Journal of Global Optimization **10:** 57–76, 1997.
© 1997 *Kluwer Academic Publishers. Printed in the Netherlands.*

Bayesian Algorithms for One-Dimensional Global Optimization

M. LOCATELLI
Università degli studi di Milano, Dipartimento di Scienze dell'Informazione, Via Comelico, 39/41-20135 Milano, Italy (email: locatelli@hermes.mc.dsi.unimi.it)

(Received 11 May 1993; accepted 29 April 1996)

Abstract. In this paper Bayesian analysis and Wiener process are used in order to build an algorithm to solve the problem of global optimization. The paper is divided in two main parts. In the first part an already known algorithm is considered: a new (Bayesian) stopping rule is added to it and some results are given, such as an upper bound for the number of iterations under the new stopping rule. In the second part a new algorithm is introduced in which the Bayesian approach is exploited not only in the choice of the Wiener model but also in the estimation of the parameter σ^2 of the Wiener process, whose value appears to be quite crucial. Some results about this algorithm are also given.

Key words: Bayesian analysis, Wiener process, stopping rule.

1. Introduction

We consider the problem of finding:

$$f^* = \max_{x \in [a,b]} f(x),$$

where $f : R \longrightarrow R$ is the objective function, assumed to be continuous, and $[a, b] \subset R$ is the feasible set. Many deterministic and probabilistic methods have been proposed to solve this problem (see [14] and [15]). Among them, those consisting in giving a stochastic model of the objective function, seem to be of particular interest. In them the objective function is seen as a particular realization of a stochastic process (see also [9]). Formally, given the stochastic process

$$\{f(x; \omega), x \in X, \omega \in \Omega\}$$

where Ω is a probability space, there exists a $\overline{\omega}$ such that the objective function $f(x)$ is equal to $f(x; \overline{\omega})$. For simplicity of investigation, the Wiener process is generally used as a stochastic process to model the objective function. The advantage of this process is mainly due to the simplicity of the formulae used to update the distribution of the random variables $f(x; \omega)$ of the process, after each observation of the objective function. Some possible critics to this choice will be discussed in Section 3. In Section 2 we give a quick description of the algorithm (see also [9] or [16]) with the introduction of a new stopping rule. We also give some results about the algorithm. Section 3 gives analogous results but for the case in which the parameter σ of the Wiener process is not *a priori* given, but only an a priori distribution is defined on it.

2. A New Stopping Rule

We first introduce some notation:

- $x_1 = a < x_2 < \cdots < x_{n-1} < x_n = b$ are the points where the function has been observed till iteration n;
- $f_1, \ldots, f_n$ are the corresponding function values;
- $f_n^* = \max\{f_1, \ldots, f_n\}$ is the record;
- $z_n = (x_1, \ldots, x_n, f_1, \ldots, f_n)$ is the vector of the information collected till iteration n;

We also recall that, if we model the function with the Wiener process with parameter σ, the distribution of the random variable $f(x)$, for $x \in [x_{i-1}, x_i]$ and conditioned on the information z_n is normal with mean:

$$\mu(x \mid z_n) = f_i \frac{x - x_{i-1}}{x_i - x_{i-1}} + f_{i-1} \frac{x_i - x}{x_i - x_{i-1}},$$

and variance:

$$v^2(x; \sigma \mid z_n) = \sigma^2 \frac{(x - x_{i-1})(x_i - x)}{x_i - x_{i-1}}$$

(see, e.g. [6], [9]).

We introduce the following function:

$$T_n(x; \sigma) = E[\max\{f(x) - f_n^*, 0\} \mid z_n] = \int_{f_n^*}^{\infty} (t - f_n^*)\, \mathrm{d}F_x(t), \tag{1}$$

where F_x is the normal distribution with mean $\mu(x)$ and variance $v(x; \sigma)$ (here and in what follows the conditioning on z_n of μ and v is understood). We can also write:

$$T_n(x; \sigma) = v(x; \sigma) \Psi \left(\frac{f_n^* - \mu(x)}{v(x; \sigma)} \right)$$

where

$$\Psi(x) = \int_x^{\infty} (t - x)\, \mathrm{d}\phi(t) = \varphi(x) - x[1 - \phi(x)], \tag{2}$$

ϕ is the standardized normal distribution and φ its density (see, e.g. [3] for more information about Ψ). We have that (1) can be interpreted as the expected gain, or the expected increment with respect to the record f_n^*, if we observe the function in point x.

In order to introduce a new stopping rule we need to introduce the loss function:

$$L(z_n; c) = -f_n^* + nc,$$

where c is the cost of an observation of the function. We want to stop the search at iteration n if the expected loss at iteration $n+1$ is greater or equal than the loss at iteration n, independently from where we put the next observation.

Therefore we have:

$$\text{stop if } \ L(z_n; c) \leq \min_{x \in [a,b]} E[L(z_n, x, f(x); c) \mid z_n]$$

or, equivalently:

$$\text{stop if } \ \max_{x \in [a,b]} T_n(x; \sigma) \leq c. \tag{3}$$

As an interpretation of this stopping rule we can say that we stop when the expected improvement over the current record is smaller than the cost of an observation of the function. It is questionable whether it is possible to find always a common unit of measure for the cost of an observation and the expected gain. All the same the form of the stopping rule is such that c can also be interpreted as a given accuracy and we stop when the expected accuracy is lower than c, as can be clearly seen in (3). The same cost structure was used, e.g. in [1]. Finally we give a quick description of the algorithm denoted with $\mathcal{A}(\sigma)$, in order to underline its dependence on the choice of the parameter σ:

at iteration n:

1. choose:
$$y = \arg \max_{x \in [a,b]} \ T_n(x; \sigma); \tag{4}$$
2. if $T_n(y; \sigma) \leq c$ then STOP otherwise go to 3.;
3. if $y \in (x_{i-1}, x_i)$ for some i, $2 \leq i \leq n$, then set $\forall \ j \geq i : x_{j+1} = x_j$ and $x_i = y$
4. evaluate f in y, set $z_{n+1} = z_n \cup (y, f(y))$ and $f^*_{n+1} = \max\{f^*_n, f(y)\}$;
5. go to the next iteration.

The rule for the selection of the next point at which the function has to be observed is given by (4) and it is called the *one-step* optimal approach, in which the next point is chosen in an optimal way assuming that it will be the last point at which the function will be observed. It would also be possible to think about *k-step* look-ahead rules, $k > 1$, in which at any iteration we should find where to put the next k observations in order to maximize the expected improvement over the current value of the record f^*_n (see, e.g. [9]). The problem is that the formulae become rather cumbersome even for $k = 2$ and it is not clear if this increase in difficulty carries better performance of the algorithm.

Now we give some results about the algorithm. First we need the following lemma which shows that the algorithm never puts observations "too close" to each other:

LEMMA 1. *Assuming that the function has been observed in points* $x_1, \ldots, x_n$, *then:*

$$\forall x \ : \ \exists i, \ 1 \leq i \leq n \ \textit{ such that } \quad | \ x - x_i \ | \leq \bar{c} \ : \quad T_n(x;\sigma) \leq c,$$

where

$$\bar{c} = \left(\frac{c\sqrt{2\pi}}{\sigma} \right)^2 .$$

Therefore the algorithm will never observe the function at these points.

Proof. First we note that:

$$T_n(x;\sigma) \leq v(x;\sigma)\Psi(0) = \frac{v(x;\sigma)}{\sqrt{2\pi}}$$

We now set $\Delta p = x_i - x$. Then we have:

$$\frac{v(x;\sigma)}{\sqrt{2\pi}} = \left(\frac{\sigma}{\sqrt{2\pi}} \right) \sqrt{\frac{(\Delta x - \Delta p)\Delta p}{\Delta x}}$$

We want to show that if $\Delta p < \bar{c}$ then the given limitation from above of the expected gain is smaller than c. It is equivalent to prove that:

$$\frac{\Delta p(\Delta x - \Delta p)}{\Delta x} \leq \left(\frac{c\sqrt{2\pi}}{\sigma} \right)^2$$

$$\frac{\Delta p(\Delta x - \Delta p)}{\Delta x} \leq \bar{c}$$

$$\Delta p(\Delta x - \Delta p) \leq \bar{c}\Delta x$$

We observe that if $\Delta p < \bar{c}$, being $(\Delta x - \Delta p) < \Delta x$, the inequality is true. That means the algorithm can not choose the next point in a position at a distance that is smaller than $\bar{c}$ from a point in which the function has been already observed. □

Now it is easy to give an upper bound for the number of iterations with the given stopping rule. We can not put a new observation at a distance smaller than $\bar{c}$ from points where the function has been already observed. That means we will certainly stop in, at most, $\frac{b-a}{\bar{c}}$ iterations. So we have proved the following theorem:

THEOREM 1. *The stopping rule*

$$\textit{stop if } \max_{x \in [a,b]} T_n(x;\sigma) \leq c$$

causes the algorithm to stop in a finite number of steps bounded from above by:

$$n^* = \frac{b-a}{\bar{c}} = \frac{\sigma^2(b-a)}{(c\sqrt{2\pi})^2}.$$

In particular it is possible to show that there exists at least one function (for example the constant one), for which the number of iterations has the order of magnitude of n^* for c tending to 0.

Indeed it can be shown for the constant function that the first $2^n + 1$ points divide $[a, b]$ into 2^n equal subintervals and the maximum expected gain in any subinterval is

$$\frac{v(\frac{x_{i-1}+x_i}{2}; \sigma)}{\sqrt{2\pi}} = \frac{\sigma}{2}\sqrt{\frac{\Delta x}{2\pi}}.$$

Then after n steps we have $\Delta x = O\left(\frac{b-a}{n}\right)$ and therefore the expected gain will be less than c if

$$n = O\left(\frac{(b-a)\sigma^2}{4(c\sqrt{2\pi})^2}\right) = O\left(\frac{1}{4}n^*\right).$$

Lemma 1 could also be exploited to show that:

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} (x_i - x_{i-1}) = 0, \tag{5}$$

i.e. the set of points at which the function is observed if the algorithm is never stopped, is dense in $[a, b]$. It is then immediate to prove consistency of the algorithm, i.e.:

$$\lim_{n\to\infty} f_n^* = f^*$$

(see, e.g. [2] or [16] for proofs of consistency).

3. The Case σ^2 Not a priori Given

Till now we have worked with the hypothesis of σ^2 *a priori* known. Now we want to turn to the more realistic situation in which σ^2 is not known and must be estimated through the observations (see also [4] and [5]). We want to point out the reason for this further development. The choice of σ is crucial for the good behaviour of the algorithm as it has been seen through experimentation (see the appendix). It is possible to understand that by studying the behaviour of the algorithm for $\sigma \to 0$ and $\sigma \to \infty$. In the first case all the observations tend to be concentrated around the best current point, so that we reduce to a local search and we lose any characteristic of globality of the algorithm. In the second case we reduce to a bisection method, where the point dividing the subinterval of maximum lenght is chosen and which has not any good local characteristic and does not take into account information given from function values. So we should avoid values of σ too small or too big. The problem is that "small" and "big" are not absolute concepts but are relative to the form of the function. A possible solution is to evaluate f in some points and find an estimate of σ (for example through an estimator inspired to the M.L.E., see [12]). The problem is that this kind of algorithm is generally meant for functions whose evaluation implies big costs, so that it would seem preferable to avoid supplementary evaluations. Here we take a Bayesian approach, by building a prior model both on the objective function and on the parameter of the Wiener process. The proposed algorithm is adaptive in the sense that exploits the information given by function values to give information about σ through the updating of a probability distribution function on σ. For other considerations about the algorithm and for a practical test which shows what has been said above, see also [8].

3.1. Updating formulae

We will give σ^2 not an *a priori* value but an *a priori* distribution which will be updated after every observation. We consider a sample $X_1, \ldots, X_n$ from a normal distribution with known mean m and variance σ^2 . It is well known that

$$Z = \frac{\sum_{i=1}^{n}(X_i - m)^2}{\sigma^2}$$

has a χ^2 distribution with n degrees of freedom (see, for example, [3]). As a consequence, we will assume that the *a priori* distribution of σ^2 is such that the distribution of $\frac{s_0}{\sigma^2}$ is a χ^2 with $a_0 = 2$ degrees of freedom, where $s_0 > 0$ is a value *a priori* fixed. It is important to see how the distribution of σ^2 is updated after every observation. We shall proof that $\frac{s_n}{\sigma^2}$ is distributed as a χ^2 with a_n degrees of freedom; this is true for $n = 0$ and will be demonstrated to be true for every n by induction, together with the formulae to update s_n and a_n. We remember that if $y = g(x)$, g is a one-to-one function and X has density $f_X(x)$, then Y has density

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y)) \tag{6}$$

for every y inside the range of the function g (see [10]). If we set

$$x = \frac{s_n}{\sigma^2},$$

then we have that the density of σ is proportional to:

$$\frac{1}{\sigma^{a_n+1}} e^{-\frac{1}{2}\frac{s_n}{\sigma^2}}.$$

Now using Bayes formula we obtain

$$g(\sigma \mid f(x_n)) \propto \frac{1}{\sigma^{a_n+1}}(e^{-(1/2)(s_n/\sigma^2)})\frac{1}{\sigma}e^{-(1/2\sigma^2)[(f(x_n)-\mu(x_n))/v(x_n;1)]^2} =$$

$$\frac{1}{\sigma^{a_n+2}} e^{-(1/2\sigma^2)[s_n + (\frac{f(x_n)-\mu(x_n)}{v(x_n;1)})^2]},$$

which means that $\frac{s_{n+1}}{\sigma^2}$ is distributed as a χ^2 with a_{n+1} degrees of freedom, where

$$a_{n+1} = a_n + 1$$

$$s_{n+1} = s_n + \left[\frac{f(x_n) - \mu(x_n)}{v(x_n; 1)}\right]^2 .$$

Now we can write a new algorithm, which we denote with $\mathcal{A}(s_0)$ to underline its dependence on s_0. The algorithm is equal to $\mathcal{A}(\sigma)$ but for the substitution of the expected gain $T_n(x;\sigma)$ with the expected gain:

$$E^{\sigma|z_n}[T_n(x;\sigma)],$$

or, equivalently from the point of view of the choice of the points at which the function is observed, with the function:

$$G_n(x;s_0) = \frac{1}{\sqrt{s_n}} E^{\sigma|z_n}[T_n(x;\sigma)],$$

which has been used in the practical tests. It is possible to obtain an explicit formula for the expected gain through heavy but elementary and non interesting computations (see [7]).

One question might arise at this point: what is the advantage of removing the dependence of the algorithm on the parameter σ, if, in doing this, we introduce a new parameter s_0 ?

The fact is that while a bad choice for σ has a constant influence on the algorithm through all the iterations, a bad choice for s_0 can be adaptively corrected by the observations. In other words σ is a fixed estimate while s_0 is only an initial guess which can be corrected. Therefore we expect that $\mathcal{A}(s_0)$ is less sensitive to the choice of the parameter. In appendix D this conjecture is made more clear through an example.

3.2. Finiteness of the algorithm and lack of consistency

Now we give an upper bound of the number of iterations after which the given stopping rule causes the algorithm to stop. We can prove the following theorem:

THEOREM 2. *Given the stopping rule*

$$\textit{stop if} \max_{x\in[a,b]} G_n(x;s_0) \le c,$$

then the algorithm stops in at most:

$$n^* = \frac{(b-a)p^2}{c^2}$$

iterations, where:

$$p = \sup \left\{\frac{1}{2}\gamma\left(\frac{n+2}{2}\right)\right\}.$$

Proof. See Appendix A.

In order to prove the lack of consistency of the algorithm we need to introduce a lemma:

LEMMA 2. *At any step and with any a priori distribution the maximum of the expected gain in the interval* $[x_{i-1}, x_i]$ *should be searched for, under the hypothesis* $\Delta f = f(x_i) - f(x_{i-1}) > 0$, *in the interval* $[\overline{x}, x']$, *where:*

$$\begin{aligned} \overline{x} &= \frac{x_{i-1} + x_i}{2} \\ x^{'} &= \overline{x} + \frac{\Delta f}{2(f_n^* - \overline{f})} \frac{\Delta x}{2} \\ \overline{f} &= \frac{f_{i-1} + f_i}{2}. \end{aligned} \tag{7}$$

Proof. We do not give the details of the proof, which are trivial (see [7]). By examinining the derivative:

$$T_n^{'}(x;\sigma) = v^{'}(x;\sigma)\Psi\left(\frac{f_n^* - \mu(x)}{v(x;\sigma)}\right) +$$

$$+v(x;\sigma)\Psi^{'}\left(\frac{f_n^* - \mu(x)}{v(x;\sigma)}\right)\left(\frac{f_n^* - \mu(x)}{v(x;\sigma)}\right)^{'},$$

we see that, for any value of σ, it is positive in $[x_{i-1}, \overline{x}]$ and negative in $[x', x_i]$, where x' is the point which minimizes:

$$\left(\frac{f_n^* - \mu(x)}{v(x;1)}\right)$$

Then we have:

$$\forall\, \sigma \text{ and } \forall\, x \in [x_{i-1}, \overline{x}] \;:\; T_n(\overline{x};\sigma) \geq T_n(x;\sigma),$$

and:

$$\forall\, \sigma \text{ and } \forall\, x \in [x', x_i] \;:\; T_n(x';\sigma) \geq T_n(x;\sigma).$$

Therefore the maximum of the expected gain is in $[\overline{x}, x']$. □

We also need an observation. Let t be a positive integer and:

$$\Gamma\left(\frac{t}{2}\right) = \int_0^\infty \left(\frac{1}{2}\right)^{t/2} y^{t/2-1} e{-1/2ydy}.$$

It is possible to find, e.g. in [10], the formulae for Γ but we can remind for instance that if t is even then $\Gamma\left(\frac{t}{2}\right) = \left(\frac{t}{2} - 1\right)!$ We have that:

OBSERVATION 1. *At iteration n, we have:*

$$E[\sigma] = \sqrt{s_n}\frac{\sqrt{\pi}}{\sqrt{2}}\gamma\left(\frac{a_n}{2}\right),$$

where:

$$\gamma\left(\frac{a_n}{2}\right) = \frac{\Gamma\left(\frac{a_n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{a_n}{2}\right)}.$$

Proof. Let $y = \frac{s_n}{\sigma^2}$, which has a distribution which is a χ^2 with a_n degrees of freedom. We have:

$$E[\sigma] = \sqrt{s_n}E\left[\frac{1}{\sqrt{y}}\right],$$

and:

$$E^y\left[\frac{1}{\sqrt{y}}\right] = \int_0^\infty \frac{1}{\sqrt{y}}\frac{1}{\Gamma(a_n/2)}(1/2)^{a_n/2}(y)^{a_n/2-1}e^{-(1/2)y}\,\mathrm{d}y =$$

$$\int_0^\infty \frac{1}{\Gamma(a_n)(2)^{a_n/2}}y^{(a_n-1)/2-1}e^{-(1/2)y}\,\mathrm{d}y =$$

$$\frac{\Gamma((a_n-1)/2)}{\sqrt{2}\Gamma(a_n)}\int_0^\infty \frac{1}{\Gamma((a_n-1)/2)(2)^{(a_n-1)/2}}y^{(a_n-1)/2-1}e^{-(1/2)y}\,\mathrm{d}y.$$

The last integral has an argument which is a χ^2 density, so it is equal to 1. So we have:

$$E^y\left[\frac{1}{\sqrt{y}}\right] = \frac{\Gamma((a_n-1)/2)}{\sqrt{2}\Gamma(a_n/2)}.$$

which proves the observation. □

Now we want to show that the method is not consistent, that is the sequence $\{f_n^*\}$ doesn't always converge to the maximum of the function. An example of that is given by this function $f : [0, 2P] \to [0, 2]$:

$$f(x) = \begin{cases} 8x & \text{for } 0 \le x \le \frac{1}{4} \\ \frac{(4-16Q)x+8Q-1}{2Q} & \text{for } \frac{1}{4} \le x \le \frac{1}{2} \\ \frac{1}{Q}x & \text{for } \frac{1}{2} \le x \le Q \\ 1 & \text{for } Q \le x \le 2P \end{cases} \tag{8}$$

where Q is the third point at which the function is observed, being the first two points:

$$x_0 = 0 \quad x_1 = 2P$$

Remembering Lemma 2, we must have $P \leq Q < 2P$. Actually it can be seen that as we increase P, Q tends to be more and more close to P.

We notice that

$$f(x) = \mu(x \mid z_2) \text{ for } \frac{1}{2} \leq x \leq 2P.$$

Now we let the algorithm run till we have to observe the function in the interval $(0, 1]$ at a certain step n_1. Since we have:

$$\forall\, n < n_1\; f(x) = \mu(x \mid z_n) \text{ for } x \geq 1,$$

then s_n remains constant. In case we never observe f in the interval $(0, 1]$, then the sequence $\{f_n^*\}$ will converge to 1 instead of the maximum 2 attained in $x = \frac{1}{4}$. Otherwise we indicate with x'' the first point in $(0, 1]$ in which the function is observed (at step n_1). We should notice that, since $\mu(x \mid z_{n_1-1})$ is increasing between 0 and the lowest point greater than 1 in which the function has been observed before step n_1, we must have $x'' \geq \frac{1}{2}$ (see Observation 1). Therefore $f(x'') = \mu(x'' \mid z_{n_1-1})$ and $f(x'') \leq \frac{1}{P}$. At step n_1 we have in the intervals between Q and $2P$:

$$G_n(x; s_0) = \frac{1}{\sqrt{s_n}} v(x; 1) E[\sigma\Psi(0)] = \frac{1}{\sqrt{s_n}} \frac{1}{\sqrt{2\pi}} v(x; 1) E[\sigma].$$

The maximum is in the middle point of the intervals and, in view of Observation 1 , has the value:

$$\left(\frac{\sqrt{\Delta x}}{4}\right) \gamma(a_{n_1}), \tag{9}$$

where Δx can be at worst of the order of $\frac{1}{n_1}$ (indeed n_1 observations between Q and $2P$ divide this interval in $n_1 + 1$ subintervals so that the ratio between the width of any two subintervals can be only 1 or 2 for any n_1 and so the dimension of any subinterval is between $\frac{1}{2n_1}$ and $\frac{2}{n_1}$). We now consider the interval $0 \leq x \leq x''$. We have that:

$$G_n(x; s_0) \leq \frac{1}{\sqrt{s_{n_1}}} v\left(\frac{x''}{2}; 1\right) E\left[\sigma\Psi\left(\frac{1 - \mu(x')}{\sigma v(x'; 1)}\right)\right],$$

where x' is given in Lemma 2 and the upper bound is due to the maximization of $v(x; 1)$ and the minimization of the argument of Ψ in the interval. From (2) we also have

$$G_n(x; s_0) \leq \frac{1}{\sqrt{s_{n_1}}} v\left(\frac{x''}{2}; 1\right) E\left[\sigma\varphi\left(\frac{1 - \mu(x')}{\sigma v(x'; 1)}\right)\right],$$

where the right side of the inequality is equal to:

$$v\left(\frac{x''}{2}; 1\right) \frac{\gamma(a_{n_1})}{2} \left(\frac{1}{[1 + t^2]^{(a_{n_1}-1)/2}}\right),$$

with:

$$t = \frac{1-\mu(x^{'})}{\sqrt{s_{n_1}}v(x^{'};1)} > 0.$$

Therefore $G_n(x;s_0)$ is limited from above by a quantity of this kind:

$$h\gamma(a_{n_1})\frac{1}{[1+t^2]^{(a_{n_1}-1)/2}},$$

where h is a constant. For n_1 great enough this is lower than the expected value in the intervals between Q and $2P$. So the next point will be chosen between Q and $2P$ or between $x^{''}$ and Q but not between 0 and $x^{''}$. By induction it is possible to show that this is still true for any $n \geq n_1$ and so we will never observe the function between 0 and $x^{''}$ nor, therefore, between 0 and $\frac{1}{2}$ where it attains its maximum. We observe that in order to have n_1 great enough, we should take P great enough. Indeed, by observing that $\forall\ i \leq n_1,\ \mu(x \mid z_i)$ is non decreasing, we must have

$$x_2 = Q \geq P,\ x_3 \geq \frac{P}{2},\ \ldots,\ x_i \geq \frac{P}{2^{i-2}},\ \ldots$$

and therefore $n_1 > \log P$.
Then finally we have the following observation:

OBSERVATION 2. *If the next point in which we observe the function is the one which maximizes $G_n(x;s_0)$, the method is not consistent, that is functions exist for which the sequence $\{f_n^*\}$ does not converge to the maximum of these functions.*

3.3. A WAY TO RECOVER CONSISTENCY

The reason for the lack of consistency for the previous function is the equality of $f(x)$ and $\mu(x \mid z_i)$ in $[\frac{1}{2}, 2P]$. That implies s_n remains constant while a_n increases, so the distribution of σ will become more and more concentrated in the neighbourhood of 0, reducing the uncertainty and the expected gain not only in $[x^{''}, 2P]$, but also in $[0, x^{''}]$ where $\mu(x \mid z_i)$ is different from $f(x)$ and the function is never observed. Getting more deeply inside the reason for the lack of consistency, we should note that the concentration of the distribution of σ around 0 is due to the fact that we considered a "regular" function as the path of a Wiener process, where these paths are with probability 1 extremely irregular (for example almost nowhere differentiable, oscillating a lot, etc.; for more information see [11]) The modification below is made necessary by this lack of fit of the model to the most common real situations. On the other hand it seems to be quite difficult to find models which fit the real cases and as handy as the Wiener process. In what follows we will try to remove the lack of consistency but before we want to underline that the concentration of the distribution of σ around 0 is not so bad from a practical point of view. A σ close to 0 means that the algorithm performs a local search around the best point as it has been said before, and a final local search around the record seems to be a good way to end the algorithm. Of course this final search should not start too early or too late and that depends on the choice of s_0 as it is more extensively shown in the appendix. Now we see how we can remove the lack of consistency. The main problem is that the set of point $\{x_i\}$ at which the function is observed according to the algorithm is not dense, that is:

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} (x_i - x_{i-1}) > 0.$$

A possible way to avoid this problem is to change the distribution of σ so that it can not become lower than a fixed $\epsilon > 0$. We assume that the *a priori* distribution of σ is such that $\frac{s_0}{\sigma^2}$ has a distribution with density:

$$f_{\frac{s_0}{\sigma^2}}(x) = \begin{cases} \frac{1}{\beta\left(\frac{a_0}{2}\right)} \left(\frac{1}{2}\right)^{a_0/2} x^{a_0/2-1} e^{-(1/2)x} & \text{for } 0 \leq x \leq \frac{s_0}{\epsilon} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\beta\left(\frac{a_0}{2}\right)$ is the normalizing constant. We also introduce the following hypothesis:

$$\begin{aligned} &\exists\, L \;:\; \forall x_1, x_2 \in [a,b] \\ &\frac{\mid f(x_1) - f(x_2) \mid}{\sqrt{\mid x_1 - x_2 \mid}} \leq L \end{aligned} \tag{11}$$

We can prove the following theorem:

THEOREM 3. *If the a priori distribution of σ is such that (10) holds and if (11) holds too, then the algorithm is consistent.*

Proof. See Appendix B.

We also mention another possible way of recovering consistency. While leaving unchanged the distribution of σ and therefore forgetting the above development, we could change the loss function and choose the point which minimizes the expected loss. The new loss function takes into account whether $[a,b]$ has been well globally explored or not:

$$L(z_n; c) = -f_n^* + nc + \epsilon \max_i (x_i - x_{i-1}),$$

with $\epsilon > 0$. However in this case the risk is that after a certain step the algorithm behaves like bisection. It could be interesting to study what happens if we substitute ϵ with $\epsilon_n \to 0$ at a certain rate, which could avoid the drawback outlined above.

4. Conclusion

In this paper the problem of global optimization has been faced by giving a stochastic model of the objective function, which has been seen as a particular realization of a stochastic process, the Wiener process, and by defining a loss function, which, through a Bayesian analysis, has been employed both to choose the point at which the function should be observed at any step, and to define a stopping rule. The resulting algorithm tries to get as much information as possible from the values of the function in the points already selected for the observation, in order to choose the next point at which the function can be observed. This approach is useful when the cost of an observation is high so that it is worthwhile to spend a good deal of resources for an accurate selection of the points at which the function is observed. Some issues concerning the algorithm have been analyzed such as the upper bound of the number of iterations of the algorithm with the proposed stopping rule. Moreover, the Bayesian approach has been exploited not only in the choice of the Wiener model but also in the estimation of σ, the parameter of the process; previous approaches found in the literature were inspired to classical M.L.E.. Results similar to the case of the fixed parameter have been obtained, but an additional hypothesis about f and a modification of the a priori distribution have been necessary to prove consistency. We also have discussed the problems connected with the use of the Wiener process which, because of its properties, does not seem to fit the most common real situations. Another possible critic is the difficulty of extension to the multidimensional case. A possible answer can be the use of Peano maps to transform a multidimensional problem in a one-dimensional problem. They are used to transform the multidimensional problem in one-dimensional with a rather oscillating objective function, which is not so bad since Wiener paths oscillate a lot, but there is a loss of information in the transformation, for example points which are close may be transformed in points which are far from each other. For a deeper discussion about the subject see [13].

Appendix

A. Proof of Theorem 3

Proof. We set $a_0 = 2$ and $s_0 > 0$. We have:

$$G_n(x; s_0) \leq \frac{1}{\sqrt{s_n}} \frac{v(x;1)}{\sqrt{2\pi}} E^{\sigma}[\sigma]$$

and in view of Observation 1:

$$G_n(x; s_0) \leq \frac{v(x;1)}{2} \gamma \left(\frac{a_n}{2} \right) .$$

Notice that $a_n = n + 2$. Using the asymptotic results for $n!$, we see that:

$$\gamma \left(\frac{n+2}{2} \right) = O \left(\frac{1}{\sqrt{n}} \right) ,$$

so that $\gamma\left(\frac{n+2}{2}\right)$ is convergent and must have a finite superior extreme p. So, independently from n, the expected gain is limited from above by

$$v(x;1)p.$$

Now we are in the same situation of the case σ^2 *a priori* known and it is easy to show that we will stop in a number of steps which is not greater than:

$$n^* = \frac{(b-a)p^2}{c^2}$$

(to show this remember the proofs of Lemma 1 and Theorem 1 for the case σ a priori known).

B. Proof of Theorem 3

We first need two lemmas:

LEMMA 3. *If (11) holds, then:*

$$s_n \leq s_0 + 4nL^2.$$

Proof. We observe that the term

$$\left[\frac{f(x_n)-\mu(x_n)}{v(x_n;1)}\right]^2$$

can also be written in the following way:

$$\left[\frac{f(x_n)(x_n-x_{i-1})+f(x_n)(x_i-x_n)-f_i(x_n-x_{i-1})-f_{i-1}(x_i-x_n)}{\sqrt{\Delta x}\sqrt{(x_n-x_{i-1})(x_i-x_n)}}\right]^2 =$$

$$= \left[\frac{(f(x_n)-f_i)(x_n-x_{i-1})+(f(x_n)-f_{i-1})(x_i-x_n)}{\sqrt{\Delta x}\sqrt{(x_n-x_{i-1})(x_i-x_n)}}\right]^2,$$

which can be bounded from above by:

$$\leq \left[\left|\frac{(f(x_n)-f_i)}{\sqrt{x_i-x_n}}\right|\sqrt{\frac{(x_n-x_{i-1})}{\Delta x}} + \left|\frac{(f(x_n)-f_{i-1})}{\sqrt{x_n-x_{i-1}}}\right|\sqrt{\frac{(x_i-x_n)}{\Delta x}}\right]^2,$$

which, in view of (11) is not greater than $[L+L]^2 = 4L^2$.

Therefore:

$$s_n \leq s_{n-1} + 4L^2 \leq s_0 + 4nL^2.$$

Let $t_n = \frac{a_n}{2}$, where a_n is a positive integer and:

$$\beta(t_n) = \int_0^{\frac{s_n}{2\epsilon}} \left(\frac{1}{2}\right)^{t_n} y^{t_n-1}e^{-y}\,\mathrm{d}y.$$

We can prove that:

LEMMA 4. *For n big enough:*

$$\frac{\beta(t_n - 1/2)}{\beta(t_n)} \leq 1.$$

Proof. After the change of variable $x = \frac{1}{2}y$, what we have to prove is that:

$$\int_0^{\frac{s_n}{2\epsilon}} x^{(t_n-3/2)} e^{-x}\,\mathrm{d}x \leq \int_0^{\frac{s_n}{2\epsilon}} x^{t_n-1} e^{-x}\,\mathrm{d}x. \tag{12}$$

We observe that $\frac{s_n}{2\epsilon} \geq \frac{s_0}{2\epsilon}, \forall\, n$, so we can choose ϵ in a way that $\forall\, n :\ p_n = \frac{s_n}{2\epsilon} > 1$. We can rewrite (12) in this way:

$$\int_0^1 (x^{(t_n-3/2)} - x^{t_n-1})e^{-x}\,\mathrm{d}x \leq \int_1^{p_n} (x^{t_n-1} - x^{(t_n-3/2)})e^{-x}\,\mathrm{d}x, \tag{13}$$

Now we observe that the right side of the inequality can be rewritten in this form:

$$\int_1^{p_n} x^{(t_n-3/2)}(x^{1/2} - 1)e^{-x}\,\mathrm{d}x$$

which is greater than

$$K_n = \int_1^{p_n} (x^{1/2} - 1)e^{-x}\,\mathrm{d}x > K > 0.$$

The left side of the inequality can be rewritten in this way:

$$\int_0^{1-\delta} x^{(t_n-3/2)}(1 - x^{1/2})e^{-x}\,\mathrm{d}x + \int_{1-\delta}^1 x^{(t_n-3/2)}(1 - x^{1/2})e^{-x}\,\mathrm{d}x. \tag{14}$$

The first integral of (14) is limited from above by

$$\int_0^{1-\delta} (1-\delta)^{(t_n-3/2)} e^{-x}\,\mathrm{d}x$$

which is lower than $\frac{K}{2}$ for t_n and then n great enough. The second one is limited from above by

$$\int_{1-\delta}^1 e^{-x}$$

which converges to 0 for $\delta \to 0$ and then is smaller than $\frac{K}{2}$ for δ small enough. But that means if we choose δ small enough and n great enough, the inequality (13) is true.

Now we are ready to prove the theorem:

Proof. It is possible to show (see (6)) that σ has a distribution with density

$$f_\sigma(\sigma) \propto \frac{1}{(\sigma)^{a_0+1}} e^{-\frac{1}{2}\frac{s_0}{\sigma^2}} \quad \text{for } \sigma^2 \geq \epsilon$$

and equal to 0 anywhere else. Using Bayes' formula we obtain that the distribution of σ given the observation $f(x_0)$ is proportional to

$$\frac{1}{\sigma^{a_0+2}} \exp\left(-\frac{1}{2}\frac{s_0+\left(\frac{f(x_0)-\mu(x_0)}{v(x_0;1)}\right)^2}{\sigma^2}\right), \tag{15}$$

for $\sigma^2 \geq \epsilon$. If we set

$$\begin{array}{l} s_1 = s_0 + \left[\frac{f(x_0)-\mu(x_0)}{v(x_0;1)}\right]^2 \\ a_1 = a_0 + 1 \end{array}$$

we can express (15) as

$$\frac{1}{\sigma^{a_1+1}} e^{-\frac{1}{2}\frac{s_1}{\sigma^2}} \text{ for } \sigma^2 \geq \epsilon,$$

from which it is possible to show that $\frac{s_1}{\sigma^2}$ has a distribution with density:

$$f_{\frac{s_1}{\sigma^2}}(x) = \begin{cases} \frac{1}{\beta(\frac{a_1}{2})}\left(\frac{1}{2}\right)^{a_1/2} x^{a_1/2-1} e^{-(1/2)x} & \text{for } 0 \leq x \leq \frac{s_1}{\epsilon} \\ 0 & \text{otherwise} \end{cases}$$

where $\beta\left(\frac{a_1}{2}\right)$ is the normalizing constant. In general at step n we have:

$$f_{\frac{s_n}{\sigma^2}}(x) = \begin{cases} \frac{1}{\beta(\frac{a_n}{2})}\left(\frac{1}{2}\right)^{a_n/2} x^{a_n/2-1} e^{-(1/2)x} & \text{for } 0 \leq x \leq \frac{s_n}{\epsilon} \\ 0 & \text{otherwise} \end{cases}$$

where $\beta\left(\frac{a_n}{2}\right)$ is again the normalizing constant and a_n, s_n are obtained from a_{n-1}, s_{n-1} through the updating formulae:

$$\begin{array}{l} s_n = s_{n-1} + \left[\frac{f(x_n)-\mu(x_n)}{v(x_n;1)}\right]^2 \\ a_n = a_{n-1} + 1 \end{array}$$

where x_n is the point in which the function is observed at step n. We know that the expected gain at step n is:

$$E^{\sigma|z_n}\left[\sigma v(x;1)\psi\left(\frac{f_n^* - \mu(x)}{\sigma v(x;1)}\right)\right].$$

The expected gain can be limited from above by:

$$\frac{v(x;1)}{\sqrt{2\pi}} E^{\sigma}[\sigma]$$

and, in a way similar to the proof of Observation 1, we have the limit from above:

$$\sqrt{s_n}\frac{v(x;1)}{2\sqrt{\pi}}\frac{\beta_n((a_n-1)/2)}{\beta_n(a_n/2)},$$

In view of lemma 4, we can consider the limit from above:

$$H\sqrt{s_n}\frac{v(x;1)}{2\sqrt{\pi}}$$

where H is a constant. In view of lemma 3, we have that s_n cannot increase faster than n.

By contradiction we assume that:

$$\lim_n \max_{i=1,\dots,n} (x_i - x_{i-1}) \not\longrightarrow 0.$$

Therefore there exists a subinterval of $[a, b]$ in which we never observe the function. If we consider the expected gain in the middle point of this interval, the fact that σ^2 is always greater than ϵ implies that it will always be greater than $h > 0$. But the limitation from above of the expected gain shows that, at step n, the expected gain cannot be greater than h if we are at a distance of the order of $\frac{1}{\sqrt{n}}$ from points where the function has already been observed. But if we multiply this distance by the number n of points, we see that for n large enough, the result is greater than the width of $[a, b]$, which means that in any point of $[a, b]$ the expected gain is smaller than h, which is a contradiction. The conclusion is that the set $\{x_i\}$ of the points at which the function is observed is dense and then:

$$\{f_n^*\} \to f^*.$$

C. Practical Tests

For the practical tests of the algorithm we refer to six functions, which can be found in [14, page 177]:

- $f_1(x) = -\sin x - \sin\frac{10x}{3} - \log x + 0.84x,\ 2.7 \leq x \leq 7.5;$
- $f_2(x) = -\sin x - \sin\frac{2x}{3},\ 3.1 \leq x \leq 20.4;$
- $f_3(x) = \sum_{i=1}^{5} i\sin((i+1)x + i),\ -10 \leq x \leq 10;$
- $f_4(x) = -(x + \sin x)e^{-x^2},\ -10 \leq x \leq 10;$

The last two functions $f_5(x)$ and $f_6(x)$ belong to the Shekel test functions, whose form is:

- $\sum_{i=1}^{10} \frac{1}{(k_i(x-a_i)^2)+c_i},\ 0 \leq x \leq 10,$

where $0 \leq a_i \leq 10$, $1 \leq k_i \leq 3$ and $0.1 \leq c_i \leq 0.3$. The values of these parameters for the two functions f_5 and f_6 are given in [14, pp. 178–179] together with the global maximum values and the global maximum coordinates for all the six functions. In testing the algorithm $\mathcal{A}(s_0)$ on these functions it has been noticed that while it can detect the global maximum region quite quickly, it has a slow rate of convergence. It has already been observed (see again [14] and references therein) that the reason for this slowness is that the statistical model of the function is unsatisfactory to describe the function locally and therefore it is better to distinguish two phases in the algorithm:

Table I. Number of iterations in order to accomplish the given accuracy in the global maximum values and coordinates

function	$\mathcal{A}(s_0)$			P^*
f_1	$s_0 = 50$	$s_0 = 500$	$s_0 = 5000$	
	19	25	32	33
f_2	$s_0 = 50$	$s_0 = 500$	$s_0 = 5000$	
	32	38	51	37
f_3	$s_0 = 5000$	$s_0 = 25000$	$s_0 = 50000$	
	27	48	42	see comment
f_4	$s_0 = 5$	$s_0 = 50$	$s_0 = 500$	
	23	32	46	35
f_5	$s_0 = 15000$	$s_0 = 75000$	$s_0 = 150000$	
	30	37	42	42
f_6	$s_0 = 25$	$s_0 = 2500$	$s_0 = 250000$	
	19	23	47	45

1. a global phase based on the statistical model;
2. a local phase where a local search (based only on function values) is started in the interval $[x_{i-1}, x_{i+1}]$ containing the record point x_i.

The stopping rule introduced in this paper can be exploited to stop the global phase instead of the whole algorithm, i.e. we switch from the global to the local phase at iteration n if:

$$\max_{x \in [a,b]} G_n(x; s_o) \leq c.$$

From Theorem 2 we know that the local phase starts after a finite number of iterations for any $c > 0$. Decreasing c has the effect of delaying the start of the local phase and the same effect can, in general, be obtained by increasing s_0, as Table I shows. In all the tests done we used $c = 0.015$. In Table I the behaviour of the algorithm for different values of s_0 is reported. For values of s_0 lower than the first one given in the table for any function, the algorithm was unable to detect the global maximum, because the local phase started too early. Any entry of the table contains the number of iterations necessary to accomplish the accuracy $\epsilon = 10^{-6}$ both in the global maximum value and in the global maximum coordinate. In [14, p. 180] the same results are given for six different algorithms. In the last column of table 1 we inserted the results for the P^*-algorithm which has been reported to be the best among the six algorithms, according to the criterion of the number of iterations in order to accomplish the desired accuracy. For the function f_3, [14] reports a number of iterations equal to 125, but it seems that this is the number of iterations necessary to detect all three of the global maximum points of this function, while we stopped after detecting with the given accuracy only one of them.

Table II. Distribution of the observations for different values of the parameters

	σ	s_0	σ	s_0	σ	s_0	σ	s_0
Interval	15	5000	36	60000	50	200000	100	1000000
[0.00,0.25]	1	1	3	2	3	3	4	5
(0.25,0.50]	1	1	1	1	2	2	3	3
(0.50,0.75]	1	0	2	2	2	3	4	4
(0.75,1.00]	0	1	2	2	2	2	4	4
(1.00,1.25]	0	0	1	1	2	2	3	2
(1.25,1.50]	0	0	4	3	4	3	4	4
(1.50,1.75]	1	0	2	2	2	2	3	2
(1.75,2.00]	0	0	5	6	6	8	6	7
(2.00,2.25]	1	1	2	2	2	1	2	1
(2.25,2.50]	0	0	12	12	10	10	6	7
(2.50,2.75]	0	1	1	1	3	2	3	3
(2.75,3.00]	30	32	8	9	7	7	5	6
(3.00,3.25]	28	26	8	9	6	7	6	5
(3.25,3.50]	1	1	1	1	3	2	3	3
(3.50,3.75]	0	0	11	11	9	9	7	7
(3.75,4.00]	1	1	2	1	2	2	2	2

D. The Choice of the Parameter

In this section we want to show through a practical example that $\mathcal{A}(s_0)$ is less sensitive to the choice of s_0 than $\mathcal{A}(\sigma)$ to the choice of σ. We consider the problem:

$$\max_{x \in [0,4]} (6x - x^2) \sin 11x$$

This function is multimodal with a global maximum in $x \approx 2.999$ with value ≈ 8.999. In order to study the behaviour of the two algorithms for different values of s_0 and σ, we consider how they distribute the observations over $[0, 4]$, by subdividing this interval in 16 subintervals of lenght 0.25 and giving the number of observations in any of them (the total number of function evaluations is in any case 65). The results are reported in Table II. We notice that the behaviour of $\mathcal{A}(\sigma)$ for certain choices of σ is absolutely similar to the behaviour of $\mathcal{A}(s_0)$ for certain choices of s_0. It is interesting to notice that a small choice of σ as well as a small choice of s_0 causes the algorithm to reduce to a local search, while increasing σ (and s_0) increases the global component of the algorithm, i.e. the search through the whole feasible set. Also we notice that the choice of an acceptable s_0 seems to be wider than the choice of σ. We can not compare the considered interval $[15, 100]$ for σ with the interval $[5000, 1000000]$ for s_0, but we can consider the choice of s_0 as the choice of an initial guess for σ. For instance a particular choice of s_0 corresponds to the initial guess of σ given by the median of the *a priori* distribution of σ, which is easily found if we remember that $\frac{s_0}{\sigma^2}$ has a χ^2- distribution with $a_0 = 2$ degrees of freedom. We have the following initial estimates:

- $s_0 = 5000 \ => \ \sigma \approx 60.0$
- $s_0 = 60000 \ => \ \sigma \approx 207.7$
- $s_0 = 200000 \ => \ \sigma \approx 379.3$
- $s_0 = 1000000 \ => \ \sigma \approx 848.1$

These data show two interesting things. The first one is that the initial estimates are greater than the corresponding values of σ for which the behaviour of the two algorithms are similar. This fact can be explained by noticing that we are estimating a value which is actually equal to 0 if the sample path represented by the objective function is regular enough, as it is the case in our example. Generally, after some iterations in $\mathcal{A}(s_0)$, the initial estimate has been reduced and the algorithm become similar to $\mathcal{A}(\sigma)$ with a value of σ lower than the initial estimate. The second thing is that the choice of the initial estimate (and then of s_0) is much wider than the choice of σ in $\mathcal{A}(\sigma)$. Then, while still dependent on the choice of a parameter, $\mathcal{A}(s_0)$ seems to be less sensitive to this choice.

References

1. B.Betró, F.Schoen (1992), Optimal and sub-optimal stopping rules for the Multistart algorithm in global optimization, *Mathematical Programming* **57**, 445–458.
2. J.M.Calvin (1991), Consistency of a myopic bayesian algorithm for global optimization, *Technical Report,* Georgia Institute of Technology.
3. M.H.DeGroot (1970), *Optimal Statistical Decisions,* McGraw-Hill, New York.
4. A.O'Hagan (1978), Curve fitting and optimal design for prediction, *Journal of Royal Statistical Society B* **40**, 1–42.
5. A.O'Hagan (1991), Some bayesian numerical analysis, *Technical Report,* University of Nottingham.
6. H.Kushner (1962), A versatile stochastic model of a function of unknown and time varying form, *Journal of Math.Anal.Appl.* **5**, 150–167.
7. M.Locatelli (1992), *Algoritmi bayesiani per l'ottimizzazione globale*, unpublished laurea thesis.
8. M.Locatelli, F.Schoen (1993), An adaptive stochastic global optimization algorithm for one-dimensional functions, *Proceedings of APMOD93,* Budapest
9. J.Mockus (1988), *Bayesian Approach to Global Optimization,* Kluwer Academic Publishers, Dordrecht.
10. A.M.Mood, F.A.Graybill, D.C.Boes (1974), *Introduction to the Theory of Statistics,* McGraw-Hill, New York.
11. D.Revuz, M.Yor (1991), *Continuous Martingales and Brownian Motion,* Springer Verlag, Berlin.
12. E.Senkiene, A.Zilinskas (1978), On estimation of a parameter of Wiener process, *Lithuanian Mathematical Journal* **3**, 59–62.
13. R.G.Strongin (1992), Algorithms for multi-extremal mathematical programming problems employing the set of joint space-filling curves, *Journal of Global Optimization* **2**, 357–378.
14. A.Törn, A.Zilinskas (1987), *Global Optimization,* Springer Verlag, Berlin.
15. A.A.Zhigljavsky (1991), *Theory of Global Random Search,* Kluwer Academic Publishers, Dordrecht.
16. A.Zilinskas (1975), One-step Bayesian method of the search for extremum of an one-dimensional function, *Cybernetics* **1**, 139–144.