Safe and effective importance sampling

Art Owen, Yi Zhou
Stanford University

June 1998

Abstract

Importance sampling is a widely applied method for variance reduction in simulations. However, a naive application of importance sampling can result in a great increase in variance. This article surveys and extends some recently developed methods that eliminate this problem: defensive mixture sampling of Hesterberg (1995) and multiple importance sampling of Veach (1997). The method we propose is to importance sample from a mixture density using as control variates the ratios of mixture components to the mixture density. The resulting method is never much worse than importance sampling from the best of the mixture components, and can be much better. Finally, it is well known that importance sampling can be nearly perfect for nonnegative integrands. We show that this is also true for integrands taking both positive and negative values, using a modification of multiple importance sampling.

KEY WORDS: bidirectional path sampling, control variates, Monte Carlo, value at risk, variance reduction

1 Introduction

This paper presents and extends some recent variations on the fundamental idea of importance sampling. The ideas presented are elementary, and practical, but are not yet mentioned in texts on simulation, or gathered together in one place. The methods described here allow one: to exploit importance sampling without the sometimes disastrous loss of accuracy it can produce,

to effectively combine several importance distributions, and to extend the potential for extremely accurate simulations to integrands taking both positive and negative values. To do this we present and extend ideas from two dissertations (Hesterberg 1988, Veach 1997), and related publications (Hesterberg 1995, Veach & Guibas 1995).

We assume that the reader is already familiar with importance sampling, and some other standard variance reduction techniques in Monte Carlo, such as stratification and control variates. These methods are treated in introductory texts such as: Bratley, Fox & Schrage (1987), Ripley (1987), and Rubinstein (1981). Our interest in these variations on importance sampling arose while working on adaptive importance sampling. That work appears separately as Owen & Zhou (1998).

The problem we face is to compute an approximation to

$$I = \int_{(0,1)^d} f(x)dx \tag{1}$$

for some integrand f. Many problems can be turned into this form through change of variables, and other domains can be mapped onto the unit cube, so there is little loss of generality. Integrals written without a domain will be assumed to be over $(0,1)^d$.

In high dimensions it is common to compute I by simulation methods. Observations X_i, $i = 1, \ldots, n$ are chosen from the unit cube and the estimate

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{2}$$

is computed. The standard Monte Carlo method takes observations X_i from the uniform distribution on $(0,1)^d$. Quasi-Monte Carlo simulations take observations X_i so that the empirical distribution putting probability $1/n$ on each of the X_i, is closer to $U(0,1)^d$ than it would be by random sampling. This can increase the accuracy of $\hat{I}$. See Niederreiter (1992) for a description of quasi-Monte Carlo sampling.

In some problems from high energy physics and Bayesian statistics it happens that f is nearly zero or nearly constant over all but a small portion of $(0,1)^d$, and that the integral I is dominated by the contribution to f from that region. For such spiky integrands, there is no point in sampling the whole space uniformly. It is more efficient to concentrate more observations in the small region where f varies. This entails sampling from a distribution

other than the uniform one. The name "importance sampling" is applied to methods that sample from the important part of the input space.

In some other problems there may be several spikes, or several locations that we suspect might have spikes, or we may want to estimate several related integrands with possibly different spike locations. This leads us to consider importance sampling from more than one density.

As an example, in some finance applications, one might be interested in stress testing a portfolio to find the chance of severe losses. This might involve importance sampling from a distribution in which stocks have a reduced tendency to drift upwards, or even a tendency to drift downwards. Another sampling distribution would have stocks exhibiting extra volatility. For a portfolio that has hedged against these possibilities, it could be upward trending stock prices or decreased volatility that poses the greatest risk. These conditions might therefore provide two additional importance sampling densities to consider. Finally, the model believed to be most reasonable could also be sampled.

A second class of examples arises from rendering problems in computer graphics. This problem was brought to our attention by Veach & Guibas (1995) and Veach (1997). The light transport problem has as its goal the production of realistic images by sampling photon paths. Each photon starts at a light source and is reflected some number of times before finally being absorbed. Only photon paths that terminate at a detector, such as a virtual camera, contribute to the final image. This image is determined through various energy integrals over a space of photon paths. It can pay to combined multiple sampling procedures. Some paths can be generated by direct physical simulation of photons, but it is likely that most will miss the detector. Adjoint methods sample paths in reverse order starting from the detector. Bi-directional paths sample from both ends and meet in the middle. For some images it is desirable to combine versions of all three methods.

The outline of this paper is as follows. Section 2 presents a basic discussion of importance sampling and control variates and how they may be combined. Practitioners have long known that importance sampling can greatly increase the variance in some cases, where the importance sampling density has short tails. Section 3 presents defensive importance sampling, an ingenious method of Hesterberg (1995) for preventing importance sampling from back-firing. When combined with suitable control variates, defensive importance sampling produces a variance that is never much worse than the standard Monte Carlo variance, providing some insurance against the worst

effects of importance sampling. Defensive importance sampling can however be much worse than the original importance sampling. We show that using appropriate control variates with defensive importance sampling (as described by Hesterberg (1995)) is never much worse than the original non-defensive importance sampling. This provides a bound on the premium we have to pay for that insurance.

Section 4 presents the method that we generally recommend: importance sampling from a mixture of m sampling densities with m control variates, one for each mixture component. Theorem 1 shows that this method is never much worse than pure importance sampling from any single component of the mixture. At worst, the variance is multiplied by the reciprocal of the corresponding mixture probability. Section 4 also compares deterministic mixtures to random mixtures.

Section 5 presents multiple importance sampling, another ingenious method, similar to defensive importance sampling, due to Veach & Guibas (1995) and Veach (1997). It is a standard practice to weight observations in inverse proportion to their sampling probability. Multiple importance sampling can break that rule, and do so in a way that still results in an unbiased estimate of the integral. The motivating idea is that in some parts of the sample space, f may be roughly proportional to one of the sampling densities while other densities are appropriate to other parts of the space. The goal is to place greater weight on those locally most appropriate densities. Veach & Guibas (1995) present several examples of rules for weighting the samples, including their balance heuristic, which is the standard weighting. They provide two theorems on the central role of the balance heuristic, bounding the size of the possible improvement from other weightings. These theorems are compared to Theorem 1, which gives sharper performance bounds at the cost of requiring the estimation of control variate coefficients.

It is well known that importance sampling can be nearly perfect for non-negative integrands. Section 6 shows how the idea of multiple importance sampling can be used to extend this result to integrands taking both positive and negative values.

Section 7 presents three examples to illustrate the methods of this article. The first two examples feature importance sampling with a density nearly proportional to the integrand. In the first example, importance sampling has infinite variance, which can be made finite by defensive importance sampling. In the second, example importance sampling is extremely accurate, but defensive importance sampling, without control variates, destroys the accuracy.

The method of Section 4, mixture importance sampling with control variates performs as predicted by theory: it cures the infinite variance problem of the first example without losing the extreme accuracy of importance sampling in the second example. No other method we tried achieved this. The third example illustrates how importance sampling can be nearly exact even for integrands of mixed sign. Again the method of Section 4, mixture importance control variates is nearly best. Of the near best methods, it is the simplest to implement.

2 Variance reduction

This section reviews two variance reduction methods. The first is importance sampling, the second is control variates.

2.1 Importance sampling

Importance sampling begins by writing the integral as

$$I = \int \frac{f(x)}{p(x)} p(x)dx$$

where $p(x)$ is a probability density function on $(0,1)^d$. One then takes a sample X_i independently from $p(x)$ and estimates I by

$$\hat{I}^{(p)} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{p(X_i)}. \tag{3}$$

The requirements on p that allow it to be used in importance sampling are

IS-1 There must be an algorithm to draw $X_i \sim p$.

IS-2 It must be possible to evaluate p at any X_i.

Strictly speaking, condition IS-2 only needs to be that we can evaluate f/p but we assume that we can evaluate f and so being able to evaluate f/p is equivalent to being able to evaluate p.

Elementary manipulations give that

$$V(\hat{I}^{(p)}) = \frac{1}{n}\int \left(\frac{f(x)}{p(x)} - I\right)^2 p(x)dx = \frac{1}{n}\left[\int \frac{f(x)^2}{p(x)}dx - I^2\right]. \tag{4}$$

If f is nonnegative, the optimal density p is $p_o = f(x)/I$, apart from the trivial case with $I = 0$. This density gives $V(\hat{I}^{(p_o)}) = 0$. But if IS-2 is satisfied for this density, then we must already be able to compute the integrand I, and so importance sampling becomes irrelevant. It remains generally true that taking p approximately proportional to f can lead to very effective variance reductions.

The target f and density p are well matched if $f(x) \doteq Ip(x)$. Define $r(x) = f(x) - Ip(x)$. This residual has $\int r(x)dx = 0$ by construction. Furthermore we can re-write (4) as

$$V(\hat{I}^{(p)}) = \frac{1}{n}\left[\int \frac{r(x)^2}{p(x)}dx\right]. \tag{5}$$

For comparison, standard Monte Carlo sampling with X_i iid from the $U(0,1)^d$ distribution has a variance σ^2/n where

$$\sigma^2 = \int (f(x) - I)^2 dx. \tag{6}$$

2.2 Control variates

The method of control variates uses knowledge of one integral to reduce variance in the estimate of another. Suppose that we know $\int h(x)dx = I_h$. Writing $I = (I - I_h) + I_h$ we see that the "difference estimator"

$$\hat{I}_D = \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i) - h(X_i)\Big) + I_h,$$

is unbiased for I. If $I_h \neq 0$ we can write $I = (I/I_h)I_h$ and this motivates the ratio estimator

$$\hat{I}_R = \frac{\sum_{i=1}^n f(X_i)}{\sum_{i=1}^n h(X_i)} I_h.$$

A third way to use knowledge of I_h is to write $I = (I - \beta I_h) + \beta I_h$, for some scalar β, and this motivates the regression estimator

$$\hat{I}_\beta = \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i) - \beta h(X_i)\Big) + \beta I_h.$$

The difference estimator is successful to the extent that $f - h$ has smaller variance than f. The ratio estimator is most successful if f is nearly proportional to h. The regression estimator is successful if $f(x) - \beta h(x)$ is nearly

constant. The variance minimizing choice of β is easily seen to be

$$\beta^o = \frac{\int (f(x) - I)(h(x) - I_h)dx}{\int (h(x) - I_h)^2 dx}.$$

The regression estimator is most widely used because it succeeds when either ratio or difference estimation would. That is $\lim_{n\to\infty} V(\hat{I}_{\beta^o})/V(\hat{I}_D) \le 1$ and $\lim_{n\to\infty} V(\hat{I}_{\beta^o})/V(\hat{I}_R) \le 1$. See Cochran (1977).

It is customary to estimate the optimal β from the data, for example by

$$\hat{\beta} = \frac{\sum_{i=1}^n (f(X_i) - \bar{f})(h(X_i) - \bar{h})}{\sum_{i=1}^n (h(X_i) - \bar{h})^2},$$

where $\bar{f} = (1/n)\sum_{i=1}^n f(X_i)$, and $\bar{h} = (1/n)\sum_{i=1}^n h(X_i)$, and then to use the estimate

$$\hat{I}_{\hat{\beta}} = \frac{1}{n}\sum_{i=1}^n \Big(f(X_i) - \hat{\beta} h(X_i) \Big) + \hat{\beta} I_h.$$

Under mild conditions the estimate $\hat{\beta}$ converges to the optimal value β^o with $n^{1/2}(\hat{\beta} - \beta^o)$ having an asymptotic normal distribution, and

$$\lim_{n\to\infty} \frac{V(\hat{I}_{\hat{\beta}})}{V(\hat{I}_{\beta^o})} = 1. \tag{7}$$

Notice that $\hat{I}$ is not usually unbiased when the same data are used to compute $\hat{\beta}$ and $\hat{I}$. The bias is typically of order n^{-1} and so is eventually negligible when compared to the standard deviation which is of order $n^{-1/2}$.

In some applications it is desirable to avoid even so small a bias. For example, in some computer graphics problems variance produces a speckled look that is obvious to the user and can be eliminated by further sampling, while bias can produce artifacts, such as dark patches, that may not be obviously wrong. Similarly, in financial applications one might seek to avoid even a small bias, because in repeated applications it could result in a small but steady loss of money, whereas a slightly larger variance would tend to yield offsetting gains and losses in multiple uses.

The bias can be eliminated by computing

$$\hat{I} = \frac{1}{n}\sum_{i=1}^n \Big(f(X_i) - \beta_i h(X_i) + \beta_i I_h \Big)$$

where β_i is an estimate of β^o computed independently of X_i. One approach is to compute $\hat{\beta}_i$ based on some historical data $X_{-h+1}, \ldots, X_0$ that is not otherwise used in the estimation of I. Another approach is to keep a running estimate β_i based on observations X_j, $j < i$. Perhaps the best approach is to employ cross-validation, basing β_i on all observations X_j, $j \neq i$.

Given a list of control variates $h_1, \ldots, h_m$ with known integrals $I_{h_1}, \ldots, I_{h_m}$, one can use the estimator

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i) - \sum_{j=1}^{m}\beta_j h_j(X_i)\Big) + \sum_{j=1}^{m}\beta_j I_{h_j}.$$

The optimal β_j can be estimated by multiple regression, and bias corrected if so desired.

2.3 Importance sampling with control variates

It is natural to combine control variates and importance sampling. Suppose that one knows the integral $\int h(x)dx = I_h$. Then when sampling from $p(x)$ the expected value of $h(x)/p(x)$ is I_h, and so h/p may be used as a control variate. Thus the estimate

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i) - \beta h(X_i)}{p(X_i)} + \beta I_h$$

may be used. The optimal β can be estimated by regression of f/p on h/p on a sample from p.

The trivial function $h(x) = 1$ has known integral $I_h = 1$. This leads to the use of the nontrivial control variate $1/p(x)$. Hesterberg (1988) considers using this control variate as a ratio estimator yielding

$$\hat{I} = \frac{\sum_{i=1}^{n} f(X_i)/p(X_i)}{\sum_{i=1}^{n} 1/p(X_i)}. \tag{8}$$

A strong motivation for (8) is that it gives the right answer for constant f. This means that for two integrands with a known sum or difference, the corresponding estimated integrals will have that known sum or difference. This kind of consistency is achieved by weighting the observations $f(X_i)$ with weights that sum to unity.

When used as a regression estimator the result is

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i) - \hat{\beta}}{p(X_i)} + \hat{\beta}, \tag{9}$$

with $\hat{\beta}$ estimated from a regression of f/p on $1/p$.

Trotter & Tukey (1956) state that experience shows that (3) is almost always superior to (8). Hesterberg (1995) agrees with this for the estimation of rare event probabilities but for more general use recommends the regression estimator.

3 Defensive importance sampling

It is clear from variance formula (4) that $p \ll f$ can cause trouble, even if $f \doteq Ip$. Using "short tailed" p to approximate a "long tailed" f can back-fire. A remedy known as defensive importance sampling (Hesterberg 1995) is to mix into p a very long tailed distribution. Let $p_\alpha(x) \equiv \alpha p_0(x) + (1-\alpha)p(x)$ where p is a good approximation to f/I, p_0 has wide tails, as for example the $U(0,1)^d$ distribution has, and $\alpha \in (0,1)$ is a constant. We assume in the following that p_0 is the $U(0,1)^d$ distribution, so that

$$p_\alpha(x) = \alpha + (1-\alpha)p(x).$$

For defensive importance sampling with $p_0 = U(0,1)^d$, we have $1/p_\alpha(x) \leq 1/\alpha$ from which it follows that

$$V(\hat{I}) \leq \frac{1}{n\alpha}\left[\sigma^2 + (1-\alpha)I^2\right]. \tag{10}$$

Equality holds in (10) when $f(x) = 0$ everywhere $p(x) \neq 0$. Thus defensive importance sampling provides an upper bound on the variance that holds even if $p(x)$ is very small in places. Hesterberg (1995) recommends values of α between 0.1 and 0.5.

Hesterberg (1995) also uses $1/p_\alpha(x)$ as a control variate in defensive importance sampling. For expository purposes, we prefer to take $-1/p_\alpha(x)$ as the control variate; this has no effect on the accuracy of the method. Thus let

$$\hat{I}_{\alpha,\delta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)+\delta}{p_\alpha(X_i)} - \delta.$$

Under sampling of X_i from p_α, we find $E(\hat{I}_{\alpha,\delta}) = I$, and

$$V(\hat{I}_{\alpha,\delta}) = \frac{1}{n}\left[\int \frac{(f(x)+\delta)^2}{p_\alpha(x)}dx - (I+\delta)^2\right]. \tag{11}$$

The reciprocal $1/p_\alpha$ appears in numerous integrals. We introduce the shorthand notation

$$J(g) = \int \frac{g(x)}{p_\alpha(x)} dx, \tag{12}$$

for use in equations with many such integrals. The notation $J(1)$ refers to $J(g)$ for the function g that is everywhere equal to 1. For scalars a, b, $J(a + bg) = aJ(1) + bJ(g)$. We also note that $J(\alpha + (1-\alpha)p) = 1$.

The insurance property (10), applied to $f(x) + \delta$, becomes

$$V(\hat{I}_{\alpha,\delta}) \le \frac{1}{n\alpha}\Big(\sigma^2 + (1-\alpha)(I+\delta)^2\Big). \tag{13}$$

It is clear from (13) that if we could take $\delta = -I$, then we would find $V(\hat{I}_{\alpha,-I}) \le \sigma^2/n\alpha$.

The optimal value of δ would of course be even better than $\delta = -I$. The right side of (11) is a quadratic in δ with nonnegative curvature. Assuming that $p(x)$ is not equivalent to the $U(0,1)^d$ distribution, the quadratic has positive curvature $J(1) - 1$ and a unique minimum at the optimal value

$$\delta^o = \frac{I - J(f)}{J(1) - 1}. \tag{14}$$

This optimal value may be estimated from the data by regression:

$$\hat{\delta}_n = -\frac{\sum_{i=1}^n (1/p_\alpha(X_i) - \overline{1/p_\alpha})(f(X_i)/p_\alpha(X_i) - \overline{f/p_\alpha})}{\sum_{i=1}^n (1/p_\alpha(X_i) - \overline{1/p_\alpha})^2},$$

where $\overline{1/p_\alpha} = (1/n)\sum_{i=1}^n 1/p_\alpha(X_i)$ and $\overline{f/p_\alpha} = (1/n)\sum_{i=1}^n f(X_i)/p_\alpha(X_i)$.

Because $\hat{\delta}_n$ is eventually as good as δ^o which in turn is at least as good as $-I$ we find

$$\lim_{n\to\infty} nV(\hat{I}_{\alpha,\hat{\delta}_n}) \le \frac{\sigma^2}{\alpha}. \tag{15}$$

Hesterberg (1995) also obtains (15) using a comparison to the ratio estimate. Equation (15) shows that, asymptotically, defensive importance sampling with n observations and the control variate $1/p_\alpha(x)$ is never worse than ordinary Monte Carlo sampling with a sample of size $n\alpha$. Thus the recommended values 0.1 and 0.5 provide insurance against being one tenth or one half as efficient, respectively, as ordinary Monte Carlo.

3.1 Cost of defensive importance sampling

Defensive importance sampling provides insurance against doing much worse than ordinary Monte Carlo. The premium we pay for this insurance is possibly much worse performance than under the original $\alpha = 0$ importance sampling scheme. Here we show that, if one does not use $1/p_\alpha$ as a control variate, that the result can be an arbitrarily large increase in variance.

Let $p(x)$ be a density function and write $f(x) = Ip(x) + r(x)$ where $\int r(x)dx = 0$. Now suppose that $|r(x)| \le \epsilon p(x)$ for all x. Then importance sampling achieves the variance

$$\frac{1}{n}\int \frac{r(x)^2}{p(x)}dx \le \frac{\epsilon^2}{n}.$$

If we let the match-up between f and p improve by sending $\epsilon \to 0$, then the importance sampling variance decreases to zero as ϵ^2.

It is easy to construct cases in which defensive importance sampling without a control variate does not share this improvement. To avoid trivialities we take $I \neq 0$ and p a density other than $U(0,1)^d$. Suppose that

$$\int \frac{r(x)p(x)}{p_\alpha(x)}dx \ge 0. \tag{16}$$

If (16) fails to hold for some $r(x)$ it surely does hold for $-r(x)$. Now note that, for $\alpha > 0$,

$$D(\alpha) = \int \frac{p^2(x)}{p_\alpha(x)}dx - 1 > 0. \tag{17}$$

Equation (17) holds because $D(0) = D'(0) = 0$ and $D''(\alpha) > 0$. Finally, the defensive importance sampling variance is

$$\frac{1}{n}\left[\int \frac{(Ip(x) + r(x))^2}{p_\alpha(x)}dx - I^2\right] \ge \frac{D(\alpha)I^2}{n}$$

and hence does not approach 0 as $\epsilon \to 0$.

In cases where importance sampling succeeds spectacularly, defensive importance need not share the success. We show below that this failure can be fixed by using $1/p_\alpha(x)$ as a control variate.

3.2 A naive control variate

For a smooth integrand f with $\inf_x f(x) = 0$, the defensive importance sampling density p_α cannot approximate f arbitrarily closely without having α approach zero, thus losing the insurance properties (10) and (15).

But it is possible for $f+\delta$ to be a good approximation to $\alpha+(1-\alpha)p$ for some constant δ. A naive first approach is to note that when $f(x) = Ip(x)$, then ordinary importance sampling gives zero variance, and that this zero variance can be recovered in mixture sampling by taking

$$\delta = \delta^\star \equiv \frac{\alpha}{1-\alpha} I, \tag{18}$$

because

$$\int \frac{(Ip(x)+\delta^\star)^2}{p_\alpha(x)} dx - (I+\delta^\star)^2 = 0. \tag{19}$$

One might then anticipate that for $f \doteq Ip$, the naive choice $\delta^\star$ should still produce a low variance. Finally, the optimal choice δ^o should be even better.

Proposition 1 shows that defensive importance sampling with the naive value $\delta^\star$ is never much worse than the original $\delta = \alpha = 0$ importance sampling, for α not close to 1. Proposition 1 also gives an expression for the difference in variance between the naive and optimal choices for δ.

Proposition 1 *Write $f(x) = Ip(x) + r(x)$ where $\int r(x)dx = 0$. Then*

$$V(\hat{I}_{\alpha,\delta^\star}) = \frac{1}{n}\int \frac{r(x)^2}{p_\alpha(x)}dx \le \frac{1}{1-\alpha} V(\hat{I}_{0,0}) \tag{20}$$

and

$$V(\hat{I}_{\alpha,\delta^o}) = V(\hat{I}_{\alpha,\delta^\star}) - \frac{1}{n}\frac{\left(\int \frac{r(x)}{p_\alpha(x)}dx\right)^2}{\int \frac{1}{p_\alpha(x)}dx - 1}. \tag{21}$$

Proof: First note that $Ip(x) + \delta^\star = Ip_\alpha(x)/(1-\alpha)$. Therefore

$$\begin{aligned}
&\int \frac{(Ip(x)+\delta^\star+r(x))^2}{p_\alpha(x)}dx \\
= \;& \frac{I^2}{(1-\alpha)^2}\int \frac{p_\alpha(x)^2}{p_\alpha(x)}dx + 2\frac{I}{(1-\alpha)}\int \frac{r(x)p_\alpha(x)}{p_\alpha(x)}dx + \int \frac{r(x)^2}{p_\alpha(x)}dx \\
= \;& \left(\frac{I}{1-\alpha}\right)^2 + \int \frac{r(x)^2}{p_\alpha(x)}dx,
\end{aligned}$$

and so

$$\begin{aligned} V(\hat{I}_{\alpha,\delta^\star}) &= \frac{1}{n}\left[\left(\frac{I}{1-\alpha}\right)^2 + \int \frac{r(x)^2}{p_\alpha(x)}dx - (I+\delta^\star)^2\right] \\ &= \frac{1}{n}\int \frac{r(x)^2}{p_\alpha(x)}dx \\ &\le \frac{1}{n}\int \frac{r(x)^2}{(1-\alpha)p(x)}dx \\ &= \frac{1}{1-\alpha}V(\hat{I}_{0,0}). \end{aligned}$$

The right side of equation (11) is a quadratic in δ with curvature $2(J(1)-1)/n$. The value at $\delta^\star$ exceeds that at the minimum δ^o by one half the curvature times $(\delta^\star-\delta^o)^2$. Therefore

$$n(V(\hat{I}_{\alpha,\delta^\star}) - V(\hat{I}_{\alpha,\delta^o})) = (\delta^\star-\delta^o)^2(J(1)-1). \tag{22}$$

Finally,

$$\begin{aligned} \delta^\star-\delta^o &= \frac{\frac{\alpha}{1-\alpha}I(J(1)-1)+J(f)-I}{J(1)-1} \\ &= \frac{J(\alpha I+(1-\alpha)f)-I}{(1-\alpha)(J(1)-1)} \\ &= \frac{J(\alpha I+(1-\alpha)(Ip+r))-I}{(1-\alpha)(J(1)-1)} \\ &= \frac{J(r)}{J(1)-1}, \end{aligned}$$

which combined with (22) establishes (21). □

By using the estimated value $\hat{\delta}_n$ we achieve a variance that is eventually not much worse than $\sigma^2/(\alpha n)$ or $\int(r^2/p)dx/((1-\alpha)n)$. For $\alpha = 0.5$ the result is never more than twice as bad as the better of ordinary Monte Carlo and importance sampling. For $\alpha = 0.1$ the result is never more than ten times as bad as ordinary Monte Carlo and never more than 10/9 times as bad as importance sampling.

Results (15) and (20) look very symmetric, with reversed roles for p and $U(0,1)^d$, and simultaneous replacement of α by $1-\alpha$. To complete the reversal, we would have to replace the control variate $1/p_\alpha(x)$ by $p(x)/p_\alpha(x)$. But because $\alpha 1/p_\alpha + (1-\alpha)p/p_\alpha = 1$ for all x, the control variates are equivalent so long as the regression includes an intercept.

4 General mixtures

Here we extend the idea to sampling from a list of importance distributions p_j, $j = 1, \ldots, m$, not necessarily including the $U(0,1)^d$ distribution, and using control variates for them all. Take mixing proportions $\alpha_j > 0$ and $\sum_{j=1}^m \alpha_j = 1$ and use the estimator

$$\hat{I}_{\alpha,\delta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i) + \sum_{j=1}^m \delta_j p_j(X_i)}{\sum_{j=1}^m \alpha_j p_j(X_i)} - \sum_{j=1}^m \delta_j, \tag{23}$$

where α and δ are now vectors and X_i are independent draws from the mixture $\sum_{j=1}^d \alpha_j p_j(x)$. Theorem 1 below shows that $\hat{I}_{\alpha,\delta}$ is unbiased and that, for the optimal δ, the variance is never larger than what one gets from an importance sample of size $n\alpha_j$ from p_j.

Theorem 1 *Let p_j, α_j and $\hat{I}_{\alpha,\delta}$ be as above. Then*

$$E(\hat{I}_{\alpha,\delta}) = I, \tag{24}$$

and there is a choice of $\delta_1^o, \ldots, \delta_m^o$ for which,

$$V(\hat{I}_{\alpha,\delta^o}) \le \min_{1\le j\le m} \frac{1}{\alpha_j n}\left[\int \frac{f(x)^2}{p_j(x)}dx - I^2\right]. \tag{25}$$

Proof: Let $p_\alpha(x) = \sum_{j=1}^m \alpha_j p_j(x)$. To show unbiasedness, write

$$E(\hat{I}_{\alpha,\delta}) = \int \frac{f(x) + \sum_{j=1}^m \delta_j p_j(x)}{p_\alpha(x)} p_\alpha(x)dx - \sum_{j=1}^m \delta_j = I,$$

establishing (24).

To prove (25) let $\delta_1^\star = 0$, and $\delta_j^\star = I\alpha_j/\alpha_1$ for $j > 1$. Write $f(x) = Ip_1(x) + r_1(x)$ where $\int r_1(x)dx = 0$. With these values we get

$$\begin{aligned}
V(\hat{I}_{\alpha,\delta^o}) &\le V(\hat{I}_{\alpha,\delta^\star}) \\
&= \frac{1}{n}\left[\int \left(\frac{Ip_\alpha(x)/\alpha_1 + r_1(x)}{p_\alpha(x)}\right)^2 p_\alpha(x)dx - \left(I + I\frac{1-\alpha_1}{\alpha_1}\right)^2\right] \\
&= \frac{1}{n}\int \frac{r_1(x)^2}{p_\alpha(x)}dx \\
&\le \frac{1}{n\alpha_1}\int \frac{r_1(x)^2}{p_1(x)}dx \\
&= \frac{1}{n\alpha_1}\left[\int \frac{f(x)^2}{p_1(x)}dx - I^2\right].
\end{aligned}$$

Repeating the argument for $j = 2, \ldots, m$, we find that the optimal values δ_j^o satisfy (25). $\square$

Theorem 1 is of the "no regret" variety. It shows that mixture importance sampling with control variates p_j/p_α is never worse than individual importance sampling without such control variates. Therefore taking the additional observations from p_k with $k \neq j$ did not make the estimate worse. Perhaps a fairer comparison would be to individual importance sampling with the control variate $1/p_j$ or even p_k/p_j. But in applications with p_j tending to zero over much of the space, adding a non-zero control variate to f would usually make the importance sampling variance (11) much larger, and so we think the comparison in Theorem 1 is appropriate.

To find sample based estimates of the optimal values take minimizers of

$$\sum_{i=1}^{n} \left(\frac{f(X_i) + \sum_{j=1}^{m} \hat{\delta}_{jn} p_j(X_i)}{p_\alpha(X_i)} - \hat{\mu}_n \right)^2,$$

and then use the estimate

$$\hat{I}_{\alpha,\hat{\delta}} = \hat{\mu}_n - \sum_{j=1}^{m} \hat{\delta}_{jn}. \tag{26}$$

Because $\sum_j \alpha_j p_j(x)/p_\alpha(x) = 1$ for all x, the regression above will be singular. Thus in practice, one can use all but one of the p_j/p_α or use a singular value decomposition (Golub & Van Loan 1983) to compute $\hat{\delta}_{jn}$.

Hesterberg (1988) proves some results similar to, and in some ways, sharper than Theorem 1. His Theorem 6.2 proves that the ratio estimate (see equation (8)) always leads to smaller asymptotic variance with n observations from p_α than it does with $n\alpha_j$ observations from p_j. Theorem 1 cannot be proved as a corollary to his Theorem 6.2 because the ratio estimator from p_j can be worse than ordinary importance sampling from p_j. His Theorem 6.3 shows that ratio estimation with a mixture density beats taking a linear combination of ratio estimates based on independent samples from the individual p_j.

Hesterberg (1988) also makes an effort to compare mixture to single component sampling for the regression estimator and for ordinary importance sampling. He finds that there is no analogue to his Theorem 6.2 for these. Then he introduces meta-weighted versions of the estimators for which the mixture based estimates are superior to the component based estimates. The

coefficients in that estimate, as given by equation (6.29) on page 157 of Hesterberg (1988), are in our notation, carefully weighted regressions of f/p_k on $1/p_k$. We prefer the estimate (26) because it estimates the δ_j jointly by a multiple regression.

4.1 Deterministic mixture sampling

Hesterberg (1995) recommends using deterministic mixtures instead of random mixtures. This means that one takes $n_j = n\alpha_j$ observations (or an integer close to $n\alpha_j$) from the density p_j. We also choose to employ control variates proportional to $p_j(x)/p_\alpha(x)$. The same estimate is used as if a random mixture had been taken, only now it may be written

$$\hat{I}_{\alpha,\delta} = \frac{1}{n}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{f(X_{ji}) + \sum_{k=1}^{m}\delta_k p_k(X_{ji})}{p_\alpha(X_{ji})}\right) - \sum_{j=1}^{m}\delta_j, \tag{27}$$

where X_{ji} are independent and drawn from $p_j(x)$. Equation (27) treats each observation the same, no matter which distribution it came from.

The estimate is still unbiased. The advantage of a deterministic mixture is that it has smaller variance than the random mixture. For large n, the number of observations coming from each mixture component will be close to its expected value. But the extra variance induced by random mixture proportions is of the same order of magnitude as the variance at the nominal mixture values, and can be the dominant source of variance in some applications.

5 Multiple Importance Sampling

Multiple importance sampling was introduced by Veach & Guibas (1995) for rendering problems in computer graphics as discussed in Section 1. Suppose that one has importance measures $p_1, \ldots, p_m$. In the graphics application these correspond to different path sampling methods.

One cannot always tell in advance which importance method to use. It is therefore advantageous to sample from several importance measures and then combine the results. The method of multiple importance sampling takes observations X_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, m$. Here the counts n_j are not random and the observations are sampled independently.

Let $w_j(x)$, $j = 1, \ldots, m$ be a partition of unity: for every $x \in (0,1)^d$, $0 \le w_j(x) \le \sum_{j=1}^m w_j(x) = 1$. Define

$$\hat{I}_{n,w} = \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} w_j(X_{ij}) \frac{f(X_{ij})}{p_j(X_{ij})} \tag{28}$$

where the subscripts on I denote the partition of unity and the sample sizes used. The estimate $\hat{I}_{n,w}$ is unbiased under mild conditions on the supports of the function p_j and w_j. Let $S(p_j) = \{x \mid p_j(x) > 0\}$, and $S(w_j) = \{x \mid w_j(x) > 0\}$. Assume that $S(w_j) \subset S(p_j)$ and that $\cup_{j=1}^m S(p_j) = (0,1)^d$. Then $1_{x \in S(p_j)} w_j(x) = w_j(x)$ and

$$\begin{aligned}
E(\hat{I}_{n,w}) &= \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \int_{S(p_j)} \frac{w_j(x) f(x)}{p_j(x)} p_j(x) dx \\
&= \sum_{j=1}^{m} \int_{S(p_j)} w_j(x) f(x) dx \\
&= \sum_{j=1}^{m} \int_{S(w_j)} w_j(x) f(x) dx \\
&= \int f(x) \sum_{j=1}^{m} 1_{x \in S(w_j)} w_j(x) dx \\
&= I.
\end{aligned}$$

Thus multiple importance sampling is unbiased under the mild conditions that the p_j cover the whole of $(0,1)^d$ and that w_j only weights the region that p_j samples.

Multiple importance sampling includes importance sampling, stratified sampling and stratified importance sampling as special cases. The latter two cases arise when $S(p_j) \cap S(p_k) = \emptyset$ for $j \ne k$.

Veach & Guibas (1995) consider several ways of selecting w_j. Their balance heuristic takes weights

$$\bar{w}_j(x) = \frac{n_j p_j(x)}{\sum_{k=1}^m n_k p_k(x)}. \tag{29}$$

The result is that

$$\hat{I}_{n,w} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{f(X_{ji})}{\sum_{k=1}^m n_k p_k(x)}$$

$$= \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{f(X_{ji})}{p_\alpha(X_{ji})}$$

where $n = \sum_j n_j$ and p_α is $\sum_j \alpha_j p_j$ with $\alpha_j = n_j/n$. Thus the balance heuristic matches the deterministic mixture sampling algorithm of Hesterberg (1995). It matches our equation (27) with all $\delta_j = 0$, that is, without control variates.

Veach & Guibas (1995) consider choosing functions w_j to make $\hat{I}_{n,w}$ perform nearly as well as importance sampling from an individual p_j, when one of the individual p_j's happens to be especially well suited to the f at hand. They use $w_j(x)$ in an attempt to make $p_j(x)$ get high weight in those parts of the space where f is nearly proportional to p_j.

They suggest two approaches, both of which increase $w_j(x)$ in cases where $p_j(x)$ is large compared to $p_k(x), k \neq j$. Their cutoff heuristic takes

$$w_j(x) \propto n_j p_j(x) 1_{n_j p_j(x) \geq \gamma \max_k n_k p_k(x)}, \tag{30}$$

for some $0 \leq \gamma \leq 1$, and their power heuristic takes

$$w_j(x) \propto (n_j p_j(x))^\beta, \tag{31}$$

for $0 \leq \beta$. In both heuristics the weights are normalized to sum to unity. Sending $\beta \to \infty$ in the power heuristic or taking $\gamma = 1$ in the cutoff heuristic gives rise to the maximum heuristic

$$w_j(x) \propto 1_{n_j p_j(x) = \max_k n_k p_k(x)}. \tag{32}$$

Unless there are ties among the $n_j p_j$, equation (32) puts all of the weight on one of the j's.

Veach & Guibas (1995) apply variational arguments to consider good choices for the functions w_j. They prove the two theorems below.

Theorem 2 (Veach & Guibas (1995)) *Let $\bar{w}_j$ be the weight functions from the balance heuristic (29) and let w_j be any other partition of unity. If $f(x) \geq 0$, then under deterministic mixture sampling*

$$V(\hat{I}_{n,\bar{w}}) \leq V(\hat{I}_{n,w}) + \left(\frac{1}{\min_j n_j} - \frac{1}{\sum_j n_j}\right) I^2. \tag{33}$$

Their second theorem uses random mixture sampling in which pairs (j_i, X_i) are chosen as follows: j_i takes the value k with probability $\alpha_k > 0$, $k = 1, \ldots, m$ and, given the value of j_i, X_i is sampled from p_{j_i}. The (j_i, X_i) pairs are independent of each other. The estimate is

$$\hat{I}_{w,\alpha} = \frac{1}{n} \sum_{i=1}^{n} w_{j_i}(X_i) \frac{f(X_i)}{\alpha_{j_i} p_{j_i}(X_i)}. \tag{34}$$

Notice that equation (34) is well defined even if none of the mixture samples come from p_1, whereas equation (28) is undefined if $n_1 = 0$.

Theorem 3 (Veach & Guibas (1995)) *Let $\bar{w}_j$ be the weight functions from the balance heuristic (29) and let w_j be any other partition of unity. Then under random mixture sampling*

$$V(\hat{I}_{n,\bar{w}}) \le V(\hat{I}_{n,w}). \tag{35}$$

Theorem 3 shows that the balance property is optimal under random mixture sampling. More generally, Hesterberg (1988) argues (page 135) from the Rao-Blackwell theorem, that in random mixture sampling, there is no advantage to taking account of which mixture component generated the sample.

From Hesterberg (1995) we know that the balance property can only improve further under deterministic mixture sampling. It may however, lose its optimality. Perhaps some other weighting heuristic undergoes an even bigger improvement. Theorem 2 give a bound on how far from optimal the balance heuristic weights could become under deterministic importance sampling. Note that Veach & Guibas (1995) don't include the condition $f(x) \ge 0$ in the statement of their theorem, but their proof seems to use it. (Such a condition is very reasonable in ray tracing applications.)

We can use Theorems 2 and 3 to bound the inefficiency that results when one of the p_j is an especially good importance sampling choice, but we don't use it. Consider the j-heuristic which takes $w_k(x) = 1$ if $k = j$ and $w_k(x) = 0$ otherwise. This rule simply ignores all p_k except p_j. It is only a valid rule if $p_j(x) > 0$ everywhere, but we already need that condition if we're considering using p_j for ordinary importance sampling. With this heuristic we get

$$\hat{I}_{n,w=j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{f(X_{ji})}{p_j(X_{ji})} \tag{36}$$

for deterministic mixtures, which is equivalent to taking an importance sample of size n_j from p_j.

Theorem 2 gives

$$V(\hat{I}_{n,\bar{w}}) \le V(\hat{I}_{n,w=j}) + \frac{1}{n}\left(\frac{1}{\min_k \alpha_k} - 1\right) I^2, \tag{37}$$

for nonnegative f under deterministic mixture sampling. By contrast, equation (25) of Theorem 1 gives

$$V(\hat{I}_{\alpha,\delta_o}) \le V(\hat{I}_{n,w=j}), \tag{38}$$

under random mixture sampling, without requiring $f(x) \ge 0$, but requiring an estimate of the δ_j. Equation (38) gives a better guarantee than does equation (37). The guarantee comes without an assumption that $f(x) \ge 0$ and we know that the result can improve further when deterministic sampling is used.

Theorem 3 gives a bound that looks like (38) under random mixture sampling. But the estimator $\hat{I}_{n,w}$ behaves quite differently under random mixture sampling. For example, suppose that $f(x) = Ip_j(x)$. Deterministic mixture sampling with the j-heuristic gives a perfect answer. Random mixture sampling with the j-heuristic can't match the perfect answer. Under the j-heuristic with random mixtures

$$\hat{I}_{\alpha,w=j} = \frac{n_j I}{n\alpha_j}. \tag{39}$$

This has mean I but because n_j has a binomial distribution, we find

$$V(\hat{I}_{\alpha,w=j}) = \left(\frac{I}{n\alpha_j}\right)^2 n\alpha_j(1-\alpha_j) = \frac{I^2(1-\alpha_j)}{n\alpha_j}. \tag{40}$$

Thus the bound in Theorem 3 is not sharp enough to allow the balance heuristic to inherit the excellent performance of p_j, for any $\alpha_j < 1$.

6 Positivisation

Almost every course or text that discusses variance reduction methods mentions that the importance sampling variance can vanish for an integrand

$f \geq 0$, if $p = f/I$. And clearly if f is nonpositive, the optimal density becomes $-f(x)/I$, which also achieves zero variance. In general, for f of mixed sign it is true that the optimal importance sampling density is proportional to $|f(x)|$ (Kahn & Marshall 1953), although the resulting optimal variance is no longer zero.

It is less widely known that (multiple) importance sampling, can produce a zero variance for integrands taking both positive and negative values. Here we show that importance sampling can achieve zero variance for integrands of mixed sign at the small extra cost of having to sample from two different importance sampling densities. Then we show that the same can be accomplished by mixture sampling with control variates as in equation (27), for appropriate p_j.

6.1 Simple positivisation

For f having mixed sign, we write $f(x) = f_+(x) - f_-(x)$ where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. Then $I = \int f(x)dx = \int f_+(x)dx - \int f_-(x)dx$. By taking a sample of size n_+ from $p_+ \propto f_+$ and a sample of size n_- from $p_- \propto f_-$ it is possible to attain a zero variance estimate:

$$\hat{I}_\pm = \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{f_+(X_{i,+})}{p_+(X_{i,+})} - \frac{1}{n_-}\sum_{i=1}^{n_-}\frac{f_-(X_{i,-})}{p_-(X_{i,-})}. \tag{41}$$

Thus, even for integrands of mixed sign it is possible to obtain an importance sampling variance that is arbitrarily small. There is a practical difficulty in finding good approximations $p_\pm$, but there is no difficulty in evaluating $f_\pm$ at a given value of x.

The decomposition above "splits the integrand at 0". We could as easily write $f(x) = c + (f(x) - c)_+ - (f(x) - c)_-$ for some well chosen constant c. In a personal communication, Ben Fox mentioned taking some value $c < \inf_x f(x)$ so that $(f(x) - c)_- = 0$ and then applying importance sampling to $(f(x) - c)_+ = f(x) - c$.

More generally we have $f(x) = h(x) + (f(x) - h(x))_+ - (f(x) - h(x))_-$ where $h(x)$ is a function for which $\int h(x) = I_h$ is known. Then the estimate

$$\hat{I}_{h\pm} = I_h + \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{(f-h)_+(X_{i,+})}{p_+(X_{i,+})} - \frac{1}{n_-}\sum_{i=1}^{n_-}\frac{(f-h)_-(X_{i,-})}{p_-(X_{i,-})} \tag{42}$$

involves a known integral I_h and two estimated integrals of nonnegative functions.

Assuming that the samples $X_{i,+}$ are independent of the $X_{i,-}$, the variance of the estimate in (42) is

$$V(\hat{I}_{h\pm}) = \frac{1}{n_+}\left(\int \frac{(f-h)_+^2}{p_+}dx - I_{h+}^2\right) + \frac{1}{n_-}\left(\int \frac{(f-h)_-^2}{p_-}dx - I_{h-}^2\right), \quad (43)$$

where

$$I_{h+} = \int (f-h)_+ dx, \quad \text{and} \quad I_{h-} = \int (f-h)_- dx.$$

The function h acts as a control variate, though it is used in a nonstandard way. Intuitively one expects that the smaller $f-h$ is, the better the method will be. This may be true in practice, although it is theoretically possible that an increase in the size of $(f-h)_\pm$ could be more than compensated for by a more accurate approximation of $p_\pm$ to $(f-h)_\pm$.

6.2 Smooth positivisation

A function such as $(f(x)-h(x))_+$ may not be smooth even if $f(x)$ and $h(x)$ are both smooth. For some methods, such as quasi-Monte Carlo (Niederreiter 1992) this lack of smoothness can limit the accuracy of numerical integration. It is thus of interest to positivise f without losing smoothness. Define a partition of the identity by a set of smooth functions v_j, $j = 1, \ldots, p$ satisfying

$$z = \sum_{j=1}^{r} v_j(z), \quad -\infty < z < \infty. \quad (44)$$

If also, for each j we either have $v_j(z) \geq 0$ for all z, or $v_j(z) \leq 0$ for all z, we call these functions a semi-definite partition of the identity function.

Our estimate now becomes

$$\hat{I}_{h,v} = I_h + \sum_{j=1}^{r} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{v_j\left((f-h)(X_{ij})\right)}{p_j(X_{ij})} \quad (45)$$

where the p_j are densities chosen to be approximately proportional to $|v_j|$, and the random vectors X_{ij} are drawn from the density p_j and all of the X_{ij} are independent. Now

$$E(\hat{I}_{h,v}) = I_h + \sum_{j=1}^{r} \int v_j((f-h)(x))dx = I_h + \int (f-h)dx = I \quad (46)$$

and so the estimate is unbiased. Furthermore

$$V(\hat{I}_{h,v}) = \sum_{j=1}^{r} \frac{1}{n_j} \left(\int \frac{v_j^2 \left((f-h)(X_{ij}) \right)}{p_j(X_{ij})} - I_{h,v_j}^2 \right) \tag{47}$$

where

$$I_{h,v_j} = \int v_j((f-h)(x))dx.$$

In Section 6.1 we have $r = 2$ and nonsmooth functions $v_1(z) = z_+$ and $v_2(z) = -z_-$. We may replace these functions by smooth ones such as

$$v_1(z) = \frac{z}{2} + \sqrt{\eta + \left(\frac{z}{2}\right)^2} \geq 0$$

and

$$v_2(z) = \frac{z}{2} - \sqrt{\eta + \left(\frac{z}{2}\right)^2} \leq 0,$$

where $\eta > 0$. Each function, v_1 and v_2, represents a branch of an hyperbola. They can be differentiated any number of times and so are smooth. Notice that $\lim_{\eta \to 0} v_1(z) = z_+$ and $\lim_{\eta \to 0} v_2(z) = -z_-$.

6.3 Mixture sampling and positivisation

Defensive mixture sampling can be incorporated into the positivized estimates (42) and (45). Each of the r terms in (45) or each of the two terms in (42) can be estimated by defensive mixtures and control variates as in equation (27). Suppose that for term j one uses m_j mixture components. Then the estimate is

$$\begin{aligned} \hat{I} \;=\; & \sum_{j=1}^{r} \frac{1}{n_j} \sum_{k=1}^{m_j} \sum_{i=1}^{n_{jk}} \frac{v_j\Big((f-h)(X_{jki})\Big) + \sum_{k=1}^{m} \delta_{jk} p_{jk}(X_{jki})}{p_{\alpha_j}(X_{jki})} \\ & + I_h - \sum_{j=1}^{r} \sum_{k=1}^{m_j} \delta_{jk} \end{aligned} \tag{48}$$

where X_{jki} are independent draws from p_{jk} for $j = 1, \ldots, r$, $k = 1, \ldots, m_r$, $i = 1, \ldots, n_{jk}$, $p_{\alpha_j}(x)$ is the mixture $\sum_{k=1}^{m_j} \alpha_{jk} p_{jk}(x)$, and $n_{jk} = \alpha_{jk} n_j$.

Equation (48) simplifies considerably if one uses the same defensive mixture sampling on all r of the $v_j(f-h)$. Suppose we take all $n_j = n$, a

common choice $m_j = m$, common mixture components $\alpha_{jk} = \alpha_k > 0$ and the same densities $p_{jk} = p_k$. Let p_α denote the common mixture density. Finally suppose that the $n = n_1$ observations drawn for v_1 are retained and used for $j = 2, \ldots, r$. Then equation (48) reduces to

$$\hat{I} = \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k}\frac{(f-h)(X_{ki}) + \sum_{k=1}^{m}\delta_k p_k(X_{ki})}{p_\alpha(X_{ki})} + I_h - \sum_{k=1}^{m}\delta_k. \qquad (49)$$

The $v_j(f-h)$ have recombined to $f-h$ and the δ_{jk} only enter through their sums $\sum_{j=1}^{r}\delta_{jk}$ which are denoted by δ_k.

We suggest the following strategy in applications where subject matter knowledge allows it. First find a suitable proxy function h that is close to f and has known integral. Then find densities p_1 and p_2 that are nearly proportional to $(f-h)_+$ and $(f-h)_-$ respectively. Then take a third density p_3 as the uniform or some other defensive density. The density p_3 should overlap enough with p_1 and p_2 to allow effective estimation of the control variate coefficients δ_j. This procedure is illustrated by Example 3 in Section 7. Where $(f-h)_\pm$ has a small number of modes with known locations, a mixture density can be constructed for each mode. Where the positivisation required has r terms, as in equation (44), one or component densities can be designed for each term.

7 Examples

We give three examples to compare the methods described in this paper. The first two examples illustrate defensive mixture sampling and multiple importance sampling of Sections 3 through 5. The third example illustrates the positivisation ideas of Section 6. The mixture sampler of Section 4 using control variates p_j/p_α provides consistently good results.

The first example has an importance sampling density with a mode very closely matched to the integrand, but with infinite variance nonetheless. It is designed to show how defensive mixtures can protect against bad results. The second example has a very effective importance sampling density for which the original non-defensive importance sampling greatly outperforms defensive mixture sampling without control variates. In both examples the mixture sampler of Section 4 using control variates p_j/p_α performs nearly best. The third example illustrates a case, motivated by examples in finance,

where one might realistically expect to have a control variate h with a good idea of where $f > h$ and where $f < h$.

We compare 11 methods in the first two examples. Here is a brief description of each method:

IID: Let $X_1, \ldots, X_n$ be iid samples from $U(0,1)^5$. I is estimated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^{n} f(X_j).$$

IS: Let $X_1, \ldots, X_n$ be iid samples from $p(x)$. I is estimated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^{n} \frac{f(X_j)}{p(X_j)}.$$

DIS: Let $X_1, \ldots, X_n$ be iid samples from $p_\alpha(x) = \alpha + (1-\alpha)p(x)$. I is estimated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^{n} \frac{f(X_j)}{p_\alpha(X_j)}.$$

MISCV: Let $X_1, \ldots, X_n$ be iid samples from $p_\alpha(x)$. I is estimated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^{n} \frac{f(X_j) - \hat{\delta}_0 - \hat{\delta}_1 p(X_j)}{p_\alpha(X_j)} + (\hat{\delta}_0 + \hat{\delta}_1),$$

where $\hat{\delta}$'s are the estimated coefficients of linear regression of f/p_α on $1/p_\alpha$ and p/p_α, using the data $X_1, \ldots, X_n$.

BALANCE: Let $n_0 \propto \alpha$, and $n_1 \propto (1-\alpha)$. Sample $X_{10}, \ldots, X_{n_0 0}$ iid from $U(0,1)^5$, and $X_{11}, \ldots, X_{n_1,1}$ from $p(x)$.

$$\hat{I} = \frac{1}{n} \sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{f(X_j)}{p_\alpha(X_j)}.$$

CUTOFF: The cutoff point γ as in (30) is chosen to be 0.1, and $n_j = n/2$.

POWER: The power β as in (31) is chosen to be 2.0, and $n_j = n/2$.

MAXIMUM: Using Formula (32) to determine the weight function, with $n_j = n/2$.

Of the 8 methods above, DIS, MISCV and BALANCE are investigated at $\alpha = 0.1$ and at $\alpha = 0.5$, bringing the total to 11 methods.

For all examples, we report estimated asymptotic variances. They are computed by running 20 independent runs. In each run, $n = 10^5$ sample points are generated and an estimate $\hat{I}_i$ is computed. The estimated asymptotic variance is

$$n\hat{V} = \frac{n}{20}\sum_{i=1}^{20}(\hat{I}_i - I)^2. \tag{50}$$

The estimate (50) of asymptotic variance requires that we know the true integrand I. We know this in the examples we consider. We consider some cases in which the true asymptotic variance is infinite. Equation (50) gives large but finite answers there. The bias in an estimate will inflate Equation (50) by an asymptotically (in n) small factor.

7.1 Example 1

We construct this example using the beta density function

$$B(x, a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)/\Gamma(a+b)}, \quad 0 < x < 1$$

where $a > 0$ and $b > 0$ are parameters and $\Gamma(z)$ is the gamma function. The integrand is a mixture of products of beta density functions, defined as:

$$f(x) = 0.9 \times \prod_{j=1}^{5} B(x^j, 20, 20) + 0.1 \times \prod_{j=1}^{5} B(x^j, 2, 2), \tag{51}$$

and $p(x)$ is chosen to be:

$$p(x) = \prod_{j=1}^{5} B(x^j, 20, 20). \tag{52}$$

Here superscripts j denote the component of the five dimensional vector x. Clearly $I = 1$ because f is a density.

		Example 1		Example 2	
		Mean	$n\hat{V}$	Mean	$n\hat{V}$
IID		1.034	$8.58 \times 10^{+2}$	1.038	$1.06 \times 10^{+3}$
IS		0.934	$4.51 \times 10^{+2}$	1.000	5.23×10^{-8}
α	DIS	1.000	6.56×10^{-2}	1.001	1.07×10^{-1}
$\|$	MISCV	1.000	7.17×10^{-2}	1.000	7.08×10^{-8}
0.1	BALANCE	0.999	1.36×10^{-1}	1.000	4.14×10^{-3}
α	DIS	1.000	3.81×10^{-1}	1.000	1.07×10^{-1}
$\|$	MISCV	1.000	2.77×10^{-2}	1.000	9.50×10^{-8}
0.5	BALANCE	1.000	5.63×10^{-2}	1.000	2.77×10^{-2}
	CUTOFF	1.000	4.78×10^{-2}	1.000	2.08×10^{-2}
	POWER	1.000	3.61×10^{-2}	1.000	1.24×10^{-2}
	MAXIMUM	1.000	3.66×10^{-2}	1.000	1.16×10^{-2}

Table 1: Estimated integral and asymptotic variance of Examples 1 and 2 for methods described in the text. The true integral is 1. For each method 20 independent simulations with $n = 10^5$ observations each were run. The column headed Mean gives the average of those 20 estimates. The column headed $n\hat{V}$ gives the estimated asymptotic variance of the methods.

The integrand $f(x)$ has a very large spike at the center of the unit cube. The density $p(x)$ is a good match to $f(x)$ around this peak, but it has a lighter tail than the integrand. Therefore, importance sampling using $p(x)$ as sampling distribution is expected to give a large variance. In fact, the variance (equation (4)) for this p and f is infinite. The second mixture component of $f(x)$ is flatter than the first, but still spiky enough so that the uniform distribution used in defensive importance sampling does not lead to extreme accuracy.

The results for this example are given in Table 1. As expected the IID and IS methods do very badly on this integrand. DIS with either value of α does well. It does especially well with $\alpha = 0.1$. In hindsight this appears to be a lucky match between the mixture proportions used in DIS and the ones used in defining $f(x)$. For the balance heuristic, the choice $\alpha = 0.5$ outperforms $\alpha = 0.1$. The three other heuristics (power, cutoff and maximum) using $\alpha = 0.5$ do better still, with power being the best and maximum a near second. The best method on this example is MISCV with $\alpha = 0.5$.

We note that although the asymptotic variance of importance sampling

in this example is infinite, the estimated value is still finite, though large. The method has a large variance because f/p can become very large away from the center of the cube. It doesn't estimate this variance so well because it rarely samples in that region. This is not a case of "what you don't know can't hurt you". In order to get the mean about right the region where f/p is large must be sampled. The sample mean is quite far from the true value 1.0. A large systematic error in the mean is what we would expect from a biased method, and yet IS is unbiased. The reconciliation of these two facts lies in the infinite skewness of $\hat{I}$. In this example, for the vast majority of simulations, importance sampling produces an estimate that is slightly smaller than I. In rare settings it would be very much larger than I.

For the IID method, analytical calculations show a finite asymptotic variance of 447.

7.2 Example 2

The integrand is a tensor product of the univariate beta density function, with a mean zero perturbation of magnitude 0.1 added to a neighborhood of its maximum:

$$f(x) = \prod_{j=1}^{5} B(x^j, 20, 20) + 0.1 \prod_{j=1}^{5} \sin\left(60\pi(x^j - 1/3)\right) 1_{1/3 \le x^j \le 2/3}. \tag{53}$$

The true integral for this f is $I = 1$. The importance function $p(x)$ is

$$p(x) = \prod_{j=1}^{5} B(x_j, 20, 20). \tag{54}$$

As in Example 1, this importance density provides a reasonable match near the peak, but unlike Example 1 it also provides a good match away from the peak.

From Table 1 it is clear that ordinary importance sampling gives excellent performance on this problem. Of all the defensive methods considered, only multiple importance sampling with the control variates, comes close to matching this excellent performance.

Importance sampling is usually motivated by a desire to sample a spiky function and for that it is considered important to match the integrand well near the spike. The irony of Table 1 is that, for these two examples, the

performance of importance sampling is dominated by how well it matches the integrand away from the spikes.

Examples 1 and 2 illustrate the point made in Section 4 that multiple importance sampling with control variates protects against the worst outcomes of importance sampling with an acceptably small loss of efficiency in cases where importance sampling does extraordinarily well. None of the other methods considered achieved this in these two examples.

7.3 Example 3

This example is modeled after some integrands in computational finance. Suppose that $h(x)$ is a control variate whose integral I_h is known to us, and that the integrand $f(x)$ is equal to $h(x)$ as long as $h(x)$ is in the range $[A, B]$. If $h(x)$ falls below A then $f(x) = A$ and if $h(x)$ goes over B then $f(x) = B$. The finance connection is as follows: $h(x)$ may describe a financial instrument whose value is easily found in closed form or by very fast computations, but traders might want the expected value of a version of $h(x)$ subject to a ceiling of B and a floor of A.

We take

$$h(x) = 100\Big(\sum_{j=1}^{5} x^j - 1\Big), \quad \text{and,} \quad f(x) = \max\Big(\min(h(x), 300), -25\Big), \quad (55)$$

so $f(x)$ is $h(x)$ clipped to be in the range $[-25, 300]$. Our f and h are simpler than realistic finance integrands, but the clipping process is realistic. We easily find $I_h = 150$, and then using a result on the volume of a simplex we compute $I_f = 150 - 100/6! + (0.75)^5 * 75/6! = 149.8858$.

The difference $f - h$ is zero over most of the cube. It is positive near the origin $(0, 0, 0, 0, 0)$ and negative near the opposite corner $(1, 1, 1, 1, 1)$. The probability that $f \neq h$ is $(1 + .75^5)/5! \doteq 0.01$. It would be possible to achieve an importance sampling variance of zero using the simple positivisation technique in Section 6.1. We don't feel that it is realistic to be able to find the exact densities in practice that would do this. But it may be realistic in practice to know qualitatively where the integrand is likely to be positive or negative and to construct densities that over sample those regions.

Let $p_\alpha(x) = \sum_j \alpha_j p_j(x)$ be a mixture of 3 densities $p_j(x)$, for $j = 0, 1, 2$. First, $p_0(x)$ is the uniform density on $(0, 1)^5$ serving as the defensive density. Under $p_1(x)$, each x^j is independently drawn as a $N(0, \sigma_1^2)$ random variable

conditioned to be in $(0,1)$, and under $p_2(x)$, each x^j is independently drawn as a $N(1,\sigma_2^2)$ random variable conditioned to be in $(0,1)$. The constants σ_1 and σ_2 are chosen below to give rough matches to the integrand.

To find values for σ_1 and σ_2 we tried to get rough coverage of the regions in which $f \neq h$. The distance from the point $(1,1,1,1,1)$ to the plane at which h is truncated by 300 is $5^{-1/2}$. Taking $2\sigma_2 = 5^{-1/2}$ gets some points into the range where $f = h$. That is we take $\sigma_2 = 0.5 * 5^{-1/2} \doteq 0.2236$. Similarly we take $\sigma_1 = 0.75 * \sigma_2 \doteq 0.1677$. Notice that σ_2 gives a negligible probability of reaching the point $(1,1,1,1,\epsilon)$ for small $\epsilon > 0$, and so $(f-h)_-/p_1$ can take very large values. Thus we expect the naive estimate (41) to be unstable.

We use a simple non-optimal method to choose p_1, and p_2 because in practice one is likely to have to use simple and non-optimal methods. We choose mixing proportions $\alpha_0 = 0.1$, and $\alpha_1 = \alpha_2 = 0.45$. We considered the methods described below.

IID: This is the same as described at the beginning of Section 7.

CCV: This is the classical control variate method, using $h(x)$ as the control function. The estimate can be written as:

$$\hat{I} \;=\; \hat{\beta} I_h + \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \hat{\beta} h(X_i),$$

where $X_1, \ldots, X_n$ are iid from $p_0(x)$. The coefficient $\hat{\beta}$ is estimated by regressing f on h using X_i's.

PM: This is the simple positivisation method (41). We split $f(x) - h(x)$ into positive and negative parts $(f-h)_\pm$, and then apply ordinary importance sampling separately to the two parts. For $(f-h)_+$, we use $p_+ = p_1(x)$ as the importance function. Similarly, for $(f-h)_-$, we use $p_- = p_2(x)$ as the importance function. The estimator can be written as:

$$\begin{aligned}\hat{I} \;=\; & I_h + \frac{1}{n_+}\sum_{i=1}^{n_+} \frac{(f(X_{i+}) - h(X_{i+}))_+}{p_+(X_{i+})} \\ & - \frac{1}{n_-}\sum_{i=1}^{n_-} \frac{(f(X_{i-}) - h(X_{i-}))}{p_-(X_{i-})}\end{aligned}$$

where we choose $n_+ = n_- = n/2$. The coefficients $\hat{\delta}_{j+}$'s are found by multiple regression of f/p_+ on p_0/p_+ and p_1/p_+ using X_{i+}, and the

$\hat{\delta}_{j-}$'s are found by multiple regression of f/p_- on p_0/p_- and p_1/p_- using X_{i-}.

MISCV-R: Apply multiple importance sampling, with control variates p_j/p_α. The estimator is:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)-h(X_i)+\sum_{j=0}^{m}\hat{\delta}_j p_j(X_i)}{p_\alpha(X_i)}+\left(I_h-\sum_{j=0}^{m}\hat{\delta}_j\right),$$

where $m=2$, $X_1,\ldots,X_n$ are iid from p_α. The $\hat{\delta}_j$'s are estimated by multiple linear regression of $(f-h)/p_\alpha$ on p_j/p_α, for $j=0,1,2$.

MISCV-Rh: Apply multiple importance sampling, with control variates p_j/p_α and h/p_α. The estimator is:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)-\hat{\beta}h(X_i)+\sum_{j=0}^{m}\hat{\delta}_j p_j(X_i)}{p_\alpha(X_i)}+\left(\hat{\beta}I_h-\sum_{j=1}^{m}\hat{\delta}_j\right),$$

where $m=2$ and $\hat{\beta},\hat{\delta}_1,\ldots,\hat{\delta}_m$ are estimated by multiple linear regression of f/p_α on h/p_α and p_j/p_α. This is MISCV-R with an additional estimated coefficient for $h(x)$.

MISCV-D: Apply multiple importance sampling, with control variates p_j/p_α and a deterministic mixture sample. The estimator is:

$$\hat{I}_{\alpha,\delta} = \frac{1}{n}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{f(X_{ji})-h(X_{ji})+\sum_{k=1}^{m}\hat{\delta}_k p_k(X_{ji})}{p_\alpha(X_{ji})}\right)+I_h-\sum_{j=1}^{m}\delta_j,$$

where $m=3$, $X_{j1},\ldots,X_{jn_j}$ are iid from p_j. The $\hat{\delta}_j$'s are estimated by multiple linear regression of $(f-h)/p_\alpha$ on p_j/p_α, for $j=0,1,2$.

MISCV-Dh: Apply multiple importance sampling, with control variates p_j/p_α and h/p_α and a deterministic mixture sample. The estimator is:

$$\hat{I}_{\alpha,\delta} = \frac{1}{n}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\frac{f(X_{ji})-\hat{\beta}h(X_{ji})+\sum_{k=1}^{m}\hat{\delta}_k p_k(X_{ji})}{p_\alpha(X_{ji})}\right)+\hat{\beta}I_h-\sum_{j=1}^{m}\delta_j,$$

where $m=3$ and $\hat{\beta},\hat{\delta}_1,\ldots,\hat{\delta}_m$ are estimated by multiple linear regression of f/p_α on h/p_α and p_j/p_α. This is MISCV-D with an additional estimated coefficient for $h(x)$.

	IID	**PM**	**CCV**	**PMDM**
$-$Err	-1.4×10^{-1}	5.4×10^{-2}	6.7×10^{-4}	8.4×10^{-5}
$n\hat{V}$	1.9×10^{4}	3.9×10^{3}	4.2×10^{0}	5.2×10^{-2}
	MIPCV-R	**MIPCV-Rh**	**MIPCV-D**	**MIPCV-Dh**
$-$Err	3.4×10^{-4}	6.7×10^{-4}	9.8×10^{-5}	9.6×10^{-5}
$n\hat{V}$	7.5×10^{-2}	6.8×10^{-2}	3.9×10^{-2}	3.8×10^{-2}

Table 2: The average error and asymptotic variance is given for the integrand of Example 3 for methods described in the text. Each method was run 20 times and each run had a sample size of $n = 10^5$. Err $= 1/20 \sum (I_i - I)$, where I_i is the estimated integral from the ith run, and $I = 149.886$ is the true integral. Most Err's were negative, so $-$Err is reported. The asymptotic variance is labeled $n\hat{V}$

PMDM: This is the positivisation method (PM) with separate defensive mixtures. We split $f(x)-h(x)$ into positive and negative parts $(f-h)_{\pm}$, and then apply multiple importance sampling with control variates separately to the two parts. For $(f-h)_+$, we use $p_+ = 0.1 \times p_0(x) + 0.9 \times p_1(x)$ as the importance function, and use p_0/p_+ and p_1/p_+ as control variates. Similarly, for $(f-h)_-$, we use $p_- = 0.1 \times p_0(x) + 0.9 \times p_2(x)$ as the importance function, and use $p_0(x)/p_-(x)$ and $p_2(x)/p_-(x)$ as control variates. The estimator can be written as:

$$\begin{aligned} \hat{I} &= I_h + \frac{1}{n_+} \sum_{i=1}^{n_+} \frac{(f(X_{i+}) - h(X_{i+}))_+ - (\hat{\delta}_{0+} + \hat{\delta}_{1+} p_1(X_{i+}))}{p_+(X_{i+})} \\ &\quad - \frac{1}{n_-} \sum_{i=1}^{n_-} \frac{(f(X_{i-}) - h(X_{i-}))_- - (\hat{\delta}_{0-} + \hat{\delta}_{2-} p_2(X_{i-}))}{p_-(X_{i-})} \\ &\quad + \delta_{0+} + \delta_{1+} - \delta_{0-} - \delta_{1-} \end{aligned}$$

where we choose $n_+ = n_- = n/2$. The coefficients $\hat{\delta}_{j+}$'s are found by multiple regression of f/p_+ on p_0/p_+ and p_1/p_+ using X_{i+}, and the $\hat{\delta_{j-}}$ are found by multiple regression of f/p_- on p_0/p_- and p_1/p_- using X_{i-}.

For each method, we conduct 20 independent runs, each with sample size $n = 10^5$. The results are reported in Table 2. None of the methods appeared

to have appreciable bias: each Err is small compared to the square root of $n\hat{V}$. Thus we can focus our comparisons on $n\hat{V}$.

The naive positivisation method (PM) does not provide for much of an improvement over IID sampling. This is reasonable because the sampling densities used are only qualitatively similar to the optimal ones. By contrast, simply using h as a classical control variate (CCV) makes a large improvement on IID sampling. By incorporating defensive mixtures into PM, the PMDM method is not so sensitive to whether the sampling densities are close to optimal. The result is a large improvement, and much better performance than CCV.

The four MIPCV methods all perform very well, ranging from slightly better than PMDM to slightly worse, but always much better than CCV. Estimating a coefficient for h does not make much difference, but using a deterministic instead of a random mixture cuts the variance by about half. We also found that PMDM can have worse numerical condition than the MISCV methods.

Acknowledgements

This work was supported by the National Science Foundation. We thank Eric Veach for explaining some graphics concepts. Any errors in the description of graphics problems here are our own.

References

Bratley, P., Fox, B. J. & Schrage, L. E. (1987), *A Guide to Simulation (Second Edition)*, Springer-Verlag.

Cochran, W. G. (1977), *Sampling Techniques (3rd Ed)*, John Wiley & Sons.

Golub, G. H. & Van Loan, C. F. (1983), *Matrix Computations*, Johns Hopkins University Press.

Hesterberg, T. (1988), Advances in Importance Sampling, PhD thesis, Stanford University.

Hesterberg, T. (1995), 'Weighted average importance sampling and defensive mixture distributions', *Technometrics* **37**(2), 185–194.

Kahn, H. & Marshall, A. (1953), 'Methods of reducing sample size in Monte Carlo computations', *Journal of the Operations Research Society of America* **1**, 263–278.

Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, S.I.A.M., Philadelphia, PA.

Owen, A. B. & Zhou, Y. (1998), Adaptive importance sampling by mixtures of beta distributions, Technical report, Stanford University, Statistics Department.

Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*, John Wiley & Sons.

Trotter, H. F. & Tukey, J. W. (1956), Conditional Monte Carlo for normal samples, *in* H. A. Meyer, ed., 'Symposium on Monte Carlo methods', John Wiley and Sons, New York.

Veach, E. (1997), Robust Monte Carlo Methods for Light Transport Simulation, PhD thesis, Stanford University.

Veach, E. & Guibas, L. (1995), Optimally combining sampling techniques for Monte Carlo rendering, *in* 'SIGGRAPH '95 Conference Proceedings', Addison-Wesley, pp. 419–428.