MINIMUM VARIANCE MONTE CARLO IMPORTANCE SAMPLING WITH PARAMETRIC DEPENDENCE

12/18/2013

INTRODUCTION

We consider the approach to Monte Carlo importance sampling calculations proposed by Ragheb, Halton and Maynard. It avoids the occurrence of infinite variances and effective biases which arise in calculations where importance sampling is used with a moderate number of simulations.

The overall functional dependence of the second moment about the origin, which has the same behavior as the variance, is obtained over a set of simulations by proper weighting of the generated histories. Then the results based upon samples corresponding to the importance sampling parameter leading to minimum variance are adopted, and all other results are rejected. Further simulations after determining that optimum parameter will lead to further variance reduction proportional to the reciprocal of the square root of the number of simulations.

The approach corresponds to the observation that once we gain knowledge of the importance or the right optimal value of the biasing parameter, then we have solved the problem and need not proceed further in the sampling process. This means that the optimal value of the parameter and the solution, together with its variance will be obtained simultaneously at a certain point of the computation. Carrying out the calculation further will decrease the variance:

$$\frac{\sigma_{opt}^2}{N}$$

by the 1/N factor, but will not affect the factor σ_{opt}^2 any further. The value σ_{opt}^2 is assumed to have been obtained after the occurrence of statistical convergence, which results from an adequate sampling of the considered probability space.

THEORY

Consider the estimation of the integral:

$$t = \int_x g(x) f(x) dx \tag{1}$$

with variance:

$$\text{var}\, t = \int_x [g(x) - t]^2 f(x) dx \tag{2}$$

where: f(x) is a probability density function, and the variance is defined for g(x) averaged over f(x).

For an importance sampling process depending on a parameter α, with a probability density function h(x,α), we consider the integral:

$$t' = \int_x \frac{g(x) f(x)}{h(x,\alpha)} h(x,\alpha) dx \qquad (3)$$

where x represents sample points obtained from the importance function h(x,α) with parameter α.

Even though t' does not depend upon the choice of α it does in a statistical evaluation effectively depends on it. For a given number of trials, different values of the parameter α result in estimates with different variances and convergence characteristics related to the uniformity of sampling. These lead to the observed infinite variances and effective biases. One is then intesrested in knowing the values of α which leads to a minimum variance in a given calculation. Since the second moment about the origin behaves in the same way as the variance, one can simplify the analysis by considering it, rather than the variance:

$$M_2(\alpha) = \int_x \frac{g^2(x) f^2(x)}{h^2(x,\alpha)} h(x,\alpha) dx \qquad (4)$$

which now clearly depends on α, even if it were evaluated analytically instead of stochastically.

We may compute approximately by Monte Carlo $M_2(\alpha)$ for different choices of the parameter α:

$$\alpha_1, \alpha_2, \alpha_3, \ldots \ \alpha_n,$$

by generating a sample of x's for each α_i, and then draw a curve through these points $M_2'(\alpha_i)$. However, according to a theory by Frolov and Chentsov, such an approach is inconvenient since it will not result in a smooth curve. This is so because the statistical variation from each sample of α_i to another does not allow the points $M_2'(\alpha_i)$ to lie on a smooth curve, even if $M_2(\alpha)$ is smooth and the factor:

$$\frac{g^2(x) f^2(x)}{h^2(x,\alpha)},$$

depends smoothly on α.

We propose the following: Let us use a weighting factor, which amounts to using conditioned samples, and rewrite Eqn. 4 in a working form:

$$M_2(\alpha) = \int_x \frac{g^2(x) f^2(x)}{h^2(x,\alpha)} W(x,\alpha_k) h(x,\alpha) dx$$

or:

$$M_2(\alpha) = \int_x \frac{g^2(x) f^2(x)}{h(x,\alpha) h(x,\alpha_k)} h(x,\alpha_k) dx \qquad (5)$$

where:

$$W(x,\alpha_k) = \frac{h(x,\alpha_k)}{h(x,\alpha_k)} \equiv 1,$$

is a weighting factor, and α_k for k = 1, 2, …, r, are choices for the parameter α used to sample the importance function $h(x, \alpha_k)$.

We can capture the α-dependence of $M_2(\alpha)$ by the estimates:

$$\bar{M}_{2,\alpha_k}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \frac{g^2(x_{i,\alpha_k}) f^2(x_{i,\alpha_k})}{h(x_{i,\alpha_k},\alpha) h(x_{i,\alpha_k},\alpha_k)} \qquad (6)$$

for k = 1, 2, … , r.

where: N is the number of histories drawn from $h(x, \alpha_k)$,

x_{i,α_k} are sample values chosen from the probability density function h with parameter α_k. Each sampled history or pseudo random number will generate r values of x corresponding to each α_k.

Even though there exists theoretically one single graph for $M_2(\alpha)$, in practice, however, different choices of α_k do in fact lead to different $M_2(\alpha)$'s graphs for moderate number of histories because of the incomplete sampling of the probability space for certain α's. Equation 6 thus generates a surface relating M_2 to α for different choices of α_k. After a sufficient number of trials, one can infer from that generated surface the position of the minimum, and consequently the optimal value of α The result of the mean value corresponding to that optimum value of α

$$\bar{t}' = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_{i,\alpha_{opt}}) f(x_{i,\alpha_{opt}})}{h(x_{i,\alpha_{opt}},\alpha_{opt})} \qquad (7)$$

is chosen as our answer, together with its variance:

$$\text{var}\,\bar{t}' = \bar{M}_{2,\alpha_{opt}}(\alpha_{opt}) - (\bar{t}')^2 \qquad (8)$$

Other values of t' corresponding to non-optimal α's are to be rejected, since they would involve an effective bias, unless an unrealistic number of trials is used.

More formally, for a random variable ξ depending on a parameter α, and using the Radon-Nikodym derivative:

$$\frac{d\nu}{d\mu(\alpha)},$$

in:

$$\xi(\alpha)d\mu(\alpha) = \eta d\nu\ ,$$

for an importance sampling process; the working formula for the approach as Eqn. 5 reads:

$$M_2(\alpha) = \int_x \xi^2(x,\alpha)d\mu(x,\alpha)$$

or:

$$M_2(\alpha) = \int_x \xi(x,\alpha)\xi(x,\alpha_k)d\mu(x,\alpha_k) \qquad (5')$$

if:

$$\xi(x,\alpha)d\mu(x,\alpha) = \xi(x,\alpha_k)d\mu(x,\alpha_k)$$

for all x's and k = 1, 2, 3, … ,r.

For a certain choice α_0 of α_k, this specializes to:

$$M_2(\alpha) = \int_x \xi(x,\alpha)\xi(x,\alpha_0)d\mu(x,\alpha_0) \qquad (9)$$

an equation used by Spanier for generating a set of histories from sampling with a parameter value α_0, to estimate $M_2(\alpha)$ for $\alpha \neq \alpha_0$, over a multistage sampling process.

COMPUTATIONAL SCHEMES

Sequential decision-making is necessary for the application of the methodology. This does not imply a multistage process, and a single stage procedure is maintained to preserve the correlation among the samples obtained. The sequential procedure produces a decision on contracting the range of the parameter values as we gain more information about the position of the minimum, and possibly refining the mesh around the optimum as the calculation proceeds. Figure 1 suggests two computational schemes for the application of the method.

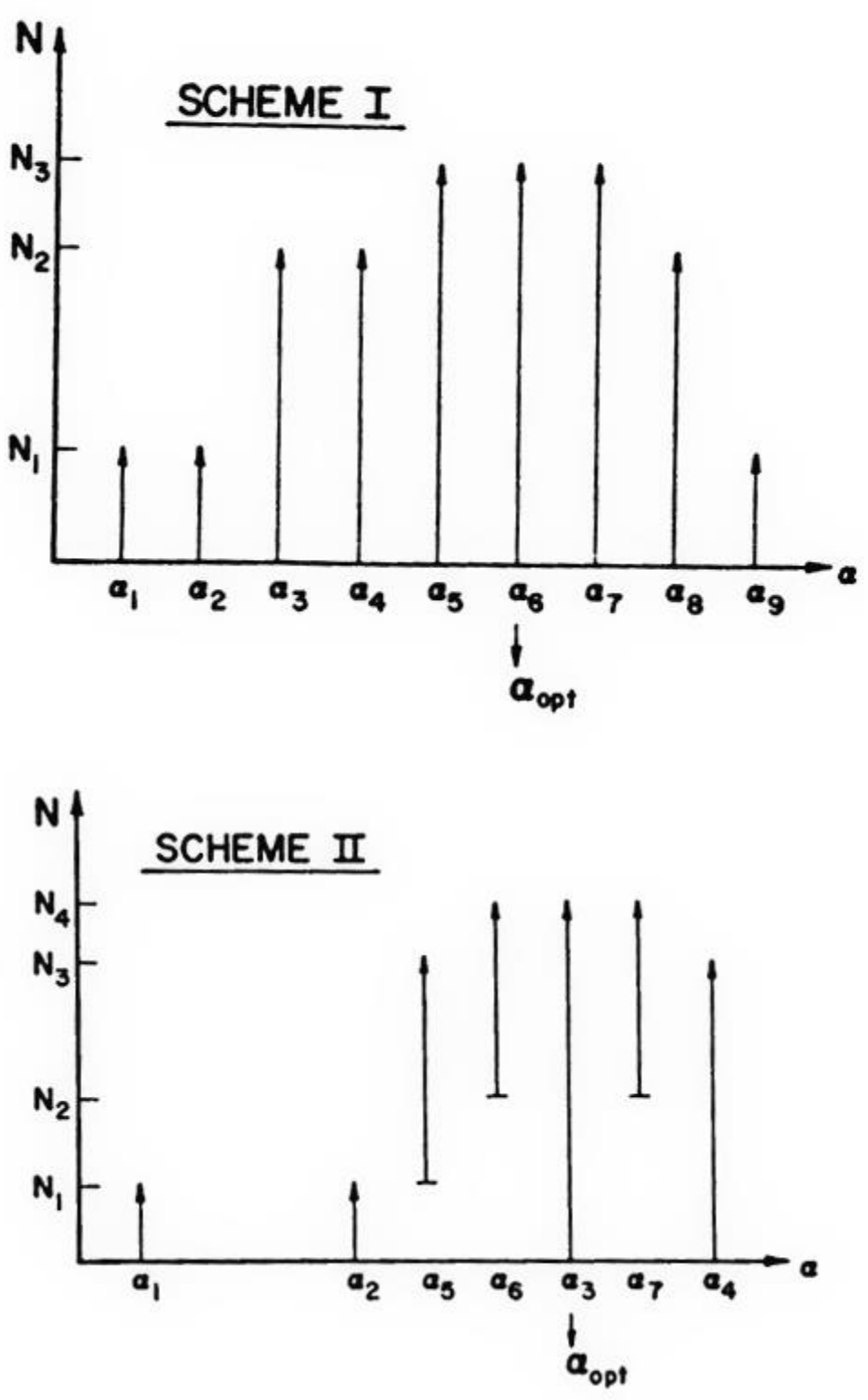

Figure 1. Computational Schemes for optimization of biasing parameter α.

APPLICATION TO THE EVALUATION OF INTEGRALS

The evaluation of integrals allows insight into the method and its application, economical computations, and the results and conclusions generated apply to other applications, such as particle transport.

We consider both the Crude Monte Carlo method as exposed by Hammerseley and Handscomb, and importance sampling with parametric dependence.

We sample $x_1, x_2, \ldots, x_n$ independent uniformly distributed pseudo-random numbers over the interval [0,1] to estimate the integral:

$$M_1 = \int_0^1 g(x) f(x) dx \tag{10}$$

where f(x) is the uniform distribution: $f(x_i) \equiv 1$.

The unbiased Crude Monte Carlo estimator of M_1 is:

$$\bar{M}_1 = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \tag{11}$$

Its variance is:

$$\frac{\sigma^2}{n} = \frac{1}{n} \int_0^1 [g(x) - M_1]^2 dx \tag{12}$$

and its theoretical standard error is thus:

$$\sigma_{\bar{M}_1} = \frac{\sigma}{\sqrt{n}} \tag{13}$$

If we consider:

$$g(x) = e^x \tag{14}$$

we get:

$$M_1 = e - 1 = 1.718281828$$
$$M_2 = \frac{e^2 - 1}{2} = 3.194528048$$
$$\sigma^2 = M_2 - M_1^2 = 0.242035608,$$

The standard error is estimated by the formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}[g(x_i) - \bar{M}_1]^2 \tag{15}$$

giving an estimate s for σ. The result is announced in the form:

$$M_1 = \bar{M}_1 \pm \frac{s}{\sqrt{n}} \tag{16}$$

Since, by the Central Limit Theorem, we expect the distribution of $\bar{M}_1$ to be approximately normal, we may say with 95 percent confidence that we are within 2 standard deviations of the mean.

Using the importance function:

$$Ce^{\alpha x},$$

the normalization:

$$C\int_0^1 e^{\alpha x}dx = 1, \Rightarrow C = \frac{\alpha}{e^{\alpha}-1}.$$

Thus, using Eqn. 3, with the normalized density function:

$$h(x,\alpha) = \frac{\alpha}{e^{\alpha}-1}e^{\alpha x}$$

and:

$$g(x) = e^x,\ f(x) \equiv 1,$$

we get:

$$M_1 = \int_0^1 \frac{e^x}{\frac{\alpha}{e^{\alpha}-1}e^{\alpha x}}[\frac{\alpha}{e^{\alpha}-1}e^{\alpha x}]\mathrm{dx} = e-1 \tag{17}$$

and:

$$M_2(\alpha) = \int_0^1 \frac{e^{2x}}{\frac{\alpha^2}{(e^{\alpha}-1)^2}e^{2\alpha x}}[\frac{\alpha}{e^{\alpha}-1}e^{\alpha x}]\mathrm{dx} = \frac{(e^{\alpha}-1)(e^{2-\alpha}-1)}{\alpha(2-\alpha)} \tag{18}$$

The minimum of $M_2(\alpha)$ occurs at α = 1, and is then equal to:

$$M_2(\alpha = 1) = (e-1)^2 = 2.952492440$$

and the variance at that optimal value of α = 1, will lead to a zero variance scheme.

We assume, as is the case in practical applications, that we have no prior knowledge about the optimal value of α, so that we seek to determine it from the computational results. The approach is applied to this model problem, and compared to Crude Monte Carlo.

COMPUTATIONAL RESULTS

Over the unit interval, the normalized probability density function:

$$pdf = h(x, \alpha_k) = \frac{\alpha_k}{e^{\alpha_k} - 1} e^{\alpha_k x}$$

has a cumulative distribution function:

$$cdf = F(x, \alpha_k) = \int_0^1 \frac{\alpha_k}{e^{\alpha_k} - 1} e^{\alpha_k x} dx = \frac{e^{\alpha_k x} - 1}{e^{\alpha_k} - 1}$$

For pseudo random numbers ρ_i uniformly distributed over the unit interval, a sample from the considered pdf will be:

$$x_{i,\alpha_k} = \frac{1}{\alpha_k} \ln[\rho_i (e^{\alpha_k} - 1) + 1].$$

Thus, for given values of α_k, k = 1, 2., … , r, one can obtain r sample points x_{i,α_k} for each generated pseudo random number ρ_i. For each α_k the dependence of $M_2(\alpha)$ upon α can be then be obtained by using Eqn. 6 for $\bar{M}_{2,\alpha_k}(\alpha)$.

Table 1 shows the behavior of the second moment when sampling from the importance function with parameters α_k as a function of α, for discrete values of α : α_i. For different numbers of histories N, the position of the minimum is underlined. We notice that for a small number of histories, the minima do not all occur at the position of the theoretical minimum (α = 1). For N = 10, and α_k = 2.0, 4.0 it occurs at α = 0.5. When the number of histories is increased to 30, 40 and 50, the whole solution stabilizes and the minima for different choices of α_k occur at α = 1.

Now that we are sure that the solution has converged statistically, we can announce our results as the first moment obtained for α_k = 1 and α = 1. The variance was estimated as:

$$\frac{n-1}{n} s^2 = (\bar{M}_2 - \bar{M}_1^2) = 1.0 \times 10^{-7}$$

and our answer can be announced as:

$$\bar{M}_1 \pm \frac{s}{\sqrt{50}} = 1.7182817 \pm 4.5\times10^{-5},$$

This can be compared with the exact solution of $M_1 = 1.7182818$.

Table 1. Second moment as a function of biasing parameter α and number of histories. α ε [0.25,4.00]

α_k	α_1 = 0.25	0.50	1.00	2.00	4.00	N
0.25	3.3512876	3.2049569	3.0391791	3.1843314	6.2900656	10
	3.4226299	3.2695809	3.0852759	3.1615611	5.7180804	20
	3.3694730	3.2303055	3.0671501	3.1645162	5.6710798	30
	3.3480686	3.2163517	3.0626664	3.1637914	5.6298111	40
	3.2448789	3.1362457	3.0204790	3.1749161	5.6630727	50
0.50	3.2559052	3.1261681	2.9905398	3.1957246	6.5441389	10
	3.3272808	3.1874427	3.0063235	3.1435372	5.8185010	20
	3.2926763	3.1635460	3.0177194	3.1461850	5.7436520	30
	3.2823820	3.1578029	3.0170411	3.1430902	5.6887521	40
	3.2067142	3.1018391	2.9929388	3.1609894	5.6991751	50
1.00	3.1295846	3.0302866	2.9524924†	3.2845908	7.2004070	10
	3.1874242	3.0712709	2.9524923	3.1479645	6.0652575	20
	3.1768045	3.0657614	2.9524921	3.1415549	5.9159886	30
	3.1771167	3.0667762	2.9524920	3.1320937	5.8679066	40
	3.1437968	3.0467231	2.9524921	3.1506930	5.7919430	50
2.00	3.1099843	3.0679785	3.1085874	3.7390714	9.0952197	10
	3.0771947	2.9984964	2.9491732	3.2842149	6.6686219	20
	3.0754017	2.9935579	2.9335215	3.2223181	6.2535771	30
	3.0738270	2.9903632	2.9271677	3.2119834	6.2775736	40
	3.0774526	2.9956231	2.9299994	3.1838111	5.9545728	50
4.00	3.8899635	3.9536145	4.2312592	5.5161900	13.9203260	10
	3.2736266	3.2400379	3.2826001	3.8400135	8.1914040	20
	3.1506004	3.0930301	3.0758384	3.4395828	6.6343833	30
	3.1444539	3.0862379	3.0668895	3.4186276	6.4665629	40
	3.0811210	3.0087624	2.9551107	3.2025470	5.7015834	50

† $M_2(\alpha=1)_{exact} = 2.952492440$

For comparative purposes, the results for the first moment using Crude Monte Carlo calculations are shown in Fig. 2. The error bars always contain the exact solution since no biasing is introduced. At n = 50, the answer was still:

$$\bar{M}_1 \pm \frac{s}{\sqrt{50}} = 1.773606 \pm 0.065788;$$

and at N = 10,000 samples it is:

$$\bar{M}_1 \pm \frac{s}{\sqrt{10,000}} = 1.720281 \pm 0.004908.$$

Both results are inferior to the importance sampling result for 50 samples. Obviously, the proposed approach involves more computational labor for each history. Assuming that the sampling of each point requires 10 times more computational effort,

and that we choose a very fine mesh of 20 x 20 = 400 points, this will lead to 4,000 times more computational effort per sample than Crude Monte Carlo. However, the theoretical standard error of the Crude Monte Carlo scheme is:

$$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.242035}}{\sqrt{n}}.$$

If the approach gives for n = 50 a standard error of 4.5 x 10^{-5}, this would have required the Crude Monte Carlo0 sampling to have used a number of histories n = 1.2 x 10^{8} histories to achieve the same level of error. Even reducing that by 4,000 times to accommodate for the extra sampling operations, the gain in efficiency will still be:

$$\frac{1.2\times10^{8}}{4,000} \simeq 3.0\times10^{4}$$

which is a considerable gain. Obviously such a large gain may not be achievable in more practical applications in which the parametric representation of the importance sampling function, together with its sampling will require much extra effort; but a gain is certainly expected.

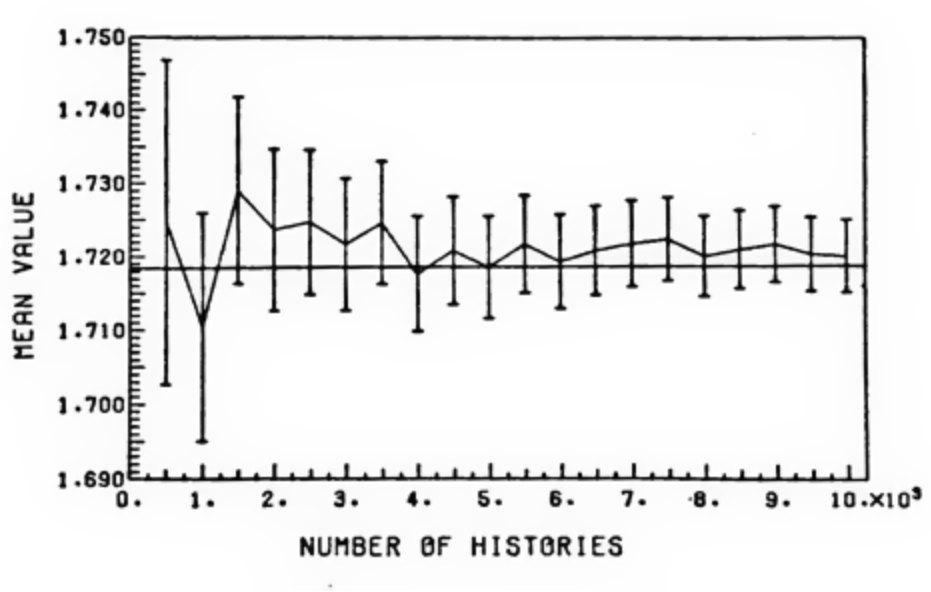

Figure 2. Crude Monte Carlo results for n = 10,000, showing error bars and exact value.

The behavior of the first moment for different α_k ε [0.4375, 4.0] is displayed in three dimensional forms for n = 100, 10,000 and 100, 000 in Figs. 3-5. Away from the optimum, an effective bias in the result can be noticed unless a large number of histories is used.

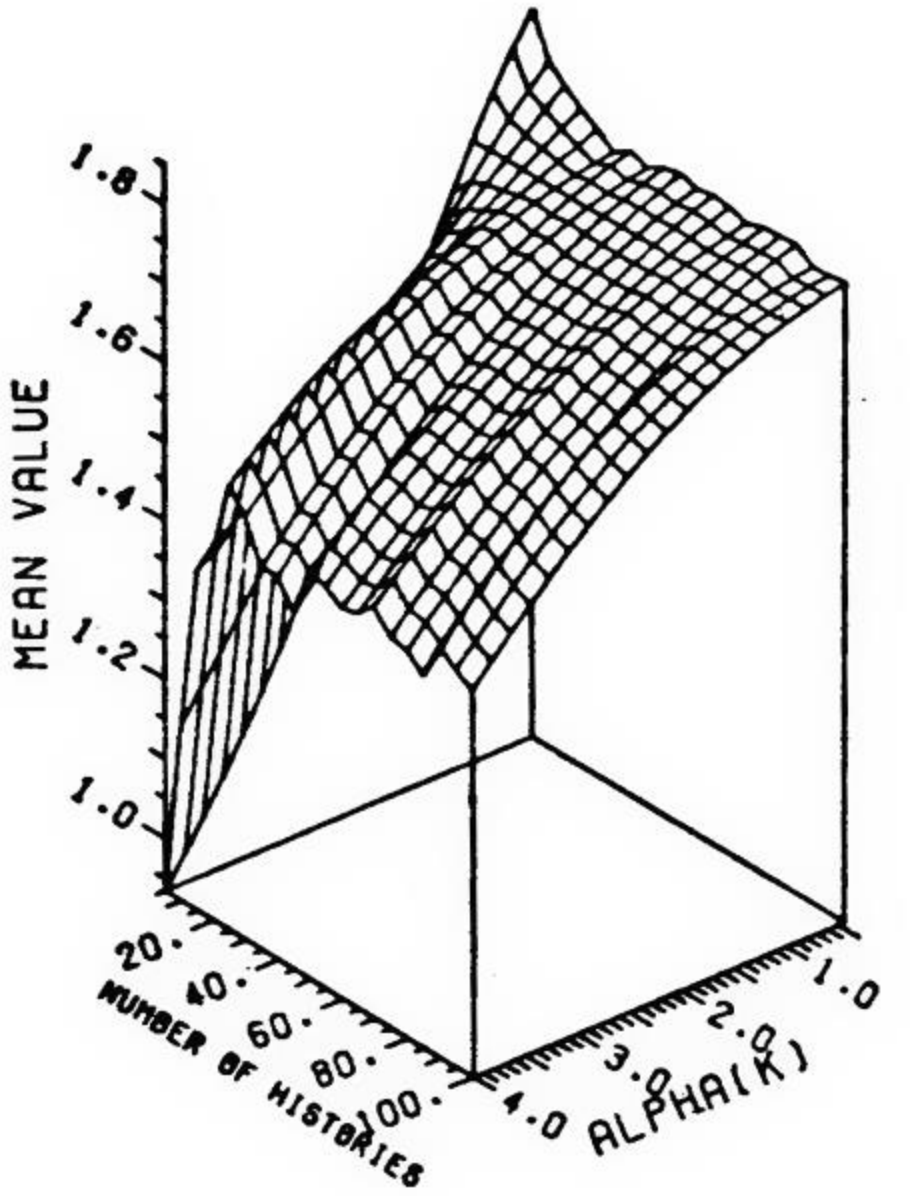

Figure 3. Mean value as a function of the biasing parameter α, for N = 100 histories.

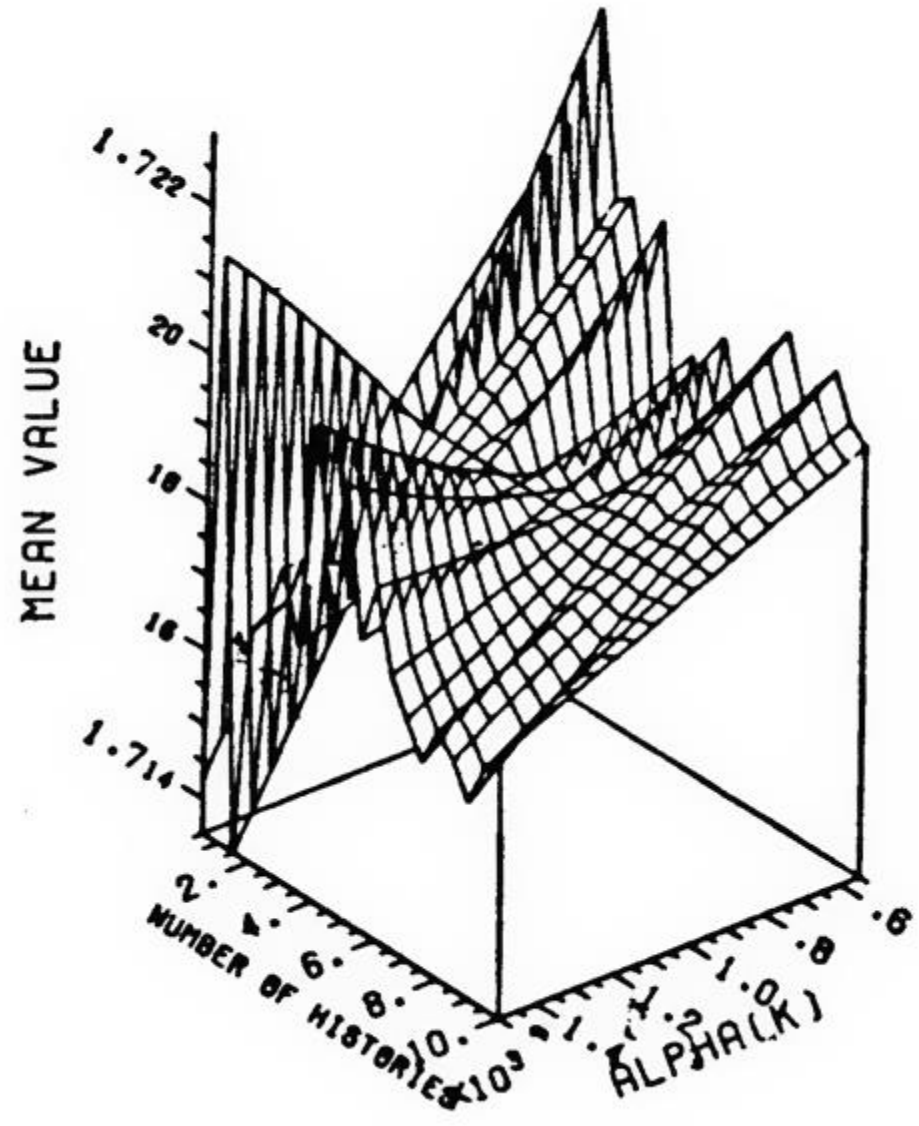

Figure 4. Mean value as a function of the biasing parameter α, for N=10,000 histories.

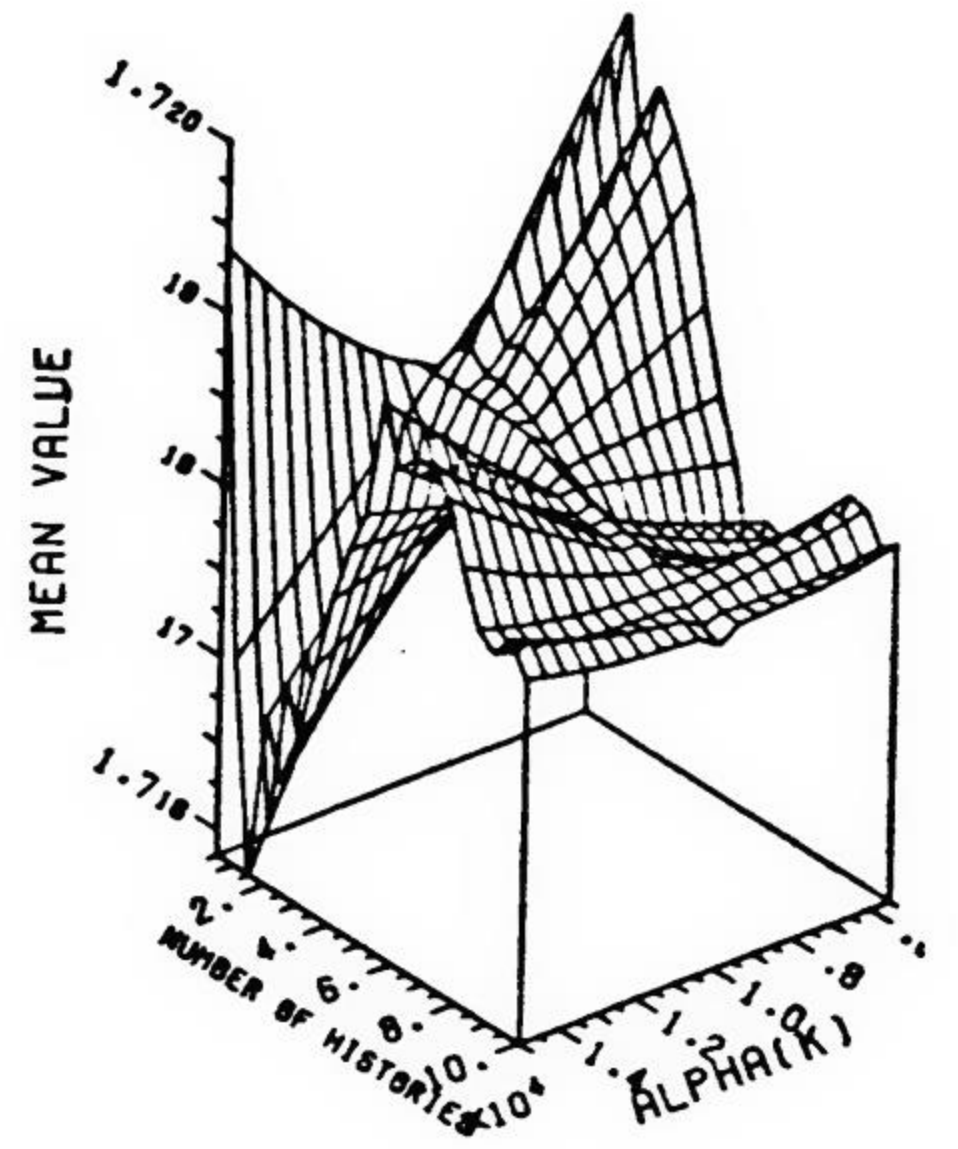

Figure 5. Mean value as a function of the biasing parameter α, for N = 100,000 histories.

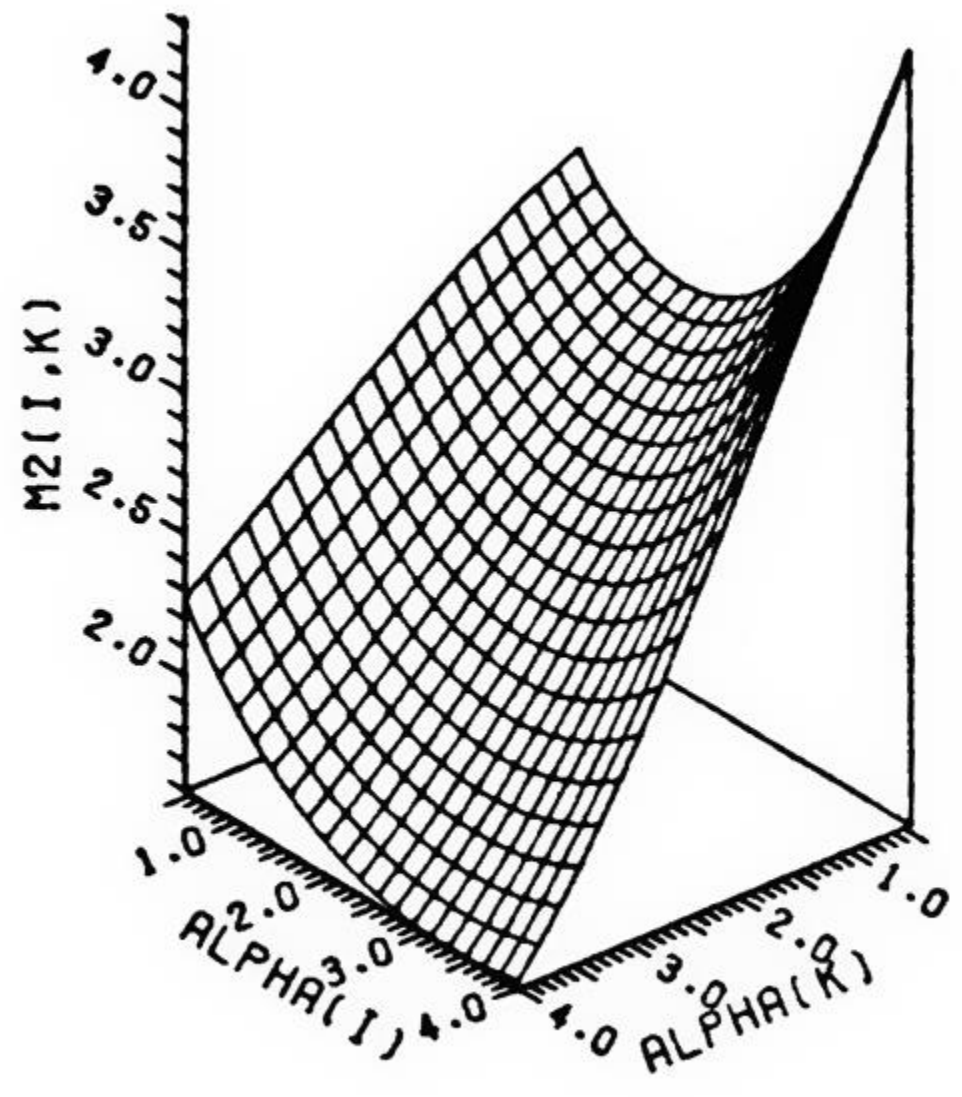

Figure 6. Second moment surface as a function of the biasing parameter α, for N = 10 histories.

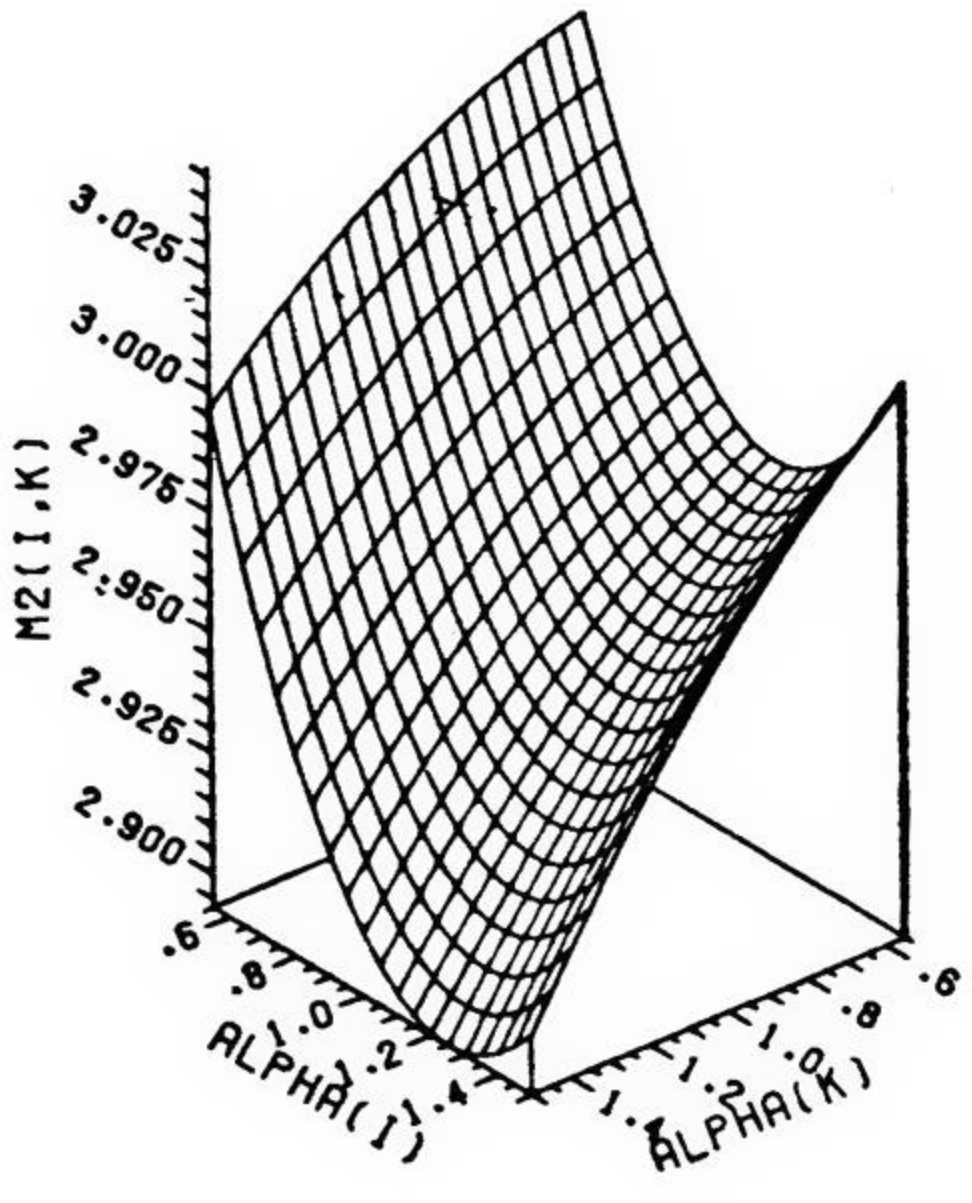

Figure 7. Second moment surface as a function of the biasing parameter α, for N = 100 histories.

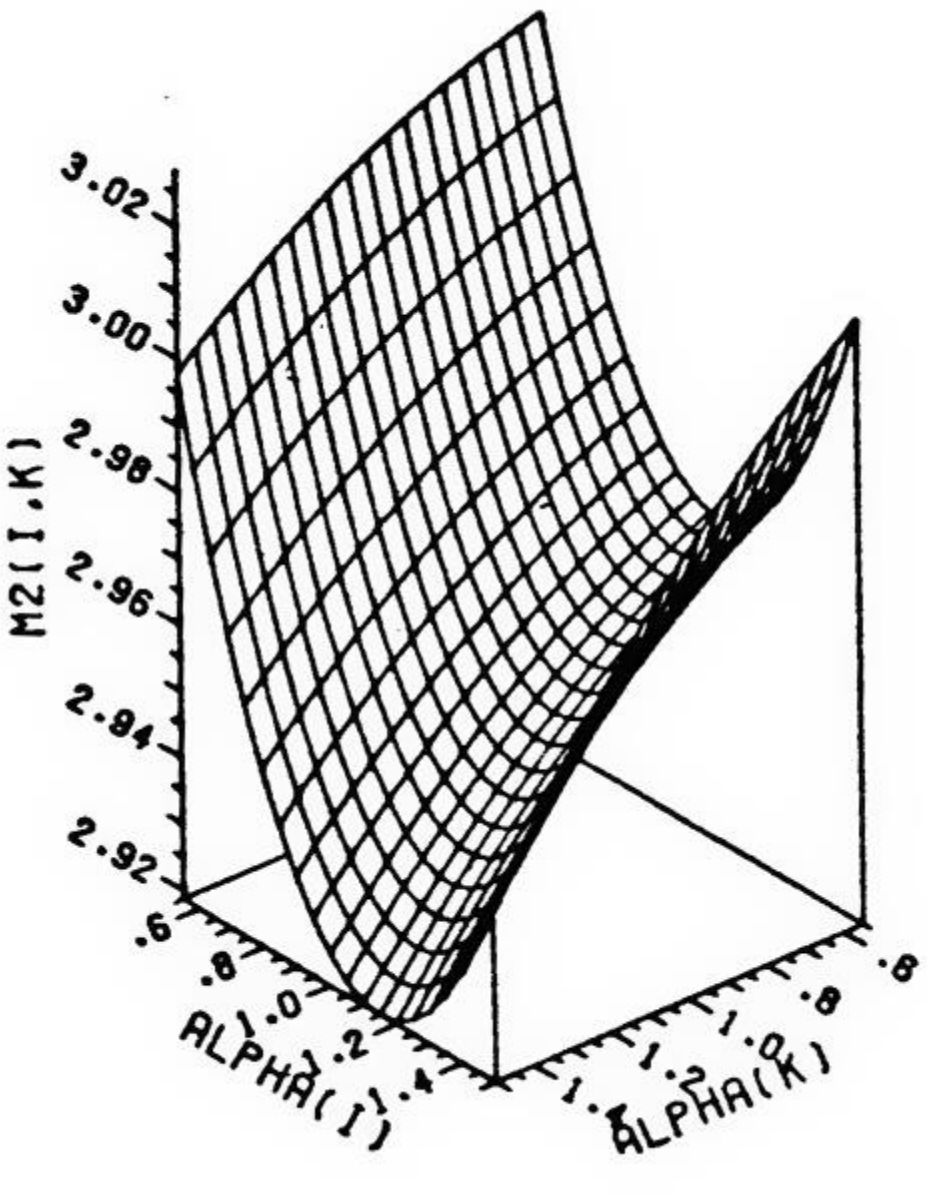

Figure 8. Second moment surface as a function of the biasing parameter α, for N = 300 histories.

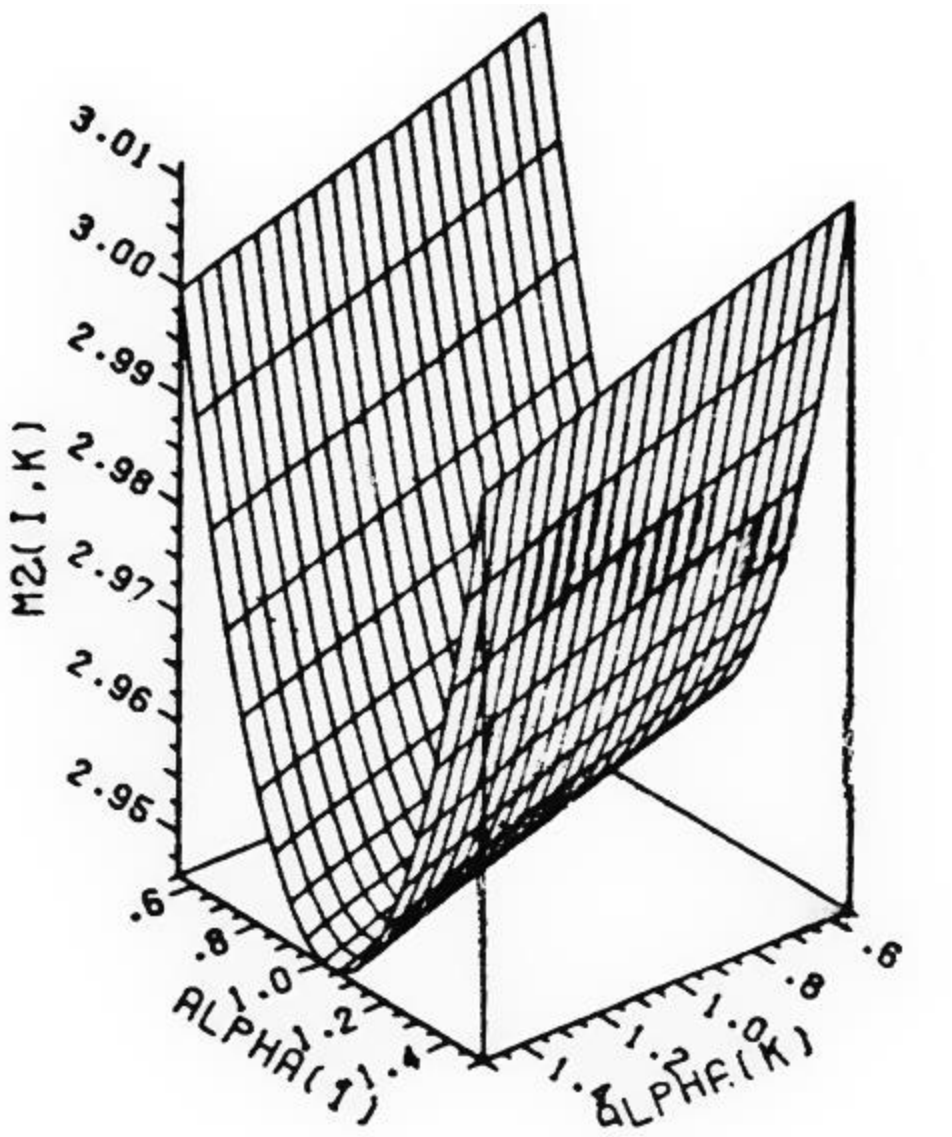

Figure 9. Second moment surface as a function of the biasing parameter α, for N = 600 histories.

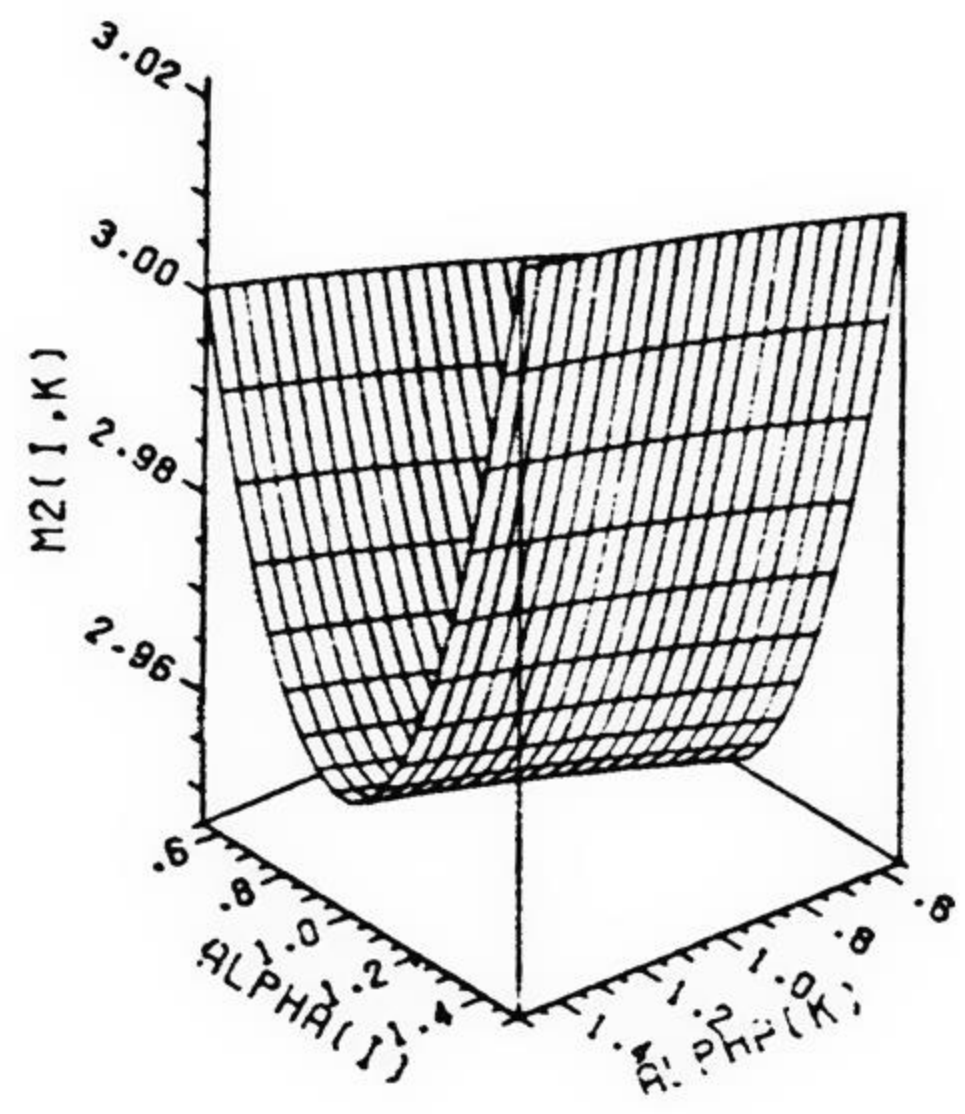

Figure 10. Second moment surface as a function of the biasing parameter α, for N = 1,000 histories.

Since our measure of the "variation" of the first moment for determining our optimal parameter value is the second moment, its behavior is investigated further in Figs. 6-10. These figures show clearly how the second moment surface fluctuates around its optimal point $\alpha_k = 1.0$ as a pivot, and how it approaches the exact smooth curve $M_2(\alpha)$ as n grows larger. Infinite variances will be obtained if we are not close to the optimum.

DISCUSSION

Minimum variance in Monte Carlo importance sampling with parametric dependence can be achieved through the generation of the second moment surface in view of identifying the optimal variance parameter. The results based upon samples corresponding to the importance sampling parameter leading to minimum variance are adopted, and all other results are rejected. Further simulations after determining that optimum parameter will lead to further variance reduction proportional to the reciprocal of the square root of the number of simulations. The approach is shown to avoid the infinite variances and effective biases that can arise when importance sampling is used. Whereas the methodology has been demonstrated for the estimation of integrals, greater usefulness of the methodology should be more drastic upon usage in particle and fluid transport applications.

The determination of the position of the optimum can be carried out by table lookout or by some steepest descent or alternating gradient method, or by genetic algorithms or simulated annealing optimization methods over the second moment surface.