ECONOMETRIC INSTITUTE

A STOCHASTIC METHOD FOR GLOBAL OPTIMIZATION

C.G.E. BOENDER, A.H.G. RINNOOY KAN

L. STOUGIE and G.T. TIMMER

REPORT 8001 /0

ERASMUS UNIVERSITY ROTTERDAM. P.O. BOX 1738. ROTTERDAM, THE NETHERLANDS

A STOCHASTIC METHOD FOR GLOBAL OPTIMIZATION

by

C.G.E. Boender, A.H.G. Rinnooy Kan,
L. Stougie and G.T. Timmer
Econometric Institute, Erasmus University Rotterdam

ABSTRACT

A stochastic method for global optimization is described and evaluated. The method involves a combination of sampling, clustering and local search, and terminates with a range of confidence intervals on the value of the global optimum. Computational results on standard test functions are included as well.

Contents

	Page
1. Introduction	2
2. Clustering	5
2.1. Density clustering	5
2.2. Single linkage clustering	6
2.3. The gradient criterion	8
2.4. Computational comparison of the two procedures	8
3. A confidence interval	10
4. Computational experiments	13
5. Concluding remarks	15
Acknowledgements	16
References	16
Appendix	18

1. INTRODUCTION

The *global optimization problem* is to find the *global optimum* of a real valued objective function. Relatively few methods have been developed to solve this problem, certainly in comparison to the multitude of nonlinear programming methods that have been designed to find *local optima*. Yet in many applications, a truly global optimum rather than an arbitrary local one is asked for. There is no need, then, to elaborate on the obvious practical applicability of global optimization, nor on the computational challenge that it offers [3,5].

All methods for global optimization assume that a bounded set $S \subset \mathbb{R}^n$, containing the global optimum in its interior, is given in advance. Thus, with $x = (x_1, \ldots, x_n)$ and $f : S \to \mathbb{R}$, the problem will be to find

$$y_* = \min_{x \in S} \{f(x)\}. \tag{1}$$

The known methods for this problem can be divided into two classes according as to whether they are of a deterministic or of a stochastic nature.

Deterministic methods find the global optimum by an exhaustive search over S. Clearly, in order to insure optimality in a finite number of steps, certain restrictions have to be put on f : it is easy to construct objective functions for which the global minimum can only be recognized on a subspace of arbitrarily small volume. A popular assumption in certain deterministic methods is that a *Lipschitz constant* is given, i.e. a constant L such that for all $x, x' \in S$

$$|f(x) - f(x')| \leq L||x - x'||. \tag{2}$$

The upper bound on the rate of change of f implied by (2) can be used in various ways to bound the size of the subspaces occurring in the search.

Unfortunately, assumption (2) is rarely verifiable in practice. In addition, the computational effort required by these methods is quite formidable: it increases exponentially with n, the *dimension* of the problem. Better computational results have been obtained with a completely different deterministic method, based on the construction - by numerical integration - of paths along which the gradient of f points in a constant direction [9].

None the less, the typical combination of severe, sometimes implicit restrictions on the objective function on one hand and at best erratic computational behaviour on the other hand is not an attractive one.

Stochastic methods start from a sample of points drawn randomly from S. These methods offer an *asymptotic* guarantee: under mild conditions on f, the probability that the global optimum will be found by a stochastic method approaches 1 as the sample size increases [6]. In addition to a sampling or *global* phase, most stochastic methods contain a *local* phase, in which the sample is manipulated to yield a candidate solution value to (1). If possible, stochastic methods should terminate by providing the user with some probabilistic information on the quality of the obtained result.

The algorithm presented in this paper is a stochastic one. Its goal is to find all the local minima that are potentially global. These local minima will be found by means of a *local search procedure*, starting from appropriately chosen points in the sample.

We cannot afford to neglect any of these local minima, but neither can we afford - for reasons of computational efficiency - to find the same one again and again. Thus, given a particular local search procedure, we define the *region of attraction* of a local minimum x* to be the set of all points in S starting from which the local search procedure will arrive at x*, and strive to initiate the local search procedure no more than once in each relevant region of attraction.

In an effort to identify these regions of attraction, the algorithm invokes a *clustering procedure* (cf. [14]). Before every application of this procedure, however, the current sample is *transformed* by temporarily removing a pre-specified percentage of the sample points whose function values are relatively high, and by performing a single *steepest descent step* from the remaining ones. This is supposed to result in the formation of groups of relatively close points, each of which surrounds a promising local minimum. The aim of the clustering procedure is to identify the *clusters* of points corresponding to these groups.

In each application of the procedure, clusters are grown around appropriately chosen *seed points*, in a way yet to be described, until all points of the transformed sample have been allocated to a cluster. If in the course of doing so, one or more local minima are found that had not been discovered before,

the sample is increased by an additional, fixed number of points. The extended sample is transformed in the way described above, and the clustering procedure is applied anew.

Thus, before each application of the clustering procedure, we have available a set X^* containing the local minima that were found so far. We also have available a set $X^{(1)}$ containing sample points to which the local search procedure has been applied unsuccessfully, in the sense that this produced a local minimum x^* that was known already. Initially, both sets are empty. As far as the choice of seed points for the current clustering phase is concerned, we start by growing clusters from all the local minima in X^*; if any points in the transformed sample then remain unclustered, we start to use the points in $X^{(1)}$ as seed points; and if any points still remain unclustered, we choose the point $x^{(1)}$ with lowest function value among those, apply the local search procedure to $x^{(1)}$ to find a local minimum x^* (cf. [14]) and grow a cluster with either x^* or $x^{(1)}$ as a seed point, depending on whether or not x^* was already known to be a member of X^*. In the former case, $x^{(1)}$ is of course added to $X^{(1)}$. We repeat this last mentioned clustering step until all points have been assigned to a cluster.

It remains to describe in more detail how to grow and terminate a cluster around a given seed point. We have developed two different methods for this purpose, both to be discussed in Section 2. In that section, we also present some computational evidence that has motivated the rejection of one of them as being marginally less accurate.

If during any clustering phase no new local minima are added to X^*, the sampling is terminated and the smallest local minimum $y^* = f(x^*)$ that has been obtained is declared to be the candidate solution to (1). As a final step, the user has to be provided with some probabilistic information on the quality of this outcome. In Section 3, we use a result from [10] to show how the two smallest function values found in the complete sample can be used to construct a range of confidence intervals on the true value of the global minimum. These confidence intervals are valid under a mild restriction on f, which is, for example, satisfied if the Hessian in the global minimum is nonsingular. As will be demonstrated in Section 3, the smallest function value y^* found during the local phase can be incorporated subsequently to provide the user of the algorithm with a final range of confidence intervals, whose right end point is now equal to y^*.

The algorithm has been tested on the standard test functions for global optimization from [6], with good results that are reported in Section 4. Concluding remarks and a brief discussion of ongoing and future research are contained in Section 5.

2. CLUSTERING

In this section we describe and evaluate two procedures, based on *density clustering* and *single linkage clustering* respectively, that have been developed to grow and terminate a cluster around a given seed point. The general structure of both procedures is identical: selected points of the transformed sample are added to the cluster one by one until a *termination criterion* is satisfied. If any points of the transformed sample then remain unclustered, we proceed as described in Section 1.

The choice of proper termination criteria for the clustering procedures requires some assumption about the distribution over S from which the sample has been drawn. We shall be assuming here that the sample is generated from a *uniform* distribution over S.

2.1. Density clustering

In the first method, based on *density clustering*, a cluster will correspond to the points in a subset T of S of stepwise increasing volume. The cluster will be terminated if the number of points added to the cluster in a single step drops below a threshold level that is calculated in advance.

Let us assume first that the seed point of the cluster is a local minimum x^*. As motivated in Section 1, T should then ideally have the same shape as the region of attraction around x^*. It is difficult to characterize these regions in general. However, if we define

$$L(y) = \{x \mid x \in S,\ f(x) \leq y\},$$

it can be proved for every steepest descent local search procedure that the region of attraction around x^* contains every connected subset of L(y) that contains x^* as its only stationary point [4]. This suggests to let T correspond to such a subset for stepwise increasing values of y. The actual level sets L(y) may again be hard to construct, but we can approximate them by the level sets $\tilde{L}(y)$ around x^*, defined by the second order approximation $\tilde{f}$ of f around x^*:

$$\tilde{f}(x) = f(x^*) + \tfrac{1}{2}(x - x^*)^T H(x^*)(x - x^*)$$

where $H(x^*)$ is the Hessian of f in x^*. Thus, the connected subsets around x^* correspond to *ellipsoids* (cf. the use of *hyperspheres* in [14]).

As to obtaining $H(x^*)$, we note that if the local search procedure is a *quasi-Newton* one, we obtain an increasingly accurate approximation $\tilde{H}(x^*)^{-1}$ of $H(x^*)^{-1}$ in the course of the procedure. It is then a simple matter (e.g., by maintaining an LU decomposition of $\tilde{H}(x^*)^{-1}$), to obtain $\tilde{H}(x^*)$ as a good approximation of $H(x^*)$, and it is this approximation that was used in our computational experiments.

To give a complete description of the procedure, we derive a termination criterion for the growth of a cluster. Let $m(A)$ indicate the *Lebesgue measure* of $A \subseteq \mathbb{R}^n$ and define the *density* δ by

$$\delta = \frac{N}{m(S)} .$$

where N is the number of points in the current sample. If we ignore the effect of the steepest descent step, the assumption of a uniform distribution implies that δ is also the expected density within a cluster (cf. [14]). During the growth of a cluster, we increase the volume of T in each step by an amount such that as a result one new point is expected to enter the cluster. Thus, in the i-th step, we increase the volume of the ellipsoid to i/δ. Since this volume is given by

$$\frac{[2\pi(y - f(x^*))]^{n/2}}{\Gamma(1 + n/2)|H(x^*)|^{1/2}} ,$$

it follows that in the i-th step we have to check if there is at least one point x such that x is not yet clustered and

$$\pi(x - x^*)^T H(x^*)(x - x^*) \leq \left(\frac{i\Gamma(1 + n/2)|H(x^*)|^{1/2}}{\delta}\right)^{2/n} .$$

If no such x can be found, the termination criterion is satisfied.

If the seed point is a member of the set $X^{(i)}$ rather than a local minimum, the same procedure can be applied with $H(x^*)$ replaced by the unit matrix I.

2.2. Single linkage clustering

The second clustering method is based on the observation that the approximation of the regions of attraction by ellipsoids need not always be satisfactory. *Single linkage* clustering schemes [8] can be used to produce clusters of any geometrical shape.

The original single linkage method is an *agglomerative hierarchical procedure:* we start from the partition into single element subsets, and in each step we fuse two subsets E and E' whose distance (as measured by the minimal distance between a point in E and a point in E') is minimal.

We adapt this procedure to our purposes by assuming again that the seed point is a local minimum x^* and by defining the distance $d(x, x')$ between two points x and x' in the neighbourhood of this local minimum to be

$$d(x, x') = [(x - x')^T H(x^*)(x - x')]^{1/2}.$$

We have already remarked upon the availability of an estimate of $H(x^*)$ in the context of the density clustering procedure.

We now initialize a cluster C to contain the single unclustered point that is closest to x^*. In every subsequent step, we find an unclustered point x such that

$$D(x, C) = \min_{x' \in C} \{d(x, x')\}$$

is minimal, we add x to C and repeat, until $D(x, C)$ exceeds a certain threshold level in which case the termination criterion is satisfied.

To choose a proper threshold level for $D(x, C)$, we consider the probability distribution of distances between points within a cluster. In doing so, we shall again ignore the effect of the steepest descent step and assume that the points within a cluster still satisfy the original uniform distribution over S.

Under this assumption, we can estimate the probability that one point in a cluster has none of the other N-1 sample points at distance d or less by

$$\left(1 - \frac{d^n \pi^{n/2}}{\Gamma(1 + n/2)|H(x^*)|^{1/2} m(S)}\right)^{N-1}. \qquad (3)$$

Formula (3) reflects the fact that for each of the N-1 points separately the required probability can be approximated for small d by dividing the volume of an ellipsoid $(x - x')^T H(x^*)(x - x') \leq d^2$ by $m(S)$ and subtracting the result from 1. No results other than comparable asymptotic ones [7, 12] are available about the distribution of the *nearest neighbour statistic.*

The termination criterion for the current cluster is now said to be satisfied if the distance $D(x, C)$ is so large as to make the probability (3) smaller than a prespecified threshold level α. (Thus, the cluster is terminated if the hypothesis that x belongs to the uniform distribution is rejected, with α corresponding to the probability of a type I error.) The threshold level on $D(x, C)$ is therefore equal to

$$\left[\frac{\Gamma(1 + n/2)|H(x^*)|^{1/2}\, m(S)}{\pi^{n/2}} (1 - \alpha^{1/(N-1)})\right]^{2/n}$$

If the seed point is a member of $X^{(1)}$, then again the same technique can be used with I replacing $H(x^*)$.

2.3. The gradient criterion

To complete the description of the clustering procedures, we mention that in both procedures, before a point x is added to a cluster that is related to a local minimum x^*, the negative gradient at x is verified to point in the direction of x^*. More precisely, if x^* is the seed point, we approximate the derivative of f in x in the direction of x^* by

$$\frac{f(x + h(x^* - x)) - f(x)}{h||x^* - x||}$$

for small h, and reject x for the cluster if this value is positive. If $x^{(1)} \in X^{(1)}$ is the seed point, we verify in an analogous manner if the gradient at x is pointing in the same direction as the gradient at $x^{(1)}$. This modification turned out to be very useful from a computational point of view, in that it regularly allowed the identification of clusters that would have been overlooked otherwise.

2.4. Computational comparison of the two procedures

Early implementations of the two clustering procedures described in this section were the subject of limited computational experiments. Some representative results are gathered in Table 1. The procedures were coded in FORTRAN and run on the IBM 370/158 of the Delft Computer Centre. Three test functions were selected from the standard collection for global optimization listed in Table 2; for each of them, the number of relevant local minima is mentioned after their abbreviation in Table 1. For each function, both density clustering and single linkage clustering were applied to three independent samples of 500 points drawn from S. In Table 1, they are compared with respect to the number of local minima found, the number of clusters actually constructed, the number of function evaluations required excluding the initial one for each point, and the number of units of *standard*

time required for clustering and local search, where one unit of standard time corresponds to 1000 evaluations of the Shekel 5 test function in the point (4, 4, 4, 4)[6].

Table 1

		run 1		run 2		run 3	
		D	SL	D	SL	D	SL
GP(3):	l.m.f.	3	3	3	3	3	3
	c.f.	5	4	8	3	7	3
	f.e.	541	507	751	449	702	484
	u.s.t.	1.5	2.0	1.9	1.9	1.8	1.8
BR(3):	l.m.f.	3	3	3	3	3	3
	c.f.	7	3	3	3	3	3
	f.e.	550	418	430	433	424	427
	u.s.t.	1.8	1.8	1,4	1.9	1.4	2.1
S7(7):	l.m.f.	7	7	6	6	7	7
	c.f.	7	7	6	6	8	7
	f.e.	755	773	716	731	822	772
	u.s.t.	2.4	2.1	2.3	2.0	2.9	2.2

D : density clustering
SL : single linkage clustering
l.m.f. : number of local minima found
c.f. : number of clusters found
f.e. : number of required function evaluations
u.s.t. : number of units of standard time

Table 2

Test functions (cf.[6])

GP	Goldstein & Price
BR	Branin (RCOS)
H3	Hartman 3
H6	Hartman 6
S5	Shekel 5
S7	Shekel 7
S10	Shekel 10

From Table 1, it is obvious that both methods performed almost equally well. The single linkage method, however, was marginally more accurate and performed on the average a little better with respect to both performance measures. A slightly improved implementation of this method was subsequently incorporated in the final algorithm.

3. A CONFIDENCE INTERVAL

In this section we describe how a range of confidence intervals for the global minimum y_* can be obtained from the sample.

We shall assume again that the total sample consists of N n-vectors $\underline{x}$ drawn independently from a uniform distribution over S. It is intuitively clear that some regularity condition on f is required to allow the derivation of any statement about the distribution of the resulting sample of function values. Such a condition can be conveniently formulated in terms of the *distribution function* F of f, with

$$F(y) = \Pr[f(\underline{x}) \leq y].$$

Note that the fact that $\underline{x}$ is drawn from a uniform distribution implies that

$$F(y) = \frac{m(L(y))}{m(S)}, \tag{4}$$

with m and L as defined in Section 2.

An interval $\underline{I}(p)$ is called a *level-p asymptotic confidence interval* for y_* if, for sufficiently large sample size,

$$\Pr[y_* \in \underline{I}(p)] = p.$$

Let $\underline{y}^{(1)}$ be the smallest function value from the original sample and $\underline{y}^{(2)}$ the smallest but one.

Theorem 1

If there exist positive constants ρ and K such that

$$\lim_{y \downarrow y_*} \frac{F(y)}{(y - y_*)^{1/\rho}} = K,$$

Then a level-p asymptotic confidence interval $\underline{I}(p)$ for y_* is given by

$$\underline{I}(p) = [\underline{y}^{(1)} - \frac{\underline{y}^{(2)} - \underline{y}^{(1)}}{p^{-\rho}-1}, \underline{y}^{(1)}].$$

Proof

This theorem is a special case of a result in [10]. An adapted proof is given in the Appendix. □

The following theorem can be invoked to show that the conditions of Theorem 1 hold under a mild assumption on f.

Theorem 2

If f assumes a unique global minimum at a point x_* lying in the interior of S, if f is twice differentiable in x_* and if the Hessian $H(x_*)$ is nonsingular, then the limit

$$\lim_{y \downarrow y_*} \frac{m(L(y))}{(y - y_*)^{n/2}}$$

exists and is a positive number.

Proof

See [1]. □

In view of (4), the above result implies that under the conditions of Theorem 2 a level-p asymptotic confidence interval $\underline{I}(p)$ for y_* is given by

$$\underline{I}(p) = [\underline{y}(p), \underline{y}^{(1)}]$$

with

$$\underline{y}(p) = \underline{y}^{(1)} - \frac{\underline{y}^{(2)} - \underline{y}^{(1)}}{p^{-2/n}-1}.$$

Note that $\underline{y}(p)$ is a monotone decreasing function of p, and that with probability 1

$$\lim_{p \uparrow 1} \underline{y}(p) = -\infty$$

as was to be expected.

Let us now consider the situation after the local phases of the algorithm, when clustering and local searches have yielded a number of promising local minima. The best of those will be the candidate global minimum, and since it has been obtained from the sample in an admittedly complicated but well defined way, we must consider this candidate global minimum value to be a random variable $\underline{y}^*$.

Obviously,

$$\Pr [\underline{y}^* \leq \underline{y}^{(1)}] = 1$$

and

$$\Pr [y_* \leq \underline{y}^*] = 1 \tag{5}$$

With probability 1, there will be a range of values for p such that $\underline{y}^*$ is smaller than $\underline{y}(p)$ and a range for which the opposite is true. In fact, the former situation occurs if p is smaller than the threshold value

$$\underline{p}_0 = \left(\frac{\underline{y}^{(2)} - \underline{y}^*}{\underline{y}^{(1)} - \underline{y}^*} \right)^{-n/2}$$

and the latter situation occurs if p is larger than $\underline{p}_0$.

In the first case, we simply have two probabilistic statements about y_*, one in the form of $\underline{I}(p)$ and one as in (5), with the latter dominating the former. In the second case, we know that

$$\Pr [\underline{y}(p) \leq y_* \leq \underline{y}^{(1)}] =$$

$$\Pr [\underline{y}(p) \leq y_* \leq \underline{y}^*] + \Pr [\underline{y}^* < y_* \leq \underline{y}^{(1)}] = p, \tag{6}$$

and since the second probability in (6) is 0 because of (5), we have a new level-p asymptotic confidence interval for y_*, given by $[\underline{y}(p), \underline{y}^*]$.

Thus, as a final outcome of the algorithm we obtain a candidate global minimum y^* together with an infinite range of level-p asymptotic confidence intervals (for $p > \underline{p}_0$) whose right end point is y^*. Note that for a meaningful interpretation of these confidence intervals by the user of the algorithm,

a specification of the units of measurement that f refers to is essential. In the case of the standard test functions for global optimization, such a specification is not immediately available; none the less, we do present a selection of the confidence intervals that were obtained for these test functions in Section 4.

4. COMPUTATIONAL EXPERIMENTS

The final version of the algorithm described in Sections 1, 2 and 3 has been the subject of extensive computational experiments. The test functions involved in these experiments were all those listed in Table 2.

In the tested version of the algorithm we used the single linkage clustering scheme with the threshold parameter α set equal to 0.01. Points were added to the sample in groups of 50, and the best 10 percent of the points in a current sample was retained for clustering and local search. We found that by increasing the group size to 100 an even more reliable method was obtained in the sense that every relevant minimum of the test functions was always found. This occurred, however, at the expense of a 50 percent increase in computing time.

The local search procedure used in our algorithm was the VA10AD variable metric routine from the Harwell Subroutine Library.
We also tested the SSVM variable metric routine developed by Van der Hoek [11], with almost identical results.

The algorithm was coded in FORTRAN and run on the DEC 20/50 of the Computer Institute Woudestein. Its performance was measured by the number of function evaluations required, as well as by the number of units of standard time as defined in Section 2.4. Both measurements are sensitive to the particularities of the sample at hand, and therefore the results reported actually represent the average outcome of four independent runs.

In Tables 3 and 4, we summarize these computational results, and compare them to those obtained for a few leading contenders as reported in [6].

Table 3

Number of function evaluations

Method \ Function	GP	BR	H3	H6	S5	S7	S10
Törn [14]	2499	1558	2584	3447	3679	3606	3874
De Biase [2]	378	597	732	806	620	788	1160
Price [13]	2500	1800	2400	7600	3800	4900	4400
Branin [9]	-*	-*	-*	-*	5500	5020	4860
New algorithm	398	235	235	462	567	624	755

*No results available

Table 4

Number of units standard time

Method \ Function	GP	BR	H3	H6	S5	S7	S10
Törn [14]	4	4	8	16	10	13	15
De Biase [2]	15	14	16	21	23	20	30
Price [13]	3	4	8	46	14	20	20
Branin [9]	-*	-*	-*	-*	9	8.5	9.5
New algorithm	1.5	1	1.7	4.3	3.5	4.5	7

*No results available

The results for our algorithm are obviously quite satisfactory; the algorithm also never failed to find the true global minimum. Yet, it remains difficult to arrange a fair and direct comparison of different stochastic global optimization procedures. Each of those methods has the property that the user's confidence in its final result can always be increased at the expense of increasing the sample [15]. A fair comparison of stochastic methods should therefore be based on a comparison of the costs involved in achieving a certain level of confidence. However, many algorithms do not provide any confidence information at all, and even our own attempts in that direction hardly

capture the effect on user's confidence of the elaborate local search calculations carried out to find the candidate value y*.

As remarked before, the confidence intervals produced by our algorithms are meaningless if nothing is known about the units of measurement in which the objective function is expressed. Nevertheless, we present the confidence intervals obtained for various values of p in Table 5. The values of p have been selected to be approximately equally distributed over the interval $[\max\{p_0, 0.75\}, 1.00]$. It seems reasonable to conclude that the intervals in Table 5 are on the whole not so excessively large as to preclude the applicability of this probabilistic information in a more practical context.

Table 5

Confidence intervals

	Minimum value	Pr_1	L_1	Pr_2	L_2	Pr_3	L_3	Pr_4	L_4	Pr_5	L_5
GP	3.000	0.79	42	0.83	58	0.88	85	0.92	135	0.96	296
BR	0.3978	0.79	0.75	0.83	1	0.88	1.5	0.92	2.5	0.96	5.5
H3	-3.862	0.79	2	0.83	3	0.88	4	0.92	6.5	0.96	13
H6	-3.322	0.79	2	0.83	3	0.88	4	0.92	7.5	0.96	15
S5	-10.15	0.975	2	0.98	5	0.985	10	0.99	20	0.995	50
S7	-10.40	0.975	2	0.98	5	0.985	10	0.99	20	0.995	50
S10	-10.53	0.94	2	0.95	5	0.96	10	0.98	20	0.99	50

Pr_j: the probability corresponding to the j - th confidence interval (j=1,..,5)

L_j : the lenth of the j - th confidence interval (j=1,...,5)

5. CONCLUDING REMARKS

The stochastic algorithm described in the previous sections has turned out to be a reliable and computationally attractive method for global optimization.

The main imperfection of the current design of this and all other stochastic methods is that there is no really satisfactory way to proceed if the final confidence statement is not acceptable to the user. Of course, it is always possible to enlarge the sample; asymptotically, the length of all confidence intervals will approach 0 for every $p < 1$ with probability 1. However,

if the sample is again drawn from a uniform distribution over S, all information that was obtained about the objective function so far is ignored. Ideally, the sampling distribution should be updated to reflect this information in a *Bayesian learning procedure*. Development of such a procedure, with due consideration for possibilities created in this way for an *interactive approach*, is very much a subject of future research.

A second part for future research will be the extension of this stochastic approach to *constrained global optimization*, other than through the use of penalty functions, and the extension to *combinatorial* (e.g., 0-1) *optimization*. It remains to be seen if a sampling procedure can provide a satisfactory starting point to solve these problems.

ACKNOWLEDGEMENTS

We gratefully acknowledge our useful discussions with L. de Biase, L. Brillinger, L. de Haan, G. van der Hoek and J. Koerts. This research was partially supported by NATO Special Research Grant 9.2.02 (SRG.7) and by the A.A. van Beek Fund in Rotterdam.

REFERENCES

[1] Archetti, F., Betro, B. and Steffe, S., 'A theoretical framework for global optimization via random sampling', Technical report, University of Pisa (1975).

[2] De Biase, L. and Frontini, F., 'A stochastic method for global optimization: its structure and numerical performance', p. 85-102 in [5] (1978).

[3] Dixon, L.C.W. and Szegö, G.P., *Towards global optimisation*, North Holland, Amsterdam (1975).

[4] Dixon, L.C.W., Gomulka, J. and Szegö, G.P., 'Towards global optimisation', p. 29-54 in [3] (1975).

[5] Dixon, L.C.W. and Szegö, G.P., *Towards global optimisation* 2, North Holland, Amsterdam (1978).

[6] Dixon, L.C.W. and Szegö, G.P., 'The global optimisation problem: an introduction', p. 1-18 in [5] (1978).

[7] Eberl, W. and Hafner, R., 'Die asymptotische Verteilung von Koinzidenzen', *Zeitschrift für Wahrscheinlichkeitsrechnung* 18, p. 322-332 (1971).

[8] Everitt, B., *Cluster analysis*, Heinemann, London (1974).

[9] Gomulka, J., 'Two implementations of Branin's method: numerical experience', p. 151-164 in [5] (1978).

[10] De Haan, L., 'Estimation of the minimum of a function using order statistics', Technical report Erasmus University Rotterdam (1979).

[11] Van der Hoek, G. and Dijkshoorn, M.W., 'A numerical comparison of self scaling variable metric algorithms', Technical report Erasmus University Rotterdam (1979).

[12] Kester, A., 'Asymptotic normality of the number of small distances between random points in a cube', *Stochastic Processes and their Applications* 3, p. 45-54 (1975).

[13] Price, W.L., 'A controlled random search procedure for global optimisation', p. 71-84 in [5] (1978).

[14] Törn, A.A., 'Probabilistic global optimization, a cluster analysis approach', p. 521-527 in: *Proceedings of the Second European Congress on Operations Research*, North Holland, Amsterdam (1976).

[15] Törn, A.A., 'Optimality by means of confidence', Technical report Åbo Swedish University School of Economics (1979).

APPENDIX

Proof of Theorem 1

The proof of Theorem 1 is based on the following lemma. Suppose that we have N independent random variables $\underline{Z}_1,\ldots,\underline{Z}_N$, that are all uniformly distributed over the interval [0,1]. Let $\underline{Z}^{(1)}$ be the smallest *order statistic* (the minimum) and $\underline{Z}^{(2)}$ the smallest but one.

Lemma

The statistic $\underline{Z}^{(1)}/\underline{Z}^{(2)}$ is uniformly distributed on the interval [0,1].

Proof

$$\Pr[\underline{Z}^{(1)}/\underline{Z}^{(2)} \leq z] = \Pr[\underline{Z}^{(1)} \leq z\underline{Z}^{(2)}]$$

$$= \int_0^1 \int_0^{zZ^{(2)}} N(N-1)(1-Z^{(2)})^{N-2}\, dZ^{(1)} dZ^{(2)}$$

$$= z. \qquad \square$$

We may assume the inverse of $F(y)$ to exist. The assumption of Theorem 1

$$\lim_{y \downarrow y_*} \frac{F(y)}{(y - y_*)^{1/\rho}} = K$$

can be transformed, by means of the substitution $u = y - m$, $v = F(m + u)$, into

$$\lim_{v \downarrow 0} \frac{v}{(F^{-1}(v) - y_*)^{1/\rho}} = K$$

As $K > 0$,

$$\lim_{v \downarrow 0} \frac{(F^{-1}(v) - y_*)^{1/\rho}}{v} = \frac{1}{K} \qquad (1)$$

To prove Theorem 1, we define the random variable $\underline{y}_i$ to be a transformation of the variable $\underline{Z}_i$, such that $\underline{y}_i$ has distribution function $F(y)$ $(i = 1, 2,\ldots, N)$. A transformation that satisfies this requirement is $\underline{y}_i = F^{-1}(\underline{Z}_i)$:

$$\Pr[\underline{y}_i \leq y] = \Pr[F^{-1}(\underline{Z}_i) \leq y]$$

$$= \Pr[\underline{Z}_i \leq F(y)]$$

$$= F(y) \qquad (i = 1, 2,\ldots, N)$$

As $F(t)$ is nondecreasing, $\underline{y}^{(1)} = F^{-1}(\underline{Z}^{(1)})$ is the smallest order statistic and $\underline{y}^{(2)} = F^{-1}(\underline{Z}^{(2)})$ is the smallest but one.
The random vector

$$(NK(\underline{y}^{(1)} - y_*)^{1/\rho},\ NK(\underline{y}^{(2)} - y_*)^{1/\rho})$$

is equal to

$$\left(K\frac{(\underline{y}^{(1)} - y_*)^{1/\rho}}{\underline{Z}^{(1)}}N\underline{Z}^{(1)},\ K\frac{(\underline{y}^{(2)} - y_*)^{1/\rho}}{\underline{Z}^{(2)}}N\underline{Z}^{(2)}\right).$$

Asymptotically, it has the same distribution as

$$(N\underline{Z}^{(1)},\ N\underline{Z}^{(2)})$$

because of (1).

The lemma implies that the distribution function of $\underline{Z}^{(1)}/\underline{Z}^{(2)}$ does not depend on N. Hence, for $N \to \infty$

$$\Pr\left[\frac{NK(\underline{y}^{(2)} - y_*)^{1/\rho} - NK(\underline{y}^{(1)} - y_*)^{1/\rho}}{NK(\underline{y}^{(1)} - y_*)^{1/\rho}} \geq x\right] =$$

$$\Pr\left[\frac{\underline{Z}^{(2)} - \underline{Z}^{(1)}}{\underline{Z}^{(1)}} \geq x\right] =$$

$$\Pr[\underline{Z}^{(2)}/\underline{Z}^{(1)} \geq x + 1] =$$

$$\Pr[\underline{Z}^{(1)}/\underline{Z}^{(2)} \leq \frac{1}{x+1}].$$

From the lemma, the last probability is equal to $\frac{1}{x+1}$.

Put $\frac{1}{x+1} = p$. Then $x = (1 - p)/p$, and hence asymptotically

$$\Pr\left[\frac{NK(\underline{y}^{(2)} - y_*)^{1/\rho} - NK(\underline{y}^{(1)} - y_*)^{1/\rho}}{NK(\underline{y}^{(1)} - y_*)^{1/\rho}} \geq \frac{(1 - p)}{p}\right] = p.$$

Algebraic manipulation yields

$$\Pr\left[y_* \geq \underline{y}^{(1)} - \frac{\underline{y}^{(2)} - \underline{y}^{(1)}}{p^{-\rho} - 1}\right] = p.$$

Since obviously

$$\Pr[y_* > \underline{y}^{(1)}] = 0,$$

the desired level-p asymptotic confidence interval follows immediately.

LIST OF REPORTS 1980

8000 "List of Reprints, nos 241-260, Abstracts of Reports Second Half 1979".

8001/0 "A Stochastic Method for Global Optimization", by C.G.E. Boender, A.H.G. Rinnooy Kan, L. Stougie and G.T. Timmer.