54

Computer Sciences Department
1210 West Dayton Street
The University of Wisconsin
Madison, Wisconsin 53706

MONTE CARLO INTEGRATION WITH SEQUENTIAL STRATIFICATION

by

John H. Halton
and
Edward A. Zeidman

Technical Report #61

April 1969

MONTE CARLO INTEGRATION WITH SEQUENTIAL STRATIFICATION

INTRODUCTION

Monte Carlo integration estimates, by statistical sampling techniques, the parameter

$$\theta = \int_R f(\underline{x})\, d\underline{x}, \qquad f \in L^2, \tag{1}$$

where R is the region of integration and $\underline{x}$ is a k-dimensional vector. The estimator $\hat{\theta}$ of θ and the variance of $\hat{\theta}$, $\mathrm{var}(\hat{\theta})$, depend on the sampling technique, the nature of the integrand, and the region of integration. The stratified sampling method consists of partitioning R into disjoint subregions R_i, $i = 1, \ldots, h$, so that

$$\theta = \Sigma_{i=1}^{h} \theta_i, \tag{2}$$

where

$$\theta_i = \int_{R_i} f(\underline{x})\, d\underline{x}.$$

Now each θ_i is estimated independently by $\hat{\theta}_i$ using some Monte Carlo technique, usually by crude Monte Carlo. (Note, it may be possible to evaluate some of the θ_i by direct evaluation.) If the strata, R_i, $i = 1, \ldots, h$, have already been assigned, then the variance of $\hat{\theta} = \Sigma_{i=1}^{h} \hat{\theta}_i$ will be minimized when the number of sample points for each stratum is directly proportional to the standard deviation of the

estimator for the stratum. This is the Tschuprow-Neyman theorem (see [5] equation 8.1.) If the partition of R and the number of sub-regions, h, were not choosen in advance, but instead allowed to vary, then further reduction of $\mathrm{Var}(\hat{\theta})$ is possible. Dalenius and Hodges [2] have given strong evidence that the minimization of $\mathrm{Var}(\hat{\theta})$ will be nearly achieved if

$$\mathrm{Var}(\hat{\theta}_i) = \text{constant}, \quad i = 1, \ldots, h . \tag{3}$$

Hammersley and Handscomb [4] have a detailed discussion of stratified sampling.

The most recent numerical quadrature programs use adaptive integration methods; that is, the points at which the integrand will be evaluated are chosen according to the behavior of the integrand. And in many cases, these programs use an iterative scheme which successively approximates the integral until the desired accuracy is achieved. McKeeman and Tesler [7] and Tavernini [8] have good examples of programs of this type. The difficulty is that for multiple integrals the number of function evaluations needed is N^k, where N is the number of evaluations in each direction and k is the dimensionality of the integral.

Monte Carlo integration with sequential stratified sampling is an adaptive iterative scheme which attempts to minimize the variance of the stratified Monte Carlo estimator. The scheme produces an

approximate optimal choice of strata by a recursive binary search procedure. The algorithm also yields a confidence interval for the estimate of θ less than or equal to the one desired. Halton [3] has a discussion of other sequential Monte Carlo schemes.

Monte Carlo Integration by Sequential Stratified Sampling

Description of the scheme

This paper will deal with the single variable integration over a finite domain, with the exception of a few remarks given to the multiple integral at the conclusion. A subsequent paper will deal more fully with the multiple integral [9].

The sequential stratified sampling scheme, for the estimation of $\theta = \int_A^B f(x)\,dx$, consists of determining the location of stratification points, or equivalently finding the size of each and every stratum, and the number of points to be sampled in each stratum. The number of points to be sampled per stratum for the estimate of the integral is dependent upon the integrand and the desired size and significance level of the probablistic confidence interval. A few basic definitions are in order before the exposition of the procedure.

The decision rule is used to determine whether a given stratum should be stratified. The stopping rule is used to test if more sampling in a stratum is necessary to reach the desired accuracy.

The procedure begins with drawing a sample from the entire interval of integration. The stopping rule is applied. If the condition is satisfied, the search procedure stops and the estimate of the area is calculated. (For simplicity, the authors use crude Monte Carlo). If not, the decision rule is applied. If the decision rule signifies that stratification is not advantageous, then sampling over the entire interval is continued until the stopping rule is satisfied. If the decision rule recommends stratification, the entire interval is bisected (i.e., a stratification point is chosen at the mid-point of the interval). Next the algorithm in a recursive fashion examines the left half interval (or stratum). The location of the right half interval is saved and stored in a last in-first out (LIFO) list. (In trial calculations, the length of this list has remained small.)

The procedure for the left half interval is the same as for the entire interval. The stopping rule is applied first and if necessary, the decision rule is applied. As before, if the decision is to stratify, then this stratum is bisected, storing the right half's location in the LIFO list and examining the left half. If stratification is <u>not</u> recommended the sampling continues in the entire region being examined until the stopping rule is satisfied. Afterwards, the unexamined stratum, whose location has previously been stored in the LIFO list, is explored. This process continues recursively until the entire region of integration is

observed and the estimate of the integral for each stratum is calculated. The sum of all these estimates is the estimate of the integral. (See figure 1.) This process is no more than a recursive binary search procedure. The decision rule indicates whether or not to branch, the stopping rule determines the amount of sampling at each node of the tree, and the LIFO list shows where the process should return to after the end of a branch has been reached. (See figure 2).

The Decision Rule

To facilitate the explanation of the decision rule some notation is introduced. Let

$$\theta = \int_A^B f(x)\, dx$$

and

$\hat{\theta}$ = the sequential stratified sampling estimator of θ .

Denote a stratum in $[A, B)$ by the interval $[a, b)$, where $[a, b) \subset [A, B)$. For each stratum $[a, b)$, let $c = \frac{1}{2}(a + b)$ and let:

$$(4) \qquad \begin{cases} \tau_0 = (b-a)\, f(\xi_0) \quad , & \text{where } \xi_0 \sim U(a, b) \, , \\ \tau_1 = (c-a)\, f(\xi_1) \quad , & \text{where } \xi_1 \sim U(a, c) \, , \\ \tau_2 = (b-c)\, f(\xi_2) \quad , & \text{where } \xi_2 \sim U(c, b) \, , \end{cases}$$

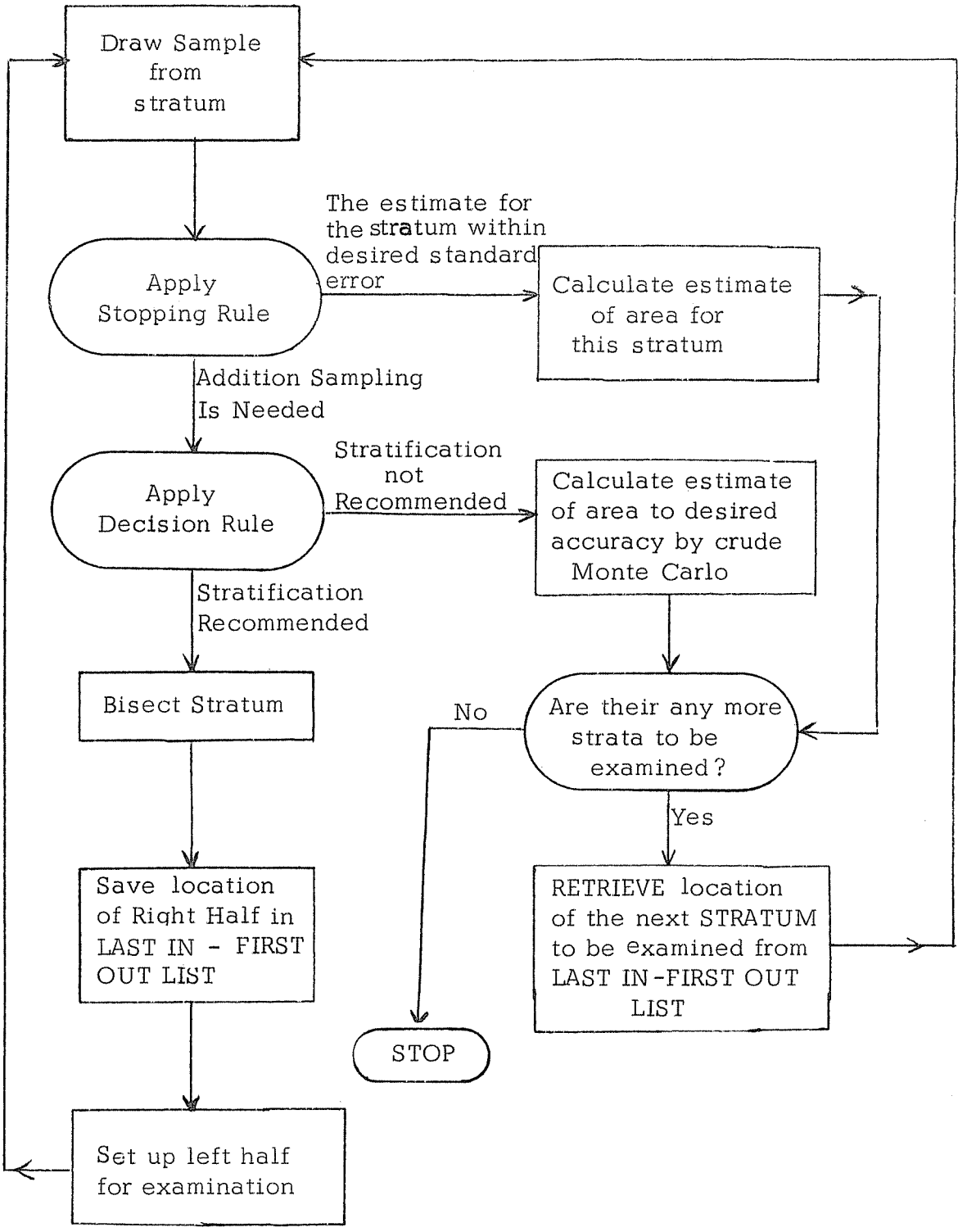


Figure 1 : Flowchart of the algorithm for a single stratum.

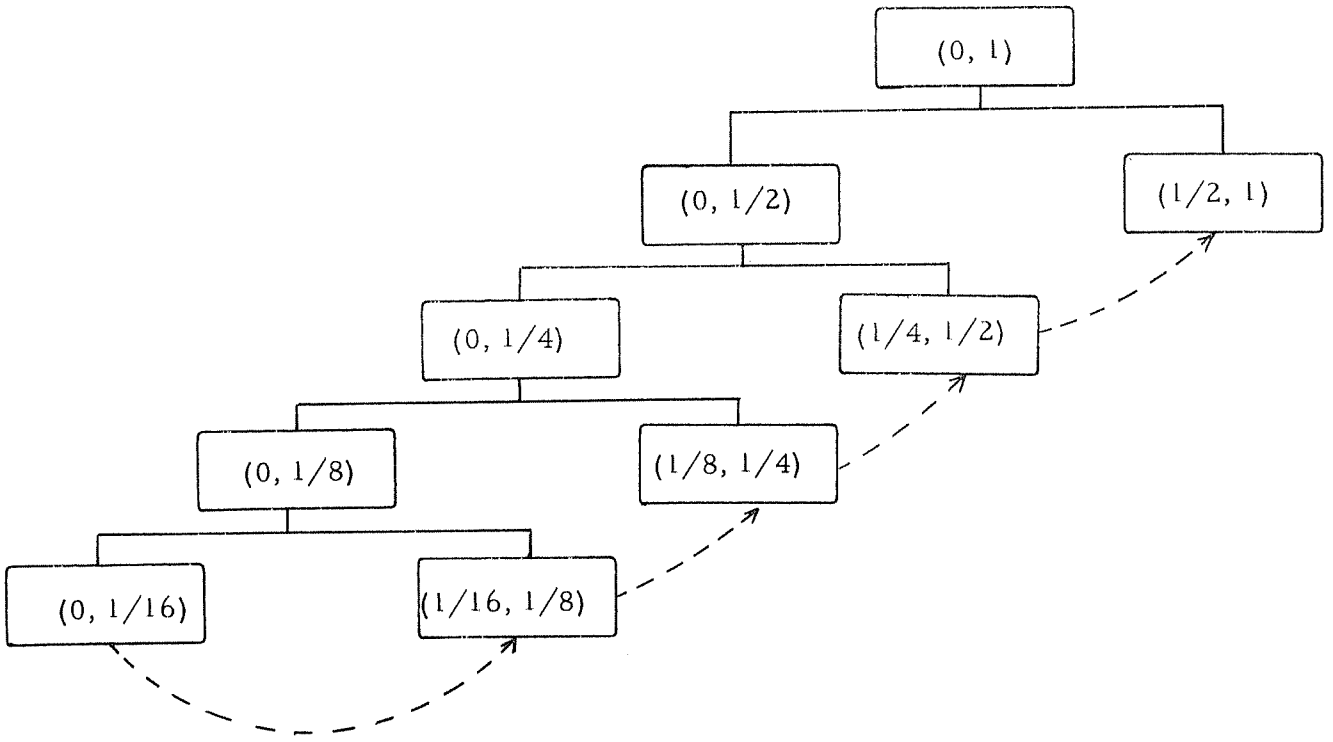


Figure 2a.

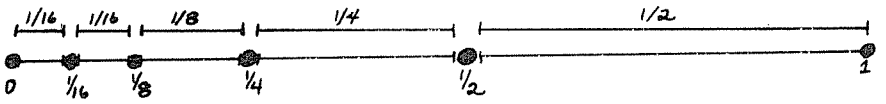


Figure 2b.

The sequential stratification scheme determined the stratification points to be 0, 1/16, 1/8, 1/4, 1/2, 1 in calculating $\int_0^1 \log x \, dx$ with error bound of 0.1 and 1% significance level. Figure 2a indicates the tree structure. The dotted line shows where to proceed after reaching the end of a branch. Figure 2b shows the corresponding strata and stratification points.

$$(5)\qquad \begin{cases} \mu_0 = E[\tau_0] = \int_a^b (b-a)\, f(x)\, \dfrac{dx}{b-a} = \int_a^b f(x)\, dx\,, \\ \mu_1 = E[\tau_1] = \int_a^c f(x)\, dx\,, \\ \mu_2 = E[\tau_2] = \int_c^b f(x)\, dx\,, \end{cases}$$

$$(6)\qquad \begin{cases} \sigma_0^2 = Var[\tau_0] = \int_a^b [(b-a)f(x)]^2 - \mu_0^2 = (b-a)\int_a^b f^2(x)dx - \mu_0^2\,, \\ \sigma_1^2 = Var[\tau_1] = (c-a)\int_a^c f^2(x)\, dx - \mu_1^2\,, \\ \sigma_2^2 = Var[\tau_2] = (b-c)\int_c^b f^2(x)\, dx - \mu_2^2\,. \end{cases}$$

To establish the final version of the decision rule the following two lemmas are needed.

Lemma 1. $\sigma_0^2 = 2(\sigma_1^2 + \sigma_1^2) + (\mu_1 - \mu_2)^2$,

where σ_0^2, σ_1^2, σ_2^2, μ_1, μ_2 are defined above.

Remark. In analysis of variance terminology, the quantity $2(\sigma_1^2 + \sigma_2^2)$ is referred to as the within stratum contribution to σ_0^2 and $(\mu_1 - \mu_2)^2$ is between strata contribution.

Proof.

$(b-a)\int_a^b f^2(x)\, dx = 2(\sigma_1^2 + \sigma_2^2) + 2(\mu_1^2 + \mu_2^2)$, by (5) and (6).

$\mu_0^2 = (\mu_1 + \mu_2)^2 = \mu_1^2 + 2\mu_1\mu_2 + \mu_2^2$, by (5).

$\sigma_0^2 = (b-a)\int_a^b f^2(x)\, dx - \mu_2^2$, by (6),

$= 2(\sigma_1^2 + \sigma_2^2) + 2(\mu_1^2 + \mu_2^2) - \mu_1^2 + 2\mu_1\mu_2 + \mu_2^2$

$= 2(\sigma_1^2 + \sigma_1^2) + (\mu_1 - \mu_2)^2$. ⊩

Lemma 2. If n_1 and n_2 are real numbers whose sum n_0 is fixed, then for $n_1 \in (0, n_0)$

$$\min_{n_1, n_2} \left\{ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right\} = \frac{1}{\hat{n}_0} (\sigma_1 + \sigma_2)^2$$

is attained when

$$\hat{n}_i = \sigma_i \frac{n_0}{\sigma_1 + \sigma_2} , \quad i = 1, 2 .$$

Proof.

The optimum values, $\hat{n}_i$, can be determined by Lagrangian multipliers, or by elementary calculus in the following manner:

$$\frac{\partial}{\partial n_1} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_0 - n_1} \right) = - \frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{(n_0 - n_1)^2} ,$$

since $n_0 = n_1 + n_2$; and so $- \frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{(n_0 - n_1)^2} = 0$ at the critical point.

Hence

$$\hat{n}_1 = \sigma_1 \frac{n_0}{\sigma_1 + \sigma_2} .$$

Since

$$\frac{\partial^2}{\partial n_1^2} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_0 - n_1} \right) = 2 \frac{\sigma_1^2}{n_1^3} + 2 \frac{\sigma_2^2}{(n_0 - n_1)^3} \geq 0 ,$$

$\hat{n}_1$ is the minimum value. We can easily find $\hat{n}_2$ by using $n_0 = n_1 + n_2$. Therefore we have

$$\frac{\sigma_1^2}{\hat{n}_1} + \frac{\sigma_2^2}{\hat{n}_2} = \frac{1}{\hat{n}_0} (\sigma_1 + \sigma_2)^2 . \qquad \#$$

The quantity $\sigma_1^2/n_1 + \sigma_2^2/n_2$ is the variance of the stratified Monte Carlo estimator of $\int_a^b f(x)\,dx$, where the interval $[a, b)$ is stratified into two strata of equal length, $[a, c)$ and $[c, b)$, and n_1, n_2 are the number of points sampled in the respective strata ([4], p. 55). For a fixed $\hat{n}_0 = n_1 + n_2$, we can minimize this variance, in practice, by choosing integers close to the theoretical values in lemma 2.

The crude Monte Carlo estimate of $\theta_i = \int_a^b f(x)\,dx$, using n_0 points, has variance σ_0^2/n_0 ([4] p. 51), where σ_0^2 is defined in (6). If a variance of V is desired, then we want $\sigma_0^2/n_0 = V$, therefore, $n_0 = \sigma_0^2/V$. If k_c is the amount of computing time for the calculation of a single crude Monte Carlo estimate, the amount of labor required for n_0 points is

$$n_0 k_c = \frac{\sigma_0^2}{V} k_c \quad . \tag{7}$$

For stratified Monte Carlo, where the stratum $[a, b)$ is divided into two equal strata, using $\hat{n}_0 = \hat{n}_1 + \hat{n}_2$ points, $\hat{n}_1$ and $\hat{n}_2$ as in lemma 2, the variance is seen to be

$$\frac{\sigma_1^2}{\hat{n}_1} + \frac{\sigma_1^2}{\hat{n}_2} = \frac{1}{\hat{n}_0}\,[(\sigma_1 + \sigma_2)^2], \tag{8}$$

upon applying lemma 2. Again if a variance V is desired, we need $\hat{n}_0 = (\sigma_1 + \sigma_2)^2/V$ points. Letting k_s be the amount of time required to evaluate each term of the stratified estimate, we find that the amount of labor required for $\hat{n}_0$ points is

$$(9) \qquad \hat{n}_0 k_s = \frac{(\sigma_1 + \sigma_2)^2}{V} k_s .$$

Therefore, the labor required in stratification is less than crude Monte Carlo when $\hat{n}_o k_s < n_0 k_c$. This is the basis for the

Decision Rule. Stratification is recommended if

$$(10) \qquad \sigma_0^2 > K(\sigma_1 + \sigma_2)^2 , \text{ where } K = \frac{k_s}{k_c} .$$

Indeed, since $\hat{n}_o k_s < n_o k_c$, we should stratify if and only if, by (7) and (9),

$$\frac{(\sigma_1 + \sigma_2)^2}{V} k_s < \frac{\sigma_0^2}{V} k_c ,$$

that is,

$$\sigma_0^2 > \left(\frac{k_s}{k_c}\right) (\sigma_1 + \sigma_2)^2 .$$

Equivalently, applying lemma 1 to (10), we get

$$(11) \qquad (\mu_1 - \mu_2)^2 > (K - 2)(\sigma_1^2 + \sigma_2^2) + 2K\sigma_1\sigma_2 .$$

Turning, now, to a graphical interpretation of the decision rule, we let n_o remain fixed and let $x = n_1/n_0 = v_1$ vary continuously. Define the function d on $[0, 1]$ by

$$(12) \qquad d(x) = \sigma_0^2 - K\left(\frac{\sigma_1^2}{x} + \frac{\sigma_2^2}{1-x}\right) .$$

Clearly, $\lim_{x \to 0} d(x) = \lim_{x \to 1} d(x) = -\infty$, when $\sigma_1^2 \neq 0$ and $\sigma_2^2 \neq 0$

Since $K > 1$, $d(x) > 0$ for some x in $[0, 1]$ if and only if the inequality (10) holds, where $v_1 = x$ and $n_0 = n_1 + n_2$. Hence, if (10) holds, then by the continuity of d, there will be an interval of values, (x_1, x_2), for v_1, in which stratification will be advantageous. In either case, the function takes on its maximum value on $[0, 1]$ when

$$(13) \qquad x = \hat{v}_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2}, \text{ where } v_1 = \frac{\hat{n}_1}{n_0}, \ \hat{n}_1 \text{ as in lemma 3 .}$$

Graphically, this is described in figure 3 .

In order to decide if the bisection of the stratum is advantageous, it is necessary to determine whether there exists any $x = v_1$ for which $d(x) > 0$. However, this search can be eliminated by determining if $d(x) > 0$ for the point $x = v_1$, where d takes its maximum value, i.e. $d(\hat{x}) = \max \{d(x)\}$, $\hat{x} = \hat{v}_1$, for all x on $[0, 1]$. Substituting (13) into (12), we get the equation

$$(14) \qquad d(\hat{x}) = \sigma_0^2 - K(\sigma_1 + \sigma_2)^2, \text{ where } \hat{x} = \frac{\sigma_1}{\sigma_1 + \sigma_2} .$$

If we apply lemma 1 to (14), then we get

$$(15) \qquad d(\hat{x}) = (\mu_1 - \mu_2)^2 + (2 - K)(\sigma_1^2 + \sigma_2^2) - 2K\sigma_1\sigma_2 .$$

Concerning the size of the labor ratio, $K = k_s/k_c$, a little calculation shows that K is such that $1 < K < 2$. The exact value

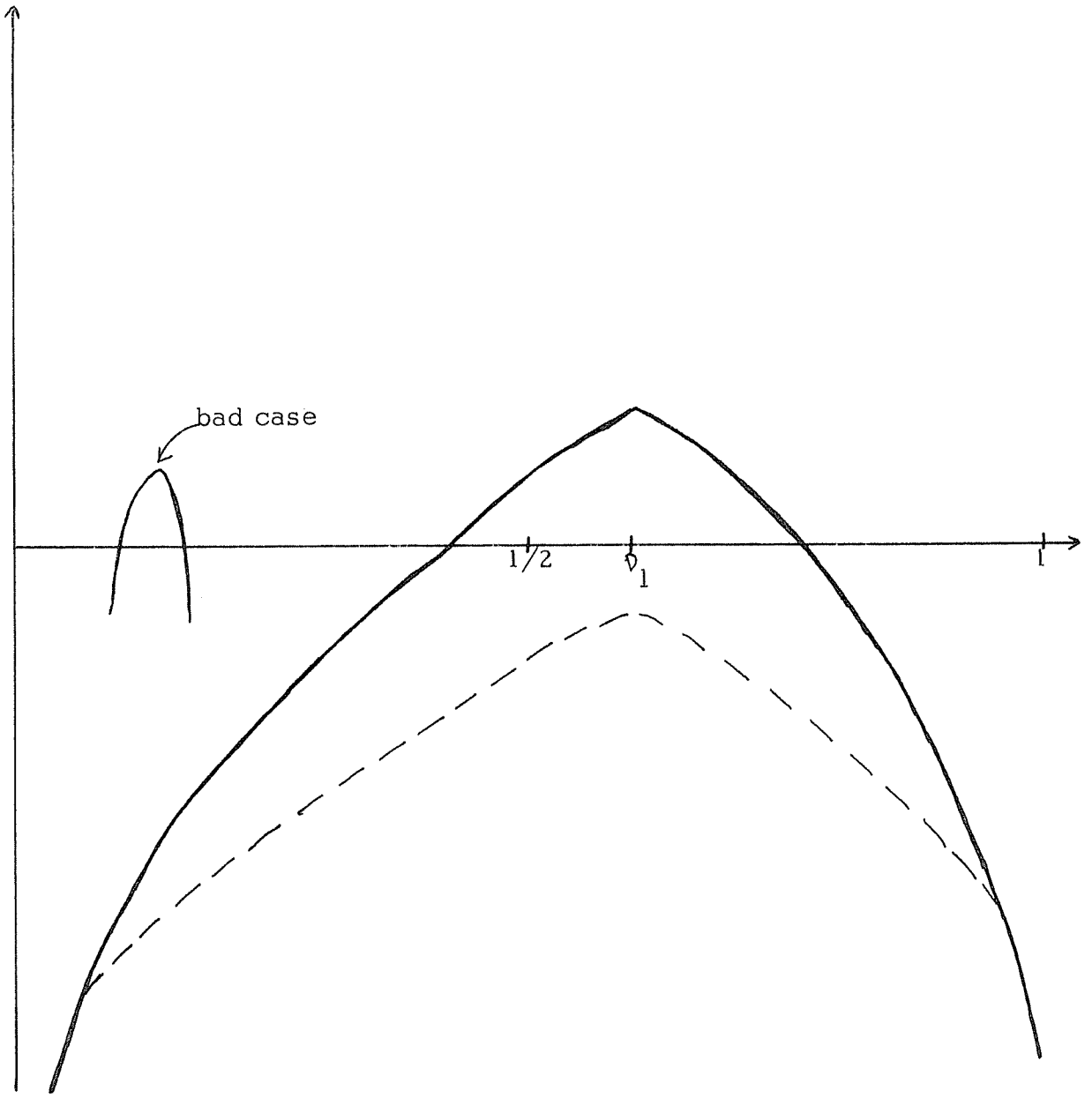


Figure 3: Graphs of $d(x) = \sigma_0^2 - K\left(\frac{\sigma_1^2}{x} + \frac{\sigma_2^2}{1-x}\right)$, $x \in [0, 1]$.

of K depends on the number of multiplications required to evaluate the integrand and the square root of σ_1^2 and σ_1^2. Since the number of evaluations of the integrand is the same for each integrand in the test, $n_0 = n_1 + n_2$, we have that as n_0 increases or as the complexity of the integrand increases, K tends toward 1. (See the appendix.)

It should be noted that since, in practice, n_1 is an integer, it is possible that $d(\hat{x}) > 0$, but that $d(\frac{n}{n_0}) < 0$ for all integers, n, in $[0, n_0]$. (This is the "bad" case referred to in figure 3.) However, the chances of such an event occurring are very small, and the probability of any resulting inefficiency being significant is negligible.

It is clear from (15), assuming n_0 and K fixed, that the stratification criterion depends upon both the difference in the means between the strata, $(\mu_1 - \mu_2)^2$ and the size of the variances, σ_1^2 and σ_2^2. Let us examine some of the possibilities.

If $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, then we have, upon substituting into (15) and using the fact that $K > 1$,

$$f(\hat{x}) = 4(1 - K)\sigma_1^2 < 0 ,$$

which implies stratification is not recommended, as expected. If $\mu_1 \neq \mu_2$ and $\sigma_1^2 = \sigma_2^2$, then stratification is advantageous if

$$(\mu_1 - \mu_2)^2 > 4(k - 1)\sigma_1^2 .$$

On the other hand, if $\mu_1 = \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$, then we <u>don't</u> stratify if

$$\sigma_2 \neq 0 \quad \text{and} \quad \left| \frac{\sigma_1}{\sigma_2} - \left(1 + \frac{2(K-1)}{2-K}\right) \right| \leq \frac{2\sqrt{K-1}}{2-K} \quad .$$

<u>The Stopping Rule</u>

The stopping rule tests to determine if additional sampling in a particular stratum is necessary to reduce the variance of the estimated to the desired accuracy. The user of the algorithm determines the desired accuracy by specifying the length of the confidence interval and the significance level he wants the estimate of the integral to have.

Let the length of the confidence interval wanted be $2e$, and the desired significance level be α. What is required, then, is that,

$$\Pr\{|\theta - \hat{\theta}| \leq e\} = 1 - \alpha \,, \tag{16}$$

where $\hat{\theta}$ is the sequential stratified estimate of

$$\theta = \int_A^B f(x)\, dx \quad .$$

For a sufficiently large number of samples, N, we can apply the central limit theorem ([6] p. 316) to (15), yielding that $\hat{\theta}$ is distributed normally with mean θ and variance σ^2/N, where $\sigma^2 = (B-A) \int_A^B f^2(x)\, dx - \theta^2$ and $N = \Sigma_{i=0}^{k} n_0^{(i)}$. Lemma 3 yields that the mean and variance of $\hat{\theta}$ are, indeed, θ and σ^2/N, respectively. Now, for all strata, $[a_i, b_i)$, in $[A, B)$, the estimators

$\hat{\theta}_i$ of $\theta_i = \int_{a_i}^{b_i} f(x)\,dx$ are independent random variables, where $[a_i, b_i)$, $i = 0, 1, \ldots, k$, for some k, the number of strata, such that

(17) $$A = a_0 < b_0 = a_1 < b_1 = a_2 < \ldots < b_k = B .$$

and

(18) $$\mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{N} = \Sigma_{i=0}^{k} \mathrm{Var}(\hat{\theta}_i) = \Sigma_{i=0}^{k} \frac{\sigma_0^{(i)2}}{n_0^{(i)}}$$

by the Bienaymé equality ([6] p. 234). Let t_α be the normal critical level for the confidence interval symmetric about the mean with a probability $1 - \alpha$, then we want

(19) $$t_\alpha \frac{\sigma}{\sqrt{N}} = e$$

The relation (18) will hold if and only if

(20) $$\frac{\sigma^2}{N} = \left(\frac{e}{t_\alpha}\right)^2 = T$$

If we assume a uniform distribution of the desired accuracy, that is,

(21) $$T_i = \frac{b_i - a_i}{B-A} T ;$$

then we get

(22) $$T = \Sigma_{i=0}^{k} T_i ,$$

since $\Sigma_{i=0}^{k} T_i = T/(B-A) \Sigma_{i=0}^{k} (b_i - a_i) = T$, by (17) .

We can now state the

Stopping Rule. Stop sampling in the i-th stratum, $[a_i, b_i)$, when,

$$\text{(23)} \qquad \mathrm{Var}(\hat{\theta}_i) = \frac{\sigma_0^{(i)2}}{n_0} = T_i \quad .$$

We have that

$$\text{(24)} \qquad \mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{N} = \Sigma_{i=0}^{k} T_i = T \ ,$$

by (18), (22), and (23), so that (20) holds. Hence we have the required accuracy, for, by (16), (19), and (24),

$$\text{(25)} \qquad \Pr\{|\theta - \hat{\theta}| \leq t_\alpha \frac{\sigma}{\sqrt{N}} = e\} = 1 - \alpha \ .$$

We can now sum up the results by the basic

Theorem. If the sequential stratification algorithm is applied to $\theta = \int_A^B f(x)\,dx$, and if the strata $[a_i, b_i)$, $i = 0, \ldots, k$, form the stratification of $[A, B)$ recommended by the procedure, then

(a) the estimator $\hat{\theta}_i$ of $\theta_i = \int_{a_i}^{b_i} f(x)\,dx$ is such that,

$$\text{(26)} \qquad \mathrm{Var}(\hat{\theta}_i) \cdot \frac{B-A}{b_i - a_i} = T \ ,$$

where T is a constant and $i = 0, \ldots, k$;

(b) given $e > 0$ and $0 < \alpha < 1$, the sequential stratification estimator of θ, $\hat{\theta}$, is such that

$$\Pr\{|\theta - \hat{\theta}| \leq e\} = 1 - \alpha; \text{ and}$$

(c) given $K + 1$, the number of strata and N, the number of samples, then the above choice of the strata, $[a_i, b_i)$, $i = 0, \ldots, k$,

yields a greater increase in efficiency over crude Monte Carlo than any other choice, by bisection process, of $k + 1$ strata.

Proof.

Part (a) follows directly from (21) and (23).

Part (b) is a restatement of (25).

Part (c) follows from the decision rule (10), by induction.

Remark. Part (c) of the theorem only holds for strata being chosen by a bisection process, that is strata of length $2^{-p}(B-A)$, where p ranges through non-negative integers. The existence of an algorithm which produces an exact optimum choice of strata is still an unsolved problem. The bisection process used in the sequential stratification method was chosen for its simplicity, but many other stratification schemes can easily be used with the algorithm.

In practice, the quantities μ_i and σ_i^2, $i = 0, 1, 2,$ are unknown for all strata $[a, b)$ of the range of integration $[A, B)$. Therefore, we need the following estimates:

(27) μ_0 is estimated by $m_0 = \frac{b-a}{n_0} \sum_{j=1}^{n_0} f(\xi_j)$, where $\xi_j \sim U(a, b)$;

(28) σ_0^2 is estimated by $s_0^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} [(b-a)f(\xi_j) - m_0]^2$.

Estimates for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are similarly defined by m_1, m_2, s_1^2, s_2^2 That these estimates are unbiased follows from

Lemma 3. Let $\mu_i, m_i, \sigma_i^2, s_i^2$ be defined as above; then

$$E[m_i] = \mu_i \qquad \text{and} \qquad E[s_i^2] = \sigma_i^2 \ , \quad i = 0, 1, 2 \ .$$

Proof. The result will be shown for $i = 0$. The other cases follow analogously.

$$E[m_0] = E[\frac{1}{n_0} \Sigma_{j=1}^{n_0} (b-a)\, f(\xi_j)\,], \qquad \text{by (27)},$$

$$= \frac{1}{n_0} \Sigma_{j=1}^{n_0} E[\tau_{0j}] = \mu_0, \qquad \text{by (4) and (5)} .$$

$$E[s_0^2] = \frac{1}{n_0 - 1} \Sigma_{j=1}^{n_0} E[(\{(b-a)f(\xi_j) - \mu_0\} - \{m_0 - \mu_0\})^2], \text{ by (28)},$$

$$= \frac{1}{n_0 - 1} \Sigma_{j=1}^{n_0} \{E[\{(b-a)f(\xi_j) - \mu_0\}^2] - E[(m_0 - \mu_0)^2]\} ,$$

since $E(m_0 - \mu_0) = 0$.

$$E[((b-a)f(\xi_i) - \mu_0)^2] = E[(\frac{1}{n_0} \Sigma_i \tau_{0i} - \mu_0)^2]$$

$$= E[\frac{1}{n_0} \Sigma_i (\tau_{0i} - \mu_0)^2] = \frac{1}{n_0^2} E[\Sigma_i (\tau_{0i} - \mu_0)]^2$$

$$= \frac{1}{n_0^2} \Sigma_i E[(\tau_{0i} - \mu_0)^2] ,$$

since cross product terms vanish.

$$E[(m_0 - \mu_0)^2] = \frac{1}{n_0^2} \cdot n_0 \cdot \sigma_0^2 = \frac{1}{n_0} \sigma_0^2$$

$$E[s_0^2] = \frac{1}{n_0 - 1} \cdot n_0 \left(\sigma_0^2 - \frac{\sigma_0^2}{n_0}\right) = \sigma_0^2 . \qquad \#$$

In practice, the decision rule (10) becomes, stratify if

(29) $$s_0^2 > K(s_1 + s_2)^2 \qquad \text{where } K = k_s/k_c ;$$

or equivalently (11) becomes,

(30) $$(m_1 - m_2)^2 > (K - 2)(s_1^2 + s_2^2) + 2k\, s_1 s_2 .$$

The last relation was introduced as the test criterion to eliminate the calculation of s_0^2. Similarly, the stopping rule (23) becomes, stop sampling in stratum $[a, b)$, when

(31) $$\frac{s_0^2}{n_0} = T_i .$$

To determine the statistical estimates m_i, s_i^2, $i = 0, 1, 2$, random samples in the stratum $[a, b)$ being considered must be drawn. Denote the initial samples by $n_0^{(0)}$, $n_1^{(0)}$, $n_2^{(0)}$, where $n_1^{(0)}$, $n_2^{(0)}$ is the number of sample points in $[a, c)$, $[c, b)$ respectively, where $c = \frac{1}{2}(a+b)$ and $n_0^{(0)} = n_1^{(0)} + n_2^{(0)}$. Since the optimum values for n_1 and n_2, as in lemma 2, are unknown (because σ_1 and σ_2 are not known until they can be estimated by s_1 and s_2), we set $n_1^{(0)} = n_2^{(0)}$ to produce the estimates $m_1^{(0)}$, $m_2^{(0)}$, $s_1^{(0)}$, $s_2^{(0)}$. Using these estimates, the decision rule (30) determines where the additional sampling is needed. The algorithm does not calculate the values $\hat{n}_1$ and $\hat{n}_2$, which minimize $\sigma_1^2/n_1 + \sigma_2^2/n_2$. What is done, is to locate the strata where the variation of the function f is

considerable, as determined by the decision rule, and concentrate the sampling on those strata to see if further stratification is necessary. This, in fact, is approximately equivalent to finding $\hat{n}_1$ and $\hat{n}_2$. To see this, apply lemma 2, to yield $\hat{n}_1/\hat{n}_2 = \sigma_1/\sigma_2$, which states that the optimum sampling strategy is to draw a sample directly proportional to the standard error of the estimate of θ for the stratum. This is precisely what the algorithm accomplishes, as is shown by the result in (26).

Results:

Example Number	Integrand	[A, B)	$\theta = \int_A^B f(x)\,dx$	$\hat{\theta}$	$\lvert\theta-\hat{\theta}\rvert$	No. of points	No. of points for crude MC	desired accuracy	3.0 std. error
1	$\log x$	[0, 1)	-1.000	-.9983	.0017	360	696	.1	.060
2	x^8	[0, 8)	14,913,080.89	14,913,725.99	645.10	12,280	6×10^7	10^4	5100.
3	$[\sin^2(x)+\cos(x^3+1)e^{\frac{x^2}{8}}]^2$	[0, 2π)	3,308.96	3,329.91	20.95	3,300	8.6×10^4	100.	67.8
4	$\sin x$	[0, 2π)	0.	.002	.002	3,431	18,900	.1	.100
5	$\sin x$	[0, 2)	1.4162	1.4136	.0026	200	407	.1	.045
6	$\frac{e^x-1}{e-1}$	[0, 1)	.418	.420	.002	600	73,620	.01	.006
7	$\frac{e^x-1}{e-1}$	[0, 1)	.418	.418	.000	3,080	736,200	.001	.000

For the above examples the sequential stratification produced the following stratifications:

Example No.	Stratification Points
1	(0, 1/16, 1/8, 1/4, 1/2, 1)
2	$(0, 2, 3, 3\frac{1}{2}, 3\frac{3}{4}, 4, \ldots, 7\frac{63}{64}, 7\frac{127}{128}, 8)$
3	$(0, \pi, 5\pi/4, 3\pi/2, 13\pi/8, 7\pi/4, 15\pi/8, 2\pi)$
4	$(0, \pi, 2\pi)$
5	(0, 1/2, 1, 2)
6	(0, 1/8, 1/4, ..., 7/8, 1)
7	(0, 1/32, ... , 31/32, 63/64, 1)

Remarks:

In each example, the actual error, $|\theta-\hat{\theta}|$, was less than the desired accuracy. The number of sample points necessary to achieve the same desired accuracy for crude Monte Carlo is given and in each case greater than that for stratification. If instead of the desired accuracy, we used the number of samples needed to achieve the actual error, the results would have been even more impressive. Note that

the integrand, $\log x$, has a singularity at $x = 0$, however this did not cause us any trouble. Examining the stratification points, we see that the length of the stratum is inversely proportional to the variation of the function, that is

$$\int_a^b |f(x) - \bar{f}|^2 \, dx, \text{ where } \bar{f} = (b-a)^{-1} \int_a^b f(x)\, dx .$$

The authors have also estimated the θ in example 2 by an antithetic variate transformation using 180 sample points and obtained an estimate with ten digit accuracy! An adaptive iterative Simpson's rule program which yielded an estimate of six digit accuracy, for the θ in example 3, ran nine times faster than the Monte Carlo program; a result which is not surprising, since Monte Carlo generally excels classical quadrature only when applied to multiple integrals of sufficiently large dimension. Examples 4 and 5, show how a change in the value B, of the range of integration $[A, B)$, affects the stratification of the interval, while examples 6 and 7 show how a change in the desired accuracy affects the stratification.

Conclusions

The advantage of using the sequential stratification procedure is in allowing the computer to search for an optimum placement of the strata. The effect of the stratification is to concentrate the sampling on the strata where the integrand has the greatest variation.

Further reduction in the variance of the estimator may be achieved, if in each stratum control variates, importance sampling, or antithetic variates were used instead of the crude Monte Carlo sampling. The decision and stopping rules used here would have to be modified, but the change is straight forward.

The extension to multiple integrals can be made by one of the following schemes. The first is to use the iterated integral representation for (1). Then, we can use the one dimensional scheme recursively to estimate θ. This method can be immediately ruled out, except when the number of dimensions is small, because of the exponential increase in the number of samples needed as the dimensionality of the integral increases.

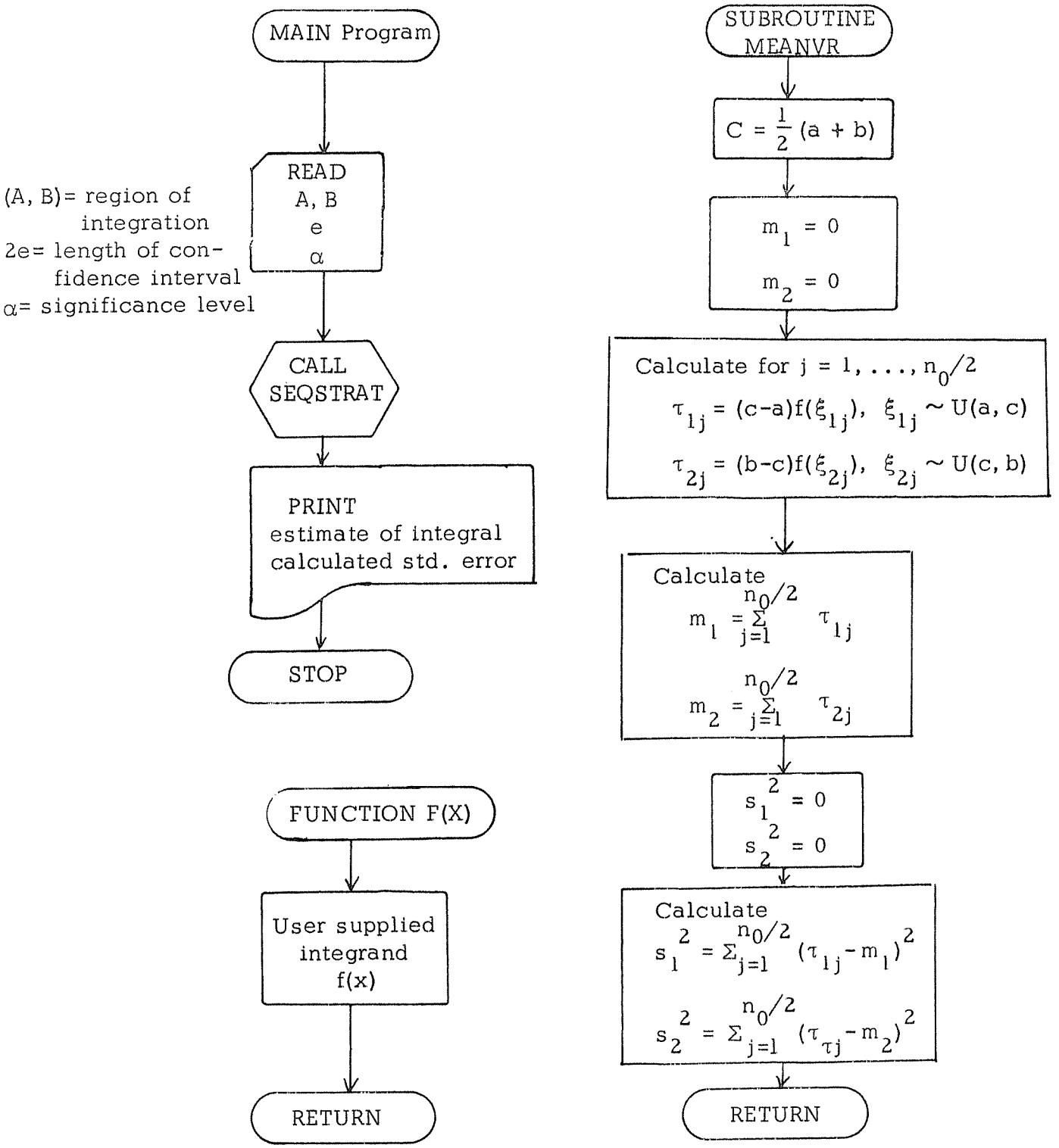
Let $\theta = \int_R f(\underline{x})\, d\underline{x}$ as in (1), where R is a k-cell, that is $R = \{\underline{x}: A_i \le x_i \le B_i,\ i = 1, \ldots, k\}$. The second method consists of stratifying a subregion, R_h, along the x_j-th's axis when $\delta_j = \max_i \delta_i$, $i = 1, \ldots, k$, and $\delta_j > 0$, where $\delta_i = \sigma_{0i}^2 - k(\sigma_{1i} + \sigma_{2i})^2$ and

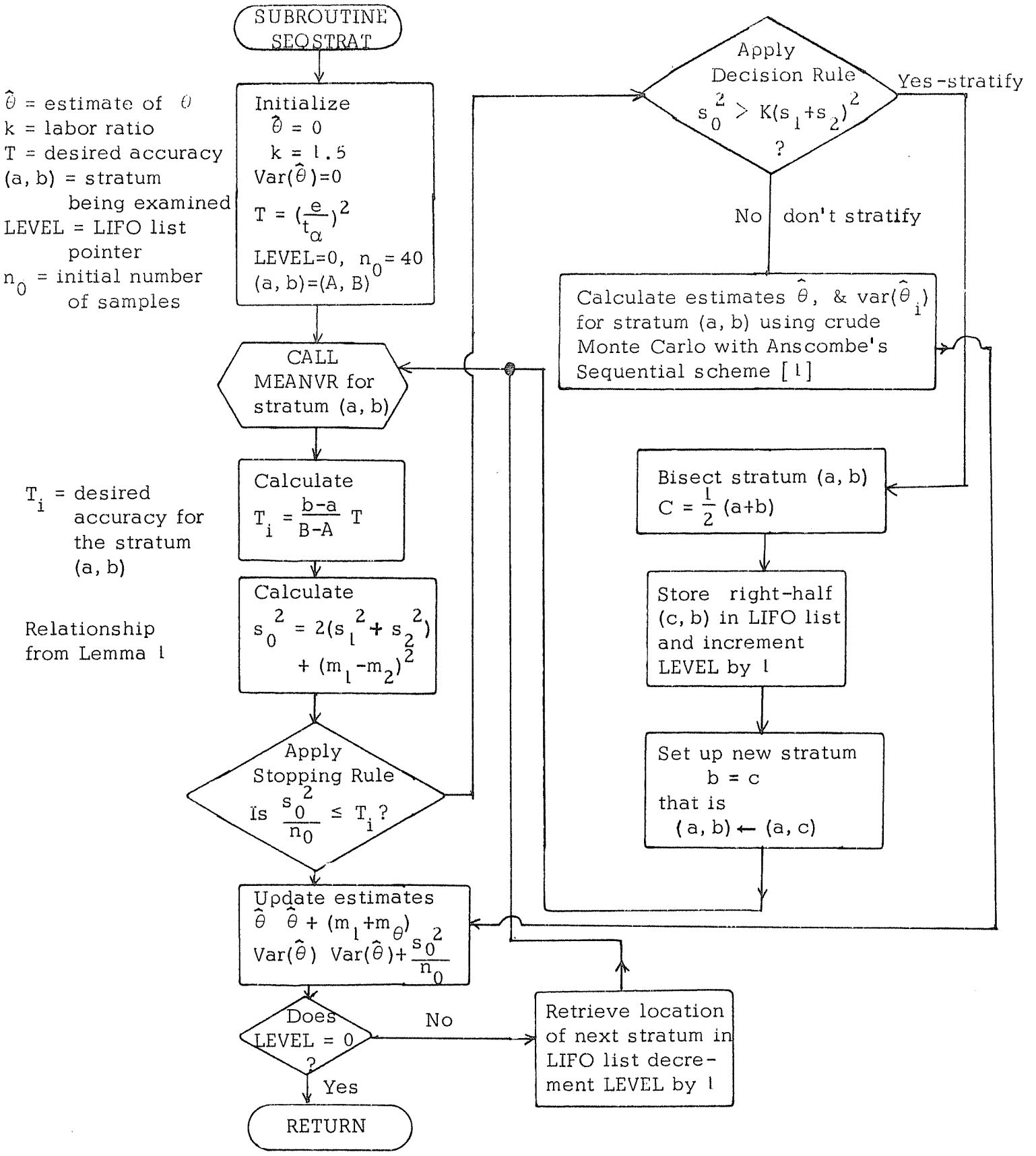
$\sigma_{0i}^2 = \text{Vol}(R_h) \int_{A_i}^{B_i} f(\underline{x})^2 dx_i - [\int_{A_i}^{B_i} f(\underline{x})\, dx_i]^2$. If $\max \delta_i \leq 0$, then we don't stratify. The third approach would be to apply the decision rule in a direction chosen randomly. This method would be necessary when the dimensionality of the integral is very large. For these two methods, we have that as the dimensionality of the integration increases, the number of sample points needed increase only linearly.

The general idea is to isolate those subregions in which the function varies considerably and concentrate our sampling there with the restriction that the search produces a stratification that reduces the amount of labor to yield the desired accuracy.

Acknowledgment

The authors wish to express their thanks to Lucio Tavernini for his very thorough explanation of his multidimensional quadrature procedure, error analysis, and results; and to the National Science Foundation, for supporting the present work under grant number GJ-171.


MAIN Program
READ
A, B
e
α
CALL
SEQSTRAT
PRINT
estimate of integral
calculated std. error
STOP
(A, B)= region of integration
2e= length of confidence interval
α= significance level
FUNCTION F(X)
User supplied
integrand
f(x)
RETURN
SUBROUTINE
MEANVR
C = 1/2 (a + b)
m1 = 0
m2 = 0
Calculate for j = 1, . . . , n0/2
τ1j = (c-a)f(ξ1j), ξ1j ~ U(a, c)
τ2j = (b-c)f(ξ2j), ξ2j ~ U(c, b)
Calculate
m1 = Σ j=1 n0/2 τ1j
m2 = Σ j=1 n0/2 τ2j
s1^2 = 0
s2^2 = 0
Calculate
s1^2 = Σ j=1 n0/2 (τ1j - m1)^2
s2^2 = Σ j=1 n0/2 (ττj - m2)^2
RETURN


SUBROUTINE SEQSTRAT
$\hat{\theta}$ = estimate of θ
k = labor ratio
T = desired accuracy
(a, b) = stratum being examined
LEVEL = LIFO list pointer
n_0 = initial number of samples
Initialize
$\hat{\theta} = 0$
k = 1.5
$Var(\hat{\theta})=0$
$T = (\frac{e}{t_\alpha})^2$
LEVEL=0, $n_0 = 40$
(a, b)=(A, B)
CALL MEANVR for stratum (a, b)
T_i = desired accuracy for the stratum (a, b)
Calculate
$T_i = \frac{b-a}{B-A} T$
Relationship from Lemma 1
Calculate
$s_0^2 = 2(s_1^2 + s_2^2) + (m_1-m_2)^2$
Apply Stopping Rule
Is $\frac{s_0^2}{n_0} \leq T_i$?
Update estimates
$\hat{\theta}$ $\hat{\theta} + (m_1+m_\theta)$
$Var(\hat{\theta})$ $Var(\hat{\theta}) + \frac{s_0^2}{n_0}$
Does LEVEL = 0 ?
No
Yes
RETURN
Retrieve location of next stratum in LIFO list decrement LEVEL by 1
Apply Decision Rule
$s_0^2 > K(s_1+s_2)^2$?
Yes-stratify
No don't stratify
Calculate estimates $\hat{\theta}$, & $var(\hat{\theta}_i)$ for stratum (a, b) using crude Monte Carlo with Anscombe's Sequential scheme [1]
Bisect stratum (a, b)
$C = \frac{1}{2}(a+b)$
Store right-half (c, b) in LIFO list and increment LEVEL by 1
Set up new stratum
b = c
that is
(a, b) ← (a, c)

APPENDIX

Calculation of the labor ratio.

Using the current convention, the labor required will be measured by the number of multiplications and divisions. First, we will find k_c, the average amount of labor for a single crude Monte Carlo estimate.

Let r be the labor in generated a pseudo-random number uniformly distributed on $[0, 1]$, f be the labor required to evaluate $f(x)$, and n be the number of samples. Then

$$k_c = \frac{2}{n} + r + f + 3 ,$$

is the average amount of labor required to calculate $\hat{\theta}$ and estimate its variance. Next, we calculate k_s, the average amount of labor for a single stratified Monte Carlo estimate.

In addition to the above definitions, let s be the amount of labor required to find the square root of a given number. Then

$$k_s = \frac{9+2s}{n} + r + f + 3 ,$$

is average amount of labor required to calculate $\hat{\theta}_s$ and estimate its variance. Therefore,

$$K = \frac{k_s}{k_c} = 1 + \frac{1}{n} \left(\frac{7 + 2s}{r + f + 3}\right) ,$$

is the labor ratio. Hence, $K \rightarrow 1$ as $n \rightarrow \infty$ and we can now find an upper bound for k.

Assuming $r = s$, $n \geq 2$, and $f \geq 1$, we have that

$$K \leq 1 + \frac{2r + 7}{2r + 8} < 2,$$

since r is finite.

REFERENCES

1. Anscombe, F. J. Sequential estimation. Journ. Roy. Statist. Soc., B, 15 (1953), 1-29.

2. Dalenius, T. and Hodges, J. L. The choice of stratification points. Skandinavisk Aktuarietidskrift, 3-4 (1957), 198-203.

3. Halton, J. H. Sequential Monte Carlo (Revised). Technical Summary Report 816 (Mathematics Research Center, U.S. Army, The University of Wisconsin, Madison, 1967).

4. Hammersley, J. M. and Handscomb, D. C. Monte Carlo Methods, (Methuen, London, 1964).

5. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. Sample Survey Methods and Theory, Vol. I, (John Wiley, New York, 1953).

6. Loève, Probability Theory, (Van Nostrand, Princeton, N. J., Third edition, 1963).

7. McKeeman, W. M. and Tesler, L. Algorithm 182, Communications of the Association for Computing Machinery, 6 (1963), 315.

8. Tavernini, L. Library Programs and Subroutines for the 3600 Computer, UWCC, 4 (1969), 3.46.

9. Zeidman, E. A., The Evaluation of Multidimensional integrals by Sequential Stratification, [In preperation].