Bayesian Methods in Global Optimization

BRUNO BETRÒ
CNR-IAMI, Via Ampère 56, I-20131 Milano, Italy

(Received: 27 February 1990; accepted: 14 September 1990)

Abstract. This paper reviews methods which have been proposed for solving global optimization problems in the framework of the Bayesian paradigm.

Keywords. Bayesian inference, stochastic processes, decision theory, stopping rules, multistart method.

1. Introduction

The global optimization problem in a numerically well-defined sense can be formulated as

$$\text{find } \hat{x} \text{ s.t. } \hat{f} = f(\hat{x}) \text{ is } \textit{close} \text{ to } f^* = \max_{x \in K} f(x) \tag{1}$$

where K is a compact set in R^N and f is a continuous function over K. Obviously, solving (1) requires the previous specification of some closeness criterion in order to evaluate the goodness of any feasible point $\hat{x}$. It is a trivial observation that if, for given positive ε, an $\hat{x}$ is sought such that

$$f^* - \hat{f} < \varepsilon\,, \tag{2}$$

then, under solely qualitative assumptions on f like continuity or differentiability up to a certain order, no algorithm can be given which terminates after a finite number of steps with a point $\hat{x}$ guaranteed to satisfy (2).

The situation can be improved if quantitative assumptions on the objective function are introduced, e.g. the validity of a Lipschitz condition with known Lipschitz constant. In such a case results can be proven ([2]) asserting that an algorithm exists leading to $\hat{f}$ as accurate as required by (2); however, the number of evaluations of f needed may be as large as

$$(L/2\varepsilon)^N\,,$$

where L is the Lipschitz constant, and this is typically a large number. The reason for that is intrinsic to the idea of guaranteeing the achievement of a prefixed accuracy, and hence of taking into account the possibility of occurrence of the worst possible case. Therefore the idea naturally arises to relax prevention against the worst possible case, usually quite pathological and never occurring in the practice, and to take care only of cases that have some nonnegligible chance to occur.

Journal of Global Optimization **1**: 1–14, 1991

A satisfactory treatment of changes of occurrence can be achieved only if the global optimization problem (1) is reformulated superimposing to the class of problems to be considered a probabilistic structure and an accuracy criterion is set up consequently. The Bayes theorem then provides the basic tool for adapting the superimposed probabilistic structure to information gained about the problem through function evaluations.

Methods derived according to this framework will be referred to as Bayesian methods. The aim of this paper is to review the different probabilistic formulations of the global optimization problem and the related Bayesian methods as yet proposed.

2. The Random Function Approach

The approach dates back to the sixties (see [22] and references in [23]) and it is based on the idea of introducing a probabilistic model for the objective function f in the form of a random function $f(x, \omega)$, where ω belongs to some measurable space Ω over which a probability measure P is defined and, for fixed x, $f(x, \omega)$ is a random variable, i.e. a measurable function of ω. The actual function to be optimized is seen as a realization $f(x)$ of $f(x, \omega)$. It is useful to assume that, for almost every ω, $f(x, \omega)$ is a continuous function of x, so that a.s.

$$\max_{x \in K} f(x, \omega)$$

exists under the hypothesis of compactness of K. For sake of brevity, the argument ω will be frequently omitted in the sequel.

Let S_n be a sequential n-step optimization strategy, that is a mapping of the function space considered over the set of n-tuples of points in K such that $S_n(f)$ produces the points

$$\begin{aligned} &x_1 \\ &x_2 = x_2(x_1, f(x_1)) \\ &\vdots \\ &x_n = x_n(x_1, f(x_1), \ldots, x_{n-1}, f(x_{n-1})) \end{aligned} \tag{3}$$

The effectiveness of S_n for the function f can be defined by the difference

$$L(S_n, f) = \max_x f(x) - f_n^*$$

where x_n is given by (3) and $f_n^* = \max_{1 \leq i \leq n} f(x_i)$, so that over the class $\{f(x, \omega)\}$ S_n will display the average effectiveness

$$\begin{aligned} E(L(S_n, f)) &= \int L(S_n, f) P(\mathrm{d}\omega) \\ &= E(\max_x f(x)) - E(f_n^*) . \end{aligned} \tag{4}$$

Then an optimal strategy S_n^* can be defined as any strategy minimizing (4), i.e.

$$E(L(S_n^*, f)) = \inf_{S_n} E(L(S_n, f))$$

or, equivalently, such that

$$E(f_n^{**}) = \sup_{S_n} E(f_n^*)\,,$$

where f_n^{**} is the largest function value attained by S_n^*.

The determination of S_n^* can take profit of a standard dynamic programming approach; introducing the vectors z_i of information available about f at step i, $z_i = (x_1\ f(x_1), \ldots, x_i, f(x_i))$, the Bellman equations are to be solved

$$\begin{aligned} u_n(z_{n-1}) &= \max_{x \in K} E(\max\{f(x), f_{n-1}^*\} \mid z_{n-1}) \\ u_i(z_{i-1}) &= \max_{x \in K} E(u_{i+1}(z_{i-1}, x, f(x)) \mid z_{i-1}) \qquad i = n-1, \ldots, 2 \\ u_1 &= \max_{x \in K} E(u_2(x, f(x)))\,. \end{aligned} \tag{5}$$

Each equation defines the next point of the strategy S_n^* as the point where the max on the right hand side is attained.

It is well known that serious problems arise from a computational point of view in dealing with Bellman equations even for moderate n; it is therefore usual practice to derive suboptimal strategies considering only a few or just use one of equations (5). For instance (*a*)

$$x_{i+1} = \arg\max_{x \in K} E(\max\{f(x), f_{n-1}^*\} \mid z_i)\,; \tag{6}$$

this strategy can be said one-step optimal as, once i points have been obtained, the $(i+1)-st$ is determined in an optimal way. Differently from (5), the influence of this choice on future choices is not taken into account.

(*b*) if, given z_i, the next points x_{i+1} and x_{i+2} are obtained according to the equations

$$\begin{aligned} u_{i+2}(z_i + 1) &= \max_{x \in K} E(\max\{f(x), f_{n-1}^*\} \mid z_{i+1}) \\ u_{i+1}(z_i) &= \max_{x \in K} E(u_{i+2}(z_i, x, f(x)) \mid z_i) \end{aligned} \tag{7}$$

then the resulting strategy can be named two-step optimal.

Clearly (6) and (7) can be applied sequentially without prefixing in advance the number of points at which one is willing to evaluate the function f. This, however, requires that some suitable stopping criterion be previously set up. Both for this aspect and with respect to implementation of (6) and (7), the situation is rather different if the problem is one-dimensional or multi-dimensional. In the former case, in fact, the Wiener process provides a manageable stochastic model of continuous multimodal functions, by which the distribution of $f(x)$ given z_i is easily obtained for each i and for each x; indeed this distribution is normal with mean

$$\mu(x \mid x_1, f_1, \ldots, x_n, f_n) = f_{i-1} \frac{x_i - x}{x_i - x_{i-1}} + f_i \frac{x - x_{i-1}}{x_i - x_{i-1}},$$
$$x \in [x_{i-1}, x_i], i = 2, \ldots, n;$$
$$= f_n, x \geq x_n, \tag{8}$$

where $f_k = f(x_k)$, $i = 1, \ldots, n$, and variance

$$\sigma^2(x \mid x_1, f_1, \ldots, x_n, f_n) = \sigma^2 \frac{(x - x_{i-1})(x_i - x)}{x_i - x_{i-1}};$$
$$x \in [x_{i-1}, x_i];$$
$$= \sigma^2(x - x_n), x > x_n. \tag{9}$$

In (8) and (9) it is assumed that x_1 is the origin of the Wiener process and that the x_i's are increasingly ordered. σ^2 is the parameter of the Wiener process which must be specified in advance. Under the Wiener model, the optimization problem on the right hand side of (6) is a multiextremal one, but the function to be maximized can be recognized as unimodal in each interval $[x_{i-1}, x_i]$.

A relevant consequence of the Wiener model for f is that the conditional distribution of $\max_x f(x)$ can be expressed by the simple formula ([19])

$$P\{\max_{x_1 \leq x \leq x_n} f(x) \leq \bar{f} \mid x_1, f_1, \ldots, x_n, f_n\}$$
$$= \prod_{i=2}^{n} P\{\max_{x_{i-1} \leq x < x_i} f(x) \leq \bar{f} \mid x_{i-1}, f_{i-1}, \ldots, x_i, f_i\},$$

where

$$P\{\max_{x_1 \leq x \leq x_n} f(x) > \bar{f} \mid x_1, f_1, \ldots, x_n, f_n\}$$
$$= \begin{cases} 1 & \bar{f} \leq \max(f_{i-1}, f_i) \\ \exp\left(-2 \dfrac{(\bar{f} - f_{i-1})(\bar{f} - f_i)}{\sigma^2(x_i - x_{i-1})}\right) & \bar{f} > \max(f_{i-1}, f_i) \end{cases}. \tag{10}$$

This result enables to deal probabilistically with the error $\max f(x) - \max_i f_i$. Algorithms based on the Wiener model and the error control provided by (10) have been proposed in [22], [32], [30], [3]; their performance has been shown to be fairly good on several test instances.

In the multidimensional case, the easiest way to model an unknown continuous function is to consider Gaussian random functions. According to this model, for each n-tuple of points $x_1, \ldots, x_n$, the joint density of the function values $f(x_1), \ldots, f(x_n)$ is multivariate normal (see, e.g., [1]) with mean and covariance matrix specified once the mean and covariance functions are given

$$\mu(x) = E\{f(x)\}$$
$$R(x, y) = E\{[f(x) - \mu(x)][f(y) - \mu(y)]\}.$$

Gaussian random functions have the attractive property that the posterior distribution of $f(x)$ given $f(x_1)$ given $f(x_1), \ldots, f(x_n)$ is still normal, with mean

$$\mu_n(x) = \mu(x) + \Sigma_n^T \Sigma_{nn}^{-1}(F_n - \mu_n) \tag{11}$$

and variance

$$\sigma_n^2(x) = \sigma^2(x) - \Sigma_n^T \Sigma_{nn}^{-1} \Sigma_n \tag{12}$$

where

$$\begin{aligned} \Sigma_n^T &= (R(x, x_1), \ldots, R(x, x_n)) \\ F_n &= ((x_1), \ldots, f(x_n)) \\ \Sigma_{nn} &\quad \text{covariance matrix of } F_n \\ \mu_n &= (\mu(x_1), \ldots, \mu(x_n)) \\ \sigma^2(x) &= R(x, x) . \end{aligned}$$

It is therefore in principle straightforward to compute the expectations required for example by (7), as they simply require to integrate with respect to the normal distribution. But a serious difficulty arises when the next point has to be selected: indeed, the maximization problem involved, which turns out to be of the form $\max_{x \in K} \phi(x, z_i)$, is just a global one and, unlike in the one-dimensional case, there is no easy way to decompose it into a number of unimodal subproblems. It has been argued however ([25]) that exact maximization is not required, so that it may be sufficient for instance to evaluate ϕ at a number m (say $m = 100$) of random points and to select x_{i+1} as the one where the best value of ϕ has been observed. Observe that it is the function ϕ to be evaluated in this procedure and not the objective function f. This implies that, if the evaluation cost of f is very high, the overhead cost can be relatively small, at least for moderate i. Conversely, looking at (11), (12), the storage and computational cost required by ϕ may become prohibitively high for large i.

Several attempts have been made to reduce such overheads. Modelling f by a Gaussian random function, the complexity of (11) and (12) is a consequence of the model updating formulas given by Bayes theorem. Simplification is possible only if one drops the consistency of the model after n observations with the model after $n-1$ observations. In other words, the idea is to consider, for each n, a Gaussian random function such that $\mu_n(x)$ and $\sigma_n^2(x)$ are given through simple expressions in x_i, $f(x_i)$, $i = 1, \ldots, n$, so that they will no longer be in general conditional means and conditional variances. Methods of this type have been introduced in [24] and [29], [31], [33], [34]. It is worth mentioning that the approach of Žilinskas is based on a number of axioms aimed at formalizing information available to a "rational" optimizer about the function behavior. It has to be finally remarked that no exact result about the distribution of the global maximum of a Gaussian random function is available in the multidimensional case. This means that, unless the number of function evaluations is fixed in

advance, the question of how to evaluate the error of an approximation $\hat{f}$ to the global optimum cannot be answered accurately.

In the same framework of a stochastic model representing the function to be optimized is the information approach, recently reviewed in ([27]). In this approach, in the case of one-dimensional optimization problems with a single global extremum x^*, a prior distribution is considered on the location of x^*, a Gaussian distribution is assumed for the increments of the objective function depending on the location of x^*, and the posterior density of x^* is obtained after a number of function evaluations. Then an estimate of x^* is obtained maximizing the posterior density, and this estimate is assumed as the next search point. Under suitable conditions the resulting algorithm has the property that the set of its accumulation points coincides with the set of local maxima. For the solution of multidimensional problems, it is proposed to transform the problem in a one-dimensional one by means of Peano maps.

3. A Probabilistic Model for the Structure of Global Optimization Problems and the Multistart Method

It is an obvious observation that a global optimization problem would be solved once all local minima were discovered. It is therefore common practice to try to reach all local maxima starting a local search from n points randomly (uniformly) drawn in K. The procedure is usually referred to as *multistart* method. If the number of local maxima is finite and apart from pathological situations, the procedure will achieve its goal with positive probability p; furthermore, p increases as the number of trials increases, so that n is usually taken large. It is conceivable that if it were possible to obtain information about p during the search procedure, then some stopping criterion could be set up avoiding to spend local searches superfluously.

Following [28] it is useful to state the problem as follows. Assume that in K there is a finite number of local maxima, say $x_1^*, \ldots, x_k^*$. Let A be a search algorithm which, starting from a point x in K leads to some point $A(x)$ in K. Define

$$X_i^* \equiv \{x \in K : A(x) = x_i^*\}, \; i = 1, \ldots, k$$

and assume that the sets X_i^* are Lebesgue measurable. Denoting by m the Lebesgue measure, assume that $m(K) = m(\cup_{i=1}^k X_i^*)$, that is the set of starting points causing A not to converge to a local maximum has null measure. The set X_i^* is called *region of attraction* of x_i^* and the quantity $\theta_i = m(X_i^*)/m(K)$ is the *share* of x_i^*.

If $X_1, \ldots, X_n$ are *i.i.d.* random variables uniformly distributed in K, then $P\{X_j \in X_i^*\} = \theta_i$, $j = 1, \ldots, n$, $i = 1, \ldots, k$. The application of A to the X_j's will lead to find N_i times the maximum x_i^*, with $N_i \geq 0$ and $\Sigma_{i=1}^k N_i = n$. Given the number of maxima k and the shares $\theta_1, \ldots, \theta_k$, then the vector $\langle N_1, \ldots, N_k \rangle$

has the multinomial distribution

$$P\{N_1 = n_1, \ldots, N_k = n_k\} = \frac{n!}{n_1! \cdots n_k!}\, \theta_1^{n_1} \cdots \theta_k^{n_k}. \tag{13}$$

As k and $\theta_1, \ldots, \theta_k$ are unknown, it is sensible to infer about them on the basis of the available observations. By (13), this problem can be seen as a problem of inference about a multinomial distribution with an unknown number of classes.

Let W be the random variable representing the number of different local maxima found after n local searches, and let the random vector $\langle N_1, \ldots, N_W \rangle$ be such that N_j represents the number of times the j-th maximum among the W has been found. Let values w, $\langle n_1, \ldots, n_w \rangle$ be observed, h_j be the number of n_i's equal to j, and $S_k[w]$ denote the set of all permutations of w different elements from $\{1, \ldots, k\}$; then the likelihood is given by

$$P\{W = w, N_1 = n_1, \ldots, N_w = n_w \mid k, \theta_1, \ldots, \theta_k\}$$
$$= \frac{n!}{\Pi_{j=1}^{n} h_j! \Pi_{i=1}^{w} n_i!} \sum_{\langle g_1, \ldots, g_w \rangle \in S_k[w]} \prod_{i=1}^{w} \theta_{g_i}^{n_i}. \tag{14}$$

Once a prior distribution on the parameter space $\Theta = \{\langle \theta_1, \ldots, \theta_k \rangle,\ k = 1, 2, \ldots, \Sigma_{i=1}^{k} \theta_i = 1,\ \theta_i \geq 0\}$ is specified, the Bayes Theorem enables to obtain the posterior distribution of the unknown parameters k and $\theta_1, \ldots, \theta_k$. If the prior distribution on Θ is given in the form

$$\mu = \sum_{j=1}^{\infty} p_j \mu_j, \tag{15}$$

where p_j is the prior probability that the true number of local maxima equals j and μ_j is the conditional *a priori* distribution of $\langle \theta_1, \ldots, \theta_j \rangle$ given that j is the true number of local maxima, then by (14) for $k \geq w$

$$P\{K = k, \langle \theta_1, \ldots, \theta_k \rangle \in \langle d\theta_1, \ldots, d\theta_k \rangle \mid W = w, N_1 = n_1, \ldots, N_W = n_w\} \tag{16}$$

$$= \frac{p_k \Sigma_{\{g_1, \ldots, g_w\} \in S_k[w]} \Pi_{i=1}^{w} \theta_{g_i}^{n_i} \mu(d\theta_1, \ldots, d\theta_k)}{\Sigma_{j=w}^{\infty} p_j \int \Sigma_{\{g_1, \ldots, g_w\} \in S_j[w]} \Pi_{i=1}^{w} \theta_{g_i}^{n_i} \mu(d\theta_1, \ldots, d\theta_j)}. \tag{17}$$

Having found w maxima, it is of interest to determine whether $w = k$ (all maxima have been located) or $w \neq k$. Assume that there is a *loss* c for having decided $w = k$ when $w < k$, and a loss C for having decided $w < k$ when $w = k$; moreover, assume that right decisions have no losses. Then standard arguments show that it is optimal to decide $w = k$ when

$$cP\{K > w \mid W = w, N_1 = n_1, \ldots, N_w = n_w\}$$
$$< CP\{K = w \mid W = w, N_1 = n_1, \ldots, N_w = n_w\}$$

or, making use of (17), when

$$c \sum_{k=w+1}^{\infty} p_k \sum_{\{g_1, \ldots, g_w\} \in S_k[w]} \prod_{i=1}^{w} \theta_{g_i}^{n_i} \mu(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_k)$$

$$< C p_w \int \prod_{i=1}^{w} \theta_i^{n_i} \mu(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_w), \tag{18}$$

and to decide $k > w$ when (18) is not satisfied.

A convenient choice for $\mu(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_k)$ is the *symmetric Dirichlet* distribution ([21]), i.e. a Dirichlet distribution with parameters all equal to some $\alpha > 0$. In particular, $\alpha = 1$ gives the uniform distribution on the k-dimensional simplex $\{\theta_i \geq 0,\ i = 1, \ldots, k\ \Sigma_{i=1}^{k}\ \theta_i = 1\}$. Under a symmetric Dirichlet distribution, in (18)

$$\text{II} \int \prod_{i=1}^{w} \theta_{g_i}^{n_i} \mu(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_k) = E\left\{\prod_{i=1}^{w} \theta_i^{n_i}\right\}$$

$$= \frac{\Gamma(k\alpha) \prod_{i=1}^{w} \Gamma(\alpha + n_i)}{(\Gamma(\alpha))^w \Gamma(k\alpha + n)}, \tag{19}$$

so that after some simplifications (18) becomes

$$c \sum_{k=w+1}^{\infty} p_k \binom{k}{w} \frac{\Gamma(k\alpha)}{\Gamma(k\alpha + n)} < C p_w \frac{\Gamma(w\alpha)}{\Gamma(w\alpha + n)}. \tag{20}$$

The optimal decision about the number of local maxima can now be obtained after the specification of the prior $\{p_k\}$. The numerical behavior of rule (20) has been thoroughly investigated by Monte Carlo simulation in [9] in the case of a truncated Poisson distribution. In the simulated situations the rule has been found accurate and robust at a satisfactory level.

If $\{p_k\}$ is taken as the improper distribution $\{p_k = \text{const}\}$, which corresponds to the idea of any number of maxima being equally likely, then, for $n \geq w + 2$ it is optimal to decide $k = w$ if the condition

$$\frac{\Gamma(n + w)\Gamma(n - w - 1)}{\Gamma(n)\Gamma(n - 1)} \leq 1 + \frac{C}{c}$$

is satisfied.

It is interesting to observe that, because of the improperness of the distribution $\{p_k\}$, an optimal rule does not exist when $n < w + 2$.

In case when $k > w$ has been decided, it may be of interest to evaluate the total share of undiscovered regions. By (17) and (19) it is easy to obtain after some manipulations that, if $\gamma = 1 - \Sigma_{i=1}^{w}\ \theta_i$, then the optimal estimate according to a quadratic loss function, is

$$\hat{\gamma} = E(\gamma \mid w, n_1, \ldots, n_w)$$

$$= \frac{\Sigma_{k=w}^{\infty}\ \alpha \frac{k - w}{n + k\alpha}\ p_k \binom{k}{w} \frac{\Gamma(k\alpha)}{\Gamma(n + k\alpha)}}{\Sigma_{k=w}^{\infty}\ p_k \binom{k}{w} \frac{\Gamma(k\alpha)}{\Gamma(n + k\alpha)}}. \tag{21}$$

Note that both in (20) and in (21) the only observed quantity involved is the number of local maxima found.

The above decision setting assumes that n is fixed. Assume now that the number of local searches is not fixed in advance, but they are started sequentially. Then after each of them is completed it is possible to decide whether to continue or not according to some criterion comparing the results obtained so far with the benefit expected from starting new searches. To be more precise, suppose that each local search has a fixed cost C, and that there is a loss A if not all the local maxima have been discovered, so that after n local searches the total loss is

$$L(w_n, K) = \begin{cases} Cn & \text{if } w_n = K \\ A + Cn & \text{if } w_n < K \end{cases} \tag{22}$$

where w_n is the number of discovered local maxima during the first n local searches. The expected posterior loss is then

$$\begin{aligned} L(w) &= E\{L(W_n, K) \mid W_n = w\} \\ &= AP\{K > w \mid W_n = w\} + cn = Aq_n(w) + cn , \end{aligned}$$

where, by (17) and (19),

$$q_n(w) = 1 - \frac{p_w \Gamma(w\alpha)/\Gamma(w\alpha + n)}{\Sigma_{k=w+1}^{\infty} p_k \binom{k}{w} \Gamma(k\alpha)/\Gamma(k\alpha + n)} .$$

The problem is now to find an optimal stopping rule N^* which minimizes $E(Y_N)$ over the class of stopping rules N, where, for $c = C/A$,

$$Y_n = q_n + cn .$$

As optimal stopping requires to look ahead for future observations given the present ones, the predictive probability

$$P\{W_{n+1} = w + 1 \mid W_n = w\}$$

has to computed. Note that, once $W_n = w$, the only possible values for W_{n+1} are w and $w + 1$.

A moment consideration shows that the above probability is just the probability of starting the next search within the region of attraction of a maximum not yet discovered, that is the expected total share of undiscovered maxima, so that, by (21),

$$P\{W_{n+1} = w + 1 \mid W_n = w\} = E(\gamma \mid W_n = w) = \hat{\gamma}(w) . \tag{23}$$

Formula (23) can now be used to show that the sequence Y_n is a submartingale ([12]) and hence an $\bar{n}$ exists such that $P\{N^* \leq \bar{n}\} = 1$. The actual construction of N^* can be achieved by backward induction (see, e.g. [13]). It is possible to derive analogous results for losses different from (22), see [10] and [11]. It has to be observed that all the proposed loss functions (including the one in the following Section) provide rather crude, although computationally convenient, approximations to "real" losses suffered by actual users; in particular the assumption that each local search has a fixed computational cost is rather naive. This is likely to be cause of some difficulties in assessing the loss function parameters. However, this is an usual situation in applications of Statistical Decision Theory (see, e.g., the discussion in [4]).

A drawback of the approach is that the function values at the maxima are not taken into account. It has been recently observed ([26]) that in the case that different maxima have different values, it is possible to incorporate ordering information about the maxima induced by their function values. Then stopping rules can be provided under losses based on the number and the shares of maxima not yet found but "better", with respect to function value, than the best maximum found so far. It is reasonable to expect that such rules have better performance in actual situations than the ones not based on the ordering. Supporting evidence is presented in [26] for standard test problems from [14].

4. A Bayesian Model for the Distribution of the Sampled Maxima

The approach presented in the previous section, even with the improvement proposed in [26], is unable to deal explicitly with function values, neglecting this way important information about the structure of the problem. A way of circumventing this situation has been proposed in [7], following ideas introduced in [5].

Let $t_i = f(x_i^*)$, $i = 1, \ldots, n$, be the optimum values sampled by the multistart methods after n steps. Assume that each local search has a fixed cost $c > 0$, expressed in the same units as f, and that the cost connected with stopping at step n is

$$L(t_1, \ldots, t_n; c) = -t_{(n)} + nc \tag{24}$$

where $t_{(n)} = \max_{i=1,\ldots,n} t_i$. The cost (24) combines the cost of a new local search with the gain corresponding to a unit increase in the maximum observed value.

The observations t_i are independent realizations of a random variable T whose distribution F is unknown (except for trivial cases). Observe that, in case the distribution were known, the problem of optimally stopping the multistart method would be solved. Indeed, under (24), the problem is the one of stopping sequential sampling from a distribution F which is referred to in the literature as *optimal sampling with recall* ([13]): when F is known, the optimal stopping rule is to stop the sampling process as soon as

$$\int_{t_{(n)}}^{\infty} (t - t_{(n)})\, \mathrm{d}F(t) \leq c\,, \tag{25}$$

or equivalently,

$$\int_{t_{(n)}}^{\infty} (1 - F(t))\, \mathrm{d}t \leq c\,. \tag{26}$$

As F is unknown, it is necessary to adopt some model of it. This has the consequence that it is no longer possible to determine the optimal stopping rule, but the model hereafter described enables to derive easily suboptimal rules which perform satisfactorily in practice.

Following the ideas of Bayesian nonparametric analysis ([16], [17], [15], [18]) a class of random distribution functions in the family of *neutral to the right* can be

introduced as a model for the unknown distribution function. A class in this family is described by the process

$$F(t) = 1 - \exp(-Y(t)) , \tag{27}$$

where $Y(t)$ is a stochastic process a.s. nondecreasing and right continuous with $\lim_{t\to-\infty} Y(t) = 0$ and $\lim_{t\to\infty} Y(t) = \infty$. In [7], the *simple homogeneous* process introduced in [18] was considered for its manageability in presence of coinciding observations, as it is the case for the optimum values generated by the multistart method. This process yields through (27) a probability measure which satisfies the basic requirements for being a suitable *a priori* probability measure over the class of distribution function ([16]): (*a*) its support is wide enough to contain all distributions of practical interest: (*b*) the posterior probability given a sample from it is computationally tractable.

It can be proved that, under a simple homogeneous process $Y(t)$, given a sample $t_1, \ldots, t_n$ from $F(t)$ as in (27), the posterior Bayesian estimate of $F(t)$ is given by

$$\begin{aligned} 1 - \hat{F}_n(t) &= E(1 - F(t) \mid t_1, \ldots, t_n) \\ &= \frac{m(t) + \lambda}{n + \lambda} \exp\left\{ - \sum_{j=1}^{n(t)} (\gamma(t_{(j)}) - \gamma(t_{(j-1)}))/(m_{j-1} + \lambda) \right\} \\ &\quad \exp\{-(\gamma(t) - \gamma(t_{(n(t))}))/(m(t) + \lambda)\} , \end{aligned} \tag{28}$$

where γ is a continuous nondecreasing function with $\gamma(-\infty) = 0$ and $\gamma(\infty) = \infty$; λ is a positive parameter; $t_{(1)}, \ldots, t_{(n_0)}$ are the increasingly ordered distinct observations in the sample;

$$\begin{aligned} m_j &= \#\{\text{observations} > t_{(j)}\} ; \\ n(t) &= \#\{\text{distinct observations} \leq t\} ; \\ m(t) &= \#\{\text{observations} > t\} ; \\ t_{(0)} &= -\infty . \end{aligned}$$

As it seems to be out of question to find an optimal stopping rule under the loss (24) and the simple homogeneous process, it is natural to consider suboptimal rules. *k*-step look ahead (*k-sla*) rules ([4]) frequently represent useful approximations to optimal stopping rules; a *k-sla* rule calls for stopping the sampling process as soon as the current cost is not greater than the cost expected if at most k further observations are taken. Under a stochastic model for F, either parametric or nonparametric, the 1-*sla* is found to be the rule which prescribes stopping as soon as

$$\int_{t_{(n)}}^{\infty} (t - t_{(n)}) \, \mathrm{d}\hat{F}_n(t) \leq c , \tag{29}$$

where $\hat{F}_n(t) = E(F(t) \mid t_1, \ldots, t_n)$, or equivalently

$$\int_{t_{(n)}}^{\infty} (1 - \hat{F}_n(t)) \, \mathrm{d}t \leq c . \tag{30}$$

Stopping is prescribed by the 2-*sla* rule when

$$\int_{t_{(n)}}^{\infty} (1 - \hat{F}_n(t))\, dt$$

$$- E^{T_{n+1}|n}\left(\min\left\{0, c - \int_{\max(t_{(n)},\, T_{n+1})} (1 - \hat{F}_{n+1}(t))\, dt\right\}\right) \leq c\,,$$

where $E^{T_{n+1}|n}$ stands for expectation with respect to the conditional distribution of T_{n+1} given the first n observation.

As $\hat{F}_n(t)$ is the predictive distribution of the next observation given the first n observations, (30) says that stopping occurs according to the 1-*sla* as soon as the expected improvement in the best sampled value determined by a further observation is not larger than c. Observe that (29) has the same form as (25) and (30) has the same form of (26), with $\hat{F}_n$ in place of F, which shows that the optimal rule when F is known is actually a 1-*sla*.

Looking at (28) and (30), it is easily seen that the 1-*sla* is easily implemented if $\int_a^{\infty} e^{-\gamma(t)/\lambda}$ is easily obtainable for any a. As γ and λ are parameters of the model, they can be chosen in accordance with this condition. It can be seen that, if $F_0(t) = E(F(t))$ is a prior *guess* for $F(t)$, then F_0, λ and γ are linked by the equation

$$\gamma(t) = -\lambda \log(1 - F_0(t))\,,$$

which is helpful for assessing the prior parameters.

Implementation of the 2-*sla* is more cumbersome, but it can be worked out numerically (see [7] for the details). It should be noted that, due to the fact that whenever a *k*-*sla* says to continue it is optimal to continue, then the 2-*sla* needs to be invoked only when the 1-*sla* says to stop.

1- and 2-*sla* have been tested for the standard test problems of [14] ([7]), for randomly generated problems in up to six dimensions and with various degrees of complexity ([6]), and for randomly generated distribution functions $F(t)$ ([8]). The results show that the percentage of failures in finding the global maximum is low, with a moderate number of local searches performed before stopping, both for the 1-*sla* and for the 2-*sla*: actually the behavior of the two rules is nearly indistinguishable. This fact leads to the conclusion that the 1-*sla* rule is, because of its simplicity and effectiveness, an attractive stopping rule for the multistart method. It is to be observed that the fact that the two rules are close does not imply that they are close to the optimal one: however the analysis performed in [8] shows that, under reasonable tuning of the prior parameters, the two approximate rules can be made close to the optimal rule built under perfect knowledge of the distribution of the sampled optimum values, that is of the very structure of the problem.

5. Conclusions

Three Bayesian approaches to global optimization have been reviewed. They

share the idea of dealing with uncertainty about the problem according to the Bayesian paradigm; each of them is however characterized by a different probabilistic formulation of the problem structure. It is worth to recall that the random function approach considers a stochastic model of the whole function to be optimized, and it is aimed at developing optimization algorithms which are optimal in some sense; in the other two approaches the modelization is restricted to features of the objective function which are relevant for the performance of the multistart algorithm, with the scope of deriving effective rules for the statistical control of the algorithm itself. Thus the random function approach is the only applicable one in situations in which the possibility of effectively performing local searches is ruled out, by very costly function evaluations or by low function regularity. When the multistart algorithm can be sensibly adopted, the approach outlined in Section 4 offers, unlike the one of Section 3, the attractive possibility of taking into account explicitly information provided by function values at the sampled extrema. However, the possibility of providing a comparative numerical evaluation of the three approaches is related to the general question of the evaluation of performances of global optimization algorithms. A satisfactory answer to this question would require the definition of a standard evaluation environment (set of test functions, local search routine, algorithm parameters, . . .) and of standard performance criteria. Unfortunately, after the early attempt in [14], no further effort has been devoted to such a definition. It is to be remarked in particular that the set of test functions of [14] is very limited in size and in complexity (few local extrema, up to six variables), so that the need is felt for the introduction of a wider and more significant set of test functions. With respect to this, the author recalls his proposal (see [5]) of exploiting the concept of *generalized metric interpolation* ([20]) for generating global optimization test functions (see also [6]), having full control of features like number of variables, number, values and locations of the extrema, degree of regularity.

References

1. Adler, R. J. (1981), *The Geometry of Random Fields*, Wiley, New York.
2. Archetti, F. and B. Betrò (1978), A Priori Analysis of Deterministic Strategies for Global Optimization, in Dixon L. C. W. and G. P. Szegö (eds.), *Towards Global Optimization II*, North Holland, Amsterdam.
3. Archetti, F. and B. Betrò (1979), A Probabilistic Algorithm for Global Optimization, *Calcolo* **16**, 335–343.
4. Berger, J. O. (1980), *Statistical Decision Theory*, Springer, Berlin.
5. Betrò, B. (1984), Bayesian Testing of Nonparametric Hypotheses and Its Application to Global Optimization, *Journal of Optimization Theory and Applictions* **42**, 31–50.
6. Betrò, B. and F. Schoen (1986), Una Tecnica Stocastica per l'Ottimizzazione Globale, in Bielli M. (ed.), *Ricerca Operativa e Informatica*, F. Angeli, Milano.
7. Betrò, B. and F. Schoen (1987), Sequential Stopping Rules for the Multistart Algorithm in Global Optimization, *Mathematical Programming* **38**, 271–286.
8. Betrò, B. and F. Schoen (1988), Optimal and Suboptimal Stopping Rules for the Multistart Method in Global Optimization, *IAMI 88.11*, CNR-IAMI, Milano.

9. Betrò, B. and R. Zieliński (1987), A Monte Carlo Study of a Bayesian Decision Rule Concerning the Number of Different Values of a Discrete Random Variable, *Communications in Statistics – Simulation and Computation* **16**, 925–938.
10. Boender, C. G. E. and A. H. G. Rinnooy Kan (1983), A Bayesian Analysis of the Number of Cells of a Multinomial Distributions, *The Statistician* **32**, 240–248.
11. Boender, C. G. E. and A. H. G. Rinnooy Kan (1987), Bayesian Stopping Rules for Multistart Global Optimization Methods, *Mathematical Programming* **37**, 59–80.
12. Boender, C. G. E. and R. Zieliński (1985), A Sequential Bayesian Approach to Estimating the Dimension of a Multinomial Distribution, *Banach Center Publications*, 16, Warsaw.
13. DeGroot, M. H. (1970), *Optimal Statistical Decisions*, McGraw-Hill, New York.
14. Dixon L. C. W. and G. P. Szegö (eds.) (1978), *Towards Global Optimization II*, North Holland, Amsterdam.
15. Doksum, K. (1974), Tailfree and Neutral Random Probabilities and Their Posterior Distributions, *Annals of Probability* **2**, 183–201.
16. Ferguson, T. S. (1973), A Bayesian Analysis of Some Nonparametric Problems, *Annals of Statistics* **1**, 209–230.
17. Ferguson, T. S. (1974), Prior Distributions on Spaces of Probability Measures, *Annals of Statistics* **2**, 615–629.
18. Ferguson, T. S. and E. G. Phadia (1979), Bayesian Nonparametric Estimation Based on Censored Data, *Annals of Statistics* **7**, 163–186.
19. Gikhman, I. I. and A. V. Skorohod (1974), *The Theory of Stochastic Processes*, Springer, Berlin.
20. Gordon, W. J. and J. A. Wixom (1978), Shepard's Method of Metric Interpolation to Bivariate and Multivariate Interpolation, *Mathematics of Computations* **32**, 253–264.
21. Johnson, N. L. and S. Kotz (1976), *Distributions in Statistics: Continuous Multivariate Distributions*, Wiley, New York.
22. Kushner, H. (1962), A Versatile Stochastic Model of a Function of Unknown and Time Varying Form, *J. of Mathematical Analysis and Applications* **9**, 379–388.
23. Mockus, J. (1975), On Bayesian Methods of Optimization, in Dixon L. C. W. and G. P. Szegö (eds.), *Towards Global Optimization*, North Holland, Amsterdam.
24. Mockus, J. (1980), The Simple Bayesian Algorithm for the Multidimensional Global Optimization, in Archetti, F. and M. Cugiani (eds.), *Numerical Techniques for Stochastic Systems*, North Holland, Amsterdam.
25. Mockus, J., V. Tiešis, and A. Žilinskas (1978), The Application of Bayesian Methods for Seeking the Extremum, in Dixon L. C. W. and G. P. Szegö (eds.), *Towards Global Optimization II*, North Holland, Amsterdam.
26. Piccioni, M. and A. Ramponi (1990), Stopping Rules for the Multistart Method When Different Local Minima Have Different Function Values, *Optimization* **21**, 697–707.
27. Strongin, R. G. (1989), The Information Approach to Multiextremal Optimization Problems, *Stochastics and Stochastics Reports* **27**, 65–82.
28. Zieliński, R. (1981), A Statistical Estimate of the Structure of Multiextremal Problems, *Mathematical Programming* **21**, 348–356.
29. Žilinskas, A. (1978), On Statistical Models for Multimodal Optimization, *Math. Operationsforsch., Ser. Statistics* **9**, 255–266.
30. Žilinskas, A. (1978), Optimization of One-Dimensional Multimodal Functions, Algorithm AS133, *Applied Statistics* **27**, 367–375.
31. Žilinskas, A. (1981), On Multimodal Minimization Algorithm Constructed Axiomatically, *Methods of Operations Research* **40**, 197–200.
32. Žilinskas, A. (1981), Two Algorithms for One-Dimensional Multimodal Optimization, *Math. Operationsforsch., Ser. Optimization* **12**, 53–63.
33. Žilinskas, A. (1982), Axiomatic Approach to Statistical Models and Their Use in Multimodal Optimization Theory, *Mathematical Programming* **22**, 104–116.
34. Žilinskas, A. (1985), Axiomatic Characterization of a Global Optimization Algorithm and Investigation of Its Search Strategy, *Operations Research Letters* **4**, 35–39.