STRATIFIED SAMPLING WITH SPHERICALLY SYMMETRIC IMPORTANCE SAMPLERS

MICHAEL EVANS
University of Toronto

TIM SWARTZ
Simon Fraser University

SUMMARY

Multivariate normal or Student importance sampling is a commonly used technique for integration problems in statistical inference. This integration approach is easy to implement, has straight-forward error estimates and is effective in a number of problems. A variety of variance reduction techniques can be considered with importance sampling. Stratified sampling is one of these and in fact, is about the only guaranteed variance reducer. We consider the implementaion of stratified sampling in the context of spherically symmetric importance samplers.

Keywords and phrases: numerical intergration, importance sampling, stratified sampling, envelope generators.

1. INTRODUCTION

The general problem we are concerned with is the approximation of an integral

$$I(f) = \int_{R^k} f(x)\,dx$$

for some function $f : R^k \to R$. While there are many methods for addressing this problem, see for example Evans and Swartz (1995a, 2000) for a fairly complete coverage of these, we focus here on the use of independent importance sampling. For this we suppose that we have a probability density w such that the support of $|f|$ is contained in the support of w and such that we can easily generate samples from w. Then the importance sampling algorithm based on w proceeds by generating a sample $x_1, \ldots, x_n$ from w and reporting the estimate

$$\hat{I}_{w,n}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{w(x_i)};$$

i.e. we average the values of f at the generated points weighted by the inverse of the importance sampling density at these points. By the strong law of large numbers this quantity converges almost surely to $I(f)$ as $n \to \infty$. The variance of the estimate is given by σ_w^2/n where

$$\sigma_w^2 = Var_w\left[\frac{f}{w}\right] = \int_{R^k} \frac{f^2(x)}{w(x)}\,dx - I(f)^2 .$$

If σ_w^2 is finite then we can estimate it by

$$s_w^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f(x_i)}{w(x_i)} - \hat{I}_{w,n}(f)\right)^2 .$$

Further if σ_w^2 is finite then a Central Limit Theorem, see for example Evans and Swartz (2000), establishes that

$$\frac{\hat{I}_{w,n}(f) - I(f)}{s_w/\sqrt{n}} \to^D Z$$

where $Z \sim N(0,1)$ and this gives us the basic method of assessing the accuracy of $\hat{I}_{w,n}(f)$.

The central issue with importance sampling is the choice of the importance sampler w. If a poor choice is made σ_w^2 may be large or even infinite. Of course we want to choose w so that σ_w^2 is as small as possible. Actually when $k > 1$ there are not many candidates for the choice of w, at least when actual practice is considered. A typical strategy is to transform the integral so that all variables vary without restriction over R^1 and then choose a multivariate Student$_k(\lambda, \mu, \Sigma)$ distribution for w for some choice of $\mu \in R^k, \Sigma \in R^{k\times k}$ positive definite and $\lambda > 0$ with $\lambda = \infty$ indicating the $N_k(\mu, \Sigma)$ distribution. We try to match this choice as closely as possible to the integrand f. For example, if f is nonnegative and unimodal then we first calculate the mode $\hat{\mu}$ of f and

$$\hat{\Sigma} = \left(\frac{-\partial^2 \log f(\hat{\mu})}{\partial x_i \partial x_j}\right)^{-1}$$

the inverse Hessian matrix of $\log f$ at $\hat{\mu}$. Taking w to be a

$$\text{Student}_k\left(\lambda, \hat{\mu}, \frac{\lambda + k}{\lambda}\hat{\Sigma}\right)$$

distribution matches the mode and the Hessian matrix of the log of the importance sampler to the corresponding quantities for f. In some statistical contexts f may be thought of as being asymptotically multivariate normal and so Laplace's approximation would be appropriate for approximating I; see, for example, Evans and Swartz (2000) for a discussion of this approximation. A deficiency in asymptotic approximations is that they provide no means of assessing accuracy. So a natural alternative to Laplace's approximation, which allows for error assessment, is to use importance sampling via a

$$\text{Student}_k\left(\lambda, \hat{\mu}, \frac{\lambda-2}{\lambda}\hat{\Sigma}\right)$$

distribution for w where now we are matching approximate means and variance matrices. Of course this choice requires that $\lambda > 2$ whereas matching Hessian matrices does not impose such a restriction. In neither case have we discussed how to choose λ. Basically the idea is to choose λ low enough so that, hopefully, the variances of all our estimators are finite. The only method available for ensuring that an appropriate value of λ has been chosen is to study the tail behavior of f and this is typically quite difficult. So it is customary to simply choose λ fairly low. We note that in both of these cases we can make the transformation $x \to y = \hat{\Sigma}^{-1/2}(x - \hat{\mu})$ in $I(f)$ so that the relevant importance sampler is now a $\text{Student}_k\,(\lambda, 0, cI)$ density for some choice of λ and c. This class of importance samplers is spherically symmetric; i.e. if $X \sim w$ and $Q \in R^{k\times k}$ is orthogonal then also $Y = QX \sim w$.

When f is not nonnegative we try to obtain a factorization $f = mg$ where $g \geq 0$ is integrable and choose w as above but use g instead of f to guide our choice of w; e.g. g could be the unnormalized posterior in a Bayesian integration problem. When f is multimodal then a different strategy must be employed; see Evans and Swartz (2000) for a discussion of this situation.

Of course there is generally no guarantee that the choice of a particular w will work in the sense of giving adequate approximations in reasonable computing times. But we suppose for the remainder of this discussion that such a choice has been made and then consider how we can improve on the algorithm. This leads us to the consideration of various variance reduction techniques. The three main variance reduction methods are systematic sampling (sometimes referred to as antithetic variates), control variates and stratified sampling. Of these stratified sampling is the only one that guarantees a reduction in variance so it is a natural approach to consider first. Despite this there is little literature on using stratified sampling with Student importance sampling and we subsequently point out a plausible reason for this. This paper presents a significant improvement of some developments with respect to stratified sampling that were first presented in Evans and Swartz (1995b). For a full discussion of variance reduction methods see Evans and Swartz (2000).

It is important to note that in many statistical problems; e.g. Bayesian inference, our interest lies in approximating a ratio of integrals of the form

$$R(m) = \frac{I\,(mf)}{I\,(f)}$$

where $f \geq 0$ and $m : R^k \to R^1$; i.e. $R(m)$ is the posterior expectation of m. Typically there are many such functions m of interest and so it makes sense to choose a single importance sampler w matched to the unnormalized posterior

f. Then based on a sample $x_1, \ldots, x_n$ from w we estimate $R(m)$ by

$$\hat{R}_{w,n}(m) = \frac{\hat{I}_{w,n}(mf)}{\hat{I}_{w,n}(f)}.$$

An application of the delta theorem together with some conditions, see Evans and Swartz (2000), gives that $\hat{R}_{w,n}(m)$ is asymptotically normal with mean $R(m)$ and asymptotic variance

$$\frac{\sigma^2_{R(m),w}}{n} = \frac{1}{nI^2(f)} \left\{ \begin{array}{c} Var_w\left[\frac{mf}{w}\right] + I^2(mf)I^{-2}(f)Var_w\left[\frac{f}{w}\right] \\ -2I(mf)I^{-1}(f)Cov_w\left[\frac{mf}{w}, \frac{f}{w}\right] \end{array} \right\}.$$

This quantity is estimated by substituting the sample analogues of the various quantities into the formula.

In Section 2 we discuss stratified sampling in general. In Section 3 we discuss the implementation of stratified sampling when using spherically symmetric importance samplers. In Section 4 we discuss the relationship between stratified sampling with spherically symmetric importance samplers and other approaches to integration.

2. STRATIFIED SAMPLING

The technique of *stratified sampling* is based on a decomposition of the importance sampler w as

$$w = p_1 w_1 + \cdots + p_l w_l$$

where $w_1, \ldots, w_l$ are densities for which we have generating algorithms, with $\text{supp}(w) = \cup_{i=1}^{l} \text{supp}(w_i)$, $\text{supp}(w_i) \cap \text{supp}(w_j) = \phi$ when $i \neq j$ and

$$p_i = \int_{\text{supp}(w_i)} w(x)\, dx$$

for $i = 1, \ldots, l$. The sets $\text{supp}(w_i)$ are the *strata* and of course $\sum_{i=1}^{l} p_i = 1$. Stratified sampling occurs when we sample a *predetermined* number n_i times from w_i and then estimate

$$\int_{\text{supp}(w_i)} f(x)\, dx$$

by $\hat{I}_{w_i,n_i}$. The estimate of I is then given by

$$\hat{I}_{strat}(f) = \sum_{i=1}^{l} \hat{I}_{w_i,n_i}.$$

By the strong law of large numbers $\hat{I}_{strat}(f)$ converges almost surely to $\int_{R^k} f(x)\,dx$ as $\min(n_1, \ldots, n_l) \to \infty$. We have that

$$\sigma^2_{strat} = \sum_{i=1}^{l} \frac{\sigma^2_{w_i}}{n_i}$$

and we estimate this using

$$s^2_{strat} = \sum_{i=1}^{l} \frac{s^2_{w_i}}{n_i}$$

where $s^2_{w_i}$ is the sample estimate of $\sigma^2_{w_i}$. Provided that each $\sigma^2_{w_i} < \infty$ we have that

$$\frac{\hat{I}_{strat}(f) - I(f)}{s_{strat}} \to^D Z$$

where $Z \sim N(0,1)$ as $\min(n_1, \ldots, n_l) \to \infty$.

The idea behind stratified sampling is that we can allocate sample resources more carefully by deciding ahead of time to place more sample values in regions where the importance sampler w is not very good at mimicking f and so improve the accuracy of the approximation. In practice it is difficult to determine a suitable stratification but if we choose any particular stratification we have the following result. For a given overall sample size $n = \sum_{i=1}^{l} n_i$, if we take $n_i = np_i$ then

$$\sigma^2_{strat} = \frac{1}{n} \sum_{i=1}^{l} \frac{\sigma^2_{w_i}}{p_i} \leq \frac{\sigma^2_w}{n}.$$

This choice of the n_i is known as *proportional allocation.* So we need never do any worse than importance sampling based on w if we allocate the n sample values appropriately amongst the strata. This is the basis for our statement that stratified sampling is a guaranteed variance reducer. In contrast the other variance reduction methods mentioned in Section 1 can result in increased variance when compared to straight importance sampling.

It is easy to show, see Evans and Swartz (2000), that with proportional allocation,

$$\frac{\sigma^2_w}{n} = \sigma^2_{strat} + \frac{Var\left[E\left[f(X)/w(X)|i\right]\right]}{n}.$$

Therefore to get the maximum benefit of stratification with proportional allocation we want to decompose w so that $Var\left[E\left[f(X)/w(X)|i\right]\right]$ is as large as possible. This is just the variance of the l numbers

$$\frac{1}{p_i} \int_{supp(w_i)} f(x)\,dx$$

with respect to the probability distribution given by the p_i.

In fact with a particular stratification one can do better than proportional allocation. A well-known result establishes that the smallest variance is achieved by the *Tschuprow-Neyman allocation* given by

$$n_i = n \frac{\sigma_{w_i}}{\sum_{j=1}^{l} \sigma_{w_j}}$$

and the minimized variance is

$$\frac{\left(\sum_{j=1}^{l} \sigma_{w_j}\right)^2}{n} \leq \frac{\sigma_w^2}{n}.$$

For a proof of these results see, for example, Evans and Swartz (2000). This does not lead directly to a practical algorithm, however, as the σ_{w_i} are typically unknown. There are several approaches to getting around this problem. One possibility is to estimate the σ_{w_i} in a pilot study, say based on proportional allocation, before running importance sampling.

It is also possible to estimate a ratio of integrals R using stratification. Using the same stratification for the numerator and denominator let

$$I_i(mf) = \int_{supp(w_i)} m(x) f(x)\, dx$$

with estimate $\hat{I}_{w_i,n_i}(mf)$. Then we estimate $R(m)$ by

$$\hat{R}_{strat}(m) = \frac{\sum_{i=1}^{l} \hat{I}_{w_i,n_i}(mf)}{\sum_{i=1}^{l} \hat{I}_{w_i,n_i}(f)}.$$

If $I(f) \neq 0$ then, under certain conditions as outlined in Evans and Swartz (2000), $\hat{R}_{strat}(n_1, \ldots, n_l)$ has asymptotic variance

$$I^{-2}(f) \sum_{i=1}^{l} \frac{1}{n_i} \left\{ \begin{array}{c} Var_{w_i}\left[\frac{mf}{w_i}\right] + I_i^2(mf) I_i^{-2}(f) Var_{w_i}\left[\frac{f}{w_i}\right] \\ -2 I_i(mf) I_i^{-1}(f) Cov_{w_i}\left[\frac{f}{w_i}, \frac{mf}{w_i}\right] \end{array} \right\}.$$

We estimate this by substituting in the appropriate sampling analogues.

3. STRATIFIED SAMPLING WITH SPHERICALLY SYMMETRIC IMPORTANCE SAMPLERS

Of immediate practical importance in the implementation of stratified sampling is the choice of the stratification. For several reasons we cannot choose this arbitrarily. For example, we must have an algorithm to generate from the w_i and moreover we require the values of the p_i to implement the proportional allocation. Clearly when using a Student$_k$ $(\lambda, 0, cI)$ distribution it would be very difficult to use a stratification based upon rectangles for these reasons. As we will see, however, after overcoming one significant problem, it is relatively straight-forward to stratify this distribution using its spherical symmetry.

We now discuss how to implement stratified sampling for a general spherically symmetric importance sampler w. For such a density there is a function $g : R^k \to R^1$ such that

$$w(x) = g(||x||)$$

and

$$h(r) = \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)} r^{k-1} g(r)$$

is a density on $[0, \infty)$. Further if U is distributed uniformly on S^{k-1}, the unit sphere in R^k, statistically independent of $r \sim h$ then $X = rU \sim w$.

Now we decompose w using the spherical symmetry. For this we choose $0 = r_0 < r_1 < \cdots < r_{l-1} < r_l = \infty$ and let

$$\begin{aligned} A_1 &= \{x : r_0 \le ||x|| < r_1\}, \\ A_2 &= \{x : r_1 \le ||x|| < r_2\}, \\ &\vdots \\ A_l &= \{x : r_{l-1} \le ||x|| < r_l\}; \end{aligned}$$

i.e. a sequence of annular rings centered at 0 that partition R^k. Then w_i is just w conditioned to the annulus A_i and, making the transformation $x \to (r, U)$, we have that

$$\begin{aligned} p_i &= \int_{A_i} w(x)\, dx \\ &= \int_{r_{i-1}}^{r_i} \int_{S^{k-1}} r^{k-1} g(r)\, dU\, dr \\ &= \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)} \int_{r_{i-1}}^{r_i} r^{k-1} g(r)\, dr. \end{aligned}$$

This requires us to evaluate l one-dimensional integrals to obtain the p_i for the proportional allocation. This is not a significant computational problem. In fact if we choose w to be in the multivariate Student family then these integrals are available to us from tabulations of the Chisquare distribution when $\lambda = \infty$ and from tabulations of the F distribution otherwise.

To implement stratified sampling for w, however, we must also have an algorithm to generate from each w_i. We can do this by generating U uniform on S^{k-1} statistically independent of r which is generated from h conditioned to the interval (r_{i-1}, r_i). Generating U is straight-forward as we can generate $Z \sim N_k(0, I)$ and put $U = Z/||Z||$; see, for example, Eaton (1983). It is the generation of r, however, where the essential difficulty in implementing this procedure lies. In general this is not an easy generating problem even when w is multivariate normal. The naive method of generating r according to h, which presumes the existence of such an algorithm as well, and then selecting only those values that fall in (r_{i-1}, r_i) to reflect the conditioning, is hopelessly inefficient for even moderate l.

In the past few years, however, good methods have been developed for generating from a fairly general density h. These are called envelope methods in Evans and Swartz (2000) and their origins lie in the methods developed for log-concave densities in Gilks and Wild (1992) and extended in Hörmann (1995) and Evans and Swartz (1998). For this we break the density h into s parts on $(-\infty, c_1), [c_1, c_2), \ldots, [c_{s-1}, \infty)$ and suppose that we have monotone transformations $T_i : [0, \infty) \to R^1$ for $i = 1, \ldots, s$ and such that $T_i \circ h$ is concave when T_i is increasing or $T_i \circ h$ is convex when T_i is decreasing. The concavity or convexity of the transformed density leads to easily constructed upper and lower envelopes for h. There are also some restrictions placed on the transformations so that the upper and lower envelopes lead to densities that can be easily generated from via inversion but there is a rich class of possible transformations; e.g. power transformations. As shown in Evans and Swartz (1998, 2000) these conditions lead to an easy to implement, highly efficient rejection generator for h and also for h conditioned to an interval with no additional difficulty. For example, for Chisquare(α) distributions with $\alpha \geq 2$ we can take $s = 1$ and $T_1 = \log$ and when $\alpha < 2$ we can find a single power transformation $T_1(f) = f^p$ that suffices. For $F(\alpha, \beta)$ distributions two power transformations, one for each tail, suffice to construct the generator. The reader is referred to Evans and Swartz (2000) for more details on the construction of these generators. In effect these methods permit the relatively easy implementation of stratified sampling with spherically importance samplers. Of course the density h cannot be just anything but all of the standard choices can be handled by this approach.

We now consider a specific numerical example.

Example *Stratified Student importance sampling.*

For this example we consider a Bayesian analysis of a linear statistical model which requires that integrals of dimension 10 be approximated. The particular Bayesian analysis we consider coincides with the conditional analysis of the model as presented in Fraser (1979). The statistical model that we use is specified as follows. Suppose that we observe the data (X, y) where $X \in R^{45\times 9}$ has $x_{ij} = 1$ for $5(j-1)+1 \leq i \leq 5j$ and 0 otherwise, and $y \in R^{45}$. The statistical model is then specified by the equation

$$\mathbf{y} = X\beta + \sigma z$$

where $\beta \in R^9, \sigma \in (0, \infty)$ and the error $z \in R^{45}$ is a sample of 45 from a distribution in the family $\{\text{Student}^*(\nu) : \nu \in (2, \infty)\}$ where Student$^*(\nu)$ denotes the Student(ν) distribution standardized to have variance 1. Therefore the density of z is given by

$$g_\nu(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)}\left(1 + \frac{z^2}{\nu-2}\right)^{-\frac{\nu+1}{2}} \frac{1}{\sqrt{\nu-2}}$$

for $-\infty < z < \infty$. Therefore we have 9 location parameters β_i, $1 \le i \le 9$, a scale parameter σ and a shape parameter ν. In an application all these parameter values are unknown but we will treat ν as being known here and make the assignment $\nu = 3$. The restriction in the degrees of freedom for the Student distribution ensures that all response variables y_i have a mean and finite variance. We note that this model corresponds to 5 independent observations from each of the 9 distributions specified by (β_i, σ) for $i = 1, \ldots, 9$. We denote the j-th observation from the i-th distribution by y_{ij}.

We place the improper prior $\pi(\beta, \sigma) = 1/\sigma$ on (β, σ). Then the posterior density for the 10-dimensional parameter (β,σ) is proportional to

$$\prod_{i=1}^{9}\prod_{j=1}^{5}\frac{1}{\sigma}g_\nu\left(\frac{y_{ij}-\beta_1 x_{i1}-\cdots-\beta_9 x_{i9}}{\sigma}\right)$$

and we note that this is a reasonably complicated function. Actually in this particular case some integrations can be carried out in closed form but we choose to ignore this here as this is not characteristic of such problems. The integrals we are approximating are therefore 10-dimensional. It makes sense to make the transformation from σ to $\log\sigma$ so that the variable in the integration is unconstrained in R^1. This more closely conforms to the behavior of an approximating multivariate Student density. With this transformation, and letting $\theta = (\theta_1, \ldots, \theta_9, \theta_{10}) = (\beta_1, \ldots, \beta_9, \log\sigma)$, we see that we want to approximate integrals of the form

$$\begin{aligned} I(mf) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} m(\theta_1, \ldots, \theta_{10}) f(\theta_1, \ldots, \theta_{10})\, d\theta_1 \cdots d\theta_{10} \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} m(\theta_1, \ldots, \theta_{10}) \exp\{-9\, n\, \theta_{10}\} \times \\ &\prod_{i=1}^{9}\prod_{j=1}^{n} g_v\left(\frac{y_{ij}-\theta_i}{\exp(\theta_{10})}\right)\, d\theta_1 \cdots d\theta_{10} \end{aligned}$$

with

$$f(\theta_1, \ldots, \theta_{10}) = \exp\{-9\, n\, \theta_{10}\}\prod_{i=1}^{9}\prod_{j=1}^{n} g_v\left(\frac{y_{ij}-\theta_i}{\exp(\theta_{10})}\right).$$

To obtain a specific numerical example we constructed a data set by fixing values for the parameters and then generating the z_{ij}. For this we specified $\beta_i = 0$ for every $i \ne 1$, $\beta_1 = 1.8$, $\sigma = 1$ and generated the z_{ij} from a Student$^*(3)$ distribution.

For these integration problems we used a

$$\text{Student}_k\left(\lambda, 0, \frac{\lambda-2}{\lambda}I\right)$$

importance sampler with $\lambda = 5$ after standardizing via the mode $\hat{\mu}$ of the posterior and the inverse of the Hessian matrix $\hat{\Sigma}$ of $\log f$. Representative results for

i	$R(\theta_i)$	$\hat{R}(\theta_i)$
1	2.043	2.043 (.001)
2	0.095	0.096 (.023)
4	0.018	0.017 (.127)
10	-0.073	-0.073 (.017)

Table 1: Importance sampling estimates and estimated absolute coefficients of variation for the Example where R denotes the exact value and $\hat{R}$ denotes the importance sampling estimate.

approximating the posterior expectations of several of the θ_i using importance sampling with this w are reported in Table 1 for a Monte Carlo sample size of $N = 10^5$. These computations require about 50 seconds of CPU time. The estimates are accurate but the estimated absolute coefficient of variation for $\hat{R}(\theta_4)$ is quite high. We tried these computations with different choices for λ in the range 1 to 20 but not much difference was noticed in the results. Overall this importance sampler was reasonably successful here.

We then considered the use of stratified sampling with this importance sampler. When $X \sim$ Student$_k \left(\lambda, 0, \frac{\lambda-2}{\lambda}I\right)$we have that

$$r^2 = \|X\|^2 \sim k\frac{\lambda-2}{\lambda}F(k,\lambda).$$

Therefore, because of the rotational symmetry of the Student$_k \left(\lambda, 0, \frac{\lambda-2}{\lambda}I\right)$ distribution, we can generate a value from the Student$_k \left(\lambda, 0, \frac{\lambda-2}{\lambda}I\right)$ conditioned to A_i by generating r^2 from the $kF(k,\lambda)$ distribution conditioned to $\left(r_{i-1}^2, r_i^2\right)$, generating $U \sim Uniform\left(S^{k-1}\right)$ and setting $X = rU$. To generate from the $kF(k,\lambda)$ distribution conditioned to $\left(r_{i-1}^2, r_i^2\right)$ we used the envelope methods discussed above. For the specific transformations employed see Evans and Swartz (1998). We chose $l = 25$ strata with $p_1 = p_2 = \cdots = p_{25}$. We started with the proportional allocation with a sample of 100 being taken from each stratum and then switched to the Tschuprow-Neyman allocation for the remaining sampling. In this example, the average efficiency increase when estimating the $R(\theta_i)$ over $i = 1, \ldots, 10$ using stratified sampling was 2.8 when compared to straight Student importance sampling. So stratified sampling was quite useful in improving the performance of the importance sampler here.

In this same example we also tried different boundary points based on the observation that the estimators of the component strata $\hat{I}_j(m)$ had greatly different standard errors. We therefore began with 2 strata having equal probability and generated a small sample of size 100 in each stratum. We then split the stratum having the largest standard error into 2 sub-strata of equal probability. This splitting process was continued until $l = 25$ strata were constructed. At this point we switched to Tschuprow-Neyman allocation. However, this strategy did not result in any meaningful variance reductions.

4. CONCLUSIONS

We have shown that approximate integration via spherical stratified sampling is now a feasible technique and moreover that it leads to meaningful improvements in computational efficiencies when compared to straight importance sampling. The more the integrand deviates from the spherical symmetry of the importance sampler w the more useful the technique will be. Still in high-dimensional situations where the integrand deviates sharply from w we can expect that importance sampling, with or without spherical stratification, will not lead to accurate approximations within feasible computation times. In such a context there are other variance reduction techniques that can be combined with spherical stratification to improve performance. In particular symmetrizing an integrand with respect to a finite subgroup of the orthogonal group has the effect of making the integrand more like the importance sampler while leaving the value of the integral invariant. Full details concerning this technique can be found in Evans and Swartz (2000).

There are alternative approaches to approximate integration that are very close in spirit to spherical stratified sampling. For example, in many such problems most of the variation in the behavior of the integrand occurs radially; i.e. the standardized integrand is roughly spherically symmetric. In such a context it would seem to make sense to use a more accurate integration approach for the radial direction. For example, we could employ Simpson's rule or a more accurate Gauss-Laguerre rule for the radial direction. A disadvantage of this is that it introduces a small systematic error in the Monte Carlo estimates; i.e. they are no longer consistent. If this error is small relative to the accuracy desired in our approximation then of course this is not a problem. The technique of randomized quadrature avoids this inconsistency. With this technique the points and weights of a quadrature rule are randomly generated from some distribution such that the randomly generated rule integrates polynomials up to a certain degree (or some other class of functions) exactly. Genz and Monahan (1998a, 1998b, 1999) discuss randomized spherical-radial rules. A disadvantage of this technique is that it requires the solution to an increasingly difficult generation problem as the maximum degree of the polynomials integrated exactly increases. Further for all the higher-order quadrature approaches it is not clear how well they handle discontinuous integrands; e.g. when computing probabilities of regions, whereas stratified sampling seems better adapted for this. Still there is nothing definitive that can be said here at this point about the best approach and this warrants further investigation.

REFERENCES

Eaton, M.L. (1983). Multivariate Statistics. John Wiley, New York.

Evans, M. and Swartz, T. (1995a). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. Statistical Science, 10, 254-272.

Evans, M. and Swartz, T. (1995b). Bayesian integration using multivariate Student importance sampling. Computing Science and Statistics 27 (M.M. Meyer and J.L. Rosenberger, eds.). Interface Foundation of North America, Inc., Fairfax Station, VA, 456-461.

Evans, M. and Swartz, T.B. (1998). Random variable generation using concavity properties of transformed densities. Journal of Computational and Graphical Statistics 7, 514-528.

Evans, M. and Swartz, T.B. (2000). Approximating Integrals Via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford.

Fraser, D.A.S. (1979). Inference and Linear Models. McGraw-Hill, New York.

Hörmann, W. (1995). A rejection technique for sampling from T-concave distributions. ACM Transactions on Mathematical Software 21, 182-193.

Genz, A. and Monahan, J. (1998a). Stochastic integration rules for infinite regions. SIAM Journal of Scientific Computation 19, 426-439.

Genz, A. and Monahan, J. (1998b). Stochastic methods for multiple integrals over unbounded regions. Math. and Comp. in Simul. 47, 287-298.

Genz, A. and Monahan, J. (1999). A stochastic algorithm for high dimensional multiple integrals over unbounded regions with Gaussian weight. To appear in J. Compl. Appl. Math.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337-348.