Proceedings of the 1997 Winter Simulation Conference
ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

COMPARISON OF MONTE CARLO AND DETERMINISTIC METHODS FOR NON-ADAPTIVE OPTIMIZATION

Hisham A. Al-Mharmah

Industrial Engineering Department
University of Jordan
Amman 11942, JORDAN

James M. Calvin

Department of Computer and Information Science
New Jersey Institute of Technology
Newark, NJ 07102-1982, U.S.A.

ABSTRACT

In this paper we compare the average performance of Monte Carlo methods for global optimization with non-adaptive deterministic alternatives. We analyze the behavior of the algorithms under the assumption of Wiener measure on the space of continuous functions on the unit interval. In this setting we show that the primary strength of the Monte Carlo methods (compositeness) is outweighed by the primary weakness (random gap size) when compared to efficient deterministic methods.

1 INTRODUCTION

The purpose of this paper is to analyze and compare the average performance of different non-adaptive algorithms for approximating the global minimum of one-dimensional real-valued functions defined on the unit interval. The average performance of an algorithm is defined as the expected difference between the observed minimum value and the actual global minimum. Brownian motion will be used as a probabilistic model for the one-dimensional continuous functions, and the objective function is to be taken as one realization of a Brownian motion process. We use an average-case framework, which can be thought of as averaging the error over many independent realizations of the algorithm on different objective functions. The concept of average optimality is more useful in this setting than a worst case analysis where the error can be arbitrarily large unless stringent conditions are placed on the set of objective functions, such as convexity or Lipschitz continuity.

To motivate the idea of average optimality consider the problem of verifying standard geometric tolerance specifications on a machined part, where the functional requirements or assembly conditions require that the entire surface must lie within two envelopes of ideal shape. A Coordinate Measuring Machine (CMM) is to be used in order to take several measurements on the part surface, and these measurements will be used to accept or reject the part. A CMM is a computer controlled device that consists of a programmable contact probe and a means of positioning the probe in three-dimensional space relative to the surface of a machined part. The machine will be used to take only a finite number of measurements on the part surface, but a decision about the acceptability of the entire part must be inferred from this sample. Considering the one-dimensional scenario where the number of measurements and their locations are to be found along one cross section, it may not be realistic to assume that the contours of all parts satisfy some strong regularity property (such as Lipschitz continuity). Therefore the set of measurements produced by any algorithm may, for a particular part, fail to identify arbitrarily large deviations from the tolerances. However, we can hope to design an algorithm so that the difference between measured and actual deviations is small on average.

In this paper we compare the average error of composite non-adaptive Monte Carlo-based algorithms as compared to composite deterministic algorithms. A composite non-adaptive algorithm is one that maintains its form as the number of observations increases; see Zhigljavsky (1991). If we denote the set of observations made by an algorithm up to time n by $T_n = \{t_1, t_2, \ldots, t_n\}$, then we will call an algorithm composite if $T_{n+1} \supset T_n$. A consequence of compositeness is that there is no need to determine in advance how many observations are to be taken in order to construct the observation set. In contrast, non-composite algorithms do not adapt gracefully as the number of observations changes. An example is the "uniform grid" algorithm (non-composite) that takes equally spaced observations; if a total of n observations are to be made, they are placed at $1/n$, $2/n$, …, 1. However, if the number of observations is increased to $n+1$, there is no way to add an observation point so as to maintain a uniform grid.

In the next section we introduce the problem and the terminology. In Section 3 we establish some background results concerning the distribution of the minimizer and path decomposition for Brownian motion. In Section 4 we compare the average performance of different Monte Carlo and deterministic algorithms.

2 NOTATION

Given the set of continuous real-valued functions defined on the unit interval $C([0,1])$, let $X \in C([0,1])$ be the objective function to be minimized. Also, let $M = \min\{X(t); t \in [0,1]\}$ denote its global minimum, and let $T = \inf\{t \leq 1 : X(t) = M\}$ be the (first) location where the minimum is attained. To approximate M we assume that we are allowed to observe the function X at n locations. Let $t_1, t_2, \ldots, t_n$ be the observation sites in $[0,1]$. Our goal is to compare the average performance of different algorithms based on their average approximation error $E(\Delta_n)$, where

$$\Delta_n = \min_{1 \leq i \leq n} X(t_i) - M.$$

Suppose that we have some prior knowledge about the relative likelihood of various functions, and that we can formalize this knowledge in a form of a probability measure μ on $C([0,1])$. Consequently, we can view any function $X \in C([0,1])$ as a sample path of a stochastic process. Hence,

$$E(\Delta_n) = \int_{X \in C([0,1])} \left(\min_{1 \leq i \leq n} X(t_i) - M \right) d\mu(X). \tag{1}$$

The Wiener measure on $C([0,1])$ will be taken as the probability distribution; i.e., X is taken to be a sample path of a Brownian motion process. It is natural to use a Gaussian measure, such as the Wiener measure, as a model for a random objective function that has multiple local minima with positive probability (see Wasilkowski 1992). Brownian motion is one of only a few non-trivial stochastic processes for which the distribution of the minimum is even known.

Our comparison will rely on results concerning processes and random variables associated with the 3-dimensional Bessel process. The 3-dimensional Bessel process is the diffusion process that is identical in law to the modulus of a 3-dimensional Brownian motion. A 3-dimensional Bessel bridge from $(0,0)$ to (t,y) is a 3-dimensional Bessel process starting from 0 at time 0 "conditioned to take the value y at time t"; see Revuz and Yor (1991). Define a "two-sided Bessel process" R by

$$R(t) = \begin{cases} R_1(t) & \text{if } t \geq 0, \\ R_2(-t) & \text{if } t \leq 0, \end{cases} \tag{2}$$

where R_1 and R_2 are two independent 3-dimensional Bessel processes. A new a random variable W will appear in the limit results we are going to present in Section 3. This random variable is defined as

$$W = \min_{i=0,\pm1,\pm2,\ldots} R(i+U), \tag{3}$$

where U is a uniformly distributed random variable on the unit interval independent of R. Also, $E(W) = -\zeta(1/2)/\sqrt{2\pi}$ (see Asmussen et al. 1995), where ζ is Riemann's zeta function. Finally, we will use $\Rightarrow$ to denote convergence in distribution; i.e., $X_n \Rightarrow X$ means that $Ef(X_n) \to Ef(X)$ as $n \to \infty$ for all bounded continuous functions f.

3 PROBABILITIES AND PATH DECOMPOSITION

For the Brownian motion prior, let $f(t;x,y)$ be the density of the first time the process reaches the level y given that $X(0) = x$. These densities are given by (see Karlin and Taylor 1975)

$$f(t;x,y) = \frac{\mid y - x \mid}{\sqrt{2\pi t^3}} \exp\left(- \frac{(y-x)^2}{2t} \right). \tag{4}$$

Theorem 1, proved in Imhof (1984), expresses the joint density of M, T, and $X(1)$, as the product of first hitting time densities.

Theorem 1 *For $x \geq y, 0 \geq y$, and $0 \leq t \leq 1$,*

$$P(M \in dy, X(1) \in dx, T \in dt) \\ = f(t;0,y) f(1-t;x,y)\, dy\, dx\, dt.$$

The marginal density ξ of T is the "arc-sine" density as shown in Feller (1971)

$$P(T \in dt) = \xi(t) = \frac{1}{\pi\sqrt{t(1-t)}}, \quad 0 < t < 1. \tag{5}$$

The following result is a special case of a general result of Fitzsimmons, Pitman, and Yor (1992) that decomposes the path of a diffusion at the minimum (this result generalizes an earlier result of Williams (1974)).

Theorem 2 *Given $(M = y, T = t, X(1) = x)$ $(0 < t < 1, y < x)$, the process $\{X(t+u) - y\}_{0 \leq u \leq 1-t}$ is a 3-dimensional Bessel bridge from 0 at time 0 to $x - y$ at time $1 - t$, independent of $\{X(t-u) - y\}_{0 \leq u \leq t}$, which is a 3-dimensional Bessel bridge from 0 at time 0 to $-y$ at time t.*

Let H be the cumulative density function of the Beta$(2/3, 2/3)$ distribution,

$$\begin{aligned} H(t) &= I(2/3, 2/3, t) \\ &= \mathcal{B}(2/3, 2/3)^{-1} \int_{s=0}^{t} \frac{ds}{[s(1-s)]^{1/3}}, \end{aligned}$$

for $0 \leq t \leq 1$, where $\mathcal{B}$ denotes the beta and I the incomplete beta function. Let $h(t) = H'(t) = \left(\mathcal{B}(2/3, 2/3)[t(1-t)]^{1/3}\right)^{-1}$ denote the corresponding probability density function. The following result, which is proved in Al-Mharmah and Calvin (1997), is used to determine the least upper bound and the largest lower bound on the convergence of the sequence $E(\sqrt{n}\Delta_n)$ for a composite algorithm as shown in the next section. The proof of this result relies on both Theorem 1 and Theorem 2.

Theorem 3 *Let $\{n_k : k \geq 1\}$ be a sequence of integers such that*

$$2^k \leq n_k < 2^{k+1}, \qquad k \geq 1,$$

and

$$t_{n_k} \to \tau \in (0, 1)$$

as $k \to \infty$. Let

$$\beta(T) = \begin{cases} 1 & \textit{if } T \geq \tau, \\ 2 & \textit{if } T < \tau. \end{cases}$$

Then

$$\sqrt{2^k \beta(T) h(T)}\, \Delta_{n_k} \Rightarrow W \tag{6}$$

as $k \to \infty$, where W is defined in equation (3), and

$$\begin{aligned} E\left(\sqrt{n_k}\Delta_{n_k}\right)^{\prime} \quad \to \quad & \sqrt{1 + H(\tau)}\, E(W) \frac{\mathcal{B}(2/3, 2/3)^{3/2}}{\pi} \\ & \times \left(1 - \left(1 - 2^{-1/2}\right) H(\tau)\right), \end{aligned}$$

as $k \to \infty$.

4 ERROR ANALYSIS

In the case of Brownian motion, it is shown in Calvin (1995) that if observations form a deterministic equi-spaced grid, then the error is about 82% as large as if the points are chosen at random uniformly over the unit interval. However, if new observations are to be added the uniformity of the deterministic grid will not hold at all times. One might expect, for example, that if the grid is such that $2k$ points are equi-spaced to the left of $1/2$ and k points are equi-spaced to the right of $1/2$, then choosing $3k$ points at random uniformly over the interval might give a smaller error on average.

Al-Mharmah and Calvin (1996) studied randomized non-adaptive algorithms, and found that the optimal distribution for placing points on the unit interval is Beta$(2/3, 2/3)$. The corresponding limiting normalized mean error for this random algorithm is

$$E\left(\sqrt{n}\Delta_n\right) \to \frac{1}{\pi\sqrt{2}} \mathcal{B}(2/3, 2/3)^{3/2} \approx 0.6623, \tag{7}$$

as $n \to \infty$. This distribution gives a slightly better convergence rate than choosing the sites according to the distribution of the maximizer, which is the arcsine distribution. Calvin (1996) showed that the n quantiles of the Beta$(2/3, 2/3)$ distribution (non-composite algorithm) are optimal within the class of deterministic non-adaptive algorithms. The average normalized error for the deterministic version is about 82% of that for the random algorithm (with the number of observations n predetermined). The advantage of the random algorithms is compositeness, and the disadvantage is the random gaps (the largest gap in uniformly distributed points is of order $\log(n)/n$, and because of length-biased sampling, the large gaps are more likely to contain the minimizer).

Consider the simple composite deterministic algorithm that chooses the following sequence of observation points (label the points as d_1, d_2, $d_3, \ldots$):

$$1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{1}{16}, \ldots. \tag{8}$$

(Recall that $X(0) = 0$, so that in effect we start off with the observation at 0.) If the number of observations is a power of 2, then the points form an equi-spaced grid, which one would expect to be efficient. Otherwise, the grid has some intervals twice as wide as others, which is clearly inefficient. The composite algorithm that we describe here uses the images of the above algorithm under a continuous transformation of the unit interval. This corresponds to making the points the quantiles of a beta distribution, which is in a sense optimal, as shown in Calvin (1996). Therefore, define an algorithm by $t_0 = 1 = H^{-1}(1)$, and

$$t_n = H^{-1}(d_n), \quad n \geq 1. \tag{9}$$

Thus the t_n's are the images under H^{-1} of the grid points defined in equation (8). We will show that this algorithm dominates the best random composite algorithm in the limit.

The sequence $E(\sqrt{n}\Delta_n)$ does not converge in this setting. However, we can determine its least upper bound and greatest lower bound as shown in Theorem 4. The proof of this theorem is found in Al-Mharmah and Calvin (1997), which is based on the limits of certain subsequences of $E(\sqrt{n}\,\Delta_n)$ shown in Theorem 3.

Theorem 4 *Under the algorithm described in this section,*

$$\begin{aligned}\limsup_{n\to\infty} E\left(\sqrt{n}\,\Delta_n\right) &= E(W)\frac{\mathcal{B}(2/3,2/3)^{3/2}}{\pi} \\ &\quad \times\frac{4\sqrt{2}-2}{3\sqrt{2}}\sqrt{\frac{2\sqrt{2}-1}{3\sqrt{2}-3}} \\ &\approx 0.5705,\end{aligned}$$

and

$$\liminf_{n\to\infty} E\left(\sqrt{n}\,\Delta_n\right) = E(W)\frac{\mathcal{B}(2/3,2/3)^{3/2}}{\pi} \approx 0.5457. \tag{10}$$

Therefore, as shown in the above theorem, a composite deterministic algorithm has a better average performance than the optimal random algorithm, see equation (7).

5 CONCLUSIONS

As shown in Al-Mharmah and Calvin (1996), the best performance among algorithms that choose observations independently from a fixed probability distribution is obtained from the beta distribution with parameters $(2/3, 2/3)$. Comparing this result with Theorem 4 shows that the deterministic composite algorithm defined in equations (8) and (9) has a better limiting performance in the sense that the lim sup of the normalized mean error is approximately 0.5705, considerably less than the limit in equation (7). In this case, the benefit of deterministic gaps outweighs the penalty of grid non-uniformity when n is not a power of 2.

ACKNOWLEDGMENTS

The second author was supported in part by the National Science Foundation under grant DMI-9500173.

REFERENCES

Al-Mharmah, H., and J. Calvin. 1996. Optimal random non-adaptive algorithm for optimization of Brownian motion. *Journal of Global Optimization* 8(1):81–90.

Al-Mharmah, H., and J. Calvin. 1997. A composite deterministic non-adaptive algorithm dominates the best random one. Submitted for publication.

Asmussen, S., P. Glynn, and J. Pitman. 1995. Discretization error in simulation of one-dimensional reflecting Brownian motion. *The Annals of Applied Probability* 5:875–896.

Calvin, J. 1995. Average performance of non adaptive algorithms for global optimization. *Journal of Mathematical Analysis and Applications* 191:608–617.

Calvin, J. 1996. Asymptotically optimal non-adaptive algorithms for minimization of Brownian motion. In *The Mathematics of Numerical Analysis,* ed. J. Renegar, M. Shub, and S. Smale, 157-163. American Mathematical Society, Lectures in Applied Mathematics, Vol. 32.

Feller, W. 1971. An introduction to probability theory and its applications, II, New York: Wiley.

Fitzsimmons, P. J., J. W. Pitman, and M. Yor. 1992. Markovian bridges: Construction, Palm interpretation, and splicing. In *Seminars in Stochastic Processes,* ed. E. Cinlar, K. L. Chung, and R. K. Getoor, Birkhäuser, Boston.

Imhof, J. P. 1984. Density factorization for Brownian motion meander, and the three-dimensional Bessel process, and applications. *Journal of Applied Probability* 21:500-510.

Karlin, S., and H. Taylor. 1975. A first course in stochastic processes, New York: Academic Press.

Revuz, D., and M. Yor. 1991. *Continuous martingales and Brownian motion.* Berlin: Springer-Verlag.

Wasilkowski, G. W. 1992. On average complexity of global optimization problems. *Mathematical Programming* 57:313-324.

Williams, D. 1974. Path decomposition and continuity of local time for one-dimensional diffusions, I, *Proceeding of London Mathematical Society*, Ser. 3, 28:738–68.

Zhigljavsky, A. 1991. *Theory of global random search*, Dordrecht: Kluwer.

AUTHOR BIOGRAPHIES

HISHAM A. AL-MHARMAH is an Assistant Professor in the Department of Industrial Engineering at the University of Jordan. He received a B.S. in Civil Engineering and an M.S. in Industrial Engineering from the University of Jordan in 1986 and 1990, respectively. He joined the School of Industrial and Systems Engineering at the Georgia Institute of Technology in 1991, and received a Ph.D. in Industrial Engineering in 1993. His research interests include applied probability and stochastic optimization.

JAMES M. CALVIN is an Assistant Professor in the Department of Computer and Information Science at the New Jersey Institute of Technology. He received a Ph.D. in Operations Research from Stanford University. His research interests include simulation output analysis and global optimization.