Journal of Global Optimization **8:** 81–90, 1996.

Optimal Random Non-Adaptive Algorithm for Global Optimization of Brownian Motion

HISHAM AL-MHARMAH and JAMES M. CALVIN*
School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, U.S.A. (email: calvin@isye.gatech.edu)

(Received: 16 May 1994; accepted: 17 April 1995)

Abstract. In this paper we study random non-adaptive algorithms for finding the maximum of a continuous function on the unit interval. We compare the average performance of different algorithms under the assumption of Wiener measure on the space of continuous functions. Placing the observations independently according to a Beta(2/3,2/3) density function is shown to be the optimal random non-adaptive algorithm. The performance is compared with other random and deterministic non-adaptive algorithms.

Key words: Global optimization, average performance, Brownian motion.

1. Introduction

In this paper we study non-adaptive algorithms for global optimization of continuous functions defined on the unit interval. These algorithms (also called passive algorithms) make no use of previously observed function values in choosing the next observation site. We confine our attention to the sub-class of such algorithms that results from choosing each observation site independently according to a fixed probability distribution. We call such algorithms random non-adaptive algorithms.

Our primary purpose is to describe the random non-adaptive algorithm that is optimal in an average sense. Our criterion for optimality is the expected difference between the global maximum and the maximum observed value after n observations. The expectation is taken with respect to the Wiener measure on the continuous functions; i.e., we view the function to be optimized as a sample path of a standard Brownian motion. There are two main reasons for our choice of Brownian motion as a model for the objective function. First, Brownian motion arises frequently as a model in diverse fields and many tools are available for its analysis. Brownian motion is one of only a few non-trivial stochastic processes for which the distribution of the maximum is even known. In our investigation of optimization algorithms, the assumption of Brownian motion allows us to carry out the calculations needed to obtain sharp results. The other reason for assuming Brownian motion (no doubt a partial consequence of the first) is that Brownian

* Research supported in part by the National Science Foundation under grant DDM-9010770.

motion has motivated the construction of many global optimization algorithms; an early example is given by Kushner (1964), and other examples are described by Törn and Žilinskas (1989). The average performance of some deterministic non-adaptive algorithms under the Wiener measure has been studied (see Ritter 1990 and Calvin 1994). While our primary goal is to determine the optimal non-adaptive random algorithm, we are also motivated by a desire to understand the performance differences between random and deterministic algorithms.

We use the average performance criterion because the class of objective functions (continuous functions on the unit interval) is too general to permit an interesting worst-case analysis. The average performance of non-adaptive algorithms is of interest because they are the easiest algorithms to implement and their performance gives a lower bound on performance to be expected from adaptive algorithms.

The next section introduces the problem and the notation. Section 3 describes the optimal random algorithm, and Section 4 contains concluding remarks.

2. Notation

Given a continuous real-valued function f defined on the unit interval, let t^* be a global maximizer and let $f^* = f(t^*) = \max\{f(t); t \in [0,1]\}$ denote its global maximum. Throughout this paper we will assume that we are allowed to make n observations of a function f to approximate its global maximum f^*. Let $t_1, t_2, \ldots, t_n$ be the observation sites in $[0,1]$. Denote the maximum of the n observed values by

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

and let t_n^* be a site where the function takes the value M_n. Our goal is to choose the sites in such a way that M_n is a good approximation to f^*.

In this paper we limit consideration to the class of random non-adaptive algorithms that locate the observation sites independently according to a fixed probability density function. Thus we identify algorithms with probability densities on the unit interval. For algorithm $\mathcal{A}$ (i.e., probability density $\mathcal{A}$), the approximation error random variable after n observations is defined by

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

Since we are interested in average performance, we compare different algorithms based on the expectation of the error random variable; i.e., we put a probability μ on some class of functions $\mathcal{F}$. Equivalently, we view the objective function $f \in \mathcal{F}$ as a sample path of a stochastic process and compare the average performance of different algorithms by comparing their average errors, $E\Delta_n^{\mathcal{A}}$, where

$$E(\Delta_n^{\mathcal{A}}) = \int_{f \in \mathcal{F}} (f^* - M_n(f)) \mathrm{d}\mu(f).$$

The Wiener measure will be taken as the probability distribution on $\mathcal{F} = C([0,1])$. For any $t \in [0,1]$, $f(t)$ has a normal distribution with mean 0 and variance t, and for any

$$0 \leq t_0 \leq t_1 \leq \ldots \leq t_k \leq 1,$$

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, with $f(t_i) - f(t_{i-1})$ normally distributed with mean 0 and variance $t_i - t_{i-1}$.

3. Optimal Random Algorithm

If the observations are chosen independently according to a distribution that is supported by the unit interval, then $\sqrt{n}E(\Delta_n)$ will converge to some constant that depends on the distribution. In this section we will find the distribution that minimizes the limiting constant; we call this the asymptotically optimal random non-adaptive algorithm. We use the qualifier "asymptotically" since the algorithm need not be optimal for a fixed finite number of observations.

The simplest random non-adaptive algorithm allocates the observations independently and uniformly on the unit interval. Calvin (1994) showed that for the uniform density,

$$E[\Delta_n^{uniform}] = \frac{1}{(2n)^{1/2}} + O(1/n). \tag{1}$$

The deterministic analog of this algorithm is the uniform grid algorithm where the observations are placed at equally spaced locations over the unit interval. Ritter (1990) showed that this method is optimal among non-adaptive algorithms for $n = 2$, but it is not optimal in general. (For $n = 3$, he showed that the optimal sites are approximately $t_1 = 0.3$, $t_2 = 0.7$ and $t_3 = 1$.) The non-optimality of the deterministic uniform grid algorithm suggests that a random algorithm can do better than to choose the sites uniformly distributed. The arcsine density, given by

$$h(t) = \frac{1}{\pi\sqrt{t(1-t)}}, \quad 0 < t < 1, \tag{2}$$

is a natural candidate for the optimal density since it is the density of the maximizer of a Brownian motion path. This density does give a better convergence rate than the uniform density, although we will show in this section that even faster convergence can be attained.

The key tools used to find the optimal random algorithm are furnished by Lemmas 3.2 and 3.3 below, which show that the limiting distribution of the suitably normalized error random variable over the unit interval given that the maximum is in subinterval T is the same as the limiting distribution of the normalized error over T. It follows that in the limit, the properly normalized expected error conditioning on the global maximum being in T depends only on the observation density within T.

The results in this section make use of properties of Brownian meander and the three-dimensional Bessel process. Roughly speaking, a Brownian meander on the interval $[0,T]$ is a Brownian motion "conditioned to be positive on $(0,T]$", while the three-dimensional Bessel process can be thought of as a Brownian motion "conditioned to be positive on $(0,\infty)$". Precise descriptions and properties are given in Imhoff (1984) and Revuz and Yor (1991).

Let $\{Y(t) : t \geq 0\}$ be a 3-dimensional Bessel process, and define an independent Poisson process with intensity 1 and points of increase $\{T_1, T_2, T_3, \ldots\}$. Set

$$Z = \min_{i\geq 1} Y(T_i). \tag{3}$$

The random variable Z will play an important role in the following lemmas. First we derive its distribution.

LEMMA 3.1. *For Z defined by (3),*

$$P(Z \leq y) = \frac{\left(1 - e^{-\sqrt{2}y}\right)^2}{1 + e^{-2\sqrt{2}y}} \tag{4}$$

for $y \geq 0$.

Proof. Let $L_y = \sup\{t : Y(t) = y\}$. Since Y is transient, $L_y < \infty$ a.s. The process $X(t) = Y(L_y - t)$, $0 \leq t \leq L_y$ has the same law as $\{B(t) : 0 \leq t \leq T_0\}$, where B is a Brownian motion starting at y and run until it hits zero; see Revuz and Yor (1991). The problem is therefore reduced to that of determining the law of the minimum of a Markov chain W_n that has the transition law of Brownian motion, sampled at exponentially distributed intervals and killed on hitting 0. Specifically, let r be the transition function of W. Then

$$\begin{aligned} r(y,z) &= \int_{t=0}^{\infty} e^{-t} P_y\left(B(t) \in dz, T_0 > t\right)\, \mathrm{d}t \\ &= \int_{t=0}^{\infty} e^{-t}\frac{1}{\sqrt{2\pi t}}\left[\exp\left(-\frac{(y-z)^2}{2t}\right) - \exp\left(-\frac{(y+z)^2}{2t}\right)\right] \mathrm{d}t \\ &= \begin{cases} \sqrt{2}\exp(-\sqrt{2}z)\sinh(\sqrt{2}y) & 0 < y < z, \\ \sqrt{2}\exp(-\sqrt{2}y)\sinh(\sqrt{2}z) & y > z > 0. \end{cases} \end{aligned}$$

Let $V(y) = P(Z \leq y)$, $y > 0$, and let $\tau \sim Exp(1)$. Then

$$V(y) = P_y\left(B(\tau) \in (0,y]\right) + P_y\left(B(\tau) > y\right) V(y), \tag{5}$$

and so

$$V(y) = \frac{\int_{z=0}^{y} r(y,z)\mathrm{d}z}{1 - \int_{z=y}^{\infty} r(y,z)\mathrm{d}z} = \frac{\left(1 - e^{-\sqrt{2}y}\right)^2}{1 + e^{-2\sqrt{2}y}}, \tag{6}$$

which completes the proof. ⊣

LEMMA 3.2. *Let $\{X(t) : 0 \le t \le \tau\}$ be a Brownian meander on $[0, \tau]$, and let $\{t_1, t_2, \cdots\}$ be independent and uniformly distributed over $[0, \sigma]$, where $0 < \sigma < \tau$. Let $U_n = \min_{1 \le i \le n} X(t_i)$ and $V = \min_{\sigma \le s \le \tau} X(s)$. Then*

$$\left(\frac{n}{\sigma}\right)^{1/2} \min\{U_n, V\} \Rightarrow Z \text{ as } n \to \infty, \tag{7}$$

where Z is defined by (3) and $\Rightarrow$ denotes convergence in distribution.

The random variables U_n and V are depicted in Figure 1.

Proof. Let $\{\tau_i : i \ge 1\}$ be the points of increase of a Poisson process with intensity 1, independent of X. Then, since $\{\tau_i/\tau_{n+1} : i = 1, \ldots, n\}$ has the same joint distribution as the order statistics of n points chosen independently and uniformly over the unit interval, U_n has the same distribution as $\widetilde{U}_n$ defined by

$$\widetilde{U}_n = \min\{X(\tau_1\sigma/\tau_{n+1}), X(\tau_2\sigma/\tau_{n+1}), \ldots, X(\tau_n\sigma/\tau_{n+1})\}. \tag{8}$$

For each $n \ge 1$, set

$$Y^n(t) = \begin{cases} \left(\frac{\tau_{n+1}}{\sigma}\right)^{1/2} X(\sigma t/\tau_{n+1}), & 0 \le t \le \tau_{n+1}, \\ \left(\frac{\tau_{n+1}}{\sigma}\right)^{1/2} X(\sigma), & t > \tau_{n+1}. \end{cases} \tag{9}$$

The processes Y^n converge in distribution to the Bessel process Y as $n \to \infty$, and rewriting (8) in terms of the Y^n's shows that U_n has the same distribution as

$$\widetilde{U}_n = \left(\frac{\tau_{n+1}}{\sigma}\right)^{-1/2} \min\{Y^n(\tau_1), Y^n(\tau_2), \ldots, Y^n(\tau_n)\}. \tag{10}$$

Now $(n/\tau_{n+1})^{1/2} \Rightarrow 1$, $\sqrt{n}\,V \to \infty$ with probability one, and because $Y^n \Rightarrow Y$,

$$\begin{aligned}\left(\frac{n}{\sigma}\right)^{1/2} \widetilde{U}_n &= \left(\frac{n}{\sigma}\right)^{1/2} \left(\frac{\sigma}{\tau_{n+1}}\right)^{1/2} \min\{Y^n(\tau_1), Y^n(\tau_2), \ldots, Y^n(\tau_n)\} \\ &= \left(\frac{n}{\tau_{n+1}}\right)^{1/2} \min\{Y^n(\tau_1), Y^n(\tau_2), \ldots, Y^n(\tau_n)\} \Rightarrow Z \end{aligned} \tag{11}$$

as $n \to \infty$. Therefore, $(n/\sigma)^{1/2} \min\{U_n, V\} \Rightarrow Z$, as was to be shown. ⊣

LEMMA 3.3. *Suppose that the observations sites are chosen independently according to the density g, where g is a simple function of the form*

$$g(t) = \sum_{i=1}^{m} c_i 1_{\{s_{i-1} < t \le s_i\}}, \quad 0 \le t \le 1, \ c_i \ge 0, \tag{12}$$

and $0 = s_0 < s_1 < s_2 < \cdots < s_m = 1$. Then for each $i = 1, 2, \ldots, m$,

$$\sqrt{n} E(\Delta_n | s_{i-1} < t^* < s_i) \to (2c_i)^{-1/2} \tag{13}$$

as $n \to \infty$.

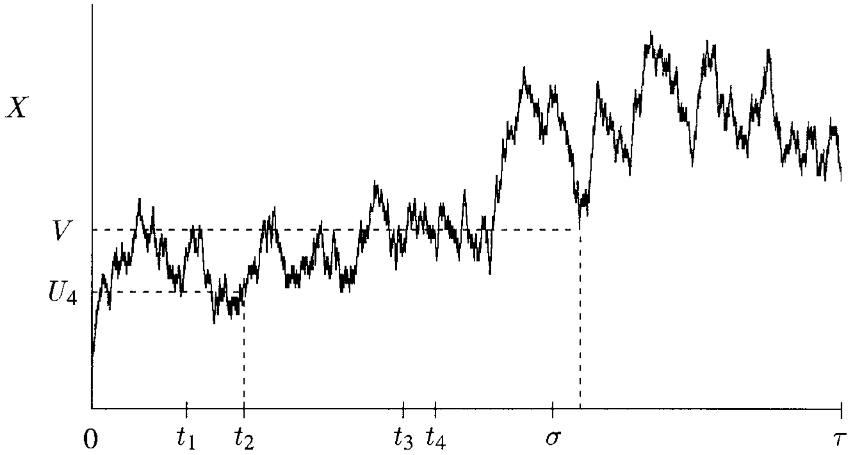

Fig. 1. Brownian meander process.

Proof. Define a mapping $H : C([0,1]) \to C([-1,1])$ by

$$Hf(t) = \begin{cases} \frac{f(t^*)-f(t^*(1+t))}{\sqrt{t^*}} & -1 \leq t \leq 0, \\ \frac{f(t^*)-f(t^*+(1-t^*)t)}{\sqrt{1-t^*}} & 0 \leq t \leq 1, \end{cases} \tag{14}$$

where

$$t^* = \inf\{t : f(s) \leq f(t) \ \forall s \in [0,1]\}. \tag{15}$$

If $t^* \in \{0,1\}$, then set $Hf = 0$. Note that t^* has a continuous distribution, and so the probability that it is equal to one of the s_i's is zero. Since we are interested in the expected error, we can (and do) ignore the possibility of the event $t^* \in \{s_i : 1 \leq i \leq n\}$ in the remainder of the proof.

If f is a Brownian motion, then Hf is equal in distribution to a "two-sided" Brownian meander (this transformation and result are described in Denisov 1984); that is, the sections of the trajectory of a suitably normalized Brownian motion to either side of the global maximizer are independent Brownian meanders. Intuitively, Brownian meander is Brownian motion conditioned to stay above 0, so the difference between the global maximum and the Brownian motion before and after the global maximum are independent Brownian meanders. The transformation H normalizes the two meanders so that they are both over a unit interval.

Let i be such that $s_{i-1} < t^* < s_i$, and let k_n^+ and k_n^- represent the number of the first n observations that are contained in the interval (s_{i-1}, s_i) and are below

t^* and above t^*, respectively:

$$\begin{aligned} k_n^- &= \#\{t_j,\, 1 \le j \le n : s_{i-1} < t_j < t^*\}, && (16) \\ k_n^+ &= \#\{t_j,\, 1 \le j \le n : t^* < t_j < s_i\}. && (17) \end{aligned}$$

Let Δ_n^- be the minimum value of the Brownian meander on $[-1, 0]$ at the points $\{(t_k - t^*)/t^* : t_k \le t^*\}$ (the image of the observation sites to the left of t^* under the transformation (14)), and similarly let Δ_n^+ be the minimum of the Brownian meander on $[0, 1]$ at the points $\{(t_k - t^*)/(1 - t^*) : t_k \ge t^*\}$. Then

$$\Delta_n = \min\left(\sqrt{t^*}\,\Delta_n^-, \sqrt{1 - t^*}\,\Delta_n^+\right). \qquad (18)$$

Therefore,

$$\begin{aligned} \sqrt{n}\,\Delta_n &= \min\left(\left(\frac{n(t^* - s_{i-1})}{k_n^-}\frac{k_n^- t^*}{t^* - s_{i-1}}\right)^{1/2}\right. && (19) \\ &\left.\Delta_n^-, \quad \left(\frac{n(s_i - t^*)}{k_n^+}\frac{k_n^+(1 - t^*)}{s_i - t^*}\right)^{1/2} \Delta_n^+\right). && (20) \end{aligned}$$

By the strong law of large numbers,

$$\lim_{n\to\infty} \frac{n(t^* - s_{i-1})}{k_n^-} = \lim_{n\to\infty} \frac{n(s_i - t^*)}{k_n^+} = c_i^{-1} \qquad (21)$$

with probability one, and by Lemma 3.2, each of the terms

$$\left(\frac{k_n^- t^*}{t^* - s_{i-1}}\right)^{1/2} \Delta_n^-, \quad \left(\frac{k_n^+(1 - t^*)}{s_i - t^*}\right)^{1/2} \Delta_n^+ \qquad (22)$$

converge in distribution to independent copies of the random variable Z described in (3). The expected minimum of two independent copies of Z is $1/\sqrt{2}$, and the conclusion of the lemma follows. ⊣

We are now ready for the main result which gives the optimal random non-adaptive algorithm under the Brownian motion assumption.

THEOREM 3.1. *The asymptotically optimal random algorithm for Brownian motion is given by the Beta distribution $Beta(2/3, 2/3)$; i.e., the points are chosen independently according to the density $g(t)$ given by*

$$g(t) = \mathcal{B}(2/3, 2/3)^{-1}[t(1 - t)]^{-1/3}, \quad 0 < t < 1, \qquad (23)$$

where $\mathcal{B}$ is the beta function,

$$\mathcal{B}(x, y) = \int_{t=0}^{1} t^{x-1}(1 - t)^{y-1}\mathrm{d}t. \qquad (24)$$

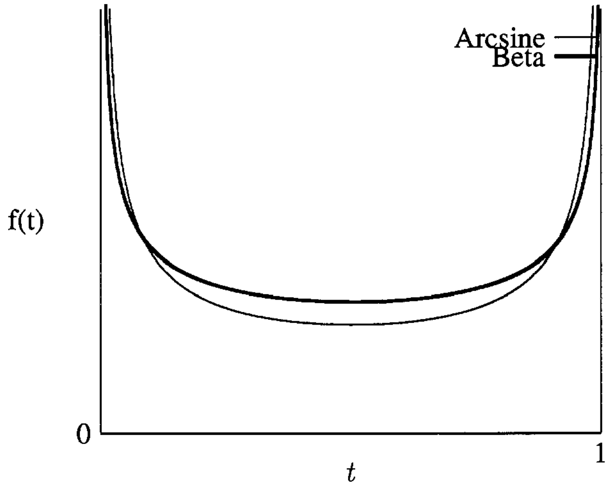

Fig. 2. Comparison between arcsine and beta(2/3,2/3) densities.

Proof. Fix an integer m and suppose that the observation sites are chosen independently according to the simple density g, given by

$$g(t) = \sum_{i=1}^{m} c_i \, 1_{\{s_{i-1}<t\leq s_i\}}, \quad 0 \leq t \leq 1,$$

where $0 = s_0 < s_1 < s_2 < \cdots < s_m = 1$. By Lemma 3.3,

$$\begin{aligned} \sqrt{n}E(\Delta_n) &= \sum_{i=1}^{m} \sqrt{n}E(\Delta_n|s_{i-1} < t^* < s_i)P(s_{i-1} < t^* < s_i) \\ &\rightarrow \frac{1}{\sqrt{2}} \sum_{i=1}^{m} c_i^{-1/2} P(s_{i-1} < t^* < s_i). \end{aligned}$$

Minimizing the above expression subject to the constraint that

$$\sum_{i=1}^{m} c_i(s_i - s_{i-1}) = 1$$

(that is, that g is a density), gives the optimal values as

$$c_i \propto \left(\frac{P(s_{i-1} < t^* < s_i)}{s_i - s_{i-1}} \right)^{2/3}.$$

The optimality of these c_i's follows from the Karush–Kuhn–Tucker conditions. Taking the limit as $\max_i |s_i - s_{i-1}| \rightarrow 0$ results in the optimal density

$$g(t) \propto h(t)^{2/3},$$

where h is the arcsine density. Therefore,

$$g(t) = \mathcal{B}(2/3, 2/3)^{-1}[t(1-t)]^{-1/3}, \quad 0 < t < 1.$$

$\dashv$

Having a beta other than arcsine as the optimal random algorithm contradicts the intuition that the best average performance would result from choosing observation sites according to the distribution of the maximizer. Figure 2 compares the arcsine and beta(2/3,2/3) densities, showing that the optimal distribution is "flatter" (closer to uniform) than the arcsine density.

The next Theorem compares the limiting normalized expected errors for the three non-adaptive algorithms discussed in this paper.

THEOREM 3.2. *Let Δ_n^{beta}, $\Delta_n^{arcsine}$, and $\Delta_n^{uniform}$ represent the errors after n observations chosen according to the Beta(2/3,2/3), arcsine, and uniform distributions, respectively. Then*

$$\begin{aligned}
\sqrt{n}E[\Delta_n^{beta}] &\to \frac{1}{\pi\sqrt{2}}\mathcal{B}(2/3, 2/3)^{3/2} \approx 0.662281, \\
\sqrt{n}E[\Delta_n^{arcsine}] &\to \frac{1}{\sqrt{2\pi}}\mathcal{B}(3/4, 3/4) \approx 0.675978, \\
\sqrt{n}E[\Delta_n^{uniform}] &\to \frac{1}{\sqrt{2}} \approx 0.707107.
\end{aligned}$$

The above results are interesting to compare with the convergence rate for the deterministic algorithm that places n points equally spaced. Calvin (1994) showed that for this algorithm,

$$\sqrt{n}E[\Delta_n] \to \frac{1 + C/2}{\sqrt{2\pi}} \approx 0.5826, \tag{25}$$

where

$$C = \int_{t=1}^{\infty} \frac{t - \lfloor t \rfloor}{t^{3/2}} \mathrm{d}t \approx 0.9207. \tag{26}$$

Thus the convergence is significantly faster with deterministic equal spacing.

An algorithm is called *composite* if it maintains its features when going from n to $n + 1$ observations (see Zhigljavsky 1991). The random non-adaptive algorithms are all composite, while the deterministic non-adaptive algorithm with equally spaced points is clearly not. Therefore, while the average performance for a fixed value of n is significantly better for the deterministic algorithm, the compositeness of the random algorithms will give them a relative advantage if the number of observations n is not fixed in advance.

4. Conclusions

The random non-adaptive algorithm that is optimal in the average sense for the Wiener measure is to take independent observations according to the Beta(2/3,2/3) distribution. This distribution gives a slightly better convergence rate than choosing the sites according to the distribution of the maximizer, which is the arcsine distribution. An informal explanation is that locally (at location t) the error decreases at rate $(ng(t))^{-1/2}$, where $g(t)$ is the observation density at t. In this sense, there are diminishing returns from increasing the observation density. Therefore, while more observations are placed where the probability of the maximizer is higher, the increase is less than that of the probability of the maximizer. How this heuristic reasoning extends to other probabilities is currently under investigation.

The difference in convergence rate between the different random non-adaptive algorithms is small compared with the improvement gained by using a deterministic non-adaptive algorithm. An important advantage of the random algorithms is that they are composite, unlike a deterministic algorithm. Our results suggest that perhaps the advantage of compositeness may not outweigh the efficiency advantage of deterministic algorithms. This topic will be pursued elsewhere.

References

Calvin, J. (1994), Average performance of passive algorithms for global optimization, in *J. Math. Anal. Appl.* **191**, 608–617.

Denisov, I. V. (1984), A random walk and a Wiener process near a maximum, *Theor. Prob. Appl.* **28**, 821–824.

Imhof, J.-P. (1984), Density factorizations for Brownian motion, meander and the three dimensional Bessel process, and applications, *J. Appl. Prob.* **21**, 500–510.

Kushner, M. J. (1964), A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise, *Journal of Basic Engineering* **86**, 97–106.

Revuz, D. and M. Yor (1991), *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin.

Ritter, K. (1990), Approximation and optimization on the Wiener space, *Journal of Complexity* **6**, 337–364.

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

Zhigljavsky, A. (1991), *Theory of Global Random Search*, Kluwer, Dordrecht.