ADAPTIVE IMPORTANCE SAMPLING IN MONTE CARLO INTEGRATION

Man-Suk Oh
Department of Statistics
Purdue University
West Lafayette, IN, USA

and

James O. Berger
Department of Statistics
Purdue University
West Lafayette, IN, USA

Technical Report #89–19C

Department of Statistics
Purdue University

October 1989

△

ADAPTIVE IMPORTANCE SAMPLING IN MONTE CARLO INTEGRATION

Man-Suk Oh * James O. Berger

Purdue University and Duke University

Aug 25, 1989

Abstract

An Adaptive Importance Sampling (AIS) scheme is introduced to compute integrals of the form $E\varphi = \int \varphi(\theta)f(\theta)d\theta / \int f(\theta)d\theta$ as a mechanical, yet flexible, way of dealing with the selection of parameters of the importance function. AIS starts with a rough estimate for the parameters λ of the importance function $g_\lambda \in \mathcal{G}$, and runs importance sampling in an iterative way to continually update λ using only linear accumulation. Consistency of AIS is established. The efficiency of the algorithm is studied in two examples and found to be substantially superior to ordinary importance sampling.

*Research was supported by the National Science Foundation, grants DMS–8702620 and DMS–8717799, and by a David Ross Fellowship from Purdue University.

1 Introduction

1.1 Monte Carlo Importance Sampling

A variety of statistical (and other) problems reduce to calculation of integrals of the form

$$E\varphi = \frac{\int \varphi(\theta) f(\theta) d\theta}{\int f(\theta) d\theta}, \tag{1}$$

where $\theta \in R^p$, $\varphi(\theta)$ is a measurable function, and $f(\theta)/\int f(\theta)d\theta$ is a density function. For instance, in Bayesian analysis $E\varphi$ could be any posterior quantity such as a posterior moment, probability of a set, predictive distribution, or marginal density; here $f(\theta)$ would be the product of the likelihood function $L(\theta|x)$ and prior $\phi(\theta)$. If $\varphi = (\varphi_1, \ldots, \varphi_\ell)' \in R^\ell$, (1) is to be defined as the componentwise expectation of φ.

In many practical situations, the integrals in equation (1) are analytically intractable and require numerical integration. When θ is high dimensional (say > 5) the preferred method of numerical integration is Monte Carlo Integration with Importance Sampling (briefly, Importance Sampling, from here on). General discussions of Importance Sampling are given in Hammersley and Handscomb (1964), Davis and Rabinowitz (1975), and Rubinstein (1985). Kloek and van Dijk (1978) considered Importance Sampling in the context of statistics and econometric problems. Modifications and variance reduction techniques have been studied in van Dijk (1983, 1984, 1987), van Dijk and Kloek (1980, 1983), van Dijk et al (1985), Bauwens and Richard (1985), Geweke (1986, 1988), Stewart (1979, 1983, 1984). Important related studies are presented in Naylor and Smith (1982, 1983, 1988), Smith et al (1985), Tierney and Kadane (1986).

Importance Sampling proceeds by choosing a probability density function g, called the importance function, defining $w(\theta) = f(\theta)/g(\theta)$, and replacing the equation (1) by

$$\begin{aligned} E\varphi &= \frac{\int \varphi(\theta) w(\theta) g(\theta) d\theta}{\int w(\theta) g(\theta) d\theta} \\ &= \frac{E_g \varphi w}{E_g w}, \end{aligned} \tag{2}$$

where the subscript g in E_g indicates that the expectation is taken with respect to the density g. Then, draw i.i.d. random samples $\theta_1, \ldots, \theta_n$ from $g(\theta)$, and approximate (2) by

$$\hat{E}^B\varphi = \frac{\Sigma_{i=1}^n \varphi(\theta_i)w(\theta_i)}{\Sigma_{i=1}^n w(\theta_i)}. \tag{3}$$

The superscript B in $\hat{E}^B\varphi$ is used to denote (Basic) Importance Sampling (BIS), to distinguish it from Adaptive Importance Sampling (AIS) that will be discussed later.

Under mild conditions, $\hat{E}^B\varphi$ converges to $E\varphi$ with probability one, and is approximately normally distributed with mean $E\varphi$ and variance σ^2/n, where

$$\begin{aligned} \sigma^2 &= \frac{1}{(\int f(\theta)d\theta)^2}\int(\varphi(\theta) - E\varphi)^2 w^2(\theta)g(\theta)d\theta & (4) \\ &= \frac{1}{(\int f(\theta)d\theta)^2}(Var_g(\varphi w) - 2(E\varphi)Cov_g(\varphi w, w) \\ &\quad +(E\varphi)^2 Var_g(w)); & (5) \end{aligned}$$

the subscript g in Var_g, Cov_g indicates that variances and covariances are taken with respect to the density g. The performance of Importance Sampling clearly depends on σ^2 and n, and σ^2 depends on the importance function g. Thus, the goal is to choose a g which yields small σ^2.

1.2 Desirable Properties of the Importance Function

There is no easy prescription for choosing a good importance function even though it is the key issue in Importance Sampling. Here are some typically desirable properties of an importance function. First, it should have convenient Monte Carlo properties, i.e., it should be easy to generate random drawings from. For example, common choices are multivariate normal and student t densities, from which random variate generation is straightforward. Second, the tails of g should not be sharper than the tails of f. Otherwise, $\hat{E}^B\varphi$ may have a large variance or even fail to converge; an example is given in Berger (1985), section 4.9.

Third, g should mimic f well. The importance function that minimizes the variance σ^2/n of $\hat{E}^B\varphi$ is proportional to

$$|\varphi(\theta) - E\varphi|f(\theta); \tag{6}$$

see Rubinstein (1981). Since determination of $E\varphi$ is the original problem, however, it is not practical to use this as the importance function. Also, one typically desires to estimate $E\varphi$ for several φs at the same time, using a single importance function for efficiency. It is often reasonable to choose g proportional to f, instead of (6). Indeed, in statistical problems φ is often fairly flat over the region where f is concentrated (at least when the sample size is moderate or large), so that a density proportional to f is nearly optimal. But even finding a density function g that is similar to f and from which random variates can be easily generated is often difficult. Thus, there is a trade off between convenient Monte Carlo properties and mimicry of f.

A convenient way to choose an importance function g is to first choose a parametric family of density functions, $\mathcal{G} = \{g_\lambda; \lambda \in \Lambda\}$, which satisfy the first and second desirable properties above. Then, select the parameters λ of g to match some features of f to achieve the third property above. For instance, if λ consists of the mean and covariance matrix of g, then these could be chosen to match the estimated mean and covariance matrix of f, with some modifications, if necessary.

1.3 Adaptive Importance Sampling

The idea of Adaptive Importance Sampling (AIS) was given in Kloek and van Dijk (1978), van Dijk (1984). Naylor and Smith (1988) used iterative ideas in Gaussian quadrature, and suggested the same for Importance Sampling. Here, AIS is developed as a mechanical, yet flexible, way of dealing with the selection of the parameters λ of the importance function $g \in \mathcal{G} = \{g_\lambda; \lambda \in \Lambda\}$. It starts with a rough guess for λ, and runs Importance Sampling in an iterative way to continually update λ. At each stage AIS estimates $E\varphi$ and λ by *pooling* the estimates obtained from BIS, using i.i.d. random drawings generated from the importance function of the current stage, and estimates obtained from AIS in the previous stage. The process stops when the desired accuracy is reached. Discussion and comparison with other adaptive schemes will be given in section 2.4.

1.4 Preview of Results

In section 2.1, the AIS algorithm is given. Section 2.2 discusses reasonable stopping rules to use for AIS. The additional computations needed by AIS are discussed in section 2.3, and argued to typically be minimal. In section 2.4, further discussion of AIS, together with comparison with Kloek and van Dijk's adaptive scheme will be given. Illustrative examples and numerical comparisons are given in section 2.5. Consistency and approximate normality of estimates of AIS are established in section 3, using martingale limit theorems. Conclusions and comments are given in section 4.

2 Adaptive Importance Sampling

Let $\mathcal{G} = \{g_\lambda; \lambda \in \Lambda\}$ be a parametric family of density functions which have convenient Monte Carlo properties and have tails no sharper than f. Assume that $\lambda = (\lambda_1, \ldots, \lambda_m)'$ and that it would be desirable to choose λ equal to $E\xi = \int \xi(\theta) f(\theta) d\theta / \int f(\theta) d\theta$, where $\xi(\theta) = (\xi_1(\theta), \ldots, \xi_m(\theta))$, in making g_λ a good approximation to f. For instance, choosing $\xi_1(\theta) = \theta$ would mean that it would be desirable to have λ_1 equal to the mean of f. Of course, $E\xi$ will itself typically be unknown (indeed, frequently some of the $\xi_i(\theta)$, $1 \leq i \leq m$, will be the target $\varphi(\theta)$), in which case the natural thing to do is to choose λ to be an estimate of $E\xi$.

2.1 Algorithm

Assume that $\varphi = (\varphi_1, \ldots, \varphi_\ell)'$.

Stage 0: Choose a stopping rule (see section 2.2) and an initial estimate $\lambda^{(0)}$ of $E\xi$. Let $g^{(0)}$ be $g_{\lambda^{(0)}}$. (Often $\lambda^{(0)}$ is chosen by likelihood methods.)

Stage 1: Draw n_1 i.i.d. random drawings $\theta_1^{(1)}, \ldots, \theta_{n_1}^{(1)}$ from $g^{(0)}$. Let

$$w^{(1)}(\theta) = f(\theta)/g^{(0)}(\theta) \tag{7}$$

and define a functional

$$N^{(1)}(h) = \sum_{i=1}^{n_1} h\left(\theta_i^{(1)}\right) w^{(1)}\left(\theta_i^{(1)}\right). \tag{8}$$

Compute $N^{(1)}(\xi_i)$, $i = 1, \ldots, m$, $N^{(1)}(\varphi_i)$, $i = 1, \ldots, \ell$, and $N^{(1)}(1)$. Also, one might need to compute some statistics needed for a stopping rule; this will be discussed in the next section. Check a desired stopping rule to see if AIS should end. If so, go to "Conclusion". If not, then set λ equal to

$$\lambda^{(1)} = \left(\frac{N^{(1)}(\xi_1)}{N^{(1)}(1)}, \ldots, \frac{N^{(1)}(\xi_m)}{N^{(1)}(1)}\right)'. \tag{9}$$

Let $g^{(1)} = g_{\lambda(1)}$ and go to the next stage.

$\vdots$

Stage k. Draw n_k i.i.d. random drawings $\theta_1^{(k)}, \ldots, \theta_{n_k}^{(k)}$ from $g^{(k-1)} = g_{\lambda(k-1)}$. Let

$$w^{(k)}(\theta) = f(\theta)/g^{(k-1)}(\theta) \tag{10}$$

and define a functional

$$N^{(k)}(h) = \sum_{i=1}^{n_k} h\left(\theta_i^{(k)}\right) w^{(k)}\left(\theta_i^{(k)}\right). \tag{11}$$

Compute $N^{(k)}(\xi_i)$, $i = 1, \ldots, m$, and $N^{(k)}(\varphi_i)$, $i = 1, \ldots, \ell$, and $N^{(k)}(1)$. Also, compute any necessary statistics for use in the stopping rule. If the stopping rule is not satisfied, then set λ equal to

$$\lambda^{(k)} = \left(\frac{\sum_{j=1}^{k} N^{(j)}(\xi_1)}{\sum_{j=1}^{k} N^{(j)}(1)}, \ldots, \frac{\sum_{j=1}^{k} N^{(j)}(\xi_m)}{\sum_{j=1}^{k} N^{(j)}(1)}\right)' \tag{12}$$

and go to the next stage with $g^{(k)} = g_{\lambda(k)}$.

$\vdots$

Conclusion. When the stopping rule yields "stop AIS" (at the kth stage, for instance), estimate $E\varphi$ by

$$\hat{E}^{(k)}\varphi = \left(\frac{\sum_{j=1}^{k} N^{(j)}(\varphi_1)}{\sum_{j=1}^{k} N^{(j)}(1)}, \ldots, \frac{\sum_{j=1}^{k} N^{(j)}(\varphi_\ell)}{\sum_{j=1}^{k} N^{(j)}(1)}\right). \tag{13}$$

Note that the AIS estimates of $E\xi$ and $E\varphi$ are obtained by linear accumulation of the $N^{(j)}(\cdot)$. This linear accumulation of statistics from previous stages has several advantages.

First, it is cheap; see section 2.3 for details. Second, the sample sizes in the stages can be small or moderate, while still accumulating to give high accuracy overall. The advantage of small or moderate sample sizes in the stages is, of course, quicker adaptation of g_λ to f. Note that even a completely adaptive scheme, with each $\theta_i^{(j)}$ defining a new stage, is possible.

It should be mentioned that, often, there will be a need to convert $\lambda^{(k)}$ to a convenient form for utilization at the next stage. For example, when λ is to be matched to an estimate of the covariance matrix, Σ, of f, it is often necessary to change $\lambda^{(k)}$ to the form $T^{(k)}T^{(k)'}$, where $T^{(k)}$ is a lower triangular matrix, to efficiently generate random variates from $g^{(k)} = g_{\lambda(k)}$. Such considerations may argue for keeping the number of stages moderate.

2.2 Stopping Rules for AIS

First, consider the case of scalar φ. Under mild conditions, $\hat{E}^{(k)}\varphi$ converges to $E\varphi$ and $\lambda^{(k)}$ to $\lambda^* = E\xi$ with probability one $\Big($ as $n^{(k)} = \Sigma_{j=1}^{k} n_j \to \infty\Big)$, and $\hat{E}^{(k)}\varphi$ is approximately normally distributed with mean $E\varphi$ and variance $\sigma^2/n^{(k)}$, where σ^2 is given by equations (4) and (5) with $g = g_{\lambda *}$. This will be established in section 3.

Consider first the simple stopping rule: stop after a fixed number of stages with fixed sample sizes. Then, σ^2 can be estimated by

$$\hat{\sigma}^{2(k)} = \frac{1}{(\overline{w}^{(k)})^2}\Big(\hat{var}^{(k)}(\varphi w) - 2(\hat{E}^{(k)}\varphi)\hat{cov}^{(k)}(\varphi w, w) + \left(\hat{E}^{(k)}\varphi\right)^2 \hat{var}^{(k)}(w)\Big), \qquad (14)$$

where

$$\overline{w}^{(k)} = \sum_{j=1}^{k} N^{(j)}(1)/n^{(k)} \qquad (15)$$

$$\hat{var}^{(k)}(\varphi w) = \sum_{j=1}^{k} N^{(j)}\left(\varphi^2 w^{(j)}\right)/n^{(k)} - \left(\hat{E}^{(k)}\varphi\right)^2\left(\overline{w}^{(k)}\right)^2 \qquad (16)$$

$$\hat{cov}^{(k)}(\varphi w, w) = \sum_{j=1}^{k} N^{(j)}\left(\varphi w^{(j)}\right)/n^{(k)} - \hat{E}^{(k)}\varphi\left(\overline{w}^{(k)}\right)^2 \qquad (17)$$

$$\hat{var}^{(k)}(w) = \sum_{j=1}^{k} N^{(j)}\left(w^{(j)}\right)/n^{(k)} - \left(\overline{w}^{(k)}\right)^2. \qquad (18)$$

Note that the *new* sums in (16), (17), and (18) can be linearly accumulated between stages exactly as in the BIS algorithm.

Now, suppose one wants to guarantee that

$$P\left(\frac{|\hat{E}^{(k)}\varphi - E\varphi|}{|E\varphi|} \leq \varepsilon\right) \approx 1 - \eta. \tag{19}$$

Then, from approximate normality, it follows that $n^{(k)}$ should be chosen such that

$$\frac{\sigma^2}{n^{(k)}|E\varphi|^2} \leq \left(\frac{\varepsilon}{c(\eta)}\right)^2, \tag{20}$$

where $c(\eta)$ is the $(1-\eta/2)$th quantile of the standard normal distribution. A method for deciding when to stop AIS immediately suggests itself. Replace σ^2 and $E\varphi$ in (20) by $\hat{\sigma}^{2(k)}$ and $\hat{E}^{(k)}\varphi$, respectively, and stop AIS when

$$\frac{\hat{\sigma}^{2(k)}}{n^{(k)}|\hat{E}^{(k)}\varphi|^2} \leq \left(\frac{\varepsilon}{c(\eta)}\right)^2. \tag{21}$$

When $\varphi = (\varphi_1, \ldots, \varphi_\ell)'$ is a vector, one could calculate $\hat{\sigma}^{2(k)}$ for each φ_i and stop when (21) is simultaneously satisfied for all components. This can be expensive, however, if ℓ is large. A rough surrogate for the stopping rule in such a situation is to replace $\hat{\sigma}^{2(k)}/|\hat{E}^{(k)}\varphi|^2$ in (21) by $v\hat{a}r^{(k)}(w)/\left(\overline{w}^{(k)}\right)^2$. This yields the rule: stop if

$$\frac{v\hat{a}r^{(k)}(w)}{n^{(k)}\left(\overline{w}^{(k)}\right)^2} \leq (\varepsilon/c(\eta))^2. \tag{22}$$

Note that $v\hat{a}r^{(k)}(w)/\left(\overline{w}^{(k)}\right)^2$ is the term of (14) which, when scaled by $\left(\hat{E}^{(k)}\varphi\right)^2$, does not involve φ.

2.3 The Additional Calculations In AIS

Suppose that φ is a vector of the form $(\varphi_1, \ldots, \varphi_\ell)'$ and that n random variates are drawn in BIS. Then BIS requires roughly n computations of $\varphi_1(\theta), \ldots, \varphi_\ell(\theta)$, $f(\theta), g(\theta)$, and $2ln$ additions and multiplications.

In addition to the computations needed in BIS, AIS requires extra computations in updating the parameters. In stage 1 of the algorithm in section 2.1, the computation of $N^{(1)}(\xi_i)$, $i = 1, \ldots, m$, requires the extra n_1 computations of $\xi_i(\theta)$, $i = 1, \ldots, m$, and mn_1 multiplications and additions. Equation (9) requires an extra $2m$ additions and m multiplications(divisions). Thus, if AIS stops at the kth stage and $n^{(k)} = n$, then the total extra work of AIS is roughly n computations of $\xi_i(\theta)$, $i = 1, \ldots, m$, plus $m(n+2k)$ additions and $m(n+k)$ multiplications, and possibly k conversions as mentioned at the end of section 2.1.

The additions and multiplications are typically cheap compared to the other computations, such as generation of random variates and computation of $f(\theta)$, $\varphi(\theta)$ and $g^{(j)}(\theta)$. Note that the computations of the $\xi_i(\theta)$ are typically cheap because the $\xi_i(\theta)$ are often linear or polynomial functions. Finally, when (say) h of the φ_i are equal to λ_i (as is often the case when the λ_is are moments), then the computations related to $\lambda_1, \ldots, \lambda_h$ are not extra anymore.

2.4 Discussion and Comparison with Other Adaptive Schemes

2.4.1 Kloek and van Dijk's Adaptive Scheme

Kloek and van Dijk's adaptive scheme (K-D's) is the same as AIS, except that it doesn't pool the estimates at the current stage with the estimates from the previous stages. Indeed, at the kth stage it estimates $E\varphi$ by $N^{(k)}(\varphi)/N^{(k)}(1)$ (while AIS estimates $E\varphi$ by $\Sigma_{j=1}^{k} N^{(j)}(\varphi)/\Sigma_{j=1}^{k} N^{(j)}(1)$). Since estimates at each stage of K-D's scheme are weighted averages of the form (3), involving only i.i.d. random drawings, it is easy to establish convergence. Also, only random drawings from the current, supposedly most accurate, importance function are used at each stage. But because there is no accumulation of random drawings, the sample sizes in the stages must be large to have good accuracy, as van Dijk (1984) pointed out.

A reasonable way to running K-D's scheme would be to have only a few stages with

greatly increasing sample sizes in the stages, so that the final (most accurate) stage receives most of the observations. There are, however, two inefficiencies in choosing greatly increasing sample sizes. The first is that one can *overshoot* the desired accuracy by having a too-large final stage. The second disadvantage of this mode of operation is that it typically requires an interaction with the statistician, to choose the sample size needed at the next stage. In contrast, AIS, operating with stages of fixed moderate sample size, can operate *on automatic.* Numerical example of K-D's scheme and AIS, with gradually increasing sample sizes in the stages, are given in section 2.5.

2.4.2 Naylor and Smith's Adaptive Scheme

Naylor and Smith's adaptive scheme (N-S) is similar to K-D's, but it updates not only the parameters of the importance function but also the form of the importance function; see Naylor and Smith (1988). Again, however, this typically requires interaction with the statistician to choose a new importance function and sample size for the next stage. If feasible, this is highly desirable, but there may be many situations in which complete automation is required.

Even within Naylor and Smith's scheme, there is a possible role for AIS: one can have a small number of stages of N-S, in each stage of which one uses AIS to estimate the information needed to update the importance function for the next N-S stage. Again, it is because AIS can be automated that such is possible.

2.5 Examples and Numerical Comparisons

2.5.1 Example 1

In order to be able to easily compare several Importance Sampling methods, we start by considering a simple two dimensional situation. Suppose $f(\theta)$ is

$$f(\theta) = .25f_1(\theta) + .75f_2(\theta), \tag{23}$$

where $\theta \in R^2$ and f_1, f_2 are the density functions corresponding to the

$$N\left(\begin{bmatrix}0.0\\0.0\end{bmatrix}, \begin{bmatrix}1.0 & 0.8\\0.8 & 1.0\end{bmatrix}\right) \text{ and } N\left(\begin{bmatrix}2.1\\2.1\end{bmatrix}, \begin{bmatrix}1.0 & 0.0\\0.0 & 1.0\end{bmatrix}\right)$$

distributions, respectively. Assume that one desires to find the mean μ, variance Σ, and the probability $P = P(\theta \leq (2.0, 5.0)')$. Thus, let φ be the vector $(\varphi_1, \ldots, \varphi_6)' = (\theta_1, \theta_2, \theta_1^2, \theta_1\theta_2, \theta_2^2, I(\theta \leq (2.0, 5.0)'))'$, where θ_1, θ_2 are the 1st and 2nd elements of θ, respectively. Defining

$$\varphi^\dagger = (\varphi_1, \varphi_2)', \quad \varphi^{\dagger\dagger} = \begin{pmatrix}\varphi_3 & \varphi_4\\ \varphi_4 & \varphi_5\end{pmatrix} \tag{24}$$

clearly

$$\mu = E\varphi^\dagger, \quad \Sigma = E\varphi^{\dagger\dagger} - \left(E\varphi^\dagger\right)\left(E\varphi^\dagger\right)', \quad P = E\varphi_6. \tag{25}$$

As shown in Fig. 1, f is skewed in the direction of $(0, 0)'$. Since the skewness is not severe, use of a standard unimodal symmetric importance function may be reasonable. We chose a multivariate t form with 1 degree of freedom for the importance function because of its thick tails and simplicity in random variate generation (see section 1.2). Thus, the class of importance functions we consider is

$$\mathcal{G} = \{g_{\lambda^\dagger, \lambda^{\dagger\dagger}}(\theta) \propto |\lambda^{\dagger\dagger}|^{-1/2}(1 + (\theta - \lambda^\dagger)'\lambda^{\dagger\dagger -1}(\theta - \lambda^\dagger))^{-3/2}, \lambda^\dagger \in R^2, \lambda^{\dagger\dagger} \text{ is p.d.}\}, \tag{26}$$

where

$$\lambda^\dagger = (\lambda_1, \lambda_2)' \text{ and } \lambda^{\dagger\dagger} = \begin{pmatrix}\lambda_3 & \lambda_4\\ \lambda_4 & \lambda_5\end{pmatrix}. \tag{27}$$

(As usual, define $\lambda = (\lambda_1, \ldots, \lambda_5)'$).

It is natural to let $\lambda^\dagger$, the location parameter for $g_{\lambda^\dagger, \lambda^{\dagger\dagger}}$, be an estimate of the mean, μ, of f. (We could of course calculate the mean of f in closed form here, but we are investigating the algorithm.) Thus, in the algorithm, choose $\xi^\dagger(\theta) = (\xi_1(\theta), \xi_2(\theta))' = \theta$. And in Oh (1989), it is argued that, if one has an estimate $\hat{\Sigma}$ of the covariance matrix for a density f, then a good choice of $\lambda^{\dagger\dagger}$ in $g_{\lambda^\dagger, \lambda^{\dagger\dagger}}$ is (for two dimensions) $\lambda^{\dagger\dagger} = (.65)\hat{\Sigma}$. Since the covariance

matrix of f can be estimated in the algorithm by approximating $E(\theta\theta' - (E\xi^{\dagger})(E\xi^{\dagger})')$, the suggested choice for the remaining $\xi_i(\theta)$ in the algorithm is

$$\xi^{\dagger\dagger}(\theta) = \begin{pmatrix} \xi_3(\theta) & \xi_4(\theta) \\ \xi_4(\theta) & \xi_5(\theta) \end{pmatrix} = (.65)\Big(\theta\theta' - \Big(E\xi^{\dagger}\Big)\Big(E\xi^{\dagger}\Big)'\Big). \tag{28}$$

Note that $\xi^{\dagger} = \varphi^{\dagger}$ and $\xi^{\dagger\dagger} = (.65)\left(\varphi^{\dagger\dagger} - \left(E\varphi^{\dagger}\right)\left(E\varphi^{\dagger}\right)'\right)$, so that there will be no extra calculations in updating λ.

As a final preliminary to apply the AIS algorithm, we need to choose sample sizes in each stage and a stopping rule. The sample sizes that will be used are 200 in the first stage and 100 in all subsequent stages. The stopping rule used is that specified by (22), with $\eta = .05, \varepsilon = .01$; thus we will stop when

$$\hat{var}^{(k)}(w) / \left\{ n^{(k)} \left(\overline{w}^{(k)}\right)^2 \right\} \leq 2.60308 \times 10^{-5}. \tag{29}$$

In our comparisons, we will also apply BIS in this framework. Of course, in BIS, λ is not updated.

At stage 0 of the AIS algorithm, we need preliminary $\lambda^{\dagger(0)}$, and $\lambda^{\dagger\dagger(0)}$. It is common to initiate Importance Sampling with likelihood estimates (here the mode, $\hat{\theta}$, and minus inverse Hessian, $-\hat{I}^{-1}$, of f) to approximate the mean μ and covariance matrix Σ of f. Thus, we shall set $\lambda^{\dagger(0)} = \hat{\theta}$ and $\lambda^{\dagger\dagger(0)} = (-.65)\hat{I}^{-1}$; here these turn out to be

$$\lambda^{\dagger(0)} = \begin{pmatrix} 2.0 \\ 2.0 \end{pmatrix}, \lambda^{\dagger\dagger(0)} = \begin{pmatrix} 1.30 & 1.26 \\ 1.26 & 1.30 \end{pmatrix}, \tag{30}$$

defining the stage 1 importance function $g^{(0)}$.

At stage 1, we drew $n_1 = 200$ random variates from $g^{(0)}$ and from these random variates computed

$$N^{(1)}(1) = \sum_{i=1}^{200} w^{(1)}\left(\theta_i^{(1)}\right) = 188.0 \tag{31}$$

$$N^{(1)}\left(\varphi^{\dagger}\right) = N^{(1)}\left(\xi^{\dagger}\right) = \sum_{i-1}^{200} \theta_i^{(1)} w^{(1)}\left(\theta_i^{(1)}\right) = (329.8, 341.5)' \tag{32}$$

$$N^{(1)}\left(\varphi^{\dagger\dagger}\right) = \sum_{i-1}^{200}\theta_i^{(1)}\theta_i^{(1)\prime}w^{(1)}\left(\theta_i^{(1)}\right) = \begin{pmatrix} 841.1 & 656.7 \\ 656.7 & 851.8 \end{pmatrix} \tag{33}$$

$$N^{(1)}(\varphi_6) = \sum_{i=1}^{200} w^{(1)}\left(\theta_i^{(1)}\right) I\left(\theta_i^{(1)} \leq (2.0, 5.0)'\right) = 106.0. \tag{34}$$

To check the stopping rule (22), we also needed to calculate

$$\begin{aligned} \hat{var}^{(1)}(w)/\left\{n^{(1)}(\overline{w}^{(1)})^2\right\} &= \frac{1}{n^{(1)}}\left\{\frac{N^{(1)}\left(w^{(1)}\right)}{n^{(1)}} - \left(\frac{N^{(1)}(1)}{n^{(1)}}\right)^2\right\}\Big/\left(\frac{N^{(1)}(1)}{n^{(1)}}\right)^2 \\ &= 4.16. \end{aligned} \tag{35}$$

Since $\hat{var}^{(1)}(w)/\{n^{(1)}(\overline{w}^{(1)})^2\} > 2.60308 \times 10^{-5}$, we did not stop, and hence calculated

$$\lambda^{\dagger(1)} = \left(\frac{329.8}{188.0}, \frac{341.5}{188.0}\right)' = (1.75, 1.82)' \tag{36}$$

$$\lambda^{\dagger\dagger(1)} = 0.65\left\{\begin{pmatrix} \frac{841.1}{188.0} & \frac{656.7}{188.0} \\ \frac{656.7}{188.0} & \frac{851.6}{188.0} \end{pmatrix} - \lambda^{\dagger(1)}\lambda^{\dagger(1)\prime}\right\} = \begin{pmatrix} 0.91 & 0.10 \\ 0.10 & 0.91 \end{pmatrix}. \tag{37}$$

These defined the stage 2 importance function $g^{(1)} = g_{\lambda^{\dagger(1)},\lambda^{\dagger\dagger(1)}}$.

AIS continued in this example to the 320th stage, at which point $\hat{\mu}, \hat{\Sigma}, \hat{P}$ (the estimates of μ, Σ, P, respectively) were obtained from $\hat{E}^{(320)}\varphi$, using the relations in (25). The results are given in the Table 1; $\hat{\mu}_1, \hat{\mu}_2$ are estimates of the 1st and 2nd elements of $\hat{\mu}$, respectively, and $\hat{\sigma}_1^2, \hat{\sigma}_{12}, \hat{\sigma}_2^2$ are estimates of the 1st diagonal, off diagonal, and 2nd diagonal elements of $\hat{\Sigma}$, respectively. The values in the column 'True Values' are the exact values of μ, Σ, and P, which can be determined analytically in this example. And $\hat{var}(\hat{\mu}_1)$, $\hat{var}(\hat{\mu}_2)$, and $\hat{var}(\hat{P}(\theta \leq (2., 5.)'))$ are the estimates of $var(\hat{\mu}_1)$, $var(\hat{\mu}_2)$, and $var(\hat{P}(\theta \leq (2., 5.)'))$ of the form $\hat{\sigma}^{2(k)}/n^{(k)}$, where $\hat{\sigma}^{2(k)}$ is given in (14). Unfortunately, $\hat{var}(\hat{\sigma}_1^2)$, $\hat{var}(\hat{\sigma}_{12})$, and $\hat{var}(\hat{\sigma}_2^2)$ cannot be calculated from (14), since σ_1^2, σ_2^2, and σ_{12} are not representable as $E\varphi$ for some $\varphi(\theta)$. However, from Cramer (1946, Sec. 28.4), analogous formulas can be derived for $\hat{var}(\hat{\sigma}_1^2), \hat{var}(\hat{\sigma}_2^2)$, and $\hat{var}(\hat{\sigma}_{12})$, and were used in this section.

To judge the effect of the updating of λ, Table 1 also lists the corresponding results for BIS (i.e., when BIS was run until the stopping rule (22) was satisfied). AIS achieved the specified accuracy (29) after many fewer iterations than BIS and was substantially faster.

Table 1: Comparison of BIS and AIS

	BIS	AIS	True Values
n	57,600	33,000	
$\hat{\mu}_1$	1.5934	1.5837	1.5750
$\hat{\mu}_2$	1.5986	1.5953	1.5750
$\hat{var}(\hat{\mu}_1)/\hat{\mu}_1^2$	5.8135×10^{-5}	5.1728×10^{-5}	
$\hat{var}(\hat{\mu}_2)/\hat{\mu}_2^2$	5.7792×10^{-5}	5.0939×10^{-5}	
$\hat{\sigma}_1^2$	1.8396	1.8007	1.8269
$\hat{\sigma}_{12}$	1.0252	1.0094	1.0269
$\hat{\sigma}_2^2$	1.8379	1.8138	1.8269
$\hat{var}(\hat{\sigma}_1^2)/(\hat{\sigma}_1^2)^2$	1.5315×10^{-4}	1.1729×10^{-4}	
$\hat{var}(\hat{\sigma}_{12})/(\hat{\sigma}_{12})^2$	1.1587×10^{-3}	3.8156×10^{-4}	
$\hat{var}(\hat{\sigma}_2^2)/(\hat{\sigma}_2^2)^2$	1.5470×10^{-4}	1.1470×10^{-4}	
$\hat{P}(\theta \leq (2.,5.)')$	0.5830	0.5834	.5919
$\hat{var}(\hat{P}(\theta \leq (2.,5.)'))/(\hat{P}(\theta \leq (2.,5.)')^2$	2.6591×10^{-5}	4.0216×10^{-5}	
$\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$	2.6029×10^{-5}	2.5957×10^{-5}	
time (sec)	295.0	169.3	

Note that the mode and minus inverse Hessian (the starting values of λ for AIS and the permanent values of λ for BIS) are quite different from the actual mean and covariance matrix.

Next, AIS and K-D's adaptive scheme are compared for a fixed number of stages and a fixed sample size in each stage. Note that in K-D's scheme the estimates are not pooled across stages. In both schemes we chose 7 stages, with sample sizes in the stages of 500, 1000, 1500, 2500, 3500, 4500, and 6500, respectively; thus the total sample size is 20000 and the sample sizes in the stages are gradually increasing. The results are given in Table 2. Since we forced both schemes to take 20,000 samples, the times of computation are quite similar. Note however, that AIS yields answers with variances that are about 1/3 those of K-D's scheme.

Observe, from Table 1 and Table 2, that $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ is about 1/2 of $\hat{var}(\hat{\mu}_i)/(\hat{\mu}_i)^2$, $i = 1, 2$. On the other hand, $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ differs from $\hat{var}(\hat{\sigma}_i^2)/(\hat{\sigma}_i^2)^2$, $i = 1, 2$, and $\hat{var}(\hat{\sigma}_{12})/(\hat{\sigma}_{12})^2$ by factors ranging from 4 to 14. Thus, use of the stopping rule (22) with $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ seems to be a reasonable surrogate for (20) in estimation of means, but less so for estimation of variances.

It is of interest to see how the $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$, $k = 1, \ldots, 7$, behave as a function of k. Table 3 gives $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ for each stage. Clearly, it is decreasing much faster in AIS than in K-D's scheme.

2.5.2 Example 2

As a more realistic example, we consider an example from Fong (1987), which analyzed a complete block design of Stenstrom (given in SAS (1985), p. 487) using a hierarchical Bayesian model. Data $\mathbf{y}$ was assumed to follow the model

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K_j, \tag{38}$$

where $I = 3$, $J = 4$, $K_1 = K_2 = K_3 = 2$, $K_4 = 1$, and $\epsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \tau^2)$, for unknown τ^2. Suppose $\mu_{ij} = \mu + \alpha_i + \beta_j$, μ has $N(33.5, 9.0)$ for the prior density, and the prior distributions

Table 2: Comparison of K-D's scheme and AIS

	K-D's	AIS	True Values
$\hat{\mu}_1$	1.6081	1.5947	1.5750
$\hat{\mu}_2$	1.6154	1.6034	1.5750
$\hat{var}(\hat{\mu}_1)/\hat{\mu}_1^2$	2.4549×10^{-4}	8.3722×10^{-4}	
$\hat{var}(\hat{\mu}_2)/\hat{\mu}_2^2$	2.4590×10^{-4}	8.2812×10^{-4}	
$\hat{\sigma}_1^2$	1.7785	1.7762	1.8269
$\hat{\sigma}_{12}$	0.9775	0.9905	1.0269
$\hat{\sigma}_2^2$	1.7467	1.7967	1.8269
$\hat{var}(\hat{\sigma}_1^2)/(\hat{\sigma}_1^2)^2$	5.9480×10^{-4}	1.9964×10^{-4}	
$\hat{var}(\hat{\sigma}_{12})/(\hat{\sigma}_{12})^2$	1.8978×10^{-3}	6.5488×10^{-4}	
$\hat{var}(\hat{\sigma}_2^2)/(\hat{\sigma}_2^2)^2$	6.2941×10^{-4}	1.9446×10^{-4}	
$\hat{P}(\theta \leq (2.,5.)')$	0.5764	0.5811	.5919
$\hat{var}^{(k)}[\hat{P}(\theta \leq (2.,5.)')]/(\hat{P}(\theta \leq (2.,5.)'))^2$	2.0793×10^{-4}	6.7478×10^{-5}	
$\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$	1.2905×10^{-4}	4.3350×10^{-5}	
time (sec)	94.0	97.1	

Table 3: $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$

stage	sample size	K-D's	AIS
1	500	2.6450×10^{-3}	2.6450×10^{-3}
2	1000	8.9219×10^{-4}	6.9655×10^{-4}
3	1500	5.9062×10^{-4}	3.2073×10^{-4}
4	2500	3.5022×10^{-4}	1.6739×10^{-4}
5	3500	2.4863×10^{-4}	1.0008×10^{-4}
6	4500	1.8711×10^{-4}	6.5152×10^{-5}
7	6500	1.2905×10^{-4}	4.3350×10^{-5}

of α_i and β_j, given τ_α^2, τ_β^2, and τ^2, are

$$\alpha_i \overset{i.i.d.}{\sim} N(0, \tau_\alpha^2),\ \beta_j \overset{i.i.d.}{\sim} N(0, \tau_\beta^2); \tag{39}$$

the second stage prior density of $(\tau_\alpha^2, \tau_\beta^2, \tau^2)$ is

$$\begin{aligned} \pi(\tau_\alpha^2, \tau_\beta^2, \tau^2) &= \pi(\tau_\alpha^2|\tau^2)\pi(\tau_\beta^2|\tau^2)\pi(\tau^2) && (40) \\ &= \frac{1}{(\tau_\alpha^2+\tau^2)(\tau_\beta^2+\tau^2)\tau^2}\,. && (41) \end{aligned}$$

Determining the probabilities $P(\beta_j$ is the largest$|\ \mathbf{y})$ is of great interest in ranking and selection. In Fong (1987), it is shown that calculation of these probabilities reduces to determining the posterior expectations of

$$\psi_j(\tau^2, \tau_\alpha^2, \tau_\beta^2) = E^\beta E^{\phi_j} \prod_{s \neq j} \Phi\left(\frac{\phi_j - u_s}{\sqrt{V_s}}\right), j = 1, \ldots, J, \tag{42}$$

where

$$\begin{aligned} \beta &\sim N(u^*, V^*), \phi_j \sim N(u_j, V_j) && (43) \\ u^* &= \tilde{y}_{...} - \left[1 + (\tau_\alpha^2 + 9.0I)\tau_b^{-2}\right]^{-1}(\tilde{y}_{...} - 33.5) && (44) \\ V^* &= [I((\tau_\alpha^2 + 9.0I)^{-1} + \tau_b^{-2})]^{-1} && (45) \end{aligned}$$

$$\tilde{y}_{...} = \frac{\Sigma_{i=1}^{I}\Sigma_{j-1}^{J}\Sigma_{k=1}^{K_j} y_{ijk}/(\tau^2 + IK_j\tau_\beta^2)}{I\tau_b^{-2}} \tag{46}$$

$$\tau_b^2 = \left(\sum_{j=1}^{J} \frac{1}{\tau^2/K_j + I\tau_\beta^2}\right)^{-1} \tag{47}$$

$$u_j = \overline{y}_{.j.} - \frac{\tau^2/K_j}{\tau^2/K_j + I\tau_\beta^2}(\overline{y}_{.j.} - \beta) \tag{48}$$

$$V_j = \frac{(\tau^2/K_j)\tau_\beta^2}{\tau^2/K_j + I\tau_\beta^2},\ j = 1, \ldots, J, \tag{49}$$

$\overline{y}_{.j.}$ is a constant obtained from data $\mathbf{y}$, and Φ is the standard normal cdf.

To compute the posterior expectations, first transform $\tau^2, \tau_\alpha^2, \tau_\beta^2$ to

$$\theta_1 = log\tau^2, \quad \theta_2 = log\tau_\alpha^2, \quad \theta_3 = log\tau_\beta^2, \tag{50}$$

since this transformation often makes the posterior density function less skewed and changes the ranges of the variables to $(-\infty, \infty)$. Let $\theta = (\theta_1, \theta_2, \theta_3)'$ and $\varphi_j(\theta) = \psi_j(e^{\theta_1}, e^{\theta_2}, e^{\theta_3}), j = 1, \ldots, J$. The transformed posterior density can be shown to be

$$\begin{aligned} P(\theta|\mathbf{y}) \propto \quad & e^{-(IN_k+I+J-1)\theta_1/2}(e^{\theta_1} + N_k e^{\theta_2})^{-(I-1)/2} \\ \times \quad & \prod_{j=1}^{J}\left(\frac{e^{\theta_1}}{K_j} + Ie^{\theta_3}\right)^{-1/2}\left(A(e^{\theta_1}, e^{\theta_3}) + \frac{1}{e^{\theta_2} + 9.0I}\right)^{-1/2} \\ \times \quad & exp\Big[-\frac{1}{2}\left(\frac{S_3}{e^{\theta_1}} + \frac{N_k S_1}{e^{\theta_1} + N_k e^{\theta_2}} + I\sum_{j=1}^{J}\frac{(\overline{y}_{.j.} - \tilde{y}_{...}^2)^2}{e^{\theta_1}/K_j + Ie^{\theta_3}}\right) \\ + \quad & \frac{I(\tilde{y}_{...} - 33.5)^2 A(e^{\theta_1}, e^{\theta_3})}{1 + (e^{\theta_2} + 9.0I)A(e^{\theta_1}, e^{\theta_3})}\Big]\frac{e^{\theta_2+\theta_3}}{(e^{\theta_1} + e^{\theta_2})(e^{\theta_1} + e^{\theta_3})}, \end{aligned} \tag{51}$$

where

$$A(e^{\theta_1}, e^{\theta_3}) = \sum_{j=1}^{J}\frac{1}{e^{\theta_1}/K_j + Ie^{\theta_3}}, \tag{52}$$

S_1, S_3 are constants obtained from $\mathbf{y}$, $N_k = \Sigma_{j=1}^{J} K_j$, and $\tilde{y}_{...}$ is given in (46).

As in the previous example, a multivariate t form with 7 degrees of freedom is chosen for the importance function, so

$$\mathcal{G} = \{g_{\lambda\dagger,\dagger\dagger}(\theta) \propto |\lambda^{\dagger\dagger}|^{-1/2}(1 + \frac{1}{7}(\theta - \lambda^\dagger)'\lambda^{\dagger\dagger -1}(\theta - \lambda^\dagger))^{-5}, \lambda^\dagger \in R^3, \lambda^{\dagger\dagger} \text{ is p.d.}\}. \tag{53}$$

Again, estimated moments, $\hat{\mu}$ and $\hat{\Sigma}$, of the posterior yield values $\lambda^{\dagger} = \hat{\mu}$ and $\lambda^{\dagger\dagger} = (.8609)\hat{\Sigma}$ for the parameters of the importance function (.8609 being the adjustment factor recommended in Oh (1989), for multivariate t density function with 7 degrees of freedom in three dimensional problems). Thus, $\xi^{\dagger}(\theta) = \theta$ and $\xi^{\dagger\dagger}(\theta) = .8609(\theta\theta' - E\xi^{\dagger}E\xi^{\dagger'})$ will be used in the AIS algorithm.

A sample size of 200 was selected for each stage, and the stopping rule (21) with $\eta = .05$ and $\varepsilon = .2$ was chosen for each element of φ; thus sampling stopped when

$$\frac{\hat{var}(\hat{E}^{(k)}\varphi_j)}{(\hat{E}^{(k)}\varphi_j)^2} \leq 1.04123 \times 10^{-2}, \text{ for } j = 1, \ldots, 4. \tag{54}$$

Note that $\hat{var}(\hat{E}^{(k)}\varphi)$ is of the form $\hat{\sigma}^{2(k)}/n^{(k)}$, where $\hat{\sigma}^{2(k)}$ is given in (14) with $\varphi = \varphi_j$. Finally, the importance function at stage 0 was chosen to have $\lambda^{\dagger(0)} = \hat{\theta}$ (the posterior mode) and $\lambda^{\dagger\dagger(0)} = (-.8609)\hat{I}^{-1}$($-\hat{I}^{-1}$ being minus the inverse Hessian); these were determined using maximum likelihood methods.

With the stopping rule mentioned above, AIS stopped after 53 stages ($n = 10600$), 9 times earlier than BIS which stopped after 470 stages ($n = 94000$). However, $\hat{var}(\hat{E}\varphi_i)/(\hat{E}\varphi_i)^2, i = 1, 2, 3$, in BIS were about 1/3 of those in AIS because large $\hat{var}(\hat{E}\varphi_4)/(\hat{E}\varphi_4)^2$ controlled the stopping rule. Thus, to compare both schemes it would be more reasonable to run BIS with 160 stages ($n = 32000$) and AIS with 53 stages so that most variances are roughly the same. The results are given in Table 4.

There is no regular relationships between $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ and $\hat{var}\hat{E}\varphi_i/(\hat{E}\varphi_i)^2, i = 1, \ldots, 4$, in this example. And because the $\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$ are for the most part smaller than the $\hat{var}\hat{E}\varphi_i/(\hat{E}\varphi_i)^2$, use of the stopping rule (22) would have caused the scheme to stop much earlier, resulting in less accuracy of the $\hat{E}\varphi_i, i = 1, \ldots, 4$.

It is of interest to compare the mode and minus inverse Hessian (which are not only the parameter values used in BIS, but are also the estimates that would result from a likelihood approach to the problem) with AIS estimates of the mean and covariance matrix. Tables 5 and 6 show that there are substantial differences between the mode and mean, and between minus the inverse Hessian and the variance of θ_2.

Table 4: Comparison of BIS and AIS in Example 2

	BIS	AIS
n	32,000	10,600
time (sec)	1033.0	361.0
$\hat{\mu}$	(1.354, 1.619, 1.667)	(1.445, 2.021, 1.759)
$\hat{\Sigma}$	.129	.121
	.195, 1.844	.021, 2.438
	.126, .144, .938	-.015, .038, 1.375
$\hat{E}\varphi_1$	.0400	.0376
$\hat{E}\varphi_2$	.0411	.0419
$\hat{E}\varphi_3$	.9176	.9177
$\hat{E}\varphi_4$	.0037	.0028
$\hat{var}(\hat{E}\varphi_1)/(\hat{E}\varphi_1)^2$	1.1640×10^{-3}	6.8301×10^{-4}
$\hat{var}(\hat{E}\varphi_2)/(\hat{E}\varphi_2)^2$	4.7175×10^{-4}	7.0058×10^{-4}
$\hat{var}(\hat{E}\varphi_3)/(\hat{E}\varphi_3)^2$	3.8291×10^{-6}	3.5244×10^{-6}
$\hat{var}(\hat{E}\varphi_4)/(\hat{E}\varphi_4)^2$	4.5000×10^{-2}	1.0365×10^{-2}
$\hat{var}^{(k)}(w)/\{n^{(k)}(\overline{w}^{(k)})^2\}$	2.2558×10^{-5}	1.4106×10^{-5}

Table 5: Mode and mean of posterior

i	1	2	3
$\hat{\theta}_i$	1.355	1.619	1.667
$\hat{\mu}_i$	1.445	2.021	1.759

Table 6: Minus inverse Hessian and covariance matrix of posterior

(i, j)	(1, 1)	(1, 2)	(1, 3)	(2, 2)	(2, 3)	(3, 3)
$\hat{-I^{-1}}$	.129	.195	.126	1.844	.144	.938
$\hat{\Sigma}$	.121	.021	-.015	2.438	.038	1.375

3 Convergence

For $\hat{E}^{(k)}\varphi$ to be a meaningful estimate of $E\varphi$, it should be consistent. Also, to measure the accuracy of $\hat{E}^{(k)}\varphi$ it is helpful to have an asymptotic distribution for $\hat{E}^{(k)}\varphi$. In BIS, since all the random variates used to compute $\hat{E}^{B}\varphi$ are i.i.d., the asymptotic theory for $\hat{E}^{B}\varphi$ is trivial. But in AIS, the importance function at stage k depends on previous random variates, so asymptotic theory for dependent random variables must be used. Section 3.2 establishes consistency of $\hat{E}^{(k)}\varphi$ under reasonable assumptions, while section 3.3 establishes asymptotic normality.

3.1 Notations and Assumptions

For convenience, we let $E^*(\cdot)$ represent expectation over random variates $\theta_i^{(j)}, i = 1, \ldots, n_j$, $j = 1, 2, \ldots$ (E^* is just a notational device; we will only actually be taking expectations over finite sets of $\{\theta_i^{(j)}\}$.) Let $\theta_i^{(j)}$, for $i = 1, \ldots, n_j, j = 1, 2, \ldots$, be real valued Borel measurable random variables, and let $\mathcal{F}_k$ be the σ-field generated by $\underline{\theta}^{(1)}, \ldots, \underline{\theta}^{(k)}$, where $\underline{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_{n_j}^{(j)}), j = 1, 2, \ldots$ Finally define, for any measurable function φ,

$$t_j^{\varphi}(\theta_i^{(j)}) = \varphi(\theta_i^{(j)}) w^{(j)}(\theta_i^{(j)}) - \int \varphi(\theta) f(\theta) d\theta, \; i = 1, \ldots, n_j, \; j = 1, 2, \ldots \tag{55}$$

$$T_j^{\varphi}(\underline{\theta}^{(j)}) = \sum_{i=1}^{n_j} t_j^{\varphi}(\theta_i^{(j)}), \; j = 1, 2, \ldots, . \tag{56}$$

The following assumptions will be used in this section.

Assumption I: $g^{(k)}, k = 1, 2, \ldots$, have the same support as f.

Assumption II: $E\varphi$ exists.

Assumption III: $w^{(k)}$ is bounded by a constant M for all $k = 1, 2, \ldots$.

Assumption IV: $E\xi$ exists.

Assumption V: g_λ is a continuous function of λ.

These assumptions are fairly innocuous, except for Assumption III, which is often not satisfied for an unconstrained family of density function $\mathcal{G}$. For example, $\mathcal{G} = \{g_{\mu,\Sigma}; g_{\mu,\Sigma}$ is a $\mathcal{T}_1(\mu, \Sigma)$ density function, $\mu \in R^p, \Sigma$ is p.d.$\}$ will typically lead to unbounded $w^{(k)}$ by letting $|\Sigma| \to \infty$ (so that the denominator of $w^{(k)}$ goes to zero). Choosing $\mathcal{G} = \{g_\lambda, \lambda \in \Lambda\}$, where Λ is a compact set and the tails of g_λ are heavier than those of f, typically does guarantee assumption III. Use of a constrained Λ could be harmful to the efficiency of AIS, however, and the consistency theorem is undoubtedly actually true with much weaker assumptions. In practice, therefore we suggest choosing unconstrained $\mathcal{G}$, but monitoring λ to see that $\lambda^{(k)}$ doesn't wander off to the boundaries or infinity.

3.2 Consistency

Lemma 3.1 *Under assumptions I and II,*

$$(i). \qquad \text{Given } \mathcal{F}_{j-1},\ t_j^\varphi(\theta_i^{(j)}),\ i = 1, \ldots, n_j \text{ are i.i.d.} \tag{57}$$

$$(ii). \qquad E^*(t_j^\varphi(\theta_i^{(j)})) = E^*(E^*(t_j^\varphi(\theta_j^{(j)})|\mathcal{F}_{j-1})) = 0. \tag{58}$$

Proof. Because, given $\underline{\theta}^{(l)}, l = 1, \ldots, j-1$, the $\theta_i^{(j)}, i = 1, \ldots, n_j$, are i.i.d. random variates from $g^{(j-1)}(\theta)$, result (i) is obvious. Now,

$$\begin{aligned} E^*(t_j^\varphi(\theta_i^{(j)})|\mathcal{F}_{j-1}) &= E^*(\varphi(\theta_i^{(j)})w^{(j)}(\theta_i^{(j)})|\mathcal{F}_{j-1}) - \int \varphi(\theta)f(\theta)d\theta \\ &= \int (\varphi(\theta_i^{(j)})(f(\theta_i^{(j)})/g^{(j-1)}(\theta_i^{(j)}))g^{(j-1)}(\theta_i^{(j)})d\theta_i^{(j)} - \int \varphi(\theta)f(\theta)d\theta \\ &= \int (\varphi(\theta_i^{(j)})f(\theta_i^{(j)})d\theta_i^{(j)} - \int \varphi(\theta)f(\theta)d\theta \\ &= 0. \end{aligned}$$

Hence,

$$E^*(t_j^\varphi(\theta_i^{(j)})) = E^*(E^*(t_j^\varphi(\theta_i^{(j)})|\mathcal{F}_{j-1})) = 0. \quad \square \tag{59}$$

Lemma 3.2 *Under assumption II,*

$$E^*((T_j^{\varphi}(\underline{\theta}^{(j)}))^2|\mathcal{F}_{j-1}) = n_j\Big\{\int \varphi^2(\theta)w^{(j)}(\theta)f(\theta)d\theta - (\int \varphi(\theta)f(\theta)d\theta)^2\Big\}. \tag{60}$$

Proof.

$$\begin{aligned} E^*((T_j^{\varphi}(\underline{\theta}^{(j)}))^2|\mathcal{F}_{j-1}) &= E^*((\sum_{i=1}^{n_j} t_j^{\varphi}(\theta_i^{(j)}))^2|\mathcal{F}_{j-1}) \\ &= \sum_{i=1}^{n_j} E^*((t_j^{\varphi}(\theta_i^{(j)}))^2|\mathcal{F}_{j-1}) \text{ by Lemma 3.1} \\ &= n_j\{\int \varphi^2(\theta)w^{(j)}(\theta)f(\theta)d\theta - (\int \varphi(\theta)f(\theta)d\theta)^2\} \end{aligned}$$

since, given $\mathcal{F}_{j-1}$, $\theta_i^{(j)}, i = 1, \ldots, n_j$, are from $g^{(j-1)}$. □

Theorem 3.1 *Suppose that assumptions I, II and III hold and that φ has finite second moment. Then*

$$\hat{E}^{(k)}\varphi \xrightarrow{a.s.} E\varphi \quad as \quad n^{(k)} \longrightarrow \infty. \tag{61}$$

Proof. Given in the appendix. □

Corollary 3.1 *Suppose that assumptions I, II, IV, and V hold and that ξ has finite second moment. Let $\lambda^{(k)} = \hat{E}^{(k)}\xi$, $\lambda^* = E\xi$; then, for each θ,*

$$g^{(k)}(\theta) \xrightarrow{a.s.} g_{\lambda^*}(\theta) \quad as \quad n^{(k)} \longrightarrow \infty. \tag{62}$$

Proof. From Theorem 3.1, $\lambda^{(k)} \xrightarrow{a.s.} \lambda^*$. The convergence (62) follows from Chow and Teicher (1978, Cor. 2, p. 67). □

3.3 Approximate Normality

As mentioned in section 1.1, $\hat{E}^B\varphi$ is approximately normally distributed when the sample size is large enough. Here, it will be shown that $\hat{E}^{(k)}\varphi$ has the same property.

Theorem 3.2 *Suppose assumptions I – V hold, φ has finite fourth moment, and $\Sigma_{j=1}^{k} n_j^2/(n^{(k)})^2 \to 0$ as $k \to \infty$ (which is true, for instance, if all $n_j \leq N$). Then*

$$a) \qquad \sqrt{n^{(k)}}(\hat{E}^{(k)}\varphi - E\varphi) \xrightarrow{d} N(0, \sigma^2), \tag{63}$$

$$b) \qquad \hat{\sigma}^{2(k)} \xrightarrow{a.s.} \sigma^2. \tag{64}$$

where $\hat{\sigma}^{2(k)}$ is given in (14) and σ^2 is given by (4) and (5) with $g^* = g_{\lambda^*}$ for g.

Proof. Given in the appendix. □

4 Conclusions and Generalizations

The advantage of AIS include its efficiency, its potential for being automated, and its flexibility. Its efficiency arises from the possibility of repeatedly improving the importance function during the simulation, from the utilization of all Monte Carlo observations in the computation, and from the utilization of cheap linear operations to update the importance function. The automatic nature of AIS is attractive in that once the class of importance functions is chosen, and the sample sizes and desired accuracy specified, there is no need for further interaction with the statistician.

Finally, AIS is flexible in terms of its features, allowing any class of parameterized importance functions (with linearly estimable parameters) and arbitrary sample sizes in each stage. When the initial inputs are quite uncertain and/or the extra cost of AIS is relatively small, one can make the sample size in each stage very small (even one) so that parameters of the importance function are updated often. It is even possible to have the choice of the sample size for the next stage depend on the current estimates; for instance, if the importance function has stabilized and there is still a long way to go to achieve the desired accuracy, one might choose a larger sample size for the next stage.

Combination of AIS with other adaptive (nonadaptive) schemes is possible. For example, it might be advantageous to run the first few stages with moderate sample sizes, and then

begin a new AIS with the updated importance function, ignoring the random drawings from the first few stages in actual calculation of the integral. This would prevent the possibly bad initial stages from contaminating the estimates. As another example, one could obtain information about f using AIS, and based upon this information choose a new form of the importance function (a new family of density functions) if the current form is not a good fit. AIS could then be utilized with this new family, and the process repeated as necessary.

APPENDIX

Proof of Theorem 3.1

Let $X_j = T_j^1(\underline{\theta}^{(j)})$. Then

i. $E^*(X_j|\mathcal{F}_{j-1}) = \Sigma_{i=1}^{n_j} E^*(t_j^1(\theta_i^{(j)})|\mathcal{F}_{j-1}) = 0$, for all j, by Lemma 3.1.

ii. Claim:

$$\Sigma_{j=1}^{\infty} \frac{1}{(n^{(j)})^2} E^*(X_j^2) < \infty$$

.

Proof:

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{1}{(n^{(j)})^2} E^*(X_j^2) &= \sum_{j=1}^{\infty} \frac{1}{(n^{(j)})^2} E^*(E^*(X_j^2|\mathcal{F}_{j-1})) && (65)\\
&= \sum_{j=1}^{\infty} \frac{n_j}{(n^{(j)})^2} E^*(\int w^{(j)}(\theta) f(\theta) d\theta - (\int f(\theta) d\theta)^2) \\
&\quad \text{by Lemma 3.2} && (66)\\
&\leq \sum_{j=1}^{\infty} \frac{n_j}{(n^{(j)})^2} \int f(\theta) d\theta (M - \int f(\theta) d\theta) && (67)\\
&\quad \text{by assumption III.} && (68)
\end{aligned}$$

But

$$\frac{n_j}{(n^{(j)})^2} = \frac{n^{(j)} - n^{(j-1)}}{(n^{(j)})^2} = \int_{n^{(j-1)}}^{n^{(j)}} \frac{1}{(n^{(j)})^2} dx \leq \int_{n^{(j-1}}^{n^{(j)}} \frac{1}{x^2} dx \qquad (69)$$

$$\sum_{j=2}^{m} \frac{n_j}{(n^{(j)})^2} \leq \int_{n^{(1)}}^{n^{(m)}} \frac{1}{x^2} dx = \frac{1}{n^{(1)}} - \frac{1}{n^{(m)}}, \tag{70}$$

$$\sum_{j=2}^{\infty} \frac{n_j}{(n^{(j)})^2} \leq \frac{1}{n^{(1)}} < \infty. \tag{71}$$

Q.E.D. □

iii. $n^{(1)} < n^{(2)} < \ldots \to \infty$.

¿From i, ii, and iii above and Feller (1971, Thm 3, p. 243), $\frac{\Sigma_{j=1}^{k} X_j}{n^{(k)}} \xrightarrow{a.s.} 0$, hence

$$\frac{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} w^{(j)}(\theta_i^{(j)})}{n^{(k)}} \xrightarrow{a.s.} \int f(\theta) d\theta.$$

Similarly,

$$\frac{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} \varphi(\theta_i^{(j)}) w^{(j)}(\theta_i^{(j)})}{n^{(k)}} \xrightarrow{a.s.} \int \varphi(\theta) f(\theta) d\theta. \tag{72}$$

The theorem follows from Chow and Teicher (1978, Cor. 2, p. 67). □

Proof of Theorem 3.2 a)

Let

$$r_{ij}^{(k)} = a_1 \frac{t_j^1(\theta_i^{(j)})}{\sqrt{n^{(k)}}} + a_2 \frac{t_j^{\varphi}(\theta_i^{(j)})}{\sqrt{n^{(k)}}} = \frac{1}{\sqrt{n^{(k)}}} t_j^{a_1 + a_2 \varphi}(\theta_i^{(j)}), \tag{73}$$

for arbitrary constants a_1, a_2 and let $S_{kl} = \Sigma_{j=1}^{l} \Sigma_{i=1}^{n_j} r_{ij}^{(k)}$. Then,

i). Claim: $\{(S_{k\ell}, \mathcal{F}_\ell),\ 1 \leq \ell \leq k,\ k \geq 1\}$ is a 0-mean square integrable martingale.

Proof: By Lemma 3.1, given $\mathcal{F}_{j-1}, r_{ij}^{(k)}, i = 1, \ldots, n_j$, are i.i.d. and

$$E^*(r_{ij}^{(k)} | \mathcal{F}_{j-1}) = 0. \tag{74}$$

Thus,

$$E^*(S_{kj} | \mathcal{F}_{j-1}) = S_{kj-1} + E^*(\sum_{i=1}^{n_j} r_{ij}^{(k)} | \mathcal{F}_{j-1}) = S_{k,j-1}, \tag{75}$$

from which it follows by induction on j that

$$E^*(S_{kl}) = E^*(S_{k1}) = E^*(\sum_{i=1}^{n_1} r_{i1}^{(k)}) = 0. \tag{76}$$

By Lemma 3.2 and assumption III,

$$\begin{aligned} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^2|\mathcal{F}_{j-1}) &= \frac{1}{n^{(k)}}E^*((\sum_{i=1}^{n_j} t_j^{a_1+a_2\varphi}(\theta_i^{(j)}))^2|\mathcal{F}_{j-1}) = \frac{1}{n^{(k)}}E^*((T_j^{a_1+a_2\varphi}(\underline{\theta}^{(j)}))^2|\mathcal{F}_{j-1}) \\ &= \frac{n_j}{n^{(k)}}\Big\{\int(a_1+a_2\varphi(\theta))^2 w^{(j)}(\theta)f(\theta)d\theta - (\int(a_1+a_2\varphi(\theta))f(\theta)d\theta)^2\Big\} \\ &\leq \frac{n_j}{n^{(k)}}\Big\{M\int(a_1+a_2\varphi(\theta))^2 f(\theta)d\theta - (\int(a_1+a_2\varphi(\theta))f(\theta)d\theta)^2\Big\} \\ &= \frac{n_j}{n^{(k)}}M^* < \infty. \end{aligned} \tag{77}$$

Now,

$$\begin{aligned} E^*(S_{kl}^2) &= E^*[E^*((S_{k,l-1}+\sum_{i=1}^{n_l} r_{il}^{(k)})^2|\mathcal{F}_{l-1})] \\ &= E^*[S_{k,l-1}^2 + 2S_{k,l-1}E^*(\sum_{i=1}^{n_l} r_{il}^{(k)}|\mathcal{F}_{l-1}) + E^*((\sum_{i=1}^{n_l} r_{il}^{(k)})^2|\mathcal{F}_{\ell-1})] \\ &= E^*(S_{k,l-1}^2) + E^*(E^*((\sum_{i=1}^{n_l} r_{il}^{(k)})^2|\mathcal{F}_{l-1})) \text{ (by (74))} \\ &\vdots \\ &= E^*(\sum_{j=1}^{l} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^2|\mathcal{F}_{j-1})) \\ &\leq E^*(\sum_{j=1}^{l}\frac{n_j}{n^{(k)}}M^*) \text{ (by (77))} \\ &\leq M^* < \infty. \quad \square \end{aligned} \tag{78}$$

ii). Claim: Let

$$\sigma_1^2 = \int\frac{f^2(\theta)}{g(\theta)}d\theta - (\int f(\theta)d\theta)^2 \tag{79}$$

$$\sigma_{12} = \int\frac{\varphi(\theta)f^2(\theta)}{g(\theta)}d\theta - \int\varphi(\theta)f(\theta)d\theta\int f(\theta)d\theta \tag{80}$$

$$\sigma_2^2 = \int\frac{\varphi^2(\theta)f^2(\theta)}{g(\theta)}d\theta - (\int\varphi(\theta)f(\theta)d\theta)^2. \tag{81}$$

Then, $\sum_{j=1}^{k} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^2|\mathcal{F}_{j-1}) \xrightarrow{a.s.} a_1^2\sigma_1^2 + 2a_1a_2\sigma_{12} + a_2^2\sigma_2^2$.

Proof: By Corollary 3.1, $(a_1+a_2\varphi(\theta))^2 w^{(j)}(\theta)f(\theta) \xrightarrow{a.s.} (a_1+a_2\varphi(\theta))^2 f^2(\theta)/g(\theta)$. And $(a_1+a_2\varphi(\theta))^2 w^{(j)}(\theta)f(\theta) \leq M(a_1+a_2\varphi(\theta))^2 f(\theta)$, by assumption III. By assumption, φ has finite second moment, so that the Lebesgue Dominated Convergence Theorem yields,

$$\int(a_1+a_2\varphi(\theta))^2 w^{(j)}(\theta)f(\theta)d\theta \xrightarrow{a.s.} \int(a_1+a_2\varphi(\theta))^2 f^2(\theta)/g(\theta)d\theta.$$

Hence,

$$\begin{aligned} E^*((t_j^{a_1+a_2\varphi}(\theta_i^{(j)}))^2|\mathcal{F}_{j-1}) &= \int(a_1+a_2\varphi(\theta))^2 w^{(j)}(\theta)f(\theta)d\theta - (\int(a_1+a_2\varphi(\theta))f(\theta)d\theta)^2 \\ &\xrightarrow{a.s.} \int(a_1+a_2\varphi(\theta))^2 f^2(\theta)/g(\theta)d\theta - (\int(a_1+a_2\varphi(\theta))f(\theta)d\theta)^2 \\ &= a_1^2\sigma_1^2 + 2a_1a_2\sigma_{12} + a_2^2\sigma_2^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{j=1}^{k} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^2|\mathcal{F}_{j-1}) &= \sum_{j=1}^{k}\sum_{i=1}^{n_j} E^*((r_{ij}^{(k)})^2|\mathcal{F}_{j-1}) \\ &\quad \text{by (74) and that } r_{ij}^{(k)} \text{ and i.i.d. given } \mathcal{F}_{j-1} \qquad (82) \\ &= \frac{1}{n^{(k)}}\sum_{j=1}^{k}\sum_{i=1}^{n_j} E^*((t_j^{a_1+a_2\varphi}(\theta_k^{(j)}))^2|\mathcal{F}_{j-1}) \qquad (83) \\ &\longrightarrow a_1^2\sigma_1^2 + 2a_1a_2\sigma_{12} + a_2^2\sigma_2^2 \text{ by Rudin (1964, p. 70).} \end{aligned}$$

iii). Claim:

$$\sum_{j=1}^{k} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^4|\mathcal{F}_{j-1}) \to 0, \text{ as } n^{(k)} \to \infty.$$

Proof:

$$\begin{aligned} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^4|\mathcal{F}_{j-1}) &= E^*(\sum_{i=1}^{n_j}(r_{ij}^{(ki)})^4 + \sum_{i,m=1,i\neq m}^{n_j} r_{ij}^{(k)}(r_{mj}^{(k)})^3 \\ &\quad + \sum_{i,m=1,i\neq m}^{n_j}(r_{ij}^{(k)})^2(r_{mj}^{(k)})^2|\mathcal{F}_{j-1}). \end{aligned}$$

By (74) and that $r_{ij}^{(k)}$ are i.i.d. given $\mathcal{F}_{j-1}, E^*(r_{ij}^{(k)}(r_{mj}^{(k)})^3|\mathcal{F}_{j-1}) = 0$. Now,

$$\begin{aligned} E^*((r_{ij}^{(k)})^4|\mathcal{F}_{j-1}) &= \frac{1}{n^{(k)^2}}E^*((t_j^{a_1+a_2\varphi}(\theta_i^{(j)}))^4|\mathcal{F}_{j-1}) \qquad (84) \\ &= \frac{1}{n^{(k)^2}}\Big\{\int(a_1+a_2\varphi)^4(w^{(j)})^3 f - 4\int(a_1+a_2\varphi)^3(w^{(j)})^2 f\cdot\int(a_1+a_2\varphi)f \\ &\quad +6\int(a_1+a_2\varphi)^2 w^{(j)} f\cdot(\int(a_1+a_2\varphi)f)^2 \qquad (85) \\ &\quad -4(\int(a_1+a_2\varphi)f)(\int(a_1+a_2\varphi)f)^3 + (\int(a_1+a_2\varphi)f)^4\Big\} \qquad (86) \\ &\leq \frac{1}{n^{(k)^2}}\Big\{M^3\int(a_1+a_2\varphi)^4 f + 4M^2\int|a_1+a_2\varphi|^3 f\cdot\int|a_1+a_2\varphi|f \qquad (87) \\ &\quad +6M\int(a_1+a_2\varphi)^2 f\cdot(\int(a_1+a_2\varphi)f)^2 - 3(\int(a_1+a_2\varphi)f)^4\Big\} \qquad (88) \\ &\quad \text{by assumption III} \qquad (89) \\ &\leq \frac{1}{n^{(k)^2}}M^{**} < \infty. \qquad (90) \end{aligned}$$

Similarly, $E^*((r_{ij}^{(k)})^2(r_{mj}^{(k)})^2|\mathcal{F}_{j-1}) \leq M^{***}/(n^{(k)})^2 < \infty$. Thus,

$$\begin{aligned} \sum_{j=1}^{k} E^*((\sum_{i=1}^{n_j} r_{ij}^{(k)})^4|\mathcal{F}_{j-1}) &\leq \frac{1}{n^{(k)^2}} \sum_{j=1}^{k} \{n_j M^{**} + \frac{1}{2} n_j(n_j-1)M^{***}\} \quad (91) \\ &= \frac{1}{2n^{(k)^2}} \sum_{j=1}^{k} \{n_j(2M^{**} - M^{***}) + (n_j)^2 M^{***}\} \\ &\longrightarrow 0, \quad (92) \end{aligned}$$

by assumption.

¿From i, ii, iii, and Hall and Heyde (1980, Cor. 3.1),

$$S_{kk} \xrightarrow{d} N(0, a_1^2\sigma_1^2 + 2a_1a_2\sigma_{12} + a_2^2\sigma_2^2).$$

Because a_1, a_2 are arbitrary constants,

$$\frac{1}{\sqrt{n^{(k)}}} \begin{pmatrix} \Sigma_{j=1}^{k} \sum_{i=1}^{n_j} t_j^1(\theta_i^{(j)}) \\ \Sigma_{j=1}^{k} \sum_{i=1}^{n_j} t_j^{\varphi}(\theta_i^{(j)}) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right). \quad (93)$$

The theorem follows from Bickel and Doksum (1977, p. 461, A.14.18). □

Proof of Theorem 3.2 b)

Let

$$X_{ij} = (\varphi(\theta_i^{(j)}))^2(w^{(j)}(\theta_i^{(j)}))^2 - \int \varphi^2(\theta) w^{(j)}(\theta) f(\theta) d\theta. \quad (94)$$

Then, by analogous arguments to Lemma 3.1 and 3.2, it can be shown that given $\mathcal{F}_{j-1}$, X_{ij} are i.i.d. with $E^*(X_{ij}|\mathcal{F}_{j-1}) = 0$, and

$$\begin{aligned} E^*((\sum_{i=1}^{n_j} X_{ij})^2|\mathcal{F}_{j-1}) &= E^*(\sum_{i=1}^{n_j} X_{ij}^2 \mathcal{F}_{j-1}) \\ &= n_j \left\{ \int \varphi^4(\theta)(w^{(j)}(\theta))^3 f(\theta) d\theta - \left(\int \varphi^2(\theta) w^{(j)}(\theta) f(\theta) d\theta \right)^2 \right\} \\ &\leq n_j \left\{ M^3 \int \varphi^4(\theta) f(\theta) d\theta - M^2 \left(\int \varphi^2(\theta) f(\theta) d\theta \right)^2 \right\}. \end{aligned}$$

Thus,

$$\Sigma_{j=1}^{\infty} \frac{1}{(n^{(j)})^2} E^*((\Sigma_{i=1}^{n_j} X_{ij})^2) = \Sigma_{j=1}^{\infty} \frac{1}{(n^{(j)})^2} E^*[E^*((\Sigma_{i=1}^{n_j} X_{ij})^2|\mathcal{F}_{j-1})] < \infty.$$

By Feller (1971, Thm 3, p. 243),

$$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}X_{ij}}{n^{(k)}} \xrightarrow{a.s.} 0, \text{ as } n^{(k)} \to \infty. \tag{95}$$

But $\varphi^2(\theta)w^{(j)}(\theta)f(\theta) \leq M\varphi^2(\theta)f(\theta)$ and $E\varphi^2$ exists. By Corollary 3.1 and the Lebesgue Dominated Convergence Theorem,

$$\int \varphi^2(\theta)w^{(j)}(\theta)f(\theta)d\theta \xrightarrow{a.s.} \int \varphi^2(\theta)f^2(\theta)/g(\theta)d\theta. \tag{96}$$

Hence, by Rudin (1964, p. 70),

$$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}\int \varphi^2(\theta)w^{(j)}(\theta)f(\theta)d\theta}{n^{(k)}} \xrightarrow{a.s.} \int \varphi^2(\theta)f^2(\theta)/g(\theta)d\theta. \tag{97}$$

¿From (94), (95), and (97),

$$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}\varphi^2(\theta_i^{(j)})(w^{(j)}(\theta_i^{(j)}))^2}{n^{(k)}} \xrightarrow{a.s.} \int \frac{\varphi^2(\theta)f^2(\theta)}{g(\theta)}d\theta,$$

as $n^{(k)} \to \infty$. Similarly,

$$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}\varphi(\theta_i^{(j)})(w^{(j)}(\theta_i^{(j)}))^2}{n^{(k)}} \xrightarrow{a.s.} \int \frac{\varphi(\theta)f^2(\theta)}{g(\theta)}d\theta \tag{98}$$

$$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(w^{(j)}(\theta_i^{(j)}))^2}{n^{(k)}} \xrightarrow{a.s.} \int \frac{f^2(\theta)}{g(\theta)}d\theta. \tag{99}$$

By Theorem 3.1, $\overline{w}^{(k)} \to \int f(\theta)d\theta$, and $\hat{E}^{(k)}\varphi \to E\varphi$. Thus,

$$\hat{var}^{(k)}(w) \xrightarrow{a.s.} \sigma_1^2, \quad \hat{cov}^{(k)}(\varphi w, w) \xrightarrow{a.s.} \sigma_{12}, \quad \hat{var}^{(k)}(\varphi w) \xrightarrow{a.s.} \sigma_2^2,$$

where $\hat{var}^{(k)}(w)$, $\hat{cov}^{(k)}(\varphi w, w)$, $\hat{var}^{(k)}(\varphi w)$ are defined in (16), (17), (18), respectively. The theorem follows. □

References

[1] Ash, B. (1972), Real Analysis and Probability, *Academic Press, New York.*

[2] Bauwens, W. and Richards, J.F. (1985), A 1-1 Poly-t Random Variable Generator with Application to Monte Carlo Integration, *J. of Econometrics*, **29**, 19–46.

[3] Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis, 2nd ed., *Springer-Verlag, New York.*

[4] Bickel, P.J. and Doksum, K.A. (1977), Mathematical Statistics, *Holden-Day Inc., San Francisco.*

[5] Billingsley, P. (1979), Probability and Measure, *Wiley, New York.*

[6] Chow, Y.S. and Teicher, H. (1978), Probability Theorem, *Springer-Verlag, New York.*

[7] Cramer, H. (1946), Mathematical methods of statistics, *Princeton University, N.J.*

[8] Davis, P.J. and Rabinowitz, P. (1975), Methods of Numerical Integration, *Academic Press, New York.*

[9] Feller, W. (1971), An Introduction to Probability Theory and Its Applications, 3rd ed., *John Wiley and Sons, New York.*

[10] Fong, D.K.H. (1987), Ranking and Estimation of Exchangeable Means in Balanced and Unbalanced Models: A Bayesian Approach, *Ph. D. Thesis, Dep. of Stat., Purdue University, W. Lafayette, IN 47907.*

[11] Geweke, J. (1986), Bayesian Inference In Econometric Models Using Monte Carlo Integration, *Dep. of Econometrics, Duke Univ., Durham, N.C.*

[12] Geweke, J. (1988), Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference, *J. of Econometrics*, **38**, 73–90.

[13] Hall, P. and Heyde, C.C. (1980), Martingale Limit Theory and Its Application, *Academic Press, New York.*

[14] Hammersley, J.M. and Handscomb, D.C. (1964), Monte Carlo Methods, *Methuen, London.*

[15] Kloek, K. and van Dijk, H.K. (1978), Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo, *Econometrica*, **46**, 1–20.

[16] Naylor, J.C. and Smith, A.F.M. (1982), Application of a Method for the Efficient Computation of Posterior Distributions, *Appl. Stat.*, **31**, 214–225.

[17] Naylor, J.C. and Smith, A.F.M. (1983), A Contamination Model in Clinical Chemistry: an Illustration of a Method for the Efficient Computation of Posterior Distributions, *The Statistician*, **32**, 82–87.

[18] Naylor, J.C. and Smith, A.F.M. (1988), Econometric Illustrations of Novel Numerical Integration Strategies for Bayesian Inference, *J. of Econometrics*, **38**, 103–125.

[19] Oh, M.S. (1989), Statistical Multiple Integration by Monte Carlo Importance Sampling, Ph.D. Thesis, *Purdue University.*

[20] Rubinstein, R.Y. (1981), Simulation and the Monte Carlo Method, *Wiley, New York.*

[21] Rudin, W. (1964), Principles of Mathematical Analysis, *McGraw-Hill Book Co., New York.*

[22] SAS Institute Inc. (1985), SAS User's Guide: Statistics, *SAS Institute Inc., Cary, NC 5th Ed.*

[23] Smith, A.F.M. et al (1985), The Implementation of the Bayesian Paradigm, *Commun. Stat. – Theor. Meth.*, **14**, 1079–1102.

[24] Stewart, L. (1979), Multiparameter Univariate Bayesian Analysis, *Journal of the American Statistical Association*, **76**, 684–693.

[25] Stewart, L. (1983), Bayesian Analysis Using Monte Carlo Integration – A Powerful Methodology for Handling Some Difficult Problems, *The Statistician*, **32**,. 195–200.

[26] Stewart, L. (1984), Bayesian Analysis Using Monte Carlo Integration With an Example of the Analysis of Survival Data, *Lockheed Palo Alto Research Lab., Palo Alto, CA.*

[27] Tierney, L. and Kadane, J.B. (1986), Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, **81**, 82–86.

[28] van Dijk, H.J. and Kloek, T. (1980), Further Experience in Bayesian Analysis Using Monte Carlo Integration, *J. of Econometrics*, **14**, 307–328.

[29] van Dijk, H.K. and Kloek, T. (1983a), Monte Carlo Analysis of Skew Posterior Distributions: an Illustrative Econometric Example, *The Statistician*, **32**, 216–223.

[30] van Dijk, H.K. and Kloek, T. (1983b), Experiments with Some Alternatives for Simple Importance Sampling in Monte Carlo Integration, *Bayesian Statistics 2, North-Holland, Amsterdam*, 511–530.

[31] van Dijk, H.K. (1984), Posterior Analysis of Econometric Models Using Monte Carlo Integration, *Reproduktie Woudestein, Erasmus Universiteit Rotterdam.*

[32] van Dijk, H.K. and Kloek, T. and Boender, C.G.E. (1985), Posterior Moments Computed by Mixed Integration, *Journal of Econometrics*, **29**, 3–18.

[33] van Dijk, H.K. (1987), Some Advances in Bayesian Estimation Methods Using Monte Carlo Integration, *Econometric Institute, Erasmus Univ., Rotterdam, and Center for Operations Research and Econometrics (CORE), Universite Catholique de Louvain.*

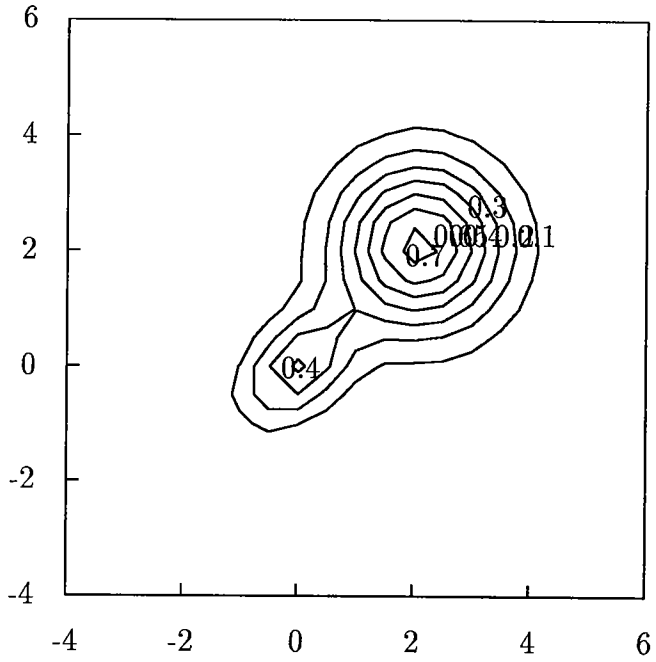

Figure 1: Contour map of f