Proceedings of the 1994 Winter Simulation Conference
ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila

AVERAGE PERFORMANCE OF MONTE CARLO AND QUASI-MONTE CARLO METHODS FOR GLOBAL OPTIMIZATION

James M. Calvin

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332, U.S.A.

ABSTRACT

Passive algorithms for global optimization of a function choose observation points independently of past observed values. We study the average performance of two common passive algorithms, where the average is with respect to a probability on a function space. We consider the case where the probability is on smooth functions, and compare the results to the case where the probability is on non-differentiable functions. The first algorithm chooses equally spaced observation points, while the second algorithm chooses the observation points independently and uniformly distributed. The average convergence rate is derived for both algorithms.

1 INTRODUCTION

We consider the problem of locating the maximum of a real-valued function defined on the unit interval by observing the value of the function at a set of observation points. In this paper we consider *passive* algorithms that make no use of prior information in choosing the next observation site. The two algorithms we study are the *uniform grid* algorithm which takes equally spaced observations and the *random* algorithm which chooses the points as independent uniformly distributed random variables. Our purpose is to analyze and compare the average performance of these two algorithms. Our criterion for error is the difference between the global maximum and the maximum observed value up to time n.

This paper complements a previous study (Calvin 1993) that compared the same two passive algorithms under the assumption of a probability on non-differentiable functions (Brownian motion). The average performance of passive algorithms for smooth functions is of particular interest since many efficient sequential optimization methods are based on the assumption of a smooth objective function. The average performance of passive algorithms gives a base performance level with which to compare the average performance of more sophisticated adaptive algorithms.

The next section introduces the problem and the notation. The convergence properties for smooth functions are investigated in Section 3, and Section 4 gives an analysis of the average performance under the assumption of Wiener measure.

2 BACKGROUND

A starting point for defining global optimization procedures is a specification of a set $\mathcal{F}$ that contains the function f to be maximized. If the set $\mathcal{F}$ is small enough algorithms can be constructed that have an acceptable worst-case performance. Given a continuous real-valued function f defined on the unit interval $[0, 1]$, let $f^* = \max_{0 \leq t \leq 1} f(t)$ denote the global maximum of the function. Suppose we are allowed to choose n points $t_1, t_2, \ldots, t_n$ in the unit interval at which to observe the value of the function. Denote the maximum of the n function values by

$$M_n = \max_{1 \leq i \leq n} f(t_i). \tag{1}$$

Our goal is to choose the sites in such a way that M_n is a good approximation to f^*, where we define the approximation error by

$$\Delta_n^{\mathcal{A}}(f) = f^* - M_n \tag{2}$$

for algorithm $\mathcal{A}$. An example of a worst-case result (see Törn and Žilinskas 1989) is for $\mathcal{F}$ the Lipschitz continuous functions f for which $|f(x) - f(y)| \leq L|x - y|$. In this case the uniform grid algorithm ($t_k = (k-1)/(n-1)$) is worst-case optimal, with

$$\sup_{f \in \mathcal{F}} \Delta_n^G(f) = L/n. \tag{3}$$

(Here the superscript G signifies the uniform grid algorithm.) If the set $\mathcal{F}$ is too large for a worst case analysis (for example, if $\mathcal{F}$ is all continuous functions), then another criterion that can be used is that of average performance. A probability measure P is put on $\mathcal{F}$, and we use the average error

$$E\Delta_n^A(f) = \int_{f\in\mathcal{F}} (f^* - M_n)\, dP(f) \qquad (4)$$

to compare algorithms.

In this paper we will analyze only algorithms that choose all the sites without using knowledge of previously observed function values. Such algorithms are called passive or non-adaptive. The uniform grid algorithm is defined by $t_i = (i-1)/(n-1)$ for $i = 1, 2, \ldots, n$, and the random algorithm is defined by choosing the sites independently and uniformly distributed. When referring to a specific algorithm, we will write Δ_n^G for the uniform grid algorithm and Δ_n^R for the random algorithm. In the following sections we derive the normalized limiting distribution of Δ_n for the two algorithms under certain probabilistic assumptions.

3 PROBABILITIES ON SMOOTH FUNCTIONS

A typical setting for studying global optimization problems in one dimension is to consider functions defined on the unit interval. We will find it convenient to consider the space $C(\mathcal{T})$ of continuous periodic functions defined on the unit interval (equivalently, continuous functions on the circle group, which we take to be $[0,1)$ with arithmetic modulo 1). As a model for an unknown function confronted by a global optimizer, it is natural to assume that the probability distribution corresponds to a strictly stationary process; i.e.,

$$P\left(\cap_{i=1}^n \{f(t_i+\tau) \in B_i\}\right) = P\left(\cap_{i=1}^n \{f(t_i) \in B_i\}\right) \qquad (5)$$

for any n, τ, and Borel sets $B_1, \ldots, B_n$. (All arithmetic involving function arguments is understood to be modulo 1.)

There are several advantages to considering periodic functions. First, it allows us to deal with unconstrained instead of constrained optimization and still work with a compact set for the function's domain. Second, the distribution of the maximizer of a strictly stationary process on the circle is obviously uniformly distributed. This is not the case for strictly stationary processes on the line; the location of the maximizer of a stationary process over a finite interval is typically concentrated near the endpoints. Much of our analysis will rest on the assumption that the maximizer is uniformly distributed.

An example of a class of stationary continuous functions on $\mathcal{T}$ is the class of stationary Gaussian processes. The covariance function of a stationary Gaussian process on the circle is of the form

$$K(t) = \sum_j e^{2\pi ijt} p_j, \qquad (6)$$

for some probability $\{p_j : j \in \mathcal{Z}\}$. A real-valued process has the representation

$$f(t) = \sum_j u_j \cos(2\pi jt) + v_j \sin(2\pi jt), \qquad (7)$$

where the u_j, v_j are mutually orthogonal random variables with $u_j, v_j \sim N(0, p_j)$. The function f will be smooth, for example, if all but finitely many of the p_j's are zero in (6).

The question of average-optimal passive algorithm is non-trivial even for the case of stationary Gaussian processes. It is tempting to think that choosing observations according to a uniform grid is optimal for stationary processes. This is not the case, however, as the following example illustrates.

Example 1 Let $p_2 = 1$ in (6), so that (7) gives

$$f(t) = u_2 \cos(4\pi t) + v_2 \sin(4\pi t), \qquad (8)$$

or equivalently $f(t) = A\cos(4\pi(t-\theta))$, where θ is uniformly distributed between 0 and 1 and A has the Rayleigh distribution with density $xe^{-x^2/2}$ for $x \geq 0$; see Leadbetter et al. (1983). Since $f(0) = f(1/2)$, it is clear that the optimal algorithm for two observations does not observe at $0 = 1$ and $1/2$ (recall that we are "wrapping" the interval around a circle, so a uniform grid of 2 points is 0 and 1/2) since either observation contains the information of the other. □

Here we list the assumptions we make on the function class $\mathcal{F}$ and the probability P on $\mathcal{F}$. We take the probability distribution P to make the process strictly stationary. As a consequence the global maximizer x^* is uniformly distributed on the unit interval (or circle). For $y \geq 0$, set

$$G(y) = \lambda\{x : f^* - f(x) \leq y\}, \qquad (9)$$

where λ denotes Lebesgue measure. We assume that

$$\lim_{t\downarrow 0} \frac{G(ty)}{G(t)} = y^\alpha, \qquad (10)$$

for some $\alpha > 0$.

We will compare the two algorithms based on a fixed sample function f, and so our notation does not show the dependence of G and α on f. The randomness is in the choice of random observations sites in the case of the random algorithm, and in the choice of a random offset or phase for the uniform grid algorithm.

3.1 Uniform Grid Algorithm

In this section we analyze the average performance of the uniform grid algorithm. We begin with the limiting distribution of the (suitably normalized) random variable Δ_n^G with the function f (and thus G and α) fixed. Set $a_n = G^{-1}(1/n)$.

Theorem 1 *For $f \in \mathcal{F}$,*

$$\frac{\Delta_n^G(f)}{a_n} \Rightarrow \Delta^G \tag{11}$$

as $n \to \infty$, where

$$P(\Delta^G \le y) = \begin{cases} 1 & y \ge 1, \\ y^\alpha & y < 1, \end{cases} \tag{12}$$

and $\Rightarrow$ denotes weak convergence.

Proof: By assumption of a unique global maximizer, there exists a number $\bar{y}$ such that for $y \le \bar{y}$, the set $\{x : f^* - f(x) \le y\}$ is an interval containing x^*. For the uniform grid, if $y \le \bar{y}$,

$$P(\Delta_n^G \le y) = \begin{cases} 1 & G(y) \ge 1/n, \\ nG(y) & G(y) < 1/n. \end{cases} \tag{13}$$

Therefore,

$$P(\Delta_n^G/a_n \le y) = \begin{cases} 1 & G(a_n y) \ge 1/n, \\ nG(a_n y) & G(a_n y) < 1/n. \end{cases} \tag{14}$$

By (10),

$$\frac{G(a_n y)}{G(a_n)} = \frac{G(a_n y)}{1/n} \to y^\alpha,$$

and so

$$P(\Delta_n^G/a_n \le y) \to \begin{cases} 1 & y^\alpha \ge 1, \\ y^\alpha & y^\alpha < 1, \end{cases} \tag{15}$$

which is the desired result.

3.2 Random Algorithm

In this section we will analyze the average performance of the algorithm that chooses the observation sites independently and uniformly distributed. In this case Δ_n^R is the minimum of independent, identically distributed random variables with distribution function G, which belongs to the minimum domain of attraction of the Weibull distribution. Let Δ_α^R denote a random variable with the Weibull distribution with parameters 1 and α; that is,

$$P(\Delta^R \le x) = 1 - e^{-x^\alpha}, \quad x \ge 0. \tag{16}$$

The following result is a routine application of extreme value theory; see for example Theorem 3.3, p. 241 in Barlow and Proschan (1975).

Theorem 2 *For $f \in \mathcal{F}$,*

$$\frac{\Delta_n^R}{a_n} \Rightarrow \Delta_\alpha^R. \tag{17}$$

The modes of convergence of the two algorithms are different; the uniform grid algorithm converges for each $f \in \mathcal{F}$, while the random algorithm converges with probability one for each $f \in \mathcal{F}$.

If we compare the expected errors of the two limiting distributions derived in this section, we obtain that the limiting ratio of expected errors for the random and deterministic algorithms is

$$\frac{\alpha+1}{\alpha^2}\Gamma\left(\frac{1}{\alpha}\right). \tag{18}$$

In particular, if f is locally quadratic in a neighborhood of the global maximizer (so $\alpha = 1/2$), the deterministic grid is asymptotically 6 times as efficient as the random algorithm.

4 PROBABILITIES ON NON-SMOOTH FUNCTIONS

In the following sections we derive the mean of Δ_n for the two algorithms under the assumption that f has the Wiener distribution; i.e., f is taken to be a sample path of a Brownian motion process (we view $\{f(t) : t \in [0,1]\}$ as a stochastic process). The Wiener measure on $C([0,1])$ is characterized as follows. For each $t \in [0,1]$, $f(t)$ has the normal distribution with mean 0 and variance t, and for any

$$0 \le t_0 \le t_1 \le \cdots \le t_k \le 1, \tag{19}$$

the random variables $f(t_1) - f(t_0), f(t_2) - f(t_1), \ldots, f(t_k) - f(t_{k-1})$ are independent. It follows that the random variables $f(t_i) - f(t_{i-1})$ are normally distributed with mean 0 and variance $t_i - t_{i-1}$. Furthermore, the global maximum, f^*, has the same distribution as the absolute value of a standard normal random variable; i.e.,

$$P(f^* \le x) = \sqrt{\frac{2}{\pi}} \int_{y=0}^{x} e^{-y^2/2}\, dy, \qquad x > 0. \tag{20}$$

The mean is therefore $E(f^*) = \sqrt{2/\pi}$.

We now analyze the average performance of the uniform grid algorithm under the Wiener measure. The convergence of the uniform grid algorithm is clear; for any $f \in C([0,1])$, $M_n \uparrow f^*$ as $n \to \infty$.

The key to the following analysis is the fact that with the uniform grid algorithm, the maximum observed up to time n, M_n, is the maximum of a random walk with normally distributed increments.

The following two results are proved in Calvin (1993).

Theorem 3 *For the uniform grid algorithm,*

$$E(\Delta_n^G) = \frac{1 + \mathcal{C}/2}{\sqrt{2\pi n}} + O(1/n), \tag{21}$$

where

$$\mathcal{C} = \int_{t=1}^{\infty} \frac{t - \lfloor t \rfloor}{t^{3/2}} dt \approx 0.9207. \tag{22}$$

Theorem 4 *For the random algorithm,*

$$E(\Delta_n^R) = \frac{1}{\sqrt{2n}} + O(1/n). \tag{23}$$

Finally, we consider the mode of convergence of Δ_n^R to 0. Clearly, we can not make as strong a statement as we made for the uniform grid algorithm. However, for the random algorithm, $\Delta_n^R \rightarrow 0$ with probability one for every continuous function.

5 CONCLUSIONS

For the probabilistic analysis of the global optimization problem, several decisions must be made regarding sensible probabilities on problem instances. As we have seen, assumptions such as smoothness of the functions have a great impact on the average performance of algorithms. We have taken the view that in order to gain insight into the relative performance of general global optimization algorithms it is sensible to assume that the probability corresponds to a strictly stationary process.

We have analyzed only passive algorithms that make no use of past observed values in choosing new observation points. Instead of addressing the question of optimal passive algorithm for a particular probability (which is non-trivial) we simply compare the performance of two simple, commonly used passive algorithms; the simplest examples from the classes of deterministic and random algorithms. Aside from the relative efficiency, the random algorithm has perhaps one advantage in that the total number of observations need not be determined in advance for it to be employed effectively. A uniform grid can not be maintained as more observations are made.

Comparing the asymptotic expected error for the two algorithms for Brownian motion, the relative efficiency of the random algorithm compared with the uniform grid algorithm is approximately 82%; that is, the ratio of the expected errors converges to approximately 0.82 as the number of observations grows. This relative efficiency may seem surprisingly high. In contrast, the uniform grid algorithm is relatively more efficient for smooth functions.

REFERENCES

Barlow, R., and F. Proschan, 1975. *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, New York.

Calvin, J. 1993. Average performance of passive algorithms for global optimization of Brownian motion. (to appear in *Journal of Mathematical Analysis and Applications*).

Leadbetter, M., G. Lindgren and H. Rootzén, 1983. *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.

Törn, A., and A. Žilinskas, 1989. *Global Optimization.* Springer-Verlag, Berlin.

AUTHOR BIOGRAPHY

JAMES M. CALVIN is an assistant professor of Industrial and Systems Engineering at Georgia Institute of Technology. His interests include average complexity of global optimization and regenerative methods for simulation output analysis.