LA--1342

LOS ALAMOS SCIENTIFIC LABORATORY

OF THE

UNIVERSITY OF CALIFORNIA

CONTRACT W-7405-ENG. 36 WITH

U. S. ATOMIC ENERGY COMMISSION

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

1
LOS ALAMOS SCIENTIFIC LABORATORY

of the

UNIVERSITY OF CALIFORNIA

Report Written:
October 13, 1951

LA-1342

Report Distributed:

APR 4 1952

W. 22714

THE EVALUATION OF DEFINITE INTEGRALS, AND A
QUASI-MONTE-CARLO METHOD BASED ON THE
PROPERTIES OF ALGEBRAIC NUMBERS

by

R. D. Richtmyer

PHYSICS

This document is
PUBLICLY RELEASABLE

Larry C. Williams
Authorizing Official

Date: *04/12/2007*

ABSTRACT

A formula is given for the approximate evaluation of multiple definite integrals using the ergodic property of a certain transformation of the unit cube into itself. Estimates of the rate of convergence are made for sufficiently smooth integrand. The work was motivated by a belief, that appeared at one time justified, that a substantial improvement of the accuracy of Monte Carlo method would result from use of the principles described herein. Although that belief proved groundless in numerical tests, it is deemed worthwhile to give this report of the work. Part of the theoretical development, the work of L. G. Peck, will be reported independently. (10)

1. INTRODUCTION - THE MONTE CARLO METHOD

In the usual Monte Carlo method of calculating the properties of a neutron chain reaction, one is concerned with a large collection or ensemble of neutron chains, each starting with a single neutron, and all developing under identical conditions, determined by the distribution of fissionable and scattering material. The chains differ from each other, even though they are all started in an identical manner, because the occurrence and characteristics of the nuclear events are determined by chance. The probability laws governing these events are presumed known and one wishes to determine the statistical properties of the ensemble of chains.

For this purpose one constructs numerical descriptions mathematically of a large number of representative chains by assigning arbitrary values to the various quantities that determine the sequence of events constituting the chains. The values are assigned at random with probabilities distributed in accordance with the aforementioned probability laws. Specifically, the assignment is made in the following way: there is supposed given an inexhaustible supply of "random" numbers; i. e., an infinite sequence of numbers,

$$x_1, x_2, x_3, \dots, x_i, \dots \quad (1)$$

uniformly distributed in the interval $(0, 1)$, of which each is statistically independent of the numbers preceding it in the sequence. These numbers are used one by one and then discarded. When in the

course of constructing the numerical description of one of the chains, it becomes necessary to assign a value to a quantity y for which a probability distribution $P(y)$ is given (by this is meant: $P(y_0)$ is the probability of $y < y_0$ and, in particular,

$$P(-\infty) = 0, P(+\infty) = 1), \quad (2)$$

then the value assigned to y is y_0 given by the equation:

$$P(y_0) = x_1, \quad (3)$$

where x_1 is the next unused number of the sequence (1). For computing purposes, the relation (3) is expressed in the form:

$$y_0 = F(x_1), \quad (4)$$

where $F(x)$ is the inverse of the function $P(y)$. The construction of each chain is continued until all the events occurring up to a given time ("census time") have been determined. Then the next chain is constructed, and so on.

(In practice, the calculations for different chains are often interspersed, the events being taken in more or less chronological order, regardless of which chain they belong to; but the final result is the same as that described above, except for an irrelevant permutation of the random numbers.)

When the descriptions of sufficiently many representative chains have been prepared in this way, average properties of the chain reaction can be obtained by statistical analysis of them. Suppose, for example, that the problem is to find the average value of the total kinetic energy E of all neutrons in the system at the end of a

specified time after the introduction of the initiating neutron. We compute the value of E for each representative chain from its numerical description and average the result for all the chains.

The value of E (or of any quantity whose average value is desired) for a given representative chain depends on the values of those "random" numbers of the sequence (1) that entered into its description. That is, we can write,

$$E = f(x_i, x_{i+1}, x_{i+2}, \dots), \quad (5)$$

for some i . The function f must, in general, be regarded as a function of infinitely many variables, because one cannot tell in advance how many numbers of the sequence (1) will be required for the description of a given chain up to the end of the specified time interval, unless one is willing to adopt some ruse such as assigning a minimum time between collisions for a neutron; for otherwise an indefinitely large number of events may, in principle, take place in the interval. If it turns out, for a given chain, when the values of the appropriate x 's are examined, that N random numbers suffice for determining the chain, then, for those particular values of the first N arguments in (5), the function f is independent of the values of the remaining, infinitely many, arguments.

In general, the function f is very complicated, and cannot be easily written down, but in any case it is clear that the objective of the Monte Carlo calculation is an approximate evaluation, by use of the random numbers, of the integral

$$\bar{E} = \int_0^1 dx_1 \int_0^1 dx_2 \dots f(x_1, x_2, \dots) \quad (6)$$

over the unit cube in a space of infinitely many dimensions. Each chain is associated with a point in this unit cube. $\bar{E}$ is the limit towards which the average of E must tend as the number of mathematical representants in the ensemble tends to infinity, because the variables in (6) are random numbers uniformly and independently distributed in the interval $(0, 1)$. The representative chains of Monte Carlo are associated with points randomly distributed in the unit cube.

One naturally wonders what is meant, numerically, by the purely formal equation (6), with its infinitely multiple integral. The expression on the right of that equation can no doubt be given a precise interpretation in many ways; a simple way that is suggested by procedures used in actual Monte Carlo practice is the following. We first regard f as a function of only k independent variables, whose values are then repeated over and over again cyclically, to provide the infinitely many arguments appearing in F . That is, we write

$$f_k(x_1, \dots, x_k) = f(x_1, \dots, x_k, x_1, \dots, x_k, \dots), \quad (7)$$

and then interpret (6) as being equivalent to the equation,

$$\bar{E} = \lim_{k \rightarrow \infty} \int_0^1 dx_1 \int_0^1 dx_2 \dots \int_0^1 dx_k f_k(x_1, \dots, x_k). \quad (8)$$

It may be surmised, on physical grounds, that the limit in (8) is approached quite rapidly in practice (this can be subjected to

empirical test). If k is of the order of 50 or 100, the representative chain constructed from the arguments in f will not consist of purely random events; but the correlations introduced by repetition of the arguments will generally be between quite distantly related and unlike events; and it is in the nature of a chain reaction, at least in a system whose dimensions are comparable to the mean free path, that what a neutron does is not strongly influenced by what one of its ancestors was doing many generations ago.

Viewed in this light, Monte Carlo is essentially a method for approximate numerical evaluation of highly multiple integrals, in cases where direct evaluation by conventional methods is out of the question.

In this report we propose an alternative, though similar, method of evaluation of such integrals. It involves again the construction of a sequence of points in the unit cube, with uniform distribution in the limit, and then approximating the integral by the average of the values of the integrand at the first N points of the sequence. The novel feature is that these points are not distributed at random, but in a particular way so as to approach a uniform distribution more rapidly (in a certain sense) than a random sequence would. We will show that for a sufficiently smooth integrand, the average over the first N points approaches the value of the integral more rapidly than for a random sequence.

Integrands met in Monte Carlo calculations are generally not smooth, and are, in fact, generally discontinuous. In particular,

if $f(x_1, \dots, x_k)$ is the number of particles present at census time in the chain associated with the point $(x_1, \dots, x_k)$, then clearly f is a step function, having only integral values. Unfortunately, our theorems have not been proved for such integrands. But, acting on the surmise that some similar theorem may apply, we have proposed an alternative method of treating problems like neutron diffusion. It involves again the construction of a sequence of representative chains, and, in fact, the whole calculational routine is the same as for Monte Carlo, except for the manner of choosing the sequence (1). For a fixed value of k , a given chain of the ensemble is associated with a particular point of our special sequence of points in the k -dimensional unit cube, and the coordinates of this point are used in order (with repetitions if necessary) until the description of the chain is complete. Then the next point of the sequence is used for constructing the description of the next chain, and so on.

It seemed unlikely that we could ascertain by purely theoretical considerations whether this method is really superior to Monte Carlo in practice, and consequently we resorted to numerical experimentation. We applied the proposed method to the calculations of the properties of an actual multiplicative system for a case in which these properties can be obtained also by analytic methods. Our result was essentially negative, as will be shown in section 5, below: the rate of convergence of our method did not seem to be appreciably better than that to be expected from Monte Carlo methods. It may be that no theorem

similar to ours on evaluation of multiple integrals (which so far has been proved only for smooth integrands) holds for the type of integrand met in stochastic problems, or it may be that the superior rate of convergence only begins to take hold for much larger sample sizes than those used in the numerical experimentation. In any case, the interest in our method as a practical competitor to Monte Carlo would seem to be very slight. As a method for numerical evaluation of multiple definite integrals, it may conceivably still have value in certain cases.

We report first, in the next two sections, the theoretical considerations up to a certain point. They were then carried further by Peck, whose investigations will be referred to only briefly here, but will be reported in detail in a separate communication. The present write-up is essentially a progress report on the method and is being written now in spite of the incomplete status and perhaps doubtful promise of the work, simply because we do not see, at the moment, how to make further advance in this direction.

2. EVALUATION OF DEFINITE INTEGRALS

We consider the integral,

$$I = \int_0^1 dx_1 \int_0^1 dx_2 \dots \int_0^1 dx_k f(x_1, x_2, \dots, x_k) \quad (9)$$

over the k -dimensional unit cube. According to the Monte Carlo

method for approximate evaluation of I , we group the numbers of the sequence (1) into groups of k numbers each, which may be denoted by,

$$x_1^n, x_2^n, \dots, x_k^n, \quad (n = 0, 1, 2, \dots), \quad (10)$$

and write:

$$I_N = \frac{1}{N} \sum_{n=0}^{N-1} f(x_1^n, x_2^n, \dots, x_k^n). \quad (11)$$

Then,

$$\lim_{N \rightarrow \infty} I_N = I \quad (12)$$

for a sufficiently smooth integrand, f (piece-wise differentiability suffices) and, indeed, according to probability theory,

$$|I_N - I| = O\left(\frac{1}{\sqrt{N}}\right), \text{ as } N \rightarrow \infty \quad (13)$$

However, the occurrence of the square root in (13) means, roughly, that if one wishes to calculate I to three figure accuracy, N must be at least of the order of 10^6 ; and this results in a rather uncomfortably large amount of calculation, at least by machines available at the moment of writing.

In stochastic problems, the fluctuations of I_N about its limiting value I have their counterpart in real physical fluctuations, and it is sometimes of interest to study these fluctuations per se. But, if one is interested solely in the value of I , these fluctuations are disadvantageous. Moreover, if one is interested in the fluctuations, one can usually characterize them by one or more statistical

quantities (root-mean-square averages and the like) that can also be expressed as integrals of the form (9).

Because of the slow convergence of I_N to I implied by (13), we propose an alternative procedure in which equation (11) is retained and in which the sequence of points described by the coordinates (10) is again uniformly distributed in the unit cube. However, the successive points, for $n = 0, 1, 2, \dots$, are not statistically independent but are chosen in a systematic way such as to result in a considerably more rapid convergence of I_N to I by reduction of the fluctuations.

Specifically, let $\xi_1, \xi_2, \dots, \xi_k$ be k irrational real numbers, linearly independent over the rationals. (That is, they are real, irrational numbers such that a linear combination of them with rational coefficients can vanish only if all the coefficients vanish.) For any real number x , let $\{x\}$ denote the fractional part of x -- that is, what is left after subtracting off the greatest whole number that does not exceed x . Then, according to a theorem of H. Weyl*, the sequence of points of which the n^{th} point has coordinates

$$(x_1^n, x_2^n, \dots, x_k^n) = (\{n \xi_1\}, \{n \xi_2\}, \dots, \{n \xi_k\}) \quad (14)$$

* See Oskar Perron, Irrationalzahlen, Chelsea Publishing Company, 1948, pages 162 ff.

is uniformly distributed in the unit cube, so that with this choice of the arguments in equation (11), I_N again converges to I as N tends to infinity. We shall show that, at least if f is a sufficiently smooth function, the convergence is considerably more rapid than for a random distribution of points.

The optimal choice of the irrational numbers is not clear (except in the one-dimensional case, $k = 1$, for which one would use the simple trapezoid rule, Simpson's rule, or Gauss' rule, anyway) but probably they should be algebraic numbers of low degree.

In a calculation with a digital computing machine, these irrational numbers will, of course, have to be approximated by rational numbers (terminating decimals or binaries), but if the number of significant decimal digits (or binary digits) retained in the calculation is not less than approximately $2 \log_{10} N$ (or $2 \log_2 N$), the rational approximants behave, for our purpose, just like the irrationals they represent. This consideration leads one to the thought that the whole theory might better be based on the properties of rational, rather than of irrational numbers, but so far we have not seen how to do this.

Considered solely as a method for the evaluation of definite integrals, the method may be criticized on the ground that more elementary procedures, using a cubic lattice of points in the unit cube are superior (trapezoid rule, etc.). If P is the number of lattice points on an edge of the unit cube, then the total number of

points is $N = p^k$, and the rate of convergence of I_N to I is generally still more rapid than for our method, as N increases through the sequence $2^k, 3^k, 4^k, \dots$. The difficulty is that if k is as large as 20 (one would hardly dare use a smaller value in Monte Carlo problems), the smallest possible lattice requires calculations of 2^{20} values of the integrand. To quote John Von Neumann, there ought to be some method in a "price" range suitable for people who are willing to "spend" a few thousand points, but not 2^{20} points. We doubt that such a method exists for the general case of a 20-dimensional integral but for the special type of integrand met in stochastic problems there is such a method -- the Monte Carlo method -- and we are trying to find a better one of the same type. The peculiarity of the integrands in question is that although k may be large (e. g., > 20) and, in fact, is really indefinite, as noted above, the effective number of dimensions is smaller (say, 4 or 5). (The effective number of dimensions may be defined as follows: at point $(x_1, x_2, \dots, x_k)$ let s be the smallest integer such that $f(x_1, x_2, \dots, x_s, x'_{s+1}, \dots, x'_k)$ is independent of $x'_{s+1}, \dots, x'_k$. Then the average of s over the unit cube is the effective number of dimensions for the integrand f .)

Possible applications of the method to integrals other than those met in stochastic problems has not yet been explored. But, a similar method has apparently been used in work on an analogue computer (electrical version of a differential analyzer). In our

method, if T is the transformation

$$T: (x_1, \dots, x_k) \longrightarrow (\{x_1 + \xi_1\}, \dots, \{x_k + \xi_k\})$$

of the unit cube into itself, we have written:

$$I_N = \frac{1}{N} \sum_{n=0}^{N-1} (T)^n \left[f(x_1, \dots, x_k) \right]_{(x_1, \dots, x_k) = (0, \dots, 0)}.$$

For an analogue computer having one continuous variable s , the corresponding equation would be

$$I_{s_0} = \frac{1}{s_0} \int_0^{s_0} \left[T_s f(x_1, \dots, x_k) \right]_{(x_1, \dots, x_k) = (0, \dots, 0)} ds$$

where T_s is a one parameter linear group of transformations. It is understood through indirect oral reports that this method has been used for evaluation of definite integrals in several dimensions at the Argonne National Laboratory. In both the discrete and continuous cases the family of transformations must have the ergodic property.

3. RATE OF CONVERGENCE

I_N is now given by

$$I_N = \frac{1}{N} \sum_{n=0}^{N-1} f(\{n \xi_1\}, \dots, \{n \xi_k\}). \quad (15)$$

We assume that f can be expanded in a Fourier series:

$$f(x_1, \dots, x_k) = \sum_{q_1=-\infty}^{\infty} \dots \sum_{q_k=-\infty}^{\infty} a(q_1, \dots, q_k) e^{2\pi i(q_1 x_1 + \dots + q_k x_k)}. \quad (16)$$

For any x , $\exp 2\pi i \{x\} = \exp 2\pi i x$; consequently,

$$I_N = \sum_{q_1, \dots, q_k} a(q_1, \dots, q_k) \frac{1}{N} \sum_{n=0}^{N-1} e^{2\pi i n (q_1 \xi_1 + \dots + q_k \xi_k)} \quad (17)$$

From the formula for the Fourier coefficients, clearly $a(0, 0, \dots, 0) = I$.

The factor multiplying $a(0, \dots, 0)$ in (17) is unity, therefore, we can form $I_N - I$ from (17) by merely omitting the term with $q_1 = q_2 = \dots = q_k = 0$. This omission will be denoted by a prime on the summation sign.

That is,

$$I_N - I = \frac{1}{N} \sum'_{q_1, \dots, q_k} a(q_1, \dots, q_k) \frac{e^{2\pi i N (q_1 \xi_1 + \dots + q_k \xi_k)} - 1}{e^{2\pi i (q_1 \xi_1 + \dots + q_k \xi_k)} - 1}, \quad (18)$$

or, by the triangle inequality,

$$|I_N - I| \leq \frac{1}{N} \sum'_{q_1, \dots, q_k} |a(q_1, \dots, q_k)| \left| \frac{\sin \pi N (q_1 \xi_1 + \dots + q_k \xi_k)}{\sin \pi (q_1 \xi_1 + \dots + q_k \xi_k)} \right|. \quad (19)$$

For any real x ,

$$|\sin \pi N x| \leq 1, \quad (20)$$

$$|\sin \pi x| \geq 2|x - \langle x \rangle|, \quad (21)$$

where $\langle x \rangle$ denotes the integer nearest to x ; and therefore, finally,

$$|I_N - I| \leq \frac{1}{2N} \sum'_{q_1, \dots, q_k} \frac{|a(q_1, \dots, q_k)|}{|q_1 \xi_1 + \dots + q_k \xi_k - p|}, \quad (22)$$

where p denotes the integer nearest to $q_1 \xi_1 + \dots + q_k \xi_k$.

If the summation in (22) is convergent, we have $|I_N - I| = O(1/N)$,

which is a distinct improvement on equation (13), that holds for the

Monte Carlo method. The denominator in (22) can become arbitrarily small (but not zero, because $\xi_1, \dots, \xi_k$ are independent irrationals) for certain values of $q_1, \dots, q_k$. But, for suitable choice of the irrationals, we can put a lower bound on the size of the denominator for given magnitudes of $q_1, \dots, q_k$. Then, if the Fourier coefficients, $a(q_1, \dots, q_k)$ of the function f tend to zero sufficiently rapidly as the magnitudes of $q_1, \dots, q_k$ increase, the summation in (22) converges. The question then arises as to how close $q_1\xi_1 + \dots + q_k\xi_k$ can come to an integer.

What is needed is a generalization of Liouville's theorem that if ξ is an algebraic number of degree n , then there exists a positive constant C , such that for any integers q, p , not both zero,

$$|q\xi - p| \geq \frac{C}{|q|^{n-1}}. \quad (23)$$

A theorem of the required sort is derived in the next section and states that if $\xi_1, \dots, \xi_k$ are linearly independent irrational numbers in an algebraic field of degree n , where necessarily $n \geq k+1$, there exists a positive constant C , depending only on $\xi_1, \dots, \xi_k$, such that

$$|q_1\xi_1 + \dots + q_k\xi_k - p| \geq \frac{C}{\left(\sum_{j=1}^k |q_j|\right)^{n-1}}, \quad (24)$$

where $q_1, \dots, q_k, p$ are any integers, not all zero. In this case, (22) gives:

$$|I_N - I| \leq \frac{1}{2CN} \sum_{q_1, \dots, q_k} \left| \left(\sum_j |q_j| \right)^{n-1} a(q_1, \dots, q_k) \right|. \quad (25)$$

The result is that if the function $f(x_1, \dots, x_k)$ is sufficiently smooth to insure convergence of the summation in (25), we have

$$|I_N - I| = O\left(\frac{1}{N}\right). \quad (26)$$

The smoothness requirement on the function f is rather severe. In the one-dimensional case ($k=1$), with ξ a quadratic irrational ($n=2$), it is required that the derivative of $f(x)$ have an absolutely convergent Fourier series; whereas, in the multidimensional case, high order partial derivatives of f are also required to have absolutely convergent Fourier series.

This result can be improved. In the first place, in the one-dimensional case with ξ a quadratic irrational, a result due to Ostrowski* shows that the method can be applied when $f(x)$ is a step function, and then

$$|I_N - I| = O\left(\frac{\ln N}{N}\right). \quad (27)$$

This is a very satisfactory result, so far as the one-dimensional case goes, because no worse than simply-discontinuous integrands are likely to be met in practice, and the result (27) is almost as good as (26) and much better than (13). In the second place, examination of the

* Abhandlungen aus dem mathematischen Seminar der Hamburgischen Universität 1, 77-98 (1922).

theorem of the next section indicates, as will be shown there, that considerable improvement should be obtainable by making use of two rather deep properties of algebraic number fields. These ideas led to a separate investigation by L. G. Peck which culminated in a theorem, to be published elsewhere, that gives what is probably the strongest result obtainable with current algebraic techniques. It results also in the relation $|I_N - I| = O(\frac{1}{N})$ under restrictions on f that are considerably less stringent than those of this report, though still too stringent for Monte-Carlo-like problems.

4. SOME THEOREMS ON ALGEBRAIC NUMBERS

Theorem I

Let $\xi_1, \xi_2, \dots, \xi_k$ be linearly independent irrational algebraic integers in a real field K of degree n (where, clearly, $n \geq k+1$). Then there exists a constant $C > 0$ such that if $q_1, \dots, q_k, p$ are any rational integers, not all zero, then

$$\left| q_1 \xi_1 + q_2 \xi_2 + \dots + q_k \xi_k - p \right| \geq \frac{C}{\left(\sum_{j=1}^k |q_j| \right)^{n-1}}. \quad (28)$$

Proof

Denote the conjugates of ξ_j by $\xi_j^{(1)}, \dots, \xi_j^{(n)}$, where $\xi_j = \xi_j^{(1)}$, $j = 1, 2, \dots, k$. Construct the algebraic integer

$$\pi = q_1 \xi_1 + \dots + q_k \xi_k - p$$

and note that π and its conjugates are all $\neq 0$. Denote its conjugates

by $\pi^{(1)}, \dots, \pi^{(n)}$, where $\pi = \pi^{(1)}$. The product of the conjugates,

$$N[\pi] = \pi^{(1)} \cdot \pi^{(2)} \cdot \dots \cdot \pi^{(n)}, \quad (29)$$

is the norm of π and is, therefore, a rational integer. $N[\pi] \neq 0$;

consequently

$$|N[\pi]| \geq 1$$

or

$$|\pi^{(1)}| \geq \frac{1}{|\pi^{(2)}| \cdot |\pi^{(3)}| \cdot \dots \cdot |\pi^{(n)}|}. \quad (30)$$

Now let $0 < \epsilon < 1$ and restrict discussion to those sets of values of the rational integers $q_1, \dots, q_k, p$ for which

$$|\pi^{(1)}| = |q_1 \xi_1 + \dots + q_k \xi_k - p| < \epsilon. \quad (31)$$

Then

$$|\pi^{(m)}| = \left| \pi^{(1)} + \sum_{j=1}^k q_j (\xi_j^{(m)} - \xi_j^{(1)}) \right|, \quad (32)$$

and, by the triangle inequality,

$$|\pi^{(m)}| \leq \epsilon + \sum_{j=1}^k |q_j| \cdot |\xi_j^{(m)} - \xi_j^{(1)}|. \quad (32a)$$

Call

$$A = \epsilon + \max_{j, m} |\xi_j^{(m)} - \xi_j^{(1)}|. \quad (33)$$

Note that not only must one of the $q_1, \dots, q_k, p$ be different from zero, but one of the $q_1, \dots, q_k$ must be different from zero, because if we had $q_1 = q_2 = \dots = q_k = 0, p \neq 0$, the inequality (31) would contradict the condition $\epsilon < 1$. Therefore, $\sum_j |q_j| \geq 1$; consequently,

from (32),

$$|\pi^{(m)}| \leq \sum_{j=1}^k |q_j| (\xi + |\xi_j^{(m)} - \xi_j^{(1)}|) \leq A \sum_{j=1}^k |q_j|. \quad (34)$$

Note that by definition (33), A is independent of the $q_1, \dots, q_k$.

From (30) and (34),

$$|\pi^{(1)}| \geq \frac{1}{A^{n-1} \left(\sum_{j=1}^k |q_j| \right)^{n-1}}. \quad (35)$$

So far we have proved that either (35) holds or $|\pi^{(1)}| \geq \epsilon$. Therefore, call:

$$C = \min \left(\frac{1}{A^{n-1}}, \epsilon \right), \quad (36)$$

and we have the result

$$|\pi^{(1)}| \geq \frac{C}{\left(\sum_{j=1}^k |q_j| \right)^{n-1}} \quad (37)$$

Q.E.D.

Corollary

The $\xi_1, \dots, \xi_k$ in the above theorem need not be integers if they satisfy the other conditions of the theorem; because any algebraic number in K can be written as a fraction whose numerator is an integer in K and whose denominator is a rational integer. Let $m > 0$ be a common denominator of $\xi_1, \dots, \xi_k$. Then the theorem holds for the numbers $m\xi_1, \dots, m\xi_k$; divide the resulting inequality through by m and replace C/m by C . The corollary results.

In establishing the inequality (30) we have been very wasteful, because most (in a certain sense) of the integers of K have norms that are much greater than 1. Therefore, application of this theorem to the integration formulas of the preceding section gives a result that is unnecessarily weak. Peck pointed out that if the integers $1, \xi_1, \dots, \xi_k$ form the basis of an ideal, we can replace the numerator of (30) by the norm of this ideal. To improve the situation further, we must be able to characterize all integers of K (more precisely, of the modul $(1, \xi_1, \dots, \xi_k)$ having a given norm N so as to be able to determine their contribution to (25). A relation similar to (30) will now enter, with the numerator replaced by N .

What seems to be needed here is, first, the Dirichlet theory of units; for this allows us to characterize quite explicitly any class of associate integers, that is, a class of integers, any one of which generates a particular principal ideal. Second, we need to know how many principal ideals there can be having a given norm N ; this is to be found in the Dedekind theory of the distribution of ideals; or, equivalently, in the properties of the Dedekind zeta function. It is thus to be expected that by making use of these two parts of the theory of algebraic numbers we can broaden the conditions under which the result, $|I_N - I| = O(\frac{1}{N})$, of the preceding section, can be established. This program was carried out in two special cases and led to the theorems:

Theorem II

If ξ is a real quadratic irrational, there exists a constant A, depending only on ξ , such that for any function $f(x)$ and associated constants, C, ϵ , satisfying:

$$f(x) = \sum_{-\infty}^{\infty} (q) a(q) e^{2\pi i q x},$$

$$|a(q)| < \frac{C}{|q|^{1+\epsilon}},$$

$$0 < \epsilon < 1;$$

there holds, in our notation, the inequality:

$$|I - I_N| < \frac{AC}{N\epsilon^2}. \quad (38)$$

Theorem III

If $1, \xi_1, \xi_2$ form an integral basis for the real cubic field K, whose conjugate fields K', K'' are complex, there exists a constant A, depending only on ξ_1 and ξ_2 , such that for any function $f(x, y)$ and associated constants, C, ϵ , satisfying:

$$f(x, y) = \sum_{-\infty}^{\infty} (q_1, q_2) a(q_1, q_2) e^{2\pi i (q_1 x + q_2 y)},$$

$$|a(q_1, q_2)| < \frac{C}{\{|q_1| + |q_2|\}^{2+\epsilon}},$$

$$0 < \epsilon < 1;$$

there holds:

$$|I - I_N| < \frac{AC}{N\epsilon^2} \quad (39)$$

Proofs are omitted here. In each case use was made of simple aspects of the Dirichlet theorem on units and an elementary property of the Dedekind zeta-function for the field in question.

Attempts to generalize the method of proof used here to the case of a general field proved difficult and were eventually abandoned because L. G. Peck obtained the desired generalization by use of more elegant and more powerful methods. In Peck's theorem the condition on the Fourier coefficients is

$$|a(q_1, \dots, q_k)| < \frac{C}{\{|q_1| + \dots + |q_k|\}^{k+\epsilon}}, \quad (40)$$

where $k+1$ is the degree of the field.

In a sense **these** are much stronger results than those based on theorem I, because (40) implies the convergence of

$$\sum_{(q_1, \dots, q_k)} \left| \left\{ |q_1| + \dots + |q_k| \right\}^k a(q_1, \dots, q_k) \right|$$

(as required in the preceding section, see equation (25)) only for $\epsilon > k$.

But in order to handle integrals such as those encountered in stochastic problems and other applications, where the integrand can possess discontinuities, we should like to be able to derive an inequality for

$|I - I_N|$ for functions $f(x_1, \dots, x_k)$ satisfying merely

$$|a(q_1, \dots, q_k)| < \frac{C}{|q_1| + \dots + |q_k|}$$

In a sense, Peck's result covers half the gap (slightly more) between our previous result (section 3, above) and the desired result. Because

of the occurrence of ϵ^2 in the denominators of (38) and (39), and similar factors in Peck's result, one is tempted to surmise that these are essentially the strongest results that can be obtained without seriously weakening the dependence on N of the bound of $|I - I_N|$.

5. THE TRIAL CALCULATION

We considered the following multiplicative system: the particles in it are identical in all respects, so that an instantaneous state is fully characterized by the number of particles present, denoted by $k(t)$. Any particle present can divide into two particles at any time, and the probability that this will happen for any particle in the infinitesimal interval dt is taken proportional to dt . The proportionality constant is the same for all particles and independent of time, and is taken equal to unity by suitable choice of the unit of time t . One particle is present at time $t=0$, and required is the expected number of particles at "census" time, $t=T$. This expected number is denoted by $\langle k(T) \rangle$.

The properties of this system can be completely determined by standard statistical methods, using generating functions. In particular,

$$\langle k(T) \rangle = e^T \quad (41)$$

We adopt the convention of giving the two particles produced in division separate identities from the old particle and of considering that the old particle dies or disappears at the instant of division. $k(T)$ is then the number of particles in the chain whose time of birth

is $< T$ and whose time of death is $\geq T$. As we construct each chain we examine its particles one by one and each time we find a particle with the aforementioned property we increase a tally k by one. k is set equal to zero before the examination of each chain begins, so that when the examination is completed, $k=k(T)$.

During the construction of a chain we also keep a running list of the times of birth of certain particles. We examine the entries of the list one by one and ask what is the time of death of the particle represented. This is obtained by adding ~~$-lnx$~~ to the time of birth, where x is a random number between 0 and 1, because the probability of the particle's survival to time t is $\exp [-(t - \text{time of birth})]$. If the time of death is $\geq T$ we strike the particle off the list and increase the tally k by one. If the time of death is $< T$, we also strike the particle off the list but this time we add two new particles to the list having as birth times the time of death of the old particle, and we leave the tally unchanged. This is continued until the list is exhausted. The list of course starts with just one entry, namely, zero, for the time of birth of the initial particle.

The whole procedure is then repeated for other chains.

For construction of the n^{th} chain, the random numbers used are, in our method, $\{n\xi_1\}$, $\{n\xi_2\}$, The independent irrationals ξ_1 , ξ_2 , were taken to be

$$\sqrt{2}, \sqrt{3}, \sqrt{6}, \sqrt{5}, \sqrt{10}, \sqrt{15}, \sqrt{30}, \sqrt{7} \text{ etc.}$$

These were generated (255 of them) by having available the square roots of the first eight primes: then the r^{th} irrational is

$$(\sqrt{2})^{a_8} (\sqrt{3})^{a_7} (\sqrt{5})^{a_6} \dots$$

where $a_8 a_7 a_6 \dots a_1$ is the binary representation of r . (This gives the above sequence in reverse order--such an ordering happened to be convenient for the machine operation.) Provision was made for repeating the ξ 's if any chain should require more than 255 "random" numbers for its construction.

After a few thousand chains have been treated in this way, the values of $k(T)$ are averaged to get an approximation to $\langle k(T) \rangle$ and this average is compared with e^T .

The calculation as described here was coded for the SEAC, and the SEAC was put at our disposal for this purpose through the kindness of the Bureau of Standards. Several successful runs were made one night and some of the results are shown graphically in Figs. 1 and 2. The solid circles in Fig. 1 give the results of a first run in which the census time T was taken = 1. We have plotted the average of $k(T)$ versus sample size n for various values of n during the run, which included 4096 chains. The horizontal line indicates the expectation value and the curves indicate the deviations expected for a Monte Carlo calculation--if the chains of the sample were chosen at random, there would be probability $1/2$ that a given point would lie between the curves. The expected deviation is, of course, proportional to $n^{-1/2}$. It had

been hoped that our points would approach the expectation value much more rapidly than the curves, with increasing n - with a deviation perhaps more like n^{-1} . This hope was not fulfilled. The most one can say is that with about 4000 chains our deviation is possibly less by about a factor $1/2$ than Monte Carlo would give. Even this result was not confirmed in the second run. In the second run, whose results are indicated by the open circles of Figs. 1 and 2, the census time T was $= 2$, for which $\langle k(T) \rangle = e^2$, but in this run we also kept a record of the number of particles per chain at time $t=1$, and the averages of this number are indicated by the open circles on Fig. 1. This run was continued to $n = 8192$. It is seen that again the deviations are of the same order as to be expected on a purely random basis.

Machine errors can almost certainly be excluded, because of the known reliability of the machine for conditions under which it was being used, and because a large portion of one of the runs (the first) was repeated with exact reproduction of results.

The two sets of data on Fig. 1 do not agree, even though the same sequence of points in the 255-dimensional unit cube was used for the two runs. The explanation is that events in the time interval $0 \leq t \leq 1$ for a given chain are associated with different coordinates of the given point in the second run, where each chain was followed until $t=2$, because the coordinates of the point are in each case taken in sequence for the "random" numbers, but the events of the chain are not taken in chronological order in our setup of the calculation. (There is, of course, no reason why they should be.)

The strong correlation that will be noted among the deviations of successive points for a given run should occasion no surprise, because the successive samples are not independent.

We consider it paradoxical that our deviations are apparently no smaller than ones that would be obtained on a random basis, for our theorems on integrals show that at least in some respects our sequence of points in the unit cube is definitely more uniformly distributed than a random sequence. A more extensive trial calculation might throw some light on this, but was not deemed worthwhile at present, because it seems by now clear that the method, at least in its present form, has no practical superiority over Monte Carlo.

30

-30-