Optimal mixture weights in multiple importance sampling

Hera Y. He
Stanford University

Art B. Owen
Stanford University

November 2014

Abstract

In multiple importance sampling we combine samples from a finite list of proposal distributions. When those proposal distributions are used to create control variates, it is possible (Owen and Zhou, 2000) to bound the ratio of the resulting variance to that of the unknown best proposal distribution in our list. The minimax regret arises by taking a uniform mixture of proposals, but that is conservative when there are many components. In this paper we optimize the mixture component sampling rates to gain further efficiency. We show that the sampling variance of mixture importance sampling with control variates is jointly convex in the mixture probabilities and control variate regression coefficients. We also give a sequential importance sampling algorithm to estimate the optimal mixture from the sample data.

1 Introduction

Importance sampling is an essential Monte Carlo method, especially when the integrand has a singularity or when we need to work with rare events. More generally, when the integrand is concentrated in a region of small probability it pays to take more than the nominal number of points from that important region. For an expectation with respect to a nominal distribution p, we sample from an importance distribution q and counter the resulting bias by multiplying our sample values by p/q.

It can be very difficult to choose q wisely. A common strategy is to sample from multiple distributions q_j, for $j = 1, \dots, J$ corresponding to different extreme scenarios. The simplest approach is to sample from a mixture distribution that samples q_j with probability α_j. An improvement on that is to sample exactly $n_j = n\alpha_j$ observations from each q_j. We call both of these methods mixture importance sampling, with the latter being stratified mixture importance sampling. Multiple importance sampling (Veach and Guibas, 1995) is a very large family of estimators generalizing stratified mixture importance sampling.

In bidirectional path sampling for graphical rendering, the scenarios underlying the q_j might be descriptions of paths that light follows between source and viewing plane (Lafortune and Willems, 1993; Veach and Guibas, 1994). Multiple importance sampling is so important for the movie industry, that it was mentioned among Eric Veach's contributions for his technical Oscar in 2014. Similar strategies are used to sample failure modes for power systems or financial positions. In these applications, the scenarios can be all of the specific failure mechanisms of which we are aware.

Owen and Zhou (2000) combine mixture and multiple importance sampling with certain control variates derived from the sampling distributions. The combination enjoys a regret bound; essentially you cannot do worse that way than had you used only the sample values from the single best component. The identity of the best component might be unknown, or it might differ for multiple integration problems estimated on the same sample. We do not expect a better lower bound in general, because the single best importance sampler might be the only one of our J choices that has a finite variance.

Letting $n = \sum_{j=1}^{J} n_j$, the sampling fraction for q_j is $\alpha_j = n_j/n$. A standard choice is $\alpha_j = 1/J$, in which all components are sampled equally. Veach (1997) recommends this allocation, but the examples in that thesis have only moderately large J. If J is very large and only one or a small number of the mixture components is very useful, then it is inefficient to sample them equally. In this paper we show how to optimize the choice of the mixture weights α_j. We show that the variance in mixture importance sampling is a convex function of these weights. Furthermore, when control variates are used, the variance is jointly convex in the control variate parameters and the mixture weights, via a quadratic-over-linear construction (Boyd and Vandeberghe, 2004). Given an estimate $\hat{\alpha}$ of the optimal α from a preliminary sample, one can sample from the $\hat{\alpha}$ mixture to get a final estimate.

Section 2 reviews basics of importance sampling, and control variates to set our notation. It also mentions their combination in a single problem, which is less well known. Section 3 contains our main theoretical results. We prove regret bounds for two versions of mixture importance sampling with control variates. In the presence of a defensive mixture component, those methods define equivalent families of estimators. The new version has a variance which is a convex function of the mixing parameter. The variance is also jointly convex in the mixing parameter and the control variate regression parameter. Stratified mixture importance sampling reduces to the most popular multiple importance sampling method, known as the balance heuristic. Incorporating control variates into it makes the variance even smaller. Our preferred method is stratified mixture sampling with control variates because it attains the regret bounds for both the balance heuristic and stratified importance sampling. Section 4 discusses sampling based ways to choose mixture probabilities using the convexity results from Section 3. Section 5 has two numerical examples, one from a singular integrand, and one from a rare event. Section 6 has our conclusions and some discussion of the literature.

2 Importance sampling and control variates

The problem we consider is approximating

$$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

for an integrand f and probability density function p. We use notation for a continuously distributed random variable $\boldsymbol{x}$ on Euclidean space but the results generalize to other settings, such as discrete random variables, with only minor changes.

The standard Monte Carlo method for estimating μ takes a sample $\boldsymbol{x}_i \overset{\text{iid}}{\sim} p$ and produces the estimate $\hat{\mu} = (1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$. When $\sigma^2 = \int (f(\boldsymbol{x})-\mu)^2\,\mathrm{d}\boldsymbol{x} < \infty$ then $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$.

For f corresponding to rare events or singular integrands, among other problems, there is region of small probability under p that is very important to sample from. Let q be any probability density function satisfying $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. Then for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sampled from q, the importance sampling estimate

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^n \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}$$

satisfies $\mathbb{E}(\hat{\mu}_q) = \mu$. Let $D(q) = \{\boldsymbol{x} \mid q(\boldsymbol{x}) > 0\}$. The variance of $\hat{\mu}_q$ is σ_q^2/n where

$$\sigma_q^2 = \frac{1}{n}\left(\int_{D(q)} \frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} - \mu^2\right) = \int_{D(q)} \frac{(f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.$$

Sampling from the nominal distribution p has variance σ_p^2/n where $\sigma_p^2 = \int (f(\boldsymbol{x})-\mu)^2 p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sigma^2$. When $f \geqslant 0$ the optimal q is proportional to fp and it yields variance 0. Good choices for q sometimes yield $\sigma_q^2 \ll \sigma_p^2$. Poor choices can have $\sigma_q^2 = \infty$ even when q is nearly proportional to fp, because $q(\boldsymbol{x})$ appearing in the denominator of the variance integral may be small. It is often convenient to work with the following mean square,

$$\mathrm{MS}(\hat{\mu}_q) \equiv \mu^2 + \sigma_q^2 = \int_{D(q)} \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x},$$

instead of with the variance.

Now suppose that h is a vector-valued function for which we know $\mathbb{E}_p(h(x))$. That is, $\int h(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta \in \mathbb{R}^K$, where θ is known. For any $\beta \in \mathbb{R}^K$

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) - \beta^{\mathsf{T}} h(\boldsymbol{x}_i) + \beta^{\mathsf{T}}\theta$$

is an unbiased estimate of μ. Then

$$\mathrm{Var}(\hat{\mu}_\beta) = \frac{\sigma_\beta^2}{n}, \quad \text{where} \quad \sigma_\beta^2 = \int (f(\boldsymbol{x}) - \mu - (h-\theta)^{\mathsf{T}}\beta)^2\,\mathrm{d}\boldsymbol{x}.$$

The (unknown) variance-optimal β is $\beta_* = \mathrm{Var}(h(\boldsymbol{x}))^{-1}\mathrm{Cov}(h(\boldsymbol{x}), f(\boldsymbol{x}))$. A standard practice is to estimate β_* by its least squares estimate $\hat{\beta}$ and then use $\hat{\mu}_{\hat{\beta}}$. This estimator has an asymptotically negligible bias of order $O(1/n)$ as $n \to \infty$ with J fixed. While the bias is ordinarily ignored, it can be eliminated entirely by using different observations to compute $\hat{\beta}$ than those used to compute $\hat{\mu}$ or by cross-validatory schemes (Avramidis and Wilson, 1993), at the expense of increased bookkeeping.

A sample least squares estimator is typically root-n consistent: $\hat{\beta} = \beta_* + O_p(n^{-1/2})$. An error of size $n^{-1/2}$ in β raises the variance of the estimate for μ by a multiplicative factor of $1+O(n^{-1})$ because that variance is locally quadratic around β_*. Because the effect of estimated $\hat{\beta}$ is negligible, we analyze the effect of control variates as if we were using the unknown optimal coefficient β_*.

Owen and Zhou (2000) incorporate control variates into importance sampling. Our approach differs slightly from that paper which defines control variates via known integrals $\int h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, rather than known expectations. Their $h(\boldsymbol{x})$ is our $h(\boldsymbol{x})p(\boldsymbol{x})$.

We may incorporate a control variate into importance sampling via either

$$\tilde{\mu}_{q,\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\big[f(\boldsymbol{x}_i) - \beta^{\mathsf{T}} h(\boldsymbol{x}_i)\big]p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} + \beta^{\mathsf{T}}\theta, \quad \text{or}$$
$$\hat{\mu}_{q,\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\big[f(\boldsymbol{x}_i) - \beta^{\mathsf{T}} (h(\boldsymbol{x}_i) - \theta)\big]p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}.$$

Both estimators $\tilde{\mu}_{q,\beta}$ and $\hat{\mu}_{q,\beta}$ require that we can compute the ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$.

In ordinary control variates, where $q = p$, these two estimates coincide. They also coincide for $\theta = 0$, or even for $\beta^{\mathsf{T}}\theta = 0$, but in general they are different:

$$\hat{\mu}_{q,\beta} - \tilde{\mu}_{q,\beta} = \frac{\beta^{\mathsf{T}}\theta}{n}\sum_{i=1}^{n}\Big(\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} - 1\Big).$$

If $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$, then $\mathbb{E}(\hat{\mu}_{q,\beta}) = \mathbb{E}(\tilde{\mu}_{q,\beta}) = \mu$ for all $\beta \in \mathbb{R}^K$. Strictly speaking, we only need $q(\boldsymbol{x}) > 0$ whenever $h(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$ or $(h(\boldsymbol{x}) - \theta)p(\boldsymbol{x}) \neq 0$ for unbiasedness of $\tilde{\mu}_{q,\beta}$ and $\hat{\mu}_{q,\beta}$ respectively.

We find that

$$n\mathrm{Var}(\tilde{\mu}_{q,\beta}) = \int \Big(\frac{fp}{q} - \mu - \beta^{\mathsf{T}}\Big(\frac{hp}{q} - \theta\Big)\Big)^2 q\,\mathrm{d}\boldsymbol{x}, \quad \text{and} \tag{1}$$
$$n\mathrm{Var}(\hat{\mu}_{q,\beta}) = \int \Big(\frac{fp}{q} - \mu - \beta^{\mathsf{T}}\frac{(h-\theta)p}{q}\Big)^2 q\,\mathrm{d}\boldsymbol{x}. \tag{2}$$

Here and below we omit the argument $\boldsymbol{x}$ to shorten some expressions. Neither estimator universally dominates the other.

The optimal β is generally unknown. For $\widetilde{\mu}_{q,\beta}$ we see from (1) that the optimal β is the slope in a q-weighted least squares regression of fp/q on $hp/q - \theta$. For $\hat{\mu}_{q,\beta}$, and using (2), we find that the optimal β is from a q-weighted regression of fp/q on a slightly different predictor, $hp/q - \theta p/q$.

Given n observations $\boldsymbol{x}_i$ from q, we can estimate β by running the corresponding regressions. Unweighted least squares is then appropriate because the sample values are already sampled from q.

We will find the mean square formulation useful. The variance of $\tilde{\mu}_{q,\beta}$ satisfies $n\mathrm{Var}(\tilde{\mu}_{q,\beta}) = \mathrm{MS}(\tilde{\mu}_{q,\beta}) - \mu^2$ where

$$\mathrm{MS}(\tilde{\mu}_{q,\beta}) = \int \frac{[(f - \beta^{\mathsf{T}} h)p + \beta^{\mathsf{T}}\theta q]^2}{q}\,\mathrm{d}\boldsymbol{x}, \quad \text{and} \tag{3}$$

$$\mathrm{MS}(\hat{\mu}_{q,\beta}) = \int \frac{\big[f - \beta^{\mathsf{T}}(h - \theta)\big]^2 p^2}{q}\,\mathrm{d}\boldsymbol{x} \tag{4}$$

for $n\mathrm{Var}(\hat{\mu}_{q,\beta}) = \mathrm{MS}(\hat{\mu}_{q,\beta}) - \mu^2$.

Though neither estimator $\hat{\mu}_{q,\beta}$ or $\tilde{\mu}_{q,\beta}$ will always have lower variance, they do differ on some other meaningful properties. We find in Section 3 that when importance sampling from mixtures, $\tilde{\mu}_{q,\beta}$ offers a regret bound compared to the individual mixture components. The estimator $\hat{\mu}_{q,\beta}$ also offers a regret bound and moreover, it can be optimized jointly over α and β by convex optimization. That makes it possible to develop an adaptive importance sampling strategy.

3 Regret bounds and convexity

In this section we produce regret bounds describing how much variance might be increased when sampling from multiple distributions compared to using the unknown optimal sampling distribution. We begin with sampling from mixtures. Sampling a deterministic number of observations from each mixture component is then at least as good. Then we look at Veach's multiple importance sampling. Finally we prove convexity results.

3.1 Mixture importance sampling

Here we sample from a mixture of distributions $q_1, q_2, \ldots, q_J$ of the form $q_\alpha = \sum_{j=1}^J \alpha_j q_j$ where $\alpha_j \geqslant 0$ and $\sum_{j=1}^J \alpha_j = 1$. We write

$$S = \{(\alpha_1, \ldots, \alpha_J) \mid \alpha_j \geqslant 0, \sum_{j=1}^J \alpha_j = 1\}$$

for the simplex of mixture weights α and S^0 for the interior of S in which every $\alpha_j > 0$. We also write $\hat{\mu}_\alpha$ as a shorthand for $\hat{\mu}_{q_\alpha}$ when the list of q_j is fixed. We choose q_α so that $q_\alpha(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. The variance of $\hat{\mu}_\alpha$ is σ_α^2/n where

$$\sigma_\alpha^2 = \int_{D(q_\alpha)} \frac{\big(f(\boldsymbol{x})p(\boldsymbol{x}) - \mu\sum_{j=1}^J \alpha_j q_j(\boldsymbol{x})\big)^2}{\sum_{j=1}^J \alpha_j q_j(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}. \tag{5}$$

It is a good practice to include a defensive mixture component, $q_j = p$. Without loss of generality we let $q_1 = p$ in defensive mixtures. The simplex of defensive mixture weights is

$$S_{\mathrm{def}} = \{\alpha \in S \mid \alpha_1 > 0\}.$$

For a defensive mixture, $p/q_\alpha = p/(\alpha_1 p + \cdots + \alpha_J q_J) \leqslant 1/\alpha_1$. This bounded density ratio leads to the variance bound

$$\sigma_\alpha^2 \leqslant \frac{1}{\alpha_1}\bigl(\sigma_p^2 + (1-\alpha_1)\mu^2\bigr), \tag{6}$$

which follows from the $J = 2$ case in Hesterberg (1988).

While defensive importance sampling yields a variance not so much more than σ_p^2, it might yield a variance much larger than $\sigma_{q_j}^2$ for some $j > 1$. Owen and Zhou (2000) counter this problem via control variates devised from the mixture components.

The components q_j integrate to 1 because they are probability densities. For $j = 1, \ldots, J$, let

$$h_j(\boldsymbol{x}) = \begin{cases} q_j(\boldsymbol{x})/p(\boldsymbol{x}), & p(\boldsymbol{x}) > 0 \\ 0, & \text{else.} \end{cases} \tag{7}$$

If $p(\boldsymbol{x}) > 0$ whenever $q_j(\boldsymbol{x}) > 0$, then $\theta_j = \int h_j(\boldsymbol{x}) p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$. In other settings we may still know θ_j, though some special purpose argument would be required to find it.

Definition 1. Given a density p, the densities $q_1, \ldots, q_J$ and the point $\alpha \in S$ together satisfy the *support conditions* if $p(\boldsymbol{x}) > 0$ whenever any $q_j(\boldsymbol{x}) > 0$ and $\sum_{j=1}^J \alpha_j q_j(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$.

Let $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_J(\boldsymbol{x}))^\mathsf{T}$ and $\theta = (\theta_1, \ldots, \theta_J)^\mathsf{T}$. The combined control variate with importance sampling estimates are now

$$\tilde{\mu}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^n \frac{[f(\boldsymbol{x}_i) - \beta^\mathsf{T} h(\boldsymbol{x}_i)]p(\boldsymbol{x}_i)}{q_\alpha(\boldsymbol{x}_i)} + \beta^\mathsf{T}\theta, \quad \text{and} \tag{8}$$

$$\hat{\mu}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^n \frac{[f(\boldsymbol{x}_i) - \beta^\mathsf{T}(h(\boldsymbol{x}_i) - \theta)]p(\boldsymbol{x}_i)}{q_\alpha(\boldsymbol{x}_i)}. \tag{9}$$

If $\alpha \in S^0$ and $q_\alpha(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$ then $\mathbb{E}(\tilde{\mu}_{\alpha,\beta}) = \mathbb{E}(\hat{\mu}_{\alpha,\beta}) = \mu$ for $\boldsymbol{x}_i \sim q_\alpha$. Note that if we have additional control variates h_j for $J < j \leqslant K$ we can incorporate them directly into $h(\boldsymbol{x})$ and θ.

Owen and Zhou (2000) noticed that the regression to estimate β is always rank deficient for control variates defined via the sampling densities. That problem can be countered by either dropping one of the control variates or by computing the least squares estimate $\hat{\beta}$ via the singular value decomposition

(SVD). In the present formulation of the problem, the regressor $(h_1 - 1)p/q_\alpha$ in $\hat{\mu}_{\alpha,\beta}$ (equation (9)) satisfies

$$(h_1(\boldsymbol{x}) - 1)p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) = (p(\boldsymbol{x}) - p(\boldsymbol{x}))/q_\alpha(\boldsymbol{x}) = 0$$

for all $\boldsymbol{x}$. The defensive control variate is by definition zero and we simply drop it from the regression. The regressors $h_j p/q_\alpha$ in estimator $\tilde{\mu}_{\alpha,\beta}$ (equation (8)) satisfy $\sum_{j=1}^J \alpha_j h_j p/q_\alpha = \sum_{j=1}^J \alpha_j q_j/q_\alpha = 1$, at all $\boldsymbol{x}$. One might just drop the control variate with smallest sample variance, or use the SVD.

Theorem 1 below shows there always exists a β that makes the estimator $\widetilde{\mu}_{\alpha,\beta}$ nearly as good as ordinary importance sampling from the best q_j in a list.

Theorem 1. *Let $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for a probability density p. Let densities $q_1, \dots, q_J$ and $\alpha \in S$ satisfy the support conditions of Definition 1. Let $\tilde{\mu}_{\alpha,\beta}$ be the estimator given by (8) for independent $\boldsymbol{x}_i \sim q_\alpha = \sum_{j=1}^J \alpha_j q_j$ with h_j defined by (7). If β_* is any minimizer of $\mathrm{Var}(\tilde{\mu}_{\alpha,\beta})$, then*

$$\mathrm{Var}(\tilde{\mu}_{\alpha,\beta_*}) \leqslant \frac{1}{n} \min_{1 \leqslant j \leqslant J} \sigma_{q_j}^2/\alpha_j.$$

Proof. See Section 7.1. □

The quantity $\sigma_{q_k}^2/(n\alpha_k)$ is the variance we would have gotten from importance sampling using only $n\alpha_k$ observations all taken from q_k. The other variances $\sigma_{q_j}^2$ could be arbitrarily large and yet they cannot make the estimate $\widetilde{\mu}_{\alpha,\beta_*}$ worse than importance sampling with $n\alpha_k$ observations from q_k. We can improve on this result as follows.

Corollary 1. *Under the conditions of Theorem 1*

$$\min_\beta \mathrm{Var}(\tilde{\mu}_{\alpha,\beta}) \leqslant \min_{1 \leqslant j \leqslant J} \min_{\beta_j} \frac{\sigma_{q_j,\beta_j}^2}{n\alpha_j}.$$

Proof. See Section 7.3. □

Next we show a regret bound for $\hat{\mu}_{\alpha,\beta}$.

Theorem 2. *Let $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for a probability density p. Let densities $q_1, \dots, q_J$ and $\alpha \in S$ satisfy the support conditions of Definition 1. Let $\hat{\mu}_{\alpha,\beta}$ be the estimator given by (9) for independent $\boldsymbol{x}_i \sim q_\alpha = \sum_{j=1}^J \alpha_j q_j$ with h_j defined by (7). If β_* is any minimizer of $\mathrm{Var}(\hat{\mu}_{\alpha,\beta})$, then*

$$\mathrm{Var}(\hat{\mu}_{\alpha,\beta_*}) \leqslant \frac{1}{n} \min_{1 \leqslant j \leqslant J} \sigma_{q_j}^2/\alpha_j.$$

Proof. See Section 7.2. □

Corollary 2. *Under the conditions of Theorem 2*

$$\min_\beta \mathrm{Var}(\hat{\mu}_{\alpha,\beta}) \leqslant \min_{1 \leqslant j \leqslant J} \min_{\beta_j} \frac{\sigma_{q_j,\beta_j}^2}{n\alpha_j}.$$

Proof. See Section 7.3. □

In practice it is wise to stratify the sampling (Hesterberg, 1995). In stratification we take n_j observations from q_j instead of sampling iid from q_α with $\alpha_j = n_j/n$. Stratification of mixture importance sampling reduces the variance of $\hat{\mu}_\alpha$ by

$$\frac{1}{n}\sum_{j=1}^{J}\alpha_j(\mu_j-\mu)^2, \quad \text{where} \quad \mu_j = \int \frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Stratification can reduce but cannot increase the variance of $\hat{\mu}_\alpha$, $\hat{\mu}_{\alpha,\beta}$, $\tilde{\mu}_\alpha$ and $\tilde{\mu}_{\alpha,\beta}$. Thus when each $n\alpha_j$ is a positive integer, Theorems 1 and 2 as well as Corollaries 1 and 2 all apply to stratified mixture sampling. We will use $\mathrm{Var}_{\mathrm{strat}}$ to denote the variance under stratified sampling of mixture sampling estimators.

3.2 Multiple importance sampling

Multiple importance sampling (Veach, 1997, Chapter 9) takes n_j observations from q_j for $j = 1, \ldots, J$ for a total of $n = \sum_{j=1}^{J} n_j$ observations. The multiple importance sampling estimator is defined in terms of a partition of unity: J functions that add up to 1. They must obey some additional conditions given by this definition:

Definition 2. A multiple sampling partition of unity for distributions $p, q_1, \ldots, q_J$ is a set of real-valued functions $\omega_j(\boldsymbol{x})$ for which $\sum_{j=1}^{J}\omega_j(\boldsymbol{x}) = 1$ at all $\boldsymbol{x}$ with $p(\boldsymbol{x}) > 0$, and for which $\omega_j(\boldsymbol{x}) = 0$ whenever $q_j(\boldsymbol{x}) = 0$.

The multiple importance sampling estimator is

$$\check{\mu}_\omega = \sum_{j=1}^{J}\frac{1}{n_j}\sum_{i=1}^{n_j}\omega_j(\boldsymbol{x}_{ij})\frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})}{q_j(\boldsymbol{x}_{ij})}. \tag{10}$$

We easily find that $\mathbb{E}(\check{\mu}_\omega) = \mu$ when $\boldsymbol{x}_{ij} \sim q_j$ independently and ω_j satisfy Definition 2. Multiple importance sampling reduces to classic stratified sampling when the ω_j have disjoint supports. In its general form, it is like a stratification that allows overlapping strata.

Among the many choices for ω_j, Veach (1997) favors the balance heuristic

$$\omega_j^{\mathrm{BH}}(\boldsymbol{x}) = \frac{n_j q_j(\boldsymbol{x})}{\sum_{k=1}^{J} n_k q_k(\boldsymbol{x})} = \frac{n_j q_j(\boldsymbol{x})}{n q_\alpha(\boldsymbol{x})}, \quad \text{for} \quad \alpha_j = n_j/n.$$

With this choice

$$\check{\mu}_{\omega^{\mathrm{BH}}} = \sum_{j=1}^{J}\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{n_j q_j(\boldsymbol{x}_{ij})}{n q_\alpha(\boldsymbol{x}_{ij})}\frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})}{q_j(\boldsymbol{x}_{ij})} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})}{q_\alpha(\boldsymbol{x}_{ij})}. \tag{11}$$

Notice that the coefficient of $f(\boldsymbol{x}_{ij})$ does not depend on which mixture component $\boldsymbol{x}_{ij}$ was sampled from. Equation (11) shows that the balance heuristic is the same as mixture sampling with stratification, but without control variates. Therefore control variates with mixture sampling improves on the balance heuristic.

Veach (1997, Theorem 9.2) proves that the balance heuristic satisfies

$$\mathrm{Var}(\check{\mu}_{\omega^{\mathrm{BH}}}) \leqslant \mathrm{Var}(\check{\mu}_{\omega}) + \Big(\frac{1}{\min_j n_j} - \frac{1}{n}\Big)\mu^2 \tag{12}$$

for any ω satisfying Definition 2. Equation (12) is more general than the Theorems of the previous section because it covers any alternative partition of unity, not just sampling from the q_j themselves. The bound (12) also applies directly to $\hat{\mu}_\alpha$ under stratified mixture sampling. For the alternative heuristic that simply takes $\omega_j(\boldsymbol{x}) = 1$ (and all other $\omega_k = 0$), equation (12) simplifies to

$$\mathrm{Var}_{\mathrm{strat}}(\hat{\mu}_\alpha) = \mathrm{Var}(\check{\mu}_{\omega^{\mathrm{BH}}}) \leqslant \frac{\sigma_j^2}{n_j} + \Big(\frac{1}{\min_j n_j} - \frac{1}{n}\Big)\mu^2,$$

while from Theorems 1 and 2 we have

$$\min_\beta \mathrm{Var}_{\mathrm{strat}}(\tilde{\mu}_{\alpha,\beta}) \leqslant \min_{1\leqslant j\leqslant J} \frac{\sigma_j^2}{n_j}, \quad \text{and} \quad \min_\beta \mathrm{Var}_{\mathrm{strat}}(\hat{\mu}_{\alpha,\beta}) \leqslant \min_{1\leqslant j\leqslant J} \frac{\sigma_j^2}{n_j}, \tag{13}$$

for $\alpha \in S^0$. The control variates remove the multiple of μ^2 which can be arbitrarily large if one adds a large enough constant to f.

Veach (1997) remarks that the balance heuristic is not competitive when one of the component samplers is especially effective. For such low variance problems he proposes alternative heuristics, such as a power heuristic that takes $\omega_j \propto q_j^r$ with $r = 2$ singled out as a useful case. Equation (13) shows that incorporating control variates automatically yields competitive estimates for all cases including those high accuracy ones.

If J is not too large then a reasonable strategy is to take each $\alpha_j = 1/J$. The variance under either multiple or mixture importance sampling will be at most J times the best single component variance and possibly much better. When the individual σ_j^2 vary over many orders of magnitude, losing a factor of J compared to the best is acceptable. However when J is large we may want to optimize over α.

3.3 Convexity

Here we provide convexity results for the estimation variances. These results support adaptive minimization strategies where one alternates between minimizing an estimated variance and sampling from the best estimate of the optimal α. Theorem 3 covers the case of mixture importance sampling without control variates.

Theorem 3. *The quantity $\sigma_\alpha^2 = n\mathrm{Var}(\hat{\mu}_\alpha)$ in equation (5) is a convex function of $\alpha \in S^0$.*

Proof. See Section 7.5 □

To handle control variates, the optimal α depends on the regression vector β and vice versa. Theorem 4 shows that the variance of $\hat{\mu}_{\alpha,\beta}$ is jointly convex in these two vectors.

Theorem 4. *The quantity $\sigma^2_{\alpha,\beta} = n\mathrm{Var}(\hat{\mu}_{\alpha,\beta})$ in equation (9) is a jointly convex function of $\alpha \in S^0$ and $\beta \in \mathbb{R}^J$.*

Proof. See Section 7.6. □

We can apply convex optimization to choose α for $\hat{\mu}_\alpha$ or $\hat{\mu}_{\alpha,\beta}$ but not directly for $\widetilde{\mu}_{\alpha,\beta}$. Lemma 1 shows that when a defensive component is present then the set of estimators attained by the two estimator forms, $\tilde{\mu}_{\alpha,\beta}$ and $\hat{\mu}_{\alpha,\beta}$ coincide. Therefore if we choose α to minimize the variance of $\hat{\mu}_{\alpha,\beta}$ and the optimal $\alpha_* \in S_{\mathrm{def}}$, then α_* is optimal for the variance of $\tilde{\mu}_{\alpha,\beta}$ too.

Lemma 1. *Suppose that $\alpha \in S_{\mathrm{def}}$ and choose $\beta \in \mathbb{R}^J$. Let α and the densities $q_1, \dots, q_J$ satisfy the support condition of Definition 1 and let h_j be defined by (7). Then there exists a $\gamma \in \mathbb{R}^J$ such that $\hat{\mu}_{\alpha,\gamma} = \tilde{\mu}_{\alpha,\beta}$.*

Proof. See Section 7.4 □

We often have to estimate multiple integrals from the same sample values $\boldsymbol{x}_{ij}$. For integrands f_k, $k = 1, \dots, K$, let $\hat{\mu}^{(k)}_{\alpha,\beta_k}$ be the estimate (9) with $f = f_k$, $\alpha \in S$ and $\beta_k \in \mathbb{R}^J$. For weights $w_k > 0$, we might make the tradeoff via either

$$\max_{1 \leqslant k \leqslant K} w_k \mathrm{Var}(\hat{\mu}^{(k)}_{\alpha,\beta_k}) \quad \text{or} \quad \sum_{k=1}^{K} w_k \mathrm{Var}(\hat{\mu}^{(k)}_{\alpha,\beta_k}).$$

Both of these criteria inherit convexity from Theorem 4. Choosing not to use control variates corresponds to constraining $\beta_k = 0$ and the multi-integral criteria remain convex in α.

4 Choosing α

Given a sample $\boldsymbol{x}_i \overset{\text{iid}}{\sim} q_{\alpha'}$ for $\alpha' \in S^0$, we may use it to estimate an improved mixture vector α. From the proof of Theorem 4 we know that the sampling variance using $\alpha \in S^0$ satisfies

$$\sigma^2_{\alpha,\beta} + \mu^2 = \int_D \frac{\bigl(f(\boldsymbol{x})p(\boldsymbol{x}) + \sum_{J=1}^{J} \beta_j (q_j(\boldsymbol{x}) - p(\boldsymbol{x}))\bigr)^2}{q_\alpha(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x}. \tag{14}$$

An unbiased estimate of the right hand side of (14) is

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\bigl(f(\boldsymbol{x}_i)p(\boldsymbol{x}_i) + \sum_{J=1}^{J} \beta_j (q_j(\boldsymbol{x}_i) - p(\boldsymbol{x}_i))\bigr)^2}{q_\alpha(\boldsymbol{x}_i) q_{\alpha'}(\boldsymbol{x}_i)}. \tag{15}$$

The quantity in (15) is jointly convex in $\alpha \in S^0$ and $\beta \in \mathbb{R}^J$. Therefore we can use convex optimization to find the minimizers (α_*, β_*) of (15) and then generate a new sample using α_*.

Equation (15) may be used to devise adaptive importance sampling methods in which early function evaluations are used to change the value of α used to generate later ones. A full exploration of adaptive strategies is beyond the scope of this paper. Our numerical examples in Section 5 use a two stage method in which a pilot sample is used to estimate α for a final sample that is used for the estimate.

4.1 Bounding α_j away from zero

It is risky to sample with a value of α_j that is so small that we may not get a reasonable number n_j of values from q_j. We may prefer to impose constraints $\alpha_j > \epsilon$ on each sampling fraction. For $\epsilon > 0$ define

$$S^\epsilon = \Big\{(\alpha_1, \cdots, \alpha_J) \mid \alpha_j \geq \epsilon, \sum_{j=1}^J \alpha_j = 1\Big\}.$$

Theorem 5 shows that we pay a remarkably small variance penalty when requiring $\alpha \in S^\epsilon$.

Theorem 5. *Let α and β minimize $\mathrm{Var}(\tilde{\mu}_{\alpha,\beta})$ over $\alpha \in S$ and $\beta \in \mathbb{R}^J$. For $0 < \epsilon < 1/J$ let δ and γ minimize $\mathrm{Var}(\tilde{\mu}_{\delta,\gamma})$ over $\delta \in S^\epsilon$ and $\gamma \in \mathbb{R}^J$. Then*

$$\mathrm{Var}(\tilde{\mu}_{\delta,\gamma}) \leq (1 + \eta(\alpha))\mathrm{Var}(\tilde{\mu}_{\alpha,\beta}),$$

where

$$\eta(\alpha) = \frac{1 - \epsilon(J - K)}{1 - \epsilon J}, \quad \text{for} \quad K = K(\alpha) = \#\{\alpha_j < \epsilon\}.$$

The same holds for the $\hat{\mu}_{\alpha,\beta}$ and $\hat{\mu}_{\delta,\gamma}$ analogously defined.

Proof. See Section 7.7. □

The penalty factor $1 + \eta(\epsilon)$ in Theorem 5 is asymptotically $1 + K\epsilon + O(\epsilon^2)$ as $\epsilon \to 0$, and $1 + K\epsilon \leqslant 1 + (J-1)\epsilon$. The price to be paid is bounded by a term nearly linear in the number of components with $\alpha_j < \epsilon$. If for example we insist on at least k (e.g. $k = 10$) observations from each q_j, then $\epsilon = k/n$ and $1 + \eta(\epsilon) \leqslant (1 - \epsilon)/(1 - J\epsilon) = (1 - k/n)/(1 - Jk/n)$.

For any given α, a suboptimal $\hat{\beta}$ generally increases variance by a factor of $1+O(\|\hat{\beta}-\beta_*\|^2)$ because the variance function is twice continuously differentiable on $\mathbb{R}^J$. A suboptimal α can multiply the variance by $1 + O(\|\hat{\alpha} - \alpha_*\|)$. The reason for the different rate is that the optimum α_* can be on the boundary of the simplex S and there is no reason to suppose that the gradient of the variance with respect to α vanishes there.

4.2 Optimization

We found that off-the-shelf convex optimization codes we tried were slow. As a result it was worthwhile to write our own code for the specific case we need in this problem. We use standard convex optimization methods, but make modifications to improve robustness.

Our workhorse numerical optimization problem is

$$\begin{aligned} \text{minimize} \quad & f_0(\alpha,\beta) = \sum_{i=1}^{n} \frac{(Y_i - X_i^{\mathsf{T}}\beta)^2}{Z_i^{\mathsf{T}}\alpha} \\ \text{subject to} \quad & \beta \in \mathbb{R}^K, \\ & \alpha_j > \epsilon_j \geqslant 0,\ j = 1,\ldots,J, \text{ and} \\ & \sum_{j=1}^{J} \alpha_j < 1 + \eta, \text{ for } \eta \geqslant 0, \end{aligned} \tag{16}$$

which is equivalent to (15). In this notation $Y \in \mathbb{R}^n$ represents fp, $X \in \mathbb{R}^{n\times K}$ represents $q_j - p$ for $j > 1$ (and possibly some additional control variates) and $Z \in [0,\infty)^{n\times J}$ represents $q_j(\boldsymbol{x})$. We have multiplied the objective function by n. Valid problem data have X of full rank $K \leqslant n$ and no row of Z equals $(0,\ldots,0)$.

The constraints are defined by parameters $\epsilon_j \geqslant 0$ and $\eta \geqslant 0$. We assume that there is at least one feasible vector α. We do not need to impose the constraint $\sum_{j=1}^{J}\alpha_j = 1$. The way the problem is formulated makes that constraint equivalent to $\sum_{j=1}^{J}\alpha_j \leqslant 1$. As a result we do not have to mix equality and inequality constraints. We use a very small η, usually $\eta = 0$, and then renormalize the solution α to sum to one. We can choose a different ϵ_j for each α_j. For example we might want a higher lower bound on the defensive component than on the others.

We use a barrier method, Algorithm 11.1 of Boyd and Vandeberghe (2004), with function

$$f(\alpha,\beta;\rho) = f(\alpha,\beta) - \rho\sum_{j=1}^{J}\log(\alpha_j - \epsilon_j) - \rho\log\biggl(1 + \eta - \sum_{j=1}^{J}\alpha_j\biggr), \tag{17}$$

for $\rho > 0$. For each ρ, this function is jointly convex over the convex set of allowable (α,β).

We minimize (17) by a damped Newton iteration, Algorithm 9.5 of Boyd and Vandeberghe (2004) using a backtracking line search (their Algorithm 9.2). The solutions are denoted $\alpha(\rho)$ and $\beta(\rho)$. Beginning with $\rho = \rho_1$ we decrease ρ by a factor of $\phi = 2$, taking $\rho_\ell = \rho_{\ell-1}/\phi$ for $\ell \geqslant 2$, minimizing $f(\alpha,\beta;\rho_\ell)$ at step $\ell \geqslant 1$ obtaining $\alpha^{(\ell)} = \alpha(\rho_\ell)$ and $\beta^{(\ell)} = \beta(\rho_\ell)$, using $\alpha^{(\ell-1)}$ and $\beta^{(\ell-1)}$ as starting values. To start the whole algorithm we use $\beta^{(0)} = 0$ and $\alpha^{(0)}$ given by

$$\alpha_j^{(0)} = \epsilon_j + \delta, \quad \text{where} \quad \delta = \frac{1 + \eta - \sum_{j=1}^{J}\epsilon_j}{J+1}.$$

The constraints are feasible if and only if $\delta > 0$. This choice of starting value equalizes all of the $J+1$ logarithmic penalty terms. We continue decreasing ρ until we reach ρ_ℓ where $(J+1)\rho_\ell < f_0(\alpha^{(\ell)}, \beta^{(\ell)})\varepsilon$, for a tolerance $\varepsilon > 0$. At this point we have a certificate that

$$f_0(\alpha^{(\ell)}, \beta^{(\ell)}) \leqslant (1+\varepsilon) \min_{\alpha,\beta} f_0(\alpha, \beta).$$

We have found it helpful to consider preconditioning the Hessian matrix for our Newton steps. The intuition is that if α and β are on very different scales, scaling them to be comparable will reduce the conditon number of H. We choose a diagonal preconditioning matrix P, with $P_{ii} = 1$ for $1 \leqslant i \leqslant a$ and $P_{ii} = p$ for $a < i \leqslant b$, where a and b are the dimensions of α and β respectively, and the indices of H corresponding to α precede those corresponding to β. We take

$$p = \sqrt{\frac{\text{median}(\text{abs}(H_{1:a,1:a}))}{\text{median}(\text{abs}(H_{(a+1):(a+b),(a+1):(a+b)}))}}.$$

With this choice, the median absolute value of the upper left $a \times a$ block of PHP is the same as that in its lower right $b \times b$ block. If PHP has a smaller condition number than H does, then we use the preconditioning, otherwise not.

5 Examples

Here we illustrate the use of convex optimization to find an improved mixture importance sampler. We use a two step process. A preliminary sample has $\alpha_j = 1/J$ for all $j = 1, \ldots, J$. On this sample we estimate optimal α_j subject to constraints $\alpha_j \geqslant \epsilon > 0$. Then we use the second sample to estimate the desired expectation.

There are two main classes of problems where importance sampling is crucial: spiky or even singular functions, and rare events. In these examples we model the imperfect information one might have in practice when choosing q_j. Therefore we include distributions q_j that are good but not quite optimal as well as some q_j that correspond to poor choices. The method does better if it can give a small α_j to the unhelpful q_j.

5.1 Singular function

For $\boldsymbol{x} \in \mathbb{R}^d$ we take the nominal distribution to be $p(\boldsymbol{x}) = \mathcal{N}(\mu, \Sigma)$ and $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_0\|^{-\gamma}$, for some $\boldsymbol{x}_0$. We choose $\gamma < d/2$ so that our example has $\text{Var}_p(f(\boldsymbol{x})) < \infty$. For this example, we use $d = 5$, $\gamma = 2.4$, $\Sigma_{i,j} = \rho^{|i-j|}$ where $\rho = 1/2$ as well as

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

We choose a sequence of proposal densities $q_j = \mathcal{N}(\boldsymbol{x}_k, 2^{-r} I_5)$, for $k = 1, \ldots, 5$, $r = 1, \ldots, 10$, with index $j = 10(k-1) + r$, where

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \boldsymbol{x}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{x}_4 = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \boldsymbol{x}_5 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

We include a defensive component via $q_{51} = p$. Note that we have moved the defensive component from $j = 1$ to $j = J$. The only singularity is at $\boldsymbol{x}_0$. We use a safety lower bound $\epsilon = 0.1/51$ on all α_j. That is, no proposal density will get less than 10% of an equal share. This still allows the algorithm to allocate the remaining 90% of the sample effort in a way to minimize variance.

For each simulation, we use 2 batches of sizes $n_1 = 10^4$, and $n_2 = 5 \times 10^5$ respectively. After sampling the first n_1 values, we optimize (15) over α with β fixed at zero to get α_*. We also optimize (15) over both α and β getting $(\alpha_{**}, \beta_{**})$. These choices allow us to compare methods that use control variates with those that do not.

To study mixture optimization with control variates we take the second sample from $q_{\alpha_{**}}$. Our estimate of μ is based on the second sample only:

$$\hat{\mu}_{\alpha_{**},\beta} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{f(x_i)p(x_i) - \beta^{\mathsf{T}}\big(\boldsymbol{q}(x_i) - p(x_i)\mathbf{1}\big)}{q_{\alpha_{**}}(x_i)}.$$

The value of β used there is found by least squares fitted to the second sample value, instead of using β_{**}. The vector $\boldsymbol{q}(x_i)$ has elements $q_j(x_i)$ (ordinarily just $J-1$ of them) and $\mathbf{1}$ has all 1s.

To judge the effect of dropping control variates we also considered taking the second sample from q_* and using the estimate

$$\hat{\mu}_{\alpha_*} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{f(x_i)p(x_i)}{q_{\alpha_*}(x_i)}.$$

The judge the effect of non-uniform sampling, we take n_2 samples from the uniform mixture q_U which has $\alpha_j = 1/J$ for $j = 1, \ldots, J$. Then we compute

$$\hat{\mu}_{q_U} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{f(x_i)p(x_i)}{q_U(x_i)}, \quad \text{and}$$

$$\hat{\mu}_{q_U,\beta} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{f(x_i)p(x_i) - \beta^{\mathsf{T}}(\boldsymbol{q}(x_i) - p(x_i)\mathbf{1})}{q_U(x_i)}$$

Here β is estimated with these n_2 samples from the uniform mixture.

We ran 5000 replicates of these simulations. Because μ is not known we estimate the mean squared error for each method by its sample variance. The

Sampling	CV	$\hat{\mu}$	VRFmc	VRF.uis
U	No	0.173862	1.494	1.000
U	Yes	0.173871	3.273	2.190
α_*	No	0.173867	28.321	18.953
α_{**}	Yes	0.173867	33.013	22.093

Table 1: Singular function example. The estimate $\hat{\mu}$ was computed from 500,000 observations using the sampler given in the first column. Control variates were used in two of those samples. The final columns give variance reduction factors compared to plain Monte Carlo and compared to uniform mixture importance sampling with no control variates.

k	σ_r^2	Mean α_j	SD α_j
D	D	0.5263	0.028722
1	0.25	0.2721	0.029184
1	0.5	0.0912	0.037610
2	0.5	0.0065	0.005128
4	0.5	0.0039	0.002168
1	0.0625	0.0031	0.001171
1	0.125	0.0031	0.001210
1	0.03125	0.0027	0.000511
3	0.25	0.0026	0.000563
3	0.5	0.0025	0.000591

Table 2: Top 10 mixture components $\mathcal{N}(\boldsymbol{x}_k, \sigma_r^2 I_5)$ for the singular integrand in α_{**}, which uses control variates. D denotes the defensive mixture. The last columns are mean and sd of α_j over 5000 simulations.

Monte Carlo variance was found by running $N = 5 \times 10^7$ IID samples. Table 1 reports the results. We see that for the singular function, optimizing the sampling weights α reduces variance by about twenty-fold compared to using the uniform distribution on weights. Incorporating a control variate brings a modest improvement.

We also present mixture components with the top 10 values of estimated α in Table 2 and 3. About half of the sample is allocated to the defensive component in both settings. The true singularity point is at $k = 1$. That point gets chosen more often than the others, especially when no control variates are used.

5.2 Rare event

For our rare event simulation we took $\boldsymbol{x} \in \mathbb{R}^3$, $p(\boldsymbol{x}) = \mathcal{N}(\mu, I_3)$ and defined event sets

$$\mathcal{D}_i = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| > \eta_i \|\boldsymbol{z}_i\| \text{ elementwise and } \operatorname{sign}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{z}_i)\}$$

k	σ_r^2	Mean α_j	SD α_j
D	D	0.4838	0.018307
1	0.25	0.2839	0.021976
1	0.5	0.0971	0.019779
1	0.0625	0.0290	0.006382
1	0.03125	0.0099	0.002762
1	0.125	0.0050	0.005936
1	0.015625	0.0029	0.000733
3	0.25	0.0022	0.000421
3	0.5	0.0021	0.000358
1	0.0078125	0.0021	0.000353

Table 3: Top 10 mixture components $\mathcal{N}(\boldsymbol{x}_k, \sigma_r^2 I_5)$ for the singular integrand in α_*, which does not use control variates. D denotes the defensive mixture. The last columns are mean and sd of α_j over 5000 simulations.

for $i = 1, \ldots, 8$. The points $\boldsymbol{z}_i$ are all eight points of the form $(\pm\eta_i, \pm\eta_i, \pm\eta_i)$ for $\eta_i > 0$, with η_i chosen so that $\Pr(\boldsymbol{x} \in D_i) = 16^{-i} \times 10^{-3}$.

The sets $\mathcal{D}_i$ are disjoint. The integrand f is 1 if $\boldsymbol{x} \in \mathcal{D} = \cup_{i=1}^8 \mathcal{D}_i$ and is 0 otherwise. For this problem we know that

$$\mu = \mathbb{E}(f(\boldsymbol{x})) = \sum_{i=1}^{8} \Pr(\boldsymbol{x} \in \mathcal{D}_i) \doteq 6.666666665 \times 10^{-5}$$

and in a Monte Carlo sample of n observations the variance of the sample mean is $\mu(1-\mu)/n \doteq \mu/n$.

We choose $q_j = \mathcal{N}(\boldsymbol{z}_k, \sigma_r^2 I_3)$, for $k = 0, \ldots, 8$ as the mixture components, including a new point $\boldsymbol{z}_0 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^{\mathsf{T}}$ and the previous corners $\boldsymbol{z}_k, k = 1, \cdots, 8$. That is, we suppose that the user has some idea where the rare events are. But we suppose that there is little knowledge of the right sampling variance, so we take

$$\sigma_r^2 \in \left\{ \frac{1}{50}, \frac{1}{40}, \frac{1}{30}, \frac{1}{20}, \frac{1}{10}, \frac{1}{2}, 2, 10, 20, 30, 40, 50 \right\}$$

for $r = 1, \ldots, 12$ respectively. In addition to these 108 mixture components, we include the defensive component as $q_{109} = p$, once again as q_J, not q_1.

We take a pilot sample of $n_1 = 10^4$ observations to estimate mixture components with, and then we follow up with $n_2 = 100{,}000$ observations for the final estimate. We optimized with safety bound $\epsilon = 0.1/109$ for all α. This simulation was repeated 5000 times.

The results are in Table 4. For this problem we can use the actual Monte Carlo variance when computing the variance reduction factor with respect to plain Monte Carlo. The method of control variates also provides at estimate $\widehat{\mathrm{Var}}(\hat{\mu}_{\alpha,\beta})$ which comes from the regression model variance estimate of the intercept term. The last column in Table 4 shows the average over 5000 simulations

Sampling	CV	$10^5\hat{\mu}$	VRFmc	VRFuis	$\overline{\text{MSE}/\widehat{\text{VAR}}}$
U	No	6.669473	65.249	1.000	0.980
U	Yes	6.668018	69.516	1.065	0.976
α_*	No	6.667731	1663.983	25.502	1.018
α_{**}	Yes	6.667475	1676.611	25.696	1.015

Table 4: Rare event example. The estimate $\hat{\mu}$ was computed from 100,000 observations using the sampler given in the first column. Control variates were used in two of those samples. The next columns give variance reduction factors compared to plain Monte Carlo and compared to uniform mixture importance sampling with no control variates. The final column compares actual squared error with its sample estimate.

k	σ^2	Mean α_j	SD α_j
1	0.5	0.5188	0.064185
1	0.1	0.3196	0.071853
2	0.5	0.0266	0.006733
2	0.1	0.0249	0.005421
1	0.05	0.0058	0.019458
1	0.02	0.0028	0.003635
1	0.033	0.0020	0.004237
1	0.025	0.0018	0.002479
3	0.5	0.0014	0.000338
2	0.05	0.0013	0.000706

Table 5: Top 10 mixture components $\mathcal{N}(\boldsymbol{z}_k, \sigma_r^2 I_2)$ for the singular integrand in α_*, which does not use control variates. D denotes the defensive mixture. The last columns are mean and sd of α_j over 5000 simulations.

of the true mean squared error divided by its estimate. Here we see about a 25-fold gain from optimizing the mixture weights. There is almost no gain from using the control variate.

The top 10 mixture components in the rare event problem are listed in Table 5. That table has values for the mixture sampling without control variates. Including control variates gave the same ten components in the same order with nearly identical mixing probabilities. The method samples very little from the mixture components with large variances. The largest weights go on the least rare component corners $i = 1, 2, 3$.

5.3 Summary

In both of our examples there was a gain from optimizing over α but very little gain from employing the control variate. It can happen that a control variate

	n_1	n_2	U no CV	U CV	α_* no CV	α_{**} CV
Singularity	10^4	5×10^5	13.49	18.44	16.25	23.82
Rare event	10^4	10^5	4.15	6.19	9.61	23.28

Table 6: Average running times in seconds for four estimators on two examples.

makes an enormous improvement. One such case is example 2 in Owen and Zhou (2000).

Table 6 gives the average running time for the four methods we compare on the two examples. These times include running the pilot sample, any optimizations needed, and then running the final sample. The more complicated methods take longer but not enough to outweight the gain from optimizing α. In three of the examples, the gain from control variates becomes a small loss when computation time is accounted for. The exception is that for the singular function with uniform mixture sampling the control variate roughly halves the variance while taking about 37% longer, resulting in a minor efficiency gain.

6 Discussion

We have developed a second version of mixture importance sampling with control variates $\hat{\mu}_{\alpha,\beta}$ at (9). For defensive mixtures, this estimator has the same minimal variance as $\tilde{\mu}_{\alpha,\beta}$ at (8) presented in Owen and Zhou (2000). In their stratified versions these estimators enjoy both the regret bounds of the balance heuristic as well as the regret bounds of mixture importance sampling. The new estimator has a variance that is jointly convex in α and β. As a result, we can optimize a sample mean square over both of these vectors and then use the resulting α for a followup sample.

There is a large literature on adaptive importance sampling, going back at least to Marshall (1956). The cross-entropy method (Rubinstein and Kroese, 2004) is devoted to finding a single importance distribution $q(\boldsymbol{x}) = q(\boldsymbol{x};\theta)$ for a parameter $\theta \in \Theta$. It can be challenging to optimize the variance over θ and they optimize instead a (sample) entropy criterion comparing q to the optimal density proportional to fp (for $f \geqslant 0$). Cappé et al. (2008) seek to optimize a mixture $\sum_j \alpha_j q(\boldsymbol{x}, \theta_j)$ over both α_j and θ_j. They use an entropy criterion instead of variance and apply an EM style algorithm to estimate the optima. They note that the entropy is convex in the mixing parameters, but that it is not jointly convex in those parameters and the θ_j. Our methods may be used in conjunction with theirs. Given a fixed set of θ_j selected by some other method, we can jointly estimate the variance-optimal mixing parameters α_j and control variate coefficients β_j via convex optimization applied to sample values.

A multiple of μ^2/n appears in the regret bound for multiple importance sampling versus the unknown best mixture component. Using a control variate removes that term. The quantity μ^2/n is often negligible compared to the attained sample variance. For instance in plain Monte Carlo sampling of a rare

event, the variance $\mu(1-\mu)/n$ is far larger than μ^2/n and the latter may be ignored. But in some rare event simulations it is possible to obtain estimator with bounded relative error (Heidelberger, 1995). That means that for a sequence of problems with $\mu \downarrow 0$ the variance is $O(\mu^2/n)$. In problems with such very efficient estimators, removing the multiple of μ^2/n from the variance bound, as density-based control variates do, is a meaningful improvement.

Acknowlegment

This work was supported by grants DMS-0906056 and DMS-1407397 from the National Science Foundation.

References

Avramidis, A. N. and Wilson, J. R. (1993). A splitting scheme for control variates. *Operations Research Letters*, 14:187–198.

Boyd, S. and Vandeberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85.

Hesterberg, T. C. (1988). *Advances in importance sampling*. PhD thesis, Stanford University.

Hesterberg, T. C. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–192.

Lafortune, E. P. and Willems, Y. D. (1993). Bidirectional path tracing. In *Proceedings of CompuGraphics*, pages 95–104.

Marshall, A. W. (1956). The use of multi-stage sampling schemes in Monte Carlo computations. In Meyer, M. A., editor, *Symposium on Monte Carlo methods*, pages 123–40, New York. Wiley.

Owen, A. B. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Rubinstein, R. Y. and Kroese, D. P. (2004). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer, New York.

Veach, E. (1997). *Robust Monte Carlo methods for light transport simulation.* PhD thesis, Stanford University.

Veach, E. and Guibas, L. (1994). Bidirectional estimators for light transport. In *5th Annual Eurographics Workshop on Rendering*, pages 147–162.

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

7 Proofs

Theorem 1 was proved earlier by Owen and Zhou (2000). Their notation defined control variates via integrals, not expectations. We give a direct proof in the notation of the present paper to make this article self-contained.

7.1 Proof of Theorem 1

Choose $k \in \{1, 2, \dots, J\}$ and then select β with $\beta_k = 0$ and $\beta_j = -\mu\alpha_j/\alpha_k$ for $j \neq k$, giving $\sum_{\ell=1}^{J} \beta_\ell = \mu(1 - 1/\alpha_k)$. From (1) we find that

$$n\mathrm{Var}(\widetilde{\mu}_{\alpha,\beta}) = \int \frac{[fp - \mu q_\alpha - \beta^{\mathsf{T}} hp + \beta^{\mathsf{T}} \mathbf{1} q_\alpha]^2}{q_\alpha} \,\mathrm{d}\boldsymbol{x}$$

The condition on q_j yields $h_j p = q_j$, and then for this β we find that $n\mathrm{Var}(\widetilde{\mu}_{\alpha,\beta})$ is

$$\int \frac{\Big[fp - \Big(\mu\alpha_k - \sum_{\ell=1}^{J} \beta_\ell\alpha_k + \beta_k\Big)q_k - \sum_{j\neq k}\Big(\mu\alpha_j - \sum_{\ell=1}^{J} \beta_\ell\alpha_j + \beta_j\Big)q_j\Big]^2}{\sum_{j=1}^{J} \alpha_j q_j} \,\mathrm{d}\boldsymbol{x}$$
$$= \int \frac{(fp - \mu q_k)^2}{\sum_{j=1}^{J} \alpha_j q_j} \,\mathrm{d}\boldsymbol{x} \leqslant \int \frac{(fp - \mu q_k)^2}{\alpha_k q_k} \,\mathrm{d}\boldsymbol{x} = \frac{\sigma_{q_k}^2}{\alpha_k}.$$

The result follows because $\mathrm{Var}(\widetilde{\mu}_{\alpha,\beta_*}) \leqslant \mathrm{Var}(\widetilde{\mu}_{\alpha,\beta})$. □

7.2 Proof of Theorem 2

The proof is similar to that of Theorem 1. We choose $k \in \{1, 2, \dots, J\}$ and define $\beta_j = \mu(1_{j=k} - \alpha_k)$. Making this substitution leads to

$$\hat{\mu}_{\alpha,\beta} = \mu + \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i) - \mu q_k(\boldsymbol{x}_i)}{q_\alpha(\boldsymbol{x}_i)},$$

which has variance

$$\frac{1}{n}\int (fp - \mu q_k)^2/q_\alpha \,\mathrm{d}\boldsymbol{x} \leqslant \frac{1}{n\alpha_k}\int (fp - \mu q_k)^2/q_k \,\mathrm{d}\boldsymbol{x} = \sigma_{q_k}^2/n\alpha_k. \quad \square$$

7.3 Proof of Corollaries 1 and 2

The same argument underlies both proofs. We give the version for Corollary 1.

Choose $j \in \{1, \ldots, J\}$ and $\beta_j \in \mathbb{R}^J$ to minimize $\sigma^2_{q_j,\beta_j}$ and define $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \beta_j^{\mathsf{T}} h(\boldsymbol{x})$. Now we apply Theorem 1 to the integrand g and let β_* be the minimizing regression coefficient from that Theorem. Writing $g(\boldsymbol{x}) - \beta_*^{\mathsf{T}} h(\boldsymbol{x}) = f(\boldsymbol{x}) - (\beta_* - \beta_j)^{\mathsf{T}} h(\boldsymbol{x})$ we see that $\beta = \beta_* - \beta_j$ satisfies the conclusion of this Corollary. □

7.4 Proof of Lemma 1

Under the conditions of the Lemma, $\theta = \mathbf{1}$. Let $\gamma_j = \beta_j - \alpha_j \beta^{\mathsf{T}} \mathbf{1}$. Now for any $\boldsymbol{x}$ sampled from q_α, we necessarily have $q_\alpha(\boldsymbol{x}) > 0$. Then

$$\begin{aligned}
\gamma^{\mathsf{T}}(h(\boldsymbol{x}) - \mathbf{1}) p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) &= \gamma^{\mathsf{T}}(q(\boldsymbol{x}) - p(\boldsymbol{x}))/q_\alpha(\boldsymbol{x}) \\
&= \sum_{j=1}^{J} (\beta_j - \alpha_j \beta^{\mathsf{T}} \mathbf{1})(q_j(\boldsymbol{x}) - p(\boldsymbol{x}))/q_\alpha(\boldsymbol{x}) \\
&= \sum_{j=1}^{J} \beta_j q_j(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) - \beta^{\mathsf{T}} \mathbf{1} p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) - \beta^{\mathsf{T}} \mathbf{1} + \beta^{\mathsf{T}} \mathbf{1} p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) \\
&= \sum_{j=1}^{J} \beta_j q_j(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) - \beta^{\mathsf{T}} \theta,
\end{aligned}$$

after cancelling and using $\theta = \mathbf{1}$. Substituting this last expression into each term for $\hat{\mu}_{\alpha,\gamma}$ for (9) yields $\tilde{\mu}_{\alpha,\beta}$ in (8). □

7.5 Proof of Theorem 3

First we write

$$\sigma_\alpha^2 + \mu^2 = \int_D \frac{\big(f(\boldsymbol{x}) p(\boldsymbol{x})\big)^2}{\sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \tag{18}$$

where $D = \{\boldsymbol{x} \mid q_\alpha(\boldsymbol{x}) > 0\}$ is the common domain of q_α for all $\alpha \in S^0$. For fixed $\boldsymbol{x} \in D$ and $r, s \in \{1, \ldots, J\}$,

$$\frac{\partial^2}{\partial \alpha_r \partial \alpha_s} \frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2}{\sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x})} = \frac{2 f(\boldsymbol{x})^2 p(\boldsymbol{x})^2 q_r(\boldsymbol{x}) q_s(\boldsymbol{x})}{\Big(\sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x})\Big)^3}$$

and so the integrand in (18) has a positive semi-definite Hessian matrix at every $\boldsymbol{x}$. Thus the integrand is convex in α for each $\boldsymbol{x}$, and hence σ_α^2 is a convex function of α. □

7.6 Proof of Theorem 4

Similar to Theorem 3, we write

$$\sigma^2_{\alpha,\beta} + \mu^2 = \int_{D(q_\alpha)} \frac{\bigl(f(\boldsymbol{x})p(\boldsymbol{x}) - \sum_{J=1}^J \beta_j(q_j(\boldsymbol{x}) - p(\boldsymbol{x}))\bigr)^2}{\sum_{j=1}^J \alpha_j q_j(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}. \tag{19}$$

The function y^2/z is jointly convex in $z \in (0,\infty)$ and $y \in \mathbb{R}$. For fixed $\boldsymbol{x} \in D$ the integrand in (19) has the form y^2/z where y is an affine function of β and z is a positive linear function of $\alpha \in S^0$. As a result the integrand in (19) is jointly convex in (α,β) and so also is $\sigma^2_{\alpha,\beta}$. □

7.7 Proof of Theorem 5

We use Lemma 2 in our proof. It is of independent interest.

Lemma 2. *Let $\alpha,\gamma,\omega \in S$ with $\gamma = \lambda\alpha + (1-\lambda)\omega$ and $0 < \lambda \leqslant 1$. Then*

$$\min_\beta \mathrm{Var}(\tilde\mu_{\gamma,\beta}) \leqslant \frac{1}{n\lambda}\sigma^2_\alpha, \quad \textit{and}$$
$$\min_\beta \mathrm{Var}(\hat\mu_{\gamma,\beta}) \leqslant \frac{1}{n\lambda}\sigma^2_\alpha.$$

Proof. For $J = 2$, let $\widetilde{q}_1 = q_\alpha$ and $\widetilde{q}_2 = q_\omega$. Then $q_\gamma = \widetilde{\alpha}_1\widetilde{q}_1 + \widetilde{\alpha}_2\widetilde{q}_2$ for $\widetilde{\alpha} = (\lambda, 1-\lambda)$. The first part then follows from Theorem 1 using $\widetilde{\alpha}$ and $\widetilde{q}_j$ in place of α and q_j. The second part follows from Theorem 2, noting that for a nominal $\widetilde{q}_1 = q_\alpha$ the mixture weights $\widetilde{\alpha}$ include a defensive component. □

Now we prove Theorem 5 itself. First we show that for any $\alpha \in S \setminus S^\epsilon$ we can find $\delta \in S^\epsilon$ with $\mathrm{Var}(\tilde\mu_{\delta,0}) \leqslant (1+\epsilon)\mathrm{Var}(\tilde\mu_{\alpha,0})$. Recall that $K = \#\{1 \leqslant j \leqslant J \mid \alpha_j < \epsilon\}$. We have $1 \leqslant K \leqslant J-1$; the lower limit comes from $\alpha \notin S^\epsilon$ and the upper limit from $\epsilon < 1/d$.

Now define $\omega \in S^\epsilon$ with $\omega_j = \epsilon$ for $\alpha_j \geqslant \epsilon$ and $\omega_j = \Omega \equiv (1-\epsilon(J-K))/K$ for $\alpha_j < \epsilon$. Define $\lfloor\alpha\rfloor \equiv \min_j \alpha_j$ and then set $\lambda = (\Omega-\epsilon)/(\Omega - \lfloor\alpha\rfloor)$ and $\delta_j = \lambda\alpha_j + (1-\lambda)\omega_j$. By construction $0 < \lambda < 1$. If $\alpha_j \geqslant \epsilon$ then $\delta_j \geqslant \min(\alpha_j,\omega_j) = \epsilon$. Otherwise

$$\delta_j \geqslant \frac{\Omega-\epsilon}{\Omega-\lfloor\alpha\rfloor}\lfloor\alpha\rfloor + \frac{\epsilon-\lfloor\alpha\rfloor}{\Omega-\lfloor\alpha\rfloor}\Omega = \epsilon.$$

It follows that $\delta \in S^\epsilon$. Now from Lemma 2 we find

$$\mathrm{Var}(\tilde\mu_{\delta,0}) \leqslant \frac{1}{\lambda}\mathrm{Var}(\tilde\mu_{\alpha,0})$$

and

$$\frac{1}{\lambda} = \frac{\Omega-\lfloor\alpha\rfloor}{\Omega-\epsilon} \leqslant \frac{\Omega}{\Omega-\epsilon} = \frac{1-\epsilon(J-K)}{1-\epsilon(J-K)-K\epsilon} = \frac{1-\epsilon J+\epsilon K}{1-\epsilon J}.$$

We can extend this result to the $\mathrm{Var}(\tilde\mu_{\alpha,\beta})$ by the same device used in Corollary 1. Then because $\alpha_1 > 0$ when $\alpha \in S^\epsilon$, a defensive component is present, and the results then apply to $\mathrm{Var}(\hat\mu_{\alpha,\beta})$. □