ADVANCES IN IMPORTANCE SAMPLING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Timothy Classen Hesterberg
August 1988.

This version created June 29, 2003.

0.1 Abstract

Importance sampling in Monte Carlo simulation is the process of estimating a distribution using observations from a different distribution. Estimates are computed using weights that are roughly proportional to the likelihood ratio between the two distributions.

Importance sampling has been very successful as a variance reduction technique in rare event applications. It can also be applied in many other applications, as a variance reduction technique, as a means of solving a problem that is otherwise intractable, or for analyzing the performance of an estimate or a physical process under multiple input distributions using a single set of observations, as in response surface estimation or in the analysis of robust estimates.

The classical importance sampling estimate is well-suited for variance reduction in rare event applications. It fails in many other applications. The ratio and regression estimates, well-known in sampling theory, succeed in many of these cases. The latter estimates are weighted averages with weights that sum to one, while the classical estimate has weights that do not sum to one.

Accurate estimates require a reasonable sampling distribution[1] Mixture distributions are an easy way to build sampling distributions. The use of the true distribution as a component in a mixture bounds the likelihood ratio and the weights, and solves a common problem.

Sometimes good results can be obtained from a poor sampling distribution by replacing likelihood ratio weights with their expected value conditioned on a sufficient statistic.

The intent of this work is to develop methods that can be used in many applications. They do not improve on the incredible variance reductions achievable in simple applications using the classical estimate, but do allow importance sampling to be safely applied in a wider variety of applications.

[1] In later work I use the terms "design distribution" in place of "sampling distribution", and "defensive mixture distribution" in place of "mixture distribution"; see the Preface.

Contents

0.2 Acknowledgements

> The world seldom notices who teachers are; but civilization depends on what they do—Lindley Stiles

I wish to thank some of the fine teachers I have worked with, in high school and at St. Olaf College, who with their enthusiasm and creativity kindled my enthusiasm for mathematics and statistics.

I thank the faculty and fellow students at Stanford for their teaching and support, in particular Bradley Efron for his encouragement and helpful discussions in my study of bootstrapping and importance sampling, and to Vernon Johns for discussions about importance sampling.

Further thanks are due my former colleagues at Pacific Gas & Electric Company both for their encouragement (in particular to Eugene Alward for his efforts to get me home promptly to work on this) and for helpful discussions. It was Paul Gribik who asked the fateful question about importance sampling that lead to this dissertation—"You mean you normalize the weights to sum to one?" (Of course, doesn't everyone? ... No, they don't!) The Fuel Inventory Probabilistic Simulator described here is joint work with Paul, Betty Look, and Rana Glascal.

Finally, I thank my wife Bev for her patience and support during the research and writing of this work.

List of Tables

List of Figures

0.3 Preface to the 2003 Revision

This version of *Advances in Importance Sampling* was created to be placed on the web. The original version of the dissertation, created in 1988, was written in Microsoft Word. Unfortunately, current versions of Word will no longer handle the equations from the earlier versions. This version was converted to LaTeX, in order to produce higher-quality formatting.

I have resisted the temptation to make substantial changes in content or terminology from the original version. I would like to note four points where the terminology used in the original dissertation differs from later usage:

- In this document I referred to the distribution from which observations are generated when importance sampling as the "sampling distribution"; however, that term has another meanings, for the distribution of a statistic calculated from a random sample. In later work I call the distribution from which observations are generated the "design distribution".

- In later work I call the distributions for which one would like estimates "target distributions" instead of "true distributions" or "object distributions".

- In Chapter 6 I use the term "mixture sampling" for the use of a design distribution in which one component is f. In later work I call that a "defensive mixture distribution" (or "defensive mixture sampling").

- In a number of chapters I use the term "response surface analysis," for analyzing the relationship between parameters of distributions (as input) and expected values (as output). In later work I called this "sensitivity analysis".

I have corrected a number of errors in the original version; the most important of these are described in Section 0.4.

The computer files containing the appendices from the original dissertation are lost. Please contact me if you would like a hard copy of the appendices.

Some of this work has been published, in particular in Hesterberg, Tim C. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions", *Technometrics*, **37**(2), 185–194. For additional references see my home page, currently at `www.insightful.com/Hesterberg`.

0.4 Errata in the Original Dissertation

Errata in *Advances in Importance Sampling*, dissertation by Tim Hesterberg, Stanford University, Department of Statistics, 1988.

Page numbers refer to the original document. Errors are generally corrected in the 2003 version.

Page 12, there are two equations numbered "(1.11)". (In the 2003 version the second is not given a number).

Page 13–16, there are two equations each numbered "(1.13)", "(1.14)", "(1.15)", "(1.16)". (In the 2003 version the second in each case is not given a number. Only the first equation labeled (1.16) was referred to elsewhere in the text.)

Page 23, line 4, replace $\hat{\mu}_{\text{int},2}$ with $\hat{\mu}_{\text{int},2} + c$.

Page 30, inconsistent usage of u or v (or, $\mathbf{u}$ or $\mathbf{v}$).

Page 39–43, there are two sets of equations each numbered "(2.60)", "(2.61)", "(2.62)", "(2.63)", "(2.64)". (In the 2003 version the first in each case is not given a number. The second set of equations were referred to elsewhere in the text, the first was not.)

Page 43, formulae 2.66–2.68 should not have squares, e.g. replace $\hat{\sigma}^2_{\text{int}}$ with $\hat{\sigma}_{\text{int}}$.

Page 96, equation (4.35) should not have $\theta(X_i)$ in the denominator.

Page 98, equations (4.37) and (4.38) should have $L(\tau|X)$ in the denominator instead of $L(\theta|X)$.

Page 105–106, equations (4.58) and (4.59) are missing closing braces.

Page 138, equation (6.5), change "$x >$" to "$y >$".

Page 173, line 3 change $f(X)$ to $g(X)$.

Page 243, reference should begin "Luzar", not "Luzer".

Page 244, in Stewart reference, misspelled "integration".

Chapter 1

Introduction

Importance sampling in Monte Carlo simulation is the process of estimating something about a distribution using observations from a different distribution.

When X is a random variable with distribution f and $\theta(X)$ is a function of x, a simple Monte Carlo estimate of the distribution of $\theta(X)$ is obtained by generating $X_i \sim f$, $i = 1, 2, \ldots, n$, and computing $\theta_i = \theta(X_i)$. The observed values θ_i are used as an estimate of the distribution of $\theta(X)$, with equal weights. Under importance sampling each replication is drawn instead from a "sampling distribution", $X_i \sim g$, and the observed values $\theta(X_i)$, together with weights that compensate for the biased sampling method, form the distribution estimate.

The use of a sampling distribution in place of the "true" distribution may be intentional, with the sampling distribution chosen with an eye to improving the efficiency of the simulation. This is the traditional realm of importance sampling, and efficiency gains of many orders of magnitude are possible. Importance sampling is particularly valuable because the largest gains are registered in some of the most difficult simulation applications, those involving the simulation of rare events. A sampling distribution chosen to increase the frequency of rare events reduces the number of replications required to observe an adequate number of such events.

The earliest example of the use of importance sampling for rare event applications is in the estimation of probabilities of nuclear particles penetrating shields (Kahn 1950, Kahn and Marshal 1953, Booth 1986, Murthy and Indira 1986, etc.). Other areas of application include reliability estimation in the fields of digital communications (Davis 1987, Hahn & Jeruchim 1987), fault-

tolerant computers (Conway and Goyal 1987, Goyal et al. 1987, Kioussis and Miller 1983), engineering analysis, the simulation of stochastic processes (Moy 1986), study of sequential tests (Siegmund 1976), and implementation of bootstrap confidence intervals (Johns 1987).

Most successful rare-event applications have involved estimating the expected value of a distribution with a large discrete mode at zero. Hesterberg (1987) used importance sampling in a fuel inventory simulation that did not have a discrete mode at zero, and obtained moderate variance reduction using methods that will be described in this work.

Importance sampling may be done because there is no other choice, because it is impossible or impractical to generate samples from the true distribution. This includes applications in Bayesian analysis, where distribution estimation must be performed with respect to an intractable posterior distribution (Stewart 1976, 1979, 1983, Kloek and van Dijk 1978, van Dijk and Kloek 1983), and the analysis of characteristic roots of a random covariance matrix (Luzar and Olkin, 1988). Importance sampling can thus be used to solve problems that could not otherwise be solved.

Importance sampling is also valuable in applications with more than one "true" distribution. Rather than running a separate simulation for each distribution, importance sampling can be used to estimate results from many distributions in a single simulation (Beckman & McKay 1987, Tukey 1987). The number of such distributions may be infinite, as the bootstrap tilting interval (Tibshirani 1984), or in response surface estimation (Glynn & Iglehart 1987). Importance sampling can also be used for derivatives of expectations with respect to parameters of input distributions (Reiman & Weiss 1986, Glynn 1986).

In any importance sampling application a sample is obtained from the sampling distribution, but an estimate based on the true distribution is required. This estimate is made using the empirical distribution, but with different weights on each observation, which are chosen to "unbias" the results, to counteract the bias introduced by using the sampling distribution in place of the true distribution. This is done by choosing weights roughly proportional to the "weight function" (also "inverse likelihood ratio")

$$W(x) := f(x)/g(x). \tag{1.1}$$

The weight function is inversely proportional to the relative likelihood of outcomes under g compared to f, so that if a given x is half as likely to be observed under g as under f, it is given twice the weight when it is observed.

Still, within the constraint of "rough proportionality" different weights are possible. Different philosophies of importance sampling lead to different choices of weights, and different estimates.

The two broad interpretations of importance sampling are the integration and sampling interpretations. The traditional view (e.g. Hammersley and Hanscomb 1964) is that Monte Carlo simulation is a form of *integration*. To estimate the expectation of $\theta(X)$, note that

$$\begin{aligned} E_f(\theta(X)) &= \int \theta(X) f(x) dx \\ &= \int \theta(X) \frac{f(x)}{g(x)} g(x) dx \end{aligned} \tag{1.2}$$

if $g(x) > 0$ when $f(x) > 0$. Write $Y(x) := \theta(X) f(x)/g(x)$, then

$$\mu := E_f(\theta(X)) = E_g(Y(X)), \tag{1.3}$$

and μ can be estimated using a sample average $\bar{Y}$ of values from g. Importance sampling done in this way involves both a modified sampling scheme and an induced transformation from θ to Y. If Y has smaller variance (under g) than does θ (under f), the estimate is more efficient. The *integration estimate* of an expectation is

$$\hat{\mu}_{\text{int}} := n^{-1} \sum_{i=1}^{n} Y(X_i) = \bar{Y} \tag{1.4}$$

The *sampling* interpretation is that importance sampling is a sampling strategy for distribution function estimates, from which expectations and other quantities can be computed. In this interpretation importance sampling involves concentrating the sampling effort on important regions of the sampling space (in the same way that strata sizes are chosen in stratified sampling).

The sampling approach requires that distributions have total probability equal to one. Note that the integration estimate can be written as a weighted average

$$\hat{\mu}_{\text{int}} = \sum_{i=1}^{n} V_{\text{int}}(X_i) \theta(X_i) \tag{1.5}$$

where

$$V_{\text{int}}(X_i) = \frac{W(X_i)}{n} \tag{1.6}$$

is the weight assigned to observation i. This is a weighted average of observed values, with weights that do not in general add to one, though the expected value is one:

$$E_g(W) = \int W(x)g(x)dx = \int f(x)dx = 1. \tag{1.7}$$

The sampling approach requires that weights sum to one. The simplest way to do this is to use weights

$$V_{\text{ratio}}(X_i) := \frac{W(X_i)}{\sum W(X_j)}. \tag{1.8}$$

This gives the *ratio estimate*:

$$\hat{\mu}_{\text{ratio}} := n^{-1} \sum_{i=1}^{n} V_{\text{ratio}}(X_i)\theta(X_i) = \frac{\bar{Y}}{\bar{W}} \tag{1.9}$$

This is equivalent to the ratio estimate (Cochran 1977) for estimating an expectation (of Y) in the presence of a covariate (W) with known expectation.

The transformation approach and integration estimate give large efficiency improvements integrals and expected values in some applications, particularly for distributions with a large discrete mode at zero; this includes estimation of small probabilities. On the other hand, the integration estimate is not equivariant, which leads to a number of unexpected consequences, and should be avoided in many applications. Particular difficulties occur in simulations with more than one output quantity, since the same multiplicative transformation must be used for all output quantities.

The sampling approach is more general. The ratio estimate produces a distribution estimate consisting of points and weights (the integration estimate estimates expectations only), and preserves physical meaning. This estimate can be used in some applications that fall outside the traditional variance-reduction realm of the integration estimate. Furthermore, this estimate is amenable to sampling strategies that guarantee robustness (relative to not using importance sampling) for any output quantity. However, the ratio estimate is less useful as a pure variance reduction technique, with a nonzero lower bound on the possible variance reduction.

Fortunately, extensions of both methods converge to produce the same results. These estimates give the best of both worlds—distribution estimates with unit mass, equivariance, and estimates of expectations which have smaller variance than both the basic integration and sampling estimates. The most widely applicable improved estimate is the *regression estimate*.

For any of these estimates good results depend on the sampling distribution. Ad-hoc techniques are used in many examples, though a number of general techniques are available.

The most widely applicable technique is known variously as exponential tilting or exponential biasing (Kahn 1950, Clark 1966, Murthy and Indira 1986, Johns 1987), and consists of choosing sampling distributions from embedded exponential families (Siegmund 1976). We discuss reasons for the success of this method—in particular, exponential tilting is the only method (within a certain class of methods) of generating sampling distributions with weights which are a function only of an approximate sufficient statistic. This helps minimize an undesirable side effect of importance sampling, which occurs when the weights on observations have large variances.

We offer a new perspectives on the use of mixture distributions (Butler 1956, Marsaglia 1961) in the specification and generation of sampling distributions. We use the term *mixture sampling* for the use of a mixture distributions which includes the true distribution as one of the components. In applications where importance sampling is optional (where the true distribution could be used as a sampling distribution) mixture sampling bounds the weight function, and can be used to give any desired degree of robustness to the results (except for the integration estimate). This solves a problem noted by Bratley, Fox and Schrage (1983) and others of the difficulty of preventing large weights in multivariate applications. Further improvement is achieved by stratifying the mixing proportions.

Mixture distributions have more uses than in mixture sampling. Van Dijk and Kloek (1985) use mixture distributions in an attempt to build sampling distributions which closely match the posterior distribution in an example in Bayesian analysis. Other uses are in applications with multiple output quantities, where mixture distributions can be used to estimate all quantities well simultaneously, as long as an equivariant estimate is used.

Sampling distributions may be specified either in terms of the distribution of the input variable X, or by modifying the distribution of the (nominally) uniform random numbers U that are used to generate X. Moy (1965) considers the use of a parametric family of transformations of U in the simulation of stochastic processes. Booth (1986) uses an adaptive scheme, based on a mixture of the U and the X values, in nuclear particle transport calculations. We relate Moy's results to the theoretical properties of exponential tilting, and propose a general scheme for specifying sampling distributions in the U space.

Importance sampling is subject to some finite sample-size problems. In any sample of a finite number of replications, the observed distribution may not be perfectly characteristic of the true distribution. If the unobserved part of the sampling space contains a disproportionate share of extreme results, then estimates of the quality of results are too optimistic. Furthermore, if the extreme values are not symmetrically distributed about the overall mean, the sample average will be conditionally biased. This is a general problem in statistical sampling, but it is particularly troublesome in importance sampling because extreme results are defined not just in terms of the model output $\theta(X)$, but in terms of the product of $W(X)$ and $\theta(X)$. A sampling distribution which almost ignores some regions practically guarantees that those regions will not be observed in a reasonable number of replications, and that is exactly where W is large. We offer partial solutions to this problem—the use of mixture sampling, modified confidence intervals, and diagnostics based on the average value of the weight function.

There are three general parts to this paper: estimation methods, examples, and sampling methods, though this trichotomy is not strict. Chapter 2 discusses estimates, including the integration, ratio, and regression estimates. Chapter 3 covers an estimation technique useful in some applications, the replacement of weights with their conditional expected values. Chapter 4 considers four classes of applications where importance sampling is useful, including the traditional variance reduction application with and without a zero mode distribution and two newer areas, solving intractable applications and simultaneous estimation for multiple distributions. Chapters 5 and 6 discuss sampling strategies and methods.

1.1 Definitions

In this section we list a number of definitions which used throughout this work.

X is an input random variable. This may be real-valued, vector-valued, a stochastic process, or general. X has probability measure f without importance sampling, or is generated according to g when importance sampling is used.

$\mathcal{X}$ "Sample space," domain of X.

$\theta = \theta(X)$ is an output random variable and a function of X. This is usually real- or vector-valued.

$f = f(X)$ probability measure (on X) for the "true" distribution of X (without importance sampling. Unless specifically noted, results do not require that f be a density.

$g = g(X)$ probability measure (on X for the "sampling" distribution of X, under importance sampling. Unless specifically noted, results do not require that g be a density.

"dominate" In general the sampling distribution g must dominate f, i.e. $g > 0$ when $f > 0$.

"weak dominance" Sometimes g need dominate only $f|\theta|$, i.e. $g > 0$ when $f|\theta| > 0$.

$W = W(x)$ Radon-Nikodym derivative of f with respect to g. For densities, $W(x) = f(x)/g(x)$. This is also called a "weight function", or "inverse likelihood ratio".

$R = R(x)$ "Likelihood ratio" $g(x)/f(x)$. R need not be finite.

$Y = Y(x)$ $= \theta(x)W(x)$; the expected value of Y under g is the desired expected value μ.

μ Expected value of θ under f, and of Y under g. When $\theta \in \mathcal{R}^d$ the expectation is computed separately for each component, and $\mu \in \mathcal{R}^d$.

$\hat{\mu}_{\text{est}}$ Estimate of μ obtained using estimate "est" (different estimates are considered).

σ^2 Variance; if for d-dimensional quantity then σ^2 is a vector of length d containing variances for each component.

i Index of replication i in a Monte Carlo experiment.

n Number of replications in a Monte Carlo experiment.

X_i, θ_i, Y_i, W_i Values of the corresponding functions in the ith replication of a Monte Carlo experiment.

$\bar{Y}$, $\bar{W}$ Sample averages, i.e. $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

V_i "Weight" assigned to replication i for use in the analysis of a Monte Carlo experiment. Sometimes $V_i \approx W(X_i)/n$. Formally, $V_i = V(X_i; X_1, \ldots X_n)$.

$V_{i,\text{est}}$ Weight for replication i for estimate "est", $\hat{\mu}_{\text{est}} = \sum V_{i,\text{est}}\theta_i$.

$\pi_{i,\text{est}}$ "Metaweight," or "weight on the weighted observation." $\pi_{i,\text{est}}W_i = V_{i,\text{est}}$.

d Dimension of X, or of θ, depending on the context.

j Index of component j of a vector.

X_j jth component of X for vector-valued X.

X_{ij} jth component of X_i (replication i).

$I(A)$ Indicator function, $I(A) = 1$ if A is true, else $I(A) = 0$.

δ_x Distribution with a point mass on x.

$\Omega = \Omega(X)$ Expected value of $W(X)$ given $S(X)$, where S is a sufficient statistic for θ. Ω is a "conditional weight."

Φ Distribution function for a standard normal (Gaussian) distribution.

"lardimaz" "Large discrete mode at zero." A distribution which has a high probability (> 0.9 or more) of being equal to zero.

"relative efficiency" Variance of an estimate divided by the variance that would be obtained using simple random sampling.

"AVar" Asymptotic variance, n times the (highest order term in the) variance of an estimate from a sample of size n.

β Regression slope in a regression equation.

1.2 Monte Carlo Simulation

Suppose that X is a random variable with probability measure f, and that $\theta(x)$ is a random variable which depends on x. Then the distribution of $\theta(X)$ can be estimated using Monte Carlo simulation. Generate values $X_i \sim f$, $i = 1, 2, \ldots n$, and compute $\theta_i = \theta(X_i)$ for each i. The empirical distribution

formed by placing weight $1/n$ on each of the values θ_i is an estimate of the distribution of $\theta(X)$.

This is a very general method. Both X and θ can be multivariate quantities, and the dimension of neither need be fixed. It is not even necessary that the distribution F be known, only that it be possible to generate pseudo-random values according to the distribution. In the simulation of a discrete Markov chain, for example, it is not necessary to enumerate sample paths and probabilities, only to be able to generate them.

Consider a simple example. In Example 1.1, X is a Gaussian random variable with mean 0.2 and variance 1, and $\theta(x) = \arctan(x)$. The goal is to characterize the distribution of X—its mean, variance, quantiles, and density. We do this using a Monte Carlo simulation, with $n = 10,000$ replications.

Example 1.1 Distribution of $\arctan(N(0.2, 1))$
$X \sim N(0.2, 1)$, $\theta(X) = \arctan(X)$, $n = 10,000$. Estimate $E(\theta)$, $\mathrm{Var}(\theta)$, the 5%, 10%, 25%, 50%, 75%, 90%, and 95% quantiles of θ, and the density of θ. This is done using Monte Carlo simulation, using the algorithm:

1. Do $i = 1$ to $10,000$
   (a) Generate $X_i \sim N(0.2, 1)$
   (b) Compute $\theta_i = \theta(X_i) = \arctan(X_i)$
2. Compute results (mean, variance, percentiles and estimated density) using the values θ_i.

Results are given in Table 1.1, and Figure 1.1 shows the estimated density function.

In contrast to most numerical analysis algorithms, results obtained using Monte Carlo simulation are random (we allow ourselves this illusion; in fact the "random" numbers that feed the simulation are the output of a deterministic algorithm that only appears random) . Proper analysis of a Monte Carlo simulation involves reference to statistical results.

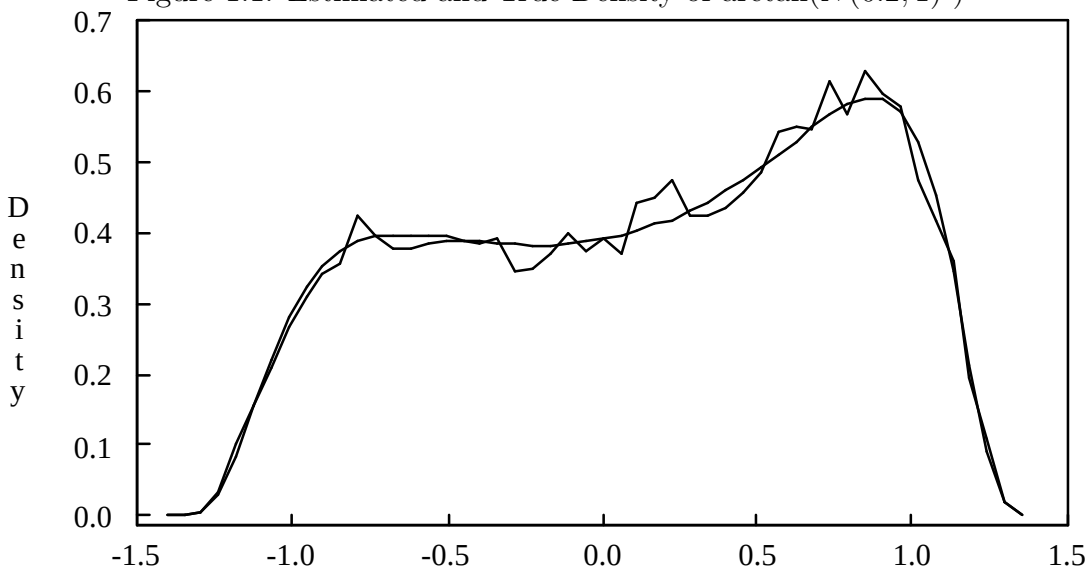
In simple Monte Carlo simulation the values θ_i are independent and identically distributed according to the distribution f_θ. Then standard statistical results apply. These include the strong law of large numbers and the central limit theorem, which describe the convergence of a sample mean to a population expected value. Write

$$\mu := E_f(\theta(X)) = E_{f_\theta}(\theta) \tag{1.10}$$

Table 1.1: Distribution of arctan($N(0.2, 1)$)

Statistic	Estimate	Standard Error	True Value
mean	0.135	(0.007)	0.131
variance	0.441	(0.004)	0.443
5%tile	-0.967	(0.007)	-0.965
10%tile	-0.818	(0.008)	-0.825
25%tile	-0.437	(0.011)	-0.443
50%tile	0.207	(0.011)	0.197
75%tile	0.721	(0.007)	0.718
90%tile	0.973	(0.005)	0.977
95%tile	1.072	(0.005)	1.074

Figure 1.1: Estimated and True Density of arctan($N(0.2, 1)$)



for the expected value of θ (if θ is vector-valued then μ is as well). The Monte Carlo estimate of μ is

$$\hat{\mu}_{\text{mc}} = n^{-1} \sum_{i=1}^{n} \theta_i = \bar{\theta} \tag{1.11}$$

Then the strong law of large numbers states that if θ_i has finite expectation, then $\hat{\mu}_{\text{mc}}$ converges to μ with probability one. If $E(|\theta|) < \infty$, then (S.L.L.N)

$$\lim_{n\to\infty} \hat{\mu}_{\text{mc}} \stackrel{a.s.}{=} \mu.$$

(*a.s.* indicates *almost surely*, i.e. with probability 1).

If in addition θ_i has finite variance, then by the Central Limit Theorem the distribution of $\hat{\mu}_{\text{mc}}$ is asymptotically normal with a standard deviation of order $O(n^{-1/2})$. If $E(\theta^2) < \infty$, then (C.L.T.)

$$\lim_{n\to\infty} \sqrt{n}(\hat{\mu}_{\text{mc}} - \mu) \sim N(0, \sigma_\theta^2). \tag{1.12}$$

(If θ is vector-valued this holds for each component.)

The central limit theorem says that a sample average converges to the population expectation at the rate $O(n^{-1/2})$. Note that the rate of convergence is independent of the dimension or structure of the application, except through the constant σ_θ. Monte Carlo simulation is particularly competitive with other numerical analysis techniques in large, complicated applications.

Still, the convergence rate is not fast, as computer algorithms go. In practical terms, it means that an extra digit of accuracy on a result requires 100 times as many replications. Speeding up the rate of convergence is a motivating factor for the development of many "variance reduction techniques," including importance sampling. Other techniques include antithetic variates, conditional expectation, stratified sampling, control functions, and stratified sampling.

Of course sample averages are not the only measure of the distribution of θ, and other statistics can be computed as well. For the most part, however, our comparison of the performance of importance sampling methods is based on their efficiency in estimating expected values using sample averages. This is reasonably general, because most simulation outputs can be expressed as expectations. Probability estimates, for example, are estimates of the expectation of an indicator function, and variance and quantile estimates can usually be approximated by sample averages of appropriate functions.

The use of sample averages is made more general by augmenting an output quantity with functions of that output. For example, to estimate both $E(\theta)$

and $P(\theta \le q)$ for real-valued θ and fixed $q \in \mathcal{R}$, define a vector $\theta^{(2)} := (\theta, I(\theta \le q))$. This augmented version of θ has component expected values equal to the two quantities to be estimated.

The restriction to sample averages provides a convenient measure for comparing the accuracy of different methods, in terms of "variance reduction". The variance of the simple estimate is σ_θ^2/n; if the use of variance reduction techniques reduces the variance to σ^2/n, with $\sigma^2 < \sigma_\theta^2$, then a variance reduction of

$$\text{variance reduction} = (1 - \sigma^2/\sigma_\theta^2) \cdot 100\% \tag{1.13}$$

has been achieved.

For a given number of replications the variance of the estimate is reduced by this percentage; alternately, the number of replications required to obtain the same quality estimate has been reduced by that percentage. An alternate scale for measuring variance reduction "efficiency", defined here as

$$\text{efficiency} := \sigma^2/\sigma_\theta^2. \tag{1.14}$$

An efficiency of 0 is perfect, 1 indicates no improvement, and > 1 indicates that a method performs worse than simple Monte Carlo sampling.

When σ_θ^2 is known, the central limit theorem can be inverted to obtain a approximate $(1 - 2\alpha) \cdot 100\%$ confidence interval for μ,

$$(\hat{\mu}_{\text{mc}} - z_\alpha \sigma_\theta,\ \hat{\mu}_{\text{mc}} + z_\alpha \sigma_\theta), \tag{1.15}$$

where z_α is the $1 - \alpha$ percentile of a Gaussian distribution. The constant σ_θ^2 can be estimated using the sample standard deviation

$$S_\theta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta_i - \bar{\theta})^2. \tag{1.16}$$

The appropriate interval is then $(\hat{\mu}_{\text{mc}} - t_{\alpha,n-1}\sigma_\theta,\ \hat{\mu}_{\text{mc}} + t_{\alpha,n-1}\sigma_\theta)$ where $t_{\alpha,n-1}$ is the $1 - \alpha$ percentile of a t distribution with $n - 1$ degrees of freedom.

The intervals are based on asymptotic considerations. We will see, however, that these asymptotic intervals are inadequate in many applications with finite sample sizes, because the joint behavior of a sample average and sample standard deviation in applications with extreme values is not adequately approximated by the normal approximations. We discuss improved intervals in Chapter 2.

The standard error estimates listed in Table 1.1 are based on the asymptotic formulas. For the mean the standard error estimate is $S_\theta/\sqrt{n}$, with S_θ defined as in (1.16). For the variance the estimate is $S_V/\sqrt{n}$, where $V_i := (\theta_i - \bar{\theta})^2$; this is based on approximating μ by $\bar{\theta}$ in the formula $\mathrm{Var}(\theta) = E((\theta - \mu)^2)$. The standard error estimates listed for quantile estimates $\hat{Q}_\alpha$ of the α quantile of the distribution of θ are $\frac{\alpha(1-\alpha)}{\sqrt{n}\hat{f}_\theta(\hat{Q}_\alpha)}$, where $\hat{f}_\theta$ is the estimated density of θ.

Confidence intervals for summary statistics other than an average are more difficult to obtain. A variety of statistical methods can be used for standard errors and confidence intervals of medians, variance estimates, standard deviation estimates, and density estimates, among others. Such methods include nonparametric methods for confidence intervals for a median, Taylor series (delta method) approximations to functions of expected values, and many more. To describe all these methods is beyond the scope of this work. We describe here only one method, which fits well into a discussion of simulation.

We can in fact use more simulation to estimate the accuracy of results of a simulation! After generating n values θ_i, select B samples of size n with replacement from the set of θ_i values. To estimate the standard error of the median estimated from the original set of θ_i, for example, compute the standard deviation of the medians from the B samples.

The method outlined so quickly in the previous paragraph is an example of the bootstrap method, implemented using Monte Carlo simulation.

1.3 Bootstrap

Given a sample of data from an unknown distribution,

$$Z_1, Z_2, \ldots Z_d \overset{i.i.d.}{\sim} F,$$

and a statistic $T(Z_1, Z_2, \ldots Z_d)$, how can the distribution of T be estimated? What are the mean and variance of T? If T is estimating something, how good is the estimate?

Suppose that T is a functional statistic (a statistic defined solely in terms of a distribution), so that $T(Z_1, Z_2, \ldots Z_d) = T(\hat{F})$ (we use the two forms of notation interchangeably), and that $T(\hat{F})$ is an estimate of $T(F)$. $\hat{F}$ is the empirical distribution function, formed by placing weight $1/d$ on each of the d points $Z_1, Z_2, \ldots Z_d$.

The questions asked in the first paragraph can now be rephrased. How good is $T(\hat{F})$ as an estimate of $T(F)$? Is $T(\hat{F})$ biased, and what is its variance? Given the sample, what is an appropriate confidence interval for $T(F)$?. These are common questions in statistical analysis. Answers are often obtained by assuming that the unknown distribution is a member of some parametric family, e.g. the family of normal distributions. This allows analytical answers or approximations to be used. Parameters of the distribution may be estimated from the data, using e.g. maximum likelihood estimation or the method of moments.

The bootstrap (Efron 1982) avoids making such assumptions about the unknown underlying distribution. Instead the bootstrap approximates the unknown distribution using the known empirical distribution $\hat{F}$, and estimates the behavior of $T(Z)$ under F by the behavior of $T(X)$ under $\hat{F}$:

$$L_F(T(Z)) \approx L_{\hat{F}}(T(X))$$

$Z = (Z_1, Z_2, \ldots Z_d)$ is the original sample, and X is a sample drawn with replacement from Z:

$$X_1, X_2, \ldots X_d \overset{i.i.d.}{\sim} \hat{F}.$$

The bias and variance of $T(\hat{F})$, as an estimate of T, are:

$$\text{bias}_F(T) = E_F(T(Z) - T(F))$$

$$\sigma^2_F(T) = E_F((T(Z) - E_F(T(Z)))^2) \tag{1.17}$$

The bootstrap estimates of these quantities, using the corresponding empirical distribution, are:

$$\widehat{\text{bias}}_{\text{boot}} = E_{\hat{F}}(T(X) - T(\hat{F})). \tag{1.18}$$

$$\hat{\sigma}^2_{\text{boot}} = E_{\hat{F}}((T(X) - E_{\hat{F}}(T(X)))^2). \tag{1.19}$$

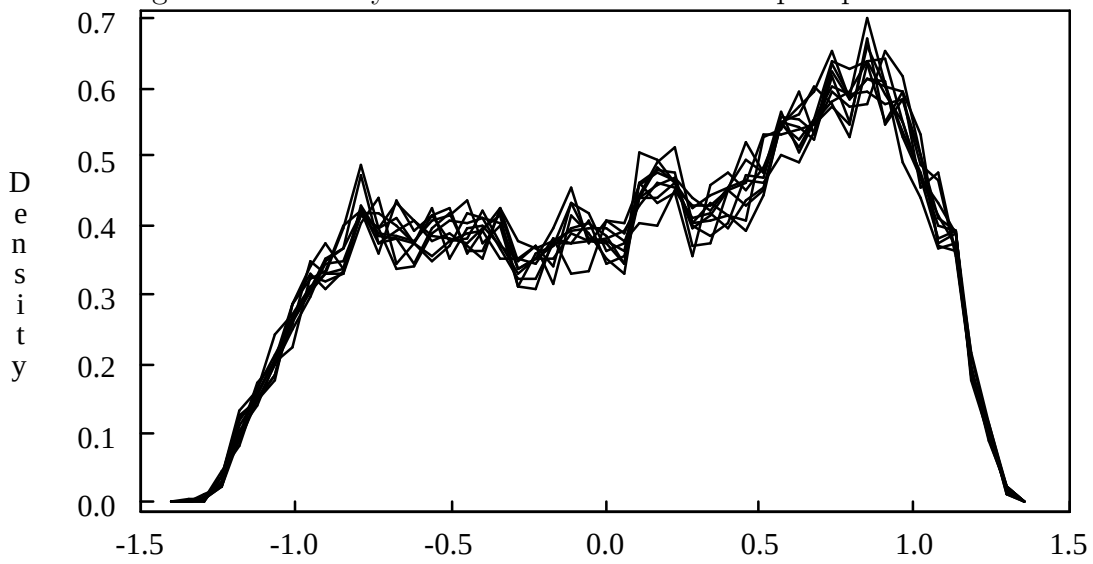
In some cases evaluation of the distribution under $\hat{F}$ of $T(X)$ can be obtained using analytical results, or in very small applications by enumerating all values that X can take on. Most often, however, the implementation of the bootstrap procedure involves Monte Carlo simulation, where samples $X_i, i = 1, 2, \ldots n$, are drawn independently with replacement from $\hat{F}$, and $T_i = T(X_i)$ is computed for each sample.

Example 1.2 Bootstrap Analysis of a Monte Carlo experiment

We return to the example described in Section 1.2, the analysis of the distribution of $\arctan(Z)$, when $Z \sim N(0.2, 1)$. A Monte Carlo experiment yielded 10,000 values of Z. Standard methods give estimates of the standard errors of the estimates of mean, variance, and quantiles. We use the bootstrap to estimate the standard error of the density estimates, by resampling from the Monte Carlo results.

The estimated densities from the first 10 bootstrap replications are shown in Figure 1.2. The true and original estimated densities are given in Figure 1.3, together with a confidence band of $\pm$ two standard errors, estimated using the standard deviations of the bootstrap distribution at each point.

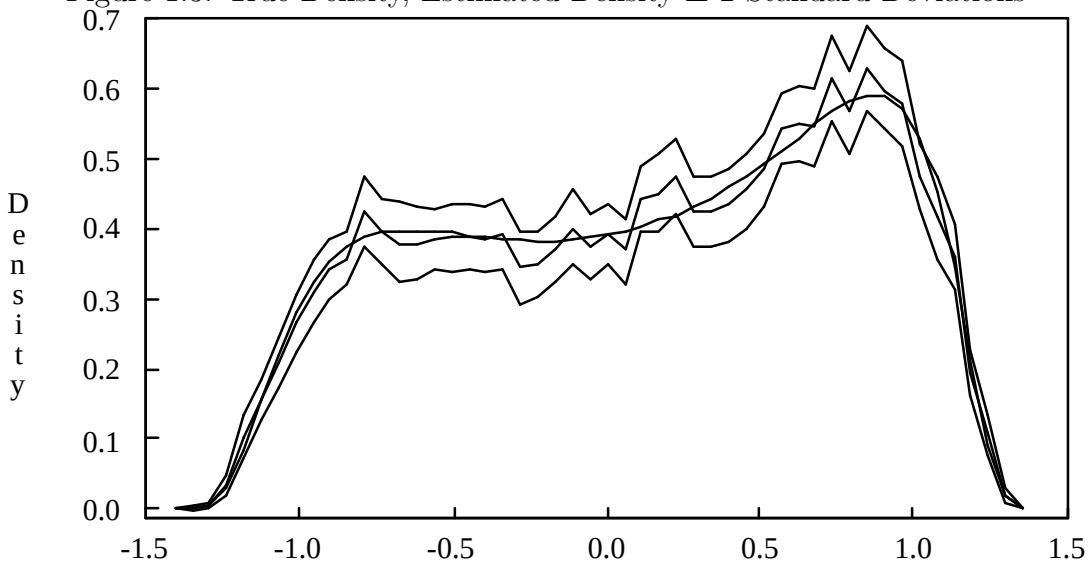
Figure 1.2: Density Estimates from 10 Bootstrap Replications



1.3.1 Bootstrap Confidence Intervals

There are three broad methods of computing confidence intervals using bootstrap simulation: standard intervals, percentile intervals, and tilting intervals. The first bootstrap confidence interval method is to use bootstrap

Figure 1.3: True Density, Estimated Density ± 2 Standard Deviations



standard deviation estimates in a confidence interval of the form

$$(T(Z) - z_\alpha \hat{\sigma}_{\text{boot}},\ T(Z) + z_\alpha \hat{\sigma}_{\text{boot}}). \tag{1.20}$$

This interval is approximately correct if both $T(Z) - T(F)$ and $T(X) - T(\hat{F})$ are approximately normally distributed with mean 0 and the same variance. The interval can be corrected based on a bias estimate, in which case the interval is

$$(T(Z) - 2\widehat{\text{bias}}_{\text{boot}} - z_\alpha \hat{\sigma}_{\text{boot}},\ T(Z) - 2\widehat{\text{bias}}_{\text{boot}} + z_\alpha \hat{\sigma}_{\text{boot}}). \tag{1.21}$$

These standard intervals require bootstrap estimates of bias and standard deviation. These are easy to estimate accurately, with 100 replications generally an adequate number. Therneau (1983) discusses variance reduction techniques for these applications. Importance sampling is only marginally useful here.

The standard intervals rely heavily on the parametric assumption that the distribution of $T(Z)$ has an approximately normal distribution, and that the variance of the distribution is constant. We hope to do better than this;

we can avoid these parametric assumptions using the other two bootstrap intervals, the bootstrap percentile and bootstrap tilting intervals (Efron 1982). For these intervals the simulations are more difficult to perform, as they require accurate quantile estimates in the extreme tails of the distribution of $T(X)$. Importance sampling is very helpful here. This will be discussed further in Chapter 4.

Chapter 2

Estimates

There are two steps in importance sampling—the choice of a sampling distribution, and the computation of results. We discuss the computation step first, because it is necessary to understand the principles behind importance sampling in order to know what is desirable in a sampling distribution.

In importance sampling a sample is obtained from the sampling distribution g, but an estimate for the true distribution is f required. This estimate is made using the empirical distribution, but with weights on the observations chosen to "unbias" the results. The weights should be roughly proportional to the "weight function"

$$W(x) := f(x)/g(x). \tag{2.1}$$

Still, within the constraint of "rough proportionality" different weights are possible, and are justified by different philosophies of importance sampling. Those philosophies, and the resulting estimates, are the subject of this chapter.

2.1 Integration Method

The classical importance sampling method is concerned with estimating the expected value of a random variable. X is a random variable with known density f, $\theta(X)$ is a function of X with an unknown distribution, and the quantity to be estimated is

$$\mu = E(\theta(X)) = \int \theta(x) f(x) dx. \tag{2.2}$$

The Monte Carlo method (without importance sampling) is to sample X_i from $f(x)$ using *simple random sampling* (SRS) and compute the average

$$\hat{\mu}_{\text{srs}} = \frac{1}{n} \sum_{i=1}^{n} \theta(X_i) \tag{2.3}$$

This estimate is unbiased for μ, and is asymptotically normally distributed with standard deviation $\sigma/\sqrt{n}$ if $\sigma^2 = \text{Var}_f(\theta(X)) < \infty$.

The classical importance sampling approach (Hammersley and Hanscomb 1964) to Monte Carlo integration is to is to modify the problem—instead of estimating $\int \theta(x) f(x) dx$, estimate $\int Y(x) g(x) dx$, where $Y(x) = \theta(x) f(x)/g(x)$, by sampling from $g(x)$ and computing the average:

$$\hat{\mu}_{\text{int}} = n^{-1} \sum_{i=1}^{n} Y(X_i) \tag{2.4}$$

This is an unbiased estimate if $g(x) > 0$ when $f(x)\theta(x) \neq 0$, and has variance $\text{Var}_g(Y(X))$. This is the "integration" estimate, or "int" for short.

In this approach the ideal sampling distribution is

$$g^*(x) = \frac{\theta(x)}{\mu} \tag{2.5}$$

(if $\theta(x) \geq 0$ for all x), which makes $\theta f/g$ constant, in which case the estimate has zero variance. This is generally impossible, even in simple applications. To paraphrase Hammersley and Hanscomb (1964), "We thus appear to have a perfect Monte Carlo method, giving the exact answer every time. This method is unfortunately useless, since to sample Y we must know g, and to determine g we must know μ, and if we already know μ we do not need Monte Carlo methods to estimate it."

If $\theta(x) \geq 0$ for all x does not hold the optimal sampling distribution is

$$g^*(x) = \frac{|\theta(x)| f(x)}{\int |\theta(z)| f(z) dz}. \tag{2.6}$$

This is no more possible to generate than is the perfect distribution (2.5).

In practice g is chosen based on physical intuition or preliminary calculations to make $\theta f/g$ more constant than θ. Computational considerations also play a part—for a simulation to run quickly it is important that it be

easy to generate random deviates from g quickly. We return to the discussion of sampling distributions in Chapters 5 and 6, and note here only that incredible variance reductions are possible but may be difficult to achieve.

The integration approach works very well in some applications, such as estimating the probability of a rare event (many authors). However variance reduction is not guaranteed, and the method can result in large variance increases. Examples are given by Bratley, Fox and Schrage (1983) and Hopmans and Kleijnen (1979).

Note that the sampling distribution is chosen not necessarily to emphasize regions where θ is extreme, but rather to induce a transformation

$$\theta \to Y = \frac{\theta f}{g}, \tag{2.7}$$

so that Y is more constant than θ. In most successful applications this is equivalent to sampling more where θ is extreme, but that is not implied by the formulation. If θ is usually large and only rarely small, the formulation implies that a good sampling distribution will sample *less* from small (extreme) θ.

There are at least four problems with the integration approach. First, a sampling distribution that induces a desirable transformation, may be hard to find, or to generate quickly.

In an application with multiple outputs it may be impossible to restructure the application in a way that works well for all output quantities simultaneously—a good sampling distribution for one may be disastrous for another. Note that every component of a multivariate output vector is multiplied by the same ratio of functions $f(x)/g(x)$, so that a sampling distribution which makes one output component more constant may make another component more variable.

The strict application of this approach can lead to inappropriate results. For example, if $\theta(X) = 1$ with probability 0.99 and $\theta = 0$ with probability 0.01, a sampling distribution which makes Y more constant than θ will sample *less* from the region where $\theta = 0$ than does the true distribution. This is a consequence of the attempt to make Y more constant, rather than sampling more often where Y is extreme.

Finally, the resulting estimate is not equivariant (Therneau 1983),

$$\theta_1 \equiv \theta_2 + c \not\Rightarrow \hat{\mu}_{\text{int},1} = \hat{\mu}_{\text{int},2} + c.$$

The benefits of integration importance sampling, as well as some of the pitfalls, can be seen in Table 2.1. The first problem is to estimate the probability that $X > z_\alpha = 2.326$, when X has a standard normal distribution. This is difficult using straight Monte Carlo integration—9,900 replications are required to reduce the standard deviation of the estimate to 10% of the true probability. Using the integration method with a sampling distribution $N(z_\alpha, 1)$ reduces the required number of replications to 268.

In this example, though, there are other quantities we wish to estimate: $P(X \leq z_\alpha)$, $E(X)$, and 1. For these quantities the results are disastrous, requiring 22,000 times as many replications as simple random sampling for estimating the probability that $X \leq z_\alpha$, for example. The method cannot even estimate a constant correctly.

Table 2.1: Gaussian Probability and Expectation
Variance of Integration Estimate

	$f = N(0,1)$ $\mathrm{Var}(\hat{\mu}_{\mathrm{srs}})$	$g = N(z_\alpha, 1)$ $\mathrm{Var}(\hat{\mu}_{\mathrm{int}})$	relative efficiency $\mathrm{Var}(\hat{\mu}_{\mathrm{int}})/\mathrm{Var}(\hat{\mu}_{\mathrm{srs}})$
$\mu_1 = P(X > z_\alpha)$	0.0099	0.000267	0.027
$\mu_2 = P(X \leq z_\alpha)$	0.0099	222.7	22,496
$\mu_3 = 1$	0.0	223.7	∞
$\mu_4 = E(X)$	1.0	1,433	1,433

$X \sim N(0,1)$, estimate $P(X > z_\alpha)$, $P(X \leq z_\alpha)$, $E(X)$ and $E(1)$, $z_\alpha = 2.326$. Using f as a sampling distribution corresponds to simple random sampling (no importance sampling). Values in this table were computed analytically and correspond to an experiment with one replication.

To be fair, the blame for this catastrophe should not be laid solely on the integration method. The sampling distribution used is very good for estimating μ_1 but is poor for estimating μ_4. Nevertheless, the two estimates presented below do equally well at estimating μ_2 as μ_1, and both estimate μ_3 correctly.

Estimating μ_3 may seem silly—why should we need to estimate the expectation of a constant? The reason is that in many simulation experiments some of the output values are nearly constants, and in other applications correct estimation of a constant is necessary for the output to be parsimonious.

That is true in this case, if you believe that $P(X > z_\alpha) + P(X \leq z_\alpha)$ should be equal to 1.

The integration method can be given another interpretation; it solves the original problem by sampling from a modified distribution and computing a weighted average, with weights chosen to correct for the sampling bias. The integration weights are

$$V_{\text{int},i} = W(X_i)/n \tag{2.8}$$

and the resulting estimate is

$$\hat{\mu}_{\text{int}} = \sum_{i=1}^{n} V_{\text{int},i}\theta(X_i)$$

Note that $E_g(W(X)) = 1$, but that $W(X)$ is a random variable. Therefore the weights used in this estimate do not in general sum to 1. This is responsible for many of the failures of this estimate, including lack of equivariance, negative variance estimates, undefined quantiles, constants estimated incorrectly, and lack of self-consistency of multivariate output.

2.2 Ratio and Regression Estimates

The next two importance sampling estimates are obtained by moving away from the view of Monte Carlo simulation as a method of integration, and instead considering Monte Carlo simulation to be a method of obtaining information about the distribution of the output variable θ. In fact the expectation of θ may all that is needed, but this is not required.

From this viewpoint it is natural to ask that the estimated distribution of θ be a probability distribution, with mass 1. We do this by assigning weights to each replication, with the sum of the weights equal to 1.

The *ratio* estimate is obtained by normalizing the weights used in the integration estimate by multiplying each weight value by the same constant. The ratio weights are:

$$V_{\text{ratio},i} = W(X_i)/\sum_{j=1}^{n} W(X_j) \tag{2.9}$$

The resulting distribution estimate has weight $V_{\text{ratio},i}$ on point θ_i, and the estimate of the expectation of $\theta(X)$ is:

$$\hat{\mu}_{\text{ratio}} = \sum_{i=1}^{n} V_{\text{ratio},i}\theta(X_i) \tag{2.10}$$

The *regression* estimate is obtained by using weights:

$$V_{\mathrm{reg},i} = W_i(1 + a(W_i - \bar{W}))/n \tag{2.11}$$

where $a = (1 - \bar{W})/\hat{\sigma}_w^2$, and $\hat{\sigma}_w^2 = n^{-1}(W_i - \bar{W})^2$. This estimate is derived by choosing "metaweights"

$$\pi_i = c_1 + c_2 W_i, \tag{2.12}$$

with c_1 and c_2 chosen such that

$$\sum_{i=1}^{n} \pi_i = \sum_{i=1}^{n} \pi_i W_i = 1. \tag{2.13}$$

The product of the meta-weights and the weight function gives regression weights, $V_{\mathrm{reg},i} = \pi_i W_i$. Where the ratio estimate is obtained by normalizing the distribution to have unit mass, this estimate is obtained by normalizing to unit mass *and* the correct expected value of W. The regression estimate is the weighted average of the θ values with respect to the product weights $V_{\mathrm{reg},i}$, which is equivalent to the weighted average of the Y values with respect to the metaweights:

$$\begin{aligned} \hat{\mu}_{\mathrm{reg}} &= \sum_{i=1}^{n} \pi_i Y_i \\ &= \sum_{i=1}^{n} V_{\mathrm{reg},i}\theta(X_i) \end{aligned} \tag{2.14}$$

The ratio and regression estimates are usually computed without explicitly computing the weights. The ratio estimate is the ratio of the sample averages of Y and W,

$$\hat{\mu}_{\mathrm{ratio}} = \bar{Y}/\bar{W}, \tag{2.15}$$

and the regression estimate is the value of the regression line of Y on W at $W = 1$,

$$\hat{\mu}_{\mathrm{reg}} = \bar{Y} - \hat{\beta}(\bar{W} - 1), \tag{2.16}$$

where $\hat{\beta}$ is the slope of the regression line.

The three estimates $\hat{\mu}_{\mathrm{int}}$, $\hat{\mu}_{\mathrm{ratio}}$, and $\hat{\mu}_{\mathrm{reg}}$ correspond to the standard, ratio and regression estimates for estimating $E(Y)$ in the presence of a covariate W with known mean (but without β fixed!); see e.g. Cochran (1977).

Both new estimates are weighted averages with weights that sum to one. They can be easily extended to estimating quantities other than expectations, and both are affine equivariant—if $\hat{\mu}(\theta)$ is the estimated expectation of θ, then $\hat{\mu}(a + b\theta) = a + b\hat{\mu}(\theta)$. Both new estimates tend to perform better in simulations with multivariate output than does the classical estimate, since they do not depend on transforming the output to be nearly constant (which can not be done using a single transformation unless all dimensions of the output are scaled versions of each other).

The performance of the new estimates in the Gaussian expectation example (Table 2.1) is shown in Table 2.2. Neither new estimate performs as well as the integration estimate for μ_1, but both perform better for the other three quantities in the experiment. The regression estimate performs better than the ratio estimate; in fact the latter is uniformly worse than using no importance sampling. This is due to the particular sampling distribution used, which is excellent for the integration estimate of μ_1, but poor for the ratio estimate.

Table 2.2: Mean Square Error of Importance Sampling Estimates

	Integration	Ratio	Regression	SRS
$\mu_1 = P(X > z_\alpha)$	0.0000065	0.00035	0.000020	0.00025
$\mu_2 = P(X \leq z_\alpha)$	2.2	0.00035	0.000020	0.00025
$\mu_3 = 1$	2.2	0	0	0
$\mu_4 = E(X)$	4.4	0.66	0.40	0.025

$X \sim f = N(0, 1)$, $g = N(z_\alpha, 1)$, $z_\alpha = 2.326$. There are 40 replications in each Monte Carlo experiment and 2000 Monte Carlo experiments. Results presented are the mean square error of the 2000 values. The last column contains reference values, which are the variance of a Monte Carlo estimate using simple random sampling (no importance sampling).

There are several ways to derive the regression estimate. The linear regression approach is one; another approach is to transform θ prior to computing the integration estimate using formula (2.4) by subtracting a constant, with the constant chosen to minimize the estimated standard error of the estimate, then adding the constant to the result; the idea is to take advantage

of the lack of equivariance of the integration estimate.

$$\hat{\mu}_{\text{reg}} = c^* + \frac{1}{n}\sum_{i=1}^{n} W_i(\theta_i - c^*) \tag{2.17}$$

where c^* minimizes $\sum W_i(\theta_i - c^*)^2$. The constant c^* turns out to be equal to $\hat{\beta}$, and the result is equivalent to the regression estimate.

The regression estimate is also equivalent to using $W(X)$ as a control function, with an estimated (not fixed in advance!) parameter:

$$\hat{\mu}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - c^*(W_i - E(W))), \tag{2.18}$$

where c^* minimizes $\widehat{\text{Var}}(Y_i - c^*W_i)$. This is equivalent to (2.17), and the result is again that $c^* = \hat{\beta}$ and the estimate is the regression estimate.

The final derivation is the approach using meta-weights, outlined above. Choose a linear function $\pi(w)$ such that $1 = \sum_i \pi_i = \sum_i \pi_i W_i$. The values π_i are "weights on the weights", chosen so that the weighted average of the W_i values has the correct mean. The product values $\pi_i W_i$ are the regression weights. The principle here is that if the average W value is too low it indicates that regions of the sampling space where W is small were overrepresented in the Monte Carlo sample. Rather than normalize the weights to sum to one by multiplying all weights by the same amount, we should place relatively more weight on the observations with large W values. Conversely, if the average W is too high relatively more weight should be placed on observations with small W values.

But why use a linear function to provide weights on the weights? A linear function is not necessarily appropriate, and can lead to bad results, including negative weights (in practice this is unlikely for reasonably large numbers of replications, and can be avoided by careful choice of sampling distributions). There are two nonlinear functional forms that seem more appropriate, which result in the "maximum likelihood" and "exponential" estimates. The estimates are similar, and the regression estimate can be interpreted as a linear approximation of either.

2.3 Nonlinear Estimates

Given a sample $\{X_i, i = 1, 2, \ldots n\}$ from a distribution g with $E_g(W(X)) = 1$, how can weights be placed on the observations so that the weighted empirical

distribution has an expected value for W equal to the known expectation of W? The weights $\pi_i = \pi(W_i)$ should be a function only of $\{W_i\}$, and should satisfy the constraints:

$$\sum_{i=1}^{n} \pi_i = 1 \tag{2.19}$$

$$\sum_{i=1}^{n} \pi_i W_i = 1 \tag{2.20}$$

so that distribution has weights that sum to one, and that the expected value of W is equal to 1. In the rest of this section all optimization is done subject to these constraints, and equations which express the weights as a function of W have parameters a and b chosen to satisfy these two constraints. In addition, a reference to weights in this section refers not to W, but to π, which describes how much weight to place on the point (W_i, Y_i).

The linear regression weights are one solution. Let the weights be a function of W of the form:

$$\pi_{\text{reg}}(W) = a + bW, \tag{2.21}$$

where a and b satisfy the distribution constraints. This functional form is obtained by minimizing the variance of the weights:

$$\text{minimize } \sum_{i=1}^{n} (\pi_i - 1/n)^2. \tag{2.22}$$

A second solution is obtained by an empirical maximum likelihood distribution. Find the maximum likelihood estimate of the distribution of W, subject to the constraint that the distribution have support only on the observed sample points. The distribution estimate can then be expressed in terms of the weight π_i assigned to each of the sample points, and the likelihood function is:

$$\text{maximize } \prod_{i=1}^{n} \pi_i \tag{2.23}$$

This results in weights of the form:

$$\pi_{\text{ml}}(W) = \frac{a}{1 - bW} \tag{2.24}$$

These weights can also be obtained by minimizing the backward Kullback-Leibler distance $\text{KL}(\mathbf{u}, \mathbf{\Pi})$ between $\mathbf{u}$ and $\mathbf{\Pi}$:

$$\text{minimize } \sum_{i=1}^{n} u_i \log(u_i/\pi_i) \tag{2.25}$$

where $u_i \equiv 1/n$, $\mathbf{u} = (u_1, u_2, \ldots u_n)$, $\mathbf{\Pi} = (\pi_1, \pi_2, \ldots \pi_n)$. $\mathbf{u}$ is the vector of weights that would be used if it were not necessary to modify the distribution to obtain the correct expected value of W.

The third solution is obtained by minimizing the forward Kullback-Leibler distance $\mathrm{KL}(\mathbf{\Pi}, \mathbf{u})$ between $\mathbf{u}$ and $\mathbf{\Pi}$:

$$\text{minimize } \sum_{i=1}^{n} \pi_i \log(\pi_i/u_i). \tag{2.26}$$

This results in weights in the form of an exponential function:

$$\pi_{\exp}(W) = ae^{bW}. \tag{2.27}$$

The estimates obtained using the latter two sets of weights are the *maximum likelihood* and *exponential* estimates, respectively.

2.3.1 Exponential Estimate as an Iterated Regression Estimate

There is another interesting derivation of the exponential weights, based on the regression estimate. A linear regression estimate of the expected value of a function at a point is usually expressed in terms of the value of the regression line or plane at that point, but can also be expressed in terms of a weighted average of the Y values used in the parameter estimation process, with weights depending on the relationship between the values of the independent variable used in parameter estimation and the values at that point. We used this principle above to find weights π_{reg} so that the weighted sample average of the W values was equal to the expected value of W, which is 1. If we apply the regression procedure repeatedly to modify the sample average of W in small increments from $\bar{W}$ to 1 we obtain weights that approximate the exponential weights, and that converge to the exponential weights as the increments are made small.

In matrix notation, the unweighted linear regression model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.28}$$

where $\mathbf{Y}$ is an $n \times k$ matrix, $\mathbf{X}$ is an $n \times 2$ matrix, $\boldsymbol{\beta}$ is a $2 \times k$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times k$ vector of random deviations which satisfy certain conditions—expected value 0, rows are independent of each other,

and each column is uncorrelated with all columns in $\mathbf{X}$. The k columns of $\mathbf{Y}$ correspond to k-dimensional output from a simulation model. The first column of $\mathbf{X}$ is a vector of ones, and the second column is $\mathbf{W}$, with W_i in row i.

The least-squares solution for the parameters is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.29}$$

The conditional estimated expected value of $\mathbf{Y}$ for a given value of W, is

$$\hat{E}(\mathbf{Y}|W) = (1, W)\hat{\boldsymbol{\beta}} = (1, W)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.30}$$

The right side of (2.30) can be rewritten as the product of a number of factors, all of which are independent of $\mathbf{Y}$, times $\mathbf{Y}$. The product is the formula for the regression weights when W is set equal to its expected value. The regression weights are:

$$\boldsymbol{\Pi}_{\text{reg}} = (1, 1)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2.31}$$

where $\boldsymbol{\Pi}_{\text{reg}}$ is a row vector. Note that the weights are independent of $\mathbf{Y}$, and of the dimension of $\mathbf{Y}$. The weights sum to 1, since the first column of $\mathbf{X}$ is a column of ones, so the sum of the weights is:

$$\boldsymbol{\Pi}_{\text{reg}}(1, 1, \ldots, 1)^T = (1, 1)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(1, 0)^T = 1 \tag{2.32}$$

The same idea can be applied in weighted regression. The weighted-least-squares solution for parameters is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\text{diag}(\mathbf{V})\mathbf{X})^{-1}\mathbf{X}^T\text{diag}(\mathbf{V})\mathbf{Y}, \tag{2.33}$$

where $\text{diag}(\mathbf{V})$ is a diagonal matrix with $\text{diag}(\mathbf{V})_{ii} = \mathbf{V}_i$, the weight given to observation i. Run a weighted regression with weight vector $\mathbf{V}$ to estimate the value of the regression line at $\bar{W}^{k+1}$. The new weights are:

$$\boldsymbol{\Pi}^{(k+1)} = (1, \bar{W}^{(k+1)})(\mathbf{X}^T\text{diag}(\mathbf{V})\mathbf{X})^{-1}\mathbf{X}^T\text{diag}(\mathbf{V}). \tag{2.34}$$

which can be expressed without matrix notation as:

$$\pi_i^{(k+1)} = V_i\left(1 + \frac{(W_i - \bar{W}^{(k)})(\bar{W}^{(k+1)} - \bar{W}^{(k)})}{\hat{\sigma}^{(k)}}\right) \tag{2.35}$$

when $\mathbf{V}^T\mathbf{W} = \bar{W}^{(k)}$ and $\hat{\sigma}^{(k)}$ is the weighted standard deviation of W values with weight vector $\mathbf{V}$. $\mathbf{\Pi}^{(k+1)}$ is new vector of weights that gives $\bar{W}^{(k+1)}$ as a weighted average of W values

$$\mathbf{\Pi}^{(k+1)}\mathbf{W} = \bar{W}^{(k+1)}. \tag{2.36}$$

Note that if the old and new weighted averages are the same the weights remain unchanged. That is, if

$$\bar{W}^{(k+1)} = \mathbf{V}^T\mathbf{W} = \bar{W}^{(k)} \tag{2.37}$$

then

$$\mathbf{\Pi}^{(k+1)} = \mathbf{V}^T. \tag{2.38}$$

We can repeat this weighted regression process, substituting $\mathbf{\Pi}^{(k)T}$ for $\mathbf{V}$ in (2.34). If we start with equal weights $\pi_i^{(0)} = \frac{1}{n}$ and choose a sequence of averages $\bar{W} = \bar{W}^{(0)} < \bar{W}^{(1)} < \bar{W}^{(2)} \ldots < \bar{W}^{(K)} = 1$ (or $\bar{W}^{(k)} > \bar{W}^{(k+1)}$ if $\bar{W} > l$) we obtain from (2.35) weights of the form

$$\pi_i^{(K)} = \frac{1}{n}\prod_{k=0}^{K-1}\left(1 + \frac{(W_i - \bar{W}^{(k)})(\bar{W}^{(k+1)} - \bar{W}^{(k)})}{\hat{\sigma}^{(k)}}\right) \tag{2.39}$$

These converge to the exponential weights (2.27) as $\max|\bar{W}^{(k+1)} - \bar{W}^{(k)}| \to 0$ (and $K \to \infty$) (proof in the appendix).

This indicates a computationally efficient way to approximate the exponential weights by a small iteration method—choose a small number of iterations K (2, 3, or 4), choose equally spaced target weighted averages $\bar{W}^{(k)} = \bar{W} + \frac{k}{K}(1 - \bar{W})$, and use (2.35) K times. This method can be applied if it is apparent that the regression weights are inadequate, if some are negative, for example. With K large enough the weights are not negative (see the previous proof in the appendix).

2.4 Influence Function

The influence function, developed in the context of robust estimation (Hampel 1968, 1974, Andrews et al. 1972), can aid our understanding of estimation methods in importance sampling.

The influence function of x is defined as

$$\mathrm{IF}(x; \theta, f, g, \mathrm{est}) = \lim_{\epsilon \to 0}\frac{\hat{\mu}_{\mathrm{est}}((1-\epsilon)g + \epsilon\delta_x) - \mu}{\epsilon} \tag{2.40}$$

where $\hat{\mu}_{\text{est}}(h)$ is the estimate of $E(Y)$ obtained using estimate "est" and distribution h, $Y = \frac{\theta f}{g}$, and δ_x is the distribution with a point mass on x.

Under certain conditions (discussed in Section 2.5)

$$\sqrt{n}\left(\hat{\mu}_{\text{est}} - \mu - n^{-1}\sum_{i=1}^{n} \text{IF}(X_i; \theta, f, g, \text{est})\right) \to 0 \tag{2.41}$$

in probability. Thus an estimate is asymptotically equivalent (to first order) to μ plus a sample average of the influence function at the observed values. In rough terms, the influence function for an observation measures the contribution that observation makes to the final result.

To define the influence function for importance sampling estimates we need to extend the definitions of the estimate to include the case where h is a distribution rather than a sample. This extension is straightforward; the definitions for the three basic estimates are:

$$\begin{aligned}
\hat{\mu}_{\text{int}}(h) &= E_h(Y) && (2.42)\\
\hat{\mu}_{\text{ratio}}(h) &= \frac{E_h(Y)}{E_h(W)} && (2.43)\\
\hat{\mu}_{\text{reg}}(h) &= E_h(Y) - \beta_h(E_h(W) - 1), && (2.44)
\end{aligned}$$

where $\beta_h = \text{Cov}_h(Y, W)/\text{Var}_h(W)$.

The influence functions for the integration, ratio, and regression estimates, and for a Monte Carlo estimate without importance sampling, are:

$$\begin{aligned}
\text{IF}_{\text{int}}(X) &= W\theta - \mu && (2.45)\\
\text{IF}_{\text{ratio}}(X) &= W(\theta - \mu) && (2.46)\\
\text{IF}_{\text{reg}}(X) &= W(\theta - \beta) + \beta - \mu && (2.47)\\
\text{IF}_{\text{srs}}(X) &= \theta - \mu && (2.48)
\end{aligned}$$

For the integration estimates (2.41) holds if g dominates $f|\theta|$; it holds for the ratio and regression estimates as well if g dominates f and W and Y have finite variance under g (proof in appendix).

The influence function of the ratio estimate is the most natural. Any single replication has an influence which is the product of the weight for that replication and the difference between θ and μ. This is the influence that the replication would have had under simple random sampling, corrected depending on whether that replication was more or less likely to be drawn

under the sampling distribution than it was under simple random sampling. This agrees nicely with the interpretation of the ratio estimate as a weighted average of results, with weights proportional to the inverse likelihood ratio.

The influence function of the ratio estimate also has the property that for any sampling distribution g, the integral of the influence function over any region of the sampling space with respect to g, is equal to the integral of the simple random sampling influence function over the same region with respect to the true distribution f, i.e.

$$\int_A \mathrm{IF}_{\mathrm{ratio}}(x)g(x) = \int_A \mathrm{IF}_{\mathrm{srs}}(x)f(x) \tag{2.49}$$

for any region A and any output θ. In effect, the structure of the application has not changed. This can be important if other variance reduction techniques are used in combination with importance sampling.

The influence function of the integration estimate reflects the transformation aspect of that estimate—the influence is the difference between the transformed value and the expected value. It is not difficult to construct examples where the results are counter-intuitive, in that small θ values cause the final result to be larger, and large θ values cause the final result to be smaller, as in Example 2.1:

Example 2.1 Influence function reversal with the integration estimate

X takes on values 1 or 2, with probability 50% each. $\theta(1) = 1$, and $\theta(2) = 2$. The sampling distribution samples $X = 1$ and $X = 2$ with probabilities 10% and 90%, respectively.

$\mathrm{IF}_{\mathrm{int}}(1) = 7/2$, and $\mathrm{IF}_{\mathrm{int}}(2) = -7/18$. If every draw in an experiment yielded $X = \theta = 2$, the estimate would be 10/9. If every draw yielded $X = \theta = 1$, the estimate would be 7/2. Low θ values cause the estimate to be higher, and vice versa.

The influence function for the regression estimate reflects the implicit use of W as a linear control function. If $\theta(X)$ is larger than expected based on the value of the regression line at $W(X)$ the influence of that replication is positive. It is possible to construct examples where some small (large) θ values are associated with large (small) influence function values for the regression estimate, but it is not possible to achieve a perfectly negative association (which would imply $\theta < \mu \Leftrightarrow \mathrm{IF} > 0$), as was the case with the integration estimate. The proof is in the appendix.

The ratio and regression estimates are equivariant, so their influence functions are not changed by the addition of a constant.

2.4.1 Influence of the Nonlinear Estimates

The influence functions for the exponential and maximum likelihood estimates are the same as for the regression estimate, if they exist, and under certain conditions the same limiting approximation (2.41) holds.

First, define general distribution versions of the two estimates

$$\hat{\mu}_{\exp}(h) = E_h(Yae^{bW}) \tag{2.50}$$

$$\hat{\mu}_{\text{mle}}(h) = E_h(Y\frac{a}{1-bW}) \tag{2.51}$$

where a and b satisfy one of

$$1 = E_h(ae^{bW}) = E_h(Wae^{bW}) \tag{2.52}$$

$$1 = E_h(\frac{a}{1-bW}) = E_h(W\frac{a}{1-bW}) \tag{2.53}$$

for the exponential and maximum likelihood estimates, respectively (the values of a and b differ for exponential and maximum likelihood methods). The influence function (2.40) is well-defined if $\bar{W} \geq 1$. If $\bar{W} < 1$ it is well-defined for the mle estimate if the distribution of W under g is bounded, and for the exponential estimate if $E_g(e^{\epsilon W}) < \infty$ for some $\epsilon > 0$. Otherwise define a modified influence function

$$\text{IF}^*(x;\theta,f,g,\text{est}) = \lim_{n\to\infty} \text{IF}(x;\theta,f,g_n,\text{est}), \tag{2.54}$$

where g_n is the same as g except that any values of W which are greater than n are truncated. The modified influence functions for the two estimates are equal to the influence function for the regression estimate.

2.5 Variance and Bias of Estimates

Under mild regularity conditions the importance sampling estimates described are consistent and asymptotically normally distributed with asymptotic variances

$$\text{AVar}(\hat{\mu}_{\text{int}}) = \text{Var}_g(Y) = \int \frac{f(x)\theta^2(x)}{g(x)} f(x) - \mu^2 \tag{2.55}$$

$$\mathrm{AVar}(\hat{\mu}_{\mathrm{ratio}}) = \mathrm{Var}_g(Y - \mu W) = \int \frac{f(x)(\theta(x) - \mu)^2}{g(x)} f(x) \tag{2.56}$$

$$\mathrm{AVar}(\hat{\mu}_{\mathrm{reg}}) = \mathrm{Var}_g(Y - \beta W) = \int \frac{f(x)(\theta(x) - \beta)^2}{g(x)} f(x) - (\mu - \beta)^2 \tag{2.57}$$

That is,

$$(\hat{\mu}_{\mathrm{est}} - \mu) \to 0 \tag{2.58}$$

a.s. and

$$\sqrt{n}(\hat{\mu}_{\mathrm{est}} - \mu) \to N(0, \mathrm{AVar}(\hat{\mu}_{\mathrm{est}})) \tag{2.59}$$

in distribution.

For the integration estimate the required conditions are weak dominance and that Y have a finite first moment for consistency and a finite second moment for asymptotic normality. For the ratio and regression estimate g must dominate f. For the ratio estimate consistency holds if Y and W have finite first moments under g, and asymptotic normality if both have finite second moments. Both consistency and asymptotic normality for the regression estimate obtain if Y and W have finite second moments under g.

Under these conditions consistency follows from the consistency of the appropriate functions of the sample moments of Y and W. Asymptotic normality is proved by approximating each of the estimates as μ plus a sample average of the influence function at the observed X values, with the influence function given in (2.40). That average is asymptotically normal by the usual central limit theorem, with variance given above. The normality of the estimate then follows since the difference between the estimate and the approximation is small. See e.g. Cramér (1946, p. 366) or Bhattacharya and Ghosh (1978), or the appendix.

Note that the regression estimate has a smaller asymptotic mean square error than the other two estimates.

2.5.1 Edgeworth Expansions

Under somewhat stronger regularity conditions an Edgeworth expansion (Cramér 1946, Bhattacharya and Ghosh 1978) holds. If (Y), (Y and W), (Y^2 and W^2) have finite absolute third moments under g, for the integration, ratio, and regression estimates respectively (i.e. Y and W have finite sixth moments for the regression estimate), then by Bhattacharya and Ghosh (1978) Theorem 1,

$$P(\sqrt{n}(\hat{\mu}_{\mathrm{est}} - \mu) \le z\sqrt{\mathrm{AVar}_{\mathrm{est}}}) = \Phi(z) + O(n^{-1/2}).$$

If in addition Cramér's condition on the characteristic function of W and Y

$$\limsup_{|t|\to\infty} E_g(\exp(i\langle t, Z\rangle)) < 1$$

holds, where $Z = (W, W^2, Y, Y^2)$ and $\langle t, Z\rangle$ is the inner product of t and Z, then a second order formal Edgeworth expansion holds, and

$$P(\sqrt{n}(\hat{\mu}_{\text{est}} - \mu) \le z\sqrt{\text{AVar}_{\text{est}}}) = \Phi(z) - \frac{\phi(z)}{\sqrt{n}}\left(\frac{\text{bias}}{\sigma} + \frac{k_3}{6\sigma^3}(z^2-1)\right) + o(n^{-1/2})$$

where bias, σ, and k_3 are the leading terms for the bias, standard deviation, and third cumulant of the distribution of $\hat{\mu}_{\text{est}}$, obtained from a delta-method calculation of the moments of the estimates. This follows immediately from Bhattacharya and Ghosh (1978) Theorem 2. The σ term is the square root of the asymptotic variance given above, and the bias term is given below (multiply by n for inclusion in 2.26). For k_3 above substitute

$$E_g((Y-\mu)^3)$$

for the integration estimate,

$$E_g((Y-\mu W)^3) - 6E_g((W-1)(Y-\mu W))E_g((Y-\mu W)^2)$$

for the ratio estimate, and

$$E((Y - \beta W - \mu + \beta)^3)$$

for the regression estimate.

2.5.2 Bias

The integration estimate is unbiased. The ratio and regression estimates are biased, with first order bias terms

$$E(\hat{\mu}_{\text{ratio}} - \mu) \approx -E_g(W(Y-\mu W))/n$$

and

$$E(\hat{\mu}_{\text{reg}} - \mu) \approx -E_g((W-1)^2(Y - \beta W - (\mu - \beta)))/n$$

These are the first order bias terms in the Edgeworth expansion of the distributions of these estimates. These are more useful than the actual biases,

which may be infinite or undefined, if e.g. there is a nonzero probability that $W = 0$ for the ratio estimate, or if the distribution of W has an atom for the regression estimate.

This bias is of low order, and the asymptotic contribution of the bias term to the mean square error of these estimates is zero. For large samples the bias is negligible compared to the standard error, and need not be a consideration in the choice of estimates. The bias can be estimated from the sample to help determine this.

All of the estimates can be "conditionally biased." If the sampling distribution is not perfect (and it never is) there may be a region that has a small probability under g, but where Y is large. The results are conditionally biased—the expected value of the estimate, given that the region is not observed, is significantly different from μ. This is not a phenomenon restricted to importance sampling; for example, 20% of the expectation is concentrated in 5% of the density of an exponential distribution so a sample of size 20 has greater than 50% chance of being conditionally biased by a substantial amount. It is especially a problem in importance sampling because importance sampling is often applied in rare event applications, and because what determines whether a given X corresponds to an extreme value is the product $\theta(X)f(X)/g(X)$, which is large if g under-samples a small region. An example of this is given in Section 6.1.2, where only four percent of all experiments of size 1000 would include any observations from a region with large $W\theta$ values under the sampling distribution considered there.

2.6 Confidence Intervals

The asymptotic normality of the integration, ratio, and regression estimates can be used to obtain approximate $(1 - 2\alpha)$ confidence intervals of the form

$$\hat{\mu}_{\text{est}} \pm z_\alpha \hat{\sigma}_{\text{est}} \tag{2.60}$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$ and $\hat{\sigma}_{\text{est}}$ is the estimated standard error of estimate $\hat{\mu}_{\text{est}}$, given by

$$\hat{\sigma}^2_{\text{int}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{2.61}$$

$$\hat{\sigma}^2_{\text{ratio}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (Y_i - W_i \hat{\mu}_{\text{ratio}})^2 \tag{2.62}$$

$$\hat{\sigma}^2_{\text{reg}} = \frac{1}{n(n-2)} \sum_{i=1}^{n} (Y_i - \hat{\beta} W_i - \bar{Y} + \hat{\beta} \bar{W})^2 \tag{2.63}$$

These intervals have limiting $(1-2\alpha)$ coverage under the same conditions as required for asymptotic normality.

These intervals are not always acceptable, and we discuss four improvements. The first is to use the percentiles of the t-distribution in place of those of the Gaussian distribution in (2.60). However, if the sample size is larger than say 30 (and it *should* be in a Monte Carlo simulation) this produces a negligible change.

The next approach is to modify the standard interval based on an Edgeworth expansion of the joint distribution of the estimate and the estimated standard error. Johnson (1978) obtains an interval of the form

$$\bar{Y} + \frac{\hat{\mu}_3}{6nS^2} \pm t_{\alpha,n-1} S/\sqrt{n} \tag{2.64}$$

for a sample of size n from a univariate population, where $\hat{\mu}_3$ is the skewness of the sample, $t_{\alpha,n-1}$ is a quantile of the t-distribution with $n-1$ degrees of freedom, and S^2 is the sample variance $\frac{1}{n-1}(Y_i - \bar{Y})^2$. This interval applies to the integration estimate, and with modifications would apply to the other estimates.

In contrast to the use of the t-correction, which makes a change in the endpoints of the interval of the order $n^{-3/2}$, Johnson's interval makes corrections of the order n^{-1}, and so is far more significant for sample sizes used in Monte Carlo simulations. The Johnson correction is particularly appropriate in applications encountered in importance sampling, where the distributions are often heavily skewed.

A third approach, which also creates an asymmetric interval and makes the same order correction as the Johnson interval, is the bootstrap BC-a interval (Efron 1987). This interval, however, is computationally expensive if there are a large number of Monte Carlo replications. DiCiccio and Tibshirani (1987) develop an approximation that eliminates much of the computational requirement.

The final approach is specific to importance sampling, and attempts to correct the common problem in importance sampling, that intervals are often the most optimistic precisely when the estimates are the worst because the observations from the sampling distribution do not represent the true distribution and the sample average $\bar{W}$ is small. This interval is

$$\hat{\mu}_{\text{est}} \pm z_\alpha \hat{\sigma}_{\text{est}} / \bar{W} \tag{2.65}$$

which is obtained by dividing the standard error estimate by the average W value.

$$\hat{\sigma}^*_{\text{int}} = \hat{\sigma}^2_{\text{int}}/\bar{W} \tag{2.66}$$
$$\hat{\sigma}^*_{\text{ratio}} = \hat{\sigma}^2_{\text{ratio}}/\bar{W} \tag{2.67}$$
$$\hat{\sigma}^*_{\text{reg}} = \hat{\sigma}^2_{\text{reg}}/\bar{W} \tag{2.68}$$

This idea could also be used in combination with Johnson's interval to obtain, e.g.

$$\bar{Y} + \frac{\hat{\mu}_3}{6nS^2} \pm \frac{t_{\alpha,n-1}S}{\sqrt{n}\bar{W}} \tag{2.69}$$

The more disjoint the true and sampling distributions are, the more likely it is that W_i will be small for any given observation, and for all observations in the sample as a whole. But note that when all W_i are small, the standard error estimates are also small (for roughly constant θ). Thus the less representative the sampling distribution is of the true distribution, the more likely it is that the sampling estimate will appear to perform extremely well.

We borrow an example from the fuel inventory simulation example which will be described further in Chapters 4 and 6. There are 2000 observations from a single sampling distribution, and we estimate the expected inventory costs for a range of distributions of gas demand. Gas demand has five normally-distributed monthly values with standard deviation 100 and serial correlation 0.2. The original sampling distribution makes a relatively minor change in these distributions, with no monthly expected value changed by more than 20.

To look at how standard errors perform we make major changes, up to ± 500 (5 standard deviations) in each monthly value. For small changes the standard error estimates appear reasonable, but for larger changes the estimated standard errors decrease. This is shown in Figure 2.1.

The reason is shown in Figure 2.2. For distributions of gas demand close to the base case, the original distribution, the average value of W is approximately equal to 1. However, for more extreme distributions the sample average W is small. This affects not only the estimates, but the usual standard error estimates (2.61-2.63).

The adjusted standard errors given in (2.66-2.67) are shown in Figure 2.3. While not perfectly satisfactory, they appear reasonable for a wider range of distributions than do the usual standard errors.

Figure 2.1: Estimated Standard Errors for Response Surface Estimation

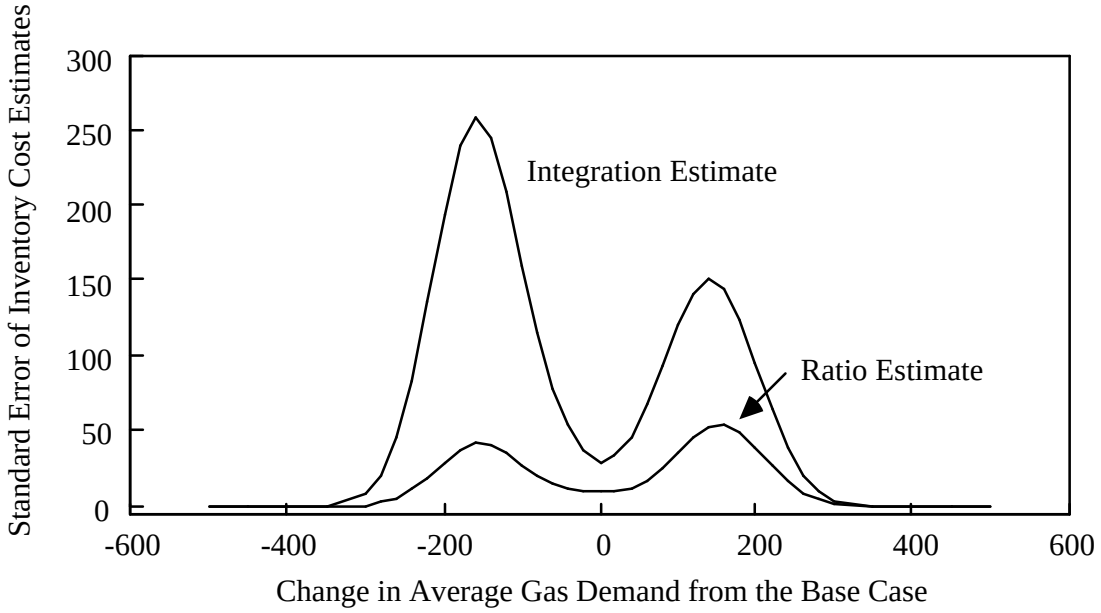
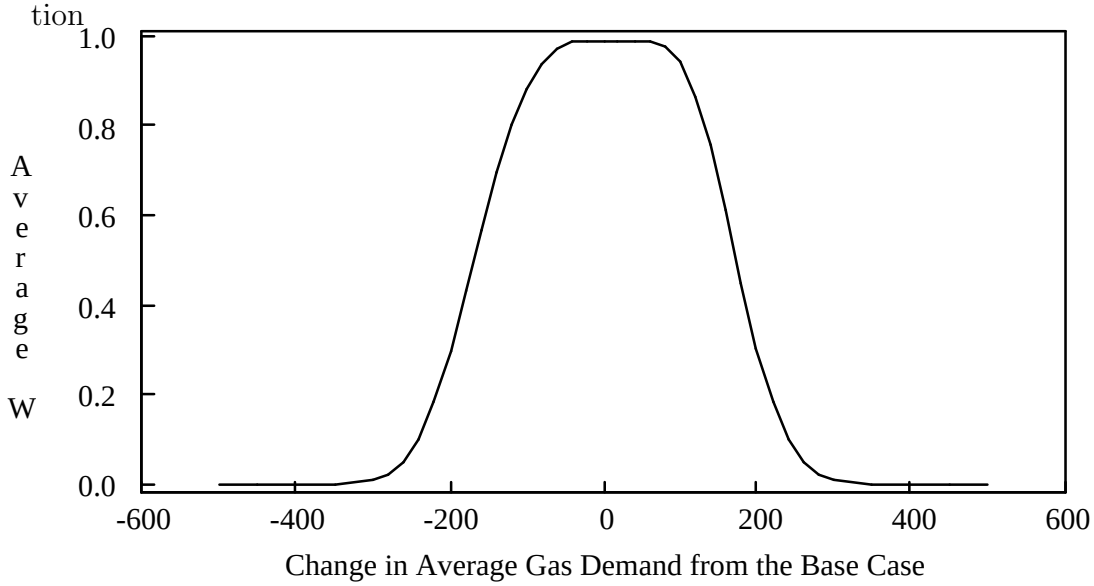


Figure 2.2: Observed Average Ratio $W = f/g$ in Response Surface Estimation



2.7 Unbiased Regression Estimate

The regression estimate as given in (2.16) is $\hat{\mu}_{\text{reg}} = \bar{Y} - \hat{\beta}(\bar{W} - 1)$. This estimate is not unbiased because of the correlation of $\hat{\beta}$ and $\bar{W}$, but can be modified to make it unbiased. The idea is to consider the formula as the average of adjusted replication results, with the modification based on the global regression estimate $\hat{\beta}$. An unbiased estimate results if the adjustment is based on a regression coefficient computed leaving out the replication to be adjusted; this eliminates the dependence (and correlation) between a replication value and the coefficient used to adjust it.

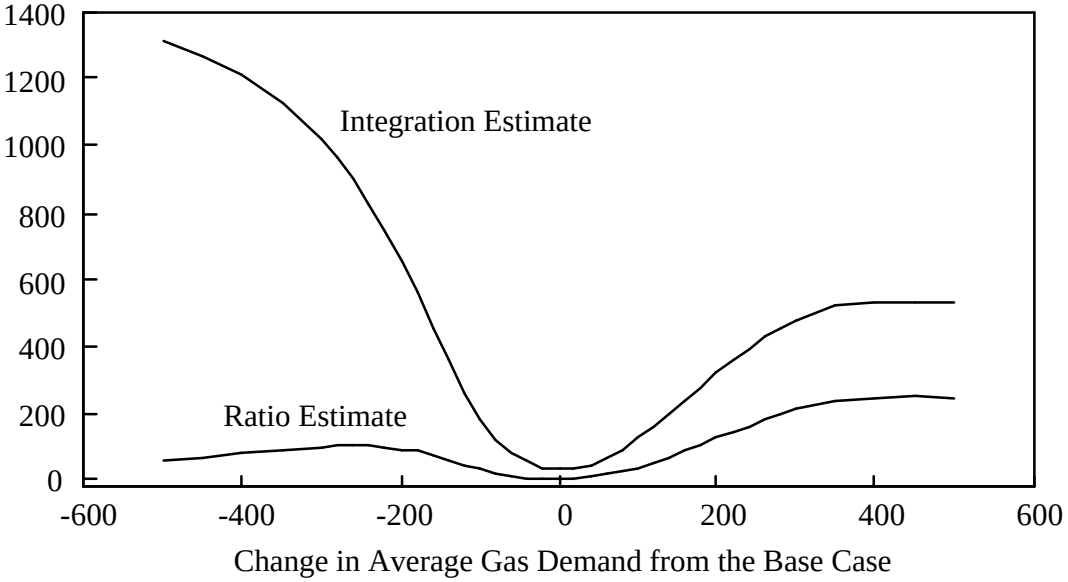
First, rewrite the regression estimate in terms of an average of adjusted Y values.

$$\hat{\mu}_{\text{reg}} = n^{-1} \sum_{i=1}^{n} Y_i - \hat{\beta}(W_i - 1) \tag{2.70}$$

The unbiased regression formula is similar:

$$\hat{\mu}_{\text{unbiased}} = n^{-1} \sum_{i=1}^{n} Y_i - \hat{\beta}_{-i}(W_i - 1) \tag{2.71}$$

Figure 2.3: Standard Error Adjusted for Average Weight



where $\hat{\beta}_{-i}$ is the regression coefficient obtained using $n-1$ replications, every replication except replication i.

Now the values $\hat{\beta}_{-i}$ and $(W_i - 1)$ are independent, so the expectation of their product is the product of their expectations, which is 0. Thus the expectation of the estimate is the same as the expectation of $\bar{Y}$, which is μ.

This derivation assumes that replications are independent. This is not the case if certain other variance reduction techniques, such as antithetic variates, are used. In that case this estimate is not unbiased.

The unbiased regression estimate shares many of the desirable properties of the usual regression estimate—it is equivariant, and has the same low asymptotic variance. It shares the principle theoretical disadvantage, that the weights can be negative, though less so than the regression estimate. It is harder to compute.

The unbiased estimate is, like all the other estimates, a weighted average

$$\hat{\mu}_{\text{unbiased}} = \sum_{i=1}^{n} V_{\text{unbiased},i}\theta_i \tag{2.72}$$

with weights:

$$V_{\text{unbiased},i} = \frac{W_i}{n}\left(1 + \frac{n}{n-1}\frac{(W_i - \bar{W})(W_i - 1)}{S^2_{-i}} + W_i A + B\right) \tag{2.73}$$

where

$$A = \sum_{i=1}^{n} \frac{1 - W_i}{S^2_{-i}} \tag{2.74}$$

$$B = \sum_{i=1}^{n} \frac{\bar{W}_{-i}(W_i - 1)}{S^2_{-i}} \tag{2.75}$$

$$S^2_{-i} = \sum_{j \neq i}^{n} (W_j - \bar{W}_{-i})^2 \tag{2.76}$$

and

$$\bar{W}_{-i} = \frac{1}{n-1}\sum_{j \neq i}^{n} W_j. \tag{2.77}$$

2.8 Computational Considerations

There are a number of computational considerations that affect the choice of importance sampling estimates. Computer memory requirements and the ease and speed of computing both estimates and standard errors of those estimates can affect the decision about which importance sampling estimate will be used.

In some applications expectation and standard error estimates should be computed using one-pass formulas, so that the results can be computed "on the run" and do not need to be stored. This can be an important consideration if many output quantities are computed from each replication, or if there are a large number of replications.

In other applications it is not important to compute a one-pass estimate. When quantities other than expectations are computed it may be necessary to have available results from all replications before doing any analysis. One-pass formulas are of less use here. Even for estimation of an expectation, the simulation may be set up so that all replications are run before any analysis is performed.

Numerically-stable one-pass formulas for the integration, ratio, and regression estimates for both the estimates and the standard error estimates are given in the appendix.

There are no one-pass formulas for the exponential, maximum-likelihood, and unbiased regression estimates; all require that the full set of results be available. The parameters for the nonlinear estimates must be computed using an iterative algorithm. Some computational efficiency is gained in the unbiased regression estimate by using updating procedures.

If a one-pass procedure is not required, then a handy shortcut is available for computing the regression estimate for a simulation with many output quantities. Rather than computing each estimate using linear regression, compute the weights once and compute a weighted average for each component. This does not produce standard error estimates.

2.9 Zero-Variance Estimates

The optimal sampling distribution, in terms of reducing the asymptotic variance of an estimate, is:

$$g^*(x) = C|\theta(x) - c|f(x) \tag{2.78}$$

where $c = 0$ and $c = \mu$ for the integration and ratio estimates respectively, and C is a normalizing constant. This is obtained using the calculus of variations (Kahn and Marshall 1953).

One attraction of the integration estimate combination is that if θ is nonnegative (or nonpositive) it can, in theory, provide a perfect (zero-variance) estimate from a single replication. The formula for such a sampling distribution is

$$g(x) = f(x)\theta(x)/\mu. \tag{2.79}$$

Unfortunately this formula involves μ, and so can not be used in practice. Still, there is potential for large reductions, and we wish to investigate this further.

Formula (2.79) requires that θ be either positive or negative. If θ takes on both positive and negative values then it is not possible to find a sampling distribution that achieves perfect variance reduction using the standard integration estimate.

If the distribution of θ is bounded either above or below the regression estimate can (theoretically) be used to obtain perfect estimates. Choose β outside the range of the distribution of θ, and use a sampling distribution of the form

$$g(x) = f(x)(\theta(x) - \beta)/(\mu - \beta) \tag{2.80}$$

so that

$$Y(x) = \mu + \beta(W(x) - 1) \tag{2.81}$$

which is a line running through the point $W = 1$, $Y = \mu$. The estimate of μ is perfect as soon as two observations have distinct W values. This is based on an idea of Therneau (1983) who suggested transforming θ prior to computing the integration estimate by subtracting a constant c, then adding the constant back to the result. This is equivalent to a regression estimate with β fixed at c.

Suppose that the distribution of θ is not bounded. Can we approach perfection by letting β go to infinity in the choice of a sampling distribution for the regression estimate? The answer depends on how heavy the tails of the distribution are.

Theorem 2.1 *Conditions for limiting zero-variance transformed regression estimate*

Suppose $\theta(X)$ has an unbounded distribution when $X \sim f$. Let $g_b(x)$ be the optimal sampling distribution for the integration estimate of $\theta - b$. Then

$$\lim_{b\to\infty} \mathrm{Var}_{g_b}(W(\theta - b)) = 0 \tag{2.82}$$

iff

$$\lim_{b\to\infty} b \int_{\theta>b} (\theta(x) - b) f(x)) = 0 \tag{2.83}$$

Corollary 2.1 *Conditions for limiting zero-variance transformed regression estimate*

If $E_f(\max(\theta, 0)^2) < \infty$ then

$$\lim_{b\to\infty} \mathrm{Var}_{g_b}(W(\theta - b)) = 0 \tag{2.84}$$

Similar results hold for $b \to -\infty$. The proofs are in the appendix.

In contrast, the ratio estimate can not be perfect. Note that the integration (regression) estimate can be perfect only if $0(\beta)$ is not within the interior of the range of θ. The corresponding value of the ratio estimate would be μ, which is possible only if θ is a constant. The minimum variance that can be achieved using the ratio estimate is

$$\left(\int |\theta(x) - \mu| f(x) dx \right)^2 \tag{2.85}$$

The mere existence of a theoretical limit, where none exists for the integration or regression estimates (if θ is bounded above or below) is not an overwhelming concern, because of the difficulty of finding perfect sampling distributions. In practice this limit is of concern, as the variances actually obtained using the integration or regression estimates in some applications are below the theoretical minimum obtainable using the ratio estimate. In Example 2.1 the relative efficiency of the integration estimate was 0.027, while the minimum achievable efficiency for the ratio estimate is 0.0396.

2.10 Comparison of Estimates

What are the advantages of each estimate? The integration estimate is unbiased, and estimates rare event probabilities well. In other applications it can perform poorly. It should not be used in applications with multivariate output.

The ratio estimate is guaranteed to give reasonable answers, in the sense that the estimate will always be within the range of observed values. Both the integration and regression estimates can give answers which are unreasonable—the integration estimate because the sum of weights does not add to 1, and the regression estimate because some of the weights can be negative. The latter problem is unlikely to occur with a reasonable number of replications if the sampling distribution is chosen so that the distribution of W does not have extremely long tails.

The single most important argument in many cases is how accurately the estimates perform. In most cases the regression estimate wins. The leading term of the asymptotic variance of the regression estimate is never larger than the same term for either other estimate, and can be significantly smaller.

The two nonlinear (exponential and maximum likelihood) estimates and the unbiased estimate are nearly equivalent to the regression estimate. Whereas the differences between the integration, ratio, and regression are of order $O_p(1/\sqrt{n})$, the differences between the regression, maximum-likelihood, and unbiased estimates are of order $O_p(1/n)$. These estimates have the same low asymptotic variance as the regression estimate. For applications with reasonably large sample sizes the extra complexity of these estimates will not provide a significant improvement over the regression estimate. If the number of replications is small, and bias is a large concern or negative weights are observed then one of these estimates may be used.

The integration, ratio, and regression estimates are all relatively easy to compute. One-pass formulas are given in the appendix.

The regression estimate is generally the estimate of choice. The fact that a large number of independent derivations all arrive at the same answer is a heuristic argument for this choice, and the lower asymptotic variance is a concrete argument.

The integration estimate is competitive in estimating expectations of distributions with a large discrete mode at zero, and must be used when the sampling distribution is zero for part of the sampling space (when $\theta = 0$).

We have discussed the different estimation formulas without regard to the choice of sampling distribution. A good choice of sampling distribution is necessary for good performance with any estimate. In addition, different estimates perform well with different sampling distributions. The sampling distribution will determine which of the integration and ratio estimates has a smaller variance, for instance (the regression estimate has the smallest asymptotic variance for any g which dominates f).

The ratio estimate should not be used without the use of mixture sampling (discussed in Chapter 6). The regression estimate can be used (relatively) safely without mixture sampling. Mixture sampling can make the integration estimate slightly more robust in some applications, and though it can not cure the lack of equivariance of the estimate, it can mitigate it to some extent.

Chapter 3

Conditional Weights

This chapter describes the use of conditional weights in importance sampling—the replacement of weights based on the inverse likelihood ratio with their expected values conditioned on a "sufficient" statistic. The intent of this procedure is to attack one of the biggest problems encountered in importance sampling, that the weights for different replications vary wildly.

In any weighted average, large variability in the weights can be detrimental to the performance of the estimate. In importance sampling some variability in the weights is desirable; the intent of importance sampling is to choose a sampling distribution that is relatively large when the output is extreme, so that weights based on the inverse likelihood ratio are small. This is the desired relationship between weights and values, but the actual correspondence is rarely perfect. It is a common observation that the weights based on the inverse likelihood ratio can vary wildly, even for the same output values, particularly in applications with multivariate input. In extreme cases this variability can lead importance sampling estimates to have infinite variance.

The problem can be understood using an ANOVA-type decomposition of the variance of weights. The asymptotic variance of the three basic estimates is:

$$\mathrm{Var}(\hat{\mu}_{\mathrm{est}}) = E((Y - cW)^2) - (\mu - c)^2 \tag{3.1}$$

where $c = 0$, μ, and β, respectively, for the integration, ratio, and regression estimates. The first term can be written as:

$$\begin{aligned} E((Y - cW)^2) &= E(E((Y - cW)^2|\theta)) \\ &= E(E((\theta - c)^2 W^2|\theta)) \end{aligned}$$

$$= \quad E((\theta - c)^2 E(W|\theta)^2) + E((\theta - c)^2 \mathrm{Var}(W|\theta)), \quad (3.2)$$

where all expectations are with respect to g. The first term can be made small by choosing g so that W is small when θ is far from c; this is where the improvement from importance sampling can occur.

The second term is the variance of W orthogonal to θ, and is counterproductive. We can make this term smaller if we steal an idea from the variance reduction technique known as "conditional Monte Carlo," and replace the weights with their expected value, conditioned on an appropriate statistic.

In conditional Monte Carlo (without importance sampling), to estimate $\mu := E_f(\theta(X))$, sample from f as usual, but compute the estimate

$$\hat{\mu}_{\mathrm{cmc}} := n^{-1} \sum_{i=1}^{n} \Psi(X_i) \quad (3.3)$$

instead of the usual sample average $\hat{\mu}_{\mathrm{mc}} = \bar{\theta}$. Here

$$\Psi(X) = \Psi(S(X)) := E(\theta(X)|S(X)) \quad (3.4)$$

is the expected value of θ, conditioned on the value of a statistic $S(X)$. This estimate is consistent, since

$$\mu = E(\theta) = E(E(\theta(X)|S(X))) = E(\Psi(X)) \quad (3.5)$$

Furthermore, the conditional estimate has smaller variance (unless $\theta = \Psi$ a.s.), since Ψ has a smaller variance than θ, as seen from the identity

$$\mathrm{Var}(\theta) = E\left(\mathrm{Var}(\theta)|S(X)\right) + \mathrm{Var}\left(E(\theta(X)|S(X))\right) \quad (3.6)$$

The latter term is the variance of Ψ. The first term is nonnegative, and represents the improvement obtained using Ψ.

We use the same conditioning principle for the weights in importance sampling. Let $S(X)$ be a sufficient statistic (defined below) for θ for which

$$\Omega(X) := E(W(X)|S(X)) = f_S(S)/g_S(S) \quad (3.7)$$

can be computed. Then use the values $\Omega(X_i)$ in place of the $W(X_i)$ values in the computation of weights for use in any of the estimates.

Here a sufficient statistic for θ is a statistic $S(X)$ such that $\theta(X)$ depends on X only through S, $\theta(X) = \theta(S(X)) = E(\theta(X)|S(X))$. Here θ is a function

of X (in contrast to the usual sufficiency context, where θ is a parameter of a parametric distribution), but the usual definition of sufficiency (that the conditional distribution of X given S be independent of θ) is trivially satisfied since θ is a function of S. The best statistic to use for S is θ itself, if the conditional expected value of the weights given θ can be computed.

Conditioning the weights gives consistent results, since

$$\begin{aligned} E_g(W(X)\theta(X)) &= E_g(E(W(X)\theta(X)|S(X))) \\ &= E_g(\theta(X)E(W(X)|S(X))) \\ &= E_g(\theta(X)\Omega(X)) \end{aligned} \tag{3.8}$$

Note that $\theta(X) = \theta(S(X))$ implies that $\theta(X) = E(\theta(X)|S(X))$. In addition, $E_g(\Omega(X)) = E_g(W(X)) = 1$.

Conditioning the weights reduces the variance of all estimates (subject to certain conditions). The asymptotic variance of the integration estimate, for example, is $\mathrm{Var}_g(\theta(X)\Omega(X))$, and

$$\begin{aligned} \mathrm{Var}_g(\theta(X)W(X)) &= E((\mathrm{Var}(\theta W)|S)) + \mathrm{Var}(E(\theta W|S)) \\ &= E(\theta^2\mathrm{Var}(W)|S) + \mathrm{Var}(\theta E(W|S)) \\ &= E(\theta^2\mathrm{Var}(W)|S) + \mathrm{Var}_g(\theta(X)\Omega(X)) \end{aligned} \tag{3.9}$$

The latter term is the variance of the new estimate, and the first term is nonnegative and is positive unless $\Omega(X) = W(X)$ a.s. when $\theta \neq 0$.

Similarly, the asymptotic variances for the ratio and regression estimates are smaller when using conditional weights. If "*" denotes the use of the conditioned weights, then

$$\begin{aligned} \mathrm{AVar}(\hat{\mu}^*_{\mathrm{ratio}}) &= \mathrm{Var}_g(\Omega(\theta - \mu)) \\ &= \mathrm{Var}_g(W(\theta - \mu)) - E_g((\theta - \mu)^2\mathrm{Var}(W|S)) \\ &= \mathrm{AVar}(\hat{\mu}_{\mathrm{ratio}}) - E_g((\theta - \mu)^2\mathrm{Var}(W|S)) \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} \mathrm{AVar}(\hat{\mu}^*_{\mathrm{reg}}) &= \mathrm{Var}_g(\Omega(\theta - \beta^*)) \\ &\leq \mathrm{Var}_g(\Omega(\theta - \beta)) \\ &= \mathrm{Var}_g(W(\theta - \beta)) - E_g((\theta - \beta)^2\mathrm{Var}(W|S)) \\ &= \mathrm{AVar}(\hat{\mu}_{\mathrm{reg}}) - E_g((\theta - \beta)^2\mathrm{Var}(W|S)) \end{aligned} \tag{3.11}$$

We demonstrate the use of conditioned weights in two examples, the modeling of errors in digital communications, and a nuclear particle transport example.

3.1 Bit Error Rate Example

This example is motivated by the use importance sampling in the evaluation of digital communication systems (Davis 1986, Hahn & Jeruchim 1987). In digital systems voltage levels are subject to random fluctuation, and an error occurs when a function of some voltages exceeds a threshold. For simplicity, the voltage fluctuations are taken to have Gaussian distributions and the system to be linear. Importance sampling is done by increasing the power of the noise, that is by increasing the variance of the distributions of the random fluctuations. This was found to work poorly in higher-dimensional applications, and newer work investigates increasing the variance of only the appropriate linear combination of voltages. We consider the use of conditioned weights as another approach.

Let f and g be d-dimensional multivariate normal distributions with mean 0 and identity covariance matrices I and $\sigma^2 I$, and let $S = \sum c_i Z_i$, where Z_i are the components of the distribution, $X = (Z_1, Z_2, \ldots Z_d)$. Let $\theta(x) = I(|S| > T)$; θ is 1 if an error occurs (the noise exceeds the threshold T), otherwise 0. The expected value of θ under f is $2(1 - \Phi(T/\sqrt{C}))$, where Φ is a standard normal distribution function and $C := \sum c_i^2$. The weights are

$$W(X) = \sigma^{-d} e^{-\alpha Z_i^2} \tag{3.12}$$

$$\Omega(X) = E_g(W(X)|S) = \frac{1}{\sigma} e^{-\alpha S^2/C} \tag{3.13}$$

where $\alpha := (1 - 1/\sigma^2)/2$.

We consider only the integration estimate in this example, because this example has an ideal structure for that estimate—equivariance is not a consideration, θ is nonzero except when a rare event occurs and the example is simple enough that it is easy to choose a sampling distribution for which $W < 1$ whenever $\theta \neq 0$. The ratio estimate would perform worse, and the regression estimate offers little improvement.

The moments of the integration estimate can be calculated analytically, yielding:

$$E_g(W\theta) = E_g(\Omega\theta) = \mu \tag{3.14}$$

$$E_g(W^2\theta^2) = 2\sigma^d \gamma^{-d} \Phi(-T\gamma/\sqrt{C}) \tag{3.15}$$

$$E_g(\Omega^2\theta^2) = 2\sigma\gamma^{-1} \Phi(-T\gamma/\sqrt{C}) \tag{3.16}$$

where $\gamma = \sqrt{2 - 1/\sigma^2}$. The factor $\sigma\gamma^{-1}$ can be interpreted as a penalty for the variance of the weights, and is larger by d times for the unconditioned weights.

Choose $d = 10$, $c_i \equiv 1$, and $T = 13.97$ (i.e. $\sqrt{10}\Phi^{-1}(1 - 5 \cdot 10^{-5})$), so the true error rate is 10^{-4}. A naive Monte Carlo estimate would require 10^6 replications to reduce the standard error of the estimate to 10% of the true value. The variances with the two importance sampling methods, and sampling parameter σ chosen separately for each method at approximately the optimal values, are:

$$\begin{aligned} &\mathrm{Var}_f(\hat{\mu}_{\mathrm{mc}}) = 1.0^{-4} \\ &\mathrm{Var}_g(\hat{\mu}_{\mathrm{int},W}) = 7.7 \cdot 10^{-6} \quad (\sigma = 1.5) \\ &\mathrm{Var}_g(\hat{\mu}_{\mathrm{int},\Omega}) = 1.7 \cdot 10^{-7} \quad (\sigma = 4.0) \end{aligned}$$

The optimal σ is much smaller for the unconditioned weights, in order to avoid a large weight variance, which dominates the variance of the resulting estimate. Increasing σ results in more observations falling in the critical region, but at the cost of increased weight variance (both in and outside the critical region). The optimal tradeoff occurs at a smaller value for the unconditioned weights.

Figure 3.1 shows the estimated efficiency as a function of the sampling parameter σ for both conditional and unconditional weights.

Figure 3.2 is a plot of the weights vs S, 200 replications with $\sigma = 1.5$. Note that for a given value of S, the weight Ω is fixed but W has large variance, and that the conditional standard deviation of W is proportional to its expectation (which is Ω).

3.2 Nuclear Shielding Example

Probably the earliest use of importance sampling (Kahn 1950) was in estimating the probability that a given particle would exit a reactor wall without being absorbed. If the shield is sufficiently thick the probability is small, and importance sampling (also "biasing" in the nuclear literature) can be used to increase the efficiency of a Monte Carlo simulation of that probability by many orders of magnitude. Other references are Kahn and Marshal (1953), Booth (1986), Murthy and Indira (1986), and Clark, F. H. (1966).

Suppose that a particle is released at a point $(x_0, 0, 0)$, $x_0 < 0$, within a three-dimensional homogeneous shield bounded on the right by the plane

Figure 3.1: Efficiency in the Bit Error Rate Experiment

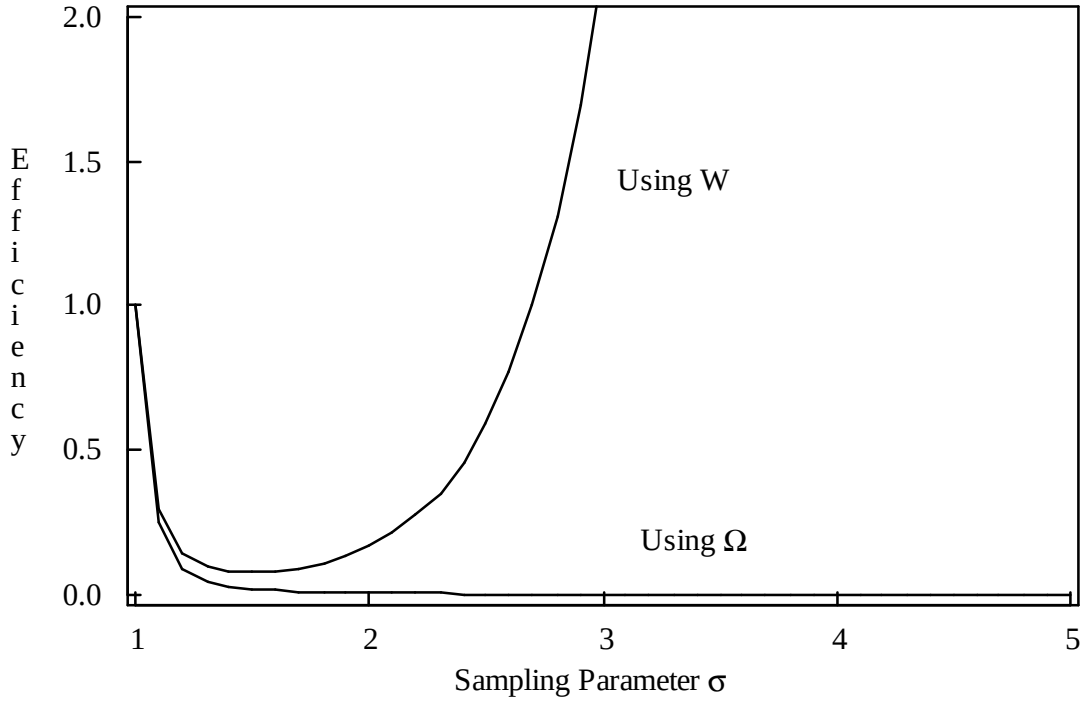
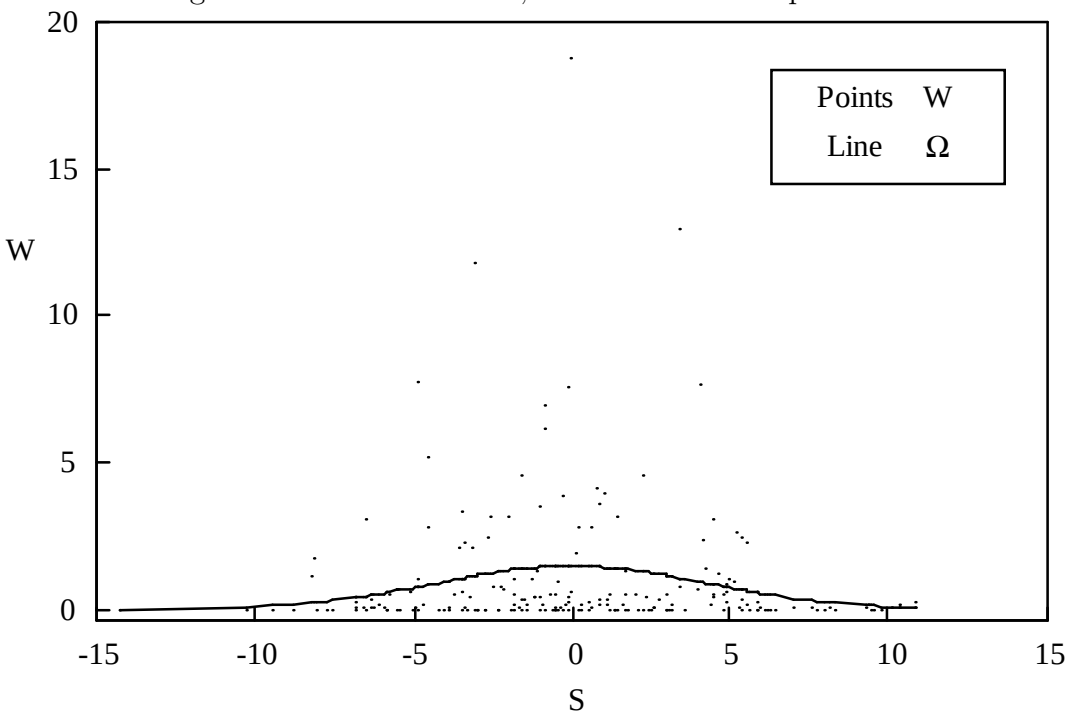
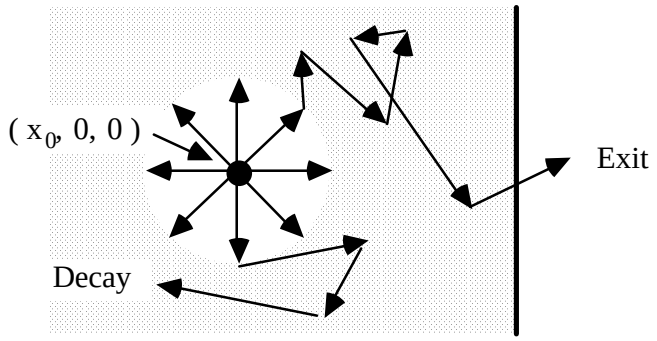


Figure 3.2: W and Ω vs S, Bit Error Rate Experiment



$x = 0$ and extending to infinity elsewhere. The particle travels a distance D_1 in an isotropic (uniform on the sphere) direction before it collides with the shield, where D_1 has an exponential distribution. The position of the particle is then (X_1, Y_1, Z_1). The particle is then absorbed with probability p, or is scattered in an isotropic direction and travels until the next collision. This continues until time T, when the particle is either absorbed or exits the shield ($X_T > 0$). Actual shield materials have different scattering properties, but we use the isotropic simplification in this example. With isotropic scattering the process (X_i, Y_i, Z_i) is a three-dimensional random walk.

Figure 3.3: Isotropic Scattering in a Shield



Monte Carlo simulation of this example involves generating sample paths for many particles (each particle is one replication) to estimate the probability of exit. We use the integration estimate in this example to estimate the probability that a particle exits the shield, so let $\theta = I(\text{exit})$. With the integration estimate it is permissible to have sampling distributions that assign 0 weights to outcomes that are necessarily zero; we follow Murthy and Indira (1986) in letting the sampling probability of absorption be zero (this requires a factor of $(1 - p)$ in the weight function) so that the simulation of a particle continues until it exits.

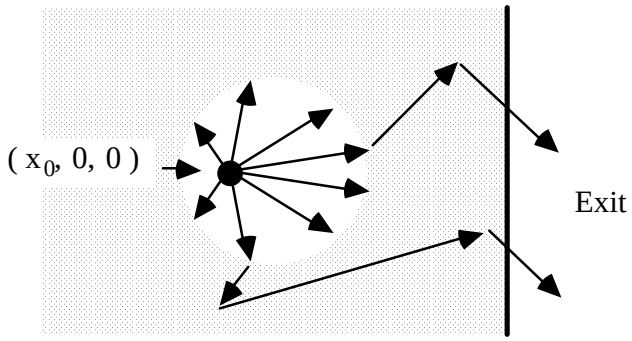
We use exponential tilting (exponential biasing), to obtain a sampling distribution for the one-step change in position:

$$g(x, y, z) = c_b e^{bx} f(x, y, z) \tag{3.17}$$

where $c_b = 2b/\log((1 + b)/(1 - b))$ is a normalizing constant, and $f(x, y, z)$

is the true distribution of a one-step change. This requires biasing both the direction and the distance of travel after a collision.

Figure 3.4: Importance Sampling Scattering



To generate variates from this distribution we first re-express (x, y, z) as (α, β, d), where $d := \sqrt{x^2 + y^2 + z^2}$, $\beta := \text{sgn}(y) \arccos(y/\sqrt{y^2 + z^2})$, and $\alpha := x/d$. d is the distance traveled, α is the cosine of the angle between the direction of travel and the direction $(1, 0, 0)$ of the closest exit point, and β is the angle in the (y, z) plane. Under f, D has an exponential distribution with mean 1, α is uniform on $(-1, 1)$ and β is uniform on $(-\pi, \pi)$.

The corresponding distributions for (α, β, d) under g are

$$g_\alpha(\alpha) = \frac{c_b I(-1 < \alpha < 1)}{2(1 - b\alpha)} \tag{3.18}$$

$$g_d(d|\alpha) = (1 - b\alpha) I(0 < d) e^{-d(1-b\alpha)} \tag{3.19}$$

$$g_\beta(\beta) = \frac{I(-\pi < \beta < \pi)}{2\pi} \tag{3.20}$$

To generate deviates, generate α first, then generate D from its conditional distribution, given α. β need not be generated at all, since it has no effect on the x coordinate of the particle.

It remains to choose b. A nearly optimal choice of b is based on p, and solves:

$$1 = \frac{c_b}{1 - p} = \frac{2b}{(1 - p) \log((1 + b)/(1 - b))} \tag{3.21}$$

This b can be found using a numerical search.

The reason for this choice of b is closely related to the idea behind conditional weights, that the weight assigned to a replication should depend only on a statistic when the outcome depends only on that statistic. In this example, the probability that a particle will exit without absorption depends only on the x coordinate of the particle, not the number and position of prior collisions. With this choice of b the inverse likelihood ratio for a single step, including the factor $(1-p)$ for the decay probability, is:

$$\begin{aligned} W(\Delta x, \Delta y, \Delta z) &:= f/g \\ &= \frac{1-p}{c_b e^{b\Delta x}} \\ &= e^{-b\Delta x} \end{aligned} \tag{3.22}$$

which depends only on Δx (the change in the x coordinate between two collisions, also write $\Delta x_t := x_t - x_{t-1}$), and the likelihood ration after T steps is

$$\begin{aligned} W_T &= \prod_{t=1}^{T} W(\Delta X_t, \Delta Y_t, \Delta Z_t) \\ &= \begin{cases} e^{-b(XT-x0)} & \text{if } X_T < 0 \\ \frac{1}{1-p} e^{-b(X_T - x_0)} & \text{otherwise} \end{cases} \end{aligned} \tag{3.23}$$

which depends only on the x-coordinate, not on the number or position of collisions prior to time T. Note that a sample path which returns to the plane $x = x_0$ has the same weight as it started with.

This sampling method and choice of b is closely related to the importance sampling scheme developed by Siegmund (1976) for exponential tilting in the study of sequential experiments.

This sampling distribution would be perfect, except for the problem of overshoot. A particle exits the shield as soon as its x coordinate is zero. In the sampling scheme we described the particle does not stop, but continues to the next collision (we are not concerned about the difference in collision rates inside and outside the shield). The weight W assigned to the particle at exit time T is

$$W(X_T, Y_T, Z_T) = \frac{1}{1-p} e^{bx_0} e^{-bX_T} \tag{3.24}$$

X_T is the overshoot; if it were zero then each particle (replication) would have the same weight function. In addition, each particle has the same value

of θ, so the product of the weight and simulation output ($Y = W\theta$ in our usual notation) would be a constant, e^{bx0}, thus the claim of perfection.

Note that the distribution of overshoot is almost insensitive to the thickness of the shield, for large shield thickness ($x_0 << 0$), since it depends only on the distribution of X_{T-1}; see Siegmund (1968) for this result in a similar example. Because of this the coefficient of variance of the estimate ($\sigma(\hat{\mu})/\mu$, where $\mu = P(\text{exit})$) is nearly independent of x_0. In contrast, the simple Monte Carlo estimate has a standard deviation proportional to the square root of the expected value ($\sigma(\hat{\mu}) = \sqrt{\mu(1-\mu)} \approx \sqrt{\mu}$), so the coefficient of variance explodes as $x_0 \to -\infty$ and $\mu \to 0$.

In the Monte Carlo experiment below, the exit probability is $2.0 \cdot 10^{-126}$ and the relative efficiency using importance sampling is $7.9 \cdot 10^{-126}$.

An improvement of 125 orders of magnitude is probably satisfactory. Still, for the sake of the argument and for what we learn that can be applied to other applications, we proceed to reduce this further.

Overshoot is the problem in this simulation (as is often the case in the study of random walks), particularly since the sampling distribution is strongly biased toward large steps in the positive x direction. For $p = 0.5$, for example, $b \approx 0.9575$. In the worst case, where $\alpha = 1$, the step size has an exponential distribution with mean value $1/(1-b) \approx 23.5$. The variance of the weights on exit is approximately 3.7 times the square of the average weight (simulation result). We attack this problem using the conditional weight method.

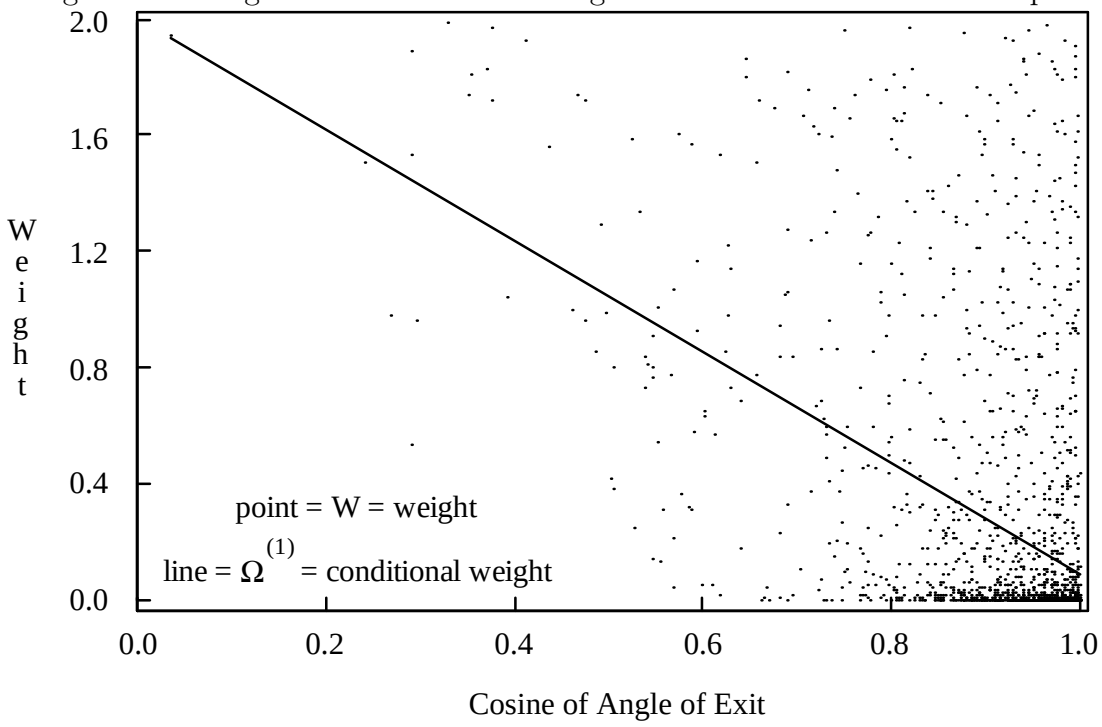
In conditioning the weights we use the same sampling scheme but compute weights differently once a particle exits. In the sequel T is the time of exit. Conditioned on X_{T-1} and α_T, the conditional weight is:

$$\begin{aligned} \Omega^{(1)} &:= E(W|\alpha_T, X_{T-1}) \\ &= \frac{1 - b\alpha_T}{1-p} e^{bx_0}. \end{aligned} \tag{3.25}$$

By using these weights the simulation reduces to the problem of finding the expected value of α_T, the angle of exit (when sampling from g). Figure 3.5 is a scatterplot of the weights from 2000 replications vs the exit angle, with the conditional weights represented by a line. It is apparent that there is considerable variance of the weights about their conditional expected value.

Can we do better than using the conditional weights defined above? Knowledge of the (unconditional) distribution of α_T would allow integration of (3.25) with respect to α_T, and we could replace $\Omega^{(1)}$ with a constant

Figure 3.5: Weights and Conditional Weights in Shield Penetration Example



weight and obtain a zero-variance estimate. That distribution is unavailable, but it is possible to use an inner Monte Carlo sampling loop to estimate

$$\Omega^{(2)} := E(W|X_{T-1}) \tag{3.26}$$

This is the integral of $\Omega^{(1)}$ with respect to the conditional distribution of α_T given X_{T-1}.

$$g_\alpha(\alpha|X_{T-1}) = \frac{C(X_{T-1})}{1-b\alpha} e^{X_{T-1}/\alpha} I(0<\alpha<1) \tag{3.27}$$

An analytical solution for $\Omega^{(2)}$ requires $C(X_{T-1})$, which in turn requires solving an intractable integral. It is possible, though, to generate random deviates from the distribution using the acceptance-rejection generation method (Knuth 1981). These deviates can be used in a small Monte Carlo experiment within each replication to obtain an estimate of $\Omega^{(2)}$.

The algorithm is:

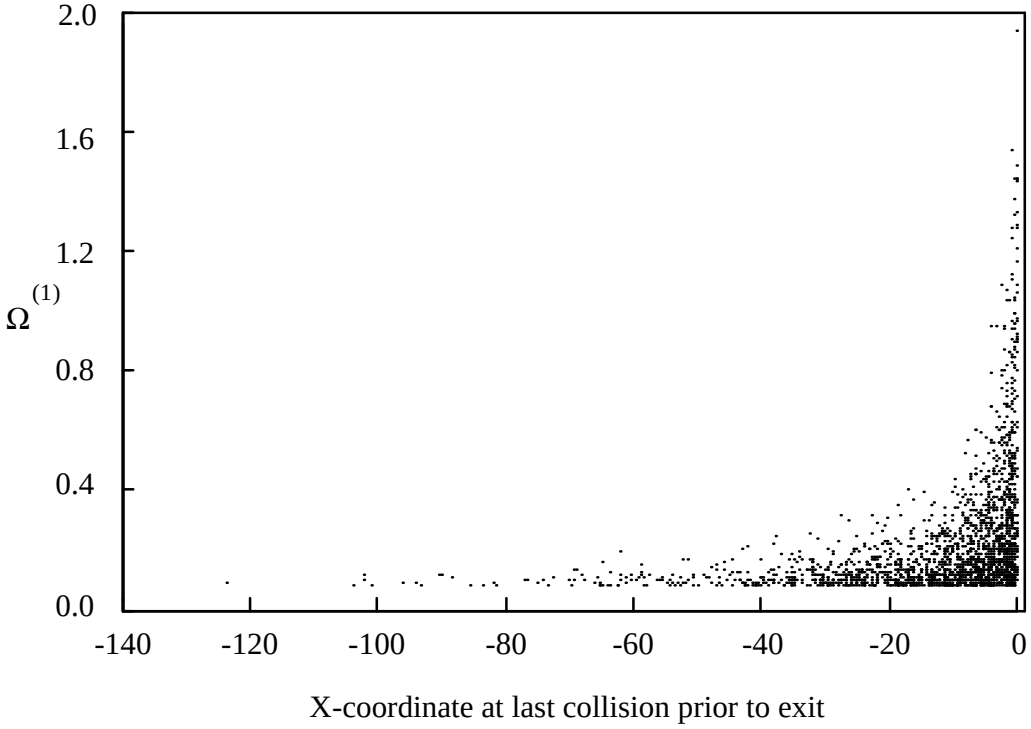
(1) Repeat for $i = 1, 2, \ldots n$ replications
  (1.1) Generate collisions until exit, save $\alpha_T^{(1)}$ and X_{T-1}
  (1.2) $W_i := \frac{1}{1-p} e^{bx_0} e^{X_T}$
  (1.3) $\Omega_i^{(1)} := \frac{1}{1-p} e^{bx_0}(1 - b\alpha_T^{(1)})$
  (1.4) Generate $\alpha_T^{(j)}$, $j = 2, 3, \ldots m \sim g_\alpha(\alpha|X_{T-1})$
  (1.5) Let $\hat{\Omega}_i^{(2)} := \frac{1}{1-p} e^{bx_0} \frac{1}{m} \sum(1 - b\alpha_T^{(j)})$
(2) $\hat{\mu}^{(1)} := n^{-1} \sum_{i=1}^n W_i$
(3) $\hat{\mu}^{(2)} := n^{-1} \sum_{i=1}^n \Omega_i^{(1)}$
(4) $\hat{\mu}^{(3)} := n^{-1} \sum_{i=1}^n \hat{\Omega}_i^{(2)}$

The use of the second, inner, loop (steps 1.4 and 1.5) is common in conditional Monte Carlo estimation when an expected value cannot be computed analytically, see e.g. Bratley, Fox & Schrage (1983). Figures 3.6 and 3.7 are scatterplots of the conditional weights $\Omega^{(1)}$ and $\hat{\Omega}^{(2)}$ against X_{T-1}. The inner sampling loop gives $\hat{\Omega}^{(2)}$ a smaller variance than $\Omega^{(1)}$ for any given value of X_{T-1}.

The variance of the estimate using the estimated conditional weights is a combination of the variance of $\Omega^{(2)}$ (which is a function of X_{T-1}) and the expected conditional variance of $\hat{\Omega}^{(2)}$.

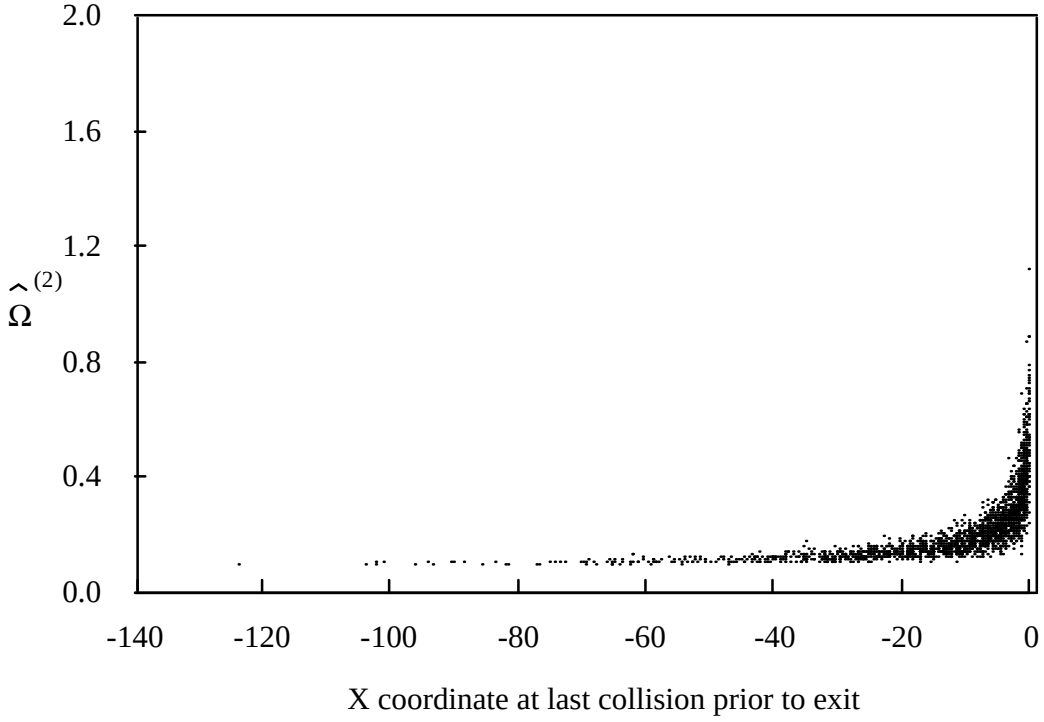
$$\begin{aligned} \mathrm{Var}(\hat{\mu}^{(3)}) &= \frac{1}{n}\left(\mathrm{Var}(\Omega^{(2)}) + E(\mathrm{Var}(\hat{\Omega}_i^{(2)}|X_{T-1}))\right) \\ &= \frac{1}{n}\mathrm{Var}(\Omega^{(2)}) + \frac{1}{mn} E(\mathrm{Var}(\hat{\Omega}_i^{(2)}|X_{T-1}, m=1)) \end{aligned} \tag{3.28}$$

Figure 3.6: Conditional Weights $\Omega^{(1)}$ in Shield Penetration Example



Choosing $m = 1$ is equivalent to not doing conditional estimation (equivalent in this context to not doing the second conditional estimate). $m > 1$ gives better accuracy. Note that the variance is greater than if n were increased to $n_2 := mn$, but that requires doing both the outer and inner loops mn times. Choosing $m > 1$ is advantageous here since an inner loop does not require generating the path prior to X_{T-1}. In this experiment a choice of $m = 10$ resulted in a overall variance 35% as large as the choice $m = 1$.

Figure 3.7: Conditional Weights $\hat{\Omega}^{(2)}$ in Shield Penetration Example



3.3 Conditioning for Multivariate Applications

In a multivariate application different weights may be used for different output quantities. In particular, any output quantity that depends only on a subset of the input values can be estimated using weights computed only

Table 3.1: Particle Scattering Experiment

	Estimate of Exit Probability $\cdot e^{-300b}$	Variance of Estimate $\cdot ne^{-600b}$	Efficiency Relate to MC $\cdot e^{-300b}$
W	0.241 (0.0104)	0.214 (0.011)	0.89 (0.020)
$\Omega^{(1)}$	0.226 (0.0045)	0.041 (0.012)	0.18 (0.006)
$\hat{\Omega}^{(2)}$	0.227 (0.0027)	0.014 (0.008)	0.063 (0.003)

Relative Efficiency

	MC	W	$\Omega^{(1)}$
W	0.89 e^{-300b}	1	
$\Omega^{(1)}$	0.18 e^{-300b}	0.20 (0.014)	1
$\hat{\Omega}^{(2)}$	0.063 e^{-300b}	0.071 (0.004)	0.63 (0.02)

$n = 2000$ replications, $x_0 = -300$, $P(\text{decay}) = 0.5$, $b = 0.957504$, isotropic scattering, exponential flight distances. The three estimates are obtained using the unconditioned weights W, weights $\Omega^{(1)}$ conditioned on X_{T-1} and α_T, and estimated weights $\hat{\Omega}^{(2)}$ conditioned on X_{T-1} using $m = 10$ inner replications. Standard errors of estimates are in parenthesis. Note that exit probabilities, variances, and efficiency relative to MC are very small; the numbers in the table must be multiplied by a factor of e^{-300b} or e^{-600b} to get the actual values.

from those input values.

This is equivalent to conditioning the weights for any output value on a sufficient statistic for that output value, the sufficient statistic being the input values that contribute to that output. The result is a more efficient estimate for that output.

In the fuel inventory example which follows in Chapter 4 the temperature in December is independent of the nuclear generation in March. To estimate the distribution of temperature or heating degree days, it is more efficient to use weights computed solely based on the temperature for the month in question than to use weights based on all input quantities.

Of course the use of different weights for different quantities is not free—it requires that different sets of weight be computed. It also can lead to inconsistencies in the output; e.g. the estimated expected value of $\theta_1 + \theta_2$ may not be equal to the sum of the individual estimates. These factors must be considered in each application before conditioning.

An example of conditioning is given in Tables 3.2 and 3.3. Here a number of quantities for January are computed using weights computed from random variables through January. This is not a full-fledged example of the conditioning principle, because a single set of weights is used for all quantities, rather than separate sets for temperature, nuclear power, etc. However it does illustrate the principle. The estimated increase in efficiency (Table 3.2 below) achieved using the three-month weights rather than the five month weights ranges between 5% and 41% (relative to the five-month weight efficiency), with an average of 12%. Results are better, though not overwhelmingly so.

Table 3.2 contains estimates of the efficiency of importance sampling using the two weights. The standard error given in this table is the standard error of the efficiency estimate. Note that the efficiency estimates are better for the three-month weights.

Table 3.3 contains the actual estimates, and the estimated standard errors of those estimates. Note that the estimated standard errors are lower for the three-month weights.

Table 3.2: Estimated Efficiency in the Fuel Inventory Example, Three- and Five-Month Weights

	Three-month Weights		Five-Month Weights	
Integration Efficiency	est	std. error	est	std. error
Jan. Oil Inventory	4.688	0.144	5.080	0.161
Jan. Gas Demand	93.949	4.115	116.533	5.053
Jan. Electric Demand	87.628	3.812	108.404	4.650
Jan. Temperature	36.951	1.466	44.210	1.748
Jan. Hydro	4.728	0.091	5.044	0.106
Jan. Nuclear	3.743	0.171	4.259	0.197
Nov-Jan Outage Prob.	0.445	0.029	0.526	0.028

	Three-month Weights		Five-Month Weights	
Ratio Efficiency	est	std. error	est	std. error
Jan. Oil Inventory	0.803	0.019	0.925	0.015
Jan. Gas Demand	1.270	0.024	1.372	0.019
Jan. Electric Demand	1.282	0.025	1.352	0.019
Jan. Temperature	1.269	0.024	1.366	0.021
Jan. Hydro	1.414	0.025	1.484	0.021
Jan. Nuclear	1.179	0.027	1.270	0.022
Nov-Jan Outage Prob.	0.531	0.029	0.602	0.025

	Three-month Weights		Five-Month Weights	
Regression Efficiency	est	std. error	est	std. error
Jan. Oil Inventory	0.387	0.019	0.651	0.025
Jan. Gas Demand	1.251	0.025	1.362	0.020
Jan. Electric Demand	1.268	0.025	1.344	0.019
Jan. Temperature	1.192	0.026	1.316	0.022
Jan. Hydro	1.119	0.021	1.293	0.017
Jan. Nuclear	1.155	0.031	1.255	0.027
Nov-Jan Outage Prob.	0.389	0.030	0.495	0.031

Table 3.3: Estimated Expectations and Standard Errors in the Fuel Inventory Example, Three- and Five-Month Weights

	Three-month Weights		Five-Month Weights	
Integration Estimate	est	std. error	est	std. error
Jan. Oil Inventory	319.013	6.091	320.141	6.340
Jan. Gas Demand	1767.473	22.100	1778.065	24.613
Jan. Electric Demand	1724.018	21.600	1733.865	24.025
Jan. Temperature	51.272	0.689	51.592	0.753
Jan. Hydro	623.515	12.591	625.437	13.005
Jan. Nuclear	222.262	3.721	223.349	3.970
Nov-Jan Outage Prob.	0.057	0.003	0.055	0.004

	Three-month Weights		Five-Month Weights	
Ratio Estimate	est	std. error	est	std. error
Jan. Oil Inventory	324.640	2.521	323.893	2.706
Jan. Gas Demand	1798.651	2.569	1798.905	2.670
Jan. Electric Demand	1754.429	2.613	1754.187	2.683
Jan. Temperature	52.177	0.128	52.197	0.132
Jan. Hydro	634.514	6.887	632.767	7.054
Jan. Nuclear	226.183	2.089	225.967	2.167
Nov-Jan Outage Prob.	0.058	0.004	0.056	0.004

	Three-month Weights		Five-Month Weights	
Regression Estimate	est	std. error	est	std. error
Jan. Oil Inventory	325.136	1.749	325.136	2.269
Jan. Gas Demand	1798.712	2.550	1798.712	2.660
Jan. Electric Demand	1754.010	2.599	1754.010	2.675
Jan. Temperature	52.218	0.124	52.218	0.130
Jan. Hydro	634.899	6.126	634.899	6.586
Jan. Nuclear	226.162	2.068	226.162	2.155
Nov-Jan Outage Prob.	0.055	0.003	0.055	0.004

Chapter 4

Some Examples

This chapter includes discussion of four general classes of applications where importance sampling is useful. The first of these is the traditional variance reduction application, in the "easy" case where the goal is to estimate the expected value of a "lardimaz" variable (one which has a large discrete mode at zero). The second also involves variance reduction, but for more general distributions, or when the simulation has multiple output variables only some of which are lardimaz.

The last two classes fall outside the traditional variance reduction framework. These can be described as applications where importance sampling is used because it is the only practical choice, though here, also, some attention to efficient estimates is worthwhile.

The third class includes applications where the true distribution can not be generated. This includes examples in Bayesian analysis, where distribution estimation must be performed with respect to an intractable posterior distribution (Stewart 1979, 1983, Kloek and van Dijk 1978, van Dijk and Kloek (1983), and the analysis of characteristic roots of a random covariance matrix (Luzar and Olkin 1988). Importance sampling can thus be used to solve problems that could not otherwise be solved.

The final class of applications is where there is more than one "true" distribution. Rather than running a separate simulation for each distribution, importance sampling can be used to estimate results from many distributions in a single simulation (Beckman & McKay 1987, Tukey 1987). The number of such distributions may be infinite, as the bootstrap tilting interval (Tibshirani 1984), or in response surface estimation (Glynn & Iglehart 1987). Importance sampling can also be used for derivatives of expectations with

respect to parameters of input distributions (Reiman & Weiss 1986, Glynn 1986).

4.1 Rare-Event Applications with Mode Zero

Most successful applications of importance sampling have been where the distribution of θ has a large discrete mode (probability close to one) at zero, a "lardimaz" distribution.

This is a large class of applications. It includes the obvious case, where θ is a physical quantity such as an outage magnitude when outages rarely occur. It also includes the estimation of rare event probabilities; the probability is the expectation of the indicator function of the event $\theta = I(\text{event})$, which is nearly always zero.

This class of applications is well-suited for importance sampling, and for the integration estimate in particular. If a sampling distribution can be found for which the relative likelihood g/f is greater than one whenever the rare event occurs, the integration estimate gives reduced variance. In addition, for the integration estimate g can be zero when θ is 0 (weak dominance suffices). This allows a natural efficiency improvement in applications such as estimating a small probability, by not sampling at all from regions where the event cannot occur.

The transformation interpretation becomes particularly simple in this case. In general the integration estimate involves an induced transformation $\theta \to Y = \theta f/g$, where g is chosen so that the transformed quantity is more constant. This can be tricky in some other examples, as we discuss in the next section; small θ values can be made into extremely large Y values by an unfortunate transformation. The lardimaz case is easy, in that zero θ values are never transformed into anything other than zero—the sampling distribution can be chosen with impunity to concentrate on the nonzero values, ignoring the zero values.

Thus, in an application with a discrete mode at zero the transformation principle of choosing a sampling distribution for its induced transformation coincides with the sampling goal of sampling heavily when θ is extreme. The preponderance of applications of this class in traditional importance sampling may be why the distinction between the two goals is not more widely known.

An additional consideration in many examples of this class is that equivariance is not a consideration. Thus one argument against the integration

estimate in general applications does not hold here. To estimate a small probability, for example, is equivalent to estimating the expected value of a function that takes on two values, zero and one; it is not necessary that the estimate increase by c if the function is increased by c, since the function is limited to two particular values.

The class of lardimaz applications includes the nuclear transport example considered by Kahn (1950), Kahn and Marshall (1953), Booth (1986), Murthy and Indira (1986) and others. Other areas include reliability estimation in the fields of digital communications (Davis 1987, Hahn & Jeruchim 1987), fault-tolerant computers (Conway and Goyal 1987, Goyal et al. 1987, Kioussis and Miller 1983), and engineering analysis, the simulation of stochastic processes (Moy 1986, Glynn and Iglehart1987), study of sequential tests (Siegmund 1976), and implementation of bootstrap confidence intervals (Johns 1987).

Applications with a large discrete mode at some known constant $c \neq 0$ can be transformed prior to computing the importance sampling estimate by subtracting the value of the mode, then adding the value to the computed estimate (Therneau 1983). The estimate is then:

$$\hat{\mu}_c = c + n^{-1} \sum_{i=1}^{n} (W_i(\theta_i - c)) \tag{4.1}$$

This is equivalent to using a regression estimate with β fixed at c. We will not call this a regression estimate, however, to avoid confusion. The transformed integration estimate behaves more like an integration estimate than like a true regression estimate (with estimated slope)—it is unbiased, is not equivariant, and is a weighted average with weights that do not add to 1.

The true regression estimate, with $\hat{\beta}$ estimated from the sample, can be used to automate the process of transformation to zero mode. Note, however, that the regression estimate requires that g dominate f to be consistent, and requires that W be bounded or have a light tail for good finite sample performance.

Other examples can sometimes be converted into lardimaz examples. Goyal et al. (1987) simulate the steady-state availability of a fault tolerant computer, using regenerative simulation. The results obtained in a regenerative simulation can be expresses as the ratio of two expectations,

$$R = \frac{E(C)}{E(T)} \tag{4.2}$$

where T is a random variable representing the time between regeneration states (where the future of a stochastic process is conditionally independent of the past), and C is a random variable representing the integral of some measure of system performance during the time between regenerations. In this case C is the number of times a computer fails during a regeneration cycle, and C is lardimaz. The natural way to estimate R is by the ratio of two estimates

$$\hat{R} = \frac{\hat{E}(C)}{\hat{E}(T)}. \tag{4.3}$$

In this case the same result is obtained by using either the integration or ratio estimate to estimate both numerator and denominator, and the influence of a pair (T, C) is $(C - RT)/E(T)$, which is not lardimaz. They estimate the two expectations independently, using different sampling distributions for each, in order to make effective use of lardimaz properties in estimating the numerator.

We will discuss some of these examples in connection with different importance sampling techniques, rather than discussing all here. For now we consider only quantile estimates for tails of distributions, show that this is a lardimaz example, and discuss the application of this to the computation of bootstrap confidence intervals.

4.1.1 Importance Sampling for Quantiles

One of the advantages of Monte Carlo simulation is that it provides information about the whole distribution of output quantities, in addition to expectation estimates. Importance sampling can be particularly useful in the estimation of extreme quantiles of distributions. We shall see that this example fits into the lardimaz framework.

Estimates of quantiles using importance sampling are generally straightforward. If $V_i, i = 1, 2, \ldots n$ are the weights from any equivariant estimate, form the empirical distribution $\hat{F}_\theta$ by placing weight V_i on the observed value θ_i,

$$\hat{F}_\theta(x) := \sum_{i=1}^{n} V_i I(\theta_i \leq x) \tag{4.4}$$

Quantiles are then defined as, e.g.

$$\hat{\xi} = \hat{F}_\theta^{-1}(\alpha) := (\min\{x : \hat{F}_\theta(x) \geq \alpha\} + \max\{x : \hat{F}_\theta(x) < \alpha\})/2 \tag{4.5}$$

If the integration estimate weights are used then modifications of these distribution and quantile estimates is in order. Remember that the integration estimate weights do not sum to 1, and unmodified use of (4.4) and (4.5) could lead to undefined estimates for extreme quantiles. For quantiles for $\alpha > 0.5$, use the quantile estimate

$$\hat{F}^{-1}_{\theta,\mathrm{up}}(\alpha) := (\min\{x : \hat{F}_{\theta,\mathrm{up}}(x) \geq \alpha\} + \max\{x : \hat{F}_{\theta,\mathrm{up}}(x) < \alpha\})/2 \qquad (4.6)$$

where

$$\hat{F}_{\theta,\mathrm{up}}(x) := \sum_{i=1}^{n} V_i I(\theta_i \leq x) + 1 - \sum_{i=1}^{n} V_i \qquad (4.7)$$

is the "upper" distribution function. For $\alpha \leq 0.5$, use the usual distribution function and its inverse.

If the distribution of θ has a nonzero density $f_\theta(\xi)$ at $\xi := F^{-1}_\theta(\alpha)$ and $E_g(W^2) < \infty$, then the limiting distribution of a quantile estimate is asymptotically normal,

$$\sqrt{n}(\hat{\xi}_{\mathrm{est}} - \xi) \to N(0, \tau^2) \qquad (4.8)$$

where

$$\tau^2 = \frac{\mathrm{AVar}(\hat{\mu}_{\mathrm{est}})}{f_\theta(\xi)^2} \qquad (4.9)$$

and $\mathrm{AVar}(\hat{\mu}_{\mathrm{est}})$ is the variance of the asymptotically normal distribution of $\hat{F}_{\theta,\mathrm{est}}(\xi)$ obtained using estimate "est" (e.g. integration, ratio, etc.). Johns (1987) gives a proof for the integration estimate under the condition that W be a single-valued function of θ. We give a proof in the appendix for the more general case that W is a function of X, and outline a proof for the ratio and regression estimates. Note that results can not be improved by choosing a sampling density which is larger at the quantile, as the denominator of 4.9 is a function only of f, not g.

This result implies that good performance in estimating a quantile is obtained by estimating $\hat{F}_{\theta,\mathrm{est}}(\xi)$ well at that quantile. Thus estimating extreme quantiles involves solving a lardimaz problem, of estimating small probabilities. This is useful, for example, in the computation of bootstrap confidence intervals.

4.1.2 Bootstrap Percentile Interval

Recall from Chapter 1 that a bootstrap distribution consists of a set of n points $T_1, T_2, \ldots, T_n$, where $T_i = T(X_i)$. The bootstrap *percentile interval*

(Efron 1982) is

$$[T_{(k)}, T_{(n-k+1)}], \tag{4.10}$$

where $T_{(k)}$ is the k'th order statistic of the set of bootstrap results $T_1, T_2, \ldots T_n$, and $k = n\alpha$. This produces an approximate two-sided $1-2\alpha$ confidence interval. If $n\alpha$ is not an integer it can either be rounded or interpolated between the two nearest order statistics.

Two variations on the bootstrap percentile interval are the "bias-corrected percentile interval" (BC) and the BC_a intervals (Efron 1982, 1987). First re-express the standard percentile interval as

$$[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)], \tag{4.11}$$

where $\hat{G}$ is the empirical distribution function for $T(X)$, based on n bootstrap replications. The BC interval is:

$$[\hat{G}^{-1}(\Phi(z_\alpha + 2z_0)), \hat{G}^{-1}(\Phi(z_{1-\alpha} + 2z_0))] \tag{4.12}$$

and the BC_a interval is:

$$[\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))] \tag{4.13}$$

where

$$z_\alpha = \Phi^{-1}(\alpha), \tag{4.14}$$

$$z_0 = \Phi^{-1}(\hat{G}(T(\hat{F}))), \tag{4.15}$$

$$z[\alpha] = z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}, \tag{4.16}$$

and a is a measure of the skewness of an application.

The intent of these two methods is to correct for biased statistics (the BC interval) or statistics which are both biased and for which the variance is a simple function of the statistic (BC_a). The BC interval is correct if there exists some transformation h such that

$$h(T(Z)) - h(T(F)) \sim N(-z_0, 1), \tag{4.17}$$

when $Z = (Z_1, \ldots Z_d)$ and $Z_j \overset{i.i.d.}{\sim} F$, for all F. That is, T has bias z_0, measured on the scale defined by the transformation h. Note that the transformation does not need to be known for the interval to be computed. The BC_a interval is based on the slightly more general formulation

$$h(T(Z)) - h(T(F)) \sim N(-z_0, (1 + ah(T(F)))^2) \tag{4.18}$$

Computation of these intervals requires accurate estimates of the appropriate quantiles—in the tails of the distributions for all intervals, and near the median for z_0. These quantiles can be hard to estimate. For example, in computing a lower one-sided 99% confidence bound for the percentile interval without the use of variance-reduction techniques, the bound corresponds to the first (1st) percentile of the distribution. To ensure (with 95% confidence) that the estimated quantile does not in fact correspond to the second (2nd) percentile requires approximately 800 replications. The same level of discrimination between the .5% and 1% levels requires approximately 1,600 replications, and between the 1% and 1.1% percentiles approximately 44,000 replications. Importance sampling can be a great help here.

The bias-correction constant z_0 is similarly hard to estimate accurately. The problem is not the lack of observations in the region of the value to be estimated (at the median of the distribution) but rather the way in which z_0 is used—a small change in z_0 has a magnified effect on the confidence interval. Unfortunately, importance sampling offers little improvement here, but there is an analytical approximation that can be used instead of a Monte Carlo estimate.

Define the empirical influence function as

$$U_j = \lim_{\Delta \to 0} \frac{T((1-\Delta)\hat{F} + \Delta\delta_j) - T(\hat{F})}{\Delta}. \tag{4.19}$$

δ_j is a point mass on X_j, and T must be a smooth function near $\hat{F}$ (Efron 1982, 1987).

In the nonparametric bootstrap

$$a = \frac{1}{6} \frac{\sum_{j=1}^{d} U_j^3}{\left(\sum_{j=1}^{d} U_j^2\right)^{3/2}} \tag{4.20}$$

Define, as well, the second-order empirical influence function as

$$V_{ij} = \lim_{\Delta \to 0} \frac{T((1-2\Delta)\hat{F} + \Delta\delta_i + \Delta\delta_j) - T(\hat{F}) - \Delta U_j - \Delta U_i}{\Delta^2} \tag{4.21}$$

The analytical approximation for z_0 is

$$z_0 = a + \frac{1}{2} \frac{\mathbf{U}'\mathbf{V}\mathbf{U}}{d|U|^3} - \frac{1}{2} \frac{\mathrm{tr}(\mathbf{V})}{d|U|}, \tag{4.22}$$

where d is the size of the original sample, $\mathrm{tr}(\mathbf{V})$ is the trace of matrix $\mathbf{V}$, and $|\mathbf{U}| = \sqrt{\mathbf{U}'\mathbf{U}}$. This approximation is based on an Cornish-Fisher expansion for the distribution of $T(\mathbf{X}^*)$, using Edgeworth approximations for the moments. It was obtained independently by Efron (1987) by another method.

To evaluate $\mathbf{U}$ and $\mathbf{V}$ use a small-delta method. Choose ϵ small, and let

$$H_j := \frac{T((1-\epsilon)\hat{F} + \epsilon\delta_j) - T(\hat{F})}{\epsilon}. \tag{4.23}$$

The H_j values give both the estimate of $\mathbf{U}$ and of the trace of $\mathbf{V}$

$$\hat{U}_j := H_j \bar{H}, \tag{4.24}$$

$$\hat{\mathrm{tr}}(\mathbf{V}) := \frac{2}{\epsilon}\sum_{j=1}^{d} H_j. \tag{4.25}$$

Now rather than using n^2 evaluations of the function $T(\cdot)$ to evaluate the term $\mathbf{U}'\mathbf{V}\mathbf{U}$, we need only two, at $(1-\epsilon)\hat{F} + \epsilon\mathbf{U}$ and $(1-\epsilon)\hat{F} - \epsilon\mathbf{U}$, where $\epsilon\mathbf{U}$ is interpreted as a signed measure $\epsilon\sum U_j\delta_j$. The estimate is

$$\mathbf{U}'\mathbf{V}\mathbf{U} \approx \frac{T((1-\epsilon)\hat{F} + \epsilon\mathbf{U}) + T((1-\epsilon)\hat{F} - \epsilon\mathbf{U}) - 2T(\hat{F})}{\epsilon^2}. \tag{4.26}$$

With a and z_0 out the way, we turn to the application of importance sampling in the evaluation of the extreme percentiles of the bootstrap distribution.

4.1.2.1 Importance Sampling for Percentile Intervals Johns (1987) uses importance sampling in the evaluation of bootstrap confidence intervals. X is a vector of length d with independent components under f, with equal probabilities $1/d$ of taking any of the values in the original set of data Z, which is also a vector of length d,

$$P_f(X_j = Z_k) = 1/d. \tag{4.27}$$

The sampling distribution also has *i.i.d.* components, with

$$P_g(X_j = Z_k) = \frac{\exp(\beta U_k)}{\sum_{l=1}^{d}\exp(\beta U_l)} \tag{4.28}$$

where U_k is the empirical influence function defined in (4.19).

To estimate the α-quantile of the bootstrap distribution of $T(X)$, the tilting parameter β is chosen so that

$$E_g(U(X_j)) = \frac{z_\alpha}{\sqrt{d}} S, \tag{4.29}$$

where S^2 is the sample variance of the U_k and z_α is the α percentile of a normal distribution. This recenters the sampling distribution at approximately the quantile to be estimated.

Johns indicates that this method results in savings of a factor of 10 or more in the number of bootstrap replications necessary to provide accurate confidence intervals.

4.2 Examples Without a Mode at Zero

The behavior of importance sampling is very different in applications without a large mode at zero than it is in lardimaz examples. In particular, the transformation $\theta \to Y = \theta f/g$ induced by the sampling distribution in the integration estimate requires not only that the sampling distribution sample relatively frequently when θ is nonzero, but rather that the sampling distribution be precisely calibrated to transform small θ values into larger Y values, while avoiding Y values which are too large. That transformation is sensitive to small changes in the relative likelihood g/f when the relative likelihood is near zero.

If the sampling distribution is not precisely calibrated large variance increases can result. An illustration of this is given in Example 4.1.

Example 4.1 Calibration of sampling distributions for transformations θ takes on values (990, 999, 1000, 1001, 1010) with probabilities (.1, .2, .4, .2, .1). Three sampling distributions are considered, given in Table 4.1.

g_1 represents a finely calibrated sampling distribution for the integration estimate. g_2 is poorly calibrated, making too much of a transformation. g_3 is a reasonable sampling distribution for the ratio estimate, in that it samples relatively often from extreme values, but it has a terrible effect on the integration estimate. The asymptotic efficiency for the three distributions, relative to no importance sampling, and for the three estimates is given in Table 4.2.

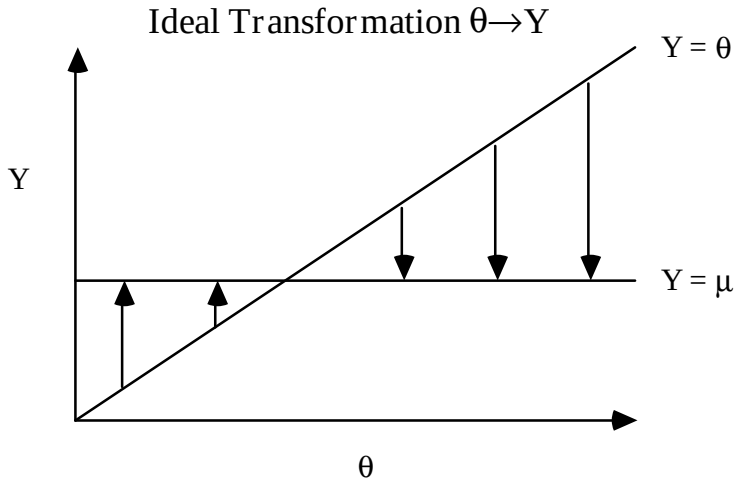


Table 4.1: True and Sampling Distributions in Example 4.1

$\theta(x)$	$f(x)$	$g_1(x)$	$g_2(x)$	$g_3(x)$
990	0.1	$0.998f$	$0.98f$	0.2
999	0.2	$0.999f$	$0.99f$	0.2
1000	0.4	$1.000f$	$1.00f$	0.2
1001	0.2	$1.001f$	$1.01f$	0.2
1010	0.1	$1.002f$	$1.02f$	0.2

Table 4.2: Efficiency for Example 4.1
Var(estimate)/Var(simple random sampling)

	f	g_1	g_2	g_3
Integration	1	0.63	2.57	1470.6
Ratio	1	1.00	1.00	0.51
Regression	1	0.21	0.21	0.51

The problem of sensitivity to the transformation is compounded when factors other than the transformation affect the choice of the sampling distribution, including ease of implementation and the need to sample for more than one output quantity or more than one "true" distribution in a single experiment.

Yet importance sampling should not be avoided altogether in these cases. What is needed is a more robust importance sampling methodology. This can be found by avoiding the integration estimate, and by using "mixture sampling" (discussed in Chapter 6) in the sampling distribution. These methods are demonstrated in the fuel inventory example, discussed next.

4.2.1 Fuel Inventory Example

The genesis of this work on importance sampling was the work the author performed for the Fuel Inventory Probabilistic Simulator (FIPS) at Pacific Gas and Electric Company (Hesterberg, 1987). FIPS is a large, complex Monte Carlo model designed to evaluate electric power plant fuel inventory operating policies, in particular, how much oil to carry in inventory at the start of a winter.

This model is representative of the kind of difficult Monte Carlo application to which importance sampling could be applied, but where it traditionally has not been. A simplified version of this model will be used throughout this work to demonstrate a number of importance sampling concepts, in choice of estimates, sampling distribution choice, and conditional weights.

The FIPS model produces hundreds of output quantities, including expected values, standard deviations, and percentiles of inventory carrying cost, outage cost, purchase costs for different fuels, inventory levels, and total cost. Some of these fit into the rare-event discrete-zero-mode framework in which importance sampling is traditionally applied; most do not, and the use of the ratio or regression estimates is required to give acceptable answers for these quantities.

There are hundreds of input random variables, the most important of which are weather, nuclear generation, and hydroelectric generation. Some of these are continuous, others are discrete, and many are correlated. There is none of the mathematical simplicity which makes the choice of sampling distributions easy in some applications. A combination of two sampling strategies is used. Mixture sampling provides a measure of robustness to the estimation of all output quantities while allowing greatly improved estimates

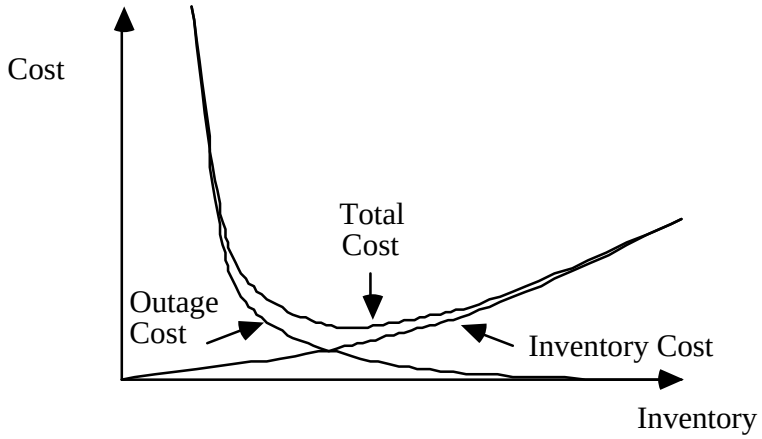
for some. Internal sampling distributions are used as a reasonable way to generate sampling distributions in the face of the complexity of the example. Both are discussed in Chapter 6.

FIPS was designed primarily to answer a single question—how much oil should PG&E carry in inventory at the start of a winter (the model has proven versatile enough to be used for other applications). The amount of oil required depends on a number of random factors, most notably temperature and the levels of hydroelectric and nuclear generation. Most years no oil is burned, other than small test burns, since oil is a fuel of last resort—hydroelectric, nuclear, geothermal, wind, and solar power, purchases from the northwest, and burning natural gas are all cheaper and cleaner than burning oil.

Fuel-related outages are very unlikely events, occurring only under a combination of unfavorable conditions severe enough to cause all available fuel sources to be exhausted. But, if such an outage does occur it is likely to continue for some time, resulting in a much larger total curtailment than in a typical storm-related outage. There is a high priority placed on avoiding such outages, by storing adequate reserves of fuel.

Still, carrying fuel in inventory is costly, and a reasonable trade-off must be found between inventory costs and outage costs. Increasing the inventory results in higher inventory costs and lower outage costs, as shown in Figure 4.1.

Figure 4.1: Inventory / Outage Cost Tradeoff



A Monte Carlo model is a natural way to handle this application. The key input values are random variables, and the output is random. A Monte Carlo model can handle the complex operating policies and interaction of the gas and electric systems, and can produce outage and operating cost distributions.

Running the model at multiple inventory levels gives an idea of the cost distributions for different inventory policies. These distributions can be used to find operating policies which minimize expected total cost, or which maximize a more general utility function.

The total cost is a convex function of the initial oil inventory, and the minimum expected cost occurs when the derivative of that curve is zero. That derivative is approximately the sum of the derivatives of the inventory cost and outage cost curves (other costs are only marginally affected by the initial oil inventory), which in turn can be approximated by

$$\frac{\partial \text{inventory cost}}{\partial \text{initial inventory}} \approx (P(\text{no outage}))(\text{inventory price}), \tag{4.30}$$

and

$$\frac{\partial \text{outage cost}}{\partial \text{initial inventory}} \approx (P(\text{outage}))(\text{outage price}), \tag{4.31}$$

where the inventory price is the costs associated with carrying a single barrel of oil in inventory for a year, and the outage price is the loss from an outage of magnitude equivalent to a single barrel of oil. Then the minimum expected cost occurs when

$$P(\text{outage}) = \frac{\text{inventory price}}{\text{outage price} - \text{inventory price}} \tag{4.32}$$

Thus a first-glance policy is to choose an inventory level which gives an outage rate determined by the relative magnitudes of the inventory price and outage price

Further refinements of the policy are possible. Outage costs are not really linear functions of the outage magnitude, as required by approximation (4.31). Some industrial users have interruptible contracts, and some additional supplies may be procured. These characteristics may be included in a Monte Carlo model relatively easily.

There is also the question of how to modify the policy to guard against model misspecification or unforeseen events, such as an oil embargo. The

shape of the curves in Figure 4.1 is characteristic of real behavior—the total cost is locally quadratic at the minimum, but then rise relatively slowly as inventories rise beyond the optimum level and sharply as inventories decrease. Thus the results indicate not only the apparent optimum inventory, but also that unmodeled random effects are more serious for low inventory levels than for high ones. Thus inventories should be above the estimated optimum, to avoid the sharp cost increase that occurs with large outages.

FIPS is designed to evaluate candidate inventory levels, rather than to find an optimum policy automatically. The key questions that FIPS considers for a given operating policy are:

- What is the likelihood of an outage?
- What is the distribution of outage magnitude?
- What is the expected value and distribution of operating cost?

In order to answer the key questions it is necessary to produce accurate estimates of the probability and distribution of outages. Unfortunately (from a modeling standpoint only!!!) such outages are rare events, and a simple Monte Carlo model would require a very large number of replications to obtain accurate answers. Importance sampling is used to obtain accurate outage answers more quickly, by biasing the distributions of the input quantities toward more unfavorable events.

If the only consideration were to obtain estimates of the outage probability, or the outage magnitude, the example would fit into the rare-event discrete-zero-mode framework for which the integration estimate works well. That is not the case here—the model produces a wide variety of output for variables, not all of which fit into that framework. Early versions of the program used the integration estimate, and produced unreasonable results, including negative variance estimates. Later versions use the ratio estimate. The regression estimate would also work, but is more difficult to implement in a one-pass fashion with so many output quantities.

4.2.1.1 Simplified Fuel Inventory Model We consider here a simplified fuel inventory model, using artificial data and simplified model logic. This model runs for five months, November through March, with all quantities aggregated on a monthly basis. The input random variables are gas and electric

demand (both from effects other than temperature), temperature, hydroelectric, and nuclear generation. Figure 4.2 contains a schematic drawing of the simple model.

The model contains two semi-autonomous balance models. Electric demands are served using all available non-fossil resources first, including hydroelectric and nuclear generation, and other sources (e.g. wind, geothermal, solar, and purchases from the northwest). Next, natural gas is burned if it is available from the gas system. Finally, burning fuel oil is the generation method of last resort. This order is consistent with actual operation, and is determined by economic and environmental reasons; non-fossil resources are generally cheaper and cleaner, followed by natural gas.

The gas balance is performed by subtracting gas demand from available gas purchases; the difference, together with inventory withdrawals, is available to the electric system. Any gas not used can be injected into gas storage, subject to a maximum injection for the month. Actual gas storage fields are large depleted oil fields in porous rock formations, which have limited maximum injection rates. Other operating policies, such as target inventory levels and gas purchase decisions based on price, are not included in the simplified model.

The model logic, followed each month, is:

1. Generate Random Variables, including base (not dependent on temperature) gas and electric demand, temperature, and hydroelectric and nuclear generation.

2. Compute heating degree days from the temperature.

3. Add the temperature load to the electric and gas demands.

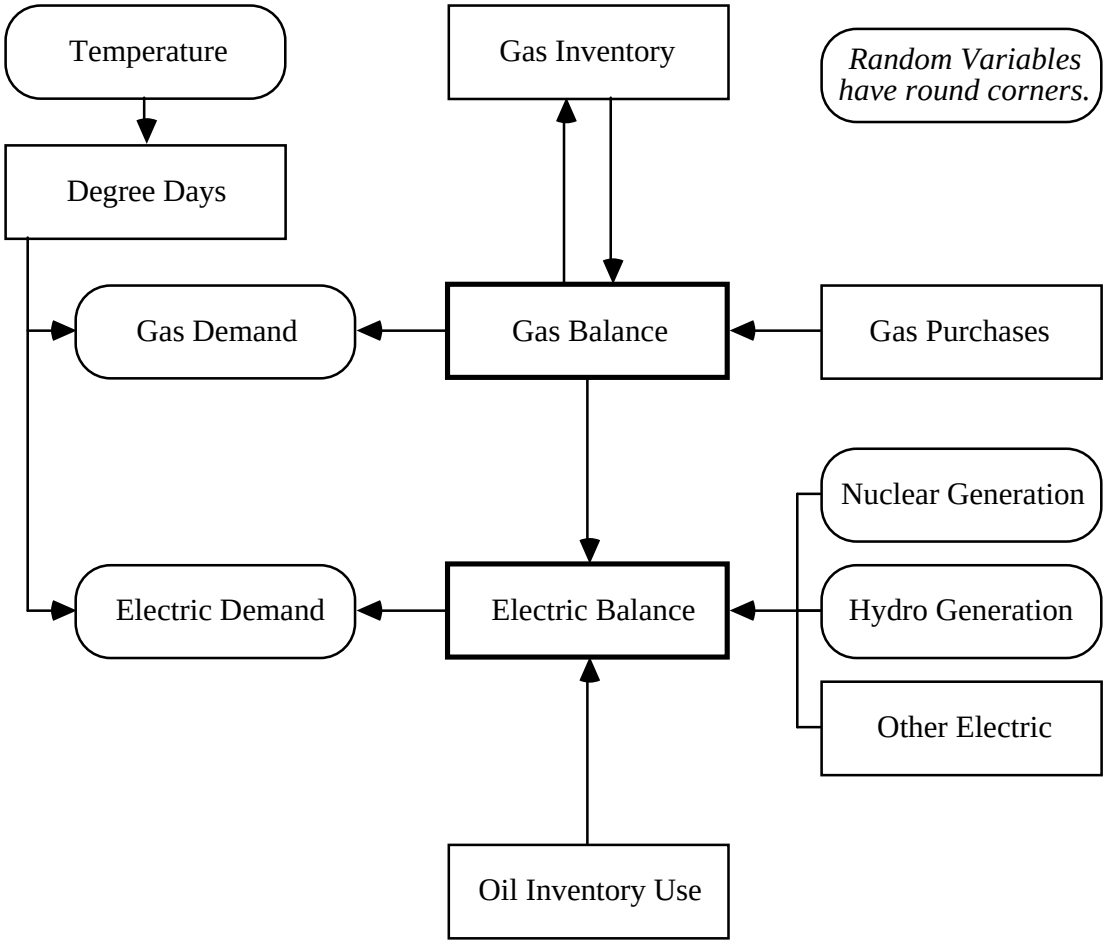
4. Perform an initial electric balance. Subtract non-fossil supplies from electric demand.

5. Perform a gas balance. Subtract gas demand from gas purchases. Satisfy electric demand with any gas available. Inject remaining gas into gas inventory, subject to the injection limit.

6. Serve any remaining electric demand with withdrawals from the oil inventory, if possible.

Output from the physical model is collected and used to compute statistics on cost and operating characteristics, including:

Figure 4.2: Fuel Oil Simulation Schematic Diagram



1. Monthly values of oil and gas inventory levels.
2. Monthly values of input random variables and degree days.
3. Overall inventory, outage, and total cost.

This model is used throughout this paper to illustrate a number of different importance sampling concepts, including confidence intervals (Chapter 2), sampling methods (Chapter 6), and a comparison of the integration, ratio, and regression estimates (next).

4.2.1.2 Fuel Inventory Results This section describes the results of a simulation of the Fuel Inventory Example, with an emphasis on comparing the three estimation formulas.

Table 4.3 contains estimates of the efficiency of the three sampling methods for selected output quantities: five levels of outage, inventory cost, outage cost, sum of inventory and outage costs, and oil inventory levels in December and March.

The sampling distribution used here is fairly conservative, aimed at improving the estimates of the most difficult quantities (outage probability, probability of large outages, and outage cost) while also producing good estimates for other quantities. Table 4.3 shows that this effort was largely successful, at least for the ratio and regression estimates. Results are encouraging, in that the greatest improvement occurs for the quantities that are the hardest to estimate.

The regression estimate is best, with efficiencies ranging from 0.092 for the probability of a large outage, to 0.63 for the final oil inventory. The ratio estimate, too, is generally successful. The integration estimate, on the other hand, performs well for estimating the outage probabilities, but ten times worse than simple random sampling for estimating average inventory levels.

Table 4.4 provides the actual estimates of the expected values, and standard errors of the estimates.

Table 4.5 gives efficiency estimates for the expectations of the input random variables. These do not need to be estimated, but if the simulation reflects their distributions accurately then further analysis can be performed, such as estimating the relationship between input variables and costs. Results are uniformly worse than simple random sampling, but are not that much worse, at least for the ratio and regression estimates. For the integration estimate, on the other hand, results are horrendous, with "efficiencies"

Table 4.3: Efficiency in Fuel Inventory Example
Var(estimate)/Var(simple random sampling)

	Estimated Efficiency			St. Error of Est. Efficiency		
	Int	Ratio	Reg	Int	Ratio	Reg
Outage Probability	0.545	0.653	0.505	0.029	0.024	0.033
$P(\text{Outage} > 100)$	0.462	0.538	0.426	0.037	0.033	0.041
$P(\text{Outage} > 300)$	0.308	0.336	0.289	0.026	0.025	0.027
$P(\text{Outage} > 500)$	0.215	0.225	0.208	0.039	0.039	0.039
$P(\text{Outage} > 700)$	0.094	0.097	0.092	0.021	0.021	0.021
Inventory Cost	9.987	0.914	0.605	0.329	0.015	0.024
Outage Cost	0.290	0.360	0.252	0.019	0.019	0.020
Total Cost	0.260	0.346	0.258	0.020	0.019	0.020
Dec. Oil Inventory	19.306	0.628	0.539	1.489	0.036	0.043
Final Oil Inventory	4.524	0.949	0.632	0.130	0.015	0.025

ranging to over one hundred (one hundred times worse than simple random sampling).

Figure 4.3 contains a plot of the estimated distribution function of outage cost, produced using the regression and integration weights. Note that the distribution function estimate for the integration weights does not go to one, because the sum of the integration weights is 0.988, not 1.00.

This is not an unusual occurrence—even with this very conservative importance sampling distribution, for which the weight function is bounded above by 2, with 2000 replications the standard deviation of the sum of the integration weights is approximately 0.014.

4.3 Intractable Examples

In addition to the usual variance reduction applications, importance sampling can be used to solve some simulation applications that present theoretical or computational difficulties without importance sampling. Two such cases are when:

- the true density is specified only up to a multiplicative constant, or when

Figure 4.3: Cumulative Distribution Function Estimates

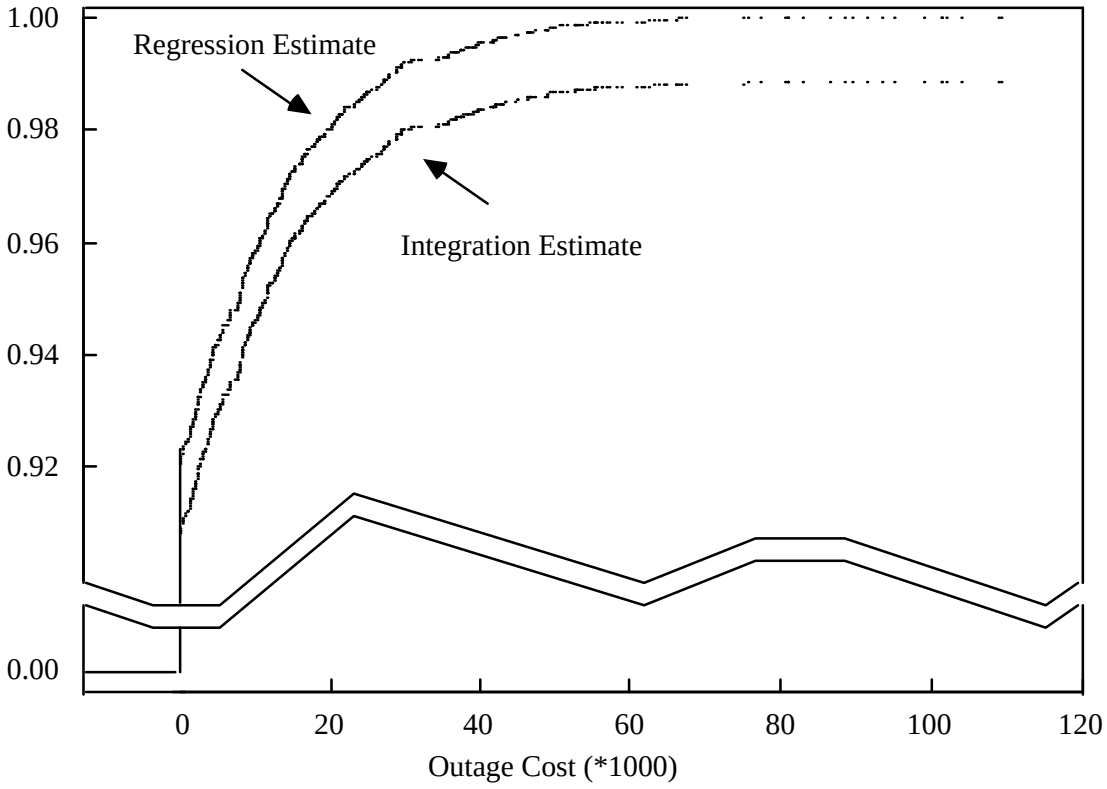


Table 4.4: Estimates in Fuel Inventory Example

	Estimated Expectation			St. Error of Est. Expectation		
	Int	Ratio	Reg	Int	Ratio	Reg
Outage Probability	0.080	0.081	0.079	0.004	0.005	0.004
$P(\text{Outage} > 100)$	0.050	0.051	0.049	0.003	0.004	0.003
$P(\text{Outage} > 300)$	0.015	0.015	0.015	0.002	0.002	0.001
$P(\text{Outage} > 500)$	0.005	0.005	0.005	0.001	0.001	0.001
$P(\text{Outage} > 700)$	0.001	0.001	0.001	0.0	0.0	0.0
Inventory Cost	1712.7	1732.9	1737.3	29.9	9.1	7.4
Outage Cost	1165.4	1179.1	1145.3	65.7	73.1	61.2
Total Cost	2878.2	2911.9	2882.7	59.7	68.9	59.5
Dec. Oil Inventory	380.6	385.0	385.4	5.8	1.0	1.0
Final Oil Inventory	308.3	311.9	313.3	6.5	3.0	2.4

- there is no practical method of generating the true distribution.

Often the two cases occur simultaneously.

4.3.1 Characteristic Roots

Luzar and Olkin study the bias and correlation of characteristic roots of a sample covariance matrix from a normal distribution. Through analytical approximations they simplify their example to one of estimating expectations with respect to the multivariate density

$$f(x) = c(d,m) \prod_{j \neq k} |x_j - x_k| \prod_{j=1}^{d} x_j^{(m-d-1)/2} e^{-(1/2)\sum x_j} \tag{4.33}$$

where d is the dimension of the original example, $x = (x_1, \ldots x_d)$, m is the sample size in the original example, and $c(d,m)$ is a normalizing constant.

If the normalizing constant is unknown then the example can be analyzed using importance sampling and the ratio estimate, for which the unknown constant cancels out. The weight function is

$$W(x) = \frac{c(d,m) f^*(x)}{g(x)} \tag{4.34}$$

Table 4.5: Efficiency for Input Quantities

	Estimated Efficiency			St. Error of Est. Efficiency		
	Int	Ratio	Reg	Int	Ratio	Reg
Nov. Gas Demand	69.660	1.335	1.328	3.264	0.020	0.020
Dec. Gas Demand	109.151	1.349	1.343	4.792	0.018	0.018
Jan. Gas Demand	116.533	1.372	1.362	5.053	0.019	0.020
Feb. Gas Demand	93.393	1.346	1.336	4.061	0.018	0.018
Mar. Gas Demand	55.127	1.357	1.351	2.276	0.018	0.018
Nov. Electric Demand	99.152	1.354	1.347	4.212	0.019	0.019
Dec. Electric Demand	103.155	1.351	1.339	4.397	0.019	0.018
Jan. Electric Demand	108.404	1.352	1.344	4.650	0.019	0.019
Feb. Electric Demand	101.788	1.361	1.353	4.481	0.019	0.019
Mar. Electric Demand	89.011	1.360	1.356	3.754	0.017	0.018
Nov. Avg Temperature	57.606	1.335	1.292	2.238	0.018	0.019
Dec. Avg Temperature	47.764	1.350	1.300	1.874	0.021	0.022
Jan. Avg Temperature	44.210	1.366	1.316	1.748	0.021	0.022
Feb. Avg Temperature	48.114	1.353	1.297	1.862	0.020	0.021
Mar. Avg Temperature	57.738	1.353	1.321	2.355	0.019	0.020
Nov. Hydro Generation	3.741	1.522	1.388	0.069	0.020	0.017
Dec. Hydro Generation	3.916	1.488	1.321	0.073	0.021	0.017
Jan. Hydro Generation	5.044	1.484	1.293	0.106	0.021	0.017
Feb. Hydro Generation	6.481	1.411	1.263	0.159	0.020	0.019
Mar. Hydro Generation	4.896	1.460	1.327	0.155	0.029	0.023
Nov. Nuclear Gen.	3.951	1.275	1.266	0.170	0.021	0.024
Dec. Nuclear Gen.	3.924	1.305	1.298	0.175	0.021	0.024
Jan. Nuclear Gen	4.259	1.270	1.255	0.197	0.022	0.027
Feb. Nuclear Gen	4.103	1.261	1.246	0.181	0.021	0.026
Mar. Nuclear Gen	4.105	1.283	1.272	0.186	0.020	0.024

where $f(x) = c(d,m)f^*(x)$, and the ratio estimate is

$$\hat{\mu}_{\text{ratio}} = \frac{\sum W(X_i)\theta(X_i)}{\sum W(X_i)} = \frac{\sum f^*(X_i)\theta(X_i)/g(X_i)}{\sum f^*(X_i)/g(X_i)} \tag{4.35}$$

In this case the normalizing constant is known, but there is no apparent efficient way to generate observations with the density (4.33). Luzar and Olkin use importance sampling, obtaining best results using a sampling density which is a product of independent chi-squared distributions (with degrees of freedom 15 for $m = 15$ and $d = 3$):

$$g(x) = c(k)^d \prod_{j=1}^{d} x_j^{k/2-1} e^{-(1/2)\sum x_j} \tag{4.36}$$

It would be possible to solve this problem without importance sampling by using the acceptance-rejection method of generating random variables (von Neumann 1951, Kennedy and Gentile 1980). This requires that it be possible to generate values from a distribution $h(x)$ which majorizes $f(x)$, i.e. $ch(x) \geq f(x)$ for all x and some $c < \infty$. Then X can be generated by the algorithm:

1. Generate $X \sim h$.
2. Generate $U \sim U(0,1)$.
3. If $U < f(x)/(ch(x))$ "accept", and output X;
   Else "reject", and return to step 1.

This requires, on average, c iterations to generate a single X value, so for efficient generation c should be small.

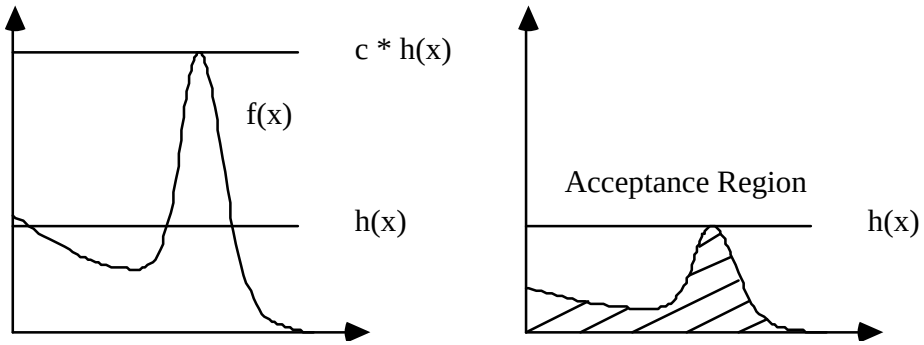
In the covariance matrix example a possible majorizing density for use in acceptance-rejection is a chi-square distribution with $m + 3d + 1$ degrees of freedom (proof in appendix), but this would be very inefficient for large d.

Importance sampling can be used as an alternate to acceptance-rejection methods, since every variable generated from X can be used, rather than using only one out of every c. Importance sampling also does require that the sampling distribution majorize the true distribution, only that it dominate.

4.3.2 Bayesian Analysis

Stewart (1976, 1983), Kloek and van Dijk (1978), van Dijk and Kloek (1983), and Bauwens (1984) discuss examples in Bayesian analysis involving nonconjugate distributions that can be solved using importance sampling.

Figure 4.4: Acceptance-Rejection Random Variable Generation



In a short departure from our usual notation, let θ be a parameter with a prior distribution with prior distribution $\xi(\theta)$ and let X be observed data. The goal is to perform some kind of analysis with respect to the posterior distribution of θ,

$$\xi(\theta|X) = \frac{\xi(\theta)L(\theta|X)}{\int \xi(\tau)L(\tau|X)d\tau} \tag{4.37}$$

where $L(\theta|X)$ is the likelihood function corresponding to X.

The estimated expected value of an output $H(\theta)$ under $\xi(\theta|X)$ is

$$E(H(\theta|X)) = \frac{\int H(\theta)\xi(\theta)L(\theta|X)d\theta}{\int \xi(\tau)L(\tau|X)d\tau} \tag{4.38}$$

In performing importance sampling in this example let the sampling distribution be $g_\theta(\theta)$, the true distribution $\xi(\theta|X)$, and the weight function $W(\theta)$ their ratio. If the denominator of (4.38) cannot be calculated explicitly then $W(\theta)$ can be calculated only up to a multiplicative factor, and the use of the ratio estimate is called for, resulting in observation weights

$$V_{\text{ratio}}(\theta_i) = \frac{\xi(\theta_i)L(\theta_i|X)/g_\theta(\theta_i)}{\sum_{k=1}^{n} \xi(\theta_k)L(\theta_k|X)/g_\theta(\theta_k)} \tag{4.39}$$

on Monte Carlo observations θ_i. The estimated expected value of $H(\theta)$ under $\xi(\theta|X)$ is

$$\hat{E}(H(\theta)|X) = \hat{\mu}_{\text{ratio}} = \sum_{i=1}^{n} V_{\text{ratio}}(\theta_i)H(\theta_i) \tag{4.40}$$

Kloek and van Dijk (1978) and Bauwens (1984) analyze this example as the ratio of two integration estimates. They consider whether it is better to attempt to choose a sampling distribution which is proportional to the numerator (and so gives a good estimate of the numerator, since the induced integration estimate transformation would make the numerator values nearly constant) or the denominator. They recommend matching the denominator, i.e. attempting to match the posterior distribution as closely as possible. The reasons (in our words) are that:

- in attempting to find a good sampling distribution for the numerator (good in the sense of the integration estimate) the variance of the denominator can explode,
- the numerator can be negative, in which case a proportional sampling distribution is not possible,
- there may be more than one output quantity being estimated, i.e. multiple equations of the form (4.38), and it is easier to optimize for a single denominator than multiple numerators.

We concur with their conclusion, though for different reasons. This example requires a ratio estimate, so attempting to find a good sampling distribution for the integration estimate is misguided; the estimate cannot be perfect, unless the numerator and denominator are proportional. A good sampling distribution here must be adequate for both the numerator and the denominator, and the posterior density is a good first choice.

In addition, we suggest the use of a sampling distribution which is formed as a mixture of component distributions

$$g_\theta(\theta) = \sum \lambda_k g_k(\theta), \tag{4.41}$$

where $\lambda_k \geq 0$ and $\sum \lambda_k = 1$. The largest component should match the posterior distribution as much as possible, but the other components can be chosen to sample more in regions where the numerator is extreme (for one or more of the output quantities). Mixture distributions are discussed further in Chapter 6.

An additional advantage of the use of importance sampling in Bayesian analysis is that it allows the Bayesian analysis to be performed for more than one prior distribution, using a single sampling distribution and a single sample (Stewart, 1979). This is an example of the use of importance sampling in analyzing applications involving multiple "true" distributions.

4.4 Multiple Object Distributions

In some simulation applications there is no single distribution that should be regarded as the "true" distribution. There may be a finite, or even an infinite number of distributions to be analyzed, which we term "object" distributions[1]. Importance sampling can be particularly useful in these applications, since it does not require that the sampling process be repeated for each object. Instead a given sample can be used as an approximation for any distribution (for which $g(x)/f(x) > 0$), as long as appropriate weights are used.

This is not a traditional interpretation of importance sampling, where there is a single object distribution and the Monte Carlo experimenter may choose a sampling distribution which provides good performance for that object distribution. Here there are multiple object distributions, and the experimenter may or may not have, or may not use, the freedom to choose a sampling distribution. Without (the use of) that freedom the name "importance sampling" is a mild misnomer, and other names have been used, such as "polysampling" (Tukey 1987). Nevertheless we use the name "importance sampling" because this makes clear the similarity between the different applications, and indicates how performance can be improved by using importance sampling ideas such as choosing a sampling distribution and choosing an estimate.

Bayesian analysis is one area with multiple input distributions, corresponding to different choices of a prior distribution. We discuss four additional areas. The first is within importance sampling, for estimating the quality of results that would have been obtained under different sampling distributions. Such analysis performed after some observations, but before the experiment is completed, can indicate ways to improve the sampling for the remaining replications.

The second is the implementation of the bootstrap tilting interval. Here there are an infinite number of object distributions, which can be quickly searched using importance sampling (Tibshirani 1984); not only does importance sampling eliminate the need for sampling from each distribution, but there is a single sampling distribution which is nearly optimal (in the usual sense of finding optimal importance sampling distributions) for all object distributions.

[1] In later work we call these "target" distributions.

The third area is response surface estimation. In many stochastic applications one is interested in the performance of the system not just for a single input distribution, but for a set of input distributions. For example we might estimate the characteristics of a queueing system for a number of different repair rates (which determine the distribution of repair times). Importance sampling can be used to analyze the system for all input distributions simultaneously (Glynn & Iglehart 1987). Beckman and McKay (1987) use importance sampling to analyze results from a single, expensive, computer simulation under a variety of input distribution assumptions.

One notable feature of importance sampling is the continuity of results obtained from different distributions. A small difference between two object distributions results in a small change in the weights assigned to different replications in a Monte Carlo sample, and results in no change in the sample values. Importance sampling thus enables distributional results to be differentiated. Reiman and Weiss (1986) use this for sensitivity analysis in stochastic applications. Glynn (1986, 1987) uses the same idea for single replications in a stochastic optimization setting.

The fourth area is the application of estimating the performance of robust estimates under a number of distributions.

4.4.1 Analysis of Importance Sampling

When importance sampling is done it is often useful to ask whether the sampling distribution used performed well, or if perhaps a different sampling distribution might be better. This knowledge can be used after an initial run to improve the distribution used for the remaining sampling (Moy 1965).

This analysis can be done without further sampling, using a second, conceptual level of importance sampling. When the true distribution is f and the sampling distribution is g, the realized values can be viewed as a sample from a third distribution h, with weights which are based on the inverse likelihood ratio

$$W^{(h:g)}(x) = h(x)/g(x), \tag{4.42}$$

as long as $g(x)/h(x) > 0$. Note the similarity to (1.1, $W(x) = f(x)/g(x)$); here h takes the place of f. This inverse likelihood ratio can be used to obtain weights for h against g using any of our usual formula. The integration, ratio, regression, exponential and maximum-likelihood weights are:

$$V_{\text{int}}^{(h:g)}(X_i) = \frac{1}{n} W_i^{(h:g)} \tag{4.43}$$

$$V_{\text{ratio}}^{(h:g)}(X_i) = W_i^{(h:g)} / \sum_{j=1}^{n} W^{(h:g)_j} \tag{4.44}$$

$$V_{\text{reg}}^{(h:g)}(X_i) = \frac{1}{n} W_i^{(h:g)} (1 - a^{(h:g)}(W_i^{(h:g)} - \bar{W}^{(h:g)})) \tag{4.45}$$

$$V_{\text{exp}}^{(h:g)}(X_i) = W^{(h:g)} a \exp(b W_i^{(h:g)}) \tag{4.46}$$

$$V_{\text{mle}}^{(h:g)}(X_i) = \frac{a W^{(h:g)}}{1 - b(W_i^{(h:g)} - \bar{W}^{(h:g)})} \tag{4.47}$$

where $a^{(h:g)} = (\bar{W}^{(h:g)} - 1)/\hat{\sigma}(W^{(h:g)})$ in the regression estimate formula and a and b are chosen in the exponential and maximum likelihood formulas so that $\sum V_i = \sum V_i / W_i^{(h:g)} = 1$.

Assume now that one of the weight formulas has been chosen, which we denote $V_i^{(h:g)} := V^{(h:g)}(X_i)$, dropping the reference to the particular formula used. We use these weights to estimate the performance of an importance sampling estimate as if we had sampled from h rather than g.

For example, the asymptotic variance of the integration estimate for estimating $\mu = E_f(\theta)$ when sampling from h (with $h(x)/f(x) > 0$) is

$$E_h((Y^{(h)} - \mu)^2), \tag{4.48}$$

where

$$W^{(h)}(x) := f(x)/h(x) \tag{4.49}$$

and

$$Y^{(h)}(x) := \theta(x) W^{(h)}(x). \tag{4.50}$$

Note that $W^{(h)}$ is the inverse likelihood ration for f against h; it is what the W function would be for a true distribution f and sampling distribution h. In contrast, $W^{(h:g)}$ is a weight function for actual sampling distribution g and conceptual sampling distribution h.

Our estimate of the variance given by (4.48) is:

$$\sum_{i=1}^{n} V_i^{(h:g)} (Y_i^{(h)} - \hat{\mu}_{\text{int}}^{(h:g)})^2. \tag{4.51}$$

Here the estimated expectation is also defined using the weights $V^{(h:g)}$:

$$\hat{\mu}_{\text{int}}^{(h:g)} = \sum_{i=1}^{n} V_i^{(h:g)} Y_i^{(h)}. \tag{4.52}$$

We can of course simplify the expectation and variance estimates by substituting the actual formula used for $V^{(h:g)}$. In the case of the integration weights $V_{\text{int}}^{(h:g)}$, the expectation estimate is the same as it was based on only f and g, and the variance estimate is:

$$n^{-1}\sum_{i=1}^{n}\frac{f^2}{gh}\theta^2 - \hat{\mu}_{\text{int}}^2. \tag{4.53}$$

This is the same as the corresponding formula for sampling from g, except that the denominator is gh rather than g^2. If h is generally larger than g when θ is large, the new sum is smaller than the corresponding sum for g; this is consistent with our prior observations that importance sampling is beneficial if the sampling density is large when θ is large.

This formula should be used with reasonable caution. It indicates that if we choose h to be larger than g at *all* observed values (not just observed values with large θ), then the apparent error variance is smaller. Taken to the extreme in adaptive distribution choice (4.53) indicates that the best sampling distribution would have support solely on points observed so far. This is less of a problem for the weight formulas other than the integration weights; since they are normalized to unit mass they do not attempt to optimize by concentrating on observed values.

Similarly, the asymptotic variance of the ratio estimate when sampling from h is

$$E_h(W^{(h)2}(\theta-\mu)^2), \tag{4.54}$$

which we estimate using:

$$\sum_{i=1}^{n} V_i^{(h:g)} W^{(h)2}(\theta - \hat{\mu}_{\text{ratio}}^{(h:g)})^2. \tag{4.55}$$

Here it is most natural to use the ratio estimate weights $V_{\text{ratio}}^{(h:g)}$, because the ratio estimate of μ is the same as the original estimate, $\hat{\mu}_{\text{ratio}}^{(h:g)} = \hat{\mu}_{\text{ratio}}$.

Finally, the asymptotic variance of the regression, exponential and maximum likelihood estimates is

$$\text{Var}_h(Y^{(h)} - \beta^{(h)}W^{(h)}), \tag{4.56}$$

where $Y^{(h)} = \theta W^{(h)}$ and $\beta^{(h)}$ is the regression slope of $Y^{(h)}$ on $W^{(h)}$ under h. Note that $\beta^{(h)}$ depends on h. The slope should be estimated using a weighted

regression using weights $V^{(h:g)}$:

$$\hat{\beta}^{(h)} = \frac{\sum_{i=1}^{n} V_i^{(h:g)}(Y_i^{(h)} - \bar{Y}^{(h)})(W_i^{(h)} - \bar{W}^{(h)})}{\sum_{i=1}^{n} V_i^{(h:g)}(W_i^{(h)} - \bar{W}^{(h)})^2} \tag{4.57}$$

and $\bar{Y}^{(h)}$ and $\bar{W}^{(h)}$ are weighted averages computed using weights $V^{(h:g)}$.

4.4.2 Bootstrap Tilting Interval

The bootstrap tilting interval was introduced by Efron (1982) as a way of obtaining a bootstrap confidence interval by performing the nonparametric analog of the inversion used in a parametric setting. Tibshirani (1984) describes an efficient way to implement this interval using importance sampling.

Let $Z_1, Z_2, \ldots Z_d \overset{i.i.d.}{\sim} F$, F unknown. In a parametric application we assume that F is a member of a parametric family, $F = F_\xi$, $\xi \in \Xi$. Let $\hat{\xi}_{\text{obs}} = T(Z_1, \ldots Z_d)$ be an estimate of ξ, where T is a functional statistic. The classical two-sided $1 - 2\alpha$ confidence interval for ξ is $[\xi_{\text{low}}, \xi_{\text{up}}]$, where

$$\begin{aligned} \xi_{\text{low}} &:= \inf\{\xi : P_{\xi_{\text{low}}}(T(X) \geq \hat{\xi}_{\text{obs}}) \geq \alpha\} \\ \xi_{\text{up}} &:= \sup\{\xi : P_{\xi_{\text{up}}}(T(X) \leq \hat{\xi}_{\text{obs}}) \geq \alpha\}. \end{aligned} \tag{4.58}$$

Here $P_\xi(A)$ is the probability of A when $X = (X_1, \ldots X_d)$, $X_j \overset{i.i.d.}{\sim} F_\xi$.

The bootstrap tilting interval is defined using the same idea, but using a nonparametric family of distributions. In principle we could let Ξ include all distributions which have support solely on the original sample points. We would then search all possible weight vectors $\mathbf{w}$ of length d to find

$$\begin{aligned} \xi_{\text{low}} &:= \inf\{T(\mathbf{w}) : P_{\mathbf{w}}(T(X) \geq \hat{\xi}_{\text{obs}}) \geq \alpha\} \\ \xi_{\text{up}} &:= \sup\{T(\mathbf{w}) : P_{\mathbf{w}}(T(X) \leq \hat{\xi}_{\text{obs}}) \geq \alpha\}. \end{aligned} \tag{4.59}$$

The probability is probability under bootstrap sampling with weights $\mathbf{w}$.

In practice it is necessary to restrict the set of weight vectors to be considered, both for efficiency reasons and because the global search is likely to lead to results that are not reasonable (e.g. a vector of weights with all its mass on a single one of the observed sample values).

Let Ξ be the family of distributions defined by placing weight

$$w_j^T := \frac{e^{tU_j}}{\sum_{k=1}^{d} e^{tU_k}} \tag{4.60}$$

on point j of the original sample, for $t \in \mathcal{R}$ and U_j is defined in (4.19).

Now find t_{low} and t_{up} which satisfy:

$$\begin{aligned} t_{\text{low}} &:= \inf\{t : P_t(T(X) \geq \hat{\xi}_{\text{obs}}) \geq \alpha\} \\ t_{\text{up}} &:= \sup\{t : P_t(T(X) \leq \hat{\xi}_{\text{obs}}) \geq \alpha\}. \end{aligned} \tag{4.61}$$

Here $P_t(A)$ is the probability of A when X is a bootstrap sample drawn with probabilities $\mathbf{w}^t$.

Then the tilting interval is

$$[\xi(t_{\text{low}}), \xi(t_{\text{up}})], \tag{4.62}$$

where $\xi(t) := T(\mathbf{w}^t)$, that is T computed on the distribution formed by placing weight w_k^t on Z_k (of the original sample).

Simultaneous solution of equations (4.60) and (4.61) is equivalent to finding weights which (locally) minimize (maximize) $T(\mathbf{w})$, subject to the restrictions that the backward Kullback-Leibler distance from $\mathbf{w}_0$ to $\mathbf{w}^t$ not be greater than $\text{KL}(\mathbf{w}_0, \mathbf{w}^t)$ and that the coverage probabilities appear correct.

Now solving (4.61) involves only a one-dimensional search, over possible t values, instead of a global search over all weight vectors. Still, the computational effort could be prohibitive, if it would require that a complete set of bootstrap samples be drawn, and corresponding ξ values evaluated, for every value of t considered in a numerical search.

The solution is to draw a single set of bootstrap samples from some sampling distribution, and to use that sample to estimate the required probabilities for any candidate distribution. The weight assigned to each of the bootstrap samples is determined using importance sampling.

A good choice for a sampling distribution is the unmodified empirical distribution function, $\mathbf{w}_0$. If we consider sampling distributions that would be obtained by exponential tilting (Chapter 6), we obtain sampling distributions of the form (4.60). Furthermore, both experience and theory suggest that the tilting parameter t should be chosen so that

$$P_w(T(X) \leq T(\hat{F})) \approx 0.5, \tag{4.63}$$

which we expect to be the case for $t = 0$, which corresponds to sampling with equal probability from the original sample (the "true" distribution here is a tilted distribution, not the original distribution).

To implement this, draw n samples X_i from the original empirical distribution, and for each, compute the sum of the empirical influence values for the sample:

$$S(X_i) = \sum_{j=1}^{d} U_{k(j)}, \tag{4.64}$$

where the jth sample value is the kth value in the original sample, $X_{i,j} = Z_{k(j)}$. The relative likelihood of the sample under the distribution $\mathbf{w}^T$ compared to $\mathbf{w}_0$ is

$$W(X_i; t) = \prod_{j=1}^{d} \frac{1/d}{w_{k(j)}^t} = \frac{\left(d^{-1} \sum_{j=1}^{d} e^{tU_j}\right)^d}{e^{tS(Xi)}}. \tag{4.65}$$

Note that the numerator is the same for all replications.

Now finding t_{low} and t_{up} that satisfy (4.61) involves a one-dimensional search over values of t, with probabilities estimated from the importance-sampling Monte Carlo estimate of the probability, i.e.

$$\hat{P}_t(T(X) \leq \hat{\xi}_{\text{obs}}) = n^{-1} \sum_{i=1}^{n} W(X_i; t) I(T(X_i) \leq \hat{\xi}_{\text{obs}}) \tag{4.66}$$

where $W(X_i)$ is the inverse likelihood ratio for replication i and tilting parameter t. Other weights can be substituted for the integration estimate weights in (4.66), but the integration method is probably the best choice in this (lardimaz) example.

4.4.3 Study of Robust Estimates

"Estimation is the art of inferring information about some unknown quantity on the basis of available data. Typically an estimator of some sort is used. The estimator is chosen to perform well under the conditions that are assumed to underly the data. Since these conditions are never known exactly, estimators must be chosen which are robust, which perform well under a variety of underlying conditions." (Andrews, et al. 1972)

So begins the introduction to the "Princeton Robustness Study," the most comprehensive study of robust estimates to date. A main focus of that work is the empirical study of the behavior of estimates using Monte Carlo simulation under a variety of distributions, from "nice" distributions like the

Gaussian to long-tailed, infinite-variance distributions like the Cauchy and slash distributions.

These Monte Carlo simulations can be time-consuming, since many robust estimates are difficult to compute (many require iterative algorithms) and because of the number of distributions to be considered. Importance sampling can be used to improve the efficiency of these simulations, in two ways. As usual, we may modify the sampling distribution used in the simulation for a single anchor distribution. The focus of this section is on the second way—the simultaneous use of observations from each sampling distribution for estimating performance under all sampling distributions.

The use of importance sampling in this example has been discussed by Tukey (1987) under the name "polysampling". Every observation, from any sampling distribution, can be regarded as an observation from each anchor distribution, as long as appropriate weights are used.

Consider the case of two distributions (the method generalizes easily). If n_f and n_g observations are sampled from distributions f and g, then all observations can be regarded as a sample from each of the distributions, with proper weights. If $w_i^f = w^f(x_i)$ is the (unnormalized) weight for replication i for distribution f, $i = 1, \ldots n$, $n = n_f + n_g$, then the set of weights w_i should satisfy (approximately)

$$E\left(\frac{\sum w_i^f I(x_i \in A)}{\sum w_i^f}\right) \approx \int_A f(x)dx. \tag{4.67}$$

Similarly, the weights w_i^g for g should satisfy

$$E\left(\frac{\sum w_i^g I(x_i \in A)}{\sum w_i^g}\right) \approx \int_A g(x)dx. \tag{4.68}$$

Tukey discusses a number of ways to obtain weights. We believe the most useful of these is to let $w^f(x)$ be the inverse likelihood ratio

$$w^f(x) = \frac{n_f + n_g}{n_f + n_g(g(x)/f(x))} \tag{4.69}$$

and similarly

$$w^g(x) = \frac{n_f + n_g}{n_f(f(x)/g(x)) + n_g}. \tag{4.70}$$

These are the usual inverse likelihood ratio formulas for true distributions f and g, respectively. If used as in formulas (4.69) and (4.70) the result is the ratio estimate.

Tukey also proposes weight functions of the form

$$w^f_{\text{Tukey}}(x) = \frac{Cn_f + Dn_g}{Cn_f + Dn_g(g(x)/f(x))} \tag{4.71}$$

and

$$w^g_{\text{Tukey}}(x) = \frac{Cn_f + Dn_g}{Cn_f(f(x)/g(x)) + Dn_g}, \tag{4.72}$$

where $C > 0$, $D > 0$. He suggests choosing C and D such that $Cn_f = Dn_g$.

We argue against this choice. It is unclear exactly what Tukey intended the weights to be from his argument; the two possibilities are:

$$w^f_i = w^f_{\text{Tukey}}(X_i) \tag{4.73}$$

or

$$w^f_i = a_i w^f_{\text{Tukey}}(X_i) \tag{4.74}$$

where

$$a_i = \begin{cases} C & \text{for } x_i \text{ from } f \\ D & \text{for } x_i \text{ from } g \end{cases} \tag{4.75}$$

Similarly, w^g_i can be defined with or without a_i.

In the first case, where weights are defined without a_i, the weights do not satisfy the defining relationship (4.68) and give estimates which are not consistent. The second case leads to a given value X_i being assigned a different weight depending on which distribution it was sampled from. Earlier Tukey argues that "such a distinction is both unnecessary and counterproductive". We concur.

Table 4.6 contains results from an experiment on the performance of five estimates for four distributions, for a sample size of 20 and 1000 replications from each distribution. The estimates are the mean, 10% trimmed mean, median, and two one-step Huber estimates. The Huber estimates are computed as a weighted average by assigning weight

$$W_i = \begin{cases} 1 & D_i \le k \\ \frac{k}{D_i} & D_i > k \end{cases} \tag{4.76}$$

to observation i, where

$$D_i \ \frac{|X_i - \text{median}|}{.67\text{MAD}}, \tag{4.77}$$

where MAD is the median absolute deviation from the median, and $k = 1$ and 2 for the Huber1 and Huber2 estimates, respectively. Any observation which is more than k times a measure of spread (scaled to be equal to the standard deviation if the distribution is Gaussian) from the median is given reduced weight.

The distributions used are the normal, Cauchy, slash (normal/uniform) and a contaminated normal ($.9N(0,1) + .1N(0,9)$, number of contaminating observations not fixed). The Cauchy and slash distributions are rescaled to make their spread more consistent with the other distributions—the Cauchy values are divided by $\sqrt{2}$ and the slash values by 2.

The results in Table 4.6 are efficiency estimates, relative to estimating each distribution separately. The estimated efficiency ranges between .32 and .74, or variance reductions ranging between 68% and 26%. The comparison is between using all 4000 replications in importance sampling to using 1000 observations without importance sampling (not 4000 observations from each anchor distribution).

Importance sampling gives substantial improvements in the quality of any single estimate. This is offset, however, by the correlation that is introduced between estimates for different distributions and the same statistic. If the criterion for selecting estimates is the total variance over all distributions the net gain from importance sampling is zero, except for a small improvement due to a stratification effect discussed in Chapter 6. Importance sampling is also difficult to apply effectively together with the variance reduction scheme employed in the Princeton robustness study.

4.4.4 Response Surface Estimation

In many stochastic applications one is interested in "response surface estimation", estimation of the performance of the system for a range of input distributions rather than a single input distribution. An example is the performance of a queueing system for a number of different repair time distributions. Importance sampling can sometimes be used to analyze the system for all input distributions simultaneously.

Let X_i, $i = 1, \ldots n$, be a sample from distribution $g(x)$, and let $f(x, \alpha)$, $\alpha \in A$. be a family of distributions for which $f(x, \alpha) > 0$ for any α implies $g(x) > 0$. Then we may use importance sampling to estimate $\mu(\alpha) := E_\alpha(\theta(X, \alpha))$, where the expectation is for sampling with respect to $f(x, \alpha)$. Any of the usual importance sampling estimates may be used,

Table 4.6: Efficiency in Robust Estimate Evaluation
Ratio of MSE with to MSE without Importance Sampling

	Integration Estimate			
	Normal	Cauchy	Slash	Contaminated
Mean	0.62	0.55	0.34	0.47
10% trim	0.64	0.67	0.32	0.60
Median	0.74	0.49	0.41	0.74
Huber2	0.66	0.51	0.42	0.57
Huber1	0.68	0.55	0.38	0.68
	Ratio Estimate			
	Normal	Cauchy	Slash	Contaminated
Mean	0.58	0.55	0.34	0.40
10% Trim	0.59	0.65	0.32	0.52
Median	0.66	0.48	0.38	0.65
Huber2	0.61	0.48	0.41	0.50
Huber1	0.62	0.53	0.36	0.59
	Regression Estimate			
	Normal	Cauchy	Slash	Contaminated
Mean	0.58	0.55	0.34	0.40
10% Trim	0.59	0.65	0.32	0.52
Median	0.66	0.48	0.37	0.64
Huber2	0.61	0.48	0.41	0.50
Huber1	0.62	0.53	0.36	0.59

though we prefer any equivariant estimate to the integration estimate to prevent changes in distribution mass from affecting estimates.

Response surface estimation is simplest if $\theta(X, \alpha) = \theta(X)$, independent of α. Then the estimates of $\mu(\alpha)$ are obtained by different weighted averages of the results obtained from a single simulation. Note that the sampling distribution may be one of the distributions under consideration, and the response surface estimation need not be planned before the simulation, but can be done after the fact, if the necessary data has been saved.

The convergence of pointwise estimates of $\mu(\alpha) := E_\alpha(\theta(X, \alpha))$ is governed by the usual convergence properties of importance sampling. Glynn & Iglehart (1987) give conditions under which the convergence of $\hat{\mu}(\alpha)$ to $\mu(\alpha)$ is uniform for α in a closed finite interval.

The convergence is not generally uniform over infinite or open intervals. Suppose that $f(x, \alpha)$ is a family of distributions such that

$$\lim_{\alpha \to m} f(x, \alpha) = 0 \tag{4.78}$$

for all x, where m is one of $-\infty$, ∞, or the endpoint of an open interval. Then the weights

$$W_i = W(X_i, \alpha) = \frac{f(X_i, \alpha)}{g(X_i)} \tag{4.79}$$

also go to zero for any sample value. If $\theta(X, \alpha)$ is independent of α, then

$$\lim_{\alpha \to m} \hat{\mu}_{\text{int}} = 0 \tag{4.80}$$

regardless of the observed θ_i. The ratio and regression estimates fare only marginally better—their limit is the value θ_k for the single replication for which

$$\lim_{\alpha \to m} \frac{W_k}{\sum W_i} = 1, \tag{4.81}$$

if there is one (in many continuous parametric families there is).

This phenomena of weights going to zero is dangerous because the *apparent* quality of performance of the estimates gets better. For the integration estimate, the usual standard error estimate is $n^{-1}\sqrt{\sum (Y_i - \bar{Y})^2}$, which goes to zero as quickly as the weights. Because of this response surface estimation using importance sampling should be used with caution, especially over long intervals. Diagnostic output should be used to indicate if the average W value is decreasing significantly below 1.

Confidence intervals for response surface estimation are discussed in Section 2.6, for the fuel inventory example. That example indicates a problem with importance sampling response surface estimation—for small deviations from the base case the confidence intervals appear reasonable, but for large deviations they are not.

Here we concentrate on a smaller range of changes in gas demand, though not so small that problems are not apparent. Gas demand has five normally-distributed monthly values with standard deviation 100 and serial correlation 0.2. The original sampling distribution makes a relatively minor change in these distributions, with no monthly expected value changed by more than 20. Changes here range to ± 200 in each monthly value. For small changes the results appear reasonable—outage costs rise and inventory costs fall with increasing gas demand. For larger changes results are unreasonable, especially for the integration estimate, which is more sensitive to changes in the average weight value. Results are presented in Figures 4.5–4.7.

Importance sampling can also be used to find derivatives of expectations with respect to changes in parameters, if both $\theta(x, \alpha)$ and $W(x, \alpha)$ are differentiable with respect to α and $E_g(|W\theta|) < \infty$. Reiman and Weiss (1986) find derivatives for the derivative of the likelihood of replications with respect to a Poisson arrival rate in the context of stochastic processes. Glynn (1986, 1987) estimates derivatives using a single replication at a time in a stochastic optimization example.

Figure 4.5: Outage Cost Response Surface Estimate

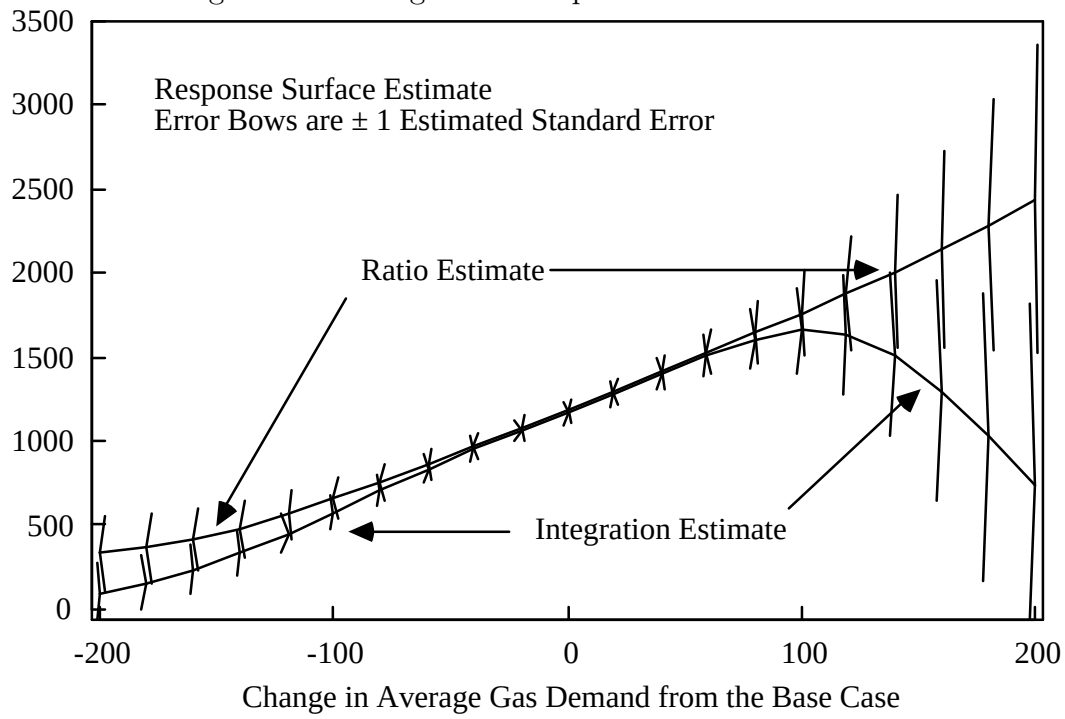


Figure 4.6: Inventory Cost Response Surface Estimate

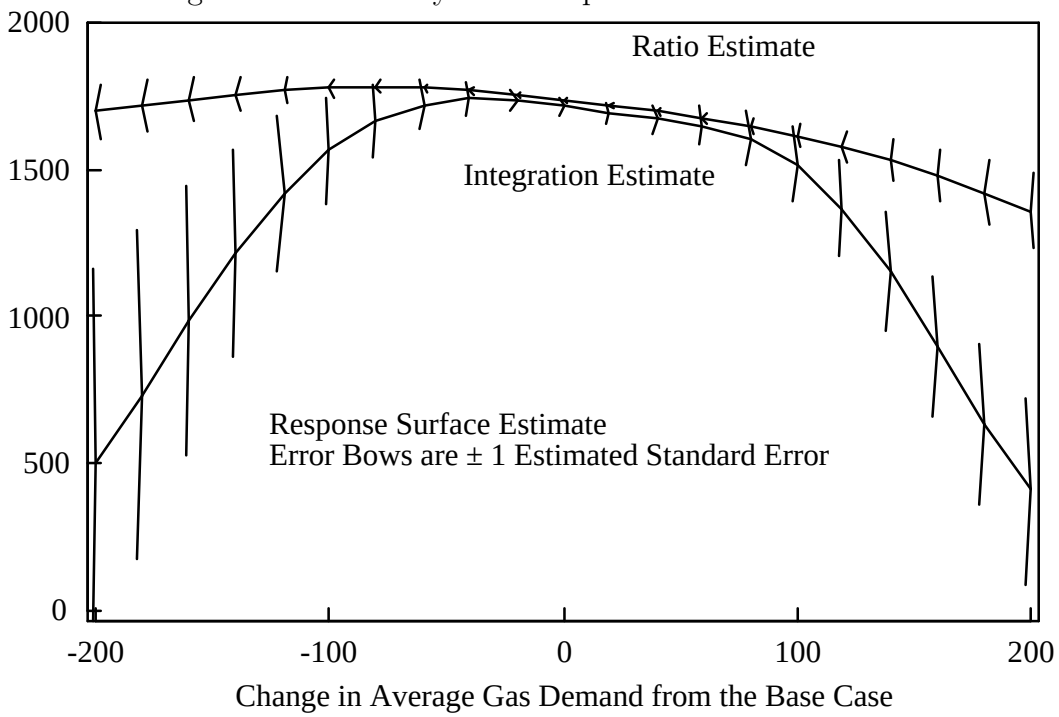
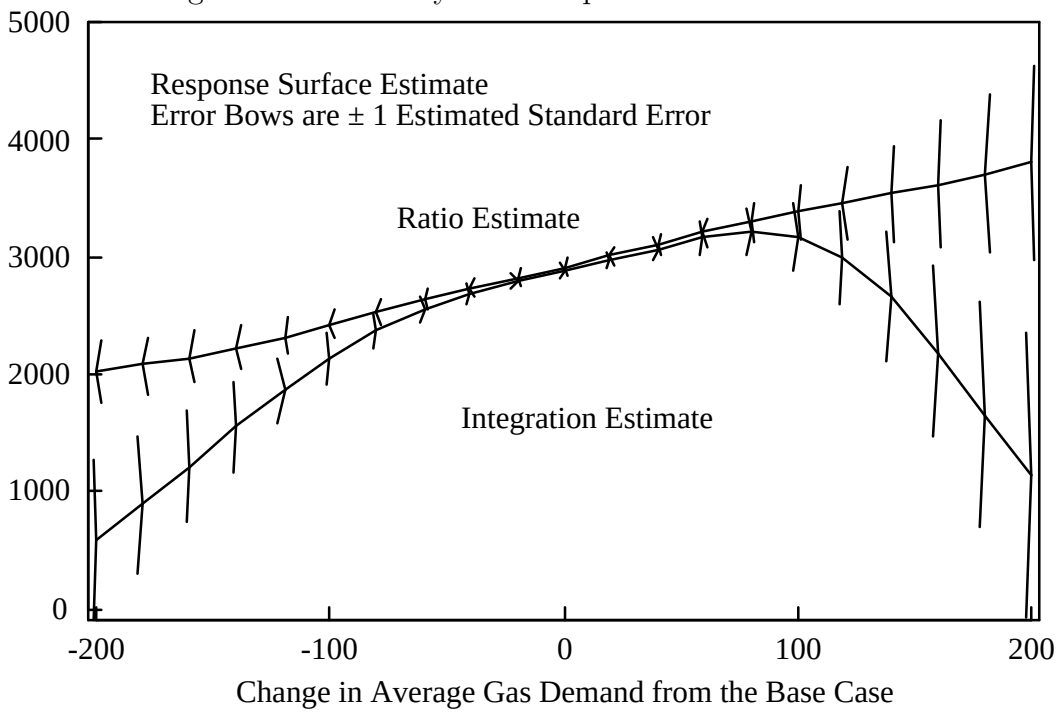


Figure 4.7: Inventory Cost Response Surface Estimate



Chapter 5

Sampling Distributions

It goes almost without saying that good results in importance sampling require good sampling distributions. Variance reductions in importance sampling are not guaranteed, and a bad sampling distribution can result in a variance increase, even an infinite variance.

Finding and generating a good sampling distribution can be difficult, especially in complicated applications with multivariate input. Indeed, some authors have recommended against using importance sampling because of the problems they have encountered with multivariate input (Wilson 1984, Bratley, Fox and Schrage 1983). We believe, however, that the problems are not insurmountable. In fact importance sampling can be made very robust using a combination of mixture sampling (discussed in Chapter 6) and any estimate except the integration estimate.

This chapter discusses requirements for a sampling distribution, and some principles useful in choosing a sampling distribution. These principles will be used in Chapter 6 in the development of specific sampling methods, including mixture sampling, generalized mixture sampling, exponential tilting, and internal distributions.

A good sampling distribution should satisfy four requirements:

- dominate the true distribution,
- it should be easy to generate random variables from the distribution,
- it should be easy to compute the likelihood ratio between the true and sampling distributions (at least up to a constant multiple), and
- the sampling distribution should give low-variance estimates.

The first requirement is that the sampling distribution g give consistent answers, which requires that g dominate f ($g > 0$ when $f > 0$), or for the integration estimate that g dominate $f|\theta|$.

The second requirement falls largely outside the scope of this work. The question of how to generate pseudo-random variables with different distributions quickly and efficiently on a computer is the subject of a great deal of research. We do consider this requirement in the discussion of some specific sampling methods.

The third requirement further limits the selection of eligible distributions. There are distributions for which fast algorithms to generate values, but for which no good algorithms to compute the density exist, e.g. stable distributions other than the Gaussian and Cauchy distributions (Chambers, Mallows, Stuck 1976). These distributions are not suitable for importance sampling.

In some cases it is possible to compute a likelihood ratio only up to a constant multiple,

$$\frac{g(x)}{f(x)} = cr(x),$$

where c is unknown. This can occur if the true distribution is known only up to a constant multiple, as is the case in some applications of importance sampling to Bayesian calculations. It can also occur if the density of the sampling distribution can be computed only up to a constant multiple. The integration and regression methods are not suitable for these proportional likelihood applications. If possible this situation should be avoided, by using sampling distributions for which densities are fully known. Otherwise only the ratio estimate can be used, for which the unknown constant cancels out.

The remainder of this chapter is devoted to a discussion of the fourth requirement—how to choose distributions with good statistical properties. We discuss two strategies for distribution selection that correspond to the integration and sampling approaches to importance sampling. Optimal distributions are generally impossible to find, so we discuss heuristic rules in distribution selection that, if followed, minimize the harm caused by departures from optimality. The heuristic rules form the basis for some specific sampling methods discussed in the next chapter.

5.1 Two Strategies

Two grand strategies that can be taken when trying to choose a sampling distribution are to:

- transform the simulation output into something nearly constant, or to
- sample extreme values more often.

Often, but not always, the two principles coincide.

The "transformation" approach underlies the integration estimate, which is based on the equality

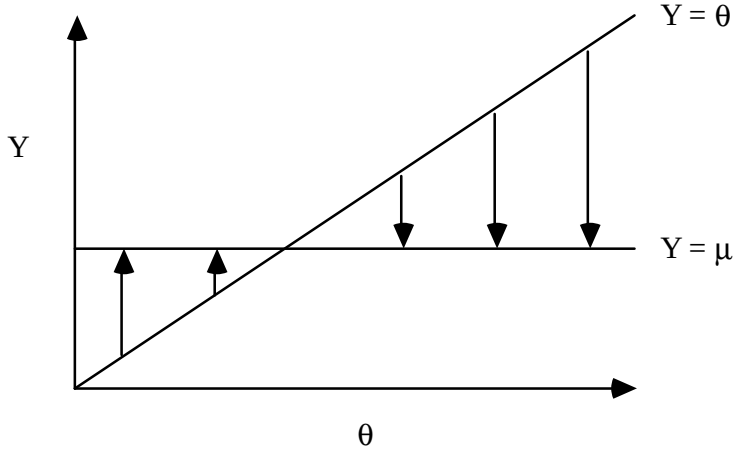
$$E_f(\theta(X)) = E_g(Y(X)) \tag{5.1}$$

where

$$Y(X) := \theta(X)\frac{f(x)}{g(x)} \tag{5.2}$$

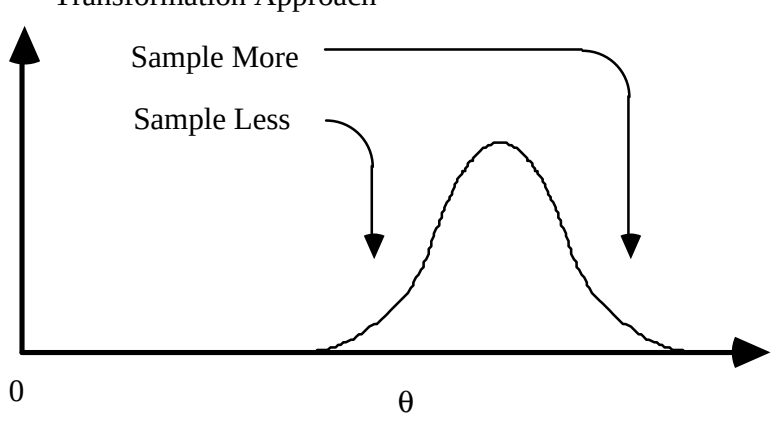
The goal is to choose a sampling distribution for which the transformed output Y has nearly constant variance.

Figure 5.1: Ideal Transformation $\theta \to Y$, Y constant



This goal is achieved by sampling less where θ is close to zero and more where θ is far from zero, as in Figure 5.2.

Figure 5.2: Ideal Sampling Distribution, Transformation Approach



The "sampling" approach is to try to sample so that extreme values are observed relatively frequently, but then given proportionately smaller weight in the computation of results. This is the same as the reasoning behind choosing the number of observations to be taken from each stratum in non-proportional stratified sampling—by concentrating the sampling effort in difficult parts of the sampling space, the quality of the estimate can be improved.

Both approaches involve sampling relatively often where $\theta(X)$ is "large," but what that means is different for the two approaches. For the transformation approach "large" means "far from zero". For the sampling approach "large" means "far from the average".

We demonstrate these two principles in Examples 5.1–5.3.

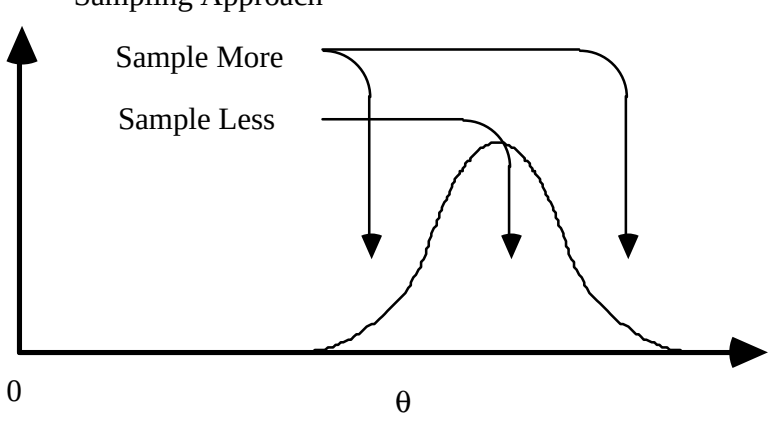
Example 5.1 (Therneau 1981)

X has a standard uniform distribution, so $f(x) = 1$ if $0 < x < 1$, otherwise $f(x) = 0$. $\theta(x) = x(1-x)$, $\mu = E_f(\theta(X)) = 1/6$. The variance of the simple Monte Carlo estimate is $1/(180n)$.

The transformation approach is to choose $g(x)$ so that $Y(x) = \theta(x)f(x)/g(x)$ is constant. This can be done in this example by choosing $g(x) = 6x(1-x)$. Then integration estimate has zero variance.

The sampling approach is to sample relatively often where $\theta(x)$ is far from

Figure 5.3: Ideal Sampling Distribution, Sampling Approach



its expected value. This occurs at the two endpoints of the interval, and to a lesser extent in the middle. The sampling distribution $g(x) = c|\theta(x) - 1/6|$ is optimal for the ratio estimate, where c is chosen so that $\int g(x)dx = 1$. This results in a 26% variance reduction.

θ and the two sampling distributions are shown in Figures 5.4–5.6.

Example 5.2 Two approaches when θ is not strictly positive

As in Example 5.1, X has a standard uniform distribution, but now $\theta(x) = x(1-x) - 1/4$. The transformation approach distribution changes to $g(x) = 12(x - 1/2)^2$. This sampling distribution is concentrated on the endpoints of the interval rather than in the middle as in Example 5.1.

The sampling approach distribution, and the variance reduction achieved, remain the same as in Example 5.1.

How do the two approaches compare? The sampling distribution obtained using the sampling approach seem more natural. If the idea behind importance sampling is to sampling relatively often from "important" regions, then it seems intuitive that importance should be defined in terms of how far from the average any particular result is, rather than how far from zero. It also is discomforting that the addition of a constant to the θ changed the optimal sampling distribution so radically for the integration estimate. On the other hand, the transformation/integration approach performs better.

Figure 5.4: θ in Therneau's Example

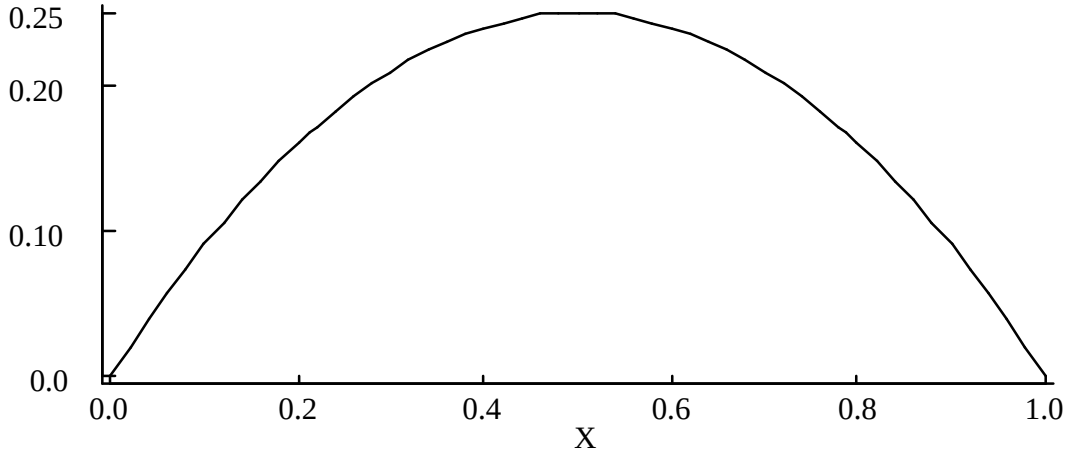


Figure 5.5: Optimal Transformation Approach Density in Therneau's Example

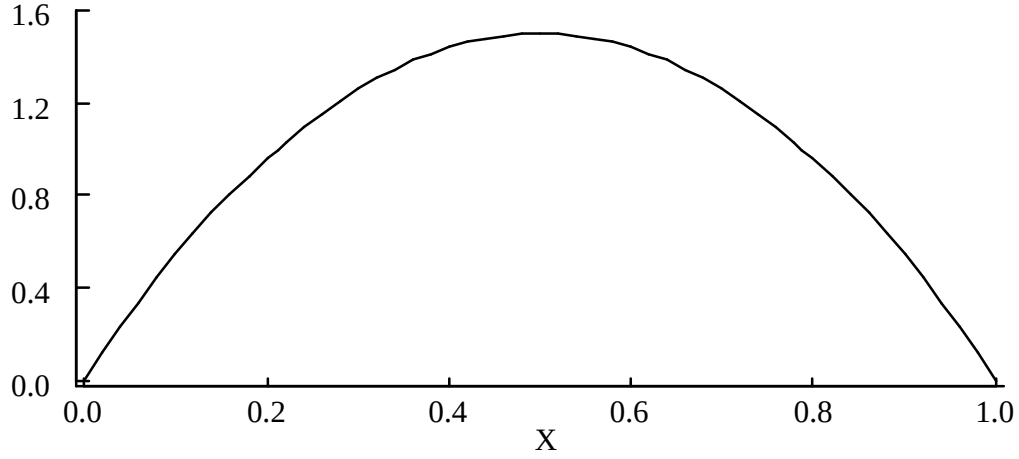
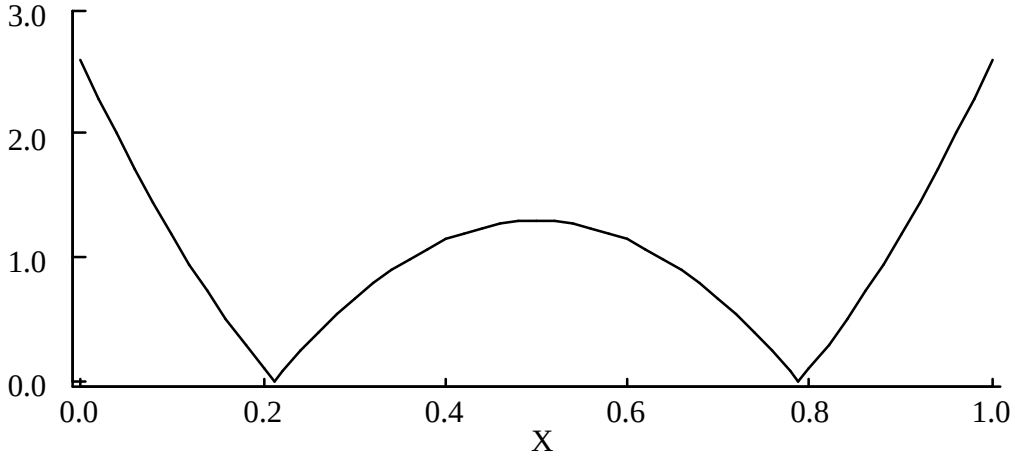


Figure 5.6: Optimal Sampling Approach Density in Therneau's Example



The difference between approaches is apparent as well in Example 5.3. In this case the sampling approach performs better.

Example 5.3 Two approaches when θ has midrange zero

X has a standard uniform distribution, $\theta(X) = I(X < 0.001) - 0.5$. The transformation approach distribution is uniform on $(0, 1)$, the same as the true distribution, because all θ values are equally far from 0.

The sampling approach distribution samples equally often from the two intervals $(0, 0.001)$ and $(0.001, 1)$, in each case sampling uniformly on the interval. The variance reduction is approximately 99.6%.

Where does the regression estimate fit into the grand strategy scheme? The ideal regression estimate is based on the generalization of the transformation approach, to find g so that Y is nearly a linear function of W, i.e. that $Y(x) = \alpha + \beta W$, at least approximately. Note, however, that β is also a function of g. The joint optimization of β and g is not generally amenable to solution. We recommend choosing a sampling distribution for one of the integration or ratio estimates, depending on the application, then using the regression estimate as an improved estimate.

5.2 Avoid Cold Spots

Given that an optimal sampling distribution can not be chosen, the next question is what characteristics a good sampling distribution has. In what ways can a sampling distribution deviate from an optimal distribution without causing grave harm?

First of all, a sampling distribution should be conservative, in the sense that it provide adequate coverage everywhere, where the coverage of a region in the sampling space is the probability of choosing an observation from that region. Undercoverage of a region by 80% means that the region is sampled only 20% as often as it should be, and the experiment as a whole could require 5 times as many observations for the same quality results. Of course, the more one region is sampled, the less another is, and some miscoverage is unavoidable; the salient goal is to avoid coverage which is far too low.

Choosing a sampling distribution is like dressing for cold weather—to minimize heat loss with a fixed amount of insulation, it is best to spread the insulation evenly everywhere (this is very much on the author's mind, due to his impending move to Minnesota). If that is not possible then the next best strategy is to avoid cold spots. It is worse to severely undersample some region of the sampling space (and sample slightly too much everywhere else) than it is to slightly undersample large regions (and oversample in a small region). Consider the following example:

Example 5.4 Sampling distributions should avoid cold spots

X has a standard uniform distribution, $f(x) = I(0 < x < 1)$. $\theta(x) = -1$ if $x < 1/2$, otherwise $\theta(x) = +1$. The optimal distribution for both the ratio and integration estimates is the uniform distribution (the best importance sampling is none at all).

Consider the choice between two sampling distributions. $g_1(x) = 1.1 - I(x < .1)$, and $g_2(x) = 0.9 + I(x < .1)$. g_1 samples too little in a small region, while g_2 samples too much in a small region. The variance of the integration estimate is $20/11$ under g_1, and is $20/19$ under g_2. We see that sampling too much in a small region is harmless, with a 5% variance increase, while sampling too little in a small region results in an 82% variance increase.

The phenomenon of cold spots can be analyzed another way. All the estimates discussed in the previous section are weighted averages, with weights approximately equal to $f(X_i)/(ng(X_i))$. If $g(x)$ is smaller than in should be in some region, then the ratio $f(x)/g(x)$ becomes too large, resulting in

large weights for replications that are sampled from that region. Variance increases occur because weighted averages are sensitive to a small number of large weights.

In Example 5.4 the optimal sampling distribution is a uniform distribution. When the optimum distribution is not uniform we can define a new measure such that a sampling distribution can be compared to a uniform distribution in that measure. The asymptotic variances for the integration and ratio estimates are:

$$\mathrm{Var}(\hat{\mu}) = \int \frac{f(x)^2(\theta(x)-c)^2}{g(x)}dx - (\mu - c)^2 \tag{5.3}$$

where $c = 0$ or $c = \mu$ for the integration and ratio estimates, respectively (or β for a regression estimate with a fixed β). The optimal sampling distribution is $g^*(x) = C|\theta(x) - c|f(x)$, where C normalizes the distribution. We can re-express the variance using the relationship between g and g^*, to obtain:

$$\mathrm{Var}(\hat{\mu}) = \int \frac{g^*(x)^2}{C^2 g(x)}dx - (\mu - c)^2 \tag{5.4}$$

The integral is minimized by choosing $g(x) = g^*(x)$, i.e., if g is uniform with respect to the g^* measure. Otherwise we define a "variance inflation factor" as:

$$\mathrm{VIF}(g) = \int \frac{g^*(x)^2}{g(x)}dx. \tag{5.5}$$

Then

$$\mathrm{Var}_g(\hat{\mu}) = \mathrm{VIF}(g)\mathrm{Var}_{g^*}(\hat{\mu}) + (\mathrm{VIF}(g) - 1)(\mu - c)^2. \tag{5.6}$$

Inspection of the form of the variance inflation factor shows that the integral is most sensitive to changes in g when g is close to zero, and the integral is large if g is close to zero over some region of the g^* measure.

In practical terms, distributions which avoid cold spots can be built using mixture distributions discussed in Chapter 6, and by choosing sampling distributions which have tails at least as heavy as the true distribution.

Inspection of (5.6) shows that the ratio estimate is the less sensitive than the integration estimate to non-optimal sampling distributions, as measured on the VIF scale. (5.6) applies to the regression estimate only if β is fixed at the same value for different sampling distributions.

5.3 Bounded Weights

A second criterion for good sampling distributions is that the inverse likelihood ratio $W(X) = f(X)/g(x)$ be bounded, say $W(x) < M$ if $f(x) > 0$ (see Bratley, Fox and Schrage, 1983, for a similar recommendation). We call this a "bounded weight" property.

The bounded weight property implies that $g(x)/f(x) > 1/M$. This is similar to the criterion that there be no cold spots, in that this implies that g not be "too small" anywhere, but now the comparison is between g and f rather than between g and g^*. This criterion can be interpreted as a minimax variant of the cold spot criterion—it ensures that for *any* θ, or component of multivariate θ, that the sampling distribution will not be worse than M times as bad as f.

The properties of bounded weights and avoiding cold spots (for a particular θ) can be satisfied simultaneously. Suppose that g_1 satisfies $g_1(x) \geq mg^*(x)$, where $m > 0$. Let $g_2(x) = 1/2(f(x) + g_1(x))$. This satisfies both $g_2(x) \geq m/2g^*(x)$ and $g_2(x) \geq 1/2f(x)$, so the weights are bounded by 2 and no spot is more than "twice as cold" under g_2 as it was under g_1.

5.3.1 Robustness

The primary advantage of bounded weights is that estimates are more robust. No matter what the relationship between X and $\theta(X)$ is, the asymptotic variance of an expectation estimate (or of a component of multivariate θ) is no worse than M times the variance that would be obtained without importance sampling, as long as either the ratio or regression estimates is used. This makes such sampling distributions particularly useful in applications where more than one quantity is being estimated. The proof is simple. For the ratio estimate the asymptotic variance is:

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{ratio}}) &= \int f(x)W(x)(\theta(x) - \mu)^2 dx \\ &\leq M \int f(x)(\theta(x) - \mu)^2 dx \\ &= M\mathrm{Var}(\hat{\mu}_{\mathrm{srs}}) \end{aligned} \tag{5.7}$$

and the corresponding result holds for the regression estimate, since its asymptotic variance is no greater than for the ratio estimate.

The same result does not hold for the integration estimate. A counterexample is trivial—if $\theta \equiv c$, the variance of the estimate is zero under simple

random sampling, or when using the ratio or regression estimates, but the integration estimate has variance $c^2\mathrm{Var}(W)$. A weaker result does hold, namely that:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\mathrm{int}}) &= \int f(x)W(x)\theta(x)^2dx - \mu^2 \\ &\leq M\int f(x)\theta(x)^2dx - \mu^2 \\ &= M\mathrm{Var}(\hat{\mu}_{\mathrm{srs}}) + (M-1)\mu^2 \end{aligned} \tag{5.8}$$

Why does the bounded weight criterion improve robustness? It ensures that no area of the sampling region is sampled less than $1/M$ times as often as it would be without importance sampling. Importance sampling always involves a tradeoff, sampling more in some regions of the sampling space and less in others, with the regions of heavy sampling chosen to reflect where "important" results are expected. If the problem is not completely understood then "important" results could occur in regions of light sampling. We illustrate this in Example 5.5.

Example 5.5 Sampling distributions for applications which may not be well understood

X_1 and X_2 are independent with uniform distributions, and $\theta(X) = I(X_1 > 0.9)+I(X_2 > 0.99)$. The expected value is $\mu = 0.11$, and the variance without importance sampling is 0.0999.

We compare two sampling distributions: $g_1(x) = 3x^2$, and $g_2(x) = 2x^2 + 1/3$, which are good distributions if $\theta(X) = I(X_1 > 0.9)$. If that were true the variance of the integration estimate would be 0.027 under g_1 and 0.037 under g_2, compared to 0.09 under f. In fact the variance of the integration estimate is infinite under g_1. Under g_2, which has $W(X) \leq 3$, the variance is 0.061, so a 39% reduction was achieved in spite of the misunderstood problem.

5.3.2 Other Advantages

There are other advantages to bounded weights:

- No observation is given excessive weight in computation of results.
- The distribution of W has finite variance.

- $\bar{W}$ tends to be close to 1.
- The integration, ratio, and regression estimates are more similar.
- The regression estimate is consistent.
- Regression weights are less likely to be negative.

These advantages, taken together, imply that results are more likely to be reasonable with bounded weights.

All the estimates we have discussed can be expressed as weighted averages, with weights roughly proportional to $W(X)$. By bounding the $W(X)$ function, we ensure that no weights are excessively large. If $W(X) \leq M$ then

$$\limsup_{n\to\infty} \frac{\max(W_i)}{\text{mean}(W_i)} \leq M \text{ a.s.} \tag{5.9}$$

The statement follows since $\text{mean}(W_i) \to 1$ a.s. (by the Law of Large Numbers) and $\max W_i \to \sup(W)$ a.s. The same statement holds if the finite sample weights $V_{\text{est},i}$ are substituted for the weight function values W_i, where "est" is one of integration, ratio, or regression.

The variance of W is finite since W is bounded above by M and below by 0, and $\text{Var}(W) \leq M - 1$. This helps ensure that the average W value is close to 1, which in turn helps ensure that the integration, ratio, and regression estimates will be similar; the three estimates are the same if $\bar{W} = 1$.

Bounded weights also helps ensure consistency and asymptotic normality of the final estimates, because if W is bounded then Y has as many finite moments as θ.

Finally, the regression weights are less likely to be negative if the distribution of W is bounded. This is primarily a phenomenon of small sample sizes and W distributions with heavy tails; the latter is generally less of a problem if W is bounded.

5.4 Memoryless Weight Function

A common denominator of many successful applications of importance sampling is that the weight function is memoryless, in the sense that the weight function depend only on an appropriate summary $S(X)$ of the input X, not how S was obtained. For example, if S is a linear combination of the input

and some function of that linear combination is a good approximation to θ, then all combinations of the input that result in the same values of S should have the relative likelihood under g compared to f, and the same weight.

This is the flip side of the conditioning method for weights discussed in Chapter 3. There the weight functions are not memoryless, but it was possible to replace the weights with their conditional expected values so that the conditioned weights are memoryless. With a memoryless weight function that conditioning is unnecessary. Note, too, that in order to condition the weights on S, S must be a sufficient statistic. That is not required for correct results in choosing a sampling distribution, though performance is best if S is an approximate sufficient statistic (some function of S should be a good approximation for θ).

The practical benefit of memoryless weights is that the variance of the weights, and the conditional variance of the weights given θ, is smaller than for a similar non-memoryless system, which tends to reduce the variance of estimates (as we saw in the discussion of conditioned weights).

Johns (1987) uses a sampling distribution with a memoryless weight function for obtaining bootstrap confidence intervals for robust location estimates. $X = (X_1, X_2, \ldots X_d)$ is a vector of length d with components distributed $\overset{i.i.d.}{\sim} f_1$ under the true distribution, where f_1 is an empirical distribution with weight $1/d$ on d points $(z_1, z_2, \ldots z_d)$ (the original sample in a bootstrap application). $\theta(X)$ is a estimate of location, such as the mean, median, or an M-estimate (Andrew, et al. 1972). An influence function argument yields an approximation for θ,

$$\theta(X) \approx S(X) := \theta(z) + \frac{1}{d}\sum_{j=1}^{d} \mathrm{IF}(X_j; f_1) \tag{5.10}$$

Johns uses a sampling distribution g_1 with *i.i.d.* component distributions formed by putting weight

$$\alpha e^{\beta \mathrm{IF}(z_j)} \tag{5.11}$$

on point z_j. $\alpha = \alpha(\beta)$ is a normalizing constant, and β determines the degree to which the distribution is "tilted". The resulting weight function is:

$$\begin{aligned} W(X) &= \prod_{j=1}^{d} \alpha \exp(-\beta \mathrm{IF}(X_j; f_1)) \\ &= \alpha^d \exp(-\beta \sum_{j=1}^{d} \mathrm{IF}(X_j; f_1)) \end{aligned}$$

$$= \alpha^d \exp(-\beta(S(X) - \theta(z)))$$

which depends on X only through S.

This sampling distribution produces a weight function which is memoryless with respect to the linear approximation for any choice of β.

The particle scattering experiment described in Chapter 3 is another example of a memoryless weight function. With the sampling distribution used there, the weight for a replication is a function of only the x-coordinate of a particle, independently of the number and position of prior collisions. Thus the system is memoryless with respect to $S(X) = (x\text{-coordinate of the particle})$. Note that it is not memoryless with respect to $S^{(2)}(X) = \min(0, x\text{-coordinate})$, since the weight function depends on both $S^{(2)}$ and the overshoot, and this shortcoming leaves room for improvement by conditioning the weights.

Siegmund (1976) discusses the use of exponential tilting in the study of sequential tests, and finds that the asymptotically optimal sampling distribution is the one that corresponds to a memoryless weight function. His example involves a random walk; $Z_1, Z_2, \ldots \overset{i.i.d.}{\sim} f_1$ with $-\infty < E_{f_1}(Z_1) < 0$ and $P_{f_1}(Z_1 > 0) > 0$. Let $S_t := Z_1 + Z_2 + \ldots + Z_t$, and for $a \leq 0 < b$ let $T := \inf\{t : t \geq 1, S_t \notin (a, b)\}$. The goal is to estimate is $P_f(S_T \geq b)$. Siegmund considers sampling distributions of the form

$$g_1(z) = \alpha e^{\beta z} f_1(z), \tag{5.12}$$

where $\alpha = \alpha(\beta)$ is a normalizing constant and β determines the amount of exponential tilting. He finds that the choice β^* that solves

$$\alpha(\beta^*) = 1 \tag{5.13}$$

is asymptotically optimal (subject to the additional conditions given there), in the sense that the relative efficiency of any other β goes to zero exponentially fast as $b \to \infty$. For β^*,

$$W = \prod \frac{f_1(Z_t)}{g_1(Z_t)} = e^{-\beta S_T} \tag{5.14}$$

which depends only on S_T.

A common thread in these three examples is that the sampling method is based on exponential tilting, and under certain conditions exponential tilting is the only method that produces memoryless weights. This is discussed further in Chapter 6.

Chapter 6

Specific Sampling Methods

We now discuss specific methods of generating sampling distributions. This discussion will not be exhaustive, due to the impossibility of recommending methods that will apply in *all* simulation applications that will be encountered. Instead we concentrate on a small number of general techniques that can be applied in many situations, including mixture sampling, mixture distributions, exponential tilting, and internal distribution specification.

We focus on methods that can be applied in difficult, multivariate applications. Finding a good sampling distribution in examples with univariate input is relatively easy. We don't generally do importance sampling, or even Monte-Carlo estimation in this case, however, as there are other techniques which are more efficient, including numerical integration and stratified sampling. Importance sampling offers the greatest benefit in applications with multivariate input where the output is a complicated function of the input, applications which are too complex to be handled by other methods.

Unfortunately the complexity of many applications makes finding an optimal sampling distribution impossible, and even finding a good one difficult. Some authors have recommended against using importance sampling because of problems they have encountered with multivariate input (Wilson 1984, Bratley, Fox and Schrage 1983).

We do not agree with their recommendations. It is possible for importance sampling to be very robust. The simplest way to achieve this is to use mixture sampling in combination with either the ratio or regression estimate.

6.1 Mixture Sampling

Mixture sampling depends on the idea of generating random variables using mixtures of distributions (Marsaglia 1961, Kennedy & Gentile 1980). Express a sampling distribution g as a linear combination of other distributions:

$$g(x) = \sum_{k=1}^{K} \lambda_k g_k(x), \tag{6.1}$$

where $\sum \lambda_k = 1$ and $\lambda_k \geq 0$. X can be generated by first choosing one of distributions, with probability λ_k for distribution k, then generating X according to g_k.

We are interested here in the special case where the sampling distribution is a mixture of the true distribution f and an alternate sampling distribution $g_0(x)$:

$$g_\lambda(x) = \lambda f(x) + (1 - \lambda) g_0(x), \tag{6.2}$$

with $0 \leq \lambda \leq 1$. We call this *mixture sampling*.

Mixture sampling is not a substitute for other methods. It still requires that an alternative distribution g_0 be available (which may itself be a mixture distribution). As usual, such a distribution should be easy to generate and should sample relatively often from regions where θ is extreme.

The use of a mixture distribution g_λ ensures that g_λ/f is not too small. Essentially we allocate a certain proportion λ of the replications to be sampled from f, while the remaining observations are from g_0. This in turn ensures that

$$W(x) = \frac{f(x)}{g_\lambda(x)} \leq \frac{1}{\lambda}, \tag{6.3}$$

which for $\lambda > 0$ enforces the bounded weight criterion discussed in Section 5.3.

The problem with multidimensional importance sampling without mixture sampling is that even if the inverse likelihood ratio f_j/g_j is reasonably small for each dimension j of a multivariate distribution, the product can become very large. For example, if X is d-dimensional with independent marginal distributions under both f and g the inverse likelihood ratio is

$$W(X) = \prod_{j=1}^{d} f_d(X_d)/g_d(X_d). \tag{6.4}$$

If $d = 50$ and $f_j/g_j \leq 2$ for all j (a reasonably small bound), the product can be as large as 2^{50}.

Mixture sampling should be done on the distribution as a whole rather than separately for each marginal distribution of multivariate input. In the previous example, let $f(X)$ and $g_0(X)$ be the product distributions with marginals f_d and g_d, respectively. By letting $g_\lambda(x) = \lambda f(x) + (1-\lambda)g_0(x)$ we obtain the bound $W(x) \leq 1/\lambda$. If instead we use $g_\lambda^{(2)}(x) = \prod(\lambda f_d(x_d) + (1-\lambda)g_d(x_d))$ we obtain only the bound $W(x) \leq 1/\lambda^D$. The difference is substantial.

Note that $W(X)$ is computed without regard to which distribution was actually used to generate X. This may seem as if we are throwing away information; why not compute W as $f(X)/g_k(X)$, where g_k is the distribution that was actually used to generate X? In fact, the distribution actually used is irrelevant, and basing W on one of the distributions rather than the mixture brings superfluous randomness to the computation of W. To see this, note that choosing a component k and generating X according to that component is distributionally equivalent to generating X according to the mixture distribution, then picking k randomly according to the conditional probabilities of the two components given X. If W is computed based on both X and k, the expected value of $W(X,k|X)$ is equal to $f(X)/g_\lambda(X)$, but the conditional distribution of W has (may have) nonzero variance.

Expressed another way, X is a sufficient statistic for W, and the Rao-Blackwell theorem indicates that replacing an estimate $(W(X,k))$ with its conditional expectation given a sufficient statistic X results in the same expected value and a lower variance.

Mixture sampling answers the criticism by Bratley, Fox, & Schrage (1983) and others of multidimensional importance sampling. Our bounded weight criterion is the same as their recommendation that $f/g \leq c$, where c is a small constant, to guard against gross misspecification of g. They warn that "Because there is no practical way to guard against gross misspecification of g, multidimensional importance sampling is risky." Mixture sampling does provide such insurance.

Mixture sampling can be used with either the ratio or regression estimate to ensure that the results are robust—the asymptotic variance of these estimates is no worse, for any component of multivariate θ, than $1/\lambda$ times the variance that would be obtained without importance sampling. The same result does not hold for the integration estimate—a simple example is if θ is

a nonzero constant, the integration estimate has nonzero variance.

The results of using mixture sampling in the Gaussian probability example considered in Chapter 2 are shown in Table 6.1. The integration estimate of μ_1 has a higher MSE for $\lambda \neq 0$, but every other combination of estimate and estimand does better with $\lambda > 0$.

The estimates become more similar as λ becomes larger, because the variance of W decreases as $\lambda \to 1$; all three estimates are the same when $\bar{W} = 1$.

Reducing the variance of W reduces the bias of the ratio and regression estimates. We can see this in the case of μ_4 in the Gaussian probability experiment, where estimated bias of the ratio estimate using the three λ values $(0, 0.1, 0.5)$ is $(0.61, 0.042, 0.0035)$ and for the regression estimate was $(0.54, 0.035, 0.0037)$.

Using a larger λ has other benefits in this example, namely in making the integration and regression estimates reasonable. The minimum (of 2000) integration estimates of $E(X)$ was $(-42.4, -1.6, -.58)$ for the three λ values. The value -42 is clearly unreasonable. The minimum regression estimate of $P(X > z_\alpha)$ is -0.03 for $\lambda = 0$, which is of course impossible.

6.1.1 Choosing the Mixing Parameter

What is a good choice of λ in the mixture sampling method? In this section we give some guidelines and examples, and a number of inequalities for the variance resulting from different mixing parameters; these bounds show that an optimal choice of λ is not critical, and that some degree of mixture sampling results in very little efficiency loss even when the alternate sampling distribution g_0 is optimal.

If the problem to be solved does not have clear mathematical structure then λ should probably not be less than 0.1, so 10% of the observations should be from the true distribution and no W is greater than 10. This small λ is a good choice in simple simulation applications as an alternative to not using mixture sampling. This makes the choice of sampling parameter more robust and is an inexpensive way to protect against the contingency that the problem is not as well understood as expected. This is demonstrated in Example 6.1.

Example 6.1 Structural Failure Analysis

A physical structure can be subjected to a number of different random

Table 6.1: Mean Square Error using Mixture Sampling in Gaussian Probability Example

	λ	Integration	Ratio	Regression
$\mu_1 = P(X > z_\alpha)$	0	.0000065	.00035	.000020
	0.1	.0000072	.000021	.0000071
.00025 for SRS	0.5	.000013	.000016	.000012
$\mu_2 = P(X \leq z_\alpha)$	0	2.2	.00035	.000020
	0.1	.056	.000021	.0000071
.00025 for SRS	0.5	.0062	.000016	.000012
$\mu_3 = 1$	0	2.2	0.0	0.0
	0.1	.056	0.0	0.0
0 for SRS	0.5	.0062	0.0	0.0
$\mu_4 = E(X)$	0	4.39	.66	.40
	0.1	.11	.12	.098
.025 for SRS	0.5	.028	.029	.027

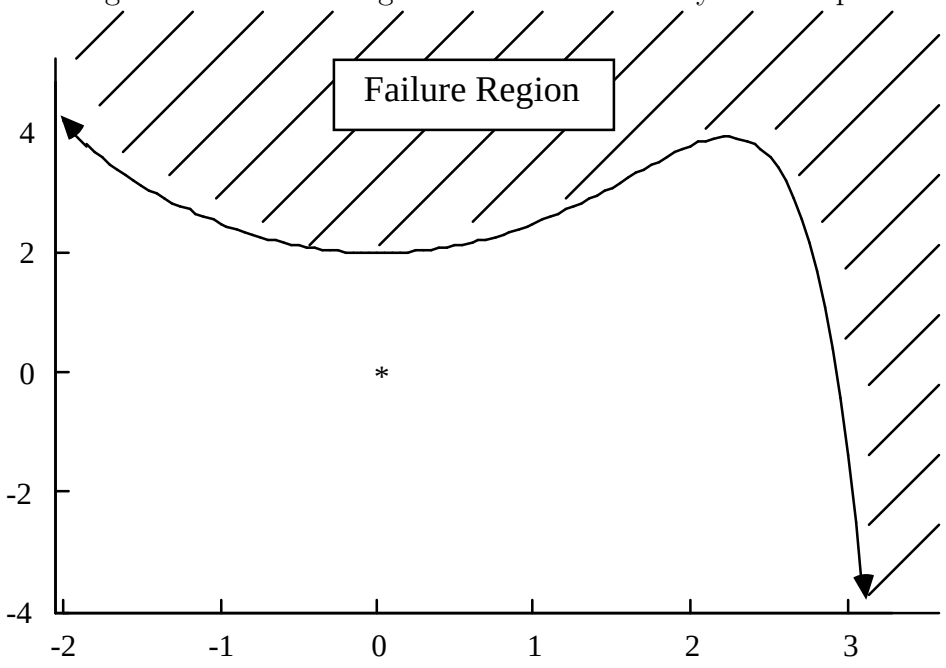
$X \sim f = N(0,1)$, $g = N(z_\alpha, 1)$, $z_\alpha = 2.326$, sample size = 1. There are 40 replications in each Monte Carlo experiment and 2000 Monte Carlo experiments. This is the same as Table 2.2, except that three values of λ are used. The reference values (variance under simple random sampling) are listed below the description of μ.

stresses, and fails if the combined stresses fall in certain regions of their domain. In the example we consider, X_1 and X_2 are independent normally distributed stresses, and the structure fails if (X_1, X_2) falls in the failure region

$$A = \left\{(x, y) : y > 2 + \frac{x^2}{2} - \frac{x^9}{1000}\right\} \tag{6.5}$$

as shown in Figure 6.1.

Figure 6.1: Failure Region in Structural Analysis Example



The most likely way for a structure to fail is for X_2 to be greater than two, so a reasonable importance sampling distribution appears to be X_1 standard normal and X_2 normal with mean 2 and variance 1.

The failure region is more complex, however, and the structure can fail for moderate X_2 values if X_1 is large.

The estimated efficiency using a number of mixing proportions is given in Table 6.2. The probability of failure is 0.0138 (standard error 0.0003). The sampling distribution $N((0, 2), I)$ is five times worse than using no impor-

Table 6.2: Efficiency in Structural Analysis Example

	Estimated Efficiency			St. Error of Efficiency Estimate		
Estimation Method:	Int	Ratio	Reg	Int	Ratio	Reg
Mixing Parameter λ						
0	4.95	5.55	4.94	0.21	0.22	0.21
0.05	0.84	0.94	0.84	0.04	0.04	0.04
0.10	0.59	0.65	0.59	0.03	0.03	0.03
0.20	0.41	0.46	0.41	0.016	0.016	0.017
0.40	0.31	0.34	0.30	0.010	0.010	0.011
0.60	0.29	0.31	0.27	0.008	0.008	0.008
1.00	1.00	1.00	1.00	0.0	0.0	0.0

tance sampling, while using a mixture proportion of 0.1 improves the results to a variance reduction of 40%.

Obtaining accurate efficiency estimates is difficult, and indicates one way in which importance sampling can fail without appearing to do so. Using the sampling distribution with $(X_1, X_2) \sim N((0, 2), I)$, results in failures occurring the lower right corner of the failure region ($X_2 < 0$) very infrequently—only 33 replications in one million. But the weight function there is very large, $W = \exp(-X_2 + 2)$, an average of 21, whereas the average W value over the whole failure region is 0.040. This is even more of a rare-event application than the original example. The biggest danger is that this method will often fail without appearing to do so— less than four percent of all experiments of size 1000 would include any observations from the lower right part of the failure region.

The way to obtain the efficiency estimates here is to use importance sampling, but instead of sampling from the "official" sampling distribution to be analyzed, sample from a distribution which emphasizes the regions where the rare events take place, and use those results to estimate the efficiency for the official sampling distribution, using methods described in Chapter 4.

The actual sampling distribution is a mixture of multivariate normal distributions, with (stratified) mixture allocations and distributions given in Table 6.3. The first distribution is the original distribution, the second the "official" sampling distribution, and the third and fourth emphasize the right side of the failure region. This is, incidentally, a better sampling distribution

Table 6.3: Sampling Distribution Used in Structural Analysis Example

Observations	Distribution
1000	N((0,0), I)
6000	N((0,2), I)
2000	N((3,-2), I)
1000	N((3,0), I)

not only for estimating the efficiency of the "official" sampling distribution, but for estimating the failure probability as well, with efficiencies of (0.08, 0.16, 0.08) using the integration, ratio, and regression estimates, respectively.

Other λ In many applications a good choice is $\lambda = 1/2$, particular in applications which do not have a discrete mode at zero. This corresponds to sampling half of the observations from each of distributions f and g_0. No single value $W(x)$ will be more than twice as large as the average (expected) value. This is also easy to implement computationally, by taking every other observation from f.

There is an optimality argument in one special case for $\lambda = 1/2$. To estimate a probability with the ratio estimate, the optimal sampling distribution has

$$r^*(x) = \begin{cases} p/2 & \theta(x) = 1 \\ (1-p)/2 & \theta(x) = 0 \end{cases} \tag{6.6}$$

where $P(\theta(X) = 1) = p$ and $g(x) = r(x)f(x)$. As $p \to 0$, the smaller value of r^* goes to 1/2. If the alternate distribution g_0 has support entirely on the rare region where $\theta = 1$ and $r(x)$ is constant over that region, then $\lambda = 1/2$ is the limiting optimal value.

6.1.1.1 Inequalities for the Mixing Parameter Fortunately it is not critical that λ be chosen exactly at the optimum value. The variance as a function of λ is a locally quadratic near the minimum, if the minimum is not at 0 or 1. We provide a number of bounds here on the variance as a function of λ.

We begin by defining a generic variance function which we use to measure

the performance of a mixing proportion. Let

$$V(\lambda, c) = \int \frac{f(x)^2(\theta(x) - c)^2}{g_\lambda(x)} dx. \tag{6.7}$$

The asymptotic variance of the integration estimates are

$$\begin{aligned} \mathrm{Var}(\hat{\mu}; \lambda) &= \int \frac{f(x)^2(\theta(x) - c)^2}{g_\lambda(x)} dx - (\mu - c)^2 \\ &= V(\lambda, c) - (\mu - c)^2 \end{aligned} \tag{6.8}$$

where $c = 0$, $c = \mu$, or $c = \beta$ for the integration, ratio, and regression estimates respectively.

The inequalities we give below are for the generic variance function. The relationship between these inequalities is that if λ^* minimizes (6.7) and

$$V(\lambda, c) \leq MV(\lambda^*, c), \tag{6.9}$$

then

$$\mathrm{Var}(\hat{\mu}; \lambda) \leq M\mathrm{Var}(\hat{\mu}; \lambda^*) + (M - 1)(\mu - c)^2 \tag{6.10}$$

The generic variance function is directly proportional to the asymptotic variance of the ratio estimate. The integration estimate is more sensitive to non-optimal sampling parameters than is the ratio estimate, at least on the scale measured by the generic variance function. It is unclear how sensitive the regression estimate is, because β is generally a function of λ, rather than being fixed.

Theorem 6.1 gives bounds for the variance function under a number of conditions. The proof is in the appendix.

Theorem 6.1 *Inequalities for the Mixing Parameter in Mixture Sampling*

Let $V(\lambda) = V(\lambda, c)$ as defined in (6.7) and let λ^* be the value which minimizes $V(\lambda)$ in the interval [0,1]. Then the following inequalities hold for $0 \leq \lambda \leq 1$:

If $V(0) < \infty$ then

$$V(\lambda) \leq \frac{V(0)}{1 - \lambda}. \tag{6.11}$$

If $V(1) < \infty$ then

$$V(\lambda) \leq \frac{V(1)}{\lambda}. \tag{6.12}$$

If $V(0) < \infty$ and $V(1) < \infty$ then

$$V(\lambda) \leq \frac{V(0)V(1)}{\lambda V(0) + (1-\lambda)V(1)}. \tag{6.13}$$

If $0 < \lambda^* < 1$ then

$$V(\lambda) \leq V(\lambda^*) \left(\frac{\lambda^{*2}}{\lambda} + \frac{(1-\lambda^*)^2}{1-\lambda} \right). \tag{6.14}$$

If $0 < \lambda^* < 1$ and $V(0) < \infty$ then

$$V(\lambda) \leq V(\lambda^*) \left(\frac{a\lambda^{*2} - \lambda(\lambda^{*2} + 2(\lambda^* - 1)(a-1))}{(1-\lambda)(\lambda(a-1) + \lambda^{*2})} \right) \tag{6.15}$$

where $a := V(0)/V(\lambda^*)$.

If $0 < \lambda^* < 1$ and $V(1) < \infty$ then

$$V(\lambda) \leq V(\lambda^*) \left(\frac{b(1-\lambda^*)^2 - (1-\lambda)(\lambda^{*2} - 2b\lambda^* + b)}{\lambda((1-\lambda)(b-1) + (1-\lambda^*)^2)} \right) \tag{6.16}$$

where $b := V(1)/V(\lambda^*)$.

6.1.1.2 Adaptive Choice of the Mixing Parameter The choice of λ may be made before the experiment begins, or can be chosen in the course of a study by performing a trial study or using more general adaptive approach. It is not necessary to run separate trials for different values of λ to choose parameters; as in the structural analysis and fuel inventory examples, it is possible to estimate efficiency for different sampling distributions from a single set of replications.

A word of caution is in order—the use of an adaptive approach (including the use of a trial study) can lead to biased results, if the sampling decisions made in later stages of the experiment are allowed to affect the weights given to earlier observations.

In the structural analysis example, for example, a trial study may show no observations in the lower right failure region, and an adaptive decision would be to decrease λ, making it even less likely that any such failures would be observed. On the other hand, if such an event occurs it would become clear that a larger λ is needed. Then in the final analysis, if all observations are

treated equally (from the trial and subsequent study), the larger λ value results in a smaller weight for the failure event than it originally had.

This bias can be avoided by computing the final result as a weighted average of results from the initial and subsequent studies, with a weighting factor fixed in advance. If the trial study has n_1 observations and the subsequent study n_2, then let $\hat{\mu} = (n_1\hat{\mu}_1 + n_2\hat{\mu}_2)/(n_1+n_2)$, or choose other weights which reflect the expectation that the second study will be more efficient than the first.

6.1.2 Mixture Sampling in the Fuel Example

$\lambda = 1/2$ is the choice used in the Fuel Inventory example. It provides a good balance between estimating outage cost, inventory cost, total cost, and outage probability. Any value of λ between .1 and .9 would be reasonable, and perform better than $\lambda = 1$ (no importance sampling) and $\lambda = 0$ (no mixture sampling). This is demonstrated in Figures 6.2–6.7 and Table 6.4, which use the methods outlined in Section 4.4.1 to estimate the efficiency that would have been obtained using other choices of λ. Note that larger λ values are generally better for estimating inventory cost, and smaller λ for estimating outage probabilities and costs, especially for estimating the probability of a large outage.

6.1.3 Stratifying Distribution Allocations

Most of the discussion of mixture sampling to this point has assumed that the sampling distribution is a true mixture distribution, with a single replication generated by choosing one component distribution randomly and then generating X according to that component distribution.

In fact, it is better to fix the number of observations to be taken from each component, say the first $n\lambda$ observations from f and the final $n(1-\lambda)$ observations from g_0. This is first of all simpler to implement, since it eliminates the random selection step in the generation process.

Second, fixing the allocations gives better estimates. We note this first in the asymptotic variance formulas. The leading terms of the variance formulas are of the form:

$$\sigma^2_{\text{random}} = \text{Var}_{g_\lambda}(Y - cW) - (\mu - c)^2, \tag{6.17}$$

Figure 6.2: Integration Method Efficiency for Cost Estimates for $0 \leq \lambda \leq 1$

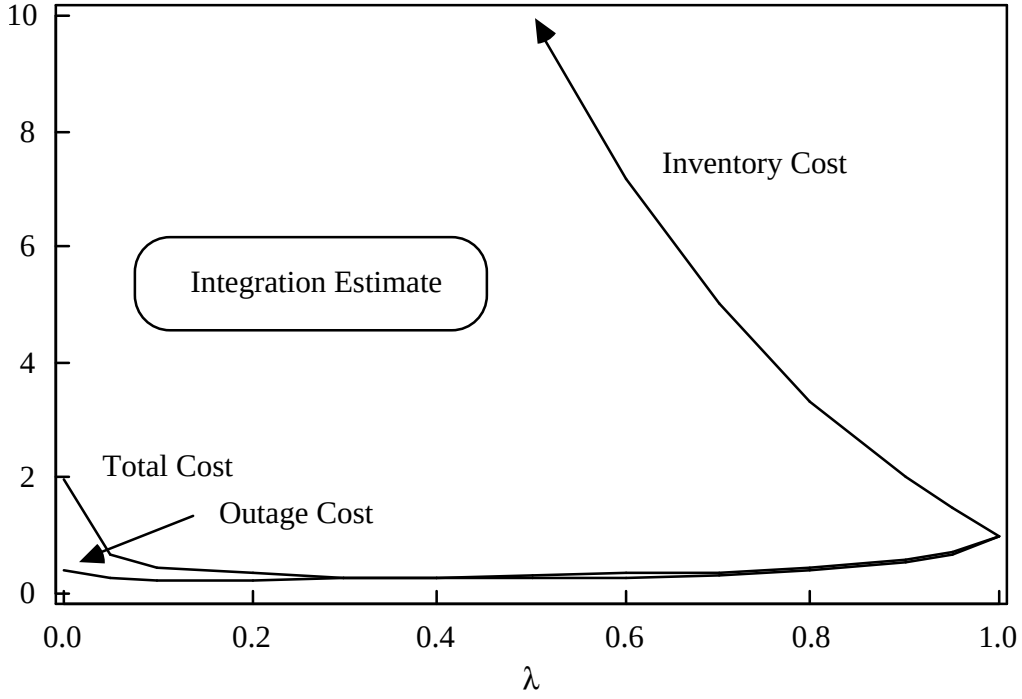


Figure 6.3: Ratio Method Efficiency for Cost Estimates for $0 \leq \lambda \leq 1$

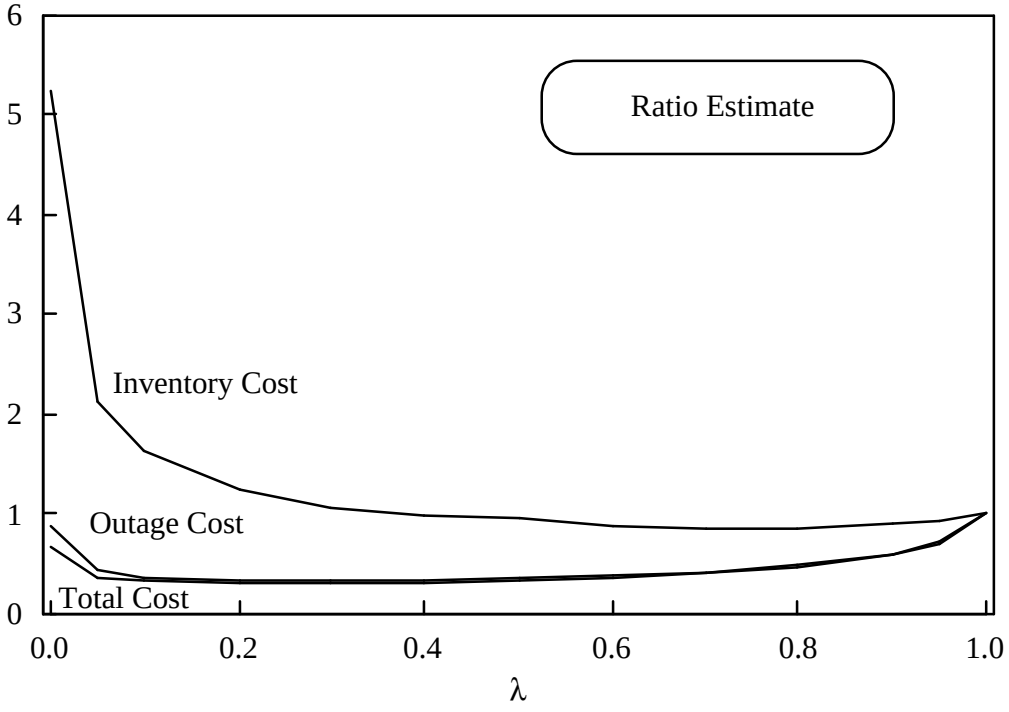


Figure 6.4: Regression Method Efficiency for Cost Estimates for $0 \leq \lambda \leq 1$

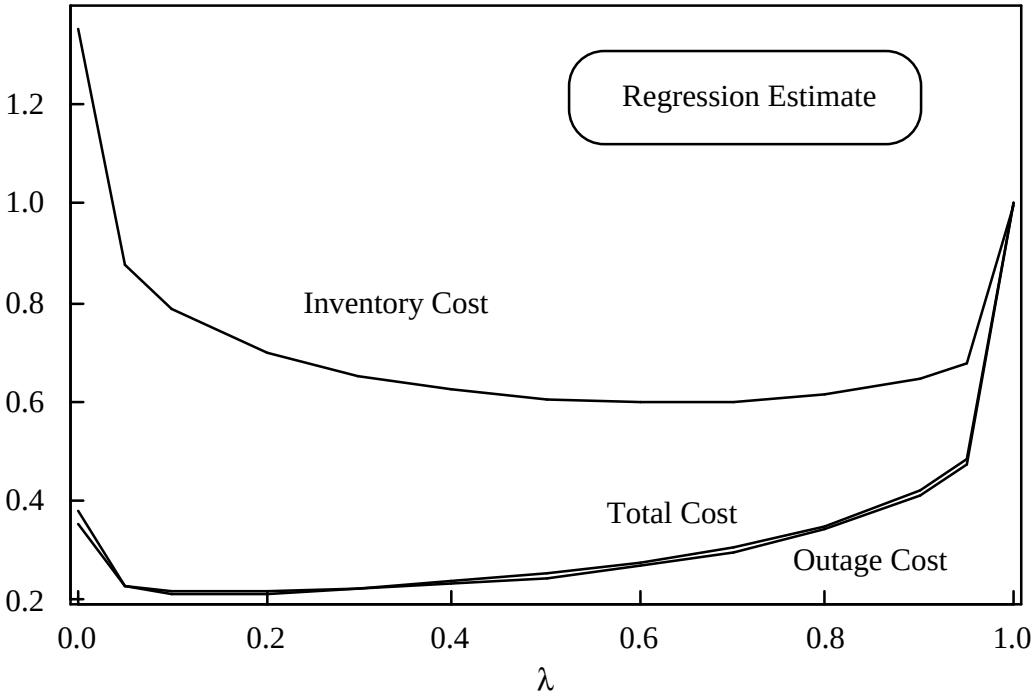


Figure 6.5: Integration Method Efficiency for Outage Estimates for $0 \leq \lambda \leq 1$

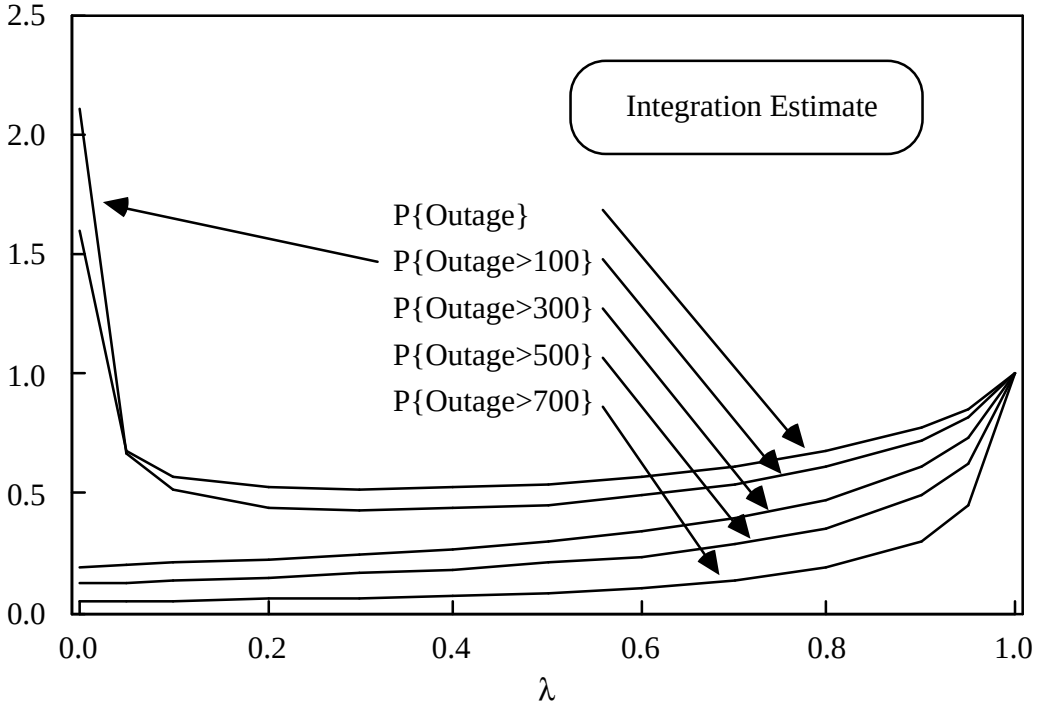


Figure 6.6: Ratio Method Efficiency for Outage Estimates for $0 \leq \lambda \leq 1$

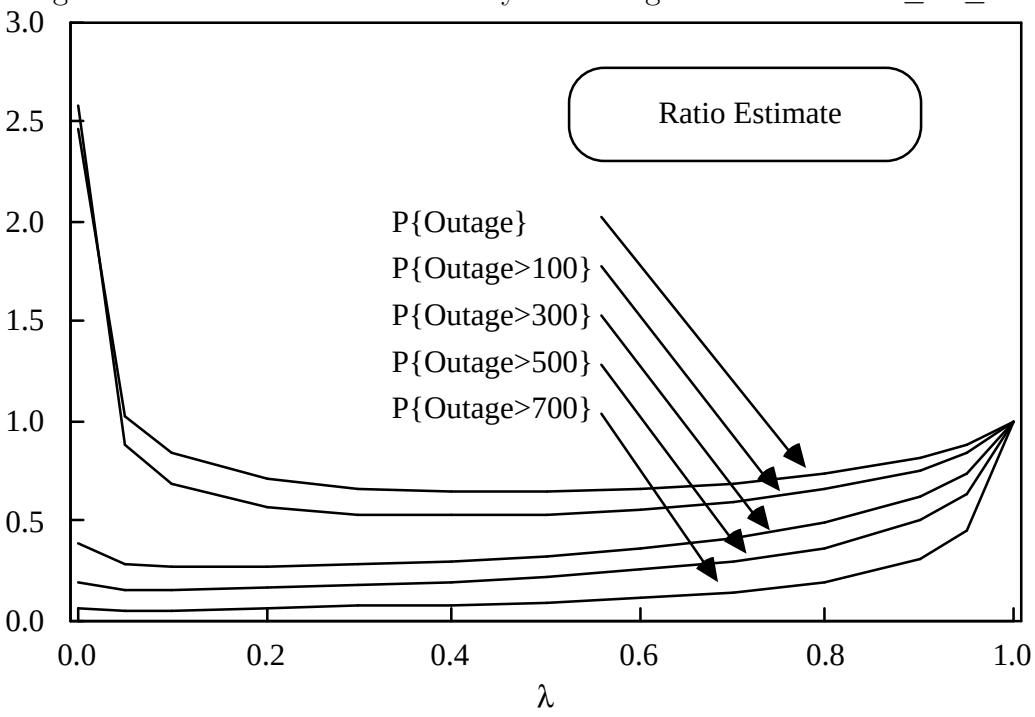


Figure 6.7: Regression Method Efficiency for Outage Estimates for $0 \leq \lambda \leq 1$

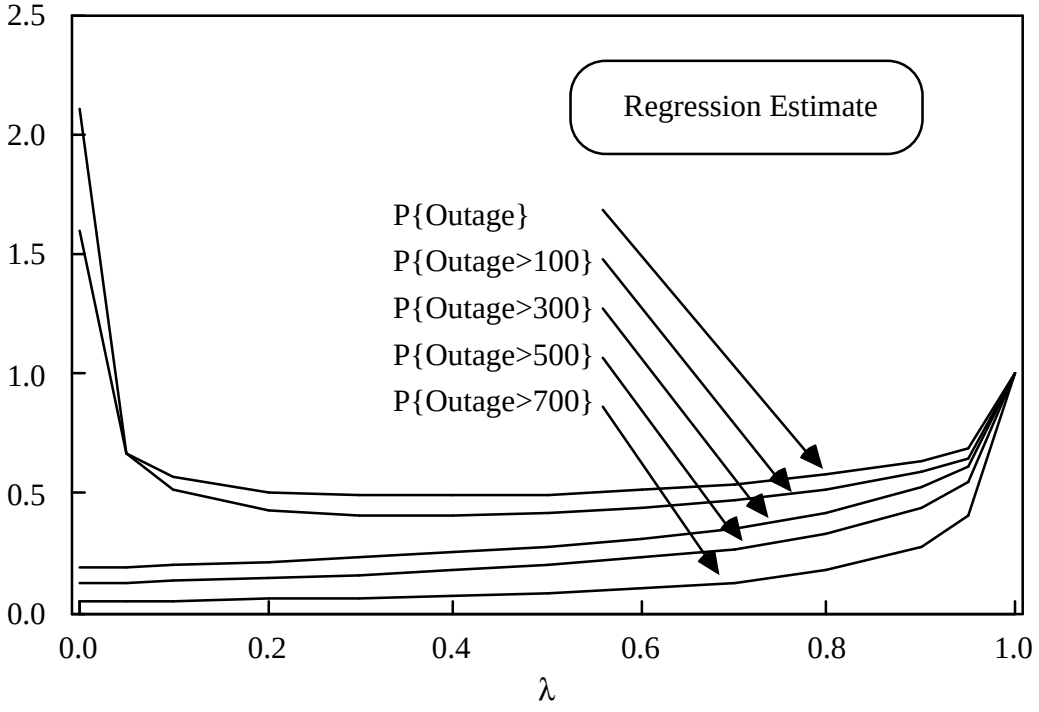


Table 6.4: Mixture Sampling Efficiency in Fuel Inventory Example
Var(estimate)/Var(simple random sampling)

Outage Cost	Estimated Efficiency			Std. Error of Efficiency Est.		
λ	Int	Ratio	Reg	Int	Ratio	Reg
0	0.381	0.876	0.381	0.186	0.148	0.186
0.05	0.233	0.435	0.229	0.039	0.032	0.041
0.1	0.223	0.375	0.216	0.026	0.022	0.027
0.2	0.227	0.338	0.214	0.020	0.018	0.021
0.3	0.241	0.331	0.221	0.019	0.018	0.020
0.4	0.261	0.338	0.233	0.019	0.018	0.020
0.5	0.284	0.351	0.247	0.019	0.018	0.020
0.6	0.322	0.380	0.271	0.020	0.019	0.021
0.7	0.370	0.420	0.300	0.020	0.019	0.021
0.8	0.443	0.484	0.342	0.020	0.019	0.022
0.9	0.572	0.600	0.412	0.019	0.018	0.023
0.95	0.694	0.713	0.472	0.016	0.016	0.024
1	1.000	1.000	1.000	0.0	0.0	0.0
Inventory Cost						
0	244.556	5.234	1.350	23.344	0.474	0.480
0.05	77.664	2.125	0.876	3.378	0.072	0.103
0.1	51.053	1.637	0.785	2.009	0.041	0.065
0.2	29.752	1.250	0.699	1.075	0.023	0.042
0.3	19.881	1.074	0.652	0.685	0.017	0.033
0.4	13.989	0.974	0.623	0.464	0.013	0.028
0.5	9.935	0.944	0.602	0.305	0.014	0.025
0.6	7.154	0.878	0.599	0.217	0.009	0.021
0.7	4.989	0.862	0.600	0.140	0.007	0.019
0.8	3.310	0.866	0.613	0.081	0.006	0.017
0.9	1.995	0.897	0.646	0.035	0.004	0.015
0.95	1.456	0.932	0.680	0.016	0.003	0.014
1	1.000	1.000	0.999	0.0	0.0	0.001

Total Cost	Estimated Efficiency			Std. Error		
λ	Int	Ratio	Reg	Int	Ratio	Reg
0	1.941	0.683	0.355	0.305	0.123	0.154
0.05	0.646	0.375	0.229	0.059	0.029	0.035
0.1	0.456	0.334	0.218	0.035	0.021	0.025
0.2	0.319	0.312	0.218	0.023	0.018	0.020
0.3	0.270	0.311	0.226	0.020	0.018	0.019
0.4	0.253	0.321	0.238	0.020	0.018	0.020
0.5	0.255	0.336	0.253	0.020	0.018	0.020
0.6	0.278	0.368	0.277	0.021	0.019	0.021
0.7	0.319	0.409	0.307	0.021	0.020	0.021
0.8	0.390	0.474	0.351	0.022	0.020	0.022
0.9	0.527	0.592	0.423	0.021	0.019	0.024
0.95	0.661	0.706	0.485	0.018	0.016	0.024
1	1.000	1.000	1.000	0.0	0.0	0.0
Outage Probability						
0	1.605	2.463	1.602	1.105	0.931	1.096
0.05	0.674	1.026	0.672	0.213	0.177	0.216
0.1	0.575	0.838	0.569	0.118	0.097	0.121
0.2	0.525	0.711	0.512	0.064	0.052	0.067
0.3	0.517	0.665	0.497	0.045	0.036	0.049
0.4	0.525	0.649	0.496	0.035	0.029	0.039
0.5	0.539	0.644	0.500	0.029	0.024	0.033
0.6	0.573	0.662	0.519	0.024	0.020	0.029
0.7	0.615	0.689	0.543	0.020	0.017	0.025
0.8	0.677	0.735	0.579	0.016	0.013	0.022
0.9	0.777	0.815	0.638	0.011	0.009	0.019
0.95	0.858	0.881	0.687	0.007	0.006	0.017
1	1.000	1.000	0.999	0.0	0.0	0.001
$P(\text{Outage} > 500)$						
0	0.130	0.190	0.129	0.031	0.027	0.031
0.05	0.135	0.159	0.134	0.031	0.030	0.031
0.1	0.140	0.159	0.139	0.032	0.031	0.032
0.2	0.153	0.167	0.151	0.034	0.033	0.034
0.3	0.168	0.180	0.165	0.035	0.035	0.035
0.4	0.187	0.197	0.182	0.037	0.037	0.037
0.5	0.210	0.220	0.203	0.038	0.038	0.039
0.6	0.243	0.252	0.233	0.041	0.041	0.042
0.7	0.289	0.296	0.273	0.043	0.043	0.044
0.8	0.359	0.365	0.334	0.044	0.044	0.045
0.9	0.493	0.498	0.447	0.042	0.041	0.045
0.95	0.629	0.633	0.555	0.035	0.034	0.041
1	1.000	1.000	1.000	0.0	0.0	0.0

where $c = 0$, $c = \mu$, and $c = \beta$ for the integration, ratio, and regression estimates, when the sample allocations are not fixed. If allocations are fixed at $n\lambda$ and $n(1-\lambda)$, then as $n \to \infty$, the first term of (6.17) becomes:

$$\begin{aligned} \sigma^2_{\text{fix}} &= \lambda \text{Var}_f(Y - cW) + (1-\lambda)\text{Var}_{g_0}(Y - cW) - (\mu - c)^2 \\ &= \sigma^2_{\text{random}} - \frac{\lambda}{1-\lambda}(\mu - c - E_f(Y - cW))^2. \end{aligned} \tag{6.18}$$

The same β is used for both components. The second term is nonnegative, so variance for the fixed allocations is no worse than and may be better than the variance for random allocations.

In addition, fixing the allocations reduces the variance of the sample average of W, and so reduces the probability or magnitude of some finite-sample problems such as negative regression weights and integration estimates outside the range of observed values.

Fixing the component allocations can be interpreted in terms of stratified sampling. Consider the augmented variable $X^* = (X, S)$, where X is the original input variable and $S = 0, 1$ according to whether X is generated from f or g. Generating X is equivalent to generating X^*, by first generating S, $P(S = 0) = \lambda = 1 - P(S = 1)$, then generating X_1^* from its marginal distribution given S; that is $X_1^* \sim f$ iff $S = 0$. The X_1^* has the same distribution as X. Now when X^* is to be generated in a simulation experiment the values of S can be stratified, resulting in fixed sample allocations. Thus it is no surprise that the stratified (fixed) allocation scheme performs better than the unstratified scheme, since this is a well-known result for stratified sampling.

All mixture sampling in this work is done with fixed allocations (though standard error estimates may not reflect this).

6.2 General Mixture Distributions

In the previous section we discussed mixtures of two distributions, where one of the distributions is the true distribution f. The mixture sampling method can be generalized to more component distributions. The generalized mixture distribution is:

$$g_\lambda(x) = \sum_{k=1}^{K} \lambda_k g_k(x) \tag{6.19}$$

Table 6.5: General Mixture Distribution in Gaussian Probability Example
Mean Square Error of Importance Sampling Estimates

	Integration	Ratio	Regression	SRS
$\mu_1 = P(X > z_\alpha)$	0.000025	0.000029	0.000024	0.00025
$\mu_2 = P(X \leq z_\alpha)$	0.0065	0.000029	0.000024	0.00025
$\mu_3 = 1$	0.0062	0	0	0
$\mu_4 = E(X)$	0.017	0.017	0.017	0.025
$\mu_5 = P(X \leq -z_\alpha)$	0.000024	0.000028	0.000023	0.00025

General mixture distribution, 50% $N(0,1)$, 25% $N(z_\alpha, 1)$, 25% $N(-z_\alpha, 1)$. $z_\alpha = 2.326$, sample size is one. There are 40 replications in each Monte Carlo experiment and 2000 Monte Carlo experiments.

where $\Lambda = (\lambda_1, \lambda_2, \ldots \lambda_k)$, $\lambda_k \geq 0$ and $\sum \lambda_k = 1$. Often one of the distributions g_k is f.

Mixture distributions can be used to create sampling distributions with nearly arbitrary forms. Van Dijk and Kloek (1985) decompose a sampling space into a large number of slabs and generate distributions within each slab which are products of uniform distributions in the short direction and mixtures of multivariate normal distributions in other directions, in attempt to build sampling distributions which closely match the posterior distribution in an example in Bayesian analysis. Slab choices, mixing parameters, and parameters of the normal distributions are estimated from a preliminary run. Mixture distributions have also been used in a structural analysis setting (personal communication, Yaacob Ibrahim).

Table 6.5 contains the results of a simulation using a general mixture distribution. Here we consider the same example as in Table 2.2, except that we also wish to estimate the probability that X is less than $-z_\alpha$. The sampling distribution is a mixture of 50% f and 25% each $N(-z_\alpha, 1)$ and $N(z_\alpha, 1)$. The results are good. Even the estimate of the expectation of X is improved, though the sampling distribution was not designed for this.

6.2.1 Multivariate Applications

Mixture distributions are particularly useful with multivariate output. If different dimensions of the output take on extreme values in different re-

gions of the sampling space, and different potential sampling distributions emphasize those regions, an appropriate mixture distribution can estimate all components of the output well in a single simulation experiment.

An alternative to a mixture distribution is to run separate simulation experiments. Using a mixture distribution has a number of advantages. First, the output is parsimonious. For example, in an experiment to estimate the performance of a service network, the average waiting time is equal to the total waiting time divided by the number of waits, which is not true if the three values are estimated in different experiments.

A second advantage is efficiency. Only one experiment need be run instead of many. Of course the results from multiple experiments could be combined by using weighted averages of the results from the different experiments, but both ways to do this are distasteful. Using a single set of weights for combining results gives bad estimates for all output quantities (each output quantity will be estimated well in one experiment but poorly in others, and bad estimates dominate). Using a different set of weights for each output, with weights chosen to minimize the estimated variance of that output, gives better estimates, but the output is not parsimonious.

Finally, for the same number of total replications, a mixture distribution produces estimates with lower (asymptotic) variances than any weighted average of results obtained from each component separately. In heuristic terms this is because deficiencies of each sampling distribution are compensated for by other components of the mixture distribution.

We use the following notation: experiment 1 consists of n_1 observations from distribution g_1, experiment 2 consists of n_2 observations from g_2, and the combined experiment consists of n observations from g, where $n = n_1 + n_2$ and $g(x) = (n_1 g_1(x) + n_2 g_2(x))/n$. For a fixed estimation method let $\hat{\mu}_1$ be an estimate derived solely from the first experiment, with variance σ_1^2/n_1 (disregarding low order terms); similarly $\hat{\mu}_2$ has variance σ_2^2/n_2. The estimate based on the combined experiment is $\hat{\mu}_c$, with variance σ_c^2/n. We write $W_1 := f/g_1$, $W_2 := f/g_2$, $W := f/g$, $Y_1 := \theta W_1$, $Y_2 := \theta W_2$, and $Y := \theta W$. All results are stated in terms of a two-component mixture distribution, but can be iteratively extended to multiple-component applications.

The variances we give for a combined experiment are not based on stratified sample allocations; such variances are lower (6.18), but this will not affect the results here because the combined experiment here is already better than separate experiments, if the conditions hold for which the stratified variance is better than the unstratified experiment.

We begin with a simple theorem which serves to illustrate some ideas. For the ratio estimate, results obtained from any distribution can be improved (or at least not made worse) by adding observations from another distribution.

Theorem 6.2 *The ratio estimate benefits from adding observations*

When using the ratio estimate, $\sigma_c^2 \leq \sigma_1^2$. If $\int(\theta-\mu)^2 g_2 > 0$ the inequality is strict.

Proof The variance terms for the single and combined experiments are:

$$\sigma_1^2 = \frac{1}{n_1}\int g_1 W_1^2(\theta-\mu)^2 = \int \frac{f^2(\theta-\mu)^2}{n_1 g_1} \tag{6.20}$$

$$\sigma_c^2 \leq \int \frac{f^2(\theta-\mu)^2}{ng} = \int \frac{f^2(\theta-\mu)^2}{n_1 g_1 + n_2 g_2} \tag{6.21}$$

Now $f^2(\theta-\mu)^2 \geq 0$, $n_1 g_1 \geq 0$, and $n_2 g_2 \geq 0$, so the integrand is never larger for σ_c^2 than it is for σ_1^2, and hence the integral is not larger. The additional condition implies that the integral is smaller. (6.21) would be an equality except for the stratification effect discussed in Section 6.1.3, but the theorem holds even without that effect.

In this instance adding n_2 additional observations from g_2 increases the denominator of the variance integrand, giving a lower variance. The new observations represent additional information which improve the quality of the estimates.

Theorem 6.3 gives a stronger result, that the combined estimate is better than any linear average of results from separate experiments. Define $\hat{\mu}_a$ to be such a linear combination

$$\hat{\mu}_a = a\hat{\mu}_1 + (1-a)\hat{\mu}_2 \tag{6.22}$$

which has variance $a^2\sigma_1^2/n_1 + (1-a)^2\sigma_2^2/n_2$.

Theorem 6.3 *The ratio estimate from mixture sampling is better than combining estimates from individual components*

When using the ratio estimate, $\sigma_c^2 \leq \sigma_a^2$, for $0 \leq a \leq 1$, with equality iff $g_2/g_1 = \zeta$ for all x for which $f(x)(\theta(x)-\mu)$ is nonzero and for some constant ζ, and $a = n_1/(n_1+\zeta n_2)$.

The proof is in the appendix.

6.2.2 Integration and Regression Estimates

Analogous results to Theorems 6.2 and 6.3 do not hold for the integration and regression estimates. Additional information can hurt, rather than help, the accuracy of those estimates.

The problem is that both the integration and regression estimates involve an induced transformation $\theta \rightarrow Y = \theta f/g$ and use of $W = f/g$ as a linear control variable. The regression estimate performs very well if the relationship between Y and W is approximately linear, and the integration estimate performs well if the relationship is approximately linear with slope 0. Now any single component g_k of a mixture distribution may be such that $Y = \theta f/g_k$ and $W = f/g_k$ have approximately a linear relationship, but the same is not necessarily true for the mixture distribution g.

We should be able to improve estimates obtained from a single distribution by adding information provided by observations from other distributions. In order to be sure of improvement, however, we need to use this information in a way that is consistent with the transformation induced by the single distribution. We can compute Y and W using g_k rather than g, where g_k is the single distribution and g is the mixture distribution which includes g_k. Then Y and W have the same structure as under g_k, but the results are biased and inconsistent—$E_g(Y) \neq \mu$, and $E_g(W) \neq 1$.

To correct for this new bias we could throw away observations, keeping any single observation with probability proportional to the likelihood of that observation under the component distribution, divided by the likelihood under the mixture distribution. This is a form of acceptance-rejection generation of random variables, and results in the remaining observations having the component distribution. By doing this we can recreate a simulation involving only one of multiple components.

Unfortunately throwing away observations randomly results in the remaining observations being a random subset of the original observations. Thus if we were to use this procedure more than once using the same set of observations, we would obtain different results. We could repeat this procedure many times and take an average, but that would be computationally expensive.

A better solution is to use modified versions of the integration and regression estimates. Use all observations from all components, but define Y and W using the component distribution g_k, and compute the moment estimates required for an estimate using weights which reflect the relative likelihood

of each observation under g_k and under g. These weights are *metaweights*, weights on the observations (θ_i, W_i), and the metaweight for any observation is proportional to the probability that it would be retained under the acceptance-rejection procedure.

We use the same notation as in the previous section, and continue to give variances for estimates using different distributions under the assumption that the distribution allocations are not stratified. The stratification effect (6.18) further reduces the variance of estimates based on multiple distributions, but that effect is not needed for the results we give here to hold.

The metaweights for replication i, based on component distribution k of the mixture distribution, are:

$$\pi_i^{(k)} = \pi^{(k)}(X_i) := \frac{g_k(X_i)}{g(X_i)}, \tag{6.23}$$

and normalized metaweights are:

$$p_i^{(k)} := \pi_i^{(k)} / \sum_{l=1}^{n} \pi_l^{(k)} \tag{6.24}$$

The Y and W values are computed as if g_k were the sampling distribution:

$$W_i^{(k)} := f(X_i)/g_k(X_i) \tag{6.25}$$

$$Y_i^{(k)} := \theta(X_i) W_i^{(k)}. \tag{6.26}$$

Henceforth if Y and W appear without superscripts they refer to Y and W computed using the mixture distribution, e.g. $W = f/g$. The weighted moments of $Y^{(k)}$ and $W^{(k)}$ are:

$$\bar{Y}^{(k:m)} := \sum_{i=1}^{n} p_i^{(k)} Y_i^{(k)} \tag{6.27}$$

$$\bar{W}^{(k:m)} := \sum_{i=1}^{n} p_i^{(k)} W_i^{(k)} \tag{6.28}$$

$$\hat{\beta}^{(k:m)} := \frac{\sum_{i=1}^{n} p_i^{(k)} (Y_i^{(k)} - \bar{Y}^{(k)})(W_i^{(k)} - \bar{W}^{(k)})}{\sum_{i=1}^{n} p_i^{(k)} (W_i^{(k)} - \bar{W}^{(k)})^2} \tag{6.29}$$

Formulas (6.30-6.32) are computed using the mixture distribution with metaweights for distribution k. We use the notation $(k : m)$ to distinguish the formulas

from the similar formulas which are based solely on results from component k.

$$\bar{Y}^{(k)} := \frac{1}{n_k}\sum_{i=1}^{n_k} Y_i^{(k)} \tag{6.30}$$

$$\bar{W}^{(k)} := \frac{1}{n_k}\sum_{i=1}^{n_k} W_i^{(k)} \tag{6.31}$$

$$\hat{\beta}^{(k)} := \frac{\sum_{i=1}^{n_k}(Y_i^{(k)} - \bar{Y}^{(k)})(W_i^{(k)} - \bar{W}^{(k)})}{\sum_{i=1}^{n_k}(W_i^{(k)} - \bar{W}^{(k)})^2} \tag{6.32}$$

This $(k:m)$ notation is not needed for the $\pi_i^{(k)}$ and $p_i^{(k)}$ values because there are no corresponding values based solely on distribution k.

The estimates using only distribution k are:

$$\hat{\mu}_{\text{int}}^{(k)} := \bar{Y}^{(k)} \tag{6.33}$$

$$\hat{\mu}_{\text{ratio}}^{(k)} := \bar{Y}^{(k)}/\bar{W}^{(k)} \tag{6.34}$$

$$\hat{\mu}_{\text{reg}}^{(k)} := \bar{Y}^{(k)} - \hat{\beta}^{(k)}(\bar{W}^{(k)} - 1) \tag{6.35}$$

The modified integration, ratio, and regression estimates are defined in terms of $\bar{Y}^{(k:m)}$, $\bar{W}^{(k:m)}$, and $\hat{\beta}^{(k:m)}$. These estimates are:

$$\hat{\mu}_{\text{int}}^{(k:m)} := \bar{Y}^{(k:m)} \tag{6.36}$$

$$\hat{\mu}_{\text{ratio}}^{(k:m)} := \bar{Y}^{(k:m)}/\bar{W}^{(k:m)} \tag{6.37}$$

$$\hat{\mu}_{\text{reg}}^{(k:m)} := \bar{Y}^{(k:m)} - \hat{\beta}^{(k:m)}(\bar{W}^{(k:m)} - 1) \tag{6.38}$$

It is instructive to compare these integration and ratio formulas to the corresponding formulas based on the whole mixture distribution. It turns out that:

$$\hat{\mu}_{\text{int}}^{(k:m)} = \bar{Y}/\bar{\pi}^{(k)} \tag{6.39}$$

$$\hat{\mu}_{\text{ratio}}^{(k:m)} = \bar{Y}/\bar{W} = \hat{\mu}_{\text{ratio}} \tag{6.40}$$

where $\bar{\pi}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \pi_i^{(k)}$

The new integration estimate is the old one, divided by the average of the metaweights (the expected value of each metaweight is one). The new estimate is based on using the average metaweight as an inverse control function. Even more surprising is the result for the ratio estimate – the new

ratio estimate is exactly the same as the old one. This estimate does not make use of any information about the structure of the components of the mixture distribution.

Now analogous results (to Theorems 6.2 and 6.3) hold for the modified integration and regression estimates. We begin with propositions giving the variance of the two estimates. The proofs are in the appendix.

Proposition 6.1 *Variance of the modified integration estimate*

The variance of the modified integration estimate (6.36) is:

$$\mathrm{Var}(\hat{\mu}_{\mathrm{int}}^{(k:m)}) = n^{-1} \int (\pi^{(k)})^2 (Y^{(k)} - \mu)^2 g(x) dx \tag{6.41}$$

where $Y^{(k)} = Y^{(k)}(x)$ and $\pi^{(k)} = \pi^{(k)}(x)$.

Note that this estimate has lower variance than the estimate based solely on observations from component k, since $(\pi^{(k)})^2(x)g(x)/n = g_k(x)\pi^{(k)}(x)/n < g_k(x)/n_k$.

Proposition 6.2 *Variance of the modified regression estimate*

The variance of the modified regression estimate (6.38) is:

$$\mathrm{Var}(\hat{\mu}_{\mathrm{reg}}^{(k:m)}) = n^{-1} \int \pi^{(k)2} (Y^{(k)} - \mu - \beta^{(k)}(W^{(k)} - 1))^2 g(x) dx. \tag{6.42}$$

Note that this estimate has lower variance than the estimate based solely on observations from component k, since $\pi^{(k)2}(x)g(x)/n = g_k(x)\pi^{(k)}(x)/n < g_k(x)/n_k$.

We now show that given any weighted average of results from single components, we can define versions of the integration and regression formulas that outperform that weighted average. The proofs of the propositions and theorems are in the appendix.

Proposition 6.3 *Variance of the combined integration estimate*

Define the combined integration estimate, with parameter a, as:

$$\hat{\mu}_{\mathrm{int}}^{(a:m)} = a\hat{\mu}_{\mathrm{int}}^{(1:m)} + (1-a)\hat{\mu}_{\mathrm{int}}^{(2:m)} \tag{6.43}$$

The asymptotic variance of $\hat{\mu}_{\mathrm{int}}^{(a:m)}$ is

$$\mathrm{Var}(\hat{\mu}_{\mathrm{int}}^{(a:m)}) = n^{-1} \int g(Y - \mu(a\pi^{(1)} + (1-a)\pi^{(2)}))^2 dx \tag{6.44}$$

Proposition 6.4 *Variance of the combined regression estimate*

Define the combined integration estimate, with parameter a, as:

$$\hat{\mu}_{\text{reg}}^{(a:m)} = a\hat{\mu}_{\text{reg}}^{(1:m)} + (1-a)\hat{\mu}_{\text{reg}}^{(2:m)} \tag{6.45}$$

The asymptotic variance of $\hat{\mu}_{\text{reg}}^{(a:m)}$ is

$$\text{Var}(\hat{\mu}_{\text{reg}}^{(a:m)}) = n^{-1}\int g(ah^{(1)} + (1-a)h^{(2)})^2 dx \tag{6.46}$$

where

$$h^{(k)}(x) = Y(x) - W(x)\beta^{(k)} - \pi^{(k)}(x)(\mu - \beta^{(1)}) \tag{6.47}$$

Theorem 6.4 *Superiority of the combined integration estimate*

Under the conditions of Proposition 6.3 the combined estimate has smaller asymptotic variance than the corresponding weighted average of results from each replication—

$$\text{Var}(\hat{\mu}_{\text{int}}^{(a:m)}) \leq \text{Var}(a\hat{\mu}_{\text{int}}^{(1)} + (1-a)\hat{\mu}_{\text{int}}^{(2)}). \tag{6.48}$$

Equality holds if $a = n_2/(n_1+n_2)$ and $g_1(x) = g_2(x)$ (except possibly on a set of measure 0).

Theorem 6.5 *Superiority of the combined regression estimate*

Under the conditions of Proposition 6.4 the combined regression estimate has smaller asymptotic variance than the corresponding weighted average of results from each replication—

$$\text{Var}(\hat{\mu}_{\text{reg}}^{(a:m)}) \leq \text{Var}(a\hat{\mu}_{\text{reg}}^{(1)} + (1-a)\hat{\mu}_{\text{reg}}^{(2)}). \tag{6.49}$$

Equality holds if $a = n_2/(n_1+n_2)$ and $g_1(x) = g_2(x)$ (except possibly on a set of measure 0).

6.3 Exponential Tilting

One of the most successful sampling techniques in importance sampling is exponential tilting, also known as exponential biasing, the use of a sampling distribution of the form:

$$g(x) = \alpha e^{\beta S(x)} f(x), \tag{6.50}$$

where S(x) is a linear combination of real-valued functions of the components of multivariate x,

$$S(x) = \sum_{j=1}^{d} s_j(x_j). \tag{6.51}$$

β determines the degree to which the sampling distribution is "tilted" away from the true distribution, and $\alpha = \alpha(\beta)$ normalizes the distribution to unit mass. The existence of such a distribution requires that $S(X)$ have a finite moment generating function (under f) at β.

For good performance S should be chosen so that some monotone function of S is a good approximation to θ,

$$\theta(X) \approx T_S(S(X)). \tag{6.52}$$

For example, to estimate a tail probability of the distribution of $\xi(X)$, let $\theta(X) = I(\xi(X) > k)$. If there exists a good linear combination of the input variables so that $\xi(X) \approx s_0 + S(X)$ for some constant s_0, then $T_S(S) = I(S > k - s_0) \approx \theta(X)$.

Exponential tilting has a number of favorable properties. It provides a convenient means of biasing the sampling distribution toward large or small values of S (and T_S and θ). If the dimension of X is fixed and the components of X are independently distributed under f, then the marginal distributions of the components of X under g are also independently distributed as

$$g_j(x_j) = \alpha_j e^{\beta s j(x)} f_j(x_j), \tag{6.53}$$

where α_j normalizes the distribution, and $\prod \alpha_j = \alpha$. This marginal independence makes it convenient to generate deviates from g.

Exponential tilting is the optimal means of biasing the sampling distribution for S, in the sense that g as in (6.53) minimizes the Kullback-Leibler

distance from g to f, subject to a constraint on the expected value of S:

$$\begin{aligned} &\min_g \int g(x) \log(\frac{g(x)}{f(x)}) \\ &\text{s.t. } \int g(x) = 1 \text{ and } \int S(x)g(x) = M(\beta) \end{aligned} \tag{6.54}$$

More importantly, exponential tilting produces sampling distributions with memoryless weights. It is the *only* sampling method that does so, when X has fixed dimension and independent components under both f and g and certain regularity conditions are met.

Theorem 6.6 *Uniqueness of Exponential Tilting for Memoryless Weights*

Let $f(x)$ and $g(x)$ have independent marginal distributions, $f(x) = \prod_{j=1}^{d} f_j(x_j)$, $g(x) = \prod_{j=1}^{d} g_j(x_j)$, $d > 1$, $g(x) > 0$ when $f(x) > 0$, $g(x) = f(x)r(S(x))$, where $S(x) = \sum_{j=1}^{d} s_j(x_j)$, s_j is a real-valued function of x_j and r is a real-valued function. If the support of each s_j is a set of consecutive lattice points with the same lattice size or if each s_j has a nonzero density on a single (possibly infinite) interval, then $g(x) = \alpha e^{\beta S} f(x)$ for some α and β.

The proof is in the appendix.

The three examples in Section 5.4 (memoryless weights) are all examples of exponential tilting, though in a slightly more general form than described above. Exponential tilting includes the case where both the true distribution and sampling distribution are members of a parametric exponential family of distributions, as in the digital communications example.

6.3.1 Exponential Tilting and Mixture Distributions

Exponential tilting can be used as components in mixture distributions, and the result still has memoryless weights. Let

$$g(x; \beta_k) = \alpha_k e^{\beta_k S(x)} f(x) \tag{6.55}$$

be a distribution generated using exponential tilting with the same S used for all $k = 1, 2, \ldots K$ ($K \leq \infty$), let $\{\lambda_k\}$ be weights with $\sum_k \lambda_k = 1$, and let

$$g(x) := \sum_k \lambda_k g(x; \beta_k) \tag{6.56}$$

be a mixture distribution. Then g has memoryless weights

$$W(x) = (\sum_k \lambda_k \alpha_k e^{\beta k S(x)})^{-1} \tag{6.57}$$

Note that this does not contradict Theorem 6.6, since when mixture distributions are used the marginal distributions are not independent under g.

The two methods can be used together to build sampling distributions of the desired form. Exponential tilting generates distributions for which the likelihood ratio between g and f is a function of only the linear function S, but the form of that function must be exponential. Mixtures increase the variety of functional forms that can be generated. In particular, if exponential tilting is used alone the likelihood ratio is exponentially small for small S (if $\beta > 0$) and the weights are exponentially large. Mixing that distribution with a small proportion of the untilted distribution bounds the weights.

The combination of exponential tilting and mixture sampling is illustrated in Figures 6.8 and 6.9. Figure 6.8 shows the likelihood ratio $g_\lambda(x)/f(x)$ as a function of S. The likelihood ratio is the sum of a constant function (f is a component in the mixture) and an exponential function. The biggest impact this combination has is on the weight function, where for small S the weights are much smaller than they would be if f were not included in the mixture; this can be seen in Figure 6.9.

The combination of exponential tilting and mixture distributions is not completely general. It is not possible to use this combination to generate sampling distributions for which $g(x) = f(x)r(S(x))$ for arbitrary likelihood ratio functions r. In particular, r must be convex. In most cases this is acceptable, since r convex implies that the sampling distribution is concentrated on one or both of the tails of the distribution (of S), which is usually desirable.

The development of a good general method like exponential tilting that can be convex would be useful in some applications, such as estimating the density of a distribution at its median, or for greater flexibility in using mixture distributions to build sampling distributions. This is an open question.

6.3.2 General Exponential Tilting

Exponential tilting may be extended in two ways, by

- letting S be a vector-valued function, and

Figure 6.8: Relative Likelihood Function for Exponential Mixture

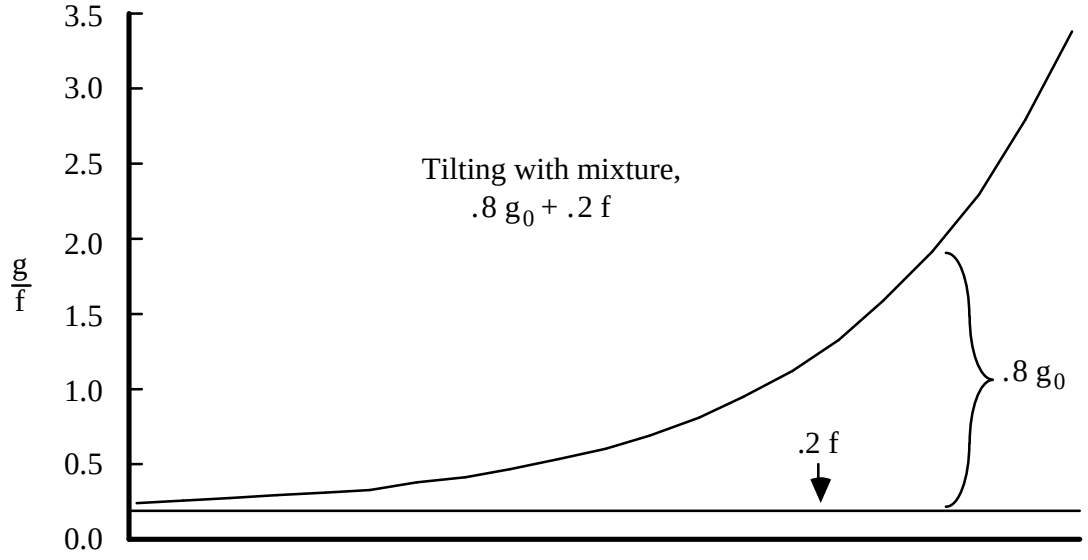
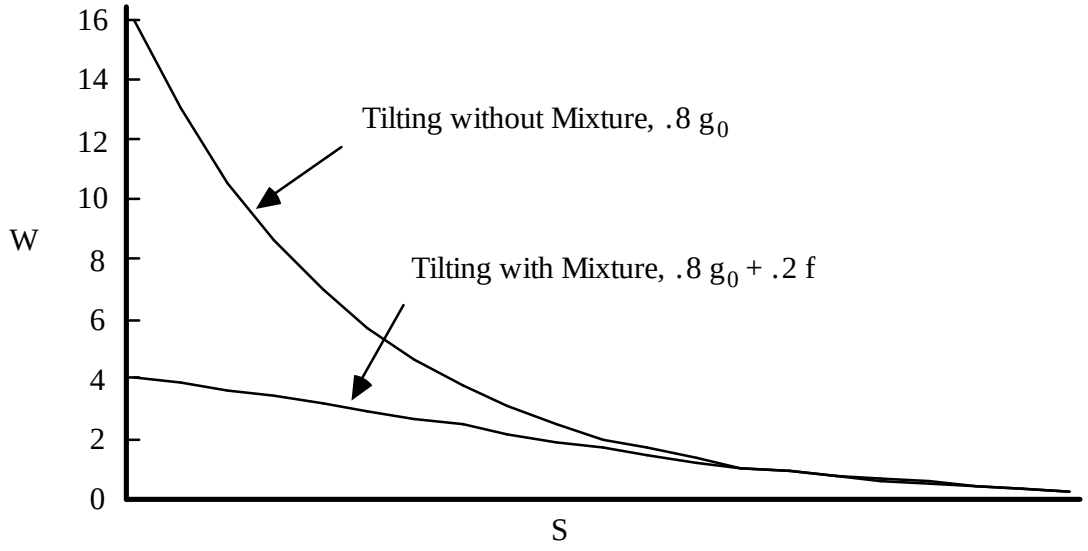


Figure 6.9: Weight Function for Exponential Mixture



- tilting from a base distribution other than f.

The combination of these extensions yields

$$g(x) = \alpha e^{\beta \cdot S(x)} g_0(x), \tag{6.58}$$

where $S(x)$ is a linear combination of vector-valued functions of the components of x,

$$S(x) = \sum_{j=1}^{d} s_j(x_j), \tag{6.59}$$

β is a vector-valued parameter, $\beta \cdot S$ is the inner product of β and S, $g_0(x)$ is a sampling distribution corresponding to $\beta = 0$, and $\alpha = \alpha(\beta)$ is real-valued and normalizes the distribution to unit mass.

The particle scattering example described in Chapter 3 is an example of tilting from a base distribution different than the true distribution. The sampling distribution for a single step is of the form:

$$g(\Delta x, \Delta y, \Delta z, d) = \alpha e^{\beta \Delta x} g_0(\Delta x, \Delta y, \Delta z, d), \tag{6.60}$$

where

$$g_0(\Delta x, \Delta y, \Delta z, d) = f(\Delta x, \Delta y, \Delta z, d) I(d = 0)/(1-p)^{I(x<0)}. \tag{6.61}$$

This is exponential tilting, but with a base distribution g_0 concentrated solely on sample paths which do not decay. The product distribution, up to exit time T, is

$$g(x, y, z) = \alpha^T e^{\beta(X_T - x_0)} (1-p)^{T-1} f(x, y, z). \tag{6.62}$$

There are two solutions for $(1-p)\alpha(\beta) = 1$, and the one we want is biased toward exit. With this choice the product of the exponentially tilted marginal distributions is a joint exponentially tilted distribution.

6.3.3 Exponential Tilting in Fixed-Dimension Applications

The easy case for exponential tilting is when the dimension of the application is fixed, and the components are independently distributed. If the marginal distributions are generated by tilting using the same tilting parameter for all marginals, as in (6.53), the joint distribution is an exponentially tilted distribution as in (6.49).

Still, there remain questions about how to use importance sampling in these applications. We must first decide what linear combination (6.51) of the input variables to base the importance sampling on (what is $s_j(X_j)$?). Second, we must decide what tilting parameter to use.

In some applications $\theta(X)$ is a monotone function of another statistic $\xi(X)$ for which a good linear approximation of the input variables exists,

$$\xi(X) \approx s_0 + \sum_{j=1}^{d} s_j(x_j). \tag{6.63}$$

This includes, e.g., estimating tail probabilities of ξ, where $\theta = I(\xi > \kappa)$. The linear combination (6.63) can form the basis for g in (6.50) and (6.51).

Johns (1987) uses a sampling distribution of this form in the evaluation of bootstrap confidence intervals. X is a vector of length d with independent components under f, with equal probabilities $1/d$ of taking any of the values in the original set of data Z, which is also a vector of length d, $P_f(X_j = Z_k) = 1/d$, and $\xi(X)$ is a robust estimate of the center of the distribution that generated the Z values. Johns uses the influence function (4.19, with T in place of ξ) to obtain an approximation of the form

$$\xi(X) \approx \xi(Z) + \frac{1}{d}\sum_{j=1}^{d} U_{k(j)}, \tag{6.64}$$

where $U_{k(j)}$ is the influence function value for Z_k when $X_j = Z_k$. The influence function approach can be used in many other applications, especially where X has *i.i.d.* components, to obtain linear approximations. In other applications physical considerations point the way to linear approximations.

For guidance in choosing the tilting parameter we turn first to Johns, who approximates the linear statistic S (6.51) by a normal distribution with mean $\xi(Z)$ and variance $\mathrm{Var}(U_j)/d$. For a normal distribution exponential tilting corresponds to changing the mean of the variable. To estimate the probability that a normal variable is greater than κ, the optimal mean for a sampling distribution (for the integration estimate) is $\kappa + 1/4\kappa + o(1/\kappa)$ as $\kappa \to \infty$. Keep the first term only. This corresponds to tilting the normal distribution so the mean of the tilted distribution is at the closest value in the "critical region."

We offer a minimax derivation of this rule (tilt so the mean matches the closest critical value) that does not depend on the normal approximation.

This derivation is based on minimizing the largest value that the weight function $W(x)$ takes on over a critical region. For estimating a probability $P(X \in A)$ the variance of the integration estimate is

$$\int_A \frac{f(x)}{g(x)} f(x)dx - P(A)^2 \tag{6.65}$$

where A is the critical region. The largest guaranteed variance reduction is obtained by minimizing f/g over the critical region.

Suppose g is obtained by exponentially tilting f using the statistic s,

$$g(x) = f(x)\frac{e^{\beta s(x)}}{\Psi(\beta)}, \tag{6.66}$$

where Ψ is the moment generation function of $s(X)$, and where s is parameterized so that $E_f(s(X)) = 0$. The minimax criterion is to choose β to maximize the minimum value of g/f over the critical region,

$$\beta^* := \arg\max_\beta \left\{ \inf_A \frac{g(x)}{f(x)} \right\} \tag{6.67}$$

Now s is real valued, and define the set

$$S_A := \{s = s(x) | x \in A\}. \tag{6.68}$$

If 0 is in the interior of the range of S_A then $\beta^* = 0$, otherwise $g(x)/f(x) < 1$ for some x. So assume 0 is not in the interior, and let κ be the point in the closure of S_A that is closest to zero. Suppose $\kappa > 0$. This implies that $\beta^* > 0$, and that the infimum of g/x over A occurs at κ, since for $\beta > 0$ $g(x)/f(x)$ is an increasing function of $s(x)$. If $\kappa < 0$ then $\beta^* < 0$, and the infimum still occurs at κ. Now solving (6.67) is equivalent to solving

$$\beta^* = \arg\max_\beta \left\{ \frac{e^{\beta\kappa}}{\Psi(\beta)} \right\}. \tag{6.69}$$

The fraction is concave in β, and differentiating with respect to β and setting equal to zero to find the maximum gives

$$\kappa = \frac{\Psi'(\beta^*)}{\Psi(\beta^*)}. \tag{6.70}$$

The fraction in (6.70) is the mean of the exponentially tilted distribution with parameter β^*, since

$$\frac{\Psi'(\beta)}{\Psi(\beta)} = \frac{\int s(x)e^{\beta s(x)}f(x)dx}{\Psi(\beta)} = E_g(s(X)) \tag{6.71}$$

This means that the minimax criterion is equivalent to tilting so the mean of the tilted distribution is at the closest point in the critical region. This rule is also used in an analytical setting by Daniels (1954) in the saddlepoint method for approximating tail probabilities.

If the ratio estimate is used the same rule for exponential tilting can be used if combined with mixture sampling. Consider the mixture distribution $g_\lambda = \lambda f + (1-\lambda)g$, where g is given by (6.66). The asymptotic variance of the ratio estimate for estimating $P(X \in A)$ is

$$\int_A \frac{f(x)}{g_\lambda(x)}(1-p)^2 f(x)dx + \int_{\neg A} \frac{f(x)}{g_\lambda(x)}p^2 f(x)dx \tag{6.72}$$

where $p = P(X \in A)$. There is no minimax criterion in this case that gives a guaranteed variance reduction, because if g_λ/f is larger than 1 somewhere it is smaller than one somewhere else, and (6.72) requires integrating over the whole sample space, in contrast to (6.65) which required integrating only over A. Still, the form of the optimal sampling distribution given by (2.79), $g^*(x) = C|\theta(x) - \mu|f(x)$, suggests that good results may be obtained by choosing β and λ by the minimax criterion

$$(\lambda^*, \beta^*) := \arg\max_{\lambda,\beta} \left\{ \inf_A \frac{g(x)}{f(x)(1-p)} \ \wedge \ \inf_{\neg A} \frac{g(x)}{f(x)p} \right\} \tag{6.73}$$

where "$\wedge$" indicates a minimum. If $0 \notin S_A$ this is solved by the same β^* as for the integration estimate,

$$\beta^* \text{ solves } E_g(s(X)) = \kappa, \tag{6.74}$$

and a mixing parameter λ^* given by

$$\lambda^* = \left(1 + \frac{1-2p}{p}\Psi(\beta^*)e^{-\kappa\beta^*}\right)^{-1}. \tag{6.75}$$

The solutions of λ^* and β^* in the case of a normal distribution, with $A = \{x|x > \kappa\}$ are given in Table 6.6, together with the resulting efficiency of the ratio estimate.

This minimax approach can be applied in other applications as well, though the mathematics may be more difficult.

Table 6.6: Minimax β^* and λ^* for a Gaussian Tail Probability, and Efficiency of the Ratio Estimate

κ	$P(X > \kappa)$	β^*	λ^*	Efficiency
0.0	0.500	0.0	1.000	1.000
0.674	0.250	0.674	0.386	0.828
1.282	0.100	1.282	0.221	0.491
1.645	0.050	1.645	0.177	0.306
2.327	0.010	2.327	0.133	0.088
2.576	0.005	2.576	0.122	0.050
3.091	0.001	3.091	0.106	0.012

6.3.4 Dependence and Random Dimension Applications

The beauty of exponential tilting in fixed-dimensional independent applications is that each component can be generated independently. When the dimension of X is not fixed, or when the components of X do not have independent distributions, then generation of each component according to its marginal (exponentially tilted) distribution may not lead to a distribution of the desired form (6.50), or do so only for fixed values of β.

Suppose, for example, that X has i.i.d. components but that the dimension D of X is random. An example of this is given by Siegmund (1976) in the study of sequential tests, discussed in Section 5.4. If the components of X are generated with the distribution

$$g_1(x_j) = \alpha e^{\beta s(x_j)} f_1(x_j), \tag{6.76}$$

for $j = 1, 2, \ldots D$, where f_1 and g_1 are the marginal true and sampling distributions for a single step. The product distribution is

$$g(x) = \alpha^D e^{\beta S(X)} f(x), \tag{6.77}$$

which depends on both D and S. Only if $\alpha = 1$ does the likelihood ratio depend solely on S. Now $\alpha = \Psi^{-1}(\beta)$, where Ψ is the moment generating function for $s(X_1)$ under f_1 , which is a strictly convex function on its domain; thus there are at most two values of β for which $\alpha = 1$, and one of these is the trivial solution $\beta = 0$ which corresponds to no importance sampling. In

this example there is only one non-trivial β for which generating marginal distributions independently gives a joint exponentially-tilted distribution.

In general applications (where X has a random number of components and/or the components are not independent) a distribution of the form (6.50) can be generated by tilting each component in turn only if the product of the normalizing constants is a single constant with probability 1.

That is, suppose $X = (X_1, X_2, \ldots X_D)$ has probability measure $f(x)$, D is a stopping time with $P_g(D < \infty) = 1$, write $x_{j-} = (x_1, \ldots x_{j-1})$, let $f_j(x_j|x_{j-})$ be the conditional distribution of X_j given x_{j-}, let $S(X) = \sum_{j=1}^{D} s_j(x_j|x_{j-})$, and let

$$g_j(x_j|x_{j-}) = \alpha_j(\beta|x_{j-})e^{\beta s_j(x_j|x_{j-})}f_j(x_j|x_{j-}), \tag{6.78}$$

where $\alpha_j(\beta|x_{j-})$ is a normalizing constant. Then

$$g(x) = \prod_{j=1}^{D} \alpha_j(\beta|x_{j-})e^{\beta S(x)}f(x) = \alpha e^{\beta S(x)}f(x) \tag{6.79}$$

for some α iff

$$\prod_{j=1}^{D} \alpha_j(\beta|x_{j-}) \equiv \alpha \text{ a.s.} \tag{6.80}$$

6.4 Internal Sampling Distributions

There are two broad ways to specify sampling distributions in importance sampling:

- specify the sampling distributions of the output, $g(X)$, or
- specify an alternate distribution for the "uniform" random numbers used to generate X.

We call the latter method "internal" sampling distribution generation.

Pseudo-random numbers needed in a Monte Carlo simulation are generated on a computer in a two-step process:

1. generate a stream of pseudo-random uniformly distributed numbers $U = \{U_1, U_2, \ldots\}$, and

2. transform the uniform deviates to obtain the desired distribution using the transformation $X = T_f(U)$.

See Knuth (1981), Kennedy & Gentle (1980) or Bratley, Fox & Schrage (1983) and their references for details on the implementation of both steps.

In importance sampling the second step can be done in two ways, by specifying a sampling distribution g and using an appropriate transformation

$$X = T_g(U), \tag{6.81}$$

or by modifying the input stream prior to using the original transformation,

$$X = T_f(\tau(U)), \tag{6.82}$$

where τ is a transformation of U. The latter method corresponds to internal distribution selection. It is often easier to implement, since τ can be a very simple function, no matter how complex the simulation. This is the primary advantage of internal sampling distributions.

An additional benefit of internal distribution generation is that the simulation program can be made modular—the main loop of the Monte Carlo program need not be aware that importance sampling is being used. Let

$$V := \tau(U). \tag{6.83}$$

The main loop of the program uses V as if it were a stream of uniform numbers; only the subroutine that produces V and the subroutine that computes weighted average results are affected. This is implemented in the FIPS program described in Chapter 4.

The relative likelihood and weight functions are now defined in terms of V, rather than in terms of X as in other importance sampling. If V has density $g_v(v)$, the weight function is

$$W(v) = \frac{1}{g_v(v)}. \tag{6.84}$$

The numerator is one because the uniform density is equal to 1.

In the sequel we consider applications where X has real-valued components, $X = (X_1, \ldots X_d)$ (with d possibly random), and each X_j requires only one uniform number for generation. The latter condition requires that variable generation techniques such as acceptance-rejection are not used. If the transformation is done separately on each dimension of U, then

$$V = \tau(U) = (\tau_1(U_1), \tau_2(U_2), \ldots \tau_d(U_d)). \tag{6.85}$$

In this case g dominates f if each τ_j is continuous, bijective, and piecewise differentiable on (0,1). The product density is:

$$g_v(v) = \prod_{j=1}^{d} g_{v_j}(v_j), \tag{6.86}$$

where

$$g_{v_j}(v_j) = \frac{1}{\tau_j'(\tau_j^{-1}(v_j))} \tag{6.87}$$

and the $'$ denotes a derivative.

Moy (1965) considered transformations of the forms

$$\tau_j(u) = \frac{\log(1 + u(\alpha_j - 1))}{\log(\alpha_j)} \tag{6.88}$$

or $\alpha_j > 1$, and

$$\tau_j(u) = u^{1/\alpha_j} \tag{6.89}$$

for $\alpha_j > 1$, which result in component densities

$$g_{v_j}(v) = \frac{\log(\alpha_j)}{\alpha_j - 1}\alpha_j^u \tag{6.90}$$

$$g_{v_j}(v) = \alpha_j v^{a_j - 1} \tag{6.91}$$

respectively. We call these Moy1 and Moy2, respectively.

Moy obtained better results with Moy1. We suspect there are two reasons for this. First, the weights obtained using the second method are unbounded, which is dangerous except in mathematically simple discrete zero mode applications. Second, the method is equivalent to exponentially tilting a distribution that doesn't fit his (or most) data well, with skewness the opposite of what is needed; we discuss this in Section 6.4.2.

In the next section we propose two families of internal sampling distributions. In contrast to Moy, we do not attempt to find a single kind of transformation that works well in every application. Instead we provide guidelines on how to choose sampling distributions within one of these families. The advantage of these families is that they are based on well-known distributions, so that the choice can reflect the experimenter's intuition about his or her application. We discuss that choice and parameter selection in Sections 6.4.2 and 6.4.3. Details of some of the distributions are given in Appendix 1.

6.4.1 Translation and Exponential Tilting Families

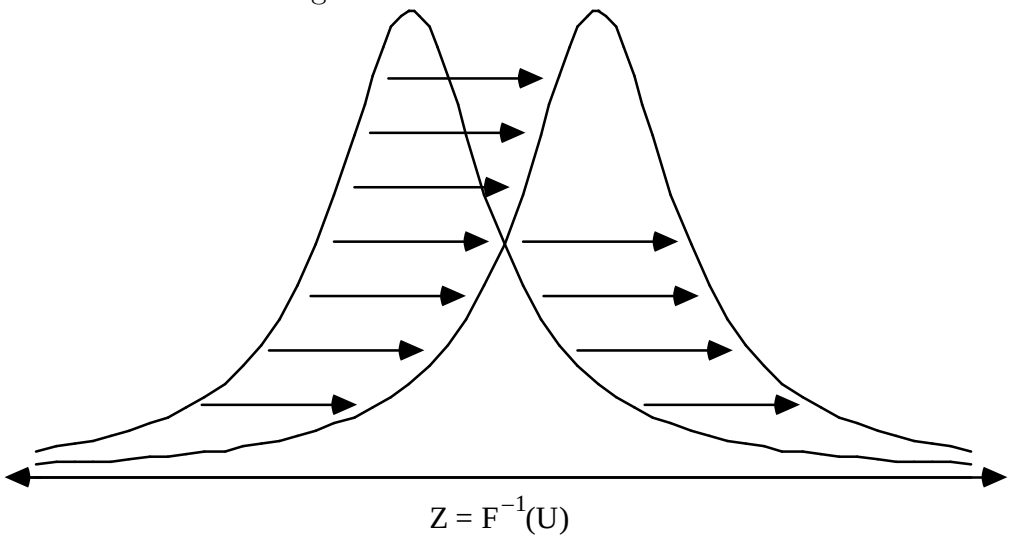
We propose two general families of internal importance sampling distributions, the "exponential tilting" and "translation" families.

The translation family (for univariate u) is defined by a distribution function F and a transformation

$$\tau(u) = F(F^{-1}(u) + \alpha) \tag{6.92}$$

for $-\infty < \alpha < \infty$, where F is a piecewise differentiable distribution function with nonzero density over the real line, $f(x) = F'(x) > 0$ a.e. for $-\infty < x < \infty$. This is equivalent to creating a random variable using the inversion technique, i.e. $Z = F^{-1}(U)$, translating that variable, and transforming back to the domain $(0, 1)$ using the original distribution function. Note that if $\alpha = 0$ no translation takes place.

Figure 6.10: Translation Method



The exponential tilting family (for univariate u) is defined by an exponential family of distributions and a transformation

$$\tau(u) = F_0(F_\alpha^{-1}(u)) \tag{6.93}$$

for $\alpha \in D(F_0)$ where F_α is the distribution function of an exponential family with parameter α,

$$dF_\alpha(z) = \frac{1}{\Psi(\alpha)} e^{\alpha z} dF_0(z), \tag{6.94}$$

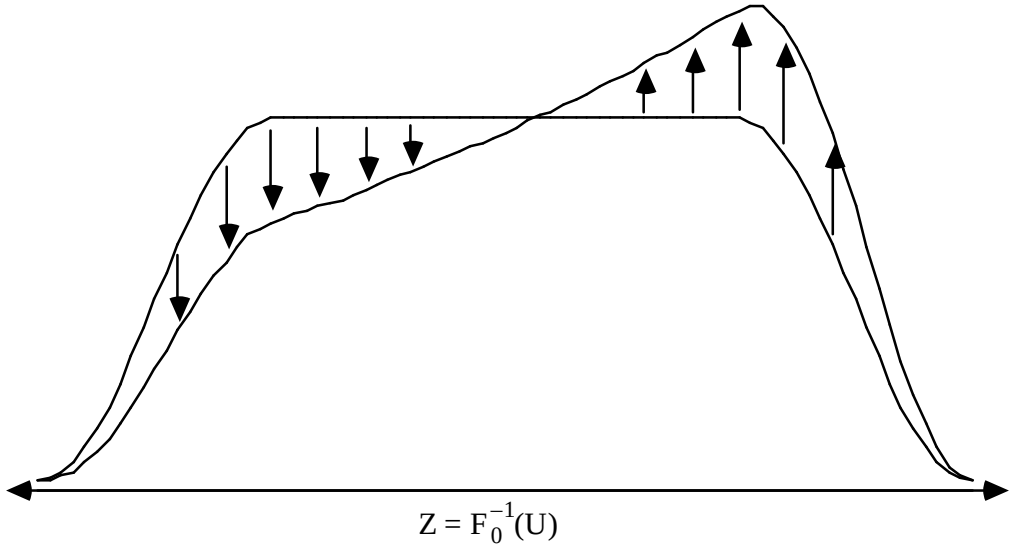
where $\Psi(\alpha) = E_{F_0}(e^{\alpha Z})$ is the moment generating function for F_0 and $D(F_0)$ is the (possibly infinite) interval for which $\Psi(\alpha) < \infty$. If $\alpha = 0$ no tilting takes place. If the base distribution has a density then (6.94) is equivalent to

$$f_\alpha(z) = \frac{1}{\Psi(\alpha)} e^{\alpha z} f_0(z); \tag{6.95}$$

if the base distribution is discrete then (6.94) reduces to

$$f_\alpha(z_j) = \frac{e^{\alpha z_j} f_0(z_j)}{\sum_k e^{\alpha z_k} f_0(z_k)} \tag{6.96}$$

Figure 6.11: Exponential Tilting Method



Standard Internal Families An advantage of using internal distributions is that the internal distributions (the "Z" values) can be chosen from well-known families of distributions. The appendix contains formulas for importance sampling using internal exponential tilting based on uniform, tilted

uniform, normal, exponential, reverse exponential (power family), gamma, reverse gamma, and general discrete distributions, and for translation for the normal, Cauchy, and logistic distributions. In each case we present the distribution and density functions for the families, the transformations, and the resulting weights.

Three of the families—the gamma, reverse gamma and tilted uniform families—allow (and require) specification of an additional parameter that determines the shape of the base distribution. The gamma family includes the exponential distribution, and the tilted uniform family includes the uniform distribution. The tilted uniform family is equivalent to a truncated exponential distribution (or truncated reverse exponential distribution).

Moy's methods are instances of exponential tilting, based on a uniform distribution (Moy1) and a reverse exponential distribution (Moy2).

6.4.2 Base Distribution Choice

To use a interior importance sampling it is necessary to choose both a distributional form and the sampling parameter α for each dimension of the input.

We return to our earlier discussion of exponential tilting for guidance in using the exponential tilting families. When using internal variables in fixed-dimensional applications with independent components the likelihood ratio is an exponential function of

$$S_1 := \sum_{j=1}^{d} \alpha_j Z_j, \tag{6.97}$$

where

$$Z_j = F_j^{-1}(V_j) \tag{6.98}$$

is the internal variable associated with input variable j. As in the earlier discussion of exponential tilting, good performance requires that a monotone function of S_1 be a good approximation to θ, $\theta(X) \approx T_S(S_1(X))$. Therefore, if there exist increasing functions $s_j(X_j)$ and a monotone transformation T_S such that $T_S(S_2) \approx \theta$ for

$$S_2 := \sum_{j=1}^{d} s_j(X_j), \tag{6.99}$$

then ideally the exponential family and parameters could be chosen so that

$$S_1 = \beta S_2 + c, \tag{6.100}$$

which implies that

$$\alpha_j Z_j = \beta s_j(X_j) + c_j \tag{6.101}$$

at least approximately, where $\beta \in \mathcal{R}$ determines the overall degree of exponential tilting and c and c_j are any constants.

This means that each base distribution should be chosen to match the shape of the corresponding s_j, and the tilting parameter chosen proportional to the ratio of standard deviations:

$$\alpha_j \propto \frac{\sigma(s_j(X_j))}{\sigma(Z_j)} \tag{6.102}$$

This indicates why Moy found that the power family transformation worked poorly. That family corresponds to a negative exponential base distribution ($f_x(x) = e^x$ for $x < 0$), which is reasonable if the influence of each variable is negatively skewed. But there, and most often, the reverse is true.

If it is necessary to err in the choice of a shape, it is better to err by being conservative and not sampling too little anywhere, especially in the tails of a distribution. This argues for the use of an internal distribution which has heavy tails on the right and light tails on the left for $\alpha > 0$ (or the converse if $\alpha < 0$); then both tails are sampled relatively often compared to a base distribution skewed the other way. If the distribution is bounded on the side of the light tail is bounded, say by using an exponential or gamma distribution, then the weights are bounded as well.

The translation method is perhaps best understood by comparison to the exponential method. First, the normal translation family is identical to the normal tilting family. Now comparing the Cauchy and logistic families to the normal shows that the families give similar transformations in the center of the distribution, but differ in the tails. The likelihood ratio for the normal family is unbounded, for the logistic family is bounded, and for the Cauchy family is not only bounded but is also redescending.

From this it appears that the logistic family has an advantage over the normal family in applications where mixture sampling will not be used, because it gives bounded weights. The Cauchy family, on the other hand, implies that the influence of an input variable is not a monotone function of the variable. Unless that is the case the Cauchy family should be avoided.

6.4.3 Parameter Choice

Once the shape of the base distributions is set the tilting parameters must be chosen. This generally breaks down into two problems—choosing the relative magnitudes of parameters for different variables, and choosing the overall degree of tilting.

We suggest basing the relative magnitudes of parameters on (6.101), so that the parameter for each input variable is proportional to ratio of standard deviations of the influence of the input variable and the corresponding internal variable.

$$\alpha_j \propto \frac{\sigma(s_j(X_j))}{\sigma(Z_j)} \tag{6.103}$$

The first standard deviation can be estimated from physical considerations or a trial study, and the second is determined by the interior distribution used. We discuss below how to modify this in the case of correlated input variables.

The overall tilting parameter can be based on physical considerations, optimized based on a trial study, or chosen to recenter the sampling distribution at a specified quantile of its distribution without importance sampling, as discussed in Section 6.3.3.

Moy (1965) uses a trial study in the study of queueing networks. All variables in a single experiment use the same transformation family, either Moy1 or Moy2. In one example he constrains the tilting parameters for all variables to be the same. That was less satisfactory in another example, where he allows different classes of variables (failures and repairs) to have different parameters. In both cases the optimal parameters are estimated from the trial study.

Moy performs the optimization by writing the variance of the integration estimate for sampling parameter α as

$$V(\alpha) = \int \dots \int \frac{\theta^2(u) f^2(u)}{g(u;\alpha)} du. \tag{6.104}$$

Differentiating with respect to α indicates that the minimum occurs when

$$\frac{\partial V(\alpha)}{\partial \alpha} = \int \dots \int \frac{\theta^2(u) f^2(u)(-\frac{\partial g(u;\alpha)}{\partial \alpha})}{g(u;\alpha)^2} du = 0. \tag{6.105}$$

This is a minimum because $g(u;\alpha)$ is concave in α. Moy evaluates this using a single sampling distribution $g(u;\alpha_1)$ (α_1 fixed) in the trial study by rewriting

(6.105) as

$$\frac{\partial V(\alpha)}{\partial \alpha} = \int \cdots \int \frac{\theta^2(u) f^2(u)(-\frac{\partial g(u;\alpha)}{\partial \alpha})}{g(u;\alpha_1) g(u;\alpha)^2} g(u;\alpha_1) du = 0 \tag{6.106}$$

and sampling from $g(u;\alpha_1)$, then finding α for which the sample average

$$\sum_{i=1}^{n} \frac{\theta_i^2 f_i^2(-\frac{\partial g_i(\alpha)}{\partial \alpha})}{g_i(\alpha_1) g_i(\alpha)^2} \tag{6.107}$$

is zero. The choice of input distributions is done partly on the basis of making (6.107) solvable. The constraint that parameters for different input variables be the same is also done with this expression in mind, to limit the search for α to a small-dimensional space.

Hesterberg (1987) uses a trial study in the Fuel Inventory Probabilistic Simulator program to estimate the standard deviations needed in (6.103). In that example all internal variables have Gaussian distributions, and a trial study is used to estimate the Spearman correlation r_j of each of the input variables (temperature, hydro, etc.) with the sum of oil use, gas use and outage magnitude. The tilting parameter for each variable is then set as

$$\alpha_j = \beta r_j \sqrt{1 - \rho_j^2} \tag{6.108}$$

where ρ_j is the Spearman correlation of variable j with the previous month's variable as specified in the example input and β determines the overall level of tilting. The square root term corrects for the correlation between variables; the Spearman correlation of an internal variable with the sum of fuel use is approximately $\sqrt{1-\rho_j^2}$ times the correlation of the corresponding physical variable with the sum, since the Z-value obtained from the internal variable enters into the Z-value for the physical variable with that factor. β is set by trial and error and experience.

The input variables in the FIPS example are very heterogeneous, so constraining parameters to be equal is unsatisfactory. The method used provides an effective way of choosing the relative values of parameters.

6.4.4 Fuel Example Parameter Choice

The parameters used in the simplified fuel example simulation are also based on (6.101), with standard deviations determined this time by a physical considerations rather than a trial study.

The relationship between input and output variables in the fuel example is for the most part nonlinear. In the worst cases, however, the magnitude of an outage is approximately determined by the difference between total demands and total supplies. Most of the random variables are either demand or supplies; only temperature is not, but at low temperatures (where outages are most likely to occur) the relationship between temperature and demand is linear. Thus for accurate simulation in the worst cases the nonlinearity of the example can be ignored, and exponential tilting can be done based on a linear combination of the input variables which represents difference between demands and supplies.

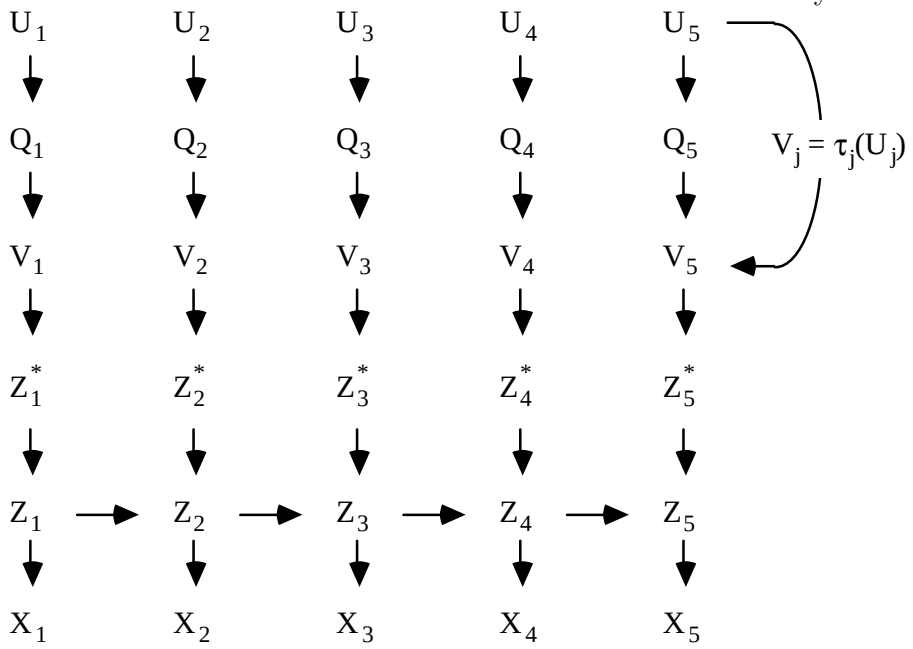
If the input variables were uncorrelated this could be done using external importance sampling, using the same tilting parameter for all input variables (modulo a sign change for demand vs. supplies, and a conversion factor for temperature). This is infeasible, however, when the input variables are specified in terms of their distributions and their correlations (except in the easy Gaussian case).

Internal importance sampling distributions are used to avoid this problem. The shape of the internal distribution is chosen to approximately match the shape of the corresponding input variable; Gaussian distributions for temperature, gas and electric demand, gamma distributions for the monthly hydroelectric generation levels, and tilted uniform variables for nuclear generation levels.

The random variable generation scheme is outlined in Figure 6.12 U is the original stream of random numbers. Q is the internal variable for which exponential tilting is done, and is not actually computed; instead V is computed directly using the relationship $V_j = \tau_j(U_j)$, where τ_j is defined on the basis of the distribution of Q_j, as described in Section 6.4.1. Q is normally distributed in this example for temperature, gas demand and electric demand, gamma for hydroelectric generation, and tilted uniform for nuclear generation. V_j is then used in the rest of the simulation as if it were uniformly distributed; Z_j^* is computed from V_j using the inverse normal transformation $Z_j^* = \Phi^{-1}(V_j)$, and is used to generating the correlation structure by $Z_j = \rho_j Z_{j-1} + \sqrt{1-\rho_j^2} Z_j^*$, and finally Z_j is used to generate X_j. The use of Q_j is a change in notation from Section 6.4.1, where Z_j referred to the internal variable; here Z_j denotes variables in the Markov process used to model the dependence between random variables in successive months.

Except for the correlation between variables, (6.101) would be satisfied

Figure 6.12: Random Variable Generation in the Fuel Inventory Example



by the choice:

$$\alpha_j = \beta \frac{\sigma(s_j(X_j))}{\sigma(Q_j)}, \tag{6.109}$$

where $\sigma(.)$ is the standard deviation under f and β determines the overall degree of tilting.

What is needed to generate here to determine the tilting parameter α_j is not $\sigma(s_j(X_j))$, but rather $\sigma(s_j^*(Z_j^*))$, the standard deviation of the contribution to the overall result from Z_j^*. That is, s_j^* satisfies (at least approximately)

$$\sum_{j=1}^{d} s_j(X_j) \approx \sum_{j=1}^{d} s_j^*(Z_j^*). \tag{6.110}$$

Now X_j is a function only of Z_j, and

$$\begin{aligned} Z_j &= \sqrt{1-\rho_j^2} Z_j^* + \rho_j Z_{j-1} \\ &= \sqrt{1-\rho_j^2} Z_j^* + \rho_j (\sqrt{1-\rho_{j-1}^2} Z_{j-1}^* + \ldots). \end{aligned} \tag{6.111}$$

Z_j^* enters the expansion for Z_k, $k \geq j$, with a term of $\sqrt{1-\rho_j^2}\rho_{j+1}\rho_{j+2}\ldots\rho_k$, where by convention $\rho_1 = 0$. Let

$$H_j := \sqrt{1-\rho_j^2}(\sigma(S_j) + \rho_{j+1}(\sigma(S_{j+1}) + \rho_{j+2}(\ldots))) \tag{6.112}$$

If $s_j(X_j)$ is linear in Z_j, then

$$H_j = \sigma(s_j^*(Z_j^*)). \tag{6.113}$$

If $s_j(X_j)$ is not linear in Z_j then (6.110) and (6.113) are only approximate equalities, and exact equalities can not be obtained; the best apparent approximation for $\sigma(s_j^*(Z_j^*))$ is still given by formula (6.113).

In the fuel example, α_j is set using (6.113), to obtain

$$\alpha_j = \beta \frac{H_j}{\sigma(Q_j)}. \tag{6.114}$$

The overall tilting parameter β is set so that the expected value of the sum of internal variables under importance sampling is approximately equal to the 95th percentile of the expected value without importance sampling. This choice is based on the reasoning discussed in Section 6.3.3, and is targeted

to estimating a probability at the 5% level. It is conservative for estimating smaller probabilities. The conservative choice is made here because the normal approximations are not perfectly satisfactory due to the heterogeneity of the input variables and the correlations between variables.

Let

$$S_3 := \sum_{j=1}^{d} \frac{H_j}{\sigma(Q_j)} Q_j, \tag{6.115}$$

then

$$\beta S_3 = \sum_{j=1}^{d} \alpha_j Q_j. \tag{6.116}$$

and exponential tilting Q_j with parameter α_j is equivalent to tilting S_3 with parameter β,

$$g_\alpha(S_3) = ce^{\beta S_3} g_0(S_3) = ce^{\beta S_3} f(S_3) \tag{6.117}$$

Now the variance of S_3 is $\sum H_j^2$. To tilt so the expected value of the tilted distribution of S_3 is at its 95th percentile, we make use of the exponential family identity:

$$\frac{\partial E_\alpha(Q)}{\partial \alpha} = \mathrm{Var}_\alpha(Q), \tag{6.118}$$

and note that

$$\mathrm{Var}_{\beta=0}(S_3) = \frac{\partial E(S_3)}{\partial \beta}_{|\beta=0} = \sum_{j=1}^{d} H_j^2 \tag{6.119}$$

Estimate the percentile using a normal approximation, and use a first-order Taylor-series approximation with respect to β for the change in the mean, and solve

$$\beta \frac{\partial E(S_3)}{\partial \beta}_{|\beta=0} = \Phi^{-1}(.95)\sqrt{\mathrm{Var}_{\beta=0}(S_3)} \tag{6.120}$$

to obtain

$$\beta = \frac{\Phi^{-1}(.95)}{\sqrt{\sum H_j^2}} \tag{6.121}$$

The errors introduced by the normal approximation and the Taylor-series expansion partially cancel.

Formulas (6.112) and (6.113) are appropriate when $Z_1, Z_2, \ldots$ form a Markov process. The result can be generalized for other covariance structures using a Cholesky decomposition. Let $\mathbf{M}$ be a covariance matrix of a multivariate normal distribution, and let $\mathbf{L}$ be the lower-triangular matrix

such that $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ (a Cholesky decomposition). Then a multivariate normal vector $\mathbf{Z}$ with covariance matrix $\mathbf{M}$ can be generated as $\mathbf{Z} = \mathbf{L}\mathbf{Z}^*$, where $\mathbf{Z}^*$ is a vector of independent standard normal random variables of the same length as $\mathbf{Z}$. Any linear combination of the $\mathbf{Z}$ variables,

$$S := \sum_k a_k \mathbf{Z}_k, \tag{6.122}$$

can be written in terms of the $\mathbf{Z}^*$ variables as

$$S = \sum_j \mathbf{Z}_j^* \sum_k a_k \mathbf{L}_{k,j}. \tag{6.123}$$

The sum over k is the coefficient on $\mathbf{Z}_j^*$. Substitute $\sigma(S_k)$ for a_k, and redefine H_j as

$$H_j := \sum_k \sigma(S_k)\mathbf{L}_{k,j}. \tag{6.124}$$

In the special case that $\mathbf{M}$ is the covariance matrix of a Markov chain of normally-distributed random variables with $\rho(Z_{j-1}, Z_j) = \rho_j$, $\mathbf{L}_{k,j} = \sqrt{1-\rho_j^2}\prod_{l=j+1}^{k}\rho_l$, and (6.124) is equivalent to (6.113).

In summary, let

$$\alpha_j = \beta\frac{\sigma(s_j(X_j))}{\sigma(Q_j)} \tag{6.125}$$

and

$$\beta = \frac{\Phi^{-1}(.95)}{\sqrt{\sigma^2(s_j(X_j))}} \tag{6.126}$$

if the input random variables are independent. Q_j is the internal variable used to generate X_j, and s_j is a function of X_j such that some function of $\sum_j s_j(X_j)$ is a good approximation to θ.

If the variables are dependent define

$$H_j := \sum_k \sigma(s_k(X_k))\mathbf{L}_{k,j} \tag{6.127}$$

where $\mathbf{L}$ is the lower triangular matrix obtained from a Cholesky decomposition of the correlation matrix of $\{s_1(X_1), s_2(X_2), \ldots s_d(x_d)\}$, and let

$$\alpha_j = \beta\frac{H_j}{\sigma(Q_j)} \tag{6.128}$$

and

$$\beta = \frac{\Phi^{-1}(.95)}{\sqrt{\sum H_j^2}}. \tag{6.129}$$

β could also be chosen using an optimization algorithm.

6.5 Dynamic Sampling

Sampling distributions may be static, where the distribution of each variable in a multivariate application is fixed in advance, or dynamic, where the conditional distribution depends on results to that point.

The use of dynamic sampling distributions is natural in many applications, particularly in the study of stochastic processes. Goyal et al. (1987) use dynamic importance sampling in the simulation of reliability of fault-tolerant computers, using a biased transition matrix until a failure is observed, then turning off the biasing.

Russian Roulette and splitting (Kioussis and Miller 1983) are complementary techniques. Think of the simulation of a single observation as a sequential process, with input random variables generated one at a time. If in the middle of this process the observation appears unlikely to produce interesting results, Russian Roulette is performed; the observation is killed with a certain probability, otherwise is given increased weight. If an observation appears likely to produce interesting results, the observation is split; two (or more) independent paths are simulated from that point and given weights which sum to the weight before splitting.

Dynamic sampling appears less useful in fixed-dimension applications. An early version of the FIPS program used a dynamic sampling method, whereby the degree of tilting was reduced for subsequent variables if the results from early variables indicated that a replication was not likely to produce an outage. It gave greatly improved results at the time, particularly in reducing the variance of the weights when outages are unlikely. However the method is unattractive, in that the marginal distribution for a variable depends on whether it is generated first or last, and there is a possibility that the tilting reduction would be performed prematurely, resulting in an outage occurring in combination with a large weight. Dynamic sampling was superseded by the development of mixture sampling, a simpler and more efficient way of achieving the same goals.

The author tried another dynamic sampling method in fixed-dimension applications, with mixed success—the method reduced the number of replications needed for a given level of accuracy, but required excessive computer time. This method is based on estimating the optimal form of a sampling distribution and generating each component of multivariate input from its marginal distribution, given the values of the previously generated components. That is, if $X = (X_1, X_2, \ldots X_d)$ and $g^*(X)$ is the estimated optimal distribution, generate X_j according to the estimated marginal distribution $g_j^*(x_j|x_1, x_2, \ldots x_{j-1})$. The estimated joint and marginal distributions were based on normal approximations to sums of independent random variables, and worked fairly well, but the method required computing the estimated marginal distribution at each step, which is computationally expensive, requiring $O(d^2)$ time in that case to generate X for a single replication, rather than the more typical $O(d)$ time. The method is also difficult to program. A combination of mixture sampling and exponential tilting is simpler and faster to implement and use.

Chapter 7

Conclusion

Importance sampling is a well-known Monte Carlo technique, known primarily as a variance reduction method, but it has other applications, as a way to solve otherwise intractable applications, and as a way to analyze results under multiple input distributions simultaneously.

The classical "integration" approach to importance sampling is well-suited to variance reduction applications that can be expressed in terms of distributions that have a large discrete mode at zero. The integration approach fails in more general applications, for expectations of distributions that do not have that form, for multivariate outcomes, for quantities other than expectations, and for comparison of multiple input distributions.

The first contribution of this work is the development of estimation methods that work well in more general applications. The "ratio" and "regression" methods, well-known in sampling theory, are particularly useful here, as both are simple ways to solve some problems of the integration estimate. The ratio and regression methods give distribution estimates with unit mass, and give equivariant expectation estimates. Other estimation formulas are described that can be applied in some small-sample applications where the regression estimate uses negative weights.

The second contribution is the use of conditional weights—replacing weights used for a distribution estimate of a quantity with their expected value given a sufficient statistic for that quantity. This reduces the undesirable component of the variance of the weights, resulting in better estimates. Future extensions of this work may prove valuable in the application of importance sampling to stochastic processes.

The third contribution is in the application of mixture distributions to

importance sampling. The use of the true distribution as one component in a mixture results in bounded weights, which allows importance sampling to be used in some applications where it would otherwise give bad results. In combination with the use of an equivariant estimate, this results in a bound on the variance increase that could be observed for any component of a multivariate application. Mixture sampling also makes the results less sensitive to mis-specifications of a sampling distribution.

The fourth contribution is also related to mixture distributions. In a multivariate application, better results can be obtained for all estimates by running a single experiment with a mixture distribution, rather than by running separate experiments for each quantity.

The fifth contribution is the extension of a general method for generating sampling distributions using transformations of the uniform random numbers used in the simulation. By interpreting such transformations as a method of exponential tilting the characteristics of desirable transformations become more clear, and appropriate transformations can be chosen. Parameter selection methods are given for both independent and dependent marginal distributions.

The intent of this work has been to develop methods that can be used in many applications. These methods do not improve on the incredible variance reductions achievable in simple applications using the classical integration method. Instead, they allow importance sampling to be profitably and safely applied in a wider variety of applications.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J, Rogers, W. H., Tukey, J. W. (1972) *Robust Estimates of Location: Survey and Advances*, Princeton University Press.

Bauwens, L. (1984), *Bayesian Full Information Analysis of Simultaneous Equation Models Using Integration by Monte Carlo*, New York: Springer Verlag.

Becker, R. A., and Chambers, J. M. (1984), *S: An Interactive Environment for Data Analysis and Graphics,* Belmont, CA: Wadsworth.

Beckman, R. J. and McKay, M. D. (1987), "Monte Carlo Estimation Under Different Distributions Using the Same Simulation," *Technometrics* 29 153-160.

Bhattacharya, R. N., and Ghosh, J. K. (1978), "On the Validity of the Formal Edgeworth Expansion," *Annals of Statistics* 6, 434-451.

Booth, T. E. (1986), "A Monte Carlo Learning/Biasing Experiment with Intelligent Random Numbers", *Nuclear Science and Engineering* 92, 465-81.

Bratley, P., Fox, B. L., and Schrage, L. E. (1983), *A Guide to Simulation,* New York: Springer Verlag.

Butler, J. W. (1956), "Machine Sampling From Given Probability Distributions", *Symposium on Monte Carlo Methods*, ed H. A. Meyer, New York: Wiley, pp. 249-264.

Chambers, J. M., Mallows, C. L., and Stuck B. W. (1976), "A Method for Simulating Stable Random Variables", *Journal of the American Statistical Association*, 71, 340-44.

Chung, K.L. (1974), *A Course in Probability,* New York: Academic Press.

Clark, F. H. (1966), "The Exponential Transform as an Importance-Sampling Device: A Review", Technical Report ORNL-RSIC-14, Oak Ridge National Laboratory.

Cochran, W. G. (1977), *Sampling Techniques,* New York: John Wiley.

Conway, A. E. & Goyal, A. (1987), "Monte Carlo Simulation of Computer System Availability/Reliability Models" *Proceedings of the Seventeenth Symposium on Fault-Tolerant Computing*, Pittsburgh, Pennsylvania, 230-235.

Cramér, H. (1945), *Mathematical Methods of Statistics*, Princeton University Press.

Daniels, H. E. (1954), "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, 25, 631-650.

Davis, B. R. (1987), "An Improved Importance Sampling Method for Digital Communication System Simulations", *IEEE Transations on Communications*, COM-34, 715-719.

DiCiccio, T. and Tibshirani, R. (1987), "Bootstrap Confidence Intervals and Bootstrap Approximations," *Journal of the American Statistical Association*, 82, 163-170.

Efron, B. (1981), "Non-parametric Standard Errors and Confidence Intervals," *The Canadian Journal of Statistics*, 9, 139-172.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1987), "Better Bootstrap Confidence Intervals", *Journal of the American Statistical Association*, 82, 171-185.

Glynn, P. W. (1986), "Stochastic Approximation for Monte Carlo Optimization", *Proceedings of the 1986 Winter Simulation Conference*, 356-364

Glynn, P. W. (1987), "Likelihood Ratio Gradient Estimation: An Overview", *Proceedings of the 1987 Winter Simulation Conference*, 366-374.

Glynn, P. W., and Iglehart, D. L. (1987), "Importance Sampling for Stochastic Simulations", Technical Report No. 49, Department of Operations Research, Stanford University.

Goyal A., Heidelberger P., and Shahabuddin P. (1987) "Measure Specific Dynamic Importance Sampling for Availability Simulations" *Proceedings of the 1987 Winter Simulation Conference*, 351-357.

Hahn, P. H., Jeruchim, M. C. (1987), "Developments in the Theory and Application of Importance Sampling", *IEEE Transations on Communications*, COM-35, 706-714.

Hammersley, J. M., and Hanscomb, D. C. (1964), *Monte Carlo Methods,* London: Methuen.

Hampel, F. (1968) "Contributions to the Theory of Robust Estimation", Ph.D. Thesis, Berkeley.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383-393.

Hesterberg, T. C. (1987), "Importance Sampling in Multivariate Problems," *Proceedings of the Statistical Computing Section, American Statistical Association 1987 Meeting*, 412-417.

Hopmans, A. and Kleijnen, J.P.C., (1979) "Importance Sampling in Systems Simulation: a Practical Failure?," *Mathematics and Computers in Simulation* 21, 209-220.

International Mathematical and Statistical Libraries, Inc. (1984), *IMSL Library User's Manual.*

Johns, M. V. (1987), "Importance Sampling for Bootstrap Confidence Intervals," Department of Statistics, Stanford University, to appear in *Journal of the American Statistical Association.*

Johnson, N. J. (1978), "Modified t Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73, 536-544.

Kahn, H. (1950), Nucleonics, 6(5), 27-37 and 6(6) 60-65.

Kahn, H. and Marshall, A. W. (1953), "Methods of Reducing Sample Size in Monte Carlo Computations," *Journal of the Operations Research Society of America*, 1, 263-278.

Kioussis, L. C., and Miller, D. R. (1983), "An Importance Sampling Scheme for Simulating the Degradation and Failure of Complex Systems During Finite Missions", *Proceedings of the 1983 Winter Simulation Conference*, 631-639.

Kloek, T. and Van Dijk, H. K. (1978) "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica* 46(1) 1-19.

Kennedy, W. J. Jr. and Gentle, J. E. (1980), *Statistical Computing*, New York: Marcel Dekker.

Kleijnen, J.P.C. (1974), *Statistical Techniques in Simulation, Part 1*, New York: Marcel Dekker.

Knuth, D. E. (1981), *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*, 2nd Edition, Reading Massachusetts: Addison Wesley.

Luzer, V. and Olkin, I. (1988), "Estimation of the Ordered Characteristic Roots of a Random Covariance Matrix," Technical Report, Department of Statistics, Stanford University.

Marsaglia, G. (1961), "Expressing a Random Variable in Terms of Uniform

Random Variables," *Annals of Mathematical Statistics*, 32, 894-898.

Moy, W. A. (1965), "Sampling Techniques for Increasing the Efficiency of Simulations of Queueing Systems", Ph.D. dissertation, Industrial Engineering and Management Science, Northwestern University.

Murthy, K. P. N. and Indira, R. (1986), "Analytical Results of Variance Reduction Characteristics of Biased Monte Carlo for Deep-Penetration Problems," *Nuclear Science and Engineering* 92, 482-487.

Reiman, M. I., and Weiss, A. (1986), "Sensitvity Analysis Via Likelihood Ratios", *Proceedings of the 1986 Winter Simulation Conference*, 285-289.

SAS Institute, Inc. (1985), *SAS User's Guide: Basics, Version 5 Edition*, Cary, NC: SAS Institute Inc.

Siegmund, D. (1976) "Importance Sampling in the Monte Carlo Study of Sequential Tests," *The Annals of Statistics* 4, 673-684.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C. and Dransfield, M., (1985) "The Implementation of the Bayesian Paradigm," *Communications in Statististics-Theory and Methods*, 14, 1079-1102.

Stewart L. (1979), "Multiparameter Univariate Bayesian Analysis" *Journal of the American Statistical Association*, 74, 684-693.

Stewart L. (1983), "Multiparameter Bayesian Inference Using Monte Carlo Integration—Some Techniques for Bivariate Analysis," *Bayesian Statistics 2*, eds J.BM. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, Elsevier Science Publishers B.V. (North-Holland), 495-510.

Therneau, T. M. (1983), "Variance Reduction Techniques for the Bootstrap," *Technical Report No. 200 (Ph.D. Thesis)* Department of Statistics, Stanford University.

Tibshirani, R. J., (1984) "Bootstrap Confidence Intervals", Technical Report LCS-3, Department of Statistics, Stanford University.

Tukey, J. W. (1987), "Configural Polysampling," *SIAM REVIEW* 29, 1-20.

Van Dijk, H.K. and Kloek, T. (1983), "Experiments with Some Alternatives for Simple Importance Sampling in Monte Carlo Integration," Bayesian Statistics 2, eds J.BM. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, Elsevier Science Publishers B.V. (North-Holland), 511-530.

Von Neumann, J. (1951), "Various Techniques Used in Connection with Random Digits", National Bureau of Standards symposium, NBS Applied Mathematics Series 12, National Bureau of Standards, Washington, D.C.

Wilson, J. R. (1984), "Variance Reduction Techniques for Digital Simulation," *American Journal of Mathematical and Management Sciences* 4, 277-312.