USE OF DIFFERENT MONTE CARLO SAMPLING TECHNIQUES

Herman Kahn

P-766

30 November 1955

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

Special Foreward to 1955 Readers

If you have any comments on form, format, or facts and if you pass them on to me in the year 1955, I will be grateful as I can probably incorporate them in the published version. If you have any in 1956 or later, keep them unless they are earthshaking.

I should add that I have attempted to be crystal clear and am very anxious to make changes where I have failed.

H. K.

Use of Different Monte Carlo Sampling Techniques [1]

Herman Kahn

The RAND Corporation, Santa Monica, California

When doing a Monte Carlo problem one focuses attention on three main topics. They are:

1. Choosing or analogizing the probability process
2. Generating sample values of the random variables on a given computing machine
3. Designing and using variance reducing techniques

This paper will discuss only the last topic. It is, of course, true that in actual problems one cannot isolate variance reduction from the first two topics. The methods which can be used to reduce variance are often sharply dependent upon the probability model and in some cases on the techniques used to generate values of the random variables. Also, the greatest gains in variance reduction are often made by exploiting specific details of the problem, rather than by routine application of general principles. However, there do seem to be some general ideas on reducing variance which can be used in many problems. In particular, six techniques seem to be most useful. They are:

1. Importance Sampling
2. Russian Roulette & Splitting
3. Use of Expected Values (combination of analytic and probabilistic methods)
4. Correlation and Regression
5. Systematic Sampling
6. Stratified Sampling (Quota Sampling)

While all of these techniques can be used in standard statistical sampling problems, the first three seem to have found particular and specialized usefulness in Monte Carlo applications as differentiated from the usual applications in ordinary sampling. This is mainly due to the fact that in a Monte Carlo problem the experimenter has complete control of his sampling procedure. If for example

---

1. This paper is a revision of the second chapter of a larger report being prepared by the author. I would also like to express my thanks to I. Mann, A. Marshall, and N. Shapiro who read and commented on a first draft. While the author is fundamentally responsible for any mistakes, he would like to point out that under the legal doctrine "of the last clear chance" the foregoing should be sued.

he wanted a green-eyed pig with curly hair and six toes and if this event had a non zero probability, then the Monte Carlo experimenter, unlike the agriculturist, could immediately produce the animal.

In order to illustrate the general nature of the techniques, we will first apply them to a very simple example--so simple, in fact, that the reader will have to exercise his imagination in order to pretend there is any problem. After we have gone through the elementary exercise, we will discuss, in some detail, how the six techniques can be applied to the Monte Carlo evaluation of definite integrals. The integral is used for the serious example, not because it is the main application of the technique — it is not (except in the generalized sense that any expected value can be calculated by an integral) — but rather because it is the application in which the ideas are most clearly exposited.

The simple example we will use is the problem of calculating the probability of obtaining a total of three when one tosses two ordinary dice. Each die is of the standard sort with six faces labeled from one to six and constructed so that each face has the same probability (1/6) of being on top.

The problem can, of course, be solved analytically. Any particular combination of the dice has a probability equal to 1/6 times 1/6 of occurring. Since there are two combinations which make three (one-two and two-one), the probability of getting a three in a random toss of the dice is 2/36 or 1/18.

In doing the above problem by Monte Carlo one would simply toss the dice[2] N times, count the number (n) of successes (threes) and then estimate the probability (p) of success by

$$\hat{p} = \frac{n}{N} \qquad (1)$$

Typically, $\hat{p}$ differs from p; that is, the estimate has a statistical error. This statistical error is usually measured[3] by the standard deviation σ, where

2. Usually one would not toss physical dice, but simulate the tosses with the aid of a table of random numbers.

3. In Monte Carlo problems the error has statistical properties which can usually be described in the following manner. The probability that the absolute value of the error will be larger than $m\sigma$ is given by the following table:

m	Prob.
.67	.50
1.00	.32
2.00	.05

σ is called the standard deviation and σ^2 the variance. The latter is usually denoted by V. A more complete explanation of statistical errors can be found in almost any elementary book on statistics.

$$\sigma = \sqrt{\frac{p(1-p)}{N}} \tag{2}$$

The percent error is then given by

$$\frac{100\sigma}{p} = 100\sqrt{\frac{1-p}{Np}} \tag{3}$$

As is well known, and as is shown by the above formula, this error goes down as the number of trials is increased. It is the purpose of Part One to illustrate other ways in which the error can be decreased.

Part One
The Trivial Example

1. Importance Sampling

If by some method we can increase the effective value of p, equation (3) shows that the percent error will be reduced. This increase in the effective value of p can be obtained very easily. We could, for example, bias the dice so that the probability that a one or a two would come up is twice as great as usual, that is 1/3 rather than 1/6. This could be done with physical dice by "loading" them, or by mathematically simulated dice by using a biased table of random numbers. If this is done then the probability of getting a three, instead of being 1/18, is four times as great or 2/9. The percent error is then cut by slightly more than a factor of two. Of course, equation (1) can no longer be used to estimate p, but

$$\hat{p} = \frac{1}{4}\frac{n}{N} \tag{4}$$

must be used instead. The 1/4 in equation (4) is called a weighting factor. By using it, the distortion introduced by the biased sampling is removed.

This illustrates the general idea of Importance Sampling--which is to draw samples from a distribution other than the one suggested by the problem and then to carry along an appropriate weighting factor which, when multiplied into the final results, corrects for having used the wrong distribution. The biasing is done in such a way that the probability of the sample's being drawn from an "interesting"[4] region is increased; the probability that it comes from

---

4. As discussed later, the words "interesting" and "uninteresting" refer to the amount of effort or interest the sensible calculator would show in the region.

an "uninteresting" region is correspondingly decreased. The reader should verify for himself that it would be correct to carry the bias to the limit; that is, the probability of getting a one or a two could be increased by a factor of three, making the probability of obtaining one of these numbers 1/2 and making the probability of obtaining any other number zero.

The above limit is not the ultimate limit. For example, if we tossed the dice one at a time, then we might want to bias the second die differently from the first one. In particular, if we were willing to let the biasing of the second die depend on the outcome of the first throw, we might consider the following scheme.

1. Increase the probability of getting a one or a two on the first die by a factor of three. This means, of course, that there will be a zero probability of getting any other numbers.
2. If the first die comes up one, increase the probability of the second die coming up two by a factor of six; if the first die comes up two, increase the probability of the second die coming up one by a factor of six.

If this scheme is followed every toss of the dice will yield a three so that the number of successes (n) will be equal to the number of trials (N). The weighting factor will be 1/3 times 1/6 or 1/18 and the estimate will be

$$\hat{p} = \frac{n}{18N} \qquad (5)$$

$$= \frac{1}{18}$$

which is exactly equal to p. We have devised a sampling procedure which has zero variance. In principle, though not in practice, it is always possible to design an Importance Sampling scheme that has zero variance.

2. Russian Roulette & Splitting[5]

Let us assume that the dice are tossed one at a time and that the cost

---

5. Both the idea and the name are due to J. von Neumann and S. Ulam.

---

of the problem is measured by the total number of tosses. Now, it is immediately clear that if the first die is tossed and if it happens to come up three or greater, it will be impossible to get a total of three, no matter how the second die comes up. Under these circumstances, there is no point in

making the second toss and we can simply record a zero for the experiment. This makes it unnecessary to toss the second die 2/3 of the time. Therefore on the average we will do 1/3 fewer tosses in an experiment.

Generally in more complicated examples where the sampling is done in stages, it is often possible to examine the sample at each stage and classify it as being in some sense "interesting" or "uninteresting". The sensible calculator is willing to spend more than an average amount of work on the "interesting" ones and contrariwise wants to spend less effort on the "uninteresting" ones. This can be done by splitting the "interesting" samples into independent branches, thus getting more of them, and by killing off some percent (in the above example 100%) of the "uninteresting" ones. The first process is Splitting and the second Russian Roulette.

The "killing off" is done by a supplementary game of chance. If the supplementary game is lost the sample is killed; if it is won the sample is counted with an extra weight to make up for the fact that some other samples have been killed. The game has a certain similarity to the Russian game of chance played with revolvers and foreheads--whence the name.

The idea of Russian Roulette & Splitting is similar in spirit to the sequential sampling schemes of quality control, though quite different in detail. It was first thought of in connection with particle diffusion problems. Particles which get into interesting regions are split into n independent sub-particles, each with one n'th of the weight of the original particle. Particles which get into uninteresting regions are, in effect, amalgamated into a fewer number of heavier particles. In this way the calculator achieves his goal of allocating his effort sensibly.

3. Use of Expected Values

If the sampling is being done in two stages, then even if we aren't clever enough to calculate the combinatorics of the whole problem, we still might be clever enough to notice that there is no point in tossing the second die; that is, once the first die is tossed, it is trivially easy to calculate the probability of obtaining a total of three. For example, when the first die comes up one, the only way we can get the three total is for the second die to come up two. This event obviously has a probability of 1/6. Similarly if the first die comes up two, the only way to get three is for the second die to be one. This event also has a probability of 1/6. Finally, all the other possibilities for the first die (three to six) have a zero probability of

giving three.

If we record the probabilities rather than toss the second die, then it is a fact that the average of these probabilities is an estimate of p. This method of doing the problem simultaneously reduces the number of tosses we need by a factor of two and decreases the variance, so that the tosses we do make are more effective.

The illustration is not artificial. In many probabilistic problems, it turns out that much of the variance or fluctuation is introduced by a part of the probabilistic problem which can be calculated analytically; the probabilistic part which is hard to calculate analytically may, in fact, not introduce much fluctuation. In these cases the sensible calculator combines analytic and probabilistic methods--calculating analytically that which is easy and Monte Carloing that which is hard.

The three techniques discussed above can be fantastically effective in realistic applications. The author is familiar with applications in which each technique has, by itself, decreased the effective variance by factors of the order of 10^4 to 10^6. In most cases this means changing the problem from one which cannot be done because it would be too expensive or lengthy to one which is easily done on modern computing machines or even by hand computers.

The three techniques which we are now going to discuss are, in general, not as effective as the three already mentioned. However, they often are very easy to use and may yield substantial improvements.

4. Correlation and Regression

In order to illustrate this technique, it will be necessary to change the example slightly. Assume, for instance, that the proprietor of one of the gaming establishments in Las Vegas wishes to change the rules in force at his dice tables. Under the current rules, if a player tosses a 2, 3, or 12, on the first throw of the dice, the player loses. If he tosses a 7 or 11, he wins, and if he tosses a 4, 5, 6, 8, 9 or 10, he will win or lose, depending on whether or not that number or a 7 comes up first in his subsequent throws.

Now let the rule change being considered be the interchange of the roles of 3 and 4 and assume that, unlike most of the proprietors in Las Vegas, the one we are considering is unsophisticated and wishes to determine by sampling what the change in his revenue will be. The obvious way to do this is to run two sets of experiments, one with the old rules and one with the new rules, and then compare the two experimentally-determined revenues. Under these

circumstances, one is subtracting two relatively large, fluctuating quantities to determine a small quantity. In general, this yields a process with a large percent error.

There is a better way to do this problem. Instead of running two independent games, the proprietor could run only one game and apply both sets of rules simultaneously to this game. In fact, he can choose to estimate the difference in revenue directly rather than the revenue that would be achieved under each set of rules.

This can be done by playing the following game:

1. Whenever a 3 comes up, continue to toss the dice until either a 3 or a 7 comes up. In the first happenstance, record a minus two, since under the old rules the customer would have lost a dollar, but under the new rules he wins one; in the second happenstance, record a zero because under both sets of rules the customer loses.
2. Follow a similar process if a 4 comes up.
3. If a number other than 3 or 4 comes up terminate the play then and there and record a zero. (Because of this rule, the effects of chance fluctuations in the proportionate number of times that the numbers 2 and 5 to 12 come up are eliminated from the comparison.)

It should be noticed that the specific game that is played is quite different from the two games that are being compared. As usual, this causes a double saving of efficiency; first because only one set of games is played, and second because the number and kinds of chance fluctuations that can affect the results are greatly reduced.

It is in fact generally true that if we wish to compare two or more situations, we can, by combining this comparison into a single problem, reduce the work substantially. Only one problem, rather than several, has to be done, and the direct estimate of the difference can usually be made more accurately than estimates of separate individual quantities.

This is a substantial virtue of the Monte Carlo method. In many complicated problems we are not actually interested in absolute values but only in comparisons. We may wish, for example, to know if Strategy A is better than Strategy B, or if Engineering Design A is better than Engineering Design B. We might, in fact, not even believe the absolute values because the idealizations

are so rough, but do believe the qualitative features implied by differences. Monte Carlo can then be used to estimate the thing which we actually desire to know and which we believe, and we can bypass the estimate of irrelevant quantities. Usually, however, we can obtain these also, but at some extra cost.

Correlated Sampling can also often be used to test the accuracy of an approximate theory. If the approximate theory happens to be an exact treatment of an idealized situation, and if the idealized situation happens to be "structurally" similar to the unidealized situation, then it is often possible to design very efficient sampling schemes to calculate the difference between the idealized and unidealized situations. The answer to the problem posed by the unidealized situation can then be obtained by adding together the results of the approximate analytic calculation and the Monte Carlo difference calculation.

5. Systematic Sampling

If we are doing a multi-stage sampling problem, it often turns out to be very easy to do the first stage systematically. For example, in our problem, if we are going to toss the dice one at a time then there is really no point in actually tossing the first die. If, for example, we were planning on getting 600 samples, we would expect on the average that each die would come up one about 100 times, two another 100, and so on. It is easy to show that we do not bias the results if we assume that the first 100 tosses of the first die actually do come up one, the second 100 tosses of this die come up two, etc. and so only toss the second die. The main advantage in doing this is that we have eliminated the error caused by the fluctuation in the proportions of ones, twos, etc. which would result if the first toss was random.

In practice, however, doing the first stage of the sampling systematically does not usually lead to substantial improvements in efficiency. Generally, in fact, it will only reduce the number of samples required by a relatively few per cent — say 5 to 30. However, it ordinarily does not cost anything to apply this technique, so that there is no point in not using it. About the only time we may not be able to use it conveniently is when we do not know in advance how big a sample we will want.

6. Stratified Sampling

This last technique is a sort of combination of Importance Sampling and

Systematic Sampling. For example, if we were only a little bit sophisticated and were doing the systematic sampling described above, we would soon notice that there is no point in considering the 400 tosses in which we had assigned the values three to six for the first toss of the die, since under these circumstances, we can never get a total of three. Therefore, we might systematically divide the sample into halves rather than sixths. In the first half we would say that the first die came up one, and in the second half that the first die came up two.

In theory, this method could be as powerful as Importance Sampling. In actual practice, the fact that you have to sample systematically turns out to decrease sharply the number of places in which it can be used. However, where it can be used, it is usually better than Importance Sampling and in any case never worse. Therefore whenever the costs of the two techniques are comparable, Stratified Sampling is preferable to Importance Sampling.

Part Two

The Evaluation of Definite Integrals[6]

The second part of the paper is concerned with the evaluation of definite integrals. Formulae will be derived which can, in principle, be used in a direct way to design sampling schemes. In practice, however, these formulae can rarely be used explicitly, because usually the Monte Carlo designer cannot obtain accurate numerical values of the variables that appear in them. It turns out though that if the Monte Carlo experimenter uses approximate values in the formulae, then the effect of using these approximations is not to bias the estimating procedure, but only to increase the statistical error of the calculation as compared with what would have been obtained if correct values had been used. Usually though the statistical error will have been decreased, at least as compared with what would be obtained if no design at all were attempted.

Perhaps the best way to describe the situation is to say that the formulae are a powerful guide to one's intuition in designing useful sampling schemes but

6. While Part Two requires somewhat more mathematical maturity to read than Part One, it does not require as much more as the casual reader is likely to think at first glance. However some skipping or a preliminary skimming may be helpful to the novice.

can rarely be used in a routine way. They are suggestive and normative rather than explicitly directive. It is, however, often possible to do subsidiary studies--Monte Carlo, analytic, or numerical--which can be used with the formulae to guide intuition.

It is probably advisable at this point to add a few words of caution. Techniques described under the headings of Importance Sampling, Russian Roulette Splitting, and Stratified Sampling have the property that in many calculations they will give a tremendous increase in efficiency if properly used; if, however, the intuition of the calculator is faulty and he does not use a reasonable design, then they can be very unreliable and actually increase the variance. The other techniques are more stable in that it is almost impossible for the experimenter to worsen the sampling variance by misusing them, even if he has a bad intuition.

We will not discuss in this paper how the experimenter can protect himself from trouble, except to say that the usual method of estimating error (calculating the sample variance and then appealing to the Central Limit Theorem) is usually satisfactory; however it can give trouble in semi-pathological though real cases.

The six techniques will be illustrated by considering the problem of estimating the integral

$$\iint_S z(x,y)f(x,y)dxdy = \overline{z(x,y)} = \bar{z} \tag{6}$$

by random sampling. $z(x,y)$ is arbitrary and

$$f(x,y) \geq 0 \text{ for } (x,y) \text{ in } S \tag{7}$$

$$\iint_S f(x,y)dxdy = 1$$

Two independent variables are shown; actually most of the discussion is unchanged if x and y are multi-dimensional variables or if one of the variables is eliminated.

$\bar{z}$ is defined as the expected value of $z(x,y)$ with respect to $f(x,y)$. $f(x,y)$ is a probability density function (to be abbreviated p.d.f.) over the space S. Mathematically speaking, this statement is equivalent to conditions (7). Roughly $f(x,y)$ can be thought of as being proportional to the joint probability that $X = x$ and $Y = y$. More properly, $f(x,y)dxdy$ is the probability that $x \leq X \leq x + dx$ and $y \leq Y \leq y + dy$.

The representation of the integral as an expected value does not in any way limit the generality of the problem; such a representation can be made for any integral.

An estimate of $\bar{z}$ can be obtained very simply by Monte Carlo. A number (N) of sample values (x_i, y_i) are picked from the p.d.f., $f(x,y)$,[7] the function $z(x,y)$

---

7. Butler's paper in this book describes ways to pick from a p.d.f.

is evaluated for each (x,y) sample, and the arithmetical average

$$\hat{z} = \frac{1}{N}\sum_{i=1}^{N} z(x_i,y_i) \tag{8}$$

of the N numbers $z(x_i,y_i)$ is calculated.

This sample average $\hat{z}$ is an estimate of $\bar{z}$. As has been mentioned (note 3 on page 2) any statistical estimate has an error which is described in terms of the variance. The variance is defined as:

$$V = \frac{1}{N}\iint[z(x,y) - \bar{z}]^2 f(x,y)dxdy = \frac{1}{N}\overline{[z(x,y) - \bar{z}]^2} \tag{9}$$
$$= \frac{1}{N}\left[\iint z^2(x,y)f(x,y)dxdy - \bar{z}^2\right] = \frac{1}{N}(\overline{z^2} - \bar{z}^2)$$

The square root of the variance is called the standard deviation and will be denoted by σ. We will loosely refer to either σ or V as the error. Let us now investigate what happens to this error in the six techniques.

1. Importance Sampling

If the integrand is multiplied and divided by an arbitrary p.d.f., $f^*(x,y)$, $\bar{z}$ is written

$$\bar{z} = \iint z(x,y)\frac{f(x,y)}{f^*(x,y)}f^*(x,y)dxdy \tag{10}$$

One can now pick from $f^*(x,y)$ instead of $f(x,y)$, and $z(x,y)f(x,y)/f^*(x,y)$ plays the role that $z(x,y)$ played previously. Therefore for (x_i,y_i) picked from $f^*(x,y)$ the estimate of $\bar{z}$ becomes

$$\hat{z}_1 = \frac{1}{N}\sum_{i=1}^{N} z(x_i,y_i)\,\frac{f(x_i,y_i)}{f^*(x_i,y_i)} \tag{11}$$

instead of $\hat{z}$ as given by equation (8). In addition to being a p.d.f., $f^*(x,y)$ is subject to the condition that it can be zero only when $z(x,y)$ or $f(x,y) = 0$. The expected value of $\hat{z}_1$ is, of course, the same as that of $\hat{z}$. Even though the expected value has not been changed, the errors of the two estimates are quite different. The variance of $\hat{z}_1$ is obtained by modifying equation (9) in the obvious way ($z(x,y)f(x,y)/f^*(x,y)$ for $z(x,y)$ and $f^*(x,y)$ for $f(x,y)$):

$$V_1 = \frac{1}{N}\iint\left[z(x,y)\frac{f(x,y)}{f^*(x,y)} - \bar{z}\right]^2 f^*(x,y)dxdy \tag{12}$$

$$= \frac{1}{N}\left[\iint \frac{z^2(x,y)f^2(x,y)}{f^*(x,y)}\,dxdy - \bar{z}^2\right]$$

and this is in general different from V given in equation (9). It is easily shown[8] that V_1 is minimized when

$$f^*(x,y) = \frac{|z(x,y)|\,f(x,y)}{\iint |z(x,y)|\;f(x,y)dxdy} \tag{13}$$

8. One applies the standard Calculus of Variations techniques to the expression

$$\iint \frac{z^2(x,y)\;f^2(x,y)}{f^*(x,y)}\,dxdy + \lambda \int f^*(x,y)dxdy$$

When $z(x,y)$ is non-negative and the optimal $f^*(x,y)$ is used, a zero variance estimate is obtained. This can be verified by substituting $f^*(x,y) = z(x,y)f(x,y)/\bar{z}$ into equation (12) or more simply by noticing that the separate estimates $z(x_i,y_i)f(x_i,y_i)/f^*(x_i,y_i)$ are actually constant and equal to $\bar{z}$ independently of what (x_i,y_i) were picked.

Of course before the optimum $f^*(x,y)$ can be obtained, $\bar{z}$ must be known. In more complicated problems even more than just the answer must be known before a zero variance sampling scheme can be designed. Under such circumstances zero variance is not a miraculous result.

The significance of the existence of zero variance estimates lies not in the possibility of actually constructing them in practice but in that they demonstrate there are no "Conservation of Cost" laws. That is, if the designer is clever, wise, or lucky he may, in choosing from the infinite number of sampling schemes available, be able to choose a very efficient one. This is in some contrast to the situation in ordinary numerical analysis. It is usually true there that once a fairly good method of doing a problem has been found, that further work or additional transformations do not reduce the cost very much, if at all. In Monte Carlo problems however we are assured that there is always a better way until we reach perfection.

When $z(x,y)$ changes sign in the area of integration, a perfect sampling scheme cannot be devised using simple Importance Sampling alone. The minimum variance using simple Importance Sampling is found to be

$$V_2 = \frac{4}{N}\left[\iint_{S_1} |z(x,y)|\,f(x,y)dxdy\right]\left[\iint_{S_2} |z(x,y)|\,f(x,y)dxdy\right] \tag{14}$$

where $z(x,y)$ is positive in S_1 and negative in S_2. This V_2 may be, but is not necessarily, small.

It is still possible, in principle, to design a perfect sampling scheme

for a general $z(x,y)$ by using correlation in addition to importance sampling. Since the author does not know of any practical application of this idea to functions $z(x,y)$ which change sign, there will be no more discussion on this point.

The name Importance Sampling was suggested by the possibility of sampling proportional to importance as shown by the theoretical zero variance estimates. In the simple case just considered, $|z(x,y)|f(x,y)$ measures the importance of the point (x,y) in the sense that it is proportional to the amount that this point contributes to $\bar{z}$, the quantity being calculated. It will turn out that this result is not general and that our use of the word, "importance", will usually carry a somewhat different connotation.

If $z(x,y) \geq 0$, the optimum Importance Sampling for $\bar{z}$ will also reduce the variance for the estimate of the higher moments, $\overline{z^n}$. But this is not necessarily true of just any $f^*(x,y)$ which reduces the variance of the estimate $\bar{z}$.

The central idea in Importance Sampling can be exploited in many different ways. Two possibilities will be mentioned. The following definitions will be needed in the discussion:

$f(x) = \int f(x,y)dy,$	the p.d.f. of the random variable x averaged over y, the so-called marginal p.d.f.
$g(y:x) = \frac{f(x,y)}{f(x)},$	the p.d.f. of y given that $X = x$, the so-called conditional p.d.f.
$\bar{z}(:x) = \int z(x,y)g(y:x)dy,$	the expected value of $z(x,y)$ given that $X = x$.
$\overline{\bar{z}(:x)} = \int \bar{z}(:x)f(x)dx = \bar{z},$	an obvious formula
$\overline{z^2}(:x) = \int z^2(x,y)g(y:x)dy,$	the expected value of $z^2(x,y)$ given that $X = x$
$\overline{\overline{z^2}(:x)} = \overline{z^2},$	follows as before.

Importance Sampling in the x space only

$\bar{z}$ can also be written

$$\bar{z} = \iint z(x,y) \frac{f(x)}{f^*(x)} g(y:x) f^*(x) dxdy \qquad (15)$$

If the (x_i, y_i) are picked out of $f^*(x,y) = g(y:x)\, f^*(x)$, an estimate of $\bar{z}$ is

$$\hat{z}_{11} = \frac{1}{N} \sum_{i=1}^{N} z(x_i, y_i) \frac{f(x_i)}{f^*(x_i)} \tag{16}$$

The variance of this estimate is

$$\begin{aligned} V_{11} &= \frac{1}{N} \left\{ \iint \left[z(x,y) \frac{f(x)}{f^*(x)} \right]^2 f^*(x)\, g(y:x)\, dxdy - \bar{z}^2 \right\} \\ &= \frac{1}{N} \left[\iint \frac{z^2(x,y)\, f^2(x)}{f^*(x)} g(y:x)\, dxdy - \bar{z}^2 \right] \\ &= \frac{1}{N} \left[\int \frac{\overline{z^2}(:x)\, f^2(x)}{f^*(x)} dx - \bar{z}^2 \right] \end{aligned} \tag{17}$$

The minimizing $f^*(x)$ is given by

$$\begin{aligned} f^*(x) &= \frac{f(x) \sqrt{\overline{z^2}(:x)}}{\int \sqrt{\overline{z^2}(:x)}\, f(x)\, dx} \\ &= \frac{f(x) \sqrt{\overline{z^2}(:x)}}{\overline{\sqrt{z^2}(:x)}} \end{aligned} \tag{18}$$

If this f^* is used, the variance becomes

$$V_{11} = \frac{1}{N} \left[\overline{\sqrt{z^2}(:x)}^2 - \bar{z}^2 \right] \tag{19}$$

The difference between this variance and that obtained by straightforward sampling is

$$V - V_{11} = \frac{1}{N} \overline{\left(\sqrt{\overline{z^2}(:x)} - \overline{\sqrt{z^2}(:x)} \right)^2} \tag{20}$$

and the variance has been reduced.

Sampling only from the x space is significant because it is common in practice to break the problem into two or more parts or stages. If this is done we must define the term "interesting" regions in a more general fashion than we did previously. Thinking of $X = x$ as a "cut" or region of the x

space, equation (18) implies that we should sample such regions with a frequency proportional to $f(x)\sqrt{\overline{z^2}(:x)}$. In other words, optimum Importance Sampling means that the a priori probability $f(x)$ of getting into a region x should be modified by the factor $\sqrt{\overline{z^2}(:x)}$.

When Importance Sampling is done in the entire space this factor is $|z(x,y)|$. In this case the word "importance" is used in the natural sense--that is, those regions are called important that make large contributions to the answer being calculated. However the natural analogue of $|z(x,y)|$, $|\bar{z}(:x)|$, is not the correct factor for optimum Importance Sampling in the x space alone. In a rough way we can say that the correct factor $\sqrt{\overline{z^2}(:x)}$ is proportional to the error rather than importance (used in the ordinary sense).

In some problems, the difference between the two definitions of importance is not as great as we may have implied, for

$$\overline{z^2}(:x) = \sigma^2(:x) + \bar{z}^2(:x) \tag{21}$$

where $\sigma^2(:x) = \overline{z^2}(:x) - \bar{z}^2(:x)$ measures the variance left in the y space.

If this variance is small compared to $\bar{z}^2(:x)$, then

$$\sqrt{\overline{z^2}(:x)} \approx \bar{z}(:x) \tag{22}$$

In many situations we may not have even a rough idea of what $\overline{z^2}(:x)$ is like, but may have a very good estimate of or approximation to $\bar{z}(:x)$. For this reason one often uses $\bar{z}(:x)$ rather than $\sqrt{\overline{z^2}(:x)}$. If this is done, then (for $\bar{z}(:x) > 0$)

$$f^*(x) = \frac{f(x)\,\bar{z}(:x)}{\bar{z}} \tag{23}$$

and the variance becomes

$$V_{12} = \frac{1}{N}\left\{\bar{z}\left[\frac{\overline{z^2}(:x)}{\bar{z}(:x)}\right] - \bar{z}^2\right\} = \frac{1}{N}\,\bar{z}\left[\frac{\sigma^2(:x)}{\bar{z}(:x)}\right] \tag{24}$$

While V_{12} is ordinarily much less than V it is easy to see that it can be large and in fact disasters are possible if care is not taken.

The problem of estimating $\overline{z^2}(:x)$ or $\bar{z}(:x)$ is crucial if Importance Sampling is to be done. It is not necessarily difficult. Previous Monte Carlo calculations, new ones perhaps, experiments, approximate calculations, or intuition can all be used to get the information. In any case the designer needs to know only the relative importance of different regions and not the absolute values.

This may make the problem much easier. If one does not have the information it is usually worthwhile to go to some effort to get reasonable estimates of $\overline{z^2}(x)$ or $\bar{z}(x)$.

Importance Sampling with a parameter

A second variation of Importance Sampling worth mentioning is restriction of the $f^*(x,y)$ to a single family of p.d.f.'s. This may be desirable to do for either computing or theoretical convenience. If such a family is represented by $h(x,y,\alpha)$, then it is desired to determine α to minimize the variance. Since $\bar{z}$ is fixed this is equivalent to minimizing (see Equation (14)):

$$\iint \left[\frac{z(x,y)\, f(x,y)}{h(x,y,\alpha)}\right]^2 h(x,y,\alpha)\,dxdy = \iint \frac{z^2(x,y)\, f^2(x,y)}{h(x,y,\alpha)}\,dxdy \qquad (25)$$

$h(x,y,\alpha)$ is, of course, subject to the usual conditions for being a p.d.f. If the form of $h(x,y,\alpha)$ is such that these conditions are satisfied for all values of α, then the optimal α is determined by

$$\iint \frac{z^2(x,y)\, f^2(x,y)}{h^2(x,y,\alpha)} \frac{\partial h(x,y,\alpha)}{\partial \alpha}\,dxdy = 0 \qquad (26)$$

One way to solve such an equation is to do a preliminary study by Monte Carlo. The function to be minimized is

$$I(\alpha) = \iint \frac{z^2(x,y)\, f^2(x,y)}{h(x,y,\alpha)}\,dxdy \qquad (27)$$

It is of course possible to evaluate (27) by Monte Carlo. It is not necessary, as might be supposed, to sample from $h(x, y, \alpha)$ when doing this evaluation because $I(\alpha)$ can be written

$$I(\alpha) = \iint \frac{z^2(x,y)\, f^2(x,y)}{h(x,y,\alpha)\, f^*(x,y)} f^*(x,y)\,dxdy \qquad (28)$$

so that an estimate of $I(\alpha)$ is

$$I(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \frac{z^2(x_i,y_i)\, f^2(x_i,y_i)}{h(x_i,y_i,\alpha)\, f^*(x_i,y_i)} \qquad (29)$$

where the (x_i,y_i) are picked out of $f^*(x,y)$. The chief application of formula (29) is when $f^*(x,y) = h(x,y,\alpha_1)$ and we wish to evaluate $I(\alpha)$ for a series of α's, say α_1, α_2, α_3.

In principle, if we chose $\alpha_2 < \alpha_1$ and $\alpha_3 > \alpha_1$ we could observe $I(\alpha_j)$ for these three values. We might then be able to see how α should be changed in order to decrease the sampling variance. We might even program this adjustment of α so that it is done automatically by the computing machine. The procedure can be dangerous, however, (see Marshall's paper) and as far as I know has never been done in this automatic fashion. The idea, however, is so intriguing that is seems worth mentioning.

It should also be clear that if we can evaluate any derivatives of $h(x,y,\alpha)$ then we could evaluate $\frac{\partial I}{\partial \alpha}$ and even maybe $\frac{\partial^2 I}{\partial \alpha^2}$ directly by Monte Carlo and use these quantities to estimate what α should be used in a subsequent calculation.

Examples

In the examples which follow here and later we will be discussing situations which may not lend themselves to being expressed in a simple fashion as integrals. More typically one would tend to describe these situations in terms of integral equations. Despite the fact that the theory of Monte Carlo applied to integral equations is in some sense richer than the one we are discussing, we draw from both for examples because most realistic examples come from the former field and because the ideas carry over at least qualitatively.

One of the major problems on which Monte Carlo can be used is to calculate the probability that nuclear particles will penetrate shields. In such a problem, the particle starts at one side of the shield and has collisions of different types with the material inside the shield, finally being either reflected backwards, absorbed in the shield, or transmitted. The calculation can be done by Monte Carlo by simulating the particle histories with the aid of random numbers. The simulation should not be faithful. For example the following types of random events increase the probability of penetration and should be emphasized at the expense of equally probable, but less "important", ones:

a. collisions resulting in a forward direction of motion
b. in most problems collisions which result in small energy losses
c. long forward jumps and short backward jumps (the exponetial transformation discussed in the paper by Beach and Theus does this)
d. survival vs. absorption (if carried to the limit, this can be looked on as an application of Use of Expected Values)

The calculator is, of course, confronted with the problem of how far to go quantitatively in biasing the sampling. In deciding this he must fall back on trying to estimate the formulae we have just derived.

Another important application of Monte Carlo is in the design or analysis of reactors. Here again we are studying the various ways in which nuclear particles--particularly neutrons--behave in matter. It is found that those which wander away from the center of the reactor will not contribute much of interest to the process. On the contrary, those neutrons which wander back toward the fissionable material are the ones which contribute most to the answer. The sampling must then be designed to sample more frequently among the second type of neutrons and less frequently among the first type of neutrons.

Applications can also be made to Operations Research problems. In one typical problem of this type we might try to calculate the vulnerability of a piece of equipment or of an airfield to some type of offensive weapon. In such cases, one often Monte Carlos, for example, the error of the missile which is doing the destruction. The distribution of errors is determined by a parameter called the CEP. Since the thing we are interested in here is destruction, if we Monte Carlo from a distribution defined from a smaller CEP than that which obtains in the real world, we will find that more of the interesting processes (hits) happen in the simulated experiment than would happen naturally. The less interesting processes (misses) are then discriminated against. This is, of course, the right kind of Importance Sampling, and is one way to do Importance Sampling with a parameter--the parameter in this case being the sampling CEP.

For another example in the same field one might consider queuing problems. In these problems one is often interested in the mean and variance of the waiting time. One then wishes to bias the sampling to emphasize long waits. This could be done by sampling from f^*'s which simulate increased traffic, increased servicing time, or increased servicing requirements.

It is worth noting that any set of samples obtained with the use of Importance Sampling is less effective in estimating certain quantities than a set that has been obtained in a straightforward fashion. For instance in the shielding problem, a sampling design which leads to an accurate estimate of the probability of penetration will be very poor for estimating the probability of reflection; in the reactor problem the suggested sampling would not be good at estimating leakage; in the vulnerability problem we will lose information about light damage and the location of misses; and finally in the queuing problem we will not get a good estimate of the idle time of the servicing facilities.

It is in fact usually (but not always) true that to design an efficient Monte Carlo calculation one must focus attention on those things one is really interested in and ignore other aspects of the problem. It may even be better to do more than one calculation than to compromise the goals of any particular design. This can be a serious disadvantage if the computation is lengthy, as the decision about what aspects to concentrate on must be made early and therefore may easily be made wrongly.

2. Russian Roulette & Splitting

This technique is essentially a sequential sampling scheme which can be used when the sampling is done in stages; in our case when first an x value is picked and then a y, For some regions of x the error is small enough or the expense of picking y and evaluating $z(x,y)$ large enough that it isn't

efficient to pick a y and evaluate a $z(x,y)$ for every such x that has been sampled, but only for a portion of them. In other regions of the x space, the error in the y space is large enough or the additional cost small enough that it pays to sample many values of y for each x value that has been picked. It may be that every x value can be classified as belonging to one region or the other.

To state the problem explicitly, we wish to determine the following:

a. Regions R_1 and R_2 which are mutually exclusive and exhaust the x space. In R_1 we are going to sample a y and evaluate $z(x,y)$ for only a portion of the x's sampled. In R_2 we are going to sample one or more y values for each x value picked.

b. The sampling in R_1 is called Russian Roulette and is done as follows. Every time a chosen x_i is in R_1 a function $q(x_i)$ is evaluated and a supplementary game of chance played. With probability $q(x_i)$ this supplementary game is won and a y_i picked and $z(x_i,y_i)$ evaluated; with probability $1 - q(x_i)$ the game is lost and no y_i is picked.

Corresponding to the two possible outcomes of the game, the i^{th} sample estimate becomes

$$z_{2i} = \begin{cases} \dfrac{z(x_i,y_i)}{q(x_i)} & \text{for a win} \\ 0 & \text{for a loss} \end{cases} \tag{30}$$

It is easy to see that the expected value of z_{2i} for fixed x_i is $\bar{z}(:x_i)$. The variance of z_{2i} for fixed x_i is

$$\sigma_2^2(:x_i) = \frac{\overline{z^2}(:x_i)}{q(x_i)} - \bar{z}^2(:x_i) \tag{31}$$

c. The sampling in R_2 is called Splitting. When an x_i is picked from R_2, a positive integral-valued function $n(x_i)$ is evaluated which gives the number of y samples to be picked. The i^{th} sample estimate then becomes the arithmetical average of the $n(x_i)$ values of $z(x_i,y_j)$, $j = 1, 2, \ldots, n(x_i)$:

$$z_{2i} = \frac{1}{n(x_i)} \sum_{j=1}^{n(x_i)} z(x_i, y_j) \tag{32}$$

Again the expected value of z_{2i} is $\bar{z}(:x_i)$.
The variance in this case is

$$\sigma_2^2(:x_i) = \frac{\sigma^2(:x_i)}{n(x_i)} \tag{33}$$

d. R_1, R_2, $q(x)$, and $n(x)$ are chosen to minimize the total cost of the problem for a fixed level of error. This is equivalent to minimizing the product of the average cost of a sample and the variance. This can be seen as follows. If the average cost is C then the total cost is NC. If the fixed error is denoted by ε then

$$N = \frac{V_2}{\varepsilon^2} \tag{34}$$

and the total cost becomes

$$NC = \frac{CV_2}{\varepsilon^2} \tag{35}$$

Since ε is constant the total cost is minimized when CV_2 is minimized.

If ordinary sampling were used the variance can be written in the form

$$V = \overline{\left[\bar{z}(:x) - \bar{\bar{z}}\right]^2} + \overline{\sigma^2(:x)} \tag{36}$$

Only the last term is affected by the Russian Roulette & Splitting with $\sigma^2(:x)$ replaced by $\sigma_2^2(:x)$ as defined in equations (31) and (33). V_2 then becomes

$$V_2 = A + \int_{R_1} \frac{\overline{z^2}(:x)}{q(x)} f(x)dx + \int_{R_2} \frac{\sigma^2(:x)}{n(x)} f(x)dx \tag{37}$$

where

$$A = \int_{R_2} \bar{\bar{z}}^2(:x) f(x)dx - \bar{\bar{z}}^2$$

The expected cost of a sample is given by

$$C = C_o + \int_{R_1} C_1(x)q(x)f(x)dx + \int_{R_2} C_1(x)n(x)f(x)dx \tag{38}$$

where C_o is the expected cost of picking an x (by definition independent of y) and $C_1(x)$ is the average additional cost of picking y and evaluating $z(x,y)$ for that x.

The R_1, R_2, $q(x)$ and $n(x)$ which minimize CV_2 are determined[9] by the set of simultaneous conditions:

---

9. Calculus of Variations again. Also, the requirement that n(x) be integral has been temporarily ignored.

$$q(x) = \frac{\lambda\sqrt{z^2(x)}}{\sqrt{C_1(x)}} \leq 1 \qquad \text{for x in } R_1$$

$$n(x) = \frac{\lambda\sigma(x)}{\sqrt{C_1(x)}} \qquad \text{for x in } R_2 \tag{39}$$

R_1 and R_2 mutually exclusive and exhaustive

where $\lambda = \sqrt{C/V_2}$.

If one substitutes equations (39) into (37) and (38), one finds that

$$\lambda = \sqrt{\frac{C_o}{A}} \tag{40}$$

At first sight finding the R_1, R_2, $q(x)$, and $n(x)$ which satisfy conditions (39) appears difficult. There is however an iterative procedure which will converge rapidly:

1. Take A as being equal to $\int_{R_1+R_2} \bar{z}^2(x)f(x)dx - \bar{z}^2$ rather than $\int_{R_2} \bar{z}^2(x)f(x)dx - \bar{z}^2$ and call this approximation A'. A' is too large by an amount $\int_{R_1} \bar{z}^2(x)f(x)dx$, but this is usually unimportant.

One reason is because R_1 is defined as the region where $\overline{z^2}(:x)$ is small and wherever $\overline{z^2}(:x)$ is small, $z^2(:x)$ is also small.

2. An approximate λ can now be estimated by using A' for A in equation (40) (or by any other reasonable method).
3. Equations (39) can now be used to define the functions $q(x)$ and $n(x)$ for the entire x space and not just the appropriate regions. Because

$$\overline{z^2}(:x) \geq \sigma^2(:x) \tag{41}$$

the everywhere defined $q(x)$ is greater than $n(x)$. If therefore we define approximate regions R_1 and R_2 by the conditions

$$\begin{aligned} &x \text{ in } R_1 \text{ if } q(x) \leq 1 \\ &x \text{ in } R_2 \text{ if } n(x) \geq 1 \end{aligned} \tag{42}$$

R_1 and R_2 will be mutually exclusive. It can be shown that it is best to put anything left over into R_2.

4. A new A' approximation to A can now be computed by using the approximate R_2 as the region of integration in the definition of A and a new λ estimated.
5. Steps 3-4 are repeated until reasonable convergence is obtained.

We can estimate (but only roughly) the improvement over straightforward sampling that is gained when optimum Russian Roulette & Splitting is used. In most problems, $C_1(x)$ is not a sensitive function of x and can be taken equal to an average C_1'. Similarly (as already mentioned) there is very little error introduced if A is taken equal to $\overline{\left[\bar{z}(:x) - \bar{z}\right]^2}$ (which we will represent by A'). Lastly we make the dubious approximation that in R_1

$$\overline{z^2}(:x) = \sigma^2(:x) \tag{43}$$

We can now write that

$$V_2 \approx A' + \overline{\sigma(:x)} \sqrt{\frac{C_1' A'}{C_o}} \tag{44}$$

$$C \approx C_o + \overline{\sigma(:x)} \frac{C_o C_1'}{A'} \tag{45}$$

$$CV_2 \approx C_o A' + 2\overline{\sigma(:x)} \sqrt{C_o C_1' A'} + C_1' \overline{\sigma(:x)}^2 \tag{46}$$

When ordinary sampling is used the product of the cost and the variance is

$$CV \approx (C_o + C_1')\left[A' + \overline{\sigma^2(:x)}\right] \qquad (47)$$

$$= C_o A' + C_o\overline{\sigma^2(:x)} + C_1' A' + C_1'\overline{\sigma^2(:x)}$$

Subtracting equation (46) from equation (47) and collecting terms

$$CV - CV_2 = (C_o + C_1')\left[\overline{\sigma(:x) - \overline{\sigma(:x)}}\right]^2 + \left[\sqrt{C_1' A'} - \overline{\sigma(:x)}\sqrt{C_o}\right]^2 \qquad (48)$$

The first term on the right side of equation (48) can be easily interpreted. $(C_o + C_1')$ is the average cost of a sample when doing straightforward sampling so the improvement is measured by comparing $\left[\overline{\sigma(:x) - \overline{\sigma(:x)}}\right]^2$ with V (i.e. the variance of $\sigma(:x)$ as compared with the variance of $z(x,y)$). There is an additional improvement given by the second term which is related to the fact that even if $\sigma(:x)$ didn't vary at all it might still pay to sample many y values for every x picked.

It should be clear to the reader that most of the examples mentioned in Importance Sampling could also be used to illustrate the use of Russian Roulette Splitting. It is instructive to contrast the two techniques. In Importance Sampling we bias the sampling in favor of "interesting" regions and try to avoid "uninteresting" regions. Sometimes though, f(x) is not given explicitly, but a complicated process for getting x values is given instead. In these circumstances we may have very good estimates of the relative "importance" of different regions and still not be able to do this biasing. We can then use Russian Roulette & Splitting which, in effect, does the same thing but ad hoc; that is, we wait to see what region is entered and then decide what the size of the sample should be.

Two examples special to the Russian Roulette Splitting technique will be given.

Application to Particle Diffusion

A special case of some interest involving three random variables (x, W, M) arises in particle diffusion problems. X is a generalized position variable which represents the position and momentum of the particle. W is a pseudo weight that is assigned to the particle, and which changes as the particle jumps from point to point. M is the final weight of the particle divided by the current weight. It is convenient to think of it the other way, as a factor which multiplies the current weight when the random walk is terminated. The

function whose expected value is desired is the final weight of the particle; so

$$z(x, w, m) = wm \tag{49}$$

where if ordinary sampling were being done the m would be independent of w. However if the w happened to be very small one would be willing to sample m rather inaccurately if it saved some cost; if it were large one would want to sample m quite accurately even if it were expensive. This is thus a natural problem on which to use Russian Roulette & Splitting.

The regions R_1 and R_2 will be defined in the (x,w) space and the decisions of Russian Roulette or Splitting concern the number of m values that are to be picked for an (x,w) value.

In discussing particles it is convenient to change the language slightly. Instead of speaking of picking n independent values of m for each (x,w) in region R_2, the particle is said to split into n independent particles each of weight w/n. Similarly in region R_1 if the particle loses the Russian Roulette it is said to have died (or disappeared); if it wins it is assigned a new weight, w/q(x,w) and its random walk continued.

In the case of most interest the p.d.f. for (X, W, M) (after the Russian Roulette & Splitting has been done) has the special form, g(m:x)f(x,w); that is the conditional p.d.f. of m is not dependent on w. (The plausibility of the assumption is discussed below.) With this assumption,

$$\overline{z^2}(:x,w) = w^2\overline{m^2}(:x) \tag{50}$$

$$\sigma^2(:x,w) = w^2\left[\overline{m^2}(:x) - \bar{m}^2(:x)\right] \tag{51}$$

Since the cost of picking an m value is independent of w, the optimum choice of q(x,w) and n(x,w) is given by

$$\begin{aligned} q(x,w) &= \lambda w \sqrt{\frac{\overline{m^2}(:x)}{C_1(x)}} \\ n(x,w) &= \lambda w \sqrt{\frac{\overline{m^2}(:x) - \bar{m}^2(:x)}{C_1(x)}} \end{aligned} \tag{52}$$

and regions R_1 and R_2 are determined as before by the appropriate inequalities. If the particle is in R_1 and happens to survive the Russian Roulette it is

assigned a new weight

$$w' = \frac{w}{q(x,w)} = \frac{1}{\lambda}\sqrt{\frac{C_1(x)}{\overline{m^2}(:x)}} \tag{53}$$

If the particle is in R_2 then each of the n independent particles is given a weight.

$$w' = \frac{w}{n(x,w)} = \frac{1}{\lambda}\sqrt{\frac{C_1(x)}{\overline{m^2}(:x) - \overline{m}^2(:x)}} \tag{54}$$

In both cases, the final weight of the particle is independent of the original weight and is a function of x only. (This seems to be, in general, one of the criteria for a good sampling scheme for particle diffusion problems.) It is because the weight of the particle after collision is independent of the weight before collision that m can be taken to be independent of w.

Truncating Sample Series

Sometimes in doing a Monte Carlo problem instead of getting a single number for the estimate from a single sample, one obtains an infinite series; more precisely each sample generates a process for calculating an infinite series term by term, and it is the sums of these series which are to be averaged in obtaining the final estimate. This occurs most often when the sampling technique in the next section, Use of Expected Values, is used.

The computer is then faced with the problem of terminating each of the sample series. This can be done by summing each series to a fixed number of terms or it can be done by summing until a term gets smaller than some previously assigned amount. Both of these methods are inefficient as one is then faced with either calculating a number of very small terms or truncating too soon and introducing an unknown bias into the estimate.

A much more effective method of terminating such sample series is made possible by the use of Russian Roulette. One can simply play the supplementary game of chance as soon as the terms in the series begin to get small. If a term fails to survive the supplementary game the series is terminated right then and there; if it survives, the weights of all subsequent terms are multiplied by the proper factor and the term by term summation continued until a

new term becomes small. In this way the series can be terminated in a completely unbiased fashion and yet very little effort is spent computing small and insignificant numbers.

3. Use of Expected Values

Sometimes it is a simple matter to evaluate $\bar{z}(:x)$. In this case the problem of evaluating

$$\bar{z} = \iint z(x,y)f(x,y)dxdy \tag{55}$$

can profitably be replaced by the problem of evaluating

$$\bar{z} = \int \bar{z}(:x)f(x)dx \tag{56}$$

In many cases this simple integral cannot be evaluated either because x is multi-dimensional or because f(x) is not actually known and the sampling of x has to be done in some complicated manner. In either case there is no point in using Monte Carlo to do that part of the problem which can be done analytically. Only the x's should be sampled--the sampling of the y space is bypassed by using the estimate

$$\hat{z}_3 = \frac{1}{N}\sum_{i=1}^{N} \bar{z}(:x_i) \tag{57}$$

The variance is easily calculated to be

$$V_3 = \frac{1}{N}\int\left[\bar{z}(:x) - \bar{z}\right]^2 f(x)dx \tag{58}$$
$$= \frac{1}{N}\overline{\left[\bar{z}(:x) - \bar{z}\right]^2}$$

This is less than the original variance by

$$V - V_3 = \frac{1}{N}\int\sigma^2(:x)f(x)dx \tag{59}$$
$$= \frac{1}{N}\overline{\sigma^2(:x)}$$

As is intuitively clear, the variance associated with the y space has been eliminated.

The reader should not conclude from the trivial nature of the explanation that the idea is not valuable. In many cases most of the variance of a problem

can be eliminated by using only a little bit of analysis. The author knows of a surprisingly large number of Monte Carlo problems where people have either not done, or failed to use, the results of a very significant but trivial piece of analysis. As a result they used sampling for all of the probabilistic parts of the problem, and had to use much larger samples than were really necessary.

Examples

Expected values can be used in all the problems mentioned under Importance Sampling. In the shielding problem, for example, instead of counting the particles which succeed in piercing the shield, we can calculate the probability after each collision that the particle would go through the shield. An unbiased estimate of the penetration is obtained by summing these probabilities. This is one way, in fact, in which the infinite series of estimates mentioned in the Russian Roulette section arises.

Another application of the same idea is in the calculation of multiple scattering effects in nuclear cross section measurements. In this calculation one simulates a scattering experiment in which the main interest is in the number of particles which enter a detector. Now the probability of a particle's entering the detector may be very small, say of the order of 10^{-4} or 10^{-6}, so that if one wished to sample in a straightforward fashion, one would practically never get any counts. It is, however, very simple to calculate the probability that the particle will enter the detector from any point in the system. If these probabilities are recorded, rather than the number of particles counted, one usually finds that the per cent variance is enormously reduced, 10^{3} being typical. The factor is larger in this case than in the previous (shielding) example because in this problem many samples tend to get into a region where the particle has a relatively high (though absolutely low) probability of getting into the detector; in the previous problem this is not true unless Importance Sampling or Russian Roulette & Splitting is also used.

This idea can be applied to more general problems than particle diffusion. For example, in vulnerability type studies, it is often true that the sampling is done in many stages and in the final stage there is given a probability p of achieving or not achieving a kill. It should be obvious to the reader at this point that it would be wrong to sample from this binomial distribution, but rather, the p's themselves should be recorded and their average used for the estimate.

A similiar situation occurs in the particle diffusion problems mentioned previously. Usually there is a non zero probability that particles will be absorbed after they have had a collision. This binomial process could be sampled, but it is generally more accurate not to sample, but simply to weight the particle with the probability of survival, multiplying all the probabilities together after it has had all its collisions. This increases the average length (number of collisions) of a history and therefore its cost, but this effect is usually dwarfed by the decrease in variance.

Another way to use expected values is to integrate a sample over the initial conditions. For example, in particle diffusion problems, it is not much work to translate, rotate, or reflect histories (thus getting new ones) and then average these translated, rotated, or reflected histories over their a priori probabilities.

In our integral example this would correspond to picking a function $y_i(x)$ out of $g(y:x)$ for the i^{th} sample and then never picking x at all but using

$$z_{3i} = \int z\left[x, y_i(x)\right] f(x)dx \tag{60}$$

as the estimate. There is a surprisingly large number of problems in which the above process can be carried through. If the pick on y is made by the standard method of solving the equation

$$\int_{-\infty}^{y_i(x)} g(y:x)dy = R_i \tag{61}$$

the above estimate corresponds to using $\bar{z}(:R_i)$. This kind of integration can be used for more than just the first stage. For instance, in plane geometry particle diffusion problems, it is possible to do the energy and angular part of the history by random sampling and then for this fixed energy-angular history, the spatial part can be done by an exact integration.

4. Correlation and Regression

There are many ways and many reasons for sampling two or more problems in a correlated fashion. We will discuss only a few of them here. As always we consider the evaluation of

$$\bar{z} = \iint z(x,y)f(x,y)dxdy \tag{62}$$

Let us assume that we know an integral

$$\bar{v} = \iint v(r,s)g(r,s)drds \tag{63}$$

where the function $v(r,s)$ and the p.d.f. $g(r,s)$ are in some unspecified way related to $z(x,y)$ and $f(x,y)$. (In most applications they will be approximations which are easy to integrate.) Assume also that sampling on (x,y) is such that it is easy or very cheap to do sampling on (r,s) also, and that in some or all parts of the calculation the same random numbers used in obtaining a value of (x,y) are also used in obtaining a value of (r,s), so that the two samples are not independent.

Instead of calculating the average value of $z(x,y)$ one can calculate the average value of

$$u(x,y,r,s) = z(x,y) - \alpha \left[v(r,s) - \bar{v}\right] . \tag{64}$$

It will be noticed that the expected value of $u(x,y,r,s)$ is equal to the expected value of $z(x,y)$ since the extra term $\alpha \left[v(r,s) - \bar{v}\right]$ has zero expected value. The variance of $u(x,y,r,s)$, however, is quite different from the variance of $z(x,y)$.

$$\begin{aligned} V_4 &= \frac{1}{N} \overline{\left[u(x,y,r,s) - \bar{u}\right]^2} \\ &= \frac{1}{N} \overline{\left\{\left[z(x,y) - \bar{z}\right] - \alpha \left[v(r,s) - \bar{v}\right]\right\}^2} \\ &= \frac{1}{N} \left\{\overline{\left[z(x,y) - \bar{z}\right]^2} - 2\alpha \overline{\left[z(x,y) - \bar{z}\right]\left[v(r,s) - \bar{v}\right]} + \alpha^2 \overline{\left[v(r,s) - \bar{v}\right]^2}\right\} \\ &= \frac{1}{N}\left(\sigma_1^2 - 2\alpha\rho\sigma_1\sigma_2 + \alpha^2\sigma_2^2\right) \end{aligned} \tag{65}$$

where σ_1^2, $\rho\sigma_1\sigma_2$, and σ_2^2 have the obvious definitions. ρ is called the correlation coefficient and, in some sense, measures the stochastic similarity of the two processes.

In many practical cases α is chosen in advance, usually to be one. If then, in some sense, the (r,s) problem is similar to the (x,y) problem, then σ_1 may be close to σ_2 and ρ close to 1. Under these circumstances the variance V_4 can be

much less than σ_1^2. How much less depends on how "close" the (r,s) problem is to the (x,y) problem.

In some cases, we can try to improve the situation by trying to choose that value of α which minimizes the variance of $u(x,y,r,s)$. Differentiating equation (65), one finds that this best value of α is

$$\alpha = \rho\frac{\sigma_1}{\sigma_2} \tag{66}$$

If this value of α is used, V_4 becomes

$$V_4 = \frac{1}{N}\sigma_1^2(1 - \rho^2) \tag{67}$$

It is clear from the above expression that the more highly the problems are correlated, the smaller one can make the variance of the difference. Of course, one does not ordinarily know $\rho\sigma_1/\sigma_2$, so it must be estimated. If the estimate is made directly from the sample being analyzed it will then turn out that equation (64) may not give an unbiased estimate of the average value of $z(x,y)$, since the expected value of $\alpha\left[v(r,s) - \bar{v}\right]$ is not necessarily zero if α is a function of (r,s).

The bias can be avoided by the following technique:

> The sample can be divided into two parts and the best estimate of $\rho\sigma_1/\sigma_2$ determined for each part separately. Each of these estimated values of α can then be used with the other half of the sample to form two estimates of $\bar{z}$. Under these circumstances the α's are independent of the terms they multiply. The two estimates are then averaged, to obtain a single unbiased estimate.

This causes a slight decrease in efficiency because two values of α are estimated and only one half of the sample is used for each estimate. The α's will therefore be a little more inaccurate than is necessary. It is not, however, usually important to estimate α very accurately and in some cases it is important to have unbiased answers, so it may be worthwhile to go to this scheme.

Elimination of the Variance Associated with $\bar{z}(x)$

A special application of the above technique occurs when one can explicitly

and easily evaluate integrals of the form

$$\bar{v} = \int v(x)f(x)dx \tag{68}$$

In this case, one replaces $v(r,s)$ by $v(x)$ and $g(r,s)$ by $f(x)$. Equation (65) then becomes

$$V_{41} = \frac{1}{N}\left\{\overline{\left[z(x,y) - \bar{z}\right]^2} - 2\overline{\left[z(x,y) - \bar{z}\right]\left[v(x) - \bar{v}\right]} + \overline{\left[v(x) - \bar{v}\right]^2}\right\} \tag{69}$$

It is easily shown that V_{41} is minimized when

$$v(x) = \bar{z}(:x) \tag{70}$$

If this $v(x)$ is used, then the variance becomes

$$V_{41} = \frac{1}{N}\overline{\sigma^2(:x)} \tag{71}$$

and the variance is reduced by

$$V - V_{41} = \frac{1}{N}\overline{\left[\bar{z}(:x) - \bar{z}\right]^2} \tag{72}$$

As will be shown in the next section on Systematic Sampling, this same reduction can be achieved in what might seem to be a much simpler fashion and by a method that does not require one to know $\bar{z}(:x)$. However, it turns out that there are complicated situations where Systematic Sampling cannot be used, so it may be convenient to have this alternative technique available.

It is worth pointing out that one doesn't need to know $\bar{z}(:x)$ exactly. Any rough approximation $v(x)$ will do so long as one knows how to compute $\bar{v}$ reasonably accurately.

Three examples illustrating the three usual reasons for doing correlations follow:

Parametric Studies

It is sometimes necessary to calculate a series of expected values where each member of the series involves a slightly different p.d.f. In this case one wishes to calculate a series of expected values of the form

$$\bar{z}_n = \iint z(x,y)f_n(x,y)dxdy \qquad n = 1, 2, \ldots, N \tag{73}$$

A simple way to do this is to pick the (x,y) once and for all out of an approximate p.d.f., $f^*(x,y)$, and then estimate the z_n by weighting with $f_n(x,y)/f^*(x,y)$. In problems where the picking is costly, but calculating the weighting factor is cheap, this process reduces the work roughly by a factor of N.

Comparing Different Bombing Strategies[10]

If a strategic or tactical bombing campaign is studied by Monte Carlo, it is customary to introduce random events of the following types:

1. Number of aircraft that abort
2. Number of aircraft shot down by area defense on the way in to the target
3. Number of aircraft that stray through navigational errors
4. Number of aircraft shot down by local defense at the target
5. Weather conditions over target (affects reconnaissance and CEP)
6. Place where bombs land
7. Damage done
8. Number of aircraft shot down by area defense on the way out of target area
9. Number of aircraft that don't get back for a miscellany of minor reasons.

Because some of the probabilities involved depend on the number of aircraft surviving to a particular stage of the strike, the above problem is non-linear. This does not prevent one from using any of the techniques suggested.

If the computer wishes to compare different bombing strategies it is often effective to use correlation to cut down the sample size required to get significant information. If the correlation is done by using the same random numbers the computer cannot use a single list of random numbers in sequence in the two problems, for they would soon get out of step. He can either throw away the excess random numbers or, what is sometimes better, save them for use on later strikes. For example, if a larger number of targets were attacked on the first strike of one of the strategies, the

---

10. This example is given in some detail because there seems to be a somewhat general misunderstanding in the operations analysis field of the applicability of the techniques in this paper.

extra random numbers that were used to determine the weather on these excess targets can be saved. If in a later strike an excess number of targets is attacked under the other strategy, the previously saved random numbers can then be used on these targets. Correlation can thus be achieved by using the same random numbers whenever the two strategies give rise to the same type of contingencies--even if they are on different strikes with different planes and targets.

Because the point is sometimes misunderstood, I would like to emphasize that when we say we are correlating the results of the same type of contingencies, the contingencies don't have to have the same detailed character. If for example in Strategy A, n_1 aircraft come up to the area defenses and in Strategy B, n_2 aircraft are used, the problems can still be correlated by using the same uniform random numbers in computing the number that survive. This is true as long as the picking is so arranged that the degree of success is a monotonic function of the uniform random numbers, so that fluctuations in the values of the uniform random numbers affect the two situations in the same qualitative way.

If the different strategies are such that a definite type of event is all-important to the comparison, then correlation by weighting may be better than by using the same random numbers. For example, if the effect of different types of defensive armament is being studied, the same kill probabilities could be used for the enemy fighters in the sampling, and weighting factors carried along to account for the differences being studied. The correlation may be higher if this is done, because exactly the same number of bombers is shot down each time, so all of the subsequent history is the same. If the correlating were done by using the same random numbers, different numbers of aircraft would be shot down and the actual progress of the two strategic campaigns might be quite different. It would still be possible to obtain correlation by using the same random numbers for the same contingencies, but it is unlikely that the correlation would be as high. Weighting will, of course, not work well if by its use one is forced to use an f^* which itself introduces a lot of variance because it is not good Importance Sampling for all the cases being considered (see next example).

Another case where weighting might be preferable to using the same

random numbers would be when two different reconnaissance devices were being compared. The possible weather situations can then be classified according to the following criteria:

1. Both devices work
2. One works and the other does not
3. Neither works

Only situation 2 makes a difference between the two devices so that in the sampling only it should be allowed to occur. If 1 and 3 occur, the sample would give zero for the estimate, so they need not be calculated; only the fraction r of time they occur is needed. Account of this is automatically taken by the weighting factors. If instead of weighting factors the same random numbers were used to do the correlating then (1-r) of the time the sample would be calculating zero and be wasted. If instead of being an all or nothing situation the devices have different probabilities of working as the weather changes, then the appropriate modification must be made in the sampling. This last is as much an example of Importance Sampling as of Correlation.

Polarization (comparing an exact and approximate theory)

In tracing γ rays through a medium it simplifies the problem greatly to assume that the γ rays are unpolarized. This assumption can be checked by doing two correlated problems, one using the exact laws and the other the approximate ones that are obtained when it is assumed that the particles are unpolarized. If weighting factors are used to do the correlating they would fluctuate wildly because the differential scattering laws are quite different in the two cases. The actual effect of polarization turns out to be quite small in most problems of interest. This is shown very effectively if the correlating is done by using the same random numbers. It then turns out, in most situations, that every time the γ ray scatters, quite different azimuthal angles are picked in the two problems but the net effect is small and is not affected very sharply.

5. Systematic Sampling

A common method of picking a sample point (x,y) out of a population described by the p.d.f. f(x,y) is to solve the equations[11]

---

11. See Butler's paper.

$$\int_{-\infty}^{x_i} f(x)dx = R_1 \tag{74}$$

$$\int_{-\infty}^{y_i} f(y:x_i)dy = R_2 \tag{75}$$

where R_1 and R_2 are independent random numbers uniformly distributed between zero and one. The expected value of any function of (x_i,y_i) will not be changed appreciably if instead of using N independent R_1's to get the samples, the N numbers $(i - 1/2)/N$, $i = 1, 2, \ldots, N$ are used.[12]

---

12. The purpose of using $(i - 1/2)/N$ is to pick systematically the midpoints of the N intervals defined by

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{1}{N}$$

Always picking the midpoint may introduce biases. These are eliminated if Systematic Sampling and random sampling are combined by using $(i - R_1)/N$ instead of $(i - 1/2)/N$. A different R_1 is used with every i.

If this is done, the estimate is

$$\hat{z}_5 = \frac{1}{N}\sum_{i=1}^{N} z(x_i,y_i) \tag{76}$$

where the x_i are deterministic. The expected value of $\hat{z}_5$ is given by

$$\bar{\hat{z}}_5 = \frac{1}{N}\sum_{i=1}^{N} \bar{z}(:x_i) \tag{77}$$

The $\bar{z}(:x_i)$ are not averaged because the x_i are systematically determined. However, from equation (74) and the use of $(i - 1/2)/N$ instead of R_1,

$$\frac{1}{N} = \int_{x_i}^{x_{i+1}} f(x)dx \tag{78}$$

$$\simeq f(x_i)\Delta x_i$$

Therefore

$$\bar{\hat{z}}_5 \approx \sum_{i=1}^{N} \bar{z}(:x_i) f(x_i) \Delta x_i \tag{79}$$

$$\approx \int \bar{z}(:x) f(x) dx$$

$$= \bar{z}$$

The variance is easily calculated.

$$V_5 = \overline{(\hat{z}_5 - \bar{\hat{z}}_5)^2} \tag{80}$$

$$= \overline{\left\{ \frac{1}{N} \sum_{i=1}^{N} \left[z(x_i, y_i) - \bar{z}(:x_i) \right] \right\}^2}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \overline{\left[z(x_i, y_i) - \bar{z}(:x_i) \right]^2}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2(:x_i)$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sigma^2(:x_i) f(x_i) \Delta x_i$$

$$\approx \frac{1}{N} \overline{\sigma^2(:x)}$$

It follows that (see equation (36))

$$V - V_5 = \frac{1}{N} \overline{\left[\bar{z}(:x) - \bar{z} \right]^2} \tag{81}$$

The variance due to the variation of $\bar{z}(:x)$ has been eliminated.

If it was convenient, in addition to using the numbers $(i - 1/2)/N$ instead of R_1 in equation (74), we could randomly sort a second set of these numbers and use this second set in place of R_2 in equation (75), thereby doing

Systematic Sampling in both the x and y spaces. If this is done, the variance is still further reduced and the improvement due to the use of Systematic Sampling becomes

$$V - V_S = \frac{1}{N}\left\{ \overline{\left[\bar{z}(:x) - \bar{z}\right]^2} + \overline{\left[\bar{z}(:i) - \bar{z}\right]^2} \right\} \qquad (82)$$

where $\bar{z}(:i)$ is defined by

$$\int_{-\infty}^{y_i(x)} g(y:x)dy = \frac{i - \frac{1}{2}}{N}$$

$$\bar{z}(:i) = \int_{-\infty}^{\infty} z\left[x, y_i(x)\right] f(x)dx \qquad (83)$$

$\bar{z}(:i)$ is a sort of analogue in the y space to $\bar{z}(:x)$. Equation (82) indicates that the variance V associated with the ordinary Monte Carlo calculation of an n-dimensional integral must be larger than the sum of n terms of the type $\overline{\left[\bar{z}(:i) - \bar{z}\right]^2}$, one for each dimension. If all the terms are of the same order of magnitude, then doing Systematic Sampling on any one variable will reduce the variance by less than V/n. While this implies that there will be no spectacular gains by doing Systematic Sampling, as already mentioned it is usually costless to use it.

The main application of Systematic Sampling is in those multi-stage problems where it is trivial to calculate the distribution of events at the first stage. In that case, the sampling should be done systematically.

6. Stratified Sampling

Stratified Sampling is a sort of combination of Systematic Sampling and Importance Sampling. One uses an $f^*(x)$ instead of $f(x)$, but does the sampling systematically. That is, the x_i are picked by solving the equation,

$$\int_{-\infty}^{x_i} f^*(x)dx = \frac{i - \frac{1}{2}}{N} \qquad i = 1, 2, \ldots, N \qquad (84)$$

and the y_i are picked at random from $g(y:x_i)$. As one would expect the sample estimates are weighted with the factor $f(x_i)/f^*(x_i)$, i.e.,

$$\hat{z}_6 = \frac{1}{N}\sum_{i=1}^{N} z(x_i,y_i)\,\frac{f(x_i)}{f^*(x_i)} \tag{85}$$

While equation (85) above is formally identical with the corresponding equation (16) in the section on Importance Sampling, the variance of $\hat{z}_6$ is quite different from that of $\hat{z}_{11}$. It is

$$V_6 = \frac{1}{N}\int \frac{\sigma^2(:x)f^2(x)}{f^*(x)}\,dx \tag{86}$$

As a result the optimum $f^*(x)$ is not proportional to $f(x)\sqrt{\overline{z^2}(:x)}$ but is

$$f^*(x) = \frac{\sigma(:x)f(x)}{\int \sigma(:x)f(x)dx} \tag{87}$$

and, $\sigma(:x)$ can be and in general is quite different from $\sqrt{\overline{z^2}(:x)}$. It is easy to find problems in which optimum Importance Sampling and Stratified Sampling are widely different. The reason for this is that in Importance Sampling one is simultaneously estimating the probability of getting into a certain x region and the expected value of $z(x,y)$ given x, while in Stratified Sampling the first quantity has, in effect, been calculated by numerical integration and only the last quantity is being estimated.

If the optimum Stratified Sampling is used then the variance becomes

$$V_6 = \frac{1}{N}\overline{\sigma(:x)}^2 \tag{88}$$

and the reduction in variance is given by

$$V - V_6 = \frac{1}{N}\left\{\overline{\left[\sigma(:x) - \overline{\sigma(:x)}\right]^2} + \overline{\left[\bar{z}(:x) - \bar{z}\right]^2}\right\} \tag{89}$$

In many cases it is convenient to take f^* as close as possible to

$$f^*(x) = \frac{\bar{z}(:x)f(x)}{\bar{z}} \tag{90}$$

The variance then becomes

$$V_{61} = \frac{1}{N}\,\bar{z}\left[\overline{\frac{\sigma^2(:x)}{\bar{z}(:x)}}\right] \tag{91}$$

As can be seen from equation (24), V_{61} is the same as would have been obtained if the sampling on $f^*(x)$ had been done randomly rather than systematically. However, it is usually better to do the sampling systematically because for a general $f^*(x)$ the difference in variance between the two techniques is

$$V_1 - V_6 = \frac{1}{N}\left[\int \frac{\bar{z}^2(:x)f^2(x)}{f^*(x)}\,dx - \bar{z}^2 - \int \frac{\sigma^2(:x)f^2(x)}{f^*(x)}\,dx\right] \qquad (92)$$

$$= \frac{1}{N}\left[\int \frac{\bar{z}^2(:x)f^2(x)}{f^*(x)}\,dx - \bar{z}^2\right]$$

$$\geq 0$$

since $V_1 - V_6$ is the variance associated with doing Importance Sampling on the integral $\int \bar{z}(:x)f(x)dx$. Therefore, where it is easy to use Stratified Sampling, it is preferable to Importance Sampling. In practice, though, the places where it can be used are essentially the same as Systematic Sampling and, as already mentioned, not too plentiful.

The single remark on the applications of Systematic Sampling also applies to Stratified Sampling; it is usually useful when it is trivial to calculate the distribution of events at the first stage. There is only the additional fact that one must have some idea of the relative importance of the various regions where importance, in this case, is measured by $\sigma(:x)f(x)$.

* * *

It is probably clear to the reader that the problems faced by the Monte Carlo experimenter in trying to cut down his statistical fluctuations are quite similar to those that are faced in almost any application of sampling. Therefore, much of the literature of statistics is relevant to the problems we have been considering. In fact, a fairly complete discussion of the last two techniques, and to a slightly lesser extent, the preceding two, can be found in many statistics textbooks; only the first two do not seem to have been discussed. For this reason, it is very valuable to have professional

statistical help in designing these calculations. However, if one has to choose between a person who is mainly interested in statistics and one who is mainly interested in the problem itself, experience has shown that, in this field at least, the latter is preferable. This last remark is not intended as a slur on statisticians, but simply to amplify a comment made earlier, that, "the greatest gains in variance reduction are often made by exploiting specific details of the problem, rather than by routine application of general principles."