

Adaptive Importance Sampling by Mixtures of Products of Beta Distributions

Art Owen
Stanford University

Yi Zhou
Goldman-Sachs

Original: January 1998
This version: October 1999

Abstract

The problem of numerically integrating spiky functions over high dimensional domains arises in computational statistics, particle physics, computer graphics and computational finance. We propose a Monte Carlo method based on adaptive importance sampling. Adaptive importance sampling methods alternate between importance sampling from a density constructed to suit the integrand, and updating the sampling density with the newly sampled data. We present a method in which the sampling density is a mixture of products of beta distributions.

1 Introduction

The problem we consider is the computation of the integral

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

Many integration 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. Monte Carlo simulations that use d independent $U(0,1)$ random variables per realization can be written as (1), even if they arise as an expectation under another distribution.

We are most interested in cases where d is moderate to large, and f is spiky. Our motivating examples come from particle physics and Bayesian statistics. Spiky integrands also arise in computational finance (Glasserman,

Heidelberger & Shahabuddin 1999, Owen & Zhou 1999), computer graphics (Veach & Guibas 1995) and reliability (Hesterberg 1995).

For $d = 1$ and smooth f , classical methods, such as those in Davis & Rabinowitz (1984), provide excellent accuracy with only a handful of function evaluations. For small d , tensor products of 1 dimensional rules work very well on smooth functions. For large d , Monte Carlo, and more recently, quasi-Monte Carlo methods Niederreiter (1992) have been extensively developed. These work by sampling the unit cube more uniformly than random points do. Some quasi-Monte Carlo rules can be randomized (Owen 1997a) to provide error estimates, and this can even improve the rate of convergence (Owen 1997b, Hickernell 1996). But, for integrands with sharply localized features, it is simply wasteful to sample the whole domain uniformly, as quasi-Monte Carlo methods do.

Spiky integrals over moderate to high dimensions are usually handled using importance sampling. For the integral (1), effective densities are nearly proportional to $|f|$. See Owen & Zhou (2000) for a survey and some extensions of importance sampling, for the case where one knows the approximate locations and shapes of the spikes in f .

This paper considers the more challenging problem of using importance sampling on f , when the number and locations of the spikes are unknown. The proposed method alternates between constructing a density that is approximately proportional to the integrand, and sampling from the most recent approximation. Our approximations are formed as mixtures of products of beta densities.

An outline of the paper is as follows. Section 2 gives a brief outline of the related literature, focussing on methods for multiple spikes. Section 3 briefly reviews importance sampling, including some recent improvements in it. Section 4 discusses adaptive importance sampling. It shows how to combine the estimates from each stage into estimates \hat{I} of I , and $\hat{V}(\hat{I})$ of $V(\hat{I})$. The combination avoids introducing sampling biases, in spite of the fact that the sampling distribution used at the $k + 1$ 'st stage depends on the data from the previous k stages. That section also presents a square root rule, which provides a nearly optimal weighting of the data from the different steps of the algorithm, under widely different assumptions on the effectiveness of the adaptation. Section 5 presents our mixtures of products of beta distribution model, and describes how we estimate the parameters in it. Section 6 gives some numerical examples and comparisons to other methods. Section 7 presents our conclusions. The Appendix proves Theorem 4 on the near optimality of the square root rule given in Section 4.3.

2 Literature survey

Here we present a brief survey of related work on integrating spiky functions. A more comprehensive survey is given in Zhou (1998). First we present those methods used in the numerical comparisons in Section 6. Then we outline some additional methods used in Bayesian calculations.

2.1 Comparison methods

The methods presented here are used in the numerical examples of Section 6. They allow the integrand to have multiple spikes, and code is readily available.

The VEGAS method (Lepage 1978) is widely used in high energy physics problems. It combines importance sampling from a product of piecewise constant density functions with stratification. A difficulty with this algorithm arises when two or more of the factors in the product density are multimodal. Then a number of spurious modes can be produced in the product that do not correspond to actual modes in the integrand.

The MISER method (Press, Teukolsky, Vetterling & Flannery 1992) estimates the average value of a function over a rectangular region. It is a recursive method, with each step splitting the region at hand orthogonally to one coordinate, into two equal, or nearly equal, subregions. Each split is made using the coordinate that, in a pilot sample, appears to most concentrate the variation in the function on one side of the split. The budgetted sample size for the region is then allocated between the two sides, with more observations taken from the side that has the greater variation. At the finest level of recursion, simple Monte Carlo sampling is used to estimate the average function value and its variance. These estimates are then propagated to the original interval. See (Press et al. 1992) for the details, especially those on variance estimation and propagation, which are based on some empirically derived heuristics.

The ADBAYES method (Berntsen, Espelid & Genz 1991) also makes adaptive splits of the integration region. The code is available from the web page of Alan Genz at Washington State University in Pullman. It applies deterministic integration rules, makes the splits in the direction with the largest fourth divided difference, and combines the errors by propagating deterministic error bounds. It is geared specifically to problems of estimating multiple integrals over the same domain from the same sample.

2.2 Methods for Bayesian calculations

There is a large literature on importance sampling for integration of spiky functions arising in Bayesian calculations. Surveys of this literature appear in Evans & Swartz (1995) and Zhou (1998).

It is common for the posterior distribution to be sharply concentrated around its mode, and this spike can easily be sharp enough that posterior moments are also spiky. Most of that literature focusses on the case of an integrand with a single spike. It is common to use distributions based on the t distribution, an early example being Kloek & van Dijk (1978).

A notable exception is Oh & Berger (1993), who employ a mixture of t distributions. They suppose that the number of spikes is known, as well as their locations and the Hessians of the integrand there. The initial parameters in their mixture model are chosen using these Hessians.

Much of the work on Bayesian importance sampling uses parametric families, especially those based on the t distribution. The univariate split t distribution can have different scale and degrees of freedom on each side of its mode. There are also multivariate t and split- t distributions in use. See Evans & Swartz (1995) for references.

Because our work uses an adaptive mixture to approximate the integrand, it is more like a nonparametric importance sampling technique. In this it is similar to Zhang (1996) who uses a kernel method. Zhang (1996)'s kernels are radially symmetric and there is one kernel function situated at each of n sample points. West (1992), West (1993) and Givens & Raftery (1996) consider clustering techniques that reduce the number of mixture components.

3 Importance sampling

This section provides a review of importance sampling, emphasizing those techniques we use later. The term “importance sampling” derives from the idea that it pays to take more sample points in the region most important to the target function.

3.1 Variance, optimality and failure

Let p be a density on $(0, 1)^d$, with $p(x) > 0$ everywhere. The integral in (1) can be written

$$I = \int \frac{f(x)}{p(x)} p(x) dx. \quad (2)$$

Then for X_i drawn independently from $p(x)$ we may estimate I by

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

Elementary manipulations give that

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

An unbiased estimate of $V(\hat{I}^{(p)})$ is

$$\hat{V}(\hat{I}^{(p)}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{f(X_i)}{p(X_i)} - \hat{I}^{(p)} \right)^2. \quad (5)$$

Let $p_o(x)$ be the optimal density, taken to be the minimizer of the variance in (4). Kahn & Marshall (1953) show that $p_o(x) \propto |f(x)|$. If $f \geq 0$ everywhere then, as is well known, $p_o = f(x)/I$ and this gives $V(\hat{I}^{(p_o)}) = 0$.

To actually use this optimal density p_o when $f \geq 0$, requires that we know I . The practical value of the finding that $p_o \propto f$ is that it guides us in finding good importance sampling densities. A good importance sampling density will be roughly proportional to f .

Less well known is that if p is nearly but not exactly proportional to f , the variance can be disastrously large. Owen & Zhou (2000) present an example where p is visually indistinguishable from a multiple of f , but yields an infinite variance because p decreases more quickly to zero than does f away from the spike. The irony is that a failure in the seemingly unimportant part of the domain can produce a disastrous increase in the variance.

Owen & Zhou (2000) show how to prevent the disastrous failure of importance sampling, by employing some control variates and mixture sampling. The method extends the defensive importance sampling method of Hesterberg (1995).

3.2 Control variates

The method of control variates (see Ripley (1987), Bratley, Fox & Schrage (1987), or Hammersley & Handscomb (1964)) exploits known values of one or more integrals to improve the estimation of I . This section considers control variates in conjunction with importance sampling.

Suppose that we know $\int h_j(x)dx = \mu_j$, $j = 1, \dots, m$. We assume that $p(x) > 0$ if any $h_j(x) > 0$, or if $f(x) > 0$. Let $\beta = (\beta_1, \dots, \beta_m)$ be a vector of real values. Under independent sampling of X_i from $p(x)$,

$$\hat{I}_{p,\beta} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i) - \sum_{j=1}^m \beta_j h_j(X_i)}{p(X_i)} + \sum_{j=1}^m \beta_j \mu_j \quad (6)$$

is an unbiased estimate of I .

The variance of $\hat{I}_{p,\beta}$ is $\sigma_{p,\beta}^2/n$, where

$$\sigma_{p,\beta}^2 = \int \left(\frac{f(x) - \sum_j \beta_j h_j(x)}{p(x)} - I + \sum_j \beta_j \mu_j \right)^2 p(x) dx. \quad (7)$$

Let β^* minimize the integral in (7), over β , for the given functions f and p . Equation (7) suggests that an estimate $\hat{\beta}$ of β^* can be found by a multiple regression (including an intercept term) of $f(X_i)/p(X_i)$ on predictors $h_j(X_i)/p(X_i)$.

Because the integral includes an intercept coefficient $\hat{\beta}_0$, the residuals will sum to zero. As a result equation (6) with $\beta = \hat{\beta}$ simplifies to

$$\hat{I}_{p,\hat{\beta}} = \hat{\beta}_0 + \sum_{j=1}^m \hat{\beta}_j \mu_j.$$

For uniform sampling over $[0, 1]^d$, Theorem 1 of Owen & Zhou (2000) simplifies to:

Theorem 1 *Suppose that there is a unique vector β^* that minimizes $\sigma_{p,\beta}^2$, and let $\hat{\beta}$ be determined by least squares as described above. Suppose further that the expectations under sampling from p of h_j^4/p^4 and $h_j^2 f^2/p^4$ exist and are finite, for $j = 1, \dots, m$. Then*

$$\hat{\beta}_j = \beta_j^* + O_p(n^{-1/2}), \quad (8)$$

for $j = 1, \dots, m$, and

$$\hat{I}_{p,\hat{\beta}} = \hat{I}_{p,\beta^*} + O_p(n^{-1}). \quad (9)$$

3.3 Mixture sampling

When importance sampling fails disastrously, it is usually because $p(x)$ is too small in some places, despite matching $f(x)$ well near the peaks. Hesterberg (1995) proposed a remedy called defensive mixture sampling. The sampling density $p(x)$ is taken to be a mixture where one component density is nearly proportional to $f(x)$ near the peak of $f(x)$. Another mixture component, such as the $U(0, 1)^d$ distribution, prevents the sampling density from getting too small anywhere.

More generally, it can be advantageous to sample from a mixture

$$p_\alpha(x) = \sum_{j=1}^m \alpha_j p_j(x) \quad (10)$$

where $\alpha_j > 0$, $\sum_{j=1}^m \alpha_j = 1$ and the p_j are densities. Each p_j might be optimized for one particular spiky feature of f , or perhaps several integrands f are being integrated from the same sample, in which case each of the different p_j may be designed to work for some of the integrands.

Owen & Zhou (2000) present the following estimator

$$\tilde{I}_{\alpha, \delta} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i) + \sum_{j=1}^m \delta_j p_j(X_i)}{p_\alpha(X_i)} - \sum_{j=1}^m \delta_j \quad (11)$$

where δ_j are scalar coefficients and X_i are iid from p_α . Under sampling from p_α the expected value of p_j/p_α is known to be 1. Thus the estimator in (11) combines importance sampling with the use of control variates.

These control variate coefficients can also be estimated by regression of f/p_α on $-p_j/p_\alpha$ would be appropriate. Because $\sum_j p_j/p_\alpha = 1$, this regression is singular. Owen & Zhou (2000) recommend the use of a singular value decomposition to compute the least squares coefficients.

Theorem 2 of Owen & Zhou (2000) is:

Theorem 2 *Let p_j , α_j and $\tilde{I}_{\alpha, \delta}$ be as above. Then*

$$E(\tilde{I}_{\alpha, \delta}) = I, \quad (12)$$

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

$$V(\tilde{I}_{\alpha, \delta^o}) \leq \min_{1 \leq j \leq m} \frac{1}{\alpha_j n} \left[\int \frac{f(x)^2}{p_j(x)} dx - I^2 \right]. \quad (13)$$

Theorem 2 shows that the combination $\tilde{I}_{\alpha,\delta}$ is never much worse than what would be obtained from any of the individual importance sampling distributions. Furthermore regression can be used to estimate this combination. The conclusion in Theorem 2 would not hold without the use of the control variates. The variance ratios between good and bad mixture components can be very large or even infinite, so the factors $1/\alpha_j$ are usually small by comparison.

The mixture sample estimator in equation (11) can be improved by taking a deterministic mixture instead of a random one. Suppose that $n_j = n\alpha_j$ is a positive integer. In practice we might have to settle for an integer close to $n\alpha_j$. The estimate

$$\hat{I}_{\alpha,\delta} = \frac{1}{n} \sum_{r=1}^m \sum_{i=1}^{n_r} \frac{f(X_{ri}) + \sum_{r=1}^m \delta_r p_r(X_{ri})}{p_\alpha(X_{ri})} - \sum_{r=1}^m \delta_r \quad (14)$$

is based on a deterministic mixture sample, with $X_{ri} \sim p_r$, $i = 1, \dots, n_r$, $r = 1, \dots, m$. It has variance no larger than the estimate $\tilde{I}_{\alpha,\delta}$ based on the random mixture sample (Hesterberg 1988).

3.4 Positivation

It is possible to obtain zero variance on general integrands, using a positivation device from Owen & Zhou (2000). 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:

$$\hat{I}_\pm = \frac{1}{n_+} \sum_{i=1}^{n_+} \frac{f_+(X_{i,+})}{p_+(X_{i,+})} - \frac{1}{n_-} \sum_{i=1}^{n_-} \frac{f_-(X_{i,-})}{p_-(X_{i,-})}. \quad (15)$$

The practical importance of (15), is that one can integrate arbitrarily well by using two importance sampling densities. One is nearly proportional to the positive part of f and one to the negative part. There is no computational difficulty in finding the positive or negative parts of the integrand.

This positivation device can be generalized. Owen & Zhou (2000) investigate integration of $(f(x) - g(x))_\pm$ by mixture importance sampling with control variates, where $g(x)$ has a known integral.

4 Adaptive importance sampling

4.1 Iteration

Our proposal is to alternate between importance sampling from a density, and re-computing the importance sampling density. We begin by considering algorithms that take a fixed number $K \geq 1$ of steps.

To begin step $k \geq 1$, we draw n_k observations $\mathcal{X}^{(k)} = (X_{k1}, \dots, X_{kn_k})$ using a rule \mathcal{R}_k that specifies the joint distribution of the X_{ki} . What we have in mind is that n_k is a fixed sample size and that $\mathcal{X}^{(k)}$ is a deterministic mixture importance sample, taking $n_{kj} = n\alpha_{kj}$ observations from the density p_{kj} , for $j = 1, \dots, m_k$. Much of what we prove below can also be extended with minimal changes to allow for random sample sizes n_k .

To conclude step k , we use observations $\mathcal{X}^{(k)}$ and possibly all the previous observations, to construct the rule \mathcal{R}_{k+1} for the next sample. What we have in mind for this step is the recomputation of the number of beta mixture components, their parameters and the mixing probabilities.

In our analysis of the algorithm, we suppose that at step k we can construct a conditionally unbiased estimate \hat{I}_k , in which the only random quantities are X_{ki} , $i = 1, \dots, n_k$ and possibly n_k itself. We also assume that we can accompany this estimate by a conditionally unbiased variance estimate $\hat{V}(I_k)$. Our unbiasedness conditions are:

$$E(\hat{I}_k | \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)}) = I, \quad (16)$$

$$E(\hat{V}(I_k) | \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)}) = V(\hat{I}_k | \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)}). \quad (17)$$

Our algorithm goes through K of these steps and then upon termination it returns the weighted estimate $\hat{I} = \sum_{k=1}^K \omega_k \hat{I}_k$ and the variance estimate $\hat{V}(\hat{I}) = \sum_{k=1}^K \omega_k^2 \hat{V}(\hat{I}_k)$. The weights ω_k are supplied to the algorithm and they must satisfy $\sum_{k=1}^K \omega_k = 1$. The entire procedure is summarized in Algorithm 1.

In practice we may tolerate a small bias, instead of imposing (16) and (17), when we are confident that such bias is negligible. The bias we have in mind is that arising from control variate coefficients estimated from the data. If one insists on eliminating even this small bias, it can be done through the method of random groups (Wolter 1985), a form of cross-validation.

4.2 Moments of the estimates

Because each step is using conditionally unbiased estimates, it follows that the quantities \hat{I} and $\hat{V}(\hat{I})$ are unbiased estimates of I and $V(\hat{I})$ respectively,

Fixed- K -AIS:

Given: f , $K \geq 1$, \mathcal{R}_1 , and $\omega_1, \dots, \omega_K$ with $\sum_{k=1}^K \omega_k = 1$

for $1 \leq k \leq K$ **do**

 Use \mathcal{R}_k to select n_k and $\mathcal{X}^{(k)} = (X_{1k}, \dots, X_{n_k})$

 Compute \hat{I}_k and $\hat{V}(\hat{I}_k)$ as functions of $\mathcal{X}^{(k)}$, satisfying (16) and (17).

 Construct \mathcal{R}_{k+1} as a function of $\mathcal{Z}^{(k)} = (\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k)})$

end for

Set $\hat{I} = \sum_{k=1}^K \omega_k \hat{I}_k$, $\hat{V}(\hat{I}) = \sum_{k=1}^K \omega_k^2 \hat{V}(\hat{I}_k)$

Deliver: \hat{I} , $\hat{V}(\hat{I})$, $\hat{I}_1, \dots, \hat{I}_K$, $\hat{V}(\hat{I}_1), \dots, \hat{V}(\hat{I}_K)$, $\mathcal{Z}^{(K)}$

Algorithm 1: This is a generic algorithm for adaptive importance sampling with a prespecified sample size. It produces unbiased estimates $E(\hat{I}) = I$ and $E(\hat{V}(\hat{I})) = V(\hat{I})$.

as stated in Theorem 3.

Theorem 3 *The estimates produced by Algorithm 1 are unbiased in that $E(\hat{I}) = I$ and $E(\hat{V}(\hat{I})) = V(\hat{I})$.*

Proof: By construction,

$$E(\hat{I}_k) = E\left(E(\hat{I}_k \mid \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)})\right) = I,$$

and so

$$E(\hat{I}) = \sum_{k=1}^K \omega_k I = I.$$

Now

$$V(\hat{I}) = \sum_{k=1}^K \sum_{l=1}^K \omega_k \omega_l \text{Cov}(\hat{I}_k, \hat{I}_l).$$

For $k < l$,

$$\text{Cov}(\hat{I}_k, \hat{I}_l) = E\left((\hat{I}_k - I)E(\hat{I}_l - I \mid \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k)})\right) = 0$$

and so it follows that

$$V(\hat{I}) = \sum_{k=1}^K \omega_k^2 V(\hat{I}_k). \quad (18)$$

Finally

$$\begin{aligned}
E(\hat{V}(\hat{I}_k)) &= E\left(E(\hat{V}(\hat{I}_k) \mid \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)})\right) \\
&= E\left(V(\hat{I}_k \mid \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)})\right) \\
&= V(\hat{I}_k) - V\left(E(\hat{I}_k \mid \mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k-1)})\right) \\
&= V(\hat{I}_k),
\end{aligned}$$

and so

$$E(\hat{V}(\hat{I})) = \sum_{k=1}^K \omega_k^2 E(\hat{V}(\hat{I}_k)) = \sum_{k=1}^K \omega_k^2 V(\hat{I}_k) = V(\hat{I}). \quad \square$$

The values of $Q_k = \sum_{r=1}^k \omega_k(\hat{I}_r - I)$ form a martingale (Williams 1991) with respect to the sigma fields defined by $\mathcal{Z}^{(k)} = (\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(k)})$. We make no formal use of martingales, but the ideas underly our methods.

4.3 Square root rule for ω_k

This section presents a simple rule for choosing the values ω_k . We propose taking $\omega_k \propto \sqrt{k}$, in the setting where all $n_k = n$. This rule is nearly optimal under some quite different assumptions on $V(\hat{I}_k)$.

If we knew $V(\hat{I}_k)$, then from equation (18) we could use

$$\tilde{\omega}_k = \frac{V(\hat{I}_k)^{-1}}{\sum_{l=1}^K V(\hat{I}_l)^{-1}}. \quad (19)$$

This choice of ω_k is optimal in that it minimizes $V(\hat{I})$.

The following model describes steady but moderate improvement as the adaptive sampling progresses. Suppose that as k increases that $V(\hat{I}_k) = \sigma^2 k^{-r_0}/n$, where σ^2 is a constant and $0 \leq r_0 \leq 1$. The lower bound $r_0 = 0$ describes a setting in which the iterations are not improving. The upper bound $r_0 = 1$ describes a setting in which quasi-Monte Carlo accuracy is being attained by adaptive Monte Carlo.

In our example, the optimal ω_k are, by (19), proportional to k^{r_0} . Suppose however that we guess incorrectly and use ω_k proportional to k^{r_1} with $0 \leq r_1 \leq 1$. Let us call the variance we get $V_{r_0, K, r_1}(\hat{I})$, so

$$V_{r_0, K, r_1}(\hat{I}) = \frac{\sigma^2 \sum_{k=1}^K k^{2r_1 - r_0}}{n \left(\sum_{k=1}^K k^{r_1}\right)^2}.$$

Theorem 4 *In the notation above,*

$$\min_{0 \leq r_1 \leq 1} \sup_{1 \leq K < \infty} \max_{0 \leq r_0 \leq 1} \frac{V_{r_0, K, r_1}(\hat{I})}{V_{r_0, K, r_0}(\hat{I})} = \frac{9}{8}, \quad (20)$$

and this minimum is attained at $r_1 = 0.5$, so that

$$\sup_{1 \leq K < \infty} \max_{0 \leq r_0 \leq 1} \frac{V_{r_0, K, 0.5}(\hat{I})}{V_{r_0, K, r_0}(\hat{I})} = \frac{9}{8}. \quad (21)$$

Proof: See Section 7. \square

Theorem 4 shows that under the square root rule, the variance is not more than $9/8 = 1.125$ times the variance of the unknown optimal rule, for any integer $K \geq 1$, when $V(\hat{I}_k)$ is proportional to k^{r_0} for some $r_0 \in [0, 1]$.

Even though the achievable variance has different rates of convergence, depending on r_0 , the square root rule always attains the optimal rate, and nearly the optimal constant. Indeed as long as $r_1 \geq r_0/2 \geq 0$, taking $\omega_k = k^{r_1} / \sum_{k=1}^K k^{r_1}$, achieves the optimal rate as $K \rightarrow \infty$. Thus if one thinks that $0 \leq r_0 \leq R$ describes the reasonable values for r_0 , then the square root rule should be replaced by one with $r_1 = R/2$.

For those cases in which one suspects that the method will have a sharp initial transient, we propose a fixed sequence of weights as follows. For $K \leq 3$ take $\omega_k = 1_{k=K}$, for $K > 3$ take $\omega_k \propto \sqrt{k} 1_{k > \sqrt{K}}$.

4.4 Sample dependent weights

Because the consequences of using incorrect fixed weights are so minimal, we prefer using a fixed set of weights to estimating a set of weights from the data. For example, taking $\omega_k \propto 1/\hat{V}(\hat{I}_k)$ makes the weights random and hence more complicated to analyze. Even if $V(\hat{I}_k)$ decreases on a smooth k^{-r_0} trajectory, the values $\hat{V}(\hat{I}_k)$ could be very noisy.

There is also the danger that a sample $\mathcal{X}^{(k)}$ which happens to miss one of the spikes in f will have a misleadingly small $\hat{V}(\hat{I}_k)$ and therefore get a particularly large weight ω_k distorting the final estimate \hat{I} . We consider this kind of bias to have much greater potential for damage than that arising from estimated control variate coefficients.

4.5 Sample dependent stopping

We have assumed that the values of K and $\omega_1, \dots, \omega_K$ are specified before sampling begins. It would be desirable to sample a random number K of

steps, stopping at the point where sufficient accuracy has been obtained, as judged by the sample variances. Unfortunately, it is easy to introduce a very large bias this way.

The following very simple example illustrates the problem. We take $d = 1$, $f(x) = 1_{0 < x < \epsilon}$ for some $0 < \epsilon < 1$. The first sample consists solely of $X_{11} \sim U(0, 1)$. If $f(X_{11}) = 0$, we stop at $K = 1$ and report $\hat{I} = 0$. If $f(X_{11}) = 1$, we continue to $K = 2$, sample $X_{21} \sim U(0, 1)$ independently of X_{11} , and report $\hat{I} = (f(X_{11}) + f(X_{21}))/2$. It is easy to see that $\int_0^1 f(x)dx = \epsilon$ but that $E(\hat{I}) = (\epsilon + \epsilon^2)/2 < \epsilon$.

For a spiky nonnegative integrand, stopping with an unusually small sample variance can be much the same as stopping with an unusually small sample mean. The example above could be reformulated with $n \geq 2$ and the decision to stop based on \hat{V}_{k_1} , without essential change.

In Section 4.2 we mentioned that $Q_k = \sum_{r=1}^k \omega_k(\hat{I}_r - I)$ form a martingale. This might lead us to believe that we can stop at k steps with an unbiased estimator. But the natural estimator $\sum_{r=1}^k \omega_k \hat{I}_r / \sum_{r=1}^k \omega_k$ does not necessarily form a martingale.

It is possible to use a limited form of sample dependent stopping. In this setting one begins with a possibly random number K_1 of steps. At the end of these steps one determines a value K_2 as well as $\omega_{K_1+1}, \dots, \omega_{K_1+K_2}$ summing to 1, samples for K_2 more steps, and combines the estimates as in Algorithm 1 using $\omega_k = 0$ for $k \leq K_1$. This procedure is similar to the practice in Markov Chain Monte Carlo methods of using a burn-in period.

5 Mixtures of products of betas

5.1 Mixture model

Our plan is to alternate between sampling from a density p that approximates the integrand f , and using the sample to find a new approximation. This plan requires a family of densities that can be sampled from, that provide a flexible set of approximations, and in which optimization is not too difficult.

We suppose that $f \geq 0$. This is no loss of generality. When f takes both signs, we treat f_{\pm} separately, as at equation (15), or as Owen & Zhou (2000) show, we can work with $(f - g)_{\pm}$ where g is a function with a known integral.

We begin with some notation. The point $x \in (0, 1)^d$ is written as $x = (x^1, \dots, x^d)$ where $0 < x^j < 1$. The function $b(z, \alpha, \beta) = z^{\alpha-1}(1-z)^{\beta-1}$, on $0 < z < 1$, is an unnormalized beta density, with parameters $\alpha > 0$ and

$\beta > 0$. The normalizing constant is $B(\alpha, \beta) = \int_0^1 b(z, \alpha, \beta) dz = B(\alpha, \beta) = \Gamma(\alpha + \beta) / (\Gamma(\alpha)\Gamma(\beta))$.

We choose to approximate the integrand by

$$p(x) = \sum_{m=0}^M \gamma_m \prod_{j=1}^d b(x^j, \alpha_{mj}, \beta_{mj}), \quad (22)$$

with $\gamma_m > 0$. This is a mixture of products of beta densities. Sampling from mixtures is straightforward, as is sampling from products.

We have mixed unnormalized densities. The m 'th mixture probability can be recovered as $\gamma_m / \prod_{j=1}^d B(\alpha_{mj}, \beta_{mj})$. Unnormalized densities are easier to differentiate with respect to their parameters than are the normalized ones, and we have found this helps when optimizing the choice of α_j and β_j .

As a reference point, taking $\alpha_{mj} = \beta_{mj} = 1$ produces the $U(0, 1)$ distribution. For defensive importance sampling we always include a $U(0, 1)^d$ mixture component, by taking $\alpha_{0j} = \beta_{0j} = 1$ for $j = 1, \dots, d$.

For large α and β , the beta distribution is approximately $N(\mu, \sigma^2)$ with $\mu = \alpha / (\alpha + \beta)$ and $\sigma^2 = \mu(1 - \mu) / (\alpha + \beta + 1)$. Thus, at least for large M , mixtures of products of betas can be as flexible as mixtures of products Gaussian densities.

The model in (22) takes the same form as the Π model of Breiman (1991). Breiman uses sums of products of smooth functions as an approximation for nonparametric regression. Our application has the additional constraint that the approximation must be a density from which we can draw a sample.

5.2 Mixture fitting

Here we describe our algorithm for picking the parameters in the mixture of products of beta densities model (22). More details are available in Zhou (1998). An algorithm of this complexity necessarily involves a number of pragmatic choices that may not be theoretically optimal.

We seek parameter values M , γ_m , α_{mj} , and β_{mj} for $m = 1, \dots, M$ and $j = 1, \dots, d$ to use in (22). We formulate the problem as one of minimizing

$$\int_{(0,1)^d} \left(f(x) - \sum_{m=1}^M \gamma_m \prod_{j=1}^d b(x^j, \alpha_{mj}, \beta_{mj}) \right)^2 dx.$$

This is numerically more convenient than attempting to minimize a direct measure of the sampling variance. The defensive mixture component with $m = 0$ and $\alpha_{0j} = \beta_{0j} = 1$ is incorporated separately.

We approximate this integral by a sum over the existing data available from steps 1 through k . We suppose that step r was a deterministic mixture sample of n_r observations $Y_{ri} = f(X_{ri})$, $i = 1, \dots, n_r$, $r = 1, \dots, k$. The sampling density is denoted by $p^{(r)}$. We usually take all $n_r = n$ as described in Section 4.3.

The sum of squared errors we work with is:

$$\frac{1}{k} \sum_{r=1}^k \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{1}{p^{(r)}(X_{ri})} \left(Y_{ri} - \sum_{m=1}^M \gamma_m \prod_{j=1}^d b(X_{ri}^j, \alpha_{mj}, \beta_{mj}) \right)^2. \quad (23)$$

Dividing by $p^{(r)}$ compensates for the sampling bias when point X_{ri} was generated.

Our algorithm is based on the Levenberg-Marquardt nonlinear least squares method applied to (23). That algorithm is appropriate for finding γ_m , α_{mj} , and β_{mj} given M . We make modifications in order to find M and to protect against numerical and sampling difficulty, as described below.

We use the constraint $\gamma_m \geq 0$. This rules out negative densities.

We add the constraint that, for each mj pair, either $\alpha_{mj} \geq 1$ or $\beta_{mj} \geq 1$. This keeps the terms in the product unimodal. This constraint might be a disadvantage in some settings, but we use it to avoid introducing a product of bimodal densities. Such a product might yield as many as 2^d modes, which we consider undesirable.

We impose an upper bound on $\alpha_{mj} + \beta_{mj}$ in order to prevent the densities from becoming too concentrated to early. A premature concentration of the densities could cause a spike in the integrand to be missed. Because the variance of a Beta(α, β) random variable is $\mu(1 - \mu)/(\alpha + \beta + 1)$ where $\mu = \alpha/(\alpha + \beta + 1)$, an upper bound on $\alpha_{mj} + \beta_{mj}$ controls the concentration. We use an upper limit of U_k increasing linearly from $U_1 = 30$ to $U_K = 300$.

We impose lower bounds $\alpha_{mj} \geq L_a = 0.05$ and $\beta_{mj} \geq L_b = 0.4$, for numerical reasons. With $\beta_{mj} < 0.4$ we get too many observations X_{ri} that should be between $1 - \epsilon$ and 1 but are generated equal to 1. Here ϵ denotes the machine epsilon. The beta density evaluates to $+\infty$ at such points. The constraint on α_{mj} is less severe because the floating point numbers are spaced more closely together near 0 than they are near 1.

Our procedure for increasing M to $M + 1$ and finding the γ_{M+1} , α_{M+1j} and β_{M+1j} begins by writing

$$\tilde{Y}_{ri} = Y_{ri} - \sum_{m=1}^M \gamma_m \prod_{j=1}^d b(X_{ri}^j, \alpha_{mj}, \beta_{mj}).$$

We then seek to minimize

$$\frac{1}{k} \sum_{r=1}^k \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{1}{p^{(r)}(X_{ri})} \left(\tilde{Y}_{ri} - \gamma_{M+1} \prod_{j=1}^d b(X_{ri}^j, \alpha_{M+1j}, \beta_{M+1j}) \right)^2. \quad (24)$$

To get starting values for this minimization we find the largest \tilde{Y}_{ri} . Suppose it is at $r = r^*$ and $i = i^*$. Then we pick α_{M+1j} and β_{M+1j} corresponding to the beta distribution with mean $X_{r^*i^*}^j$ and the smallest variance allowable under our constraints. The starting value for γ_{M+1} is the minimizer of (24) given the starting values for α_{M+1j} and β_{M+1j} .

If the starting value for γ_{M+1} is negative, we take this as an indication that the model has enough mixture components, for the amount of data at hand, and we do not increase M . Otherwise, we run Levenberg-Marquardt on (24) and then do a round of backfitting in which all parameters are reestimated, and consider increasing M again. To reduce the risk of overfitting, we did not allow M to increase by more than 3 at any iteration of adaptive importance sampling.

Some other details of the algorithm appear in Zhou (1998). These describe how the constraints were incorporated into the Levenberg-Marquardt algorithm, how the necessary derivatives were formed, and what techniques used to reduce the risk of numerical overflow and underflow.

6 Examples

For our examples, we compare the mixture of products of beta method with ordinary Monte Carlo, and some adaptive methods. The adaptive methods chosen were VEGAS (Lepage 1978), MISER (Press et al. 1992) and ADBAYES (Berntsen et al. 1991). These three methods, described in Section 2.1 can all handle multiple spikes and code was readily available.

6.1 Double gaussian integral

The following integrand,

$$\frac{1}{2} \left(\frac{1}{a\pi^{1/2}} \right)^{-d} \left[\exp \left(- \sum_{j=1}^d \left(\frac{x^j - 1/3}{a} \right)^2 \right) + \exp \left(- \sum_{j=1}^d \left(\frac{x^j - 2/3}{a} \right)^2 \right) \right]$$

with $a = 0.1$, is taken from Lepage (1978). The integral is $I = P(0 \leq N(1/3, a^2/2) \leq 1)^d$.

	MC	VEGAS	MISER	ADB	MPB
Err	0.20	-0.4996	-0.017	-0.060	0.0008
$\widehat{\text{Err}}$	0.21	0.0005	0.015	0.265	0.0007

Table 1: Results from the double Gaussian integrand. The first row gives the error $\hat{I} - I$ of each method. The second row gives the error estimate. For all but ADB, this is an estimated standard deviation. The true value of I is almost exactly 1.

We chose $d = 9$ and this gives $I \doteq 0.999989$. This case is hard for Vegas, because the optimal density in the class it uses is a product of 9 bimodal densities having $2^9 = 512$ modes. Because the integrand has only 2 modes, there is a reasonable chance that one of the genuine modes will get very few sample points and then be missed when forming the next product of importance sampling densities. Once a mode is lost, the estimate will concentrate on the other mode and converge to a value near 0.5.

The methods used all had $K = 15$ stages and $n = 10^5$ data points at each stage. Each method generated an internal error estimate. Except for ADBAYES, this was a standard error. ADBAYES uses a deterministic set of points and the error estimate is a deterministic estimate of its absolute error.

Table 1 presents the results. Vegas was seriously inaccurate on this problem. It missed one of the spikes, and the internal error estimate did not capture this. Ordinary Monte Carlo was not very accurate but had a realistic error estimate.

MISER, ADBAYES and MPB all gave realistic error estimates. Of these MPB was by far the most accurate. Because a Gaussian density can be well approximated by a beta density, it is perhaps not surprising that MPB did well on this problem.

6.2 Particle physics integral

This integrand is over $(0, 1)^7$. It describes a quantity from a Feynmann diagram. This integrand contains a number of spikes, some positive and some negative. The integrand is unbounded over $(0, 1)^7$ and the singularities dominate the value of the integral. The formula for this integrand appears in Aldins, Brodsky, Dufner & Kinoshita (1970). We thank Professor Toichuro Kinoshita of Cornell University for providing the code for this function. Professor Kinoshita also told us that the true value of the integral was

	MC	VEGAS	MISER	ADB	MPB
Err/I	0.154	0.092	0.073	0.210	0.035
$\text{Err}/\widehat{\text{Err}}$	1.163	34.0	0.750	19.5	0.394

Table 2: Results from particle physics integrand. The relative error is given in the first row. The second row is the true error divided by the error estimate. For ADBAYES the error estimate is deterministic, for the other methods it is a standard error.

determined as of 1991 to be approximately 0.371005292.

The same five methods were employed on this integrand. For mixture of products of beta sampling, the importance sampling densities were obtained by approximating $|f(x)|$ instead of approximating $f_+(x)$ and $f_-(x)$ separately. All methods were run with 2×10^6 points. For the sampling methods, that is all except ADBAYES, these were done as 20 independent runs with 10^5 points in each. The error estimate in these cases was obtained using replicates.

ADBAYES and Vegas fail to give realistic estimates of the error. The other methods all give realistic error estimates, with the mixture of products of betas being most accurate and MISER a reasonably close second.

7 Discussion

We have presented an adaptive method for numerically integrating spiky functions in high dimensions. The method alternates between sampling from a mixture of products of beta densities, and refitting such a mixture to the data at hand.

Error estimates can be obtained either by replication, or internally from a single sequence of iterations. The method proved to be effective on examples with multiple spikes, giving competitive accuracy and realistic error estimates. In these cases it succeeded without even having prior knowledge of where the spikes were.

This method can be defeated by integrands with spikes so narrow that they are not seen in the initial stages of sampling. We think that our method should generally be more effective than Vegas on problems with multiple spikes in high dimensions, because Vegas effectively models a large number of spurious spikes and can in the process lose a real one. We also have found, in our examples, that the sampling based standard errors of our

method were more reliable than the deterministic error bounds constructed by ADBAYES.

When very narrow spikes are present, they can be hard to find by sampling from the uniform distribution. In such cases one could start the algorithm by first searching for spike locations using numerical optimization, as Friedman & Wright (1981) do, and then using that information to construct a nonuniform starting density $p^{(1)}$. We expect this technique to extend the range of the method to narrower spikes, assuming that each important spike has a large enough domain of attraction for the optimization method to find it.

All of our comparisons are based on equal numbers of observations for each method. We have not yet taken account of the computational cost of fitting the mixture of beta model, nor have we explored tradeoffs between speed and accuracy for the method.

The dissertation Zhou (1998) contains some examples with a single spike. Methods that assume a single spike tended to do better on these than did the mixture of products of beta algorithm which tries to estimate the number of spikes.

Appendix: Proof of Theorem 4

The proof of Theorem 4 requires some short lemmas.

Lemma 1 *Let $Z \geq 0$ be a random variable. Then $E(Z) = \int_0^\infty P(Z > z) dz$.*

Proof: This is on page 49 of Chung (1974). Note that Chung (1974) uses the terms “positive” and “strictly positive” instead of “nonnegative” and “positive”, respectively. \square

Lemma 2 *Let X , Y and Z be random variables with $Y = f(X)$, $Z = g(X)$, where f and g are both nonnegative and nondecreasing functions of X . Then $E(f(X)g(X)) - E(f(X))E(g(X)) \geq 0$, provided that the necessary expectations are finite.*

Proof: Let $\mu_g = E(g(X))$ and $a = \min\{f(x) : g(x) \geq \mu_g\}$.

$$\begin{aligned}
& E(f(X)g(X)) - E(f(X))E(g(X)) \\
&= E(f(X)(g(X) - \mu_g)) \\
&= E\{f(X)(g(X) - \mu_g)1_{g(X) \geq \mu_g}(X)\} + E\{f(X)(g(X) - \mu_g)1_{g(X) < \mu_g}(X)\} \\
&\geq a \cdot E\{(g(X) - \mu_g)1_{g(X) \geq \mu_g}(X)\} + a \cdot E\{(g(X) - \mu_g)1_{g(X) < \mu_g}(X)\} \\
&= a \cdot E(g(X) - \mu_g) \\
&= 0. \quad \square
\end{aligned}$$

Proof of Theorem 4: Let $x, y \in [0, 1]$, take the place of r_1 and r_0 respectively, and define

$$\rho(K, x, y) = \frac{V_{y, K, x}(\hat{I})}{V_{y, K, y}(\hat{I})}.$$

The range of sums over k will always be understood to be integers from 1 to K inclusive.

The proof proceeds as follows: First we show that $\max_{0 \leq y \leq 1} \rho(K, x, y)$ takes place at either $y = 0$, or $y = 1$. If $x \geq 1/2$ this maximum is at $y = 0$ while if $x \leq 1/2$ this maximum is at $y = 1$. In either case the largest values of ρ arise as $K \rightarrow \infty$. We evaluate this limit as a function of x and note that the optimum is at $x = 0.5$.

Let E_1 denote expectation, assuming that k is a random integer in $\{1, \dots, K\}$ with probability mass function proportional to k^{2x-y} , and let E_2 denote a corresponding expectation with probability mass function proportional to k^y . Then

$$\begin{aligned}
& \frac{\partial^2}{\partial y^2} \rho(K, x, y) \\
&= \left(\sum k^x\right)^{-1} \left[\left(\sum k^{2x-y} (\log k)^2\right) \left(\sum k^y\right) \right. \\
&\quad \left. - 2 \left(\sum k^{2x-y} \log k\right) \left(\sum k^y \log k\right) + \left(\sum k^{2x-y}\right) \left(\sum k^y (\log k)^2\right) \right] \\
&= \left[E_1((\log k)^2) - 2E_1(\log k)E_2(\log k) + E_2((\log k)^2) \right] \\
&\quad \times \left(\sum k^x\right)^{-1} \left(\sum k^{2x-y}\right) \left(\sum k^y\right) \\
&\geq 0.
\end{aligned}$$

Because $\rho(K, x, y)$ is convex in y , the maximum takes place at one or other of the extremes $y = 0$ or $y = 1$. Let $\rho_i(K, x) = \rho(K, x, i)$, $i = 0, 1$. If $x > 1/2$, Lemma 2 can be used to show that $\rho_0(K, x) - \rho_1(K, x) > 0$ so for $x > 1/2$ we have $\max_y \rho(K, x, y) = \rho_0(K, x)$. Similarly for $x < 1/2$ $\max_y \rho(K, x, y) = \rho_1(K, x)$, while for $x = 1/2$ $\max_y \rho(K, x, y) = \rho_0(K, x) = \rho_1(K, x)$.

Straightforward algebra shows that $\rho_0(K + 1, x) - \rho_0(K, x) \geq 0$ when $x \geq 1/2$ and similarly that $\rho_1(K + 1, x) - \rho_1(K, x) \geq 0$ when $x \leq 1/2$. Thus

$$\sup_{1 \leq K < \infty} \max_{0 \leq r_0 \leq 1} \rho(K, x, y) = \begin{cases} \lim_{K \rightarrow \infty} \rho_0(K, x), & \text{if } x \geq 1/2, \\ \lim_{K \rightarrow \infty} \rho_1(K, x), & \text{if } x \leq 1/2. \end{cases} \quad (25)$$

For $x \geq 1/2$,

$$\lim_{K \rightarrow \infty} \rho_0(K, x) = \lim_{K \rightarrow \infty} \frac{K \sum_{k=1}^K k^{2x}}{(\sum_{k=1}^K k^x)^2} = \frac{(x+1)^2}{2x+1},$$

while for $x \leq 1/2$

$$\lim_{K \rightarrow \infty} \rho_1(K, x) = \lim_{K \rightarrow \infty} \frac{\sum_{k=1}^K k^{2x-1} \sum_{k=1}^K k}{(\sum_{k=1}^K k^x)^2} = \frac{(x+1)^2}{4x}.$$

It now follows easily that

$$\min_{0 \leq x \leq 1} \sup_{1 \leq K < \infty} \max_{0 \leq y \leq 1} \rho(K, x, y) = \sup_{1 \leq K < \infty} \max_{0 \leq y \leq 1} \rho(K, 0.5, y) = 9/8. \quad \square$$

References

- Aldins, J., Brodsky, S. J., Dufner, A. J. & Kinoshita, T. (1970), ‘Moments of the muon and electron’, *Physical Review* **D1**, 2378–2395.
- Berntsen, J., Espelid, T. O. & Genz, A. (1991), ‘An adaptive algorithm for the approximate calculation of multiple integrals’, *ACM Transactions on Mathematical Software* **17**(4), 437–451.
- Bratley, P., Fox, B. J. & Schrage, L. E. (1987), *A Guide to Simulation (Second Edition)*, Springer-Verlag.
- Breiman, L. (1991), ‘The π method for estimating multivariate functions from noisy data (disc: P145-160)’, *Technometrics* **33**, 125–143.

- Chung, K.-L. (1974), *A course in probability theory*, 2nd edn, Academic Press, New York.
- Davis, P. J. & Rabinowitz, P. (1984), *Methods of Numerical Integration (2nd Ed.)*, Academic Press, San Diego.
- Evans, M. & Swartz, T. (1995), ‘Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems’, *Statistical Science* **10**(3), 254–272.
- Friedman, J. H. & Wright, M. H. (1981), ‘A nested partitioning procedure for numerical multiple integration’, *ACM Transactions on Mathematical Software* **7**(1), 76–92.
- Givens, G. H. & Raftery, A. E. (1996), ‘Local adaptive importance sampling for multivariate densities with strong nonlinear relationships’, *Journal of the American Statistical Association* **91**(433), 132–141.
- Glasserman, P., Heidelberger, P. & Shahabuddin, P. (1999), ‘Asymptotically optimal importance sampling and stratification for pricing path-dependent options’, *Mathematical Finance* .
- Hammersley, J. & Handscomb, D. (1964), *Monte Carlo Methods*, London: Methuen.
- 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.
- Hickernell, F. J. (1996), ‘The mean square discrepancy of randomized nets’, *ACM Trans. Model. Comput. Simul.* **6**, 274–296.
- 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.
- Kloek, K. & van Dijk, H. K. (1978), ‘Bayesian estimates of equation system parameters: An application of integration by Monte Carlo’, *Econometrica* **46**, 1–20.
- Lepage, P. (1978), ‘A new algorithm for adaptive multidimensional integration’, *Journal of Computational Physics* **27**, 192–203.

- Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, S.I.A.M., Philadelphia, PA.
- Oh, M.-S. & Berger, J. O. (1993), 'Integration of multimodal functions by Monte Carlo importance sampling', *Journal of the American Statistical Association* **88**(422), 450–456.
- Owen, A. B. (1997a), 'Monte Carlo variance of scrambled equidistribution quadrature', *SIAM Journal of Numerical Analysis* **34**(5), 1884–1910.
- Owen, A. B. (1997b), 'Scrambled net variance for integrals of smooth functions', *Annals of Statistics* **25**(4), 1541–1562.
- Owen, A. B. & Zhou, Y. (1999), Advances in importance sampling, in Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigend, eds, 'Computational Finance (Proceedings of the Sixth International Conference on Computational Finance, Leonard N. Stern School of Business, January 6-8, 1999).', MIT Press, Cambridge, MA.
- Owen, A. B. & Zhou, Y. (2000), 'Safe and effective importance sampling', *Journal of the American Statistical Association* **95**. To appear.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992), *Numerical Recipes in C*, 2 edn, Cambridge.
- Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons.
- Veach, E. & Guibas, L. (1995), Optimally combining sampling techniques for Monte Carlo rendering, in 'SIGGRAPH '95 Conference Proceedings', Addison-Wesley, pp. 419–428.
- West, M. (1992), 'Modelling with mixtures', *Bayesian Statistics 4* **2**, 503–524.
- West, M. (1993), 'Approximating posterior distributions by mixtures', *Journal of the Royal Statistical Society, Ser B* **55**(2), 409–422.
- Williams, D. (1991), *Probability with Martingales*, Cambridge University Press, Cambridge.
- Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer-Verlag.
- Zhang, P. (1996), 'Nonparametric importance sampling', *Journal of the American Statistical Association* **91**(435), 1245–1253.

Zhou, Y. (1998), Adaptive Importance Sampling for Integration, PhD thesis, Stanford University.