q(\cdot)$ to minimize this expression,

$$\begin{aligned}
F_q &= \int \frac{\eta(x)^2 p(x)^2}{q(x)} dx \\
&= \int \frac{\eta(y)^2 p(y)^2}{q(y)^2} q(y) dy \\
&= \mathbb{E}\left[\frac{\eta(Y)^2 p(Y)^2}{q(Y)^2}\right] \\
&= \mathbb{E}\left[\left(\frac{|\eta(Y)| p(Y)}{q(Y)}\right)^2\right] \\
&= \int \frac{|\eta(y)| p(y)}{q(y)} q(y) dy \\
&= \int |\eta(y)| p(y) dy,
\end{aligned}$$

where we have used Jensen's inequality or the fact that for any random variable Z , $\mathbb{E}[Z^2] \geq (\mathbb{E}[Z])^2$. Furthermore we have strict equality if and only if Z is almost surely a constant. Thus F_q is minimized when $|\eta(Y)| p(Y)/q(Y)$ is almost surely a constant. However, this can hold only if $|\eta(x)| p(x)/q(x)$ is a constant. Thus the optimal choice for q is

$$q_{opt}(x) = \frac{p(x)|\eta(x)|}{\int |\eta(y)p(y)dy}. \quad (4.2)$$

Let us investigate $q_{opt}(\cdot)$. Suppose, for simplicity that $\eta(\cdot)$ is a non-negative function. Then we note that $F_{q_{opt}} = \rho \int \eta(x)p(x)dx = \rho^2$. Thus $k \text{Var}(\hat{\rho}_{q_{opt}}) = 0$ for all k . Unfortunately, this is not as wonderful as it might seem at first glance. In the first place, $p(\cdot)$ in many cases is not specified in closed form. It could be (for example) the distribution of a large sum of i.i.d. (as in the Example 4.2.1) or Markov distributed random variables. We may generate samples from it easily enough but explicit expressions for it are generally not available. Secondly, even if $p(\cdot)$ were known, the constant of proportionality is exactly ρ^{-1} , precisely the parameter that we are trying to estimate! Hence in computing the weighting factor for the importance sampling estimator of ρ , we first must know what ρ is. Clearly, we must search for other methods or criteria by which to choose a good simulation distribution $q(\cdot)$. This need to find good criteria for choosing the simulation distribution has sparked the vast amount of research in this area for the past two decades.

Let us see if we can elucidate some guidelines for choosing good practical simulation distributions. Consider the "optimal" choice in (4.2), for the case that $\eta(\cdot) = 1_{\{E\}}(\cdot)$. We think of the set E as being some "rare event." We can gain some insight from the optimal choice on what properties a good practical simulation distribution should have. First note that $q_{opt}(\cdot)$ puts all of its probability mass on the set or event E (i.e., its support is contained in or equal to E). Thus, intuitively, we want to choose the simulation distribution so that more events of interest occur. The second observation is that $q_{opt}(\cdot)$

has the same shape over the set E as the original distribution (in fact it *is* the same distribution except just scaled by ρ^{-1}). Thus if a region of E has more probability mass than another region of E under the original distribution, then the optimal choice will also have this property. We can summarize these two principles as

P1) Choose the simulation distribution so that we “hit” the rare event E of interest more often.

P2) Choose the simulation distribution so that the more likely or higher probability regions of E are hit more often during the simulation than the lower probability or less likely regions of E .

These properties have spawned a variety of ad hoc techniques for choosing the simulation distribution. By far the two most popular methods are variance scaling and mean translation.

The variance scaling method increases the “hit” probability by choosing as the simulation random variables, the original random variables multiplied by a constant. Typically the constant is greater than one and thus we are merely increasing the variance of the original distribution. Thus for some rare event E , typically this would put more probability mass on it, causing us to “hit” it more often during the simulation, which would satisfy our first property quite nicely. Whether the second property is satisfied depends on the problem. Typically, we would try to choose the variance scaling parameter to satisfy as much as possible, the second property.

Example 4.2.2. In Example 4.2.1, the variance scaling method would correspond to

$$q^*(x) = \frac{1}{a} p^*\left(\frac{x}{a}\right),$$

where a is the variance scaling parameter (if the original density $p^*(\cdot)$ has variance σ^2 , then $q^*(\cdot)$ has variance $a^2\sigma^2$).

The variance scaling method has been largely superseded by the mean translation method. This method seeks to increase the “hit” probability by adding a mean value to the input random variables.

Example 4.2.3. In Example 4.2.1, the mean translation method would correspond to

$$q^*(x) = p^*(x - m)$$

where m is the mean shift parameter (if the original density $p^*(\cdot)$ has mean m_o , then $q^*(\cdot)$ has mean $m_o + m$). In this method, almost always m is just chosen to be T , which partially explains its popularity. The scaling parameter in the variance scaling method has no such “default” choice available.

Recall that for the importance sampling estimator, we require that the support of $\eta(\cdot)q(\cdot)$ include the support of $\eta(\cdot)p(\cdot)$. We must always take this into account when choosing simulation distributions.

Example 4.2.4. Suppose for example that $\eta(x) = 1_{\{E\}}(x)$; that is, $\eta(\cdot)$ has support E . If the original density has support $[0, s]$ ($s > 0$), and $E = [s/2, s]$, then the variance scaling simulation distribution has support $[0, sa]$ (assuming $a > 0$). Thus

$$\begin{aligned} \text{support}(p \cdot \eta) &= [0, s] \cap E = E \\ &\subset \text{support}(q \cdot \eta) \\ &= [0, sa] \cap E = E, \end{aligned}$$

and thus satisfies our support requirement.

The mean shift method on the other hand has

$$\begin{aligned} \text{support}(q \cdot \eta) &= [m, s + m] \cap E \\ &= [m, s + m] \cap [s/2, s] \end{aligned}$$

which will violate the support requirement for any choice of $m > s/2$.

4.3 The Fundamental Theorem of System Simulation

Before we begin our study of how to choose good importance sampling biasing distributions, we need to consider some fundamental properties of importance sampling estimators. In this section, we consider a very general question in the field of system simulation: that is, “Should we bias the random variables at the input, at the output, or at some intermediate point of a system?.” To be a bit more specific, consider the following example.

Example 4.3.1. Suppose we are interested in estimating

$$\rho = P(S + N > a),$$

where S and N are two independent random variables with densities $p_s(\cdot)$ and $p_n(\cdot)$, respectively. Denote the sum of these two random variables as R with density denoted as $p_r(\cdot)$. We use importance sampling to estimate the value of ρ . We generate simulation random variables $S^{(1)'}, S^{(2)'}, \dots, S^{(k)'} i.i.d. with marginal density $p_{s'}$, $N^{(1)'}, N^{(2)'}, \dots, N^{(k)'} i.i.d. with marginal density $p_{n'}$. The sequences are also independent of each other. We also consider $R^{(1)'}, R^{(2)'}, \dots, R^{(k)'} i.i.d. with marginal density $p_{r'}$ which are generated by the relation $R^{(j)'} = S^{(j)'} + N^{(j)'}$.$$$

We consider two types of estimators, an input estimator and an output estimator. The “input” estimator is explicitly given as

$$\hat{p}_i = \frac{1}{k} \sum_{j=1}^k 1_{\{S^{(j)'} + N^{(j)'} > a\}} \frac{p_s(S^{(j)'})p_n(N^{(j)'})}{p_{s'}(S^{(j)'})p_{n'}(N^{(j)'})}$$

and the “output” estimator as

$$\hat{p}_o = \frac{1}{k} \sum_{j=1}^k 1_{\{R^{(j)'} > a\}} \frac{p_r(R^{(j)'})}{p_{r'}(R^{(j)'})}.$$

these estimators are unbiased. Which has lower variance?

We should note that in many situations, an output formulation of the bias distribution is impossible. If the system is very complicated, it may very well be virtually impossible to calculate the biasing distributions at the output of the system. However, it may very well be possible to calculate the distributions at some intermediate point of the system. In this chapter we take the first steps toward attempting to understand how much there is to gain or lose by using an input over an output formulation. It is essential to the theory of importance sampling in system simulation that we try to gain some understanding of the role of the bias point in Monte Carlo simulation.

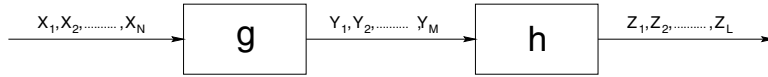


Fig. 4.1. A multi-input, multi-output system.

We are given two (Borel measurable) functions $g : \mathcal{R}^N \rightarrow \mathcal{R}^M$ and $h : \mathcal{R}^M \rightarrow \mathcal{R}^L$, which define our system as shown in Fig. 4.1. Let (X_1, X_2, \dots, X_N) be an arbitrary random vector. We consider these to be our “input random variables.” We denote their joint probability measure as P_x . Define $(Y_1, Y_2, \dots, Y_M) = g(X_1, X_2, \dots, X_N)$ which we consider to be our “intermediate random variables” with joint probability measure P_y and lastly $(Z_1, Z_2, \dots, Z_L) = h(Y_1, Y_2, \dots, Y_M)$ our “output random variables” with joint measure P_z .

Let f be a (Borel measurable) function mapping \mathcal{R}^L to \mathcal{R}^d . Suppose we are interested in the quantity

$$\begin{aligned} \rho &= \mathbb{E}[f(Z_1, \dots, Z_L)], \\ &= \mathbb{E}[f(h(Y_1, Y_2, \dots, Y_M))], \\ &= \mathbb{E}[f(h(g(X_1, X_2, \dots, X_N)))]. \end{aligned}$$

$\rho = (\rho_1, \rho_2, \dots, \rho_d)$ is of course a d -dimensional vector. The bias probability measures are always denoted with the symbol Q with a subscript to indicate which random variables are being biased, for example, Q_x, Q_y, Q_z . We assume

that the original probability measures are absolutely continuous with respect to these measures. It is enough to assume that $P_x \ll Q_x$ since this automatically implies $P_y \ll Q_y$ and $P_z \ll Q_z$. This of course guarantees the existence of the Radon–Nikodym derivatives $dP_x/dQ_x, dP_y/dQ_y, dP_z/dQ_z$ needed for our importance sampling estimators. We assume that biasing measures have the same relationship between them as do the actual measures. (We denote the biased random variables as the original random variable written with a tilde over it.) Thus, if $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N$ are generated to have measure Q_x , then $g(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)$ will have measure Q_y and $h(g(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n))$ will have measure Q_z .

Depending on at which point of the system we wish to bias, we can define various estimators of ρ . The possibilities are

$$\hat{\rho}_i = \frac{1}{k} \sum_{j=1}^k f\left(h(g(\tilde{X}_1^{(j)}, \tilde{X}_2^{(j)}, \dots, \tilde{X}_N^{(j)}))\right) \frac{dP_x}{dQ_x}(\tilde{X}_1^{(j)}, \dots, \tilde{X}_N^{(j)})$$

$$\hat{\rho}_m = \frac{1}{k} \sum_{j=1}^k f(h(\tilde{Y}_1^{(j)}, \tilde{Y}_2^{(j)}, \dots, \tilde{Y}_M^{(j)})) \frac{dP_y}{dQ_y}(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)})$$

and

$$\hat{\rho}_o = \frac{1}{k} \sum_{j=1}^k f(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)}) \frac{dP_z}{dQ_z}(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)})$$

as the input, intermediate, and output estimators, respectively, and where the superscript on a random variable indicates which one of k independent simulation runs is under consideration. Each of these estimates are d -dimensional vectors; $\hat{\rho}_i = (\hat{\rho}_{i,1}, \hat{\rho}_{i,2}, \dots, \hat{\rho}_{i,d})$, $\hat{\rho}_m = (\hat{\rho}_{m,1}, \hat{\rho}_{m,2}, \dots, \hat{\rho}_{m,d})$, and $\hat{\rho}_o = (\hat{\rho}_{o,1}, \hat{\rho}_{o,2}, \dots, \hat{\rho}_{o,d})$.

We now state the following fundamental theorem of importance sampling Monte Carlo system simulation:

Theorem 4.3.1.

$$\text{Var}(\hat{\rho}_{i,r}) \geq \text{Var}(\hat{\rho}_{m,r}) \geq \text{Var}(\hat{\rho}_{o,r}) \quad r = 1, 2, \dots, d$$

with equality for the first inequality if and only if

$$\frac{dP_x}{dQ_x}(\tilde{X}_1^{(j)}, \dots, \tilde{X}_N^{(j)}) = s_i(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)})$$

for some function s_i , and with equality for the second inequality if and only if

$$\frac{dP_y}{dQ_y}(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)}) = s_o(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)})$$

for some function s_o .

We first give a simple lemma for the importance sampling weight functions.

Lemma 4.3.1.

$$\frac{dP_y}{dQ_y}(\tilde{Y}) = \mathbb{E}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right].$$

Proof (Lemma 4.3.1). To characterize

$$\mathbb{E}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right],$$

or more precisely,

$$\mathbb{E}_{Q_y}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right],$$

first of all note that, for every bounded (measurable) function $h(y)$, we have

$$\mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})h(\tilde{Y})\right] = \mathbb{E}_{Q_y}\left[\mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right]h(\tilde{Y})\right].$$

On the other hand, we also have

$$\begin{aligned} \mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})h(\tilde{Y})\right] &= \mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})h(g(\tilde{X}))\right] \\ &= \mathbb{E}_{P_x}[h(g(\tilde{X}))] \\ &= \mathbb{E}_{P_y}[h(\tilde{Y})] \\ &= \mathbb{E}_{Q_y}\left[\frac{dP_y}{dQ_y}(\tilde{Y})h(\tilde{Y})\right]. \end{aligned}$$

Thus,

$$\mathbb{E}_{Q_y}\left[\mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right]h(\tilde{Y})\right] = \mathbb{E}_{Q_y}\left[\frac{dP_y}{dQ_y}(\tilde{Y})h(\tilde{Y})\right]$$

for all functions h . The only way that this can occur is

$$\mathbb{E}_{Q_x}\left[\frac{dP_x}{dQ_x}(\tilde{X})|\tilde{Y}\right] = \frac{dP_y}{dQ_y}(\tilde{Y}) \quad Q_y - \text{a.s.}$$

□

Proof (Proof of Theorem 4.3.1). Without loss of generality, we can just consider the relationship between the input and intermediate estimators. Also without loss of generality, we just suppose that $d = 1$, otherwise we could just work with the r th component of the estimators and have the same supposition.

Since the two estimators have the same mean, it suffices to compare the second moments of typical terms. For simplicity we write

$$\tilde{X} = (\tilde{X}_1^{(j)}, \tilde{X}_2^{(j)}, \dots, \tilde{X}_N^{(j)})$$

and

$$\tilde{Y} = (\tilde{Y}_1^{(j)}, \tilde{Y}_2^{(j)}, \dots, \tilde{Y}_M^{(j)}).$$

Thus, the typical term of the input estimator has second moment

$$\mathbb{E}[f(h(g(\tilde{X})))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] = \mathbb{E}[f(h(\tilde{Y}))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] \quad (4.3)$$

$$= \mathbb{E}[f(h(\tilde{Y}))^2 \mathbb{E}[\frac{dP_x}{dQ_x}(\tilde{X})^2 | \tilde{Y}]] \quad (4.4)$$

while the typical term for the intermediate estimator has second moment

$$\mathbb{E}[f(h(\tilde{Y}))^2 \frac{dP_y}{dQ_y}(\tilde{Y})^2] = \mathbb{E}[f(h(\tilde{Y}))^2 \mathbb{E}[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2],$$

where we have used Lemma 4.3.1.

Now observe that by Jensen's inequality,

$$\mathbb{E}[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2 \leq \mathbb{E}[\frac{dP_x}{dQ_x}(\tilde{X})^2 | \tilde{Y}].$$

Hence the general term for the input estimator has greater than equal second moment (and hence greater than or equal variance) than that of the intermediate estimator. We note also that we have equality in the Jensen's inequality if and only if $(dP_x/dQ_x)(\tilde{X})$ conditioned on \tilde{Y} is almost surely a constant (dependent possibly on \tilde{Y}). This is equivalent to $(dP_x/dQ_x)(\tilde{X}) = s(\tilde{Y})$ for some deterministic function s . This completes the proof of the theorem. \square

Remark 4.3.1. In writing (4.4), we appealed to two elementary properties of conditional expectation: Equations (34.6) and (34.4) of [8]. Equation (34.6) requires that the right-hand side of (4.3) be finite. However, if (4.3) is infinite, the theorem is trivial. To use (34.4) further requires that $\mathbb{E}[(dP_x/dQ_x)(\tilde{X})^2] < \infty$. If this is not the case, put $L_n(\cdot) = \min(dP_x/dQ_x(\cdot), n)$, and write

$$\begin{aligned}
& \mathbb{E}[f(h(\tilde{Y}))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] \\
&= \lim_{n \rightarrow \infty} \mathbb{E}[f(h(\tilde{Y}))^2 L_n(\tilde{X})^2] \\
&= \lim_{n \rightarrow \infty} \mathbb{E}[f(h(\tilde{Y}))^2 \mathbb{E}[L_n(\tilde{X})^2 | \tilde{Y}]] \\
&\geq \lim_{n \rightarrow \infty} \mathbb{E}[f(h(\tilde{Y}))^2 \mathbb{E}[L_n(\tilde{X}) | \tilde{Y}]^2] \\
&= \mathbb{E}[f(h(\tilde{Y}))^2 \lim_{n \rightarrow \infty} \mathbb{E}[L_n(\tilde{X}) | \tilde{Y}]^2] \\
&= \mathbb{E}[f(h(\tilde{Y}))^2 \mathbb{E}[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2]
\end{aligned}$$

where we use the monotone convergence theorem along with a conditional dominated convergence theorem [8, Theorem 34.2(v)]. Note also that the quantity

$$\frac{dP_x}{dQ_x}(\tilde{X})$$

is integrable since its expectation is one.

Remark 4.3.2. It is possible that the inequalities be met with equality. For example, consider the case of $h(g(x_1, \dots, x_N)) = \sum_{i=1}^N r(x_i)$, for some arbitrary function $r : \mathcal{R} \rightarrow \mathcal{R}$. We can suppose that the output estimator and the intermediate estimator are the same. Suppose $dP_x(x_1, x_2, \dots, x_N) = \prod_{i=1}^N p(x_i)$, where $p(\cdot)$ is the input probability density (or mass function if we are dealing with discrete random variables). Suppose we choose the biasing distributions to be *exponential shifts*:

$$dQ_{x,\theta}(x_1, \dots, x_N) = \prod_{i=1}^N q_\theta(x_i) = \frac{\prod_{i=1}^N p(x_i) \exp(\theta r(x_i))}{M(\theta)^N},$$

where $M(\theta) = \int p(x) \exp(\theta r(x)) dx$ is the moment generating function of the scalar random variable $r(X)$. Now note that

$$\begin{aligned}
\frac{dP_x}{dQ_x}(\tilde{X}_1, \dots, \tilde{X}_N) &= \frac{\prod_{i=1}^N p(\tilde{X}_i)}{\left(\prod_{i=1}^N p(\tilde{X}_i) \exp(\theta r(\tilde{X}_i)) \right) M(\theta)^{-N}} \\
&= \exp\left(-\theta \sum_{i=1}^N r(\tilde{X}_i)\right) M(\theta)^N \\
&= \exp(-\theta \tilde{Y}) M(\theta)^N \\
&= s_i(\tilde{Y}).
\end{aligned}$$

Thus, in this sum of i.i.d. random variables setting with exponential shift bias distributions, no performance loss is incurred by using the simpler input formulation.

4.4 Conditional Importance Sampling

Suppose we are interested in

$$\rho = \mathbb{E}[f(Z_1, Z_2)],$$

where Z_i is an \mathcal{R}^{n_i} -valued random variable for $i = 1, 2$ and $f : \mathcal{R}^{n_1} \times \mathcal{R}^{n_2} \rightarrow \mathcal{R}^d$. Denote the probability measure on $\mathcal{R}^{n_1} \times \mathcal{R}^{n_2}$ associated with the random variables (Z_1, Z_2) as P . We suppose that we wish to use importance sampling to estimate ρ and thus we have a biasing probability measure on $\mathcal{R}^{n_1} \times \mathcal{R}^{n_2}$ which we denote as Q . The usual importance sampling estimator is given by

$$\hat{\rho}_{IS} = \frac{1}{k} \sum_{j=1}^k f(\tilde{Z}_1^{(j)}, \tilde{Z}_2^{(j)}) \frac{dP}{dQ}(\tilde{Z}_1^{(j)}, \tilde{Z}_2^{(j)}).$$

Now note that by the smoothing property of conditional expectation,

$$\rho = \mathbb{E}_P[f(Z_1, Z_2)] = \mathbb{E}[\mathbb{E}_P[f(Z_1, Z_2)|Z_2]] = \mathbb{E}[g(Z_2)],$$

where we denote $\mathbb{E}_P[f(Z_1, Z_2)|Z_2]$ which is only a function of Z_2 , as $g(Z_2)$. It could very well be in certain situations that this conditional expectation g is known or is easily computable. Of course ρ is just the expectation of the g function. Thus we can use importance sampling to estimate the expectation of the g function. This leads us to the so-called *conditional importance sampling estimate* (also known in certain situations as the g -method),

$$\begin{aligned} \hat{\rho}_g &= \frac{1}{k} \sum_{j=1}^k \mathbb{E}_P[f(Z_1^{(j)}, Z_2^{(j)})|Z_2^{(j)}] \frac{dP}{dQ}(\tilde{Z}_2^{(j)}) \\ &= \frac{1}{k} \sum_{j=1}^k g(Z_2^{(j)}) \frac{dP}{dQ}(\tilde{Z}_2^{(j)}), \end{aligned}$$

where $(dP/dQ)(z_2)$ is just the Radon–Nikodym derivative of the marginal distribution of Z_2 under P with respect to the marginal distribution of Z_2 under Q . For example, it is easy to verify that

$$\mathbb{E}_Q\left[\frac{dP}{dQ}(\tilde{Z}_1, \tilde{Z}_2)|\tilde{Z}_2\right] = \frac{dP}{dQ}(\tilde{Z}_2).$$

The main result here is

Theorem 4.4.1.

$$\text{Var}(\hat{\rho}_{g,i}) \leq \text{Var}(\hat{\rho}_{IS,i}) \quad i = 1, 2, \dots, d.$$

Proof. For simplicity, we just take $d = 1$, otherwise without loss of generality, we can just consider the i th component of the estimator in isolation. As

always $k \text{Var}(\hat{\rho}_g) = F_g - \rho^2$ and of course $k \text{Var}(\hat{\rho}_{IS}) = F_{IS} - \rho^2$. We now have

$$\begin{aligned} F_g &= \int g(z_2)^2 \left(\frac{dP}{dQ}(z_2) \right)^2 dQ(z_2) \\ &= \int g(z_2)^2 \frac{dP}{dQ}(z_2) dP(z_2). \end{aligned}$$

Consider the first term in the integrand above,

$$\begin{aligned} g(z_2)^2 &= \left(\int f(z_1, z_2) dP(z_1|z_2) \right)^2 \\ &= \left(\int f(z_1, z_2) \frac{dP(z_1|z_2)}{dQ(z_1|z_2)} dQ(z_1|z_2) \right)^2 \end{aligned}$$

now applying Schwarz' inequality

$$\begin{aligned} &\leq \left(\int dQ(z_1|z_2) \right) \int f^2(z_1, z_2) \left(\frac{dP(z_1|z_2)}{dQ(z_1|z_2)} \right)^2 dQ(z_1|z_2) \\ &= \int f^2(z_1, z_2) \frac{dP(z_1|z_2)}{dQ(z_1|z_2)} dP(z_1|z_2). \end{aligned}$$

Hence

$$\begin{aligned} F_g &= \int g(z_2)^2 \frac{dP}{dQ}(z_2) dP(z_2) \\ &\leq \int \int f^2(z_1, z_2) \frac{dP(z_1|z_2)}{dQ(z_1|z_2)} dP(z_1|z_2) \frac{dP}{dQ}(z_2) dP(z_2) \\ &= \int \int f^2(z_1, z_2) \frac{dP}{dQ}(z_1, z_2) dP(z_1, z_2) \\ &= \int \int f^2(z_1, z_2) \left(\frac{dP}{dQ}(z_1, z_2) \right)^2 dQ(z_1, z_2) \\ &= F_{IS}. \end{aligned}$$

This completes the proof of the theorem. \square

4.5 Simulation Diagnostics

We typically want our importance sampling estimate $\hat{\rho}$ of a probability ρ to be within x percent accuracy with probability y . Usually this leads us to consider some measure of the *relative precision* of the estimate. Let Z be a standard Gaussian random variable. Denote the two-sided quantile of Z by $P(|Z| \leq t_y) = y$. We control the relative accuracy of our importance sampling estimates by controlling the number of simulation runs \tilde{k} . Recall

that since the importance sampling estimate $\hat{\rho}$ is unbiased, we can write $\tilde{k} \text{Var}(\hat{\rho}) = F - \rho^2$. Thus,

$$\begin{aligned}
y &= P(|\hat{\rho} - \rho| \leq \frac{x}{100}\rho) \\
&\approx P(|Z\sqrt{\text{Var}(\hat{\rho})}| \leq \frac{x}{100}\rho) \\
&= P(|Z\sqrt{\frac{F - \rho^2}{\tilde{k}}}| \leq \frac{x}{100}\rho) \\
&= P(|Z| \leq \frac{x\rho\sqrt{\tilde{k}}}{100\sqrt{F - \rho^2}}) \\
t_y &= \frac{x\rho\sqrt{\tilde{k}}}{100\sqrt{F - \rho^2}} \\
\tilde{k} &= \left(\frac{t_y 100}{x}\right)^2 \left(\frac{F}{\rho^2} - 1\right). \tag{4.5}
\end{aligned}$$

We should always set a desired level of precision and confidence before we begin a simulation. Equation (4.5) requires that we know F and ρ in order to set the number of simulation runs \tilde{k} beforehand. Obviously, we don't have this information and so we must do something else. In fact no procedure in which the run length is fixed before the simulation begins can be relied upon to produce a confidence interval that covers the true value with the desired probability level. In the author's opinion, the only practical solution to this problem is to develop some sort of sequential procedure. In other words we will, as the simulation progresses, use the simulation outputs themselves to decide when we have collected enough data and can stop the simulation.

Suppose that $\rho = P(f(Z_p) \in E)$ where Z_p is an \mathcal{S} -(a complete separable metric space¹)valued random variable (with associated probability measure P) and f is a (measurable) function mapping from S into \mathcal{R}^d .

To implement an importance sampling estimator, we generate an i.i.d. sequence of \mathcal{S} -valued random variables $\{Z_q^{(1)}, Z_q^{(2)}, \dots\}$, with associated probability measure Q . As the simulation progresses, we compute an estimate of F (in addition to the importance sampling estimate of ρ). Thus

$$\begin{aligned}
\hat{\rho}(k) &= \frac{1}{k} \sum_{j=1}^k \frac{dP}{dQ}(Z_q^{(j)}) \mathbf{1}_{\{f(Z_q^{(j)}) \in E\}} \\
\hat{F}(k) &= \frac{1}{k} \sum_{j=1}^k \left[\frac{dP}{dQ}(Z_q^{(j)}) \right]^2 \mathbf{1}_{\{f(Z_q^{(j)}) \in E\}}.
\end{aligned}$$

We can then compute

¹ We have chosen to make the domain of f a complete separable metric space (also called a Polish space). We could allow a more general topological space; all that we really require is that f be a measurable mapping.

$$k^*(k) = \left(\frac{t_y 100}{x} \right)^2 \left(\frac{\hat{F}(k)}{\hat{\rho}(k)^2} - 1 \right). \quad (4.6)$$

Some common useful values for t_y are: $t_{.80} = 1.2816$, $t_{.90} = 1.6440$, $t_{.95} = 1.96$, $t_{.99} = 2.5758$. Common values for x would be 1, 5, 10, or 20. Our criterion for stopping the simulation is as follows.

Sequential Stopping Criterion: Stop after k simulation runs if

$$k \geq k^*(k).$$

Sometimes (usually out of laziness), we wish just to run a simulation for a certain number of times k and look at the output. We can use (4.5) in another way. We might want to know what level of precision we have attained. We can then estimate an x percent level of precision with y percent confidence, by computing

$$x = 100t_y \sqrt{\frac{1}{k} \left(\frac{\hat{F}(k)}{\hat{\rho}(k)^2} - 1 \right)}.$$

Example 4.5.1. Suppose we have the following observation model for an observed sequence of random variables in a signal detection problem

$$r_i = s + N_i \quad i = 1, \dots, n,$$

where $s = -1$ under hypothesis zero (H_0), $s = +1$ under hypothesis one (H_1), and $\{N_i\}$ is an i.i.d. sequence of standard normals under either hypothesis. We process these data by computing

$$R = \frac{1}{n} \sum_{i=1}^n r_i = \frac{1}{n} \sum_{i=1}^n s + N_i = s + \bar{N}.$$

If $R > 0$, we announce H_1 is true; otherwise we announce H_0 is true. The probability of error for this receiver is

$$P(\text{error}) = P\left(\frac{1}{n} \sum_{i=1}^n N_i > 1\right) = P(\bar{N} > 1).$$

Since the N_i are i.i.d. standard normals, \bar{N} is normal mean zero, variance $1/n$. Using tables of the error function, we can easily evaluate this probability for a given n . For example, $n = 24$ gives $P(\text{error}) \approx 4.8 \times 10^{-7}$.

The importance sampling method we choose to simulate this system is mean shifting. Instead of directly simulating the standard normal noise samples $\{N_i\}$, we use the mean shifted $\{\tilde{N}_i\}$ (taken to be mean one, variance

one) random variables in an input formulation. For an output formulation we have $p_{\bar{N}}$ is normal mean zero, variance $1/n$ and $q_{\bar{N}}$ is normal mean one, variance $1/n$. Both formulations will give the same variance (the mean shift is also an exponential shift in the Gaussian setting); we use the simpler (to simulate) output formulation.

We want to investigate a bit our results for k^* in the setting of this very simple system (where we can calculate everything in closed form). We compute $\hat{\rho}_n$ and \hat{F}_n as the simulation progresses to determine the number of simulation runs we need to achieve x percent accuracy with probability y . For this example, we use $y = .9$ and x taking on the range of values $\{2.5, 5, 10, 20, 40\}$. We find that as our desired accuracy x decreases, our empirical probability \hat{y} of achieving accuracy x increases to its expected value $y = .9$ using the k^* stopping criterion. The following table shows the results of this experiment for 1000 trials.

Accuracy (x)	Achieved Probability (\hat{y})
40	.840
20	.873
10	.887
5	.893
2.5	.909

Table 4.1. Simulation results

4.6 Notes and Comments

The importance sampling idea, that of focusing on the region(s) of importance so as to save computational resources, evidently springs from a 1956 paper due to A. Marshall [59]. Importance sampling is but one of a variety of variance reduction techniques known to simulation practitioners. A very readable introduction to the subject of variance reduction in simulation is found in [67].

The notion of input versus output estimators and bias point selection was first posed (at least in the engineering literature) by P. Hahn and M. Jeruchim [37]. The basic theorem is from [14].

Conditional importance sampling estimators first appear (in the engineering literature) under the name of the g -method of R. Srinivasan [78]. The basic theorem given in the text is a generalization of his result for i.i.d. sums.

A rigorous (asymptotic) analysis of the sequential stopping rule given in the section on simulation diagnostics can be found in [62].



<http://www.springer.com/978-0-387-20078-1>

Introduction to Rare Event Simulation

Bucklew, J.

2004, XII, 268 p., Hardcover

ISBN: 978-0-387-20078-1