On a New Stochastic Global Optimization Algorithm Based on Censored Observations

FABIO SCHOEN
Department of Computer Sciences, University of Milano, via Comelico 39/41, I-20135 Milano, Italy (e-mail: schoen@ghost.dsi.unimi.it)

(Received: 28 July 1992; accepted: 29 May 1993)

Abstract. In this paper a new algorithm is proposed for global optimization problems. The main idea is that of modifying a standard clustering approach by sequentially sampling the objective function while adaptively deciding an appropriate sample size. Theoretical as well as computational results are presented.

Key words. Global optimization, Multistart algorithm, sequential stopping rules, nonparametric models, clustering techniques, censored observations.

Introduction

In this paper the problem of finding a global optimum of a continuous function f over a compact set K of $\mathbb{R}^d$ is considered, i.e. the problem of finding a value f^* such that

$$f^* = \max_{x \in K \subseteq \mathbb{R}^d} f(x) .$$

Many techniques have been proposed in the literature for dealing with this hard computational problem: for extensive surveys the reader is addressed to (Törn & Žilinskas, 1989, Betrò, 1991; Schoen, 1991; Zhigljavsky, 1991).

Among the most promising methods, in the author's opinion, a special attention deserve methods based on clustering analysis (see, for a detailed discussion, Rinnooy Kan & Timmer, 1987a,b); these methods provide a quite efficient way to implicitly explore most of the regions of attraction of different local optima without sacrificing too many local searches. Very briefly a clustering method for global optimization consists in sampling a certain number of points from K, transforming the sample in order to concentrate sampled points on regions of attractions of local optima, identifying the clusters, performing some local searches from selected points (representative of different clusters) and, possibly, iterating the whole procedure again, until a stopping criterion is satisfied.

Unfortunately, despite computational experience tends to support the feasibility of such methods, at least for moderately-sized problems, a few issues which are particularly critical for an efficient use of clustering techniques have deserved scarce attention in the literature. The first such issue is the problem of determin-

Journal of Global Optimization **4**: 17–35, 1994.

ing an appropriate sample size: choosing a sample size which is too high inevitably produces an unnecessary computational cost; on the other size, the choice of a small sample size may force, during different phases of the algorithm, the recalculation of most of the clusters identified in preceding phases. Another critical issue is the question of stopping the algorithm: a sensible criterion for deciding whether it is worthwhile to continue with a new sample is still to be developed and it might never be. This happens because the technique generally used for concentrating the sample is that of retaining a fixed percentage of those points which correspond to the highest function values; this induces a stochastic dependence on the sampled points, and thus some of the stopping rules discussed in the literature (see for example, Betrò & Schoen, 1987, 1992; Boender & Rinnooy Kan, 1987) can be no longer applied.

As a final critical remark on current global optimization methods based on clustering we should mention the fact that no measure of the computational cost incurred in performing a cluster analysis is explicitly included in the algorithms; moreover all such algorithms aim at discovering the regions of attraction of each local maximum, independently of the associated function value. In view of the fact that there never is any guarantee that such a goal is reached, it seems worthwhile to look for algorithms in which local searches are started from "promising" starting points only.

The aim of this paper is to provide a computational framework which combines some of the advantages of clustering techniques while trying to avoid some of the above mentioned difficulties. In particular, in the sampling phase, some of the observed function values are seen as *censored* observations of a local optimum; in other words the function value observed at a feasible point belonging to the region of attraction of a local optimum can be seen as a truncated observation of such an optimum. It is possible then, after a suitable stochastic model is given for the function values of local optima, to make inference on the global optimum. Such an inference is based upon two different kinds of informations: regular (or non-censored) observations which correspond to function values at local optima, as returned, e.g., by a local optimization routine; and censored observations, which are just function values at randomly chosen feasible points. The main innovation in the approach proposed in this paper consists in the introduction of a model for the objective function values at the local optima which is capable of taking into account not only, as obvious, the observed local optima, but also function values at points which are not stationary.

The main aim of this paper is to bring to attention on how it is possible to use censored information to derive inferences on the global optimum; an algorithm is eventually proposed in which such an inference mechanism is used to control the cardinality of the sample in a standard clustering method.

The paper is organized as follows: in Section 1 an overview of some techniques for inference in presence of censored observations is given; Section 2 deals with the theoretical difficulties connected with stopping a sequential sample in

presence of censored observation; an application to clustering methods is presented in Section 3, while, in Section 4, computational results are presented and discussed.

1. Stochastic Models

In this section some theoretical considerations already reported in Betrò & Schoen (1987, 1992) are recalled, and some results related to the use of censored observations introduced in Ferguson & Phadia (1979) are presented.

In Betrò & Schoen (1987, 1992) an analysis is introduced of the Multistart method. We recall here that Multistart is a Montecarlo-like method which consists in repeatedly performing local searches from randomly chosen starting points. In those papers it is recalled that, under very mild assumptions, the whole process of Multistart can be seen as the simulation of a discrete random variable T with values in the set $\mathcal{F} = \{f_1, f_2, \ldots, f_n = f^*\}$ of local optima of the objective function with discrete density function given by

$$P(T = f_i) = p_i = \frac{\mu(R_i)}{\mu(K)} \quad i = 1, \ldots, n ,$$

where $\mu(A)$ is used to denote the volume, or Lebesgue measure, of a region $A \subseteq \mathbb{R}^d$ and R_i is the region of attraction of the i-th local optimum, i.e. R_i is the subset of K characterized by the fact that a local search started from any point in R_i leads to a local optimum whose value is f_i. Obviously in any sensible application of Multistart neither the set $\mathcal{F}$ nor its cardinality is *a priori* known. The most critical point in the implementation of Multistart is the definition of an appropriate stochastic model for T upon which inference should be based. Were $\mathcal{F}$ known, then Multistart could have been easily seen as the simulation of a multinomial random variable; this approach is the starting point of Boender & Rinnooy Kan (1987), where stopping rules are provided based on a Bayesian decision-theoretic framework in which a prior distribution is imposed over n and the "shares" p_i, $i = 1, \ldots, n$. While that model is indeed sensible and provides manageable criteria for Multistart, it does not make any explicit use of the observed function values at the local optima, thus loosing a crucial information gained during the execution of the algorithm. In Piccioni & Ramponi (1990) a variant of the same approach is introduced in which function values have some influence on the overall process, by assuming that each local optimum has a different function value and thus observations can be ranked according to function values at the optima. Nevertheless there is no attempt of explicitly give a stochastic model of the values in $\mathcal{F}$; in the cited papers of Betrò & Schoen (1987, 1992) a stochastic nonparametric model for T is introduced and analyzed. We summarize here the origin and motivation of such a model.

A *random distribution function* $F_\omega(t)$ on $\mathbb{R}$ is defined as a stochastic process such that (omitting the subscript ω)

1. $F(t)$ is almost surely (a.s.) nondecreasing;
2. $F(t)$ is a.s. right-continuous;
3. $\lim_{t\to-\infty} F(t) = 0$ a.s.;
4. $\lim_{t\to+\infty} F(t) = 1$ a.s..

In other words, a random distribution function is a stochastic process whose sample paths are, a.s., probability distribution functions. All classical Bayesian statistics is based upon similar processes; in fact Bayesian statistics deals with random variables whose distribution functions depend on one or more parameters and such parameters are themselves random variables. This way, depending on the possible values of such parameters, different distribution functions are obtained.

Unfortunately, apart from trivial examples, it seems impossible to model Multistart after a *parametric* family of distribution function; this is the reason why stochastic nonparametric models have been introduced. Among the class of such models a reasonable compromise between representativeness and computational manageability is achieved through the so-called *neutral to the right* process, which is a random distribution function for which, for any choice of $t_1 < t_2$, it holds that the random variables

$$\frac{1 - F(t_2)}{1 - F(t_1)}$$

and

$$F(t) \quad t \leqslant t_1$$

are stochastically independent.

Formally a random to the right distribution function is defined as a random distribution function $F(t)$ such that

$$F(t) = 1 - \exp(-Y_t)$$

where Y_t is a stochastic process such that

1. Y_t has independent increments.
2. Y_t is a.s. non decreasing;
3. Y_t is a.s. right-continuous;
4. $\lim_{t\to-\infty} Y_t = 0$ a.s.;
5. $\lim_{t\to+\infty} Y_t = +\infty$ a.s..

The following is one of the main results for neutral to the right distribution functions.

THEOREM 1. *Let F be a random distribution function neutral to the right and let*

T be a sample of size 1 *from F; let t be any real number. Then the posterior distribution of F given any of the following events*

- $T = t$;
- $T > t$;
- $T \geqslant t$

is still a random distribution function neutral to the right.

Proof. See Ferguson & Phadia (1979). □

In other words the class of such processes is closed under conditioning either on exact observations or on right censoring. This theorem is of fundamental importance and relevance for the application to the design of a global optimization algorithm; it states that if we model function values at local optima by means of a neutral to the right process, then we still end up with a neutral to the right process after the observation of some local optimum value or even after the observation of function values at points which are not local optima. This closedness property is not enjoyed by most stochastic models.

It is worthwhile to notice here that if the local searches in Multistart are not exact, then we are left exactly with right-censored observations, i..e with the information that the local optimum which the current local search would have discovered if it had been carried out until convergence has a value which is bounded below by the current observed value. Notice that even the observed function values at randomly chosen sample points are right censored observations of local optima.

In Ferguson & Phadia (1979) a random distribution function is introduced under the name of *simple homogeneous process* which is defined as a neutral to the right process characterized by the moment generating function

$$M_t(\theta) = E(\exp(-\theta Y_t)) = \exp\left(\gamma(t) \int_0^\infty \frac{e^{-\theta z} - 1}{1 - e^{-z}} e^{-\tau z} \, dz\right)$$

where $\tau > 0$ and $\gamma(t)$ is a continuous, non decreasing function such that $\lim_{t \to -\infty} \gamma(t) = 0$ and $\lim_{t \to +\infty} \gamma(t) = +\infty$. The function γ is easily seen to be related to the "prior guess" F_0 defined as

$$F_0(t) = E(F(t))$$

through the following

$$\begin{aligned} F_0(t) &= E(1 - \exp(-Y_t)) \\ &= 1 - M_t(1) \\ &= 1 - \exp(-\gamma(t)/\tau) \, . \end{aligned}$$

Thus γ is a "parameter" which characterizes the prior expected distribution function.

One of the attractiveness of this particular class of random distribution functions is that it is relatively easy to derive *a posteriori* information based upon a sample.

Let F be a simple homogeneous process, and let T be a sample of size $k \geq 1$ from F consisting of censored as well as non censored observations; let such observations be grouped according to the following scheme:

- $u_1 < u_2 < \cdots < u_m$ are the ordered *different* sampled values (m depends on k);
- let $u_0 = -\infty$ and $u_{m+1} = +\infty$;
- let n_i, $i = 0, \ldots, m$ be the number of non censored observations in T whose value is u_i;
- let $n_i^{>}$, $i = 0, \ldots, m$ be the number of censored observations in T whose value is u_i (here, for simplicity, by censored we mean *strictly* censored observations, i.e. observation of the kind $T_i > u_i$);
- let also h_i, $i = 0, \ldots, m$ be the number of observations (censored or not) strictly greater than u_i.

Then the following holds:

THEOREM 2. *Let T be a sample of size $k \geq 1$ from a simple homogeneous process F; then the posterior expectation of $F(t)$ given T is given by*

$$
\begin{aligned}
1 - \hat{F}_k(t) &= 1 - E(F(t) \mid T) \\
&= (1 - \hat{F}_0(t))^{\tau/(h_j+\tau)} \\
&\quad \cdot \prod_{i=1}^{j} \frac{h_i + n_i^{>} + \tau}{h_i + n_i^{>} + n_i + \tau} \\
&\quad \cdot \prod_{i=1}^{j} (1 - \hat{F}_0(u_i))^{-\frac{\tau(h_{i-1}-h_i)}{(h_{i-1}+\tau)(h_i+\tau)}} \\
&\quad \cdot \mathbf{1}_{\{t \in [u_j, u_{j+1})\}} \qquad j = 0, \ldots, m
\end{aligned}
\tag{1}
$$

Proof. See Ferguson & Phadia (1979). □

It should be quite evident that all of the data required in the statement of the previous theorem is readily available during the execution of an optimization algorithm: values u_i, $i = 1, \ldots, m$ are distinct observed function values, n_i is the number of observations whose function value is u_i which satisfy the usual optimality conditions of *local* optimization; $n_i^{>}$ is the number of observations whose function value is u_i but which do not satisfy local optimality conditions. From a practical point of view, the distinction between censored and non

censored observations is based upon the return value of the local optimization routine which is used in the algorithm.

2. Application to Global Optimization

The use of stochastic nonparametric models for the values of the local optima discovered by Multistart has been extensively studied in Betrò & Schoen (1987, 1992); there the model was used in order to develop suitable stopping criteria. In particular both theoretical as well as experimental evidence supported the feasibility of stopping Multistart on the basis of a comparison between the best optimum found and the expected improvement after the next local optimization. This stopping scheme goes under the name of 1-*sla*, or *one-step look-ahead*; the stopping rule can be expressed as that rule which calls for stopping after the k-th observation if

$$\int_{u_m}^{\infty} (1 - \hat{F}_k(u))\, \mathrm{d}u \leq c \tag{2}$$

or, equivalently,

$$\int_{u_m}^{\infty} (u - u_m)\, \mathrm{d}\hat{F}_k(u) \leq c\,. \tag{3}$$

Here c is a positive constant.

The main drawback of Multistart is notoriously its inability to stop a local search if it is likely that it will lead to an already discovered local maximum. Cluster-based techniques aim exactly at reducing the number of unnecessary local searches. Here we propose a modified clustering technique in which a sample is taken of the objective function whose cardinality is not known *a priori*; the idea is to sample from the objective function until the expected improvement falls below a treshold; after the sampling phase, points are clustered and local searches started from a few selected points in the sample. A conceptual algorithmic scheme could be as follows:

1. sample uniformly one point from K;
2. perform a few steps of a local ascent algorithm;
3. update the estimate of $F(t)$ given all of the observations (exact or censored depending on whether the ascent steps have led to a local optimum or not);
4. compute $\int_{u_m}^{\infty} (1 - \hat{F}_k(u))\, \mathrm{d}u$;
5. if it is greater than c then repeat from 1;
6. otherwise perform cluster analysis and stop.

This way the cardinality of the sample is decided on the basis of a sequential

sample which is stopped when random sampling is not likely to be able to produce significantly better observations.

Unfortunately the above scheme might not lead to a practical implementation, due to the fact that the following property which is valid if all of the observations are non-censored, does not hold in the censored case:

THEOREM 3. *If all observations are non censored and* $\hat{F}_0$ *is chosen in such a way that* $\hat{F}_0(f^*) < 1$, *then, for all* $t \leq f^*$,

$$\lim_{k \to \infty} \hat{F}_k(t) = P(T \leq t) \text{ a.s.} \tag{4}$$

Moreover, under the same hypotheses, the stopping time N, *i.e. the random variable*

$$N = \min \left\{ k \geq 1 : \int_{u_m}^{\infty} (1 - \hat{F}_k(u))\, du \leq c \right\} \tag{5}$$

is finite with probability 1 *for all* $c > 0$.

Proof. See Betrò & Schoen (1992). □

In the censored case it is obvious that consistency (convergence of the estimated distribution function to the actual one) cannot in general be achieved, as one cannot hope to perfectly learn a distribution function if he can never obtain exact observations (just think of trying to estimate the probability of the outcomes of a dice given, say, only observations of the kind "the outcome is 1" or "the outcome is greater than 1"). For what concerns stopping, it should be observed that for finite stopping times it is neither necessary nor sufficient that the estimate of the distribution function is consistent.

In order to obtain a finitely convergent algorithm it seems necessary to obtain non-censored observations "sufficiently often". The following theorem gives a sufficient condition which illustrates this point.

THEOREM 14. *If*

- $\hat{F}_0(f^*) < 1$;
- $\int_u^{\infty} (1 - \hat{F}_0(t))\, dt < \infty$ *for all* $u : F(u) > 0$;
- $\lim_{k \to +\infty} n_m = +\infty$ *w.p.*1
- *there are no censored observations with value* u_m,

then, almost surely, stopping will occur in a finite number of steps for any choice of $c > 0$.

Proof. See the appendix. □

The above theorem states that, in order to guarantee that a global optimization algorithm based upon the proposed stochastic model will stop after a finite

number of function evaluation, it is sufficient to require that at least the best possible observation is non censored and that it is observed with sufficiently high frequency; the first requirement comes from the fact that if the best function value observed so far is censored, we will have a strong tendency to continue the sample in order to get an even better observation. From the assumptions of this theorem then it seems more appropriate to say that the proposed algorithm is based "also", but not exclusively, on censored observations. It is to be remarked that no published paper in the global optimization literature report any attempt of using function values in points which are not local optima for the updating of a probabilistic model.

It can be remarked here that in practice it has been observed that convergence is faster than what can be expected from the above theorem. However a growing number of exact observations seems unavoidable and, moreover, it seems that convergence for any value of c cannot be achieved by simply having a sufficient number of non-censored observations, but one has to repeatedly observe the highest one, i.e. u_m.

3. Algorithmic Details

The theoretical results of the preceding section imply that, in order to build an algorithm which stops in a finite number of steps, a sufficient number of exact observation of the highest observed value are necessary; obviously algorithms like Multistart satisfy the hypotheses of Theorem 4 if, as it is usually assumed, the region of attraction of the global optimal has non-null Lebesgue measure. On the other hand, the conceptual algorithm introduced in Section 2 does not satisfy the requirements of Theorem 4 and it is possible to actually display examples for which finite termination is not achieved.

In what follows a modification of that basic scheme is proposed which is mainly justified by the desire to reproduce, as closely as possible, the behaviour of a classical clustering technique while providing a control mechanism based upon the theoretical results of the previous section.

As a well-known and easily implemented example of clustering-based algorithm, let us consider a simplified version of the so-called Multi-Level Single-Linkage algorithm of Rinnooy Kan & Timmer (1987b) (which, strictly speaking, is not a true clustering algorithm):

1. choose an integer N^*; let $k = N^*$;
2. sample N^* points uniformly and independently from K and add them to previously sampled points;
3. let

$$r_k = \pi^{-\frac{1}{2}} \left(\sigma \mu(K) \Gamma\left(1 + \frac{d}{2}\right) \frac{\log k}{k} \right)^{\frac{1}{d}}, \tag{6}$$

where σ is a positive constant, $\mu(\cdot)$ is the Lebesgue measure, $\Gamma(\cdot)$ is the gamma function;

4. apply a local optimization routine starting from each sampled point x_i *except* if there is another sampled point x_j, such that
   (a) $f(x_j) > f(x_i)$;
   (b) $\|x_i - x_j\| \leq r_k$;
5. if a stopping condition is satisfied stop; otherwise set $k = k + N^*$ and repeat from 2.

A crucial difficulty in the implementation of this algorithm is the appropriate choice of N^* and the development of a suitable criterion for stopping.

What is particularly interesting in the present context is that the idea underlying this algorithm is that observations which fall sufficiently near to a higher valued one, say x_j, are considered as belonging to the same region of attraction of x_j. Obviously this decision, in view of the fact that $(\log k)/k$ tends to zero as $k \to \infty$, might be revised during successive iterations.

Let us call a point x_i *clustered at step* k with x_j if

- either $f(x_j) > f(x_i)$ and $\|x_i - x_j\| \leq r_k$;
- or there exist another sampled point x_s such that x_i is clustered with x_s and x_s is clustered with x_j.

In Multi-Level Single-Linkage, if a point is clustered with another one, it is also clustered with a point from which a local optimization routine has been started; let us call u_i the function value at the local optimum found by such a routine. In the following algorithm, based upon the observation that, if a clustering algorithm is stopped, no local search is started from clustered points, for the purpose of stopping it is assumed that the observation at x_i is the (non-censored) value u_i. In other words, we assign the exact value u_i to all points clustered with a single one from which a local optimization was performed leading to u_i.

The idea is now to avoid sampling in batches, i.e. we let $N^* = 1$, and to perform a single sampling cycle which is stopped when the expected gain falls below a threshold c; during the sampling phase tentative clusters are grown only around the best local optima, while a complete clustering is performed, only once, at the end of the algorithm. In order to prevent clusters to be built and destroyed too frequently, we suggest to use a conservative value for the threshold r_k: in the experiments reported in the next section, the value r_k was substituted by $r_{H\lceil k/H \rceil}$, with $H = 1000$. The algorithmic scheme is thus the following:

1. choose an integer H and a positive real c; let $k = 1$, $m = 0$, $u_m = -\infty$;
2. sample *one* point x_k uniformly and independently from K and add it to previously sampled points;
3. if $f(x_k) > u_m$ then

(a) start a local optimization routine from x_k;
(b) let u_m be the observed local optimum value;
(c) let $opt(x_k) = u_m$ and consider this value as a non-censored observation.

4. Otherwise, if $f(x_k) \leq u_m$,
   (a) if there exist a sampled point x_j such that $f(x_j) > f(x_k)$ within distance $r_{H\lceil k/H \rceil}$ from x_k, then set $opt(x_k) = opt(x_j)$, and consider this value as a non-censored observation.
   (b) Otherwise, consider the value $f(x_k)$ as a censored observation; set $opt(x_k) =$ '*undefined*';
5. update the estimate of F given by (1) on the basis of the new observation;
6. if $\int_{u_m}^{\infty} (1 - \hat{F}_k(t))\, dt > c$ then
   (a) set $k = k + 1$;
   (b) if k is a multiple of H update all of the clusters on the basis of the new treshold $r_{H\lceil k/H \rceil}$;
   (c) goto 2;
7. Otherwise, cluster all unclustered points using r_k as a treshold and start local optimization routines following the scheme of Multi-Level Single-Linkage.

The above algorithm, although quite cumbersome, implements the simple idea of sequentially deciding the sample size in order to let sampling terminate as soon as there is sufficient evidence that a high value of the objective function has been observed.

In the following section some numerical experiments will be presented in order to support the theoretical results introduced in this paper.

4. Numerical Experiments

In order to test the effectiveness of the proposed algorithm a number of numerical experiments have been performed. Here the experimental setting is presented.

The test functions used in all of the experiments are those first introduced by Betrò (1984) and successively published, with a *C* code, in Schoen (1993); the general form of the test functions is the following

$$f(x) = \frac{\sum_{i=1}^{k} f_i \prod_{j \neq i} \|x - z_j\|^{\alpha_j}}{\sum_{i=1}^{k} \prod_{j \neq i} \|x - z_j\|^{\alpha_j}}, \quad x \in \mathbb{R}^d$$

where k is a positive integer corresponding to the number of stationary points of f and $\{z_i,\ f_i,\ \alpha_i,\ i = 1, \ldots, k\}$ are parameters defining the location, value and degree of smoothness of the stationary points of f. In the experiments reported in this paper, 100 test functions were randomly generated according to the following scheme:

- $d = 2$;

- $K = [0, 1]^d$;
- for each function the number of stationary points, k, was obtained as a discrete uniform random number between 10 and 100;
- the global minimum f_* and a maximum f^* of f were generated as an ordered sample of size 2 from a uniform distribution on $[0, 100]$
- the remaining $k-2$ function values at stationary points were independently and uniformly distributed on $[f_*, f^*]$;
- the locations z_i, $i = 1, \ldots, k$ were chosen uniformly and independently on $[0, 1]^d$;
- the "smoothness parameters" α_i were chosen uniformly and independently in $[1.8, 2.2]$, thus ensuring that, at least, $f \in \mathscr{C}^1(K)$.

The local optimization routine used was LMQNBC, freely available from NETLIB, with "standard" parameter settings, i.e.

maxit	1000
η	0.25
stepmx	10.
accrcy	$10.^{-15}$

For a description of the method and the definition of the above parameters the reader is referred to Nash (1984).

For each of the 100 test functions a total of 5 independent runs of the algorithm were performed, with the following experimental setting:

- the prior guess $\hat{F}_0(t)$ was chosen as the distribution function of a random variable uniformly distributed on $[0, 100]$[1];
- $\tau = 1$
- $H = 1000$;
- after H observations the algorithm was artificially stopped (this artificial stopping actually took place only in 18 of the 500 experiments);
- no local ascent steps were performed to transform the sample;
- $\sigma = 4$ in (6);
- $c = 5$.

The choice for c was somewhat arbitrary, and chosen after a few experimentation. However it can be easily seen from the form of the stopping criterion and the definition of $\hat{F}_0$ that, with this choice of c, if the *first* observation is non censored and greater than 65.8, then the algorithm immediately stops. In other words, given the experimental setting and the knowledge of the whole class of test

[1] Notice that this is *not* the distribution function of the stationary values f_i.

functions, the "user" feels satisfied, and thus likes to stop, if he/she is offered a first observation higher than 65.8. That is, he or she thinks that a global optimum with value between 65.8 and 100, although possible, is sufficiently improbable so that he or she is not willing to pay the computational cost of further sampling. Please notice that this is only the initial situation: after a few observations, the posterior distribution might have changed in such a way that stopping will not occur even for substantially higher observations. That is, the left hand side of (2) is not monotonically decreasing with k, nor it is decreasing with u_m.

All of the experiments were made on a SUN SPARC station and the algorithm was coded in C by the author. In the figures some statistics on the computational experiments are visually reported. In all figures each of the values displayed is the average over 5 independent runs on the same objective function. In Figure 1 a comparison between the true global optimum and the (average) observed one is displayed.

It should be observed that the accuracy of the algorithm is very high; it slightly deteriorates for test functions with global optimum value greater than 70. This is due to the choice of an high value for c; however, looking also at Figure 2, it can be observed that only in rare cases the error is greater than 10%.

The fact that early stopping caused by too high a value for c is responsible for lower accuracy can be easily deduced from Figure 3.

One of the peculiarities of the algorithm proposed in this paper is that sampling

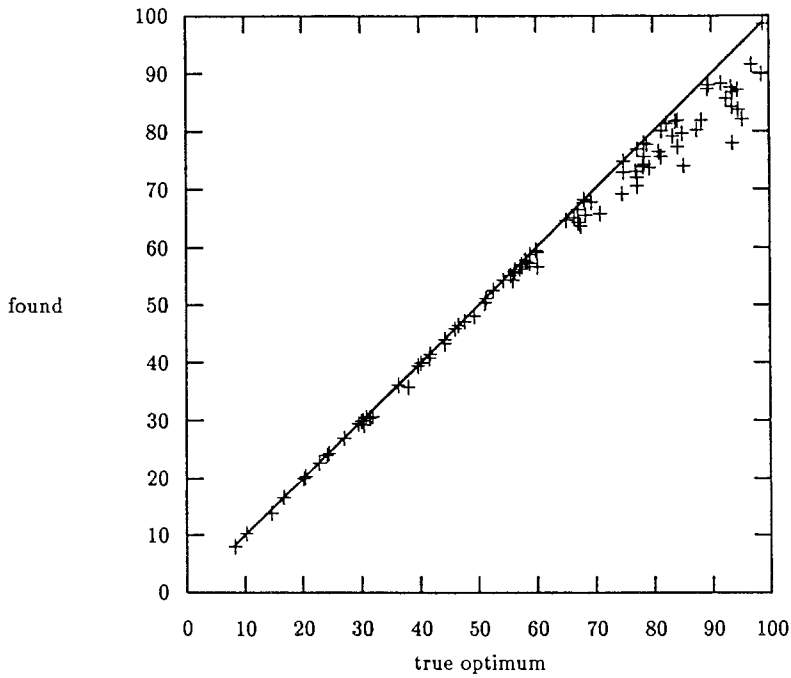

Fig. 1. Global optimum: true vs. found.

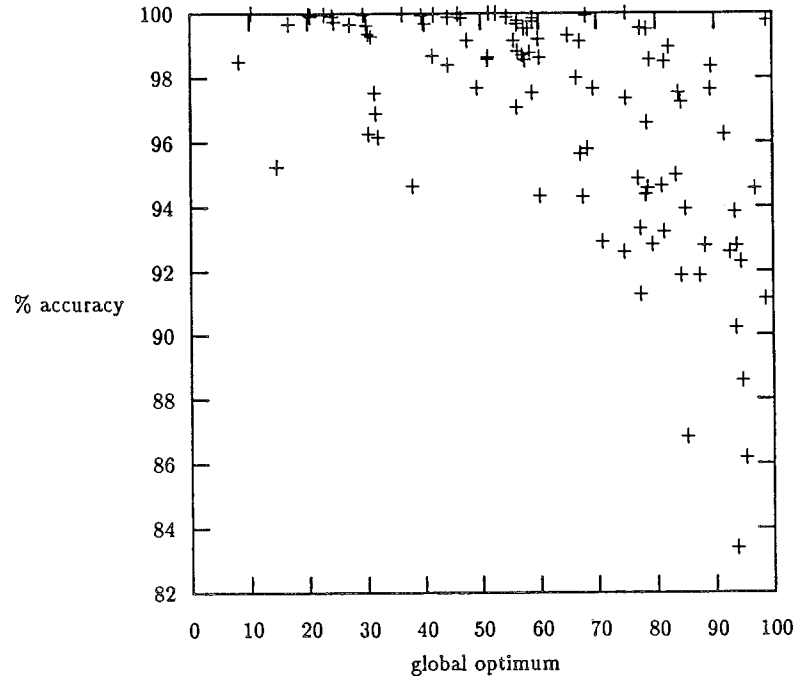

Fig. 2. Average error vs. global optimum.

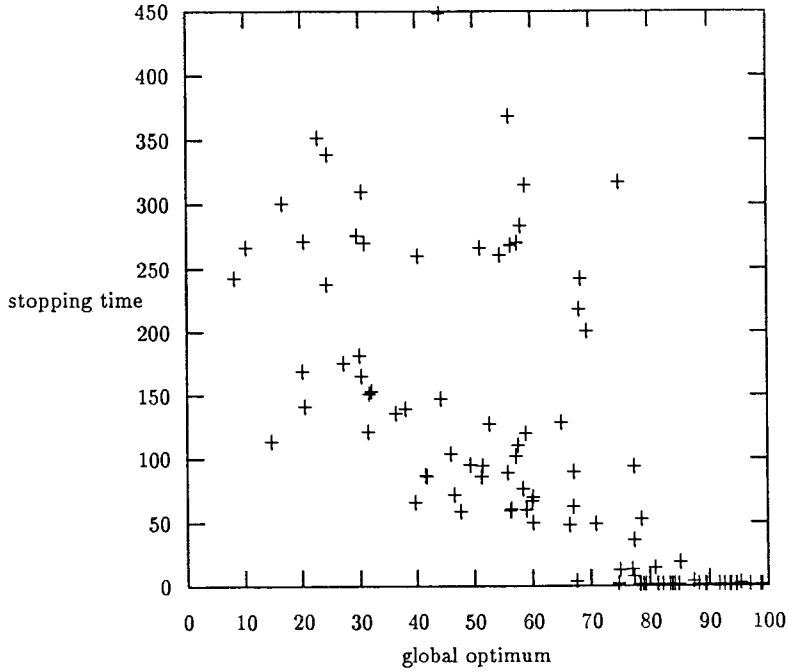

Fig. 3. Average stopping time vs. global optimum.

is continued until there is sufficient evidence that the global optimum has been observed; this is to be contrasted with other well known approaches in which sampling is carried out until there is supporting evidence to the fact that *all* of the local optima have been discovered. Obviously much is to be gained by relaxing this requirement of observing all local maxima. It can be noticed from Figure 4 that the proposed algorithm is quite insensible to the number of stationary points in the objective function.

A simple statistical analysis of the data reported in Figure 4 supports the claim of weak dependence of the accuracy of the algorithm on the number of stationary points; a simple linear regression on the data of the figure gives the line $y = -0.000330354x + 0.985914$ as the least squares approximation. This line is reported also in the picture. A standard test for the null hypotheses $\mathcal{H}_0$: *the slope is* 0 versus the alternative *the slope is not* 0 gives a value 2.55179 for the T statistics with 98 degrees of freedom. The null hypothesis is rejected at the 1% confidence level, while it is not rejected at the 0.5% level. This is a typical situation in which it is difficult to draw a clear conclusion, but it shows that, if a dependence of the accuracy of the algorithm on the number of stationary points of the objective function does exist, it certainly is very weak.

In Table I statistics are reported over the whole set of 500 runs.

The last two columns report respectively the number of local searches started during the sampling phase and the number of local searches started after the stopping condition was met: it is somewhat surprising the extremely low total number of local searches, especially when compared with the relatively high

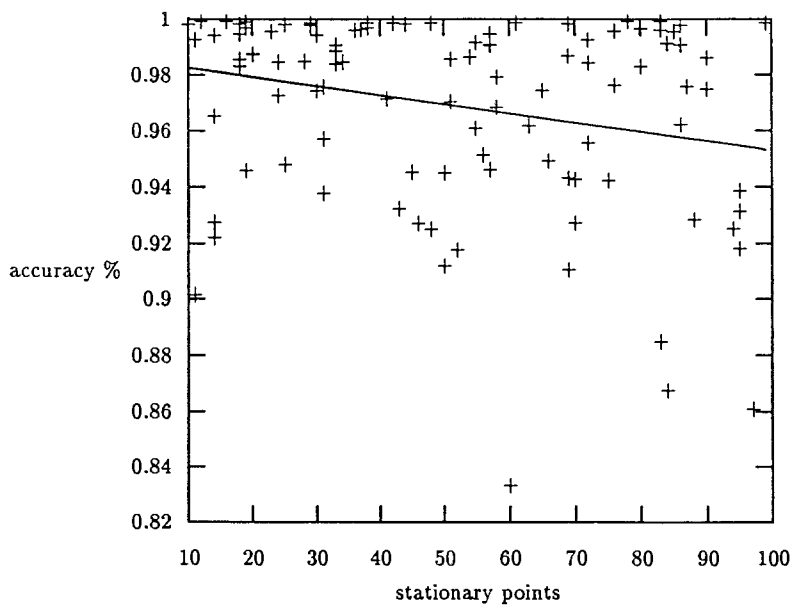

Fig. 4. Accuracy % vs. number of stationary points.

Table I. Computational results

	n. iterations	true opt.	opt. found	funct. eval.	grad. eval.
avg.	107.826	60.92603	58.53280	165.224	57.072
std. dev.	203.6142	23.97212	22.17907	236.6804	106.3163
min	1	8.15752	7.93186	6	5
max	1000	98.8236	98.7174	1223	1123
	local searches 1	local searches 2			
avg.	1.496	0.852			
std. dev.	0.702839	1.149824			
min	1	0			
max	4	6			

number of stationary points (not all of which are of course local maxima, but, it can be safely assumed, roughly half of them are indeed local maxima, thus giving an estimate of more than 20 optima for the average test function). It seems, from this initial computational results, that the algorithm succeeds in sampling, at low computational cost, the objective function, starting only few local searches from very promising starting points.

Conclusions and Future Extensions

A proper mixture of sequential statistical control of sampling and clustering algorithms has been proposed and analyzed in this paper. The resulting algorithm seems to be very efficient in terms of number of function evaluations and number of local searches performed versus accuracy. It is obvious that such an efficiency should be tested against some different algorithm; however the comparison cannot, at present, be performed due to the lack of good, public domain global optimization code. Moreover, as it is recognized by several researchers in this field, there is no single criterion on which a comparison should be based: there are conflicting objectives in the design of global optimization algorithms, some of which are accuracy, speed, memory requirements. This paper tries to suggest a possible implementation scheme which permits to get the global optimal value of functions with a moderately high number of local optima without performing too many local optimizations.

This is just the first implementation and the results here reported are only the first ones obtained. Research is still going on on this subject as well as more thorough experimentation carried out. In particular it seems worthwhile to explore the tradeoff between higher computational cost incurred by performing a few ascent steps from each sampled point and the probable improvement in final accuracy. Another issue still to be addressed concerns the possibility of using different clustering strategies. Finally, the question of finite termination, although a sufficient condition is given here, surely requires deeper understanding.

Acknowledgements

My special thanks to Dr. Bruno Betrò who first had the idea of using censored observation in a global optimization context. Sincere thanks also to the Computer Science Department of the State University of Milan, and, in particular, to the Operations Research Laboratory, where the numerical computations were carried out.

This research has been partially supported by "Progetto MURST 40% Metodi di Ottimizzazione per le Decisioni".

Appendix: Proof of Theorem 4

Proof. Stopping occurs as soon as

$$\int_{u_m}^{\infty} (1 - \hat{F}_0(t))\,\mathrm{d}t \cdot \prod_{i=1}^{m} \frac{h_i + n_i^{>} + \tau}{h_{i-1} + \tau} (1 - \hat{F}_0(u_i))^{-\frac{\tau(h_{i-1} - h_i)}{(h_{i-1}+\tau)(h_i+\tau)}} \leqslant c\,. \tag{7}$$

This equation can be rewritten in a sometimes more convenient way as

$$\int_{u_m}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(u_m)}\,\mathrm{d}t \cdot \prod_{i=1}^{m} \frac{h_i + n_i^{>} + \tau}{h_{i-1} + \tau} \left(\frac{1 - \hat{F}_0(u_i)}{1 - \hat{F}_0(u_{i-1})}\right)^{\frac{\tau}{h_{i-1}+\tau}} \leqslant c\,. \tag{8}$$

(the proof is by trivial substitution).

Let us examine the different factors in (8) separately. The quantity

$$\int_{u_m}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(u_m)}\,\mathrm{d}t$$

can be bounded, for example, in the following way:

$$\begin{aligned}\int_{u_m}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(u_m)}\,\mathrm{d}t &\leqslant \int_{u_m}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(f^*)}\,\mathrm{d}t \\ &\leqslant \int_{t_1}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(f^*)}\,\mathrm{d}t\end{aligned}$$

where, to derive the first inequality, we used the monotonicity of $\hat{F}_0$ and the hypothesis $\hat{F}_0(f^*) < 1$, while the final inequality, in which t_1 is the first observation made, descends trivially by the monotonicity of the definite integral.

The second factor in (8) is the crucial one: we have

$$\begin{aligned}\prod_{i=1}^{m} \frac{h_i + n_i^{>} + \tau}{h_{i-1} + \tau} &\leqslant \frac{n_m^{>} + \tau}{n_m + n_m^{>} + \tau} \\ &= \frac{\tau}{n_m + \tau} \to 0\,.\end{aligned}$$

Here we simply used the fact that each factor in the above product is bounded by one, as

$$\frac{h_i + n_i^{>} + \tau}{h_{i-1} + \tau} = \frac{h_i + n_i^{>} + \tau}{h_i + n_i + n_i^{>} + \tau} \leqslant 1 .$$

It can be noticed here that the hypothesis $n_m^{>} = 0$, though quite easily satisfied in practice, is not necessary: it is sufficient that, if $n_m^{>} \neq 0$, it does not grow too rapidly, i.e. $\lim_{k\to\infty} n_m / n_m^{>} = \infty$.

Finally we have

$$\left(\frac{1 - \hat{F}_0(u_i)}{1 - \hat{F}_0(u_{i-1})}\right)^{\frac{\tau}{h_{i-1}+\tau}} \leqslant 1$$

thanks to the monotonicity of $\hat{F}_0$. In summary

$$\int_{u_m}^{\infty} (1 - \hat{F}_k(t))\, \mathrm{d}t \leqslant \frac{\tau}{n_m + \tau} \int_{t_1}^{\infty} \frac{1 - \hat{F}_0(t)}{1 - \hat{F}_0(f^*)}\, \mathrm{d}t$$

which converges to zero as k grows to infinity. □

References

Betrò, B. (1984), Bayesian testing on nonparametric hypotheses and its applications to global optimization, *J.O.T.A.* **42**, 31–50.

Betrò, B. (1991), Bayesian methods in global optimization, *Journal of Global Optimization* **1**, 1–14.

Betrò, B. & Schoen, F. (1987), Sequential stopping rules for the multistart algorithm in global optimisation, *Mathematical Programming* **38**, 271–286.

Betrò, B. & Schoen, F. (1992), Optimal and suboptimal stopping rules for the multistart algorithm in global optimisation, *Mathematical Programming* **57**, 445–458.

Boender, C. & Rinnooy Kan, A. (1987), Bayesian stoping rules for multistart global optimization methods, *Mathematical Programming* **37**, 59–80.

Ferguson, T. & Phadia, E. (1979), Bayesian nonparametric estimation based on censored data, *Annals of Statistics* **7**, 163–186.

Nash, S. G. (1984), Newton-type minimization via the Lanczos method, *SIAM J. Numer. Anal.* **21**(4), 770–788.

Piccioni, M. & Ramponi, A. (1990), Stopping rules for the multistart method when different local minima have different function values, *Optimization* **21**, 697–707.

Rinnooy Kan, A. H. & Timmer, G. (1987a), Stochastic global optimization methods. Part I: clustering methods, *Mathematical Programming* **39**, 27–56.

Rinnooy Kan, A. H. & Timmer, G. (1987b), Stochastic global optimization methods. Part II: multi level methods, *Mathematical Programming* **39**, 57–78.

Schoen, F. (1991), Stochastic techniques for global optimization: a survey of recent advances, *Journal of Global Optimization* **1**, 207–228.

Schoen, F. (1993), A wide class of test functions for global optimization, *Journal of Global Optimization* **3**, 133–137.

Törn, A. & Žilinskas, A. (1989), *Global Optimization*, Lecture Notes in Computer Sciences. Springer-Verlag, Berlin.

Zhigljavsky, A. A. (1991), *Theory of Global Random Search*, Mathematics and Its Applications (Soviet Series). Kluwer Academic Publishers, Dordrecht.