A THEORETICAL COMPARISON OF THE EFFICIENCIES OF TWO CLASSICAL METHODS AND A MONTE CARLO METHOD FOR COMPUTING ONE COMPONENT OF THE SOLUTION OF A SET OF LINEAR ALGEBRAIC EQUATIONS

J. H. Curtiss

Institute of Mathematical Sciences
New York University

SUMMARY

In this paper a basis is established for a theoretical comparison of the amount of work required by three widely different methods of calculating (to a given accuracy) one component of the solution of a set of simultaneous linear algebraic equations. The equations are assumed to be in the form $\xi = H\xi + \gamma$, where γ is a given n-dimensional vector and H is a given n x n real matrix. The amount of work is measured by the number of multiplications required under the most unfavorable conditions. The three methods are (a) the Gauss elimination method, (b) the particular stationary linear iterative method defined by the recursion formula $\xi_{N+1} = H\xi_N + \gamma$, $N = 0,1,2,\ldots$, and (c) a Monte Carlo method which consists essentially of a statistical process for estimating the iterates of H. The amount of work required by the first method is proportional to n^3, where n is the order of the matrix H. The amount of work required by the second method to achieve a predetermined accuracy is given by an expression of the form $kn^2 + n$, where k is ordinarily fairly large. The amount of work required by the Monte Carlo method is given by an expression of the form $n^2 + n + b$, where b is ordinarily a very large number. If no preliminary preparations aimed at reducing b are made, then the amount of work for the Monte Carlo method is given by an expression of the form $n + b$.

+This report represents results obtained at the Institute of Mathematical Sciences, New York University, under the sponsorship of the Army Office of Ordnance Research, Contract No. DA-30-069-ORD-1257, RAD Project No. TB2-0001 (1089), Research in the field of Probability, Statistics, and Numerical Analysis (Monte Carlo Method).

The result of this varying dependence on the dimensionality of the problem is that the Monte Carlo method is theoretically more efficient than the other two methods for sufficiently large values of n. The value of n at which one method becomes more efficient than the other depends on the accuracy with which the solution is to be computed.

Upper bounds, which are actually attained in special cases, are derived in this paper for the amount of work required by the iterative and the stochastic methods. From these, break-even points on the range of the dimensionality n are calculated which serve at least as indications of the intervals of values of n which are favorable for each of the three methods. A table in Section 5 gives the favorable numerical intervals of n for various typical specifications of the problem.

A feature of the presentation is the development of a new minimum-variance arrangement of the Monte Carlo method for solving linear equations, which exploits in a simple way an initial estimate of the solution to reduce variance. The construction of the Monte Carlo method will be found in Section 3. In section 6, this Monte Carlo arrangement is adapted to the problem of inverting a matrix. Section 6 also contains derivations and comparisons of certain general linear and polynomial iterative methods for matrix inversion.

I. INTRODUCTION

Many of the problems of numerical analysis to which Monte Carlo methods have been applied belong to the following general type: A rule is given whereby each one of a set of real or complex numbers $a_1, a_2, \ldots$ can be computed. It is required to compute the sum of the series $a_1 + a_2 + \ldots$. (The series may be finite or infinite.)

The standard method of stochastic estimation ("Monte Carlo" method) for this type of problem consists in selecting numbers z_k and p_k, $k = 1,2,\ldots$, such that $z_k p_k = a_k$, $p_k \geq 0$, $\sum_k p_k = 1$. Then a random variable J is set up with the probability distribution given by $Pr(J = k) = p_k$, $k = 1,2,\ldots$.

The random variable z_J will obviously have a theoretical mean value equal to the sum of the series $a_1 + a_2 + \ldots$. The statistic

$$\bar{z}_\nu = \frac{z_{J_1} + z_{J_2} + \ldots + z_{J_\nu}}{\nu}$$

where J_1, J_2, ... are independent replicas of J, furnishes an estimator of the mean value of z_J (that is, of the solution of the problem) which has various well-known optimum statistical properties.

It is hard for some classically trained numerical analysts to see how Monte Carlo methods can ever be advantageous in such a problem. A somewhat over-simplified version of their reasoning might go as follows. Let s be chosen so that $\sum_1^s a_k$ gives a satisfactory approximation to the sum of the series. (If the series is finite, let s be the number of terms.) Since each observed value of z_J conditionally estimates just one of the numbers a_k, there will have to be at least as many terms in the summation for $\bar{z}_\nu$ as in $\sum_1^s a_k$. Indeed, because of statistical fluctuations it will probably be necessary to make very many more observations on z_J than there are terms in the finite approximation $\sum_1^s a_k$, and even then, $\bar{z}_\nu$ will probably not be as good as $\sum_1^s a_k$. This means that the count of additions alone will be very much greater for the stochastic approximation than for the straightforward method of solution; and then too, there is the trouble of setting up the z_k's and p_k's in advance, and of determining stochastically, over and over again, the value of J to use in the observations.

The case against this argument can very conveniently be stated in terms of a specific problem which is fundamental to this paper. It is the problem of calculating the i-th component of the vector $K^{N+1}\theta$, where K denotes the 2n x 2n matrix $[k_{ij}]$ and θ is the 2n-dimensional vector $(t_1,\ldots,t_{2n})$. The i-th element of $K^{N+1}\theta$, which we write as $[K^{N+1}\theta]_i$, is the sum of $(2n)^{N+1}$ terms of the type $k_{ii_1}k_{i_1i_2}\ldots k_{i_Ni_{N+1}}t_{i_{N+1}}$. The stochastic estimation [3]+ is accomplished by selecting numbers z_{ij} and p_{ij}, i,j = 1,...,2n, such that $z_{ij}p_{ij} = k_{ij}$, with $p_{ij} \geq 0$, $\Sigma_j p_{ij} = 1$ for all i. Then a family of random variables $J_o, J_1, J_2, \ldots, J_{N+1}$ is set up in such a way that it represents a Markov process (or "random walk") with states (resting places) designated by 1,2,...,2n and with stationary transition probabilities. The specification of the joint probability distribution is accomplished by assigning an arbitrary distribution to J_o (but with $Pr(J_o=i) \neq 0$), and thereafter using the equations++ $Pr(J_{k+1}=j|J_k=i) = p_{ij}$, i,j = 1,..., 2n. Finally, a random variable

$$Z = Z_{J_oJ_1\ldots J_{N+1}} = z_{J_oJ_1}z_{J_1J_2}\ldots z_{J_NJ_{N+1}}t_{J_{N+1}}$$

is set up. It is now easily seen from the definition of mean value in probability theory, that+++ $E(Z|J_o=i) = [K^{N+1}\theta]_i$. We use $(Z_1+\ldots+Z_\nu)/\nu$ as the estimator of

---

+The square brackets refer to the references at the end of the paper.

++By $Pr(z|b)$ we mean the conditional probability of the event a, given that the event b has occurred.

+++By $E(x|a)$, we mean the mean value of the conditional distribution of the random variable X, given that the event a has occurred.

$K^{N+1}\Theta]_i$, where $Z_1,\ldots,Z_\nu$ denote ν independent determinations of Z.

The various possible values of $k_{ii_1}k_{i_1i_2}\ldots k_{i_N}k_{i_{N+1}}t_{i_{N+1}}$ correspond to the values of a_k in the previous more general formulation. Let us think of these possible values as renumbered in a linear and serial order, using a single index k. There will be $s = (2n)^{N+1}$ such numbers. (They may not be all distinct.) Then the $(2n)^{N+1}$ correspondingly renumbered products $p_{ii_1}\ldots p_{i_Ni_{N+1}}$ play the role of $p_1, p_2,\ldots,p_s$ in the previous formulation, and the various possible values of the vector random variable $\underline{J} = (J_o,J_1,\ldots,J_{N+1})$ correspond to the values of J in the previous formulation.

It now begins to be evident that our formulation of the general summation problem at the beginning of the section was deceptively over-simplified. In a multi-dimensional summation problem, the following two factors may come into play on the side of a Monte Carlo method of solution:

(a) If the calculation of each a_k is a complicated one, it may be possible to arrange things so that the calculation of each observation on z_J is very much simpler than the calculation of the corresponding term a_J. This will in particular be the case if the calculation of each a_k involves the formation of a continued product, because then part of the work of calculating the product can be sidestepped in the stochastic process by using the multiplicative law of probabilities.

(b) Some of the numbers a_k in the finite approximation to Σa_k may be very unimportant and need not be represented in the statistical estimator at all. The stochastic estimation process, if properly set up, will automatically take care of this by making the appearance of the representative of a non-essential term a very rare event.

We add here, more or less as an aside, the remark that when the calculation of each particular z_J is much simpler

than that of the corresponding a_J, then the problem of the accumulation of round-off errors may not be nearly as serious for the stochastic method as it is in the direct computation.

All of these factors favoring the Monte Carlo method are present in the case of the problem $K^{N+1}\theta]_i$ and in the related problem of solving systems of linear algebraic equations. The factor (a) is usually particularly in evidence in numerical problems of stochastic origin, and indeed it is in such problems that the Monte Carlo method has had its chief successes. Matrix problems can always be thrown into a form in which they are numerically equivalent to problems with a stochastic pedigree. We shall exploit that fact in the present paper to obtain a favorable environment for the comparisons to be made.

But we cannot conclude this introduction without making a remark which is on the negative side as far as Monte Carlo methods are concerned. It certainly would seem that whenever Monte Carlo methods appear to advantage in summation problems, factor (b) above must be playing an important role, because otherwise the criticisms regarding the necessary number of addition operations required for any reasonable degree of accuracy in the Monte Carlo approach would be valid. But if that is so, why cannot a deterministic method be devised which will ignore the unimportant terms and be even more efficient? The author suspects that what we now need is a more highly developed deterministic theory of quadrature and of linear computation in many dimensions. When this becomes available, the Monte Carlo method may, at least for matrix problems, lose the very modest advantages which will be claimed for it in this paper.

II. THE NUMERICAL PROBLEM AND ITS NON-STOCHASTIC SOLUTION

Throughout the paper we shall continue to denote matrices by capital letters, and their elements by the corresponding lower case letters; thus, for example, $H = [h_{ij}]$. We represent vectors by lower case Greek letters and their components by the corresponding Roman letters; thus for

example, $\xi = (x_1, x_2, \ldots, x_n)$. We shall also find it convenient occasionally to designate the elements of a matrix by double subscripts affixed to the symbol for the matrix (thus, H_{ij} or $\left[(I-H)^{-1}H^2\right]_{ij}$). Furthermore we shall frequently designate the components of a vector by a similar subscript notation (thus $\xi = (\xi]_1, \xi]_2, \ldots, \xi]_n$). All vectors will be real and n-dimensional and all matrices will have only real elements.

By $\| H \|$ we shall mean $\max_i \sum_j |h_{ij}|$, and by $\| \xi \|$, we shall mean $\max_i |x_i|$. It is obvious that $\| H\xi \| \leq \| H \| \, \| \xi \|$ and that $\| A+B \| \leq \| A \| + \| B \|$. It is well-known+ that $\| AB \| \leq \| A \| \, \| B \|$; therefore, by induction $\| H^N \| \leq \| H \|^N$.

The numerical problem with which we shall be mainly concerned is that of solving the linear system

$$A\xi = \eta, \tag{2.1}$$

for ξ, where A is a given non-singular n x n matrix and η is a given vector. We assume that this system has been thrown into the form

$$\xi = H\xi + \gamma, \tag{2.2}$$

where $H = [h_{ij}]$ is an n x n matrix with the property that

$$\| H \| = \max_i \sum_j |h_{ij}| < 1 . \tag{2.3}$$

+See for example Courant and Hilbert [1]; p. 16, footnote.

It is beyond the scope of this paper to give an extended discussion of the methods of passing from (2.1) to (2.2), but a few general remarks on the subject are in order at this point. In the first place, it is always theoretically possible to transform the problem represented by (2.1) in this manner. Indeed, let H be any matrix whatsoever with the property (2.3). (For instance, let $H = dI$, where I is the unit matrix and d is a scalar lying between 0 and 1.) It is known that if H satisfies (2.3), then I-H cannot be singular[+]. Let M be defined by the equation $I - MA = H$. This says that $M = (I-H)A^{-1}$. Therefore M cannot be singular. The system

$$\xi = H\xi + M\eta = (I-MA)\xi + M\eta,$$

is just the same as the system

$$MA\xi = M\eta,$$

and since M is non-singular, this system is precisely equivalent (2.1).

But in practice it is not feasible to set up an arbitrary H satisfying (2.3), and then to determine M as above. The reason is that the formula $M = (I-H)A^{-1}$ presupposes a knowledge of A^{-1}, and in the presence of this the original problem becomes trivial. Thus it is natural to think of M as being chosen first, the choice being made in such a way that $I - MA = H$ has suitable properties. There are a number of different procedures in the literature of linear computation for arriving at an appropriate choice of M. For example, if A has dominant diagonal--that is, if $|a_{ii}| > \sum_{j\neq i}^{n} |a_{ij}|$, $i = 1,\ldots,n$,--then M can be chosen as the inverse of the principal diagonal matrix whose principal diagonal is that of A. This will obviously insure that $\| H\| < 1$[++].

[+]O. Taussky-Todd, [9].

[++]Further discussion will be found in any good treatise on numerical analysis which deals with iterative methods of solving linear equations. See for example Householder [7] and Milne [8]. See Forsythe [6] for further references.

There are numerous non-stochastic methods of solving (2.1)+. We shall here restrict our considerations mainly to a class of methods known as linear iterative processes. An effective Monte Carlo method for the problem (2.1) can be based on this type of process, as we presently shall see.

The general stationary linear iterative process for solving (2.1) is arrived at by throwing (2.1) into an equivalent form which has the appearance of (2.2), but with H restricted only by the requirement that its eigenvalues all lie in the unit circle. An initial estimate ξ_0 of the solution is made, and successive approximations $\xi_1, \xi_2, \ldots$, are then defined by

$$\xi_{N+1} = H\xi_N + \gamma, \quad N = 0,1,2,\ldots \qquad (2.4)$$

If ξ_∞ denotes the solution of the equations (2.2), then clearly, since $\xi_\infty = H\xi_\infty + \gamma$,

$$\begin{aligned} \xi_\infty - \xi_N &= H(\xi_\infty - \xi_{N-1}) \\ &= H^2(\xi_\infty - \xi_{N-1}) \\ &= H^N(\xi_\infty - \xi_0) \end{aligned} \qquad (2.5)$$

Thus the condition for convergence for any starting vector ξ_0 is that $\lim_{N\to\infty} H^N = 0$. The well-known necessary and sufficient condition [7] for $\lim_{N\to\infty} H^N = 0$ is that all the eigenvalues of H should lie in the unit circle, which explains the requirement on H imposed earlier in this paragraph.

But in the present discussion, we go one step beyond this requirement and insist that $\| H\| < 1$. Since $\| H^N \| \leq \| H \|^N$, this condition will certainly insure that $H^N \to 0$.

+See previous footnote for references.

For purposes of error analysis[+] it is advantageous now to introduce the residual vectors

$$(2.6)\qquad \rho_N = \gamma - (I-H)\xi_N ,$$
$$= H\xi_N + \gamma - \xi_N = \xi_{N+1} - \xi_N, \quad N = 0,1,2,\ldots$$

The vectors ρ_N are of course always computable at any stage in the iterative solution. If $\xi_N \to \xi_\infty = (I-H)^{-1}\gamma$, then obviously $\rho_N \to 0$. The converse is also true, because $(I-H)^{-1}\rho_N = \xi_\infty - \xi_N$. Thus ρ_N, or $\|\rho_N\|$, are logical measures of the error in the N-th approximation to the solution. It is to be noted from (2.4) and (2.6) that

$$(2.7)\qquad \rho_N = \xi_{N+1} - \xi_N = (H\xi_N + \gamma) - (H\xi_{N-1} + \gamma)$$
$$= H(\xi_N - \xi_{N-1}) = H\rho_{N-1}$$
$$= \ldots \qquad = H^N\rho_0 .$$

From (2.6) and (2.7) it follows that the successive approximations ξ_N generated by (2.4) can theoretically also be generated in the following manner: Select ξ_0 as before and compute ρ_0 from the definition of residual vector, $\rho_0 = \gamma - (I-H)\xi_0$. Then conduct the iterations by means of the pair of formulas

$$(2.8)\qquad \xi_{N+1} = \xi_N + \rho_N , \quad N = 0,1,\ldots ,$$

$$(2.9)\qquad \rho_{N+1} = H\rho_N , \quad N = 0,1,\ldots .$$

We note that by back substitution, we find that

[+]By error in this paper, we shall mean truncation error or statistical error or both at once. There will be no study of round-off error, nor of the effect of miscellaneous arithmetical mistakes.

$$(2.10)\quad \xi_{N+1} = \xi_0 + \rho_0 + \rho_1 + \ldots + \rho_N = \xi_0 + (I + H + \ldots + H^N)\rho_0 .$$

In actual practice, the customary running check on the accuracy of the solution consists in computing $\|\rho_N\|$ from time to time to see how small it is. The iterations are stopped when $\|\rho_N\|$ reaches a predetermined order of smallness. If (2.4) were used for the iterations, the computation of ρ_N would be done by using the formula $\rho_N = \xi_{N+1} - \xi_N$. If (2.8) and (2.9) were to be used, it would be advisable to compute test values of $\|\rho_N\|$ from the definition of ρ_N (that is, from the formula $\rho_N = \gamma - (I-H)\xi_N$), rather than to accept the values of $\|\rho_N\|$ given by (2.9). We shall come back to this point in a moment.

An <u>a priori</u> truncation error analysis, made with the purpose of estimating the number of iterations which will be required to achieve a given accuracy, can be conducted either in terms of ρ_N or in terms of the error vector $\xi_\infty - \xi_N$. If the size of $\|\rho_N\|$ is to be the criterion, then we use (2.7) to obtain the very simple estimate

$$(2.11)\quad \|\rho_N\| \leq \|H\|^N \|\rho_0\| .$$

But if the deviation of ξ_N from ξ_∞ seems to be a more appropriate or convenient measure of the truncation error, then we use the fact that $(I-H)^{-1}\rho_0 = \xi_\infty - \xi_0$. Substituting this into (2.5), we find that

$$(2.12)\quad \xi_\infty - \xi_N = H^N(I-H)^{-1}\rho_0 .$$

Since $(I-H)^{-1} = I + H + H^2 + \ldots$, it follows that $\|(I-H)^{-1}\| \leq \|I\| + \|H\| + \|H\|^2 + \ldots = (1 - \|H\|)^{-1}$. Therefore

$$(2.13)\quad \|\xi_\infty - \xi_N\| \leq \frac{\|H\|^N}{1 - \|H\|}\|\rho_0\| .$$

It is perhaps worthwhile to point out here, by way of an aside, that the inequalities (2.11) and (2.13) hold for any one of the various matrix and vector norms in common use+ , and not just for the norm $\|H\| = \max_i \sum_j |h_{ij}|$. We are using this particular norm here because of a special application it has to the Monte Carlo method, which will be brought out in the next section.

The iterative formulas (2.8) and (2.9) are (so to speak) homogeneous, and therefore look easier to use than (2.4). But (2.8) and (2.9) have the great disadvantage of being not self-correcting in case a mistake is made at one stage, whereas (2.4) does have this property. Suppose for example that for some N, ρ_{N+1} is mistakenly computed as a zero vector in using (2.9). Then since all subsequent vectors ρ_N will equal zero, it is obvious from (2.10) that ξ_N will be irretrievably wrong. But if a mistake is made in computing ξ_{N+1} by (2.4), the subsequent effect is like starting over again with another ξ_0.

Therefore we are not proposing (2.8) and (2.9) as a practical substitute for (2.4) for a non-stochastic numerical solution of the problem $\xi = H\xi + \gamma$. Our real purpose in introducing the alternative iteration formulas was to develop the representation of ξ_N given by formula (2.10). This representation seems to be an advantageous one upon which to construct a Monte Carlo solution, as we shall see in the next section.

It is not without interest, however, to point out that if the amount of work involved in a computing job is measured in any sensible way, it theoretically requires no more work to use (2.8) and (2.9) up to a predetermined value of N than to use (2.4). This assumes that check values of ρ_N will not have to be computed from the definition of residual vector from time to time in the use of (2.8) and (2.9). In this paper, we shall measure amount of work by merely counting up the number of multiplications required under the worst conditions, assuming no zero or unit elements. Additions and subtractions will be ignored. A division or reciprocation will count as a single multiplication. To arrive at ξ_N by

---

+See Householder [7], pp. 38-44.

(2.4), starting with ξ_0, requires n^2 multiplications per iteration, or Nn^2 in all. To arrive at ξ_N by (2.8) and (2.9) without computing any intermediate or final check values of ρ_N takes n^2 multiplications for ρ_0, and thereafter $(N-1)n^2$ multiplications for each iterative determination of ρ_N. So once again the count is Nn^2.

Actually, in later sections we are going to set up an error analysis which implies that ρ_0 will always have to be computed. The use of (2.4) will be tacitly assumed, so n^2 extra multiplications will have to be added to the total count.

III. STOCHASTIC SOLUTION OF THE PROBLEMS

It should be apparent from the foregoing section that even if one restricts oneself to the class of stationary linear iterative processes, there is literally an uncountably infinite number of methods for solving the problem $A\xi = \eta$, because of the different possible choices of H or M. An even more disconcerting fact than this was implicit in the discussion in the foregoing section. It can be expressed in the form of a theorem: "Given any one way of solving $A\xi = \eta$, there is always a much better way." For given any one H, a happier choice of H from the standpoint of the error analyses given by (2.11) and (2.12) always exists.

Under such circumstances it is evident that some strict ground rules are required if meaningful comparisons are to be made between methods. This statement applies even to comparisons within very special classes of classical methods, and it is especially relevant when an utterly unorthodox method such as the Monte Carlo Method is to be brought into the picture.

We propose then to adhere to the following set of rules:

(1) It will be assumed that the problem is given in the form $\xi = H\xi + \gamma$, with $\|H\| < 1$.

(2) The primary comparison will be between a Monte

Carlo Method for solving $\xi = H\xi + \gamma$ and a stationary linear iterative method of the type described in the preceding section. Both methods will be based on the particular H given in (1). The linear iterative method will be defined by the recursion relations $\xi_{N+1} = H\xi_N + \gamma$, $N = 0,1,\ldots$, or alternatively, by the formula

$$\xi_N = \xi_0 + (I+H + \ldots + H^{N-1})\rho_0, \tag{3.1}$$

where ξ_0 is the initial estimate and ρ_0 is the initial residual vector (see equation 2.10). The Monte Carlo Method will consist in effect of a statistical estimation of

$$\xi_\infty = \xi_0 + (I + H + H^2 + \ldots)\rho_0, \tag{3.2}$$

using the same H, the same ξ_0, and same ρ_0 as in (3.1).

(3) No speculation will be permited as to the existence of a better H on which to base either the stochastic or the non-stochastic method.

(4) The measure of approximation used in the case of the iterative method will be $\|\xi_\infty - \xi_N\|$. The measure of approximation used in the stochastic method will be $|\xi_\infty]_i - \bar{Z}_\nu|$, where $\xi_\infty]_i$ denotes the i-th component of the solution vector ξ_∞, and $\bar{Z}_\nu$ is its statistical estimator, based on a sample of size ν.

(5) To furnish some contact with the large family of alternative methods of solving $\xi = H\xi + \eta$, a simple direct method of solution will be brought into the comparison. For simplicity, it will be assumed that this direct method gives the exact solution. Round-off error will be ignored. The direct method which we select is the Gauss Elimination Method, because for present purposes it seems to be as good as any other and better than most.

(6) As previously stated in Section II, the amount of work required for a computation will be measured only by the

number of multiplications required in the worst cases, counting a reciprocation or division as one multiplication. In counting multiplications, the possibility of unit or zero factors is not taken into account.

(7) It will be assumed that the problem is to find only one component of the solution of $\xi = H\xi + \gamma$.

It is recognized freely that the last restriction on the comparison is a strange one. It is made because the question of efficient Monte Carlo estimation of all components of the solution simultaneously has not yet been adequately investigated. Of course, separate statistically independent estimations can be made for each of the n components of the solution. This would multiply the measure of work which we shall derive for the Monte Carlo method[+] by a factor of n. At the same time, for a given sample size N, the probability that all estimations fall within preassigned limits of error, will be smaller than it is for the estimation of a single component[++]. Therefore N should be correspondingly increased. But it is almost surely inefficient to use separate independent estimators for each component of the solution. It seems intuitively clear that data obtained in the course of estimating one component should be used again for other components[+++].

With these preliminary comments out of the way, we proceed to set up the Monte Carlo estimation of $\xi_\infty]_i$.

The standard method of estimating $K^{N+1}\theta]_i$, where $K = [k_{ij}]$ is a given 2n x 2n matrix and $\theta = (t_1,\ldots,t_{2n})$ is a 2n-dimensional vector, has already been described in Section I. Here we recapitulate it briefly. Numbers z_{ij}

[+]That is, the measure of the work for the purely stochastic part of the solution. This is represented by the third quantity in the sum on the right side of (4.5), or of (4.6), in Section IV, below.

[++]The question involved here is that of the distribution of the extreme absolute value of n normally distributed independent random variables with zero means and differing variances.

[+++]The re-use of samples to estimate various components simultaneously is discussed briefly in [4].

and p_{ij}, $i, j=1,\ldots,2n$, are chosen such that $z_{ij}p_{ij} = k_{ij}$, with $p_{ij} \geq 0$, $\sum_j p_{ij} = 1$. A Markov process with states $1,2,\ldots,2n$, and with the matrix $[p_{ij}]$ as its matrix of transition probabilities is set up. Let $J_o, J_1, \ldots, J$ be a family of random variables which represent the process, in the sense that $Pr(J_{k+1}=j \mid J_k=i) = p_{ij}$, $i,j = 1, \ldots,2n$. The random variable $Z = z_{J_o J_1} z_{J_1 J_2} \cdots z_{J_N J_{N+1}} t_{J_{N+1}}$ has the property that $E(Z \mid J_o=i) = K^{N+1}\theta]_i$.

Consider now the 2n x 2n matrix

$$K = \left[\begin{array}{c:c} H & I \\ \hdashline O & I \end{array}\right] , \tag{3.3}$$

and the vector

$$\theta = \left[\begin{array}{c} O \\ \hdashline \rho_o \end{array}\right] , \tag{3.4}$$

where H is the matrix of the equation $\xi = H\xi + \nu$, I is the n x n matrix and ρ_o is the residual vector corresponding to the initial estimate ξ_o of the solution of these equations. (That is, $\rho_o = \gamma - (I-H)\xi_o$.) Then it is easily shown that

$$K^{N+1} = \left[\begin{array}{c:c} H^{N+1} & I+H+\ldots+H^N \\ \hdashline O & I \end{array}\right] .$$

Therefore

$$K^{N+1}\theta = \left[\begin{array}{c} (I+H+\ldots+H^N)\rho_o \\ \hdashline \rho_o \end{array}\right] .$$

Our Monte Carlo solution of the equations $\xi = H\xi + \gamma$ consists of statistically estimating the i-th component of this vector $K^{N+1}\theta$, adding this statistical estimate to $\xi_o]_i$, and with reference to (2.10) or (3.1), using this sum to approximate thereby the i-th component of the solution vector ξ_∞. More specifically, we set up the numbers z_{ij} and p_{ij} for the special matrix K appearing in (3.3). For each sample random walk, represented by a determination of the vector random variable $J_o, J_1, \ldots, J_{N+1}$, made with $J_o = i$, we compute the statistic

$$\xi_0]_i + Z = \xi_o]_i + z_{J_o J_1} z_{J_1 J_2} \cdots z_{J_N J_{N+1}} t_{J_{N+1}} ,$$

in which $t_{J_{N+1}}$ is the J_{N+1}-th component of the vector θ given by (3.4). The conditional mean value of this statistic, given that $J_o = i$, is $\xi_o]_i + (I+H+\ldots+H^N)\rho_o]_i$, or $\xi_{N+1}]_i$.

The statistical estimation of ξ_{N+1} is accomplished by taking the average of ν independent determinations of the random variable $\xi_o]_i + Z$, which we denote by $\xi_o]_i + Z_1$, $\xi_o]_i + Z_2, \ldots, \xi_o]_i + Z_\nu$. This average takes the form

$$\bar{Z}_\nu = \xi_o]_i + \frac{Z_1+Z_2+\ldots Z_\nu}{\nu} .$$

Of course, $\bar{Z}_\nu$ directly estimates or approximates $\xi_{N+1}]_i$ and not the solution component $\xi_\infty]_i$. But we now eliminate the truncation error completely from consideration in the Monte Carlo solution by assuming that N, although finite, is so large that $\| \xi_\infty - \xi_{N+1} \|$ is completely negligible. From (2.12) it is obvious that this can always be done. For all practical purposes then, $\bar{Z}_\nu$ will be an estimator directly of

$$\xi_\infty]_i = \lim_{N\to\infty} K^N\theta]_i = \xi_o]_i + (I+H+H^2+\ldots)\rho_o]_i .$$

This is the i-th component of the vector which appears in (3.2).

We shall now discuss the choice of the numbers z_{ij} and p_{ij} with reference to the special matrix K now under consideration. Obviously it is necessary to choose z_{ij} and p_{ij} so that $z_{ij}p_{ij} = h_{ij}$, $i,j = 1,\ldots,n$, and $z_{i,i+n}p_{i,i+n} = 1$. Moreover, for $i > n$, $z_{ij}p_{ij} = 1$ if $i = j$ and otherwise $z_{ij}p_{ij} = 0$.

Within these limitations, there are of course an infinite number of possible choices of the numbers z_{ij} and p_{ij}. It seems likely from evidence of various types that an optimum choice, or at least a near-optimum choice, in the present instance consists in letting $p_{ij} = |h_{ij}|$, $i,j = 1,\ldots,n$, $p_{ij} = 0$, $i = 1,\ldots,n$, $j = n+1, n+2, \ldots, n+i-1, n+i+1, \ldots,2n$, $p_{ij} = 0$, $i > n$, $j \neq i$ and $p_{i,i+n} = 1 - \sum_{j=1}^{n} p_{ij}$, $i = 1,\ldots,n$, $p_{ij} = 1$, $i > n$. We must defer a complete discussion of this choice of the numbers p_{ij} to another paper which is now under preparation.

It is to be noticed that $\sum_{j=1}^{n} p_{ij} \leq \|H\| < 1$, $i = 1,\ldots,n$. With this choice of the numbers p_{ij} it follows that $z_{ij} = \pm 1$ $i,j = 1, \ldots, n$, and $z_{i,i+n} = 1/p_{i,i+n}$ for $i = 1,\ldots,n$. We henceforth shall usually drop the subscript on ρ_o and write $\rho = (r_1,\ldots,r_n)$. Then

$$Z = z_{J_oJ_1} z_{J_1J_2} \cdots z_{J_NJ_{N+1}} t_{J_{N+1}}$$

$$(3.3) \qquad = \begin{cases} 0 , & J_{N+1} = 1,2,\ldots,n \\ \pm \dfrac{r_{J_{N^+-1}}}{p_{J_{N^+-1}, n+J_{N^+-1}}} , & J_{N+1} = n+1,\ldots,2n , \end{cases}$$

where N^+ is the "duration" of the random walk represented by J_o, J_1, ... ; that is, N^+ is the number of times a state in the first n states is visited before any state in the last n states is visited.[+] (The present set-up of the Markov process makes the last n states play the role of absorbing states.)

It is to be noted at this point that

$$(3.4) \qquad |Z| \leq \frac{\|\rho\|}{\min_i p_{i,i+n}} = \frac{\|\rho\|}{1 - \max_i \sum_{j=1}^{n} p_{ij}}$$

$$= \frac{\|\rho\|}{1 - \|H\|} \quad .$$

This means that $E(Z^2|J_o = i)$ exists and is uniformly bounded for all values of N.

Let v denote the conditional variance of the random variable Z, relative to the hypothesis that $H_o = i$. This is a measure of dispersion of Z defined by $v = E\{[Z-E(Z)]^2|J_o=i\} = E(Z^2|J_o=i) - (\xi_o]_i)^2$. It is necessary for later developments to obtain an appraisal of v. The explicit formula for v is known[++], but in the present situation a rough method of appraisal which bypasses the formula will give just as good a bound for v as can be obtained from the explicit formula for v.

[+]We shall here count in the first state--the state from which the random walk starts--in computing N^+. Thus if 4 non-absorbing states are visited including the starting point and then absorption takes place, then $N^+ = 4$, and J_3 is the last one of the J's taking on one of the values 1, 2, ..., n, and J_4 is equal to $n + J_3$. This convention concerning N^+ is adopted so as to conform with the definition given in Curtiss [3], and so as to simplify later formulas slightly.

[++]See [3], p. 223.

The rough method is this. Given any random variable X distributed on the interval (-a,a), it obviously follows from the mean-value definition that $E(X^2) \leq E(a^2) = a^2$ and $E[X-E(X)]^2 \leq a^2 - [E(X)]^2$. Thus the highest value that the variance of such a random variable can have is a^2. But if the random variable has a uniform distribution on this interval, direct computation reveals that the variance is only $a^2/3$. If the distribution is somewhat bell-shaped, the variance may be much less, with zero as the greatest lower bound. Therefore, as a rough approximation, we shall in the present instance take the variance to be not greater than $a^2/2$. That is, our appraisal of v will be

$$v \leq \frac{1}{2} \frac{\|P\|^2}{(1 - \|H\|)^2} , \tag{3.5}$$

where the right side is obtained by referring to (3.4)+.

Another appraisal which will be needed relates to the mean value of the duration of N^+. It is known++ that

$$E(N^+ | J_o=i) = \sum_{j=1}^{n} (I+P+P^2+\ldots+P^N)_{ij} ,$$

where $P = [|h_{ij}|]$.

Therefore,

$$\begin{aligned} E(N^+ | J_o=i) &\leq \max_i \sum_{j=1}^{n} (I+P+P^2+\ldots)_{ij}, \\ &\leq \|I\| + \|P\| + \|P\|^2 + \ldots \\ &= \frac{1}{1 - \|P\|} . \end{aligned}$$

+If the reader prefers to work with a bound which is one hundred per cent certain not to be exceeded, he will have to comb through the remaining calculations in this paper and replace the factor 1/2 by unity wherever (3.5) is used. There are enough safety factors in our estimates, insofar as avoiding the favoring of the Monte Carlo method is concerned, so that this ought to be unnecessary.

++See Curtiss [3], p. 226.

This formula holds good for any Markov process with absorbing states and with stationary transition probabilities given by a matrix such as P. In the present case, $P = [|h_{ij}|]$, so our upper bound for the mean deviation is

$$(3.6) \qquad E(N^+|J_o = i) \leq \frac{1}{1 - ||H||} .$$

It should be noted that (3.6) becomes an equality if the sum of the elements of the i-th row of $P = [|h_{ij}|]$ is constant for $i = 1,\ldots,n$.

Incidentally, the reason for using the matrix norm $\max_i \sum_j |h_{ij}|$ instead of one of the other norms should now be apparent. The natural appraisals for both v and the mean duration seem to involve this particular norm.

The conditional variance of N^+, given that $J_o = i$, has the following bound[+] in the case $N = \infty$:

$$(3.7) \qquad \mathrm{Var}(N^+|J_o=i) \leq \frac{2}{1-||P||} - \left\{1+E(N^+|J_o=i)\right\} E(N^+J_o=i)$$

$$< \frac{2}{1-||P||} - \left\{E(N^+|J_o=i)\right\}^2 .$$

In view of certain safety factors in this formula, we shall accept the following simpler heuristic appraisal of the variance, obtained by discarding the second term in the third member of (3.7) and halving the first term:

$$(3.8) \qquad \mathrm{Var}(N^+|J_o=i) \leq \frac{1}{1 - ||H||} .$$

[+]See Curtiss [3], p. 226.

Thus our appraisal of the conditional variance of the duration is identical with our appraisal of the conditional mean value of the duration.

In concluding this section, we shall make two general remarks about the Monte Carlo method for solving $\xi = H\xi + \gamma$ developed above.

In the first place, one of the desirable features of any Monte Carlo solution is to achieve an arrangement whereby the more that is known about the solution of the problem in advance, the smaller the variance of the statistical estimator is, with zero variance attained in the presence of full knowledge of the solution. Such an arrangement has been achieved in the present case. If the solution is known in advance, then $\rho_0 = \rho = 0$, and consequently $v = 0$. The inequality (3.5) gives a bound for v which depends on the square of the norm of the zero-th residual vector, and thus the better the initial estimate or guess is, the smaller the variance is[+].

In the second place, our Monte Carlo solution has an automatic self-correction feature similar to that of the iterative method based on (2.4). If an error is made in computing the Z for any one sample walk, this erroneous Z merely is incorporated into the average of a great many other realizations of the same random variable, and its effect will ordinarily be negligible.

IV. THE A PRIORI ESTIMATION OF THE REQUIRED AMOUNT OF WORK

In the fourth of the ground rules stated at the beginning of Section III, we announced that the measure of approximation to be used in the case of the iterative method would be $||\xi_{oo} - \xi_N||$, and in the case of the stochastic method it would be $|\xi_{\infty}]_i - \bar{Z}_\nu|$. We shall now state more explicitly just how we are going to use these measures of approximation.

The general idea is that in each case, the computation is to proceed until the approximate solution is suitably close to the exact solution, and the definition of "suitably

---

[+] Another minimum variance Monte Carlo solution of the problem $A\xi = \eta$ is presented in Curtiss [3], pp. 227-231. The present arrangement seems to be simpler and somewhat easier to use in practice.

close" used in each case will be comparable. Specifically, given a small number $d > 0$, we propose that the iterations in the non-stochastic method shall be carried on until finally $\| \xi_\infty - \xi_N \| < d$, and we propose that the sampling of the Markov process shall be continued until finally $|\xi_\infty|_i - Z_\nu| < d$.

But the vector ξ_∞ is unknown. We must therefore translate our measures of error into terms of the data of the problem and the initial estimate ξ_0. To achieve a theoretical rather than empirical comparison, we shall restrict ourselves entirely to an a priori error analysis.

The error analysis and consequent appraisal of the amount of work required to achieve a given accuracy, is of necessity carried out very differently in the case of the two methods. In the case of the non-stochastic method, we base the analysis on the inequality (2.13). With an eye on this inequality, we seek the lowest value of N such that

$$\frac{\| H \|^N}{1 - \| H \|} \| \rho \| < d \quad .$$

Taking logarithms of both sides, we find that the required value of N is

$$N_0 = 1 + \left[\frac{\log \frac{d}{\| \rho \|} + \log (1 - \|H\|)}{\log \| H \|} \right] \tag{4.1}$$

where the square bracket here means "largest integer in." (The logarithms can be taken to any convenient base, as for example, 10.)

We must therefore carry out N_0 iterations of the recursion formula $\xi_{N+1} = H\xi_N + \gamma$. As pointed out at the end of Section II, each iteration counts as n^2 multiplications. However, since we have set ourselves the peculiar

problem of finding only one component of the solution vector, the last iteration (in which ξ_{N_o} is computed from ξ_{N_o-1}) can be abbreviated to just n multiplications. The formula for N_o involves $\|\rho\|$, and it seems reasonable to suppose therefore that in using this error analysis in practice, $\rho = \rho_o$ would always be computed at the outset. This would take n^2 more multiplications. Thus the grand total of the number of multiplications required a priori to achieve the inequality $\|\xi_\infty - \xi_N\| < d$ is

$$(4.2) \qquad m = (N_o-1)n^2 + n + n^2$$

$$= n^2 + n + n^2 \left[\frac{\log \frac{d}{\|\rho\|} + \log(1-\|H\|)}{\log \|H\|} \right].$$

We now attack the analogous problem for the stochastic method.

The statistical estimator $\bar{Z}_\nu$ is given by the formula

$$\bar{Z}_\nu = \xi_o]_i + \frac{Z_1+Z_2+\ldots+Z_\nu}{\nu},$$

where $Z_1, Z_2, \ldots, Z_\nu$ are ν mutually independent determinations of the random variable which appears in the right member of (3.3). The mean value of $\bar{Z}_\nu$ is $\xi_\infty]_i$ for all practical purposes. It will not be possible to adjust ν so as to assure ourselves a priori, given any $d > 0$ however small, that $\bar{Z}_\nu$ will deviate from its mean value by less than d. We must therefore have recourse to the theory of statistical estimation.

Probably the easiest way to approach the question is to demand that a priori, the probability of a deviation of less than d shall be at or above some predetermined rather high level. Specifically, we choose a small number p, and require that ν shall be taken as the lowest value for which, a priori, the following inequality holds:

$$\Pr(|\xi_{\infty}]_i - \bar{Z}_\nu| < d) > 1 - 2p \quad .$$

Now $\bar{Z}_\nu$ is a constant plus the average of ν independent, identically distributed random variables, each with a finite variance v. It therefore follows from a well-known result in probability theory called the Central Limit Theorem[+] that $(\bar{Z}_\nu - E(\bar{Z}_\nu))/v_\nu^{1/2}$ is approximately distributed according to the normal or Gaussian distribution, where v_ν denotes the conditional variance of $\bar{Z}_\nu$, given that $J_o = i$. The approximation is ordinarily very good for $\nu > 100$, and in all of our subsequent applications of this theorem we shall be dealing with values of ν much greater than this. At worst, the effect of a poor approximation would be merely to deceive us by a few one hundredths as to the value of the probability level p which is really in effect.

The variance of a constant plus a random variable is the same as the variance of the random variable alone. Therefore the variance v_ν of $\bar{Z}_\nu$ is equal to that of the average of ν independent determinations of the random variable Z. A familiar formula of statistical theory[++] then states that $v_\nu = v/\nu$, where as in Section III, v is the conditional variance of Z, given that $J_o = i$.

Putting the above facts together, we have:

$$\begin{aligned}
&\Pr(|\xi_{\infty}]_i - \bar{Z}_\nu| < d) \\
&= \Pr\left(\frac{|E(\bar{Z}_\nu) - \bar{Z}_\nu|}{v_\nu^{1/2}} < \frac{d}{v_\nu^{1/2}}\right) \\
&= \Pr\left(\frac{|E(\bar{Z}_\nu) - \bar{Z}_\nu|}{v_\nu^{1/2}} < \frac{d\nu^{1/2}}{v^{1/2}}\right) \\
&\doteq \int_{-d\nu^{1/2}/v^{1/2}}^{d\nu^{1/2}/v^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \\
&= 1 - 2\int_{d\nu^{1/2}/v^{1/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \quad .
\end{aligned}$$

[+]See e.g. [2], Chap. 17. [++]See e.g. [2] p. 345.

At this point we shall make an arbitrary decision about the level of certainty 1 - 2p which is to be demanded. If x = 2, then

$$2 \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = .0455 .$$

A level of certainty equal to 1 - 0.0455 = .9545 seems more than adequate for present purposes, in view of various other safety factors which are embodied in our appraisals.

Thus we are seeking the smallest value of ν such that

$$\frac{d\,\nu^{1/2}}{v^{1/2}} > 2 .$$

This value of ν is

$$\nu_o = \left[\frac{4v}{d^2}\right] + 1 , \tag{4.3}$$

where again [ ] means "largest integer in."

This is the number of independent sample realizations of the Markov process needed to achieve the demanded order of accuracy with a probability of at least 95.45%.

To estimate the amount of work required to attain this level of accuracy, we must make some further assumptions as to how the Monte Carlo computations will be carried out. From (3.3), we see that each sample will (almost surely) involve computing some one of the numbers $r_i/p_{i,i+n}$, $i = 1,\ldots,n$. It seems logical therefore to assume that these quantities will all be calculated in advance. It will require n^2 multiplications to compute ρ_o, given ξ_o, and thereafter it will require n multiplications to get the quotients $r_i/p_{i,i+n}$.

We must now come to an agreement as to how much work is involved in following each random walk $J_0, J_1, \ldots$ to absorption. It seems to be not unreasonable to assume that each step before absorption, and the step in which absorption takes place, always requires the equivalent of one multiplication. Of course, after absorption takes place in one of the states numbered n+1, n+2, ..., 2n, no more computations are required for the particular realization of the Markov process at hand, and a new independent sample is started. In other words, each complete random walk, represented by $J_0 \to J_1 \to J_2 \to \ldots \to J_{N^+-1} \to J_{N^+} = J_{N^+-1} + n$, will require N^+ multiplications.

In ascribing to each step the equivalent of one multiplication, we have in mind the fact that to select the value of J_{k+1}, given J_k, a pseudo-random number will presumably be generated and certain comparison operations will have to be performed.

Putting these assumptions together, we find that the total amount of work required, measured in multiplications, is

$$n^2 + n + N_1^+ + N_2^+ + \ldots + N^+_{\nu_0} , \tag{4.4}$$

where $N_1^+, \ldots, N^+_{\nu_0}$ are ν_0 independent determinations of the random variable N^+ introduced in Section III. This is a random variable whose (conditional) mean value is $n^2 + n + \nu_0 E(N^+ | J_0=i)$ and whose (conditional) variance is $\nu_0 \operatorname{Var}(N^+ | J_0=i)$.

In Section III (e.g. 3.7) we arrived at an appraisal of the magnitude of $\operatorname{Var}(N^+ | J_0=i)$ which was just the same as our appraisal of the magnitude of $E(N^+ | J_0=i)$. Interpreted very broadly, and giving our appraisals more credit for sharpness than they probably deserve, this means that if the same problem $\xi = H\xi + \gamma$, were solved over and

over by the Monte Carlo method, the amount of work could be expected to exhibit statistical fluctuations around its mean, of a magnitude as great as something like 2 or 3 times the square root of $\nu_o E(N^+|J_o=i)$. Qualitatively speaking, we can state with some assurance that the amount of work required to achieve a given accuracy would vary greatly from trial to trial if the solution by Monte Carlo methods were to be carried out over and over. Due to the effect of the Central Limit Theorem as applied to (4.4) if the Monte Carlo solution were to be carried out over and over again, about half of the time the total amount of work (excluding preliminary preparations) would be less than $\nu_o E(N^+|J_o=i)$, and about half the time it would be more than this quantity. It would practically never be more than $\nu_o E(N^+|J_o=i) + 3\left\{\nu_o E(N^+|J_o=i)\right\}^{1/2}$.

These statistical fluctuations of the amount of work constitute just one more obstacle to a comparison between stochastic and non-stochastic methods of solution of linear equations. It seems logical, however, to settle on the mean value of the random variable in (4.4) as the most suitable representative of the amount of work in the stochastic method to use for comparison purposes, and this we shall do.

Our formula for the mean number of multiplications required <u>a priori</u> to assure that $|\xi_{\infty\, i} - \overline{Z}_\nu| < d$ is thus

$$(4.5) \qquad n^2 + n + \left\{\left[\frac{4v}{d^2}\right] + 1\right\} E(N^+|J_o=i).$$

Into this, we substitute the bounds given by (3.5) and (3.6), which are in terms of the data of the problem. We finally arrived at the formula

$$(4.6) \qquad \overline{m} = n^2 + n + \frac{1}{1 - \|H\|}\left\{1 + \left[\frac{2\|\rho\|^2}{d^2(1-\|H\|)^2}\right]\right\},$$

which is our basic estimate of the (mean) amount of work required in the Monte Carlo method.

V. NUMERICAL COMPARISONS

It will now be convenient to express the number d (which measures the desired closeness of approximation) as a percentage or fraction of the norm of the initial residual vector $\rho = \rho_0$. That is because $||\rho||$ and d appear in our formulas for amount of work only in a ratio. Thus we let

$$d = r\,||\rho||\,, \qquad r > 0.$$

The goal of the non-stochastic iterative method can now be phrased as being to reduce $\|\xi_\infty - \xi_N\|$ until it becomes less than some suitably small nultiple r of the largest element (in absolute value) of the initial residual vector. The goal of the stochastic method will be to reduce $|[\xi_\infty]_i - \overline{Z}_\nu|$ by repeated sampling until it too becomes less than the same small multiple r of the largest element of $\rho = \rho_0$.

With this agreement, we recapitulate our basic formulas for measuring the amount of work. The upper bound for the amount of work required in the non-stochastic iterative method is

$$m = n^2 + n + n^2 \left[\frac{\log r + \log(1-h)}{\log h}\right], \tag{5.1}$$

where $h = \|H\|$.

The upper bound for the mean amount of work required in the stochastic method is

$$\overline{m} = n^2 + n + \frac{1}{1-h}\left\{1 + \left[\frac{2}{r^2(1-h)^2}\right]\right\}, \tag{5.2}$$

In each formula, the square brackets mean "largest integer in."

The quantity in braces in (5.2) represents ν_0, the total number of times the Markov process must be sampled.

Perhaps a more natural formulation of the goals of the iterative method and the Monte Carlo method from the purely theoretical point of view would be obtained if instead of requiring that the inequalities $\|\xi_{\infty} - \xi_N\| < r\|\rho_0\|$ and $|\xi_{\infty}]_i - \bar{z}_\nu| < r\|\rho_0\|$ shall hold (the latter with high probability), we required that the inequalities $\|\xi_{\infty} - \xi_N\| < r'\|\xi_{\infty} - \xi_0\|$ and $|\xi_{\infty}]_i - \bar{z}_\nu| < r'\|\xi_{\infty} - \xi_0\|$ shall hold for some specified $r' > 0$. These modified requirements lead to simpler estimates for the total amount of work. Proceeding in the spirit of our previous analysis, we use the relation

$$\|\xi_{\infty} - \xi_0\| = \|(I-H)^{-1}\rho_0\| \leq \frac{\|\rho_0\|}{1-h},$$

and rephrase the new requirements as follows:

$$\|\xi_{\infty} - \xi_N\| < r' \frac{\|\rho_0\|}{1-h},$$

$$|\xi_{\infty}]_i - \bar{z}_\nu| < r' \frac{\|\rho_0\|}{1-h}.$$

Substituting the right-hand members of these inequalities for d in (4.2) and (4.6) respectively, we get

$$m' = n^2 + n + n^2\left[\frac{\log r'}{\log h}\right],$$

and

$$\bar{m} = n^2 + n + \frac{1}{1-h}\left\{1 + \left[\frac{2}{r'^2}\right]\right\}.$$

Of course in practice, $\|\xi_\infty - \xi_0\|$ is itself not computable before the solution is known, so the new requirements will always have to be translated into terms of $\|\rho_0\|$ and h, just as they were in the above theoretical error analysis. This essentially reduces the new set of requirements to the old ones, with an intermediate appraisal thrown into the picture. Therefore at the expense of a slight complication in our formulas, we choose to assume that the required degree of approximation is expressed in terms of a multiple of the computable quantity $\|\rho_0\|$ rather than in terms of a multiple of the non-computable quantity $\|\xi_\infty - \xi_0\|$.

In addition to the iterative and Monte Carlo methods of solving $\xi = H\xi + \gamma$, we promised in the ground rules in Section III that a non-iterative direct method will be brought into the comparison as a sort of standard of reference. The method we propose to consider is the Gauss Elimination Method[+]. It seems to be the particular direct method best adapted to the peculiar problem to which we have addressed ourselves; namely, that of computing just one component of the solution vector.

To apply it, we might proceed as follows: We are seeking $\xi_\infty]_i$, for some fixed $i = i_0$. Permute the columns of I-H and the components of ξ so that the i_0-th column of I-H becomes the n-th one and the i_0-th component of ξ becomes the n-th one. Triangularize the (new) matrix I-H as in the first part of the Gauss elimination method, always using leading row elements as pivots. At the end of the triangularization procedure, which requires approximately $n^3/3 + n^2$ multiplications[++], the coefficient of the desired component is sitting out in the open, so to speak, at a vertex of a triangular array, with nothing but zeros for the other terms in its row. Of course at the same time γ must be suitably transformed.

It would require only about $(1/2)n^2$ more multiplications now to get the rest of the components of the solution, but for present purposes we ignore the fact that a complete solution would lie so near at hand at this point.

[+]See for example Dwyer [5], Section 6.4.

[++]The exact count depends on the order in which the arithmetic operations are carried out.

The elimination solution, as we said in Section 3, is assumed to be exact. No questions of approximation (which for large matrices in practice will indeed arise because of round-off error) will be considered here.

As indicated above, our formula for the amount of work in the direct solution is, then

$$m_g = \frac{n^3}{3} + n^2 \quad . \tag{5.3}$$

It follows from formulas (5.1), (5.2), and (5.3) that with reasonable values of h and r, the direct method will be more economical for small values of n, the non-stochastic iterative method for intermediate values of n, and the Monte Carlo method for large values of n. The formulas for the break-even points, obtained by equating our estimates for the amount of work, are as follows:

The amount of work for the Gauss elimination method, as estimated by (5.3), is less than that for the stationary linear iterative method, as estimated by (5.1), for values of n in the interval

$$1 \leq n < \frac{3a + (9a^2 + 12)^{1/2}}{2} \quad ,$$

where

$$a = \left[\frac{\log r + \log (1-h)}{\log h}\right] \quad .$$

It is greater than that for either the linear iterative method or the Monte Carlo method for values of n exceeding the right member of (5.4).

The amount of work, as estimated by (5.1), for the stationary linear iterative method is less than the mean amount of work for the Monte Carlo method, as estimated by (5.2), for values of n in the interval

$$1 \leq n < \left(\frac{b}{a}\right)^{1/2} \quad , \tag{5.5}$$

where

$$b = \frac{1}{1-h}\left\{1 + \left[\frac{2}{r^2(1-h)^2}\right]\right\} .$$

It is greater than the mean amount of work for the Monte Carlo method for values of n exceeding the right menber of (5.5).

In the accompanying table, we list numerical values of these limits, together with some related quantities, for various typical values of r and h. In one case--that in which h = 9/10, r = 1/10--the linear iterative method always requires (by our a priori estimates) more multiplications than some one or both of the other two methods. (We are referring to the mean amount as usual in the case of the Monte Carlo method). The break-even dimensionality for the Monte Carlo method was computed in this case by equating (5.1) and (5.3).

It is important to notice that the measure of work for the Monte Carlo method will increase only as n^2, and not as an^2, $a>1$. The term n^2 in (5.2) represents the work required to prepare the vector ρ_0 before the stochastic estimation procedure is begun. If one is willing to content oneself with $\xi_0 = 0$ as the initial estimate, then no multiplications whatever are needed to find $\rho = \rho_0$, and the term n^2 in (5.2) drops out. Under the circumstances, the total mean amount of work required by the Monte Carlo method increases only as the first power of n. If we also decide not to calculate the numbers $r_i/p_{i,i+n}$ in advance, but only as needed in the sampling, then all direct formal dependence of the mean amount of work in the Monte Carlo method on n disappears.

The reader should be warned not to try to check the table for consistency by assuming that two stages of a reduction in the magnitude of $\|\xi_\infty - \xi_N\|$ by an amount $r\|\rho_0\|$, using the approximate solution of $\xi = H\xi + \gamma$ obtained in the first stage as the ξ_0 for the second stage, should theoretically require just the same amount of work as a one-stage reduction in $\|\xi_\infty - \xi_N\|$ by an amount of $r^2\|\rho_0\|$. Let N_0 be the number of iterations required by the first stage. The methods

Table 1.

Favorable ranges of dimensionality for the Gauss elimination method, a linear iterative method, and the corresponding Monte Carlo method, as determined by a priori analysis.
Note: The problem is to compute only one component of the solution of $\xi = H\xi + \gamma$.

Norm of H and measure of accuracy required.	$h=\frac{5}{10}, r=\frac{1}{10}$	$h=\frac{5}{10}, r=\frac{1}{100}$	$h=\frac{5}{10}, r=\frac{1}{1000}$	$h=\frac{9}{10}, r=\frac{1}{10}$	$h=\frac{9}{10}, r=\frac{1}{100}$	$h=\frac{9}{10}, r=\frac{1}{1000}$	$h=\frac{9}{10}, r=10^{-10}$
Favorable range of dimensionality for Gauss elimination method.	$n \leq 12$	$n \leq 21$	$n \leq 33$	$n \leq 84$	$n \leq 195$	$n \leq 261$	$n \leq 720$
Favorable range of dimensionality for linear iterative method.	$13 \leq n \leq 20$	$22 \leq n \leq 151$	$34 \leq n \leq 1206$	(b)	$196 \leq n \leq 554$	$262 \leq n \leq 4794$	$n \geq 721$
Favorable range of dimensionality for Monte Carlo method.	$n \geq 21$	$n \geq 152$	$n \geq 1207$	$n \geq 85$	$n \geq 555$	$n \geq 4795$	(c)
Mean number of multiplications required by Monte Carlo method at beginning of favorable range.	2064	183,258	17,458,058	207,320	20,308,590	2,022, 996,830	(c)
Approximate time to perform multiplication in row above. (a)	3 sec.	4 min.	5 hrs.	4 min.	6 hrs.	563 hrs.	(c)

Notes: (a) Calculated at the rate of one millisecond per multiplication and rounded off to the next higher unit of time.
(b) The linear iterative method is never more favorable than both of the other two methods simultaneously in this case. For $n \leq 68$, it is more favorable than the Monte Carlo method. For $n \geq 130$ it is more favorable than the elimination method.
(c) The Monte Carlo method does not become more favorable than the iterative method in this case until a ridiculously high dimensionality, of the order of 10^{+10}, is reached. At this dimensionality, it would take 10^{13} years to perform the Monte Carlo calculations.

used to compute the table would place N_o at the smallest value compatible with

$$\frac{h^{N_o}}{1-h} < r \quad . \tag{5.6}$$

If N_o' is the number of iterations required to effect a one-stage reduction in $\|\xi_\infty - \xi_N\|$ by an amount $r^2\|\rho_o\|$, the methods used to compute the table would place N_o' at the smallest value compatible with

$$\frac{h^{N_o'}}{1-h} < r^2 \quad .$$

This inequality is the same as the following one:

$$\frac{h^{N_o'/2}}{1-h} < \frac{r}{(1-h)^{1/2}} \quad .$$

Since $(1-h)^{1/2} < 1$, it follows by comparison with (5.6) that $N_o'/2 < N_o$.

It should also be pointed out that to perform a Monte Carlo approximation in two stages would require that all components of the solution vector must be estimated in the first stage, and not just one component. The reason is that to set up the random variable Z (see (3.3)) for the second stage of estimation, all the components of the initial residual vector for this stage must be available.

VI. AN ANALOGOUS COMPARISON FOR MATRIX INVERSION

If the problem is to solve $AX = I$, where A is a given n x n matrix, a suitable modification of (3.2) on which to construct a Monte Carlo solution is as follows:

$$(6.1) \qquad X_\infty = A^{-1} = X_o + (I+H_o+H_o^2+\ldots)H_oX_o \ ,$$

where X_o is an initial estimate of A^{-1} and $H_o = I - X_oA$. If X_o is a reasonably good estimate of A^{-1}, then $\|H_o\| < 1$, and the infinite series in (6.1) converges. We assume that $\|H_o\| < 1$ throughout the remainder of this section.

We set up the numbers z_{ij} and p_{ij} in terms of the elements of H_o exactly as in Section III. Assuming that we are trying to approximate the (i,k)-th element of X_∞, we take as the ρ of formula (3.3), the k-th column of H_oX_o. The statistical estimator will now be

$$\overline{Z}_\nu = \xi_o]_i + \frac{Z_1+Z_2+\ldots+Z_\nu}{\nu} \ ,$$

where ξ_o is the k-th column of X_o.

The stationary linear iterative process which corresponds to (6.1) is given by the recursion formula

$$(6.2) \qquad X_{N+1} = H_oX_N + X_o, \quad N = 0,1,\ldots,$$

where X_N is the N-th approximation to A^{-1}. Obviously $A^{-1} = H_oA^{-1} + X_o$, so

$$A^{-1} - X_N = H_o(A^{-1} - X_{N-1}) = \ldots = H_o^N(A^{-1} - X_o) \ .$$

This is the analogue of (2.5). Since $A^{-1} - X_o = (I-H)^{-1}HX_o$, we find that

$$(6.3) \qquad A^{-1} - X_N = H_o^N(I-H_o)^{-1}H_oX_o \ .$$

This equation is the analogue of (2.12).

If we let ξ_N denote the k-th column of X_N, $N=0,1,\ldots$, then the iterations defined by (6.2) give the following sequence of approximations to ξ_∞, the k-th column of A^{-1}.

$$\xi_{N+1} = H_o \xi_N + \xi_o, \quad N=0,1,\ldots \qquad (6.4)$$

Also, from (6.3),

$$\xi_\infty - \xi_N = H_o^N (I-H_o)^{-1} \rho ,$$

where ρ is the k-th column of $H_o X_o$. This equation is formally identical with (2.12). Moreover, from it we find that

$$\|\xi_\infty - \xi_N\| \leq \frac{\|H_o\|^N}{1-\|H_o\|} \|\rho\| , \qquad (6.5)$$

which is the same as (2.13).

If we now define our problem as that of insuring that $\|\xi_\infty - \xi_N\| < d$ in the non-stochastic method, and $|\xi_\infty]_i - \bar{Z}_\nu| < d$ in the stochastic method, where $d > 0$ is preassigned, then the <u>a priori</u> error analyses become precisely the same as those given in Section IV. It requires n^3 multiplications to set up H_o, given X_o, and then n^2 more to find $\rho = H\xi_o$. However, $H\xi_o$ will be used again in the non-stochastic method to pass from ξ_o to ξ_1. The resulting formula for the total amount of work, including the preparatory work becomes in the non-stochastic case,

$$m = n^3 + n + n^2 \left[\frac{\log r + \log (1-h)}{\log h}\right] , \qquad (6.6)$$

where $h = \|H_o\|$ and $r = d/\|\rho\|$. In the stochastic case it becomes

$$\bar{m} = n^3 + n^2 + n + \frac{1}{1-h}\left\{1 + \left[\frac{2}{r^2(1-h)^2}\right]\right\} . \qquad (6.7)$$

The break-even point for the two methods is given by the formula $\{b/(a-1)\}^{1/2}$, where a and b have the same meaning as in Section V. For values of n less than this quantity, the non-stochastic method requires less work than the stochastic method, and for values of n greater than this quantity, the stochastic method requires less work than the non-stochastic method.

In comparing these formulas with (5.1), (5.2) and (5.5) it should be remembered that in arriving at the earlier work-estimates (5.1) and (5.2) for the problem $A\xi = \eta$, we assumed that the H and the γ in the equivalent form $\xi = H\xi + \gamma$ were given, and so we did not count in work required to find them. Here we did count in the work required to find our H (denoted here by H_o). (The vector γ is here ξ_o, and it comes free, so to speak.) The methods have therefore become nominally unfavorable as compared to the Gauss elimination method, which for the present problem (finding one component of the solution of $A\xi = \varepsilon_k$ where ε_k is the k-th column of I) would require rather less than $n^3/3 + n^2$ multiplications.

We can sidestep the n^3 multiplications required to get H_o, by taking X_o as a very simple matrix (maybe even $X_o = I$ if $\| I-A \| < 1$). But we should state here that the real motivation for using a linear iterative method, or one of the many orthogonalization and gradient methods, for the problem $A\xi = \eta$, or the problem $AX = I$, in place of a straightforward elimination method, usually does not lie in a theoretical count of the number of operations required in the worst cases. It lies in the fact that A may have special properties (e.g., symmetry, or many zeros) which are not suitably exploited by the elimination methods. We are completely ignoring such considerations throughout this study. Another motivation sometimes is presented by the necessity of controlling round-off error. (The Monte Carlo method looks very good from this standpoint.)

It would also be possible to construct a Monte Carlo solution on the following rearrangement of (6.1):

$$X_\infty = A^{-1} = (I+H_o+H_o^2+\ldots)X_o \quad .$$

The vector form of this equation is

$$\xi_\infty = (I + H_o + H_o^2 + \ldots)\xi_o,$$

where ξ_∞ and ξ_o have the same meaning as before. This procedure would avoid the necessity of calculating $H\xi_o$ in advance, and so the n^2 term would drop out of (6.7). The numbers z_{ij} and p_{ij}, and the random variables $J_o, J_1, \ldots$ would be set up as in Section III, but in the random variable Z, the components of ρ would be replaced by those of ξ_o, and the estimator $\overline{Z}_\nu[\xi_o]_i$ would be replaced by zero. With these changes, the estimate (3.5) of the variance of Z becomes

$$v \leq \frac{1}{2} \frac{\|\xi_o\|^2}{(1 - \|H_o\|)^2},$$

and formula (4.6) for the mean amount of work (now augmented by the calculation of H_o but decreased by the amount of work previously necessary to calculate ρ) becomes

$$\overline{m} = n^3 + n + \frac{1}{1 - \|H_o\|}\left\{1 + \left[\frac{2\|\xi_o\|^2}{d^2(1-\|H_o\|)^2}\right]\right\}.$$

The disadvantage of this arrangement is that it does not exploit the fact that v varies with the square of the norm of whatever vector is playing the role of the vector θ of Section III. Therefore the goodness of the initial estimate is here made use of to reduce the statistical fluctuations and consequent mean amount of work only through the effect it has on the value of $1/(1-\|H_o\|)$.

These remarks suggest a more general comment which is perhaps the key to all the developments in this paper. The statistical part of the amount of work required by the Monte Carlo method to achieve a given accuracy in computing one element of a solution, is independent of the dimensionality of the problem. Other known methods vary as the square and cube of the dimensionality, and those which vary as the square do so with a proportionality constant much larger than unity. Therefore if one uses the Monte Carlo method,

one can afford to make substantial preliminary preparations, involving an amount of work which varies even with the square of the dimensionality, if these preparations will substantially cut down the error in the subsequent statistical estimation procedure.

For the sake of completeness, we shall bring into the comparison a certain class of non-linear iterative processes for computing A^{-1} which theoretically converge much faster than the linear iterative process (6.2) for a given initial estimate X_0. A typical member of the class is defined by the recursion formula+

$$(6.8) \qquad X_{N+1} = (I+H_N+H_N^2+\ldots+H_N^{s-1})X_N, \quad N = 0,1,2,\ldots,$$

where $H_N = I-X_NA$, and s is some integer not less than 2. If $s = 2$, the formula becomes $S_{N+1} = (2I-X_NA)X_N$, which is mentioned in most textbooks on numerical analysis as an analogue of the Newton-Raphson method for finding the roots of non-linear single equations in scalars++.

The clue to an a priori error analysis for (6.8) lies in observing that

$$\begin{aligned} H_N = I - X_NA &= I - (I+H_{N-1}+H_{N-1}^2+\ldots+H_{N-1}^s)X_{N-1}A \\ &= I - (I+H_{N-1}+\ldots+H_{N-1}^s)(I-H_{N-1}) \\ &= H_{N-1}^s \ . \end{aligned}$$

+Our presentation of these polynomial iteration procedures will be slightly different from that usually encountered in the literature, so as to line them up with (6.1) and (6.2). The usual presentation replaces our $H_N = I-X_NA$ by $I - AX_N$. A number of references relating to these methods, as well as to all other methods discussed in this paper will be found in Forsythe [6].

++See e.g., Householder [7], pp. 56-57.

Therefore by back substitution,

$$H_N = H_o^{s^N} .$$

Now $A^{-1} - X_N = (I - X_N A)A^{-1} = H_N A^{-1}$; and $X_o A = I - H_o$, so $A^{-1} = (I - H_o)^{-1} X_o$. From all this we obtain:

$$A^{-1} - X_N = H_o^{s^N} (I-H_o)^{-1} X_o = H_o^{s^N-1} (I-H_o)^{-1} H_o X_o .$$

This equation is the analogue of (6.3). From it we get in place of (6.5),

$$\| \xi_\infty - \xi_N \| \leq \frac{\| H_o \|^{s^N-1}}{1 - \| H_o \|} \| \rho \| , \tag{6.9}$$

and this clearly represents a much faster rate of convergence than (6.5).

The difficulty is that each iteration of (6.8) requires sn^3 multiplications. Moreover in the special problem at hand--that of finding only one element of A^{-1}--the method does not appear to good advantage, because there seems to be no way to avoid computing all the elements of the matrix X_N each time, and not just the k-th column, as we did in the linear iterative method. In other words, there seems to be no direct analogue of the vector recursion formula (6.4) in the method given by (6.8).

The formula for the total amount of work required to make the right hand member of (6.9) less than $r \| \rho \|$, where $r > 0$ is preassigned, is as follows:

$$sn^3 + sn^3 \left[\frac{\log \left(1 + \frac{\log r + \log (1-h)}{\log h}\right)}{\log s} \right] \tag{6.10}$$

where, as usual, the square brackets means "largest integer in" and $h = \| H_o \|$. A study of the maximum and minimum of

this expression, considered as a function of s, reveals that $s = 2$ or $s = 3$ usually are the most advantageous values of s to use. For example, if $r = 10^{-3}$ and $h = 9/10$, then the formula (6.10) becomes

$$sn^3 \left(1 + \left[\frac{1.9465}{\log s} \right] \right) .$$

If $s = 2$, this equals $14n^3$. If $s = 3$, it equals $15n^3$. If $s = 4$, it equals $16n^3$. For higher values of s, the disadvantage becomes more pronounced.

With $r = 10^{-3}$, $h = 9/10$, $s = 2$, the amount of work required by the non-linear iterative method given by (6.4), as estimated from (6.10), is less than required by the linear method given by (6.4), as estimated from (6.6), for $n \leq 6$. It is greater for $n \geq 7$.

REFERENCES

[1] Courant, R., and Hilbert, D., Methoden der Mathematischen Physik, Berlin, 1931.

[2] Cramér, H., Mathematical Methods of Statistics, Princeton, 1951.

[3] Curtiss, J. H., "Monte Carlo" methods for the iteration of linear operators, Journal of Mathematics and Physics, vol. 32 (1954), pp. 209-232.

[4] Curtiss, J. H., Sampling methods applied to differential and difference equations; Proceedings of a Seminar on Scientific Computation, held by the International Business Machines Corporation, Endicott, New York, November, 1949; pp. 87-109.

[5] Dwyer, P. S., Linear Computations, New York, 1951.

[6] Forsythe, G. E., Tentative Classification of Methods and Bibliography on Solving Systems of Linear Equations. In "Simultaneous Linear Equations and the Determination of Eigenvalues," National Bureau of Standards Applied Mathematics Series 29, Washington, 1953.

[7] Householder, A. S., Principles of Numerical Analysis, New York, 1953.

[8] Milne, W. E., Numerical Solution of Differential Equations, New York, 1953.

[9] Taussky-Todd, O., A recurring theorem on determinants, American Mathematical Monthly, vol. 56 (1949), pp. 672-676.