Reducing the Cost of Monte Carlo Analysis of Well Logging Data

J. A. Christen, P. Jiang, L. Li, J. L. Morales-Perez, J. Spanier

July 14, 1995

1. Problem Statement

A log in the oil industry is a record of measurements taken by instruments in an attempt to determine the geology through which a drill is passing or has passed. Typically an instrument package, called a sonde, is lowered by a cable attached to a truck in which readings are recorded and analyzed. In the case of the nuclear sonde, a source of neutrons or photons is included in the instrument package, as is a collection of sensitive detectors that are used to measure radiation in and around the borehole. Interpretations of these measurements are then required to assess the properties of the surrounding material.

Because of the complex geometries, the relatively low probabilities associated with the detection of scattered radiation, and the extreme accuracies required, nuclear well logging problems pose severe challenges for any numerical method used to solve the governing transport equations. These same difficulties lead naturally to the use of Monte Carlo methods for their solution since the three dimensional, irregular geometries, the accurate representation of the source, and the need for extensive nuclear and atomic cross section data pose no particular problems for such simulations. However, the Monte Carlo solutions often entail the processing of several million particle histories, at substantial computing costs, even when only steady-state problems are to be solved. And when pulsed neutron sources are employed, the resulting time-dependence of the solution drives these computing costs even higher.

For these reasons, the Chevron Petroleum Technology Company has for the past two years sponsored a research program at The Claremont Graduate School whose objective has been to accelerate the convergence of the Monte Carlo solutions. The research, which is being carried out partly by student-faculty teams in the Graduate School's Mathematics Clinic, has involved both the development of generic improvements in the basic methodology - methods designed to be applicable to a broad class of transport problems - and techniques that attempt to take account of the special features of the well logging applications [1].

In the first category, promising new hybrid Monte Carlo methods [2] have been developed with the potential for order of magnitude acceleration of the convergence of the Monte Carlo solution. Hybrid methods combine conventional pseudorandom implementations (i.e., simulations in which pseudorandom numbers are used to make all of the random walk decisions) with quasirandom implementations (i.e., simulations in which low-discrepancy sequences are used to make the decisions) in an attempt to capture the best features of both methods. This is possible because the pseudorandom rate of convergence is independent of the average number of collisions made by random walking particles, while the quasirandom convergence rates are asymptotically (i.e., for large numbers of histories) better, except in problems requiring many collisions per particle. Two new hybrid methods, a mixed method and a scrambled method, have been studied in a series of model transport calculations and show promise of substantial reduction in error per unit computing cost. While the research program at The Claremont Graduate School continues, the work performed so far has posed a number of interesting theoretical and practical questions, and help in dealing with these was solicited at the MPI Workshop. A few of these questions are mentioned below.

1. Optimization of the implementation of the mixed and the scrambled methods pose somewhat different challenges and require somewhat different considerations. In the case of the mixed method, the idea is simply to switch from a low-dimensional quasirandom sequence to a pseudorandom sequence when an optimum number of collisions has been reached for each random walk history. This optimum number will be problem-dependent, in general. How can this optimum number be estimated in each problem?

2. In the case of the scrambled method, there is the issue of finding the theoretically most precise, yet still practical, implementation strategy. This involves finding the best way to generate independent random permutations of the integers $1,...,N$, where N is the number of histories to be simulated. Methods tested so far have relied on the use of linear congruential pseudorandom algorithms to achieve this, but are flawed theoretically. Can improvements in this implementation strategy be found?

3. Research performed so far has concentrated on developing and testing strategies capable of order of magnitude gains in rate of convergence. Only limited timing comparisons have been made to date. It is desirable now to optimize the coding associated with the new strategies and conduct accurate timing tests of the new methods and conventional methods. What will detailed timing comparisons reveal about the relative efficiencies of the new and the old methods?

4. Use of quasirandom sequences is inherently incompatible with the use of rejection sampling methods, which are extensively employed in production Monte Carlo codes, especially for the generation of angular scattering distributions. There are many ways to eliminate rejection sampling, such as determination of the functional inverse (when easily possible), through transformation of the underlying random variables, or through (brute force) direct table lookup. Which of the many techniques available should be employed in connection with the new methods?

5. Can the adjoint Monte Carlo solution be employed to any advantage in well logging problems? More generally, is it advantageous to work back and forth between direct and adjoint simulations, perhaps using tallies from one to approximate an importance function for the other?

6. It would seem useful to employ artificial detectors in well logging problems: these would be used for the purpose of generating tallies of particles fairly far from the sonde to record the energy spectrum associated with particles that had already suffered a number of collisions from the source. Especially when used in concert with focused estimation techniques - for example - with expected value estimators found by calculating the expected contribution to the physical detectors from particles still relatively removed from the sonde - such techniques would appear to provide for further reductions in run times per unit accuracy for this difficult class of problems. What will it take to incorporate such estimators in existing Monte Carlo codes?

Efforts at the Workshop concentrated mainly on Question 2 above.

2. Improvements in Implementation of the Scrambled Sequence Algorithm

Use of the scrambled hybrid sequence requires the generation of the elements of the $(M+1)\cdot N-$ dimensional matrix of $d-$ dimensional quasirandom vectors

$$\begin{array}{ccccc}
q_1 & q_2 & q_3 & \circ\ \circ\ \circ & q_N \\
q_{N+P_1(1)} & q_{N+P_1(2)} & q_{N+P_1(3)} & \circ\ \circ\ \circ & q_{N+P_1(N)} \\
q_{2N+P_2(1)} & q_{2N+P_2(2)} & q_{2N+P_2(3)} & \circ\ \circ\ \circ & q_{2N+P_2(N)} \\
\circ & \circ & \circ & \circ\ \circ\ \circ & \circ \\
\circ & \circ & \circ & \circ\ \circ\ \circ & \circ \\
\circ & \circ & \circ & \circ\ \circ\ \circ & \circ \\
q_{MN+P_M(1)} & q_{MN+P_M(2)} & q_{MN+P_M(3)} & \circ\ \circ\ \circ & q_{MN+P_M(N)}
\end{array}$$

where $P_1, P_2, ..., P_M$ are M permutations chosen "at random" from the total number of $N!$ permutations of $\{1, ..., N\}$ available. That is, we require one-to-one functions

$$P_i : \{1, ..., N\} \rightarrow \{1, ..., N\},\ i = 1, ..., M$$

randomly chosen from among the full set of permutations. In all that follows, we shall interpret the phrase "at random" to mean, at least in terms of computer implementation, "chosen pseudorandomly". Thus, a standard way of defining such permutations would be to generate pseudorandomly N variables $X_1, X_{2,...,}X_N$ in the interval [0,1) and sort them into ascending order $X_{P(1)} < X_{P(2)} < ... < X_{P(N)}$. Such a reordering of the numbers defines a permutation P of $\{1,...N\}$ and repetition of this process, using independent $N-$tuples chosen from [0,1), will produce independent pseudorandom permutations of $\{1, ..., N\}$.

However, this essentially exact method of generating independent permutations selected uniformly from among the $N!$ available suffers from a number of practical disadvantages, especially when the magnitude of N (2^{20} or more in typical transport applications) is large:

1. For each permutation, storage of all N X_i 's is required.

2. Sorting a large number of numbers is computationally complex and time-consuming.

3. While the permutations are needed to define the rows of the matrix above, the vectors themselves are utilized columnwise in the simulation. Thus, storage of the entire matrix of $(M + 1)N$ vectors seems unavoidable.

In the ideal case, it might be more convenient to have a parametrization, $P_\alpha, \alpha \varepsilon I$, of all possible permutations that lends itself to a pseudorandom selection of α. However, unless such a parametrization could be implemented more efficiently than the sorting algorithm, it could not be used in practical transport simulations.

There are, however, relatively simple algorithms that produce permutations. One very effective such algorithm is based on the use of linear congruential pseudorandom number generators modulo N. Consider the congruence

$$P(n+1) \equiv aP(n) + b \pmod{N}, \qquad n = 1, ..., N$$

where $N = 2^k$, the *seed* $P(1) = n_0$ is arbitrary, the *multiplier* $a = 4i + 1$, the *additive constant* $b = 2j + 1$, and the integers i, j are arbitrary. From the theory of such linear congruential algorithms it is known that the integers $P(n)$ do not repeat before the full cycle of N integers is generated. This assures that the function P does, indeed, define a permutation. Quite clearly, these linear congruential algorithms have three constants - n_0, a and b - that can be chosen "at random" in order to accomplish random selections of permutations.

Use of the linear congruential algorithm also enables a contraction of the total storage required for the implementation of the scrambled hybrid method inasmuch as the integers $P(n)$ are calculated by means of a function evaluation one at a time as needed. Experimentation in Claremont with pseudorandom choices of the seed n_0 have provided good results. However, such an algorithm may be visualized as the selection of numbers on a circle, and pseudorandom selection of the seed only affects the choice of the starting point on the circle. The sequencing of the points along the circle is not affected at all by changing the seed. Based on this observation, it was suggested at the Workshop that improved results might be obtained by selecting the multiplier a pseudorandomly.

This suggestion was tried by recoding programs available at Claremont to accomodate the new strategy. This was then tested on a "model" transport problem (i.e., a transport problem whose exact solution is known) that had also been studied at CGS. The results are displayed graphically in Figure 1.

Figure 1 displays the absolute value of the error as a function of the number of random walks generated. A total of 8192 histories were attempted in this initial test. The heavy solid line represents the error in a simulation in which the seed, n_0, is varied pseudorandomly, while the lighter solid line graphs the error when the multiplier, a, is varied pseudorandomly. The improvement after 8192 histories is nearly an order of magnitude: the errors are 0.00697803 and 0.00083447, respectively. For comparative purposes, Figure 1 also exhibits the results when a conventional pseudorandom simulation is used; this is displayed as the dashed line in the fiugure. The error after 8192 pseudorandomly generated histories is 0.01845264. This impressive gain in accuracy with the new scrambled implementation certainly warrants further study.

It should be pointed out that only a fraction of the total number of possible

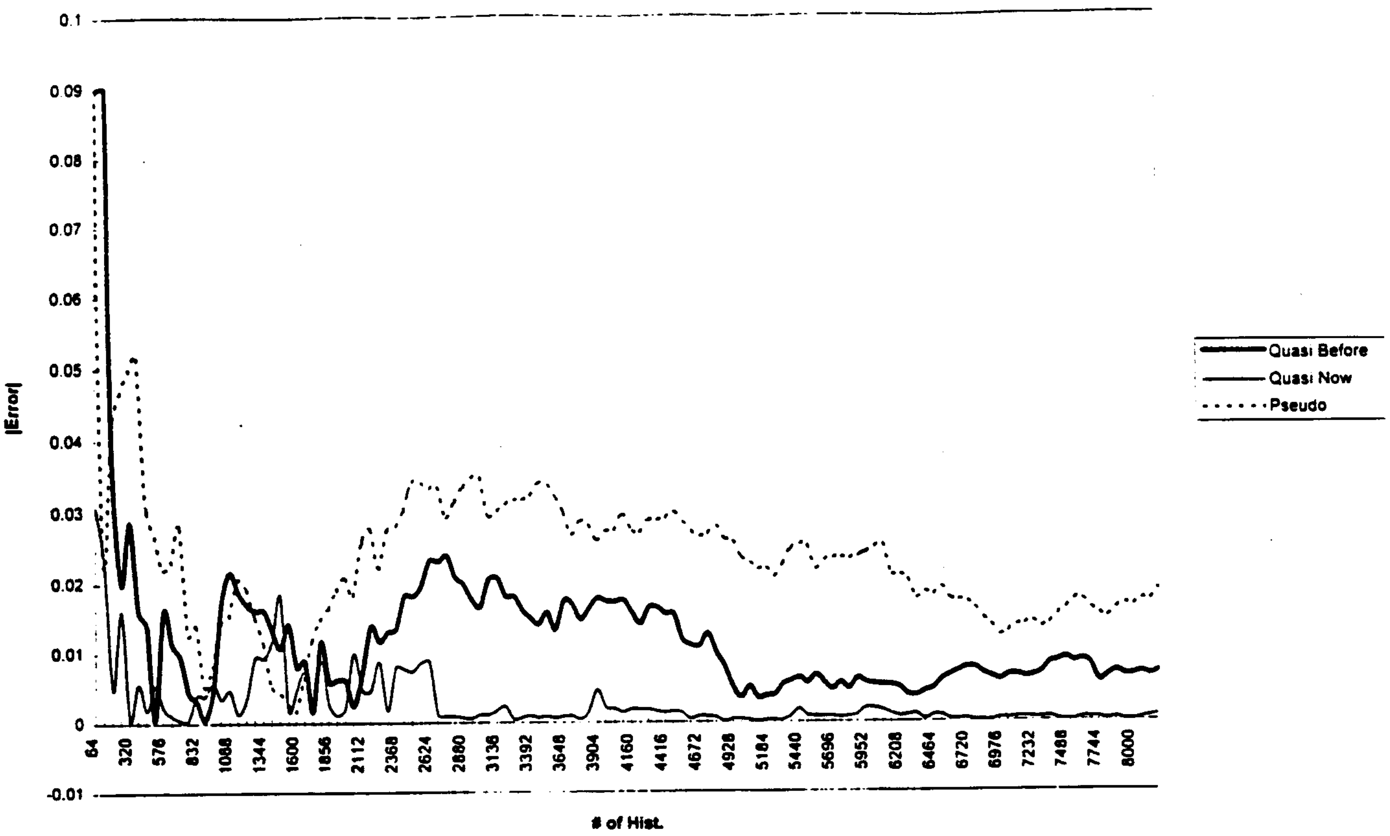

Figure 2.1: Comparison of Errors vs. Number of Histories: Model Problem

permutations can be generated in these ways using linear congruential algorithms. It is not difficult to see that selecting different integers n_0, a, and b (mod N) produces different permutations P. Thus, there are altogether $N^3/8(= N \cdot \frac{N}{4} \cdot \frac{N}{2})$ permutations accessible through such choices. While this is but a small fraction of the total number $N!$ of possible different permutations, it is hoped (and there is some evidence to support this) that the distributional properties of the permutations selected through the use of the linear congruential algorithm are quite good. This is especially important in problems for which $M \ll N$, which is typical in the transport applications of interest.

3. Ideas for Classifying Transport Problems

While the hybrid sequences were created to provide reductions generally in asymptotic rates of convergence, it is important to understand more fully the class of problems for which these methods are *certain* to provide substantial improvements, and by how much. The usual standard against which such comparisons would be made would be conventional pseudorandom Monte Carlo simulation. The key, then, would seem to be a deeper understanding of *comparable* error analyses and estimates for the rates of convergence of the scrambled hybrid method and the pseudorandom method (which presumably would go beyond the asymptotic $(\log N)^s/N$ versus $1/\sqrt{N}$ estimates), as well as *practical* methods for estimating the needed parameters (e.g., an "effective" dimension s to associate with the rate of convergence of the Neumann series of the transport equation, the probabilities p_k = exact probability that the Markov chain terminates after exactly k collisions), etc. as functions of sample size N).

A. Observation: The usual error analysis for quasirandom Monte Carlo methods is based on the Koksma-Hlawka inequality [3], which provides a *worst-case* error analysis. By contrast, the usual error analysis for pseudorandom Monte Carlo is statistical, and leads to *average-case* error estimates. However, the scrambled algorithm lends itself to conventional statistical analysis of error that could then be directly compared to the statistical analysis used for pseudorandom implementations. One need only factor the total number of histories, $N = nm$, and view the $N-$history experiment as consisting of m repetitions of a basic experiment of size n. That is, each subset of n histories is regarded as producing one sample from an approximately normal (for sufficiently large n) distribution with unknown mean μ and standard deviation σ. One can then compute the sample mean and sample standard deviation for each scrambled experiment and do the same for each pseudorandom experiment. These estimates would then provide an effective way of comparing estimates of the errors from the two implementations. Work of this sort, already begun at CGS, will continue.

B. Parameter Estimation: One might use output from each computer simulation to provide estimates of the key needed parameters. For instance, the numbers p_k could be estimated by simply counting the fraction, $\hat{p}_k$, of random walks that terminate in exactly k collisions. Then one could compute an effective dimension $\hat{s}$ as

$$\hat{s} = \sum_{k=1}^{\infty} \hat{p}_k s_k,$$

and an average second moment

$$s^2 = \sum_{k=1}^{\infty} \hat{p}_k s_k^2,$$

where s_k is the euclidean dimension associated with the kth term in the Neumann series for the transport problem. This information might then be used to see how to adjust $(\log N)^{\hat{s}}/N$ (which is much too conservative) downward to produce comparability with $1/\sqrt{N}$. This should help understand the potential for improving on pseudorandom Monte Carlo by using hybrid Monte Carlo methods such as the scrambled method.

4. Exploration of Preconditioning for Transport Simulations

Analogous to the preconditioning of direct and iterative methods for solving the linear algebraic system

$$\mathbf{x} = B\mathbf{x} + \mathbf{b}$$

we want, effectively, to replace this problem by one in which $\|B\|_1$ is reduced (since

$$(I - B)^{-1} = I + B + B^2 + \cdots$$

is a slowly convergent Neumann series when $\|B\|_1$ is close to 1). In the physical system, we think this might amount to introduction of artificial "absorption" with compensation through increased weight factors when particles survive absorption. The result would be a tradeoff between speed per history (which is lowered when absorption is increased) and accuracy per history (since the variance should increase when nonunit weights must be used to compensate for the increased absorption). This suggests studying idealized (model) problems in which the overall efficiency

$$E = RV$$

where R = run time, V = variance, provides the standard for comparison of methods with different R, V.

One idea would be to study this issue in the context of solving linear algebraic equations by Monte Carlo methods (see Chapter 2 of [4]). In a very simplified version of the general problem, consider

$$\mathbf{x} = B\mathbf{x} + \mathbf{b}$$

and assume that the problem is to estimate

$$I =< \mathbf{d}, \mathbf{x} >= x_{i_0}, \text{ i.e., } \mathbf{d} = \delta_{ii_0}.$$

Then x_{i_0} is subject to interpretation as the discrete collision density in energy state i_o. Assume for simplicity that the source $\mathbf{b}$ is normalized so that $b_i \geq 0$

and $\sum_{i=1}^{n} b_i = 1$ and that $b_{ij} \geq 0$, $\sum_{i=1}^{n} b_{ij} = 1 - p_j$, where p_j = probability of absorption is state j, b_{ij} = probability of scattering from state j to i. Let Ω be the space of all random walk histories and define the random variable

$$\Theta : \Omega \to R$$

by

$$\Theta(i_1, i_2, ..., i_k) = \delta_{i_k i_0}.$$

Then

$$\begin{aligned} E[\Theta] &= p_{i_0} b_{i_0} + p_{i_0} \textstyle\sum_{i_1} b_{i_0 i_1} b_{i_1} + p_{i_0} \sum_{i_2} b_{i_0 i_2} \sum_{i_1} b_{i_2 i_1} b_{i_1} + \cdot \\ &= p_{i_0} b_{i_0} + p_{i_0} (B\mathbf{b})_{i_0} + p_{i_0} (B^2\mathbf{b})_{i_0} + \cdots \\ &= p_{i_0} (\mathbf{b} + B\mathbf{b} + B^2\mathbf{b} + \cdots)_{i_0} \\ &= p_{i_0} x_{i_0}. \end{aligned}$$

Therefore,

$$\xi \equiv \Theta / p_{i_0}$$

is an unbiased estimator of x_{i_0} (see [4]).

To do importance sampling (see [4]) in which, for example, the transition matrix B is replaced by $\tilde{B}$ but the source vector $\mathbf{b}$ is unchanged, one would simply replace the element b_{ij} by $\tilde{b}_{ij}$ in order to determine the transition from j to i and multiply the particle weight W (initially set to 1) by the factor $b_{ij}/\tilde{b}_{ij}$. Similarly, when absorption in state k actually occurs, the weight must be multiplied by $p_k/\tilde{p}_k$, where

$$p_k = 1 - \textstyle\sum_{i=1}^{n} b_{ik} \text{ as before, and}$$
$$\tilde{p}_k = 1 - \textstyle\sum_{i=1}^{n} \tilde{b}_{ik}.$$

This weight multiplication results in a modified random variable W whose value on the Markov chain $(i_1, ..., i_k)$ is

$$W(i_1, ..., i_k) = \frac{p_{i_0}}{\tilde{p}_{i_0}} \frac{1}{p_{i_0}} \frac{b_{i_2 i_1}}{\tilde{b}_{i_2 i_1}} \cdots \frac{b_{i_k i_{k-1}}}{\tilde{b}_{i_k i_{k-1}}}.$$

It is easy to see that

$$E[W] = x_{i_0}$$

where here the expectation is taken with respect to the probability measure induced on Ω by using the matrix $\tilde{B}$, instead of B, to produce transitions from state to state.

All of the above can be generalized to include changes in the source distribution, **b**, as well as more general "detector" vectors, **d**, and more general random variables than simple binomial ones. Choosing model problems that are sufficiently simple should enable an exact calculation of the errors associated with the use of the random variables Θ and W.

Finally, one would like to impose a parametric description of the matrix $\tilde{B}$ (hence of the $\tilde{p}_j$) and study the behavior of the efficiency $E = RV$ of the method as a function of variation in this parametrization. Such a study would undoubtedly relate back to the questions encountered in Section 3 above.

References

[1] Hybrid Monte Carlo Methods Applied to Oil Well Logging Problems, CGS Mathematics Clinic Final Report to Chevron Petroleum Technology Company, May, 1995.

[2] J. Spanier, "Quasi-Monte Carlo Methods for Particle Transport Problems", Proc. Conf. on Monte Carlo Methods in Scientific Computing, Univ. Las Vegas, June 23-25, 1994 (to appear).

[3] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods,* CBMS-SIAM, 1992.

[4] J. Spanier and E.M. Gelbard, *Monte Carlo Principles and Neutron Transport Problems,* Addison-Wesley, 1969.