THE USE OF MULTI-STAGE SAMPLING SCHEMES IN MONTE CARLO COMPUTATIONS

Andrew W. Marshall

P-531

3 June 1954

The RAND Corporation

1700 MAIN ST. • SANTA MONICA • CALIFORNIA

THE USE OF MULTI-STAGE SAMPLING SCHEMES IN MONTE CARLO COMPUTATIONS*

I. Introduction

One of the available techniques of Monte Carlo computations, importance sampling, can make Monte Carlo computations much more efficient if it is possible to choose judiciously the probability distribution from which the sample observations are drawn. The details of this technique will be reviewed below. The difficulty in practice is that one is often not able to specify a priori what the more efficient probability distribution is, or even if a good choice can be made with regard to the parameter class of distributions to be used the parameter values are difficult to determine. In this situation one naturally thinks of using some sort of multi-stage sampling in which information obtained in a preliminary sample is used to determine the way in which the remainder of the sample is to be picked. Some results relating to procedures of this type are described below.

We turn now to the description of the setting of the problem and the technique of importance sampling.

---

*Presented at Symposium on Monte Carlo Methods, Gainesville, Florida, March 17, 1954.
Prepared while the author was at the University of Chicago.

II. The Setting of the Problem [1]

The problem to be discussed is the estimation of ξ, where

$$\xi = \int_{-\infty}^{\infty} g(x)\, f(x,\theta)\, dx$$

where $g(x)$ is some known function of x; e.g., x, x^2, $\sin x$, $x^2 + 3x^3$, or even

$$g(x) = 0 \qquad x \leq a$$
$$= 1 \qquad x > a.$$

$f(x,\theta)$ is the probability distribution of x, where $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of parameters determining the probability distribution. Frequently the characteristic of the problem that makes Monte Carlo methods preferable to ordinary numerical methods is the complexity of $f(x,\theta)$. Often it is not possible or convenient to write down $f(x,\theta)$ in closed form, but one is able to sample from it.

The general idea of importance sampling is the following: rather than sampling from $f(x)$, suppressing θ in our notation for the moment, it is preferable to sample from another probability distribution $h(x)$ giving

$$\xi = \int_{-\infty}^{\infty} \frac{g(x)\, f(x)}{h(x)}\, h(x)\, dx = \int_{-\infty}^{\infty} g^*(x)\, h(x)\, dx$$

where $h(x)$ is restricted to be a probability distribution such that $h(x) \neq 0$ unless $g(x)\, f(x) = 0$. There is therefore a wide choice available in choosing $h(x)$ and if we consider the estimate

$$\overline{g^*(x)} = \frac{1}{n} \sum_{i=1}^{n} g^*(x_i)$$

based upon a sample $x_1 \ldots x_n$ from $h(x)$ it is possible to make it a better estimate of ξ by a proper choice of $h(x)$ than $\overline{g(x)}$, where

$$\overline{g(x)} = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

when the sample $x_1 \ldots x_n$ is drawn from $f(x)$. Both estimates are consistent, unbiased estimates of ξ. The variance of the estimate $\overline{g^*(x)}$ is equal to

$$\sigma_n^2 = \frac{1}{n}\left[\int_{-\infty}^{\infty} \left[\frac{g(x)\, f(x)}{n(x)}\right]^2 h(x)\, dx - \xi^2\right].$$

It can be shown that the optimal $h(x)$, the $h(x)$ that minimizes the variance of its associated estimate $\overline{g^*(x)}$, is

$$h(x) = \frac{|g(x)|\, f(x)}{\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx} .$$

If $g(x) \geq 0$, $h(x) = \frac{g(x)\, f(x)}{\xi}$ is the optimum choice of $h(x)$ and the variance of the associated $\overline{g^*(x)}$ is zero. If $g(x)$ takes on both positive and negative values then a zero variance estimate is not possible unless more complicated procedures are used. It will be pointed out below it is seldom possible, or practical, to achieve these optimum results.

Before going on I wish to discuss a difficulty introduced by the application of importance sampling techniques that is glossed over in the above formulation of the problem. It often is the case that x is a very complicated random variable and that one does not pick an x from $f(x)$ but constructs x as a function of several other random variables. For example, suppose the random variable x is the statistic w_n^2,

$$w_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(y_i - \frac{2i-1}{2n}\right)^2$$

where the y's are drawn from the uniform distribution $(0,1)$ and $y_1 \le y_2 \le \dots \le y_n$. In this case x is a function of several random variables and we denote this by $x(y_1, \dots, y_n) = x(y)$. It may then be easier to construct x from the y's than to draw x from $f(x)$, or even this may be the only practical way to obtain sample values of x. The values of x obtained are drawn from $f(x)$ and thus

$$\xi = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g\left[x(y_1, \dots y_n)\right] \ell(y_1, \dots, y_n)\, dy_1 \dots dy_n .$$

The importance sampling must enter into the problem by altering the distribution of the y's, choosing them not from $\ell(y_1, \dots, y_n)$ but from say $k(y_1, \dots, y_n)$ and our estimate of ξ will be

$$\frac{1}{n} \sum_{i=1}^{n} g\left[x(y_1^{(i)}, \dots, y_n^{(i)}\right] \frac{\ell(y_1^{(i)}, \dots, y_n^{(i)})}{k(y_1^{(i)}, \dots, y_n^{(i)})} ,$$

using superscripts to denote the ith vector of sample values. This estimate is essentially $g(x_i)$ weighted by the likelihood ratio of the y's yielding x_i, and this is not necessarily the same as weighting $g(x_i)$ with $f(x_i)/h(x_i)$. If the same value of x can arise from many different vectors, y, there is no guarantee that

$$\frac{f[x(y)]}{h[x(y)]} = \frac{\ell(y)}{k(y)} .$$

Picking the y's from the distribution specified by $k(y)$ implies a distribution $h(x)$ and it is suggestive to think of the variance of the random variable $g[x(y)] \frac{\ell(y)}{k(y)}$ as being composed of two components: (1) variation in $g[x(y)] \frac{f[x(y)]}{h[x(y)]}$ and (2) variation of $\frac{\ell(y)}{k(y)}$ from $\frac{f[x(y)]}{h[x(y)]}$.

In the following discussion of the problem of finding good $h(x)$'s through multi-stage sampling procedures the explicit treatment will be in terms that imply one is sampling from $f(x)$ and $h(x)$ directly. The changes that will be made when this is not the case are clear and readers should constantly keep in mind the distinction, indicated in the above paragraph, between the two problems.

In practice, as was indicated above, one does not work with completely general classes of distribution functions but confines his choice of $h(x)$ to some parametric family of distribution functions, often confining selection still further by using some family of distributions closely related to $f(x,\theta)$. Thus in effect one makes two choices: (1) the choice of parametric family $h(x,\alpha)$, where α is the vector of parameters defining particular distributions within the family and (2) the choice of some particular parameter value, α'. It is the latter choice that will concern us most in the remainder of the discussion. The choice of the class of distributions, $h(x,\alpha)$, is a very difficult problem and only a few general rules are available for making this decision. In some problems intuition based upon the physical structure of the real problem is of help. For the moment we will consider that the choice of $h(x,\alpha)$ has been made and our problem is to choose α' in some optimal way. If $h(x,\alpha)$ is a one parameter family of distribution functions we may illustrate the problem by plotting the variance of the associated estimate of ξ as a function of α. A typical case is shown in Chart I. As shown in Chart I there will be some value of α, α_0, which is associated with the minimum variance estimate of ξ. Also there are often values of α, such as α_1, that are poles of $V(\alpha)$; i.e., $V(\alpha_1) = \infty$. For

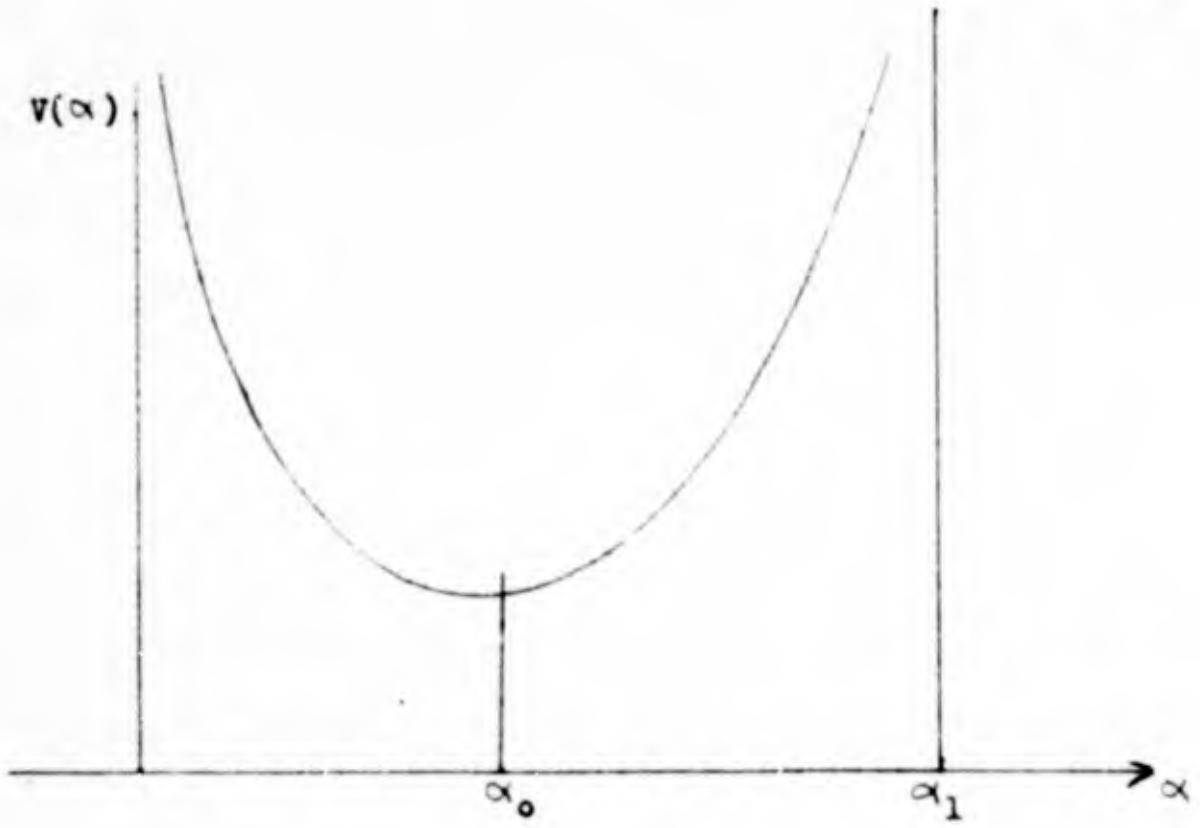

CHART I.

example,* if $f(x,\theta) = \theta\, e^{-\theta x}$, $g(x) = x$, and the class of distribution functions $h(x)$ is restricted to $h(x,\alpha) = \alpha^2 x\, e^{-\alpha x}$, then

$$V(\alpha) \begin{cases} = -\dfrac{\theta^2}{\alpha^2(2\theta-\alpha)^2} - \dfrac{1}{\theta^2} ; & 0 < \alpha < 2\theta \\ = -\infty ; & \alpha \geq 2\theta \end{cases}$$

or if $h(x,\alpha) = \alpha\, e^{-\alpha x}$, then

---

*An apology of sorts must be made for the type of examples it is possible to offer in discussions of Monte Carlo methods. In general they are of little interest themselves because the problems of most interest, as far as applications of Monte Carlo methods are concerned, are those that are not analytically tractable. All one can hope for is that the examples are revealing as far as principles are concerned.

$$V(\alpha) \begin{cases} = \dfrac{2\theta^2}{\alpha(2\theta-\alpha)^3} - \dfrac{1}{\theta^2} ; & 0 < \alpha < 2\theta \\ = \infty & ; \alpha \geq 2\theta . \end{cases}$$

In summary, then, the problem is the following: $f(x,\theta)$ is given, i.e., both the form of the distribution and the value of the parameter vector θ are known, and $g(x)$, a known function, is also given. An estimate is to be made of ξ, where

$$\xi = \int_{-\infty}^{\infty} g(x)\, f(x,\theta)\, dx .$$

The choice of $h(x,\alpha)$ has been made, but the parameter α, or rather a good value of it is not known. In order to obtain an estimate of ξ one simple suggestion would be that a sample, $x_1, \ldots, x_{n_1}$, be drawn from $f(x,\theta)$ and on this basis an estimate be made of α_o, say $\hat{\alpha}_o$, and the sampling then proceed using $h(x, \hat{\alpha}_o)$ as the distribution from which further samples are drawn. Some combined estimate based upon $\hat{\xi}_1$ and $\hat{\xi}_2$, the estimates obtained from this two stage sampling plan would be used to estimate ξ. Given a fixed cost for the computing program we wish to obtain an estimate having optimal characteristics, e.g., minimum variance or minimum risk estimates. The elements at our command are the choice of methods of estimating the best value α and the balance between the size of the initial sample and the second sample; that is, balancing the value of an improved estimate of α_o against a smaller opportunity to make use of it for estimating ξ. Of course, once the restriction of the sampling scheme to two stages is lifted other possibilities arise.

III. A Two Stage Sampling Procedure for Estimating ξ.

In this section it is proposed to discuss some aspects of the two staged sampling procedure suggested in the paragraph immediately above. The first problem to be treated will be the sampling distribution of an estimate $\hat{\alpha}_o$ of α_0, the value of α for which $V(\alpha)$ takes on its minimum value, in the case where $h(x,\alpha)$ is a single parameter family of distributions.

We have

$$\xi = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{h(x,\alpha)} h(x,\alpha)\, dx$$

$$V(\alpha) = \int_{-\infty}^{\infty} g^2(x) \frac{f(x)}{h(x,\alpha)} f(x)\, dx - \xi^2 ,$$

and (1) assume that we have

$$\frac{\partial V(\alpha)}{\partial \alpha} = \int_{-\infty}^{\infty} g^2(x) \frac{f(x)}{h^2(x,\alpha)} \left[\frac{-\partial h(x,\alpha)}{\alpha}\right] f(x)\, dx$$

$$= \int_{-\infty}^{\infty} \Psi(x,\alpha)\, f(x)\, dx,$$

$$\frac{\partial^2 V(\alpha)}{\partial \alpha^2} = \int_{-\infty}^{\infty} \frac{\partial \Psi(x,\alpha)}{\partial \alpha} f(x)\, dx,$$

and

$$\frac{\partial^3 V(\alpha)}{\partial \alpha^3} = \int_{-\infty}^{\infty} \frac{\partial^2 \Psi(x,\alpha)}{\partial \alpha^2} f(x)\, dx \quad .$$

This assumes, in effect, that for almost all x, the derivatives $\frac{\partial h}{\partial \alpha}$, $\frac{\partial^2 h}{\partial \alpha^2}$, and $\frac{\partial^3 h}{\partial \alpha^3}$ exist for every α belonging to a non-degenerate interval A.

(2) For every α in A, it is also assumed that $\left|\frac{\partial^2 \Psi(x,\alpha)}{\partial \alpha^2}\right| < H(x)$ and $\int_{-\infty}^{\infty} H(x)\, f(x)\, dx < M$, where M is independent of α.

(3) For every α in A, the integral $\int_{-\infty}^{\infty} \Psi^2(x,\alpha)\, f(x)\, dx = E_f\left[\Psi^2(x,\alpha_0)\right]$ is finite and positive.*

If we now denote α_0 as the value of α for which $\frac{\partial V(\alpha)}{\partial \alpha}\Big|_{\alpha_0} = 0$, or the minimizing value of α, an estimate of α_0 is $\hat{\alpha}_0$ the solution of the equation

$$\frac{1}{N}\sum_{i=1}^{n} \Psi(x_i,\alpha) = 0 .$$

By now the whole analogy to maximum likelihood estimation is obvious, although only in terms of the asymptotic distribution theory and its proof. $\hat{\alpha}_0$ is obviously a consistent estimate α_0 and we are only concerned here with its distribution. In order to investigate the asymptotic distribution of $\hat{\alpha}_0$ we expand $\Psi(x,\alpha)$ about α_0 and obtain [2]

$$\Psi(x,\alpha) = \Psi(x,\alpha_0) + \frac{\partial \Psi(x,\alpha_0)}{\partial \alpha}(\alpha - \alpha_0) + \frac{1}{2}\phi\, H(x)\,(\alpha - \alpha_0)^2$$

where $|\phi| < 1$. Thus the equations determining $\hat{\alpha}_0$ may be written as

$$\frac{1}{N}\sum_{i=1}^{N} \Psi(x_i,\hat{\alpha}_0) = \frac{1}{N}\sum_{i=1}^{N} \Psi(x_i,\alpha_0) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \Psi(x_i,\alpha_0)}{\partial \alpha}(\hat{\alpha}_0 - \alpha_0)$$

$$+ \frac{1}{2}\phi \sum_{i=1}^{N} H(x_i)(\hat{\alpha}_0 - \alpha_0)^2 = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} \Psi(x_i,\hat{\alpha}_0) = B_0 + B_1(\hat{\alpha}_0 - \alpha_0) + \frac{1}{2}\phi\, B_2(\hat{\alpha}_0 - \alpha_0)^2 = 0.$$

---

*We denote by $E_f(y)$ the expected value of the random variable y with respect to the probability density function f, thus we might write $\xi = E_f\left[g(x)\right]$.

The B's are random variables, being functions of random variables $x_1 \ldots, x_N$. By Khintchine's theorem B_0 converges in probability to zero, B_1 converges in probability to $\frac{\partial^2 V(\alpha_0)}{\partial \alpha^2}$, and B_2 converges to the non-negative value $E\left[H(x)\right] < M$. Thus when we rewrite the above equation as

$$\sqrt{N}\,(\hat{\alpha}_0 - \alpha_0) = \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(x_i, \alpha_0)}{-\left[B_1 + \frac{1}{2} \phi B_2 (\hat{\alpha}_0 - \alpha_0)\right]}$$

we see that the denominator converges to $\frac{-\partial^2 V(\alpha_0)}{\partial \alpha^2}$ and that the numerator is essentially the sum of independent random variables each with mean zero and variance $E_f\left[\psi^2(x, \alpha_0)\right]$, thus the central limit theorem applies and the sum $\sum_{i=1}^{N} \psi(x_i, \alpha_0)$ is asymptotically normal with mean zero and variance $NE_f\left[\psi^2(x, \alpha_0)\right]$. Therefore $\sqrt{N}\,(\hat{\alpha}_0 - \alpha_0)$ is asymptotically normal with mean zero and variance

$$\frac{E_f\left[\psi^2(x, \alpha_0)\right]}{\left[\frac{\partial^2 V(\alpha_0)}{\partial \alpha^2}\right]^2} .$$

In the case where α is a vector of parameters $(\alpha_1, \ldots, \alpha_k)$ similar results on the joint distribution of sample estimates can be obtained in a straightforward way.

We turn now to a description of an estimation procedure for ξ: a sample of size n_1 is drawn from $f(x,\theta)$ and estimates $\hat{\xi}_1$ and $\hat{\alpha}_0$ are produced,

$$\hat{\xi}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} g(x_i)$$

then a sample of size n_2 is drawn from $h(x, \hat{\alpha}_0)$ and an estimate $\hat{\xi}_2$ computed,

$$\hat{\xi}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} g(x_i) \frac{f(x_i,\theta)}{h(x_i,\alpha_0)}$$

and finally the two estimates of ξ are combined to form $\hat{\xi}$,

$$\hat{\xi} = \hat{w}_1 \hat{\xi}_1 + \hat{w}_2 \hat{\xi}_2$$

where the weights $\hat{w}_1$ and $\hat{w}_2$ are perhaps given by

$$\hat{w}_1 = \frac{\hat{\sigma}_2^2/n_2}{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$$\hat{w}_2 = 1 - \hat{w}_1$$

and

$$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=1}^{n} g^2(x_i) - \hat{\xi}_1^2$$

$$\hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^{n} g^2(x_i) \frac{f^2(x_i,\theta)}{h^2(x_i,\alpha)} - \hat{\xi}_2^2 .$$

If one were able to estimate before sampling the values of σ_1^2 and σ_2^2 as function of n_1 and n_2 a reasonable choice of n_1 and n_2 could, perhaps, be based upon the loss function $L(\xi, \hat{\xi}) = \lambda(\xi - \hat{\xi})^2$. In this case one would try to minimize by proper choice of n_1 and n_2 the expected loss (risk), $E\left[L(\xi, \hat{\xi})\right] = \lambda \operatorname{Var}(\hat{\xi}) = R(n_1, n_2)$, the expectation being taken relative to the probability distribution of $\hat{\xi}$, subject to the condition $(c_0 + c_1 n_1 + c_2 n_2) = c$, where

c_0 = initial cost of programming and cost of computing estimate $\hat{\alpha}_0$, etc., in other words all fixed costs given that two stages of sampling are to be done,

c_1 = cost per observation from $f(x,\theta)$,

c_2 = cost per observation from $h(x,\hat{\alpha}_0)$.

In general it will not be possible to evaluate $E\left[L(\xi,\hat{\xi})\right] = V(\hat{\xi}) = R(n_1,n_2)$ beforehand because the values of the required expected values, etc., will not be known. In any case $V(\hat{\xi})$ will be a quite complicated expression since $\hat{\xi}_1$, $\hat{\sigma}_1^2$, $\hat{\xi}_2$, and $\hat{\sigma}_2^2$ are not independent of one another. Although $E(\hat{\xi}_2|\hat{\alpha}_0) = \xi$, the mean value of $\hat{\xi}$ is not ξ, in general, because of the bias introduced by the various covariance between $\hat{\xi}_1$, $\hat{\sigma}_1^2$, $\hat{\xi}_2$ and $\hat{\sigma}_2^2$.* An additional difficulty is that in many cases $E_p\left[V\left[\hat{\xi}_2(\hat{\alpha}_0)\right]\right]$ is unbounded. This is easily illustrated in the case of the examples given earlier where there is always a positive, but small, probability of $\hat{\alpha}_0 \geq 2\theta$ for any finite value of n_1. The situation is illustrated in Chart II, where $P\left[\hat{\alpha}_0(n_1)\right]$ is the probability density function of $\hat{\alpha}_0$, $n_1' > n_1$. As n_1 becomes large the probability approaches one that sampling from $h(x,\hat{\alpha}_0)$ is preferable to sampling from $f(x,\theta)$. Nonetheless $E_p\left[V\left[\hat{\xi}_2(\hat{\alpha}_0)\right]\right]$ is unbounded. No reasonable man would forego the advantages of sampling from $h(x,\hat{\alpha})$ if the probability is near enough to one that it will improve his estimate of ξ despite the

---

*This bias in the estimate $\hat{\xi}$ can be removed by the use of a sampling technique called hybrid-splitting. This technique requires that two two-staged samples be taken, with n_1 and n_2 the same in each, and the weights $\hat{w}_1$ and $\hat{w}_2$ of one of the two-staged samples be used to combine the estimates $\hat{\xi}_1$ and $\hat{\xi}_2$ of the other two-staged sample. In this way the weights and the estimates are independent and the resulting weighted estimates $\hat{\xi}^{(1)}$ and $\hat{\xi}^{(2)}$ are both unbiased estimates of ξ of equal value a priori. Thus $1/2\ \hat{\xi}^{(1)} + 1/2\ \hat{\xi}^{(2)}$ is a reasonable pooled estimate, also unbiased. The removal of bias via this means is not always, or usually, costless in terms of the variance of the final estimate of ξ and these costs must be balanced against the value of the reduction of bias obtained.

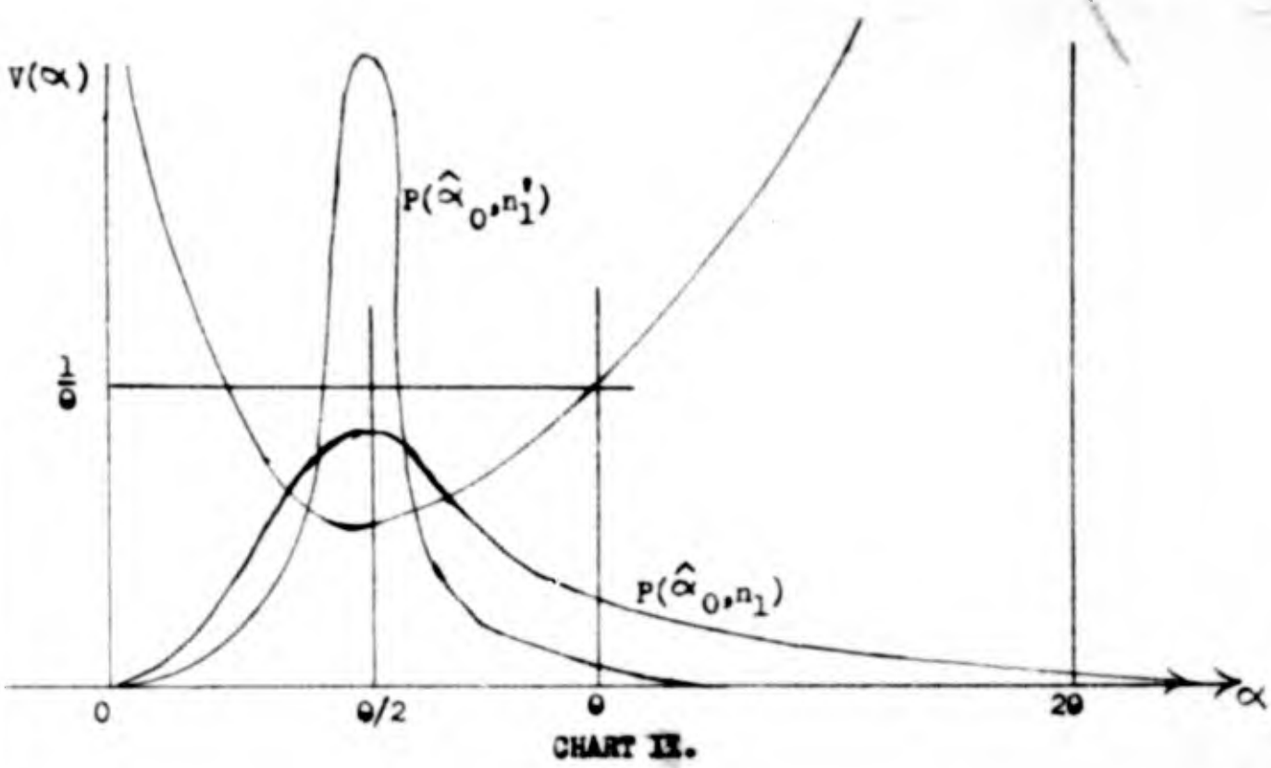

CHART II.

resulting unbounded expected value of the variance of the estimate obtained from $h(x,\hat{\alpha}_0)$. Of course, since the weight given to the estimate derived from the second sample tends to zero when its variance is large the expected or average variance of $\hat{\xi}$ is not infinite but is essentially σ_1^2/n_1. This implies that one's loss function is not really $\lambda(\xi-\hat{\xi})^2$ and as a convenient alternative I suggest that $V(\hat{\alpha})$ be replaced by a quadratic approximation about α_0 and further that the expected loss, $R(n_1,n_2)$, be simplified to

$$R(n_1,n_2) = w_1^2\, V(\hat{\xi}_1) + w_2^2\, V_{Approx}\,(\hat{\xi}_2) = \frac{w_1^2\,\sigma_1^2}{n_1} +$$

$$+ \frac{w_2^2}{n_2}\left[V(\alpha_0) + \frac{E\left[\psi^2(x,\alpha_0)\right]}{2n_1\dfrac{\partial^2 V(\alpha_0)}{\partial\alpha^2}}\right].$$

$$\sigma_1^2 = \frac{1}{n_0}\sum_{i=1}^{n_0} g^2(x_i) - \overline{g(x_i)}^2$$

$$\hat{V}(\alpha_0) = \frac{1}{n_0}\sum_{i=1}^{n_0} g^2(x_i)\frac{f(x_i,\theta)}{h(x_i,\hat{\alpha}_0)} - \overline{g(x_i)}^2$$

$$\hat{E}_f\left[\psi^2(x,\alpha_0)\right] = \frac{1}{n_0}\sum_{i=1}^{n_0}\psi^2(x_i,\hat{\alpha}_0)$$

$$\frac{\partial^2 V(\hat{\alpha}_0)}{\partial\alpha^2} = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\partial\psi(x_i,\hat{\alpha}_0)}{\partial\alpha}$$

where $\hat{\alpha}_0$ is, of course, the solution of

$$\frac{1}{n_0}\sum_{i=1}^{n_0}\psi(x_i,\alpha) = 0.$$

If $\Delta(n_0) > 0$: Continue sampling from $f(x,\theta)$ and take a sample of size n, and test again using $\Delta(n_0 + n)$

$\Delta(n_0) \leq 0$: Take sample of size $n_2(n_0)$ from $h\left[x,\hat{\alpha}_0(n_0)\right]$.

Sampling is to proceed by repeated application of this test until the decision has been made to sample from $h(x,\hat{\alpha})$, at say the mth step when $\Delta(n_0 + mn) < 0$; then sampling will proceed from $h\left[x,\hat{\alpha}_0(n_0+mn)\right]$. The final estimate of ξ is to be made as before.

If the total sample size is large (it can then be shown that the best value of n_1 will usually be large) and if the cost of making the sequence of decisions is negligible relative to the other costs of computations then this procedure should be nearly optimal, in the sense that it minimized the expected loss (risk) for fixed total cost of computation.

V. Some Examples

Some examples have been worked out in order to determine the optimal balance of n_1 and n_2. Of course, the examples are not themselves of direct interest but one would hope that they shed some light upon the best choice of n_1 and n_2 in other more difficult problems, where $g(x)$ and $f(x,\theta)$ are very similar but one cannot evaluate the required expressions. This is the most one can expect of analytic examples in Monte Carlo problems; that is, that they be suggestive of how to proceed in the more difficult cases encountered in practice.

The examples are the following: As before let

$$f(x,\theta) = \theta\, e^{-\theta x}$$

$$g(x) = x$$

so that $\xi = \int_{-\infty}^{\infty} g(x)\, f(x,\theta)\, dx = 1/\theta.$

We will examine two different choices of $h(x)$: $h(x,\alpha) = \alpha\, e^{-\alpha x}$ and $h(x,\alpha) = \alpha^2\, x\, e^{-\alpha x}$. In the case of the first of these we find that

$$\left[V(\alpha_0) + \frac{E_f\left[\psi^2(x,\alpha_0)\right]}{2n_1 \frac{\partial^2 V(\alpha_0)}{\partial \alpha^2}}\right] = \left[.185 + \frac{.356}{n_1}\right]$$

and $\alpha_0 = \theta/2$. If we let $k_1 = c_2/c_1$ and let $(c-c_0) = k_2$ then

$$R\left[n_1, n_2(n_1)\right] = \frac{1}{\theta^2}\left[\frac{\left[.185 + \frac{.356}{n_1}\right]}{.185\, n_1 + .356 + \frac{k_2 - n_1}{k_1}}\right].$$

From this we find optimal n_1's, as a function of k_1 and k_2. These

IV. An Asymptotic Minimum Expected Loss, Sequential Procedure for Estimating ξ

It is completely unlikely that optimization of two stage sampling schemes could, or should, be carried out along the lines suggested above for the reasons just mentioned. In addition, to assume that one can evaluate $R(n_1, n_2)$ implies in practical situations that one can directly compute ξ, in which case there would be no problem, or at least no Monte Carlo problem. These difficulties suggest using knowledge obtained from the initial sample observations from $f(x,\theta)$ to decide when to change over to sampling from $h(x,\hat{\alpha}_0)$, as well as determining $\hat{\alpha}_0$. A sequential scheme for accomplishing this is described below. One can think of sampling procedures that are sequential in different ways; the one discussed here is a sampling procedure sequential for the determination of the best point in the sampling for the changeover from $f(x,\theta)$ to $h(x,\hat{\alpha}_0)$. The procedure is very much like Wald's suggestion for sequential point estimation [3].

The procedure would be as follows: Choose an initial sample of size n_0, where n_0 is large enough for the asymptotic approximations involved to be reasonably accurate, and choose the value of n, the incremental sample sizes. Take a sample of size n_0 from $f(x,\theta)$ and form

$$R\left[n_0, n_2(n_0)\right] - R\left[n_0 + n, n_2(n_0 + n)\right] = \Delta(n_0) ,$$

where

$$n_2(n') = \frac{C - C_0 - C_1 n'}{C_2}$$

using sample estimates of σ_1^2, $V(\alpha_0)$, $E_f\left[\psi^2(x,\alpha_0)\right]$, and $\frac{\partial^2 V(\alpha_0)}{\partial \alpha^2}$

to evaluate these expressions. The sample estimates would be

Bibliography

[1] H. Kahn and A. W. Marshall, "Methods of Reducing Sample Size in Monte Carlo Computations," Journal of the Operations Research Society of America, Vol. 1, No. 5, pp. 263-278.

[2] H. Cramér, Mathematical Methods of Statistics, 1946, pp. 500-504.

[3] A. Wald, "Asymptotic Minimax Solutions of Sequential Point Estimation Problems," Second Berkeley Symposium on Probability and Mathematical Statistics, 1951, pp. 1-11.

BLANK PAGE

so that n_1 does not depend to any great extent upon k_1. Table 2 gives the values of $\theta^2 R(n_1, n_2)$ for optimal choices of n_1 and n_2.

Table 2 ($k_2 = 1000$)

Optimal Values of n_1, n_2, given k_1	k_1	Value of $\theta^2 R(n_1,n_2)$ for optimal choice of n_1,n_2
$n_1 = 501$, $n_2 = 499$	1	.000004
501 250	2	.000008
502 150	3	.000012
502 112	4	.000016
503 89	5	.000020

These again are to be compared with the simple random sampling value for $\theta^2 R(1000, 0)$ of .0010. For no reasonable values of k_1 is simple random sampling better than a two staged sampling procedure (if optimal).

dw

This amounts to ignoring terms for the sampling variation in the weights and the complicated covariance terms involving the variables $\hat{\xi}_1$, $\hat{\sigma}^2_1$, $\hat{\xi}_2$, and $\hat{\sigma}^2_2$. The average variance of $\hat{\xi}_2$ is approximated by taking the expected value of the quadratic approximation of $V\left[\alpha_0(n_1)\right]$ relative to the asymptotic distribution of $\hat{\alpha}_0(n_1)$, which must introduce another approximation error for finite n_1, thus we have

$$V(\hat{\alpha}_0) \approx V(\alpha_0) + \frac{\partial V(\alpha_0)}{\partial \alpha}(\hat{\alpha}_0 - \alpha_0) + \frac{1}{2}\frac{\partial^2 V(\alpha_0)}{\partial^2 \alpha}(\hat{\alpha}_0 - \alpha_0)^2$$

and

$$\int_{-\infty}^{\infty} V(\hat{\alpha}_0)\, P\left[\hat{\alpha}_0(n_1)\right] d\hat{\alpha}_0(n_1) \approx V(\alpha_0) + \frac{1}{2}\frac{\partial^2 V(\alpha_0)}{\partial^2 \alpha} V\left[\hat{\alpha}_0(n_1)\right]$$

since $\frac{\partial V(\alpha_0)}{\partial \alpha} = 0$. In evaluating the weights w_1^2 and w_2^2, σ_2^2 must also be replaced by the approximation expression given above.

All one can suggest is that this is a reasonable thing to look at when deciding upon sampling designs. If possible, one would choose n_1 and n_2 such that $E\left[L(n_1, n_2)\right] = R(n_1, n_2)$ is minimized subject to the constraint $(c_0 + c_1 n_1 + c_2 n_2) = c$. Care should be taken in any case that the n_1 is large enough for the asymptotic variance to be a reasonable approximation.

values are given in Table 1.

Table 1 ($k_2 = 1000$)

Optimal Values of n_1, n_2, given k_1	k_1	Value of $\sigma^2 R(n_1,n_2)$ for optimal choice of n_1,n_2
$n_1 = 47$, $n_2 = 953$	1	.0002
53 474	2	.0004
64 312	3	.0006
85 229	4	.0008
158 128	5	.0010

These are to be compared with $\sigma^2 R(1000, 0) = .0010$, the value for simple random sampling. We see that for $k_1 > 5$ that the two staged sampling, even if optimal, would be worse than simple random sampling because of the high relative cost of sampling from $h(x,\alpha)$.

In the case of $h(x,\alpha) = \alpha^2 \times e^{-\alpha x}$ we find that since $\alpha_0 = \theta$

$$\left[V(\alpha_0) + \frac{E_f\left[\psi^2(x,\alpha_0)\right]}{2n_1 \frac{\partial^2 V(\alpha_0)}{\partial \alpha^2}}\right] = \left[\frac{1}{n_1}\right].$$

Defining k_1 and k_2 as before; then

$$R\left[n_1, n_2(n_1)\right] = \frac{1}{\sigma^2}\left[\frac{k_1}{n_1\left[k_1 + k_2 - n_1\right]}\right].$$

If $k_2 = 1000$ we find that the optimal n_1, and n_2 as functions of k_1 are such that

$$n_1 = \frac{1000 + k_1}{2} \approx 500$$

$$n_2 = \frac{500}{k_1} - 1/2 \approx \frac{500}{k_1}$$