Mathematical Programming 39 (1987) 27-56
North-Holland

STOCHASTIC GLOBAL OPTIMIZATION METHODS PART I: CLUSTERING METHODS

A.H.G. RINNOOY KAN

Department of Industrial Engineering and Operations Research/Graduate School of Business Administration, University of California, Berkeley, CA, USA, and Econometric Institute, Erasmus University Rotterdam, The Netherlands

G.T. TIMMER

Econometric Institute, Erasmus University Rotterdam, The Netherlands, and ORTEC Consultants, Rotterdam; The Netherlands

Received 1 May 1985
Revised manuscript received 10 February 1987

In this stochastic approach to global optimization, clustering techniques are applied to identify local minima of a real valued objective function that are potentially global. Three different methods of this type are described; their accuracy and efficiency are analyzed in detail.

Key words: Global optimization, clustering, sampling methods.

1. Introduction

Let $f:\mathbb{R}^n \to \mathbb{R}$ be a continuous real valued objective function. Most nonlinear programming methods that have been developed aim for a *local optimum* (say, local minimum), i.e. a point x^* such that there exists a neighbourhood B of x^* with

$$f(x^*) \leqslant f(x) \quad \forall x \in B. \tag{1}$$

In general, however, several local optima may exist and the corresponding function values may differ substantially. The problem of designing algorithms that distinguish between these local optima and locate the best possible one is known as the *global optimization* problem, and forms the subject of this paper and its companion, Part II.

In the absence of reliable codes for the global optimization problem most problems are not modelled as such. Many problems, however, are of a global nature. This is especially true for many *technical design* problems (Dixon and Szegö, 1978b; Archetti and Frontini, 1978). *Economic applications*, where multimodal cost functions have to be minimized, have also been reported (Archetti and Frontini, 1978). Another global optimization problem often encountered in econometrics is that of locating the global maximum of a likelihood function. Thus, there is no need to dwell on the practical usefulness of quick and reliable methods to solve the global optimization problem.

The global optimization problem is to find the *global optimum* (say global minimum) x_* of a real valued objective function $f:\mathbb{R}^n \to \mathbb{R}$, i.e. to find a point $x_* \in \mathbb{R}^n$ such that

$$f(x_*) \leqslant f(x) \quad \forall x \in \mathbb{R}^n. \tag{2}$$

Unless stated otherwise, we will assume f to be twice continuously differentiable. For obvious computational reasons, one usually assumes that a set $S \subset \mathbb{R}^n$, which is convex, compact and contains the global minimum as an interior point, is specified in advance. None the less, the problem to find

$$y_* = \min_{x \in S} f(x) \tag{3}$$

remains essentially one of *unconstrained* optimization.

Any method for global optimization has to account for the fact that a numerial procedure can never produce more than approximate answers. Thus, the global optimization problem might be considered solved if, for some $\varepsilon > 0$, an element of one of the following sets has been identified (Dixon, 1978)

$$A_x(\varepsilon) = \{x \in S \mid \|x - x_*\| \leqslant \varepsilon\}, \tag{4}$$

$$A_f(\varepsilon) = \{x \in S \mid |f(x) - f(x_*)| \leqslant \varepsilon\}. \tag{5}$$

A disadvantage of the first mentioned possibility is that small perturbations in the problem data may have major effects on the location of x_* (Archetti and Betro, 1978a). A third possibility (Betro, 1981) is obtained by defining

$$\phi(y) = \frac{m(\{z \in S \mid f(z) \leqslant y\})}{m(S)}, \tag{6}$$

where $m(\cdot)$ is the *Lebesgue measure* and taking

$$A_\phi(\varepsilon) = \{x \in S \mid \phi(f(x)) \leqslant \varepsilon\}. \tag{7}$$

We note, however, that this set may contain points whose function values differ considerably from y_*.

A second problem, which is caused by the finite accuracy of numerical procedures, is that we cannot distinguish between two local minima which are very close to one another. If we define a *stationary point* of f as a point where the *gradient* $g:\mathbb{R}^n \to \mathbb{R}^n$ of f is equal to 0, then each (local) minimum is known to be a stationary point. We will assume that a positive constant ε can be specified, such that the distance between any two stationary points exceeds ε. Obviously, this implies that there can only be a finite number of stationary points in S.

Only few solution methods for global optimization have been developed so far; we refer to Dixon and Szegö (1978a, 1978b) and to Rinnooy Kan and Timmer (1984), Boender et al. (1985) for surveys. We shall be concerned with methods that incorporate *stochastic* elements. In most stochastic methods, two phases can be usefully distinguished. In the *global phase*, the function is evaluated in a number of randomly sampled points. In the *local phase*, the sample points are manipulated, e.g. by means of local searches, to yield a candidate global minimum.

Generally in turning to stochastic methods, we do sacrifice the possibility of an *absolute guarantee* of success. However, under mild conditions on the sampling distribution and on f, the probability that an element of $A_x(\varepsilon)$, $A_f(\varepsilon)$ or $A(\varepsilon)$ is sampled approaches 1 as the sample size increases (Solis and Wets, 1981). If the sample points are drawn from a *uniform* distribution over S and if f is continuous, then an even stronger result will turn out to hold: the sample point with lowest function value converges to the global minimum value with *probability* 1 (or almost surely). Thus, the global phase can yield an asymptotic guarantee with probability 1, and is therefore essential for the *reliability* of the method. However, a method that only contains a global phase will be found lacking in *efficiency*. To increase the latter while maintaining the former is one of the challenges in global optimization.

As in the case of deterministic methods, one of the questions in applying a stochastic method is when to stop. Preferably, a method of this nature should terminate with some probabilistic information on the quality of the proposed solution. Several approaches based on different assumptions about the properties of possible objective functions f and using different stochastic techniques have been proposed to design a proper *stopping rule*.

In Section 2, we review some stochastic methods and find that the most promising methods appear to be variants of the so-called *Multistart* technique where points are sampled iteratively from a *uniform distribution* over S (global phase), after which local minima will be found by applying a *local search procedure* to these points (local phase). A theoretical framework which enables the stochastic analysis of this method is developed in Boender (1984) (see also Boender and Rinnooy Kan, 1983, 1985). It turns out to be possible to develop *Bayesian estimates* of the number of local minima not yet identified and of the probability that the next local search will locate a new local minimum. By specifying the costs and the potential benefits of further experiments and weighing these against each other probabilistically, an *optimal Bayesian stopping rule* can be determined.

Multistart is still lacking in efficiency because the same local minimum may be located several times. If we define the *region of attraction* of a local minimum x^* to be the set of points in S starting from which a given local search procedure converges to x^*, then ideally, this local search procedure should be started exactly once in every region of attraction. Several new algorithms designed to satisfy this criterion are described in Section 3. The methods discussed in this section temporarily eliminate a prespecified fraction of the sample points whose function values are relatively high. The resulting *reduced sample* consists of groups of mutually relatively close points that correspond to the regions with relatively small function values. Within each group the points are still distributed according to the original uniform distribution. Thus, these groups can be identified by *clustering techniques* based upon tests on the uniform distribution. Only one local search procedure will be started in each group (Boender et al., 1980, 1982).

Unfortunately, the resulting groups do not necessarily correspond to the regions of attraction of f. It is possible that a certain group of points corresponds to a region with relatively small function values which contains several minima. Therefore, the methods which are based on the reduced sample may fail to find a local minimum although a point is sampled in its region of attraction. Methods that do not suffer from this deficiency will be dealt with in Part II of this paper (this issue, pp. 57–78). There we also discuss the computer implementation of the various global optimization methods and its theoretical properties, and we discuss the results of some computational experiments.

2. Multistart

The simplest stochastic method for global optimization consists only of a global phase. Known confusingly as Pure Random Search (Brooks, 1958; Anderssen, 1972), the method involves no more than a single step.

Pure Random Search

Step 1. Evaluate f in N points, drawn from a uniform distribution over S. The smallest function value found is the candidate solution for y_*.

In spite of its evident simplicity, Pure Random Search offers an *asymptotic guarantee* in a probabilistic sense. The proof is based on the simple observation that the probability that a uniform sample of size N contains at least one point in a subset $A \subset S$ is equal to (Brooks, 1958)

$$1 - \left(1 - \frac{m(A)}{m(S)}\right)^N . \tag{8}$$

Thus, any assumption of f guaranteeing that $m(A_f(\varepsilon))$, $m(A_x(\varepsilon))$ or $m(A_\phi(\varepsilon))$ is strictly positive will imply that Pure Random Search locates an element in the corresponding set with a probability approaching to 1 as N increases. In fact, if we let $\underline{y}_N^{(1)}$ be the smallest function value found in a sample of size N, then we can prove the following result.

Proposition 1 (Rubinstein, 1981; cf. Devroye, 1978). *If f is continuous, then $\underline{y}_N^{(1)}$ converges to the global minimum value y_* with probability* 1 *(or almost surely) with increasing N, i.e.*

$$\Pr[\lim_{N\to\infty} \underline{y}_N^{(1)} = y_*] = 1. \qquad \square \tag{9}$$

A similar guarantee will hold for all methods that follow.

The reader may well wonder to what extent an embarassingly simple method such as Pure Random Search has any advantage over an equally simplistic approach such as *Grid Search*, in which the function is evaluated in each point of a regular grid over S. The relative merits of these naive stochastic and deterministic strategies have been extensively analyzed (Sukharev, 1971; Ivanov, 1972; Anderssen and

Bloomfield, 1975; Archetti and Betro, 1978b). The net result of these analyses is that the points of the random sample cover S more efficiently (according to several probabilistic criteria) than the grid points do, at least if the dimension of the problem is not too low. In the studies mentioned above the methods are evaluated according to the distance between the global minimum and the sample or grid point closest to it. The advantage of Pure Random Search becomes more evident through an argument in Sobol (1982). Here, it is observed that for many functions some of the variables (in the n-dimensional space S) hardly affect the function value, in which case the distribution of the sample or grid points in the subspace defined by the remaining variables is of primary interest. However, it is not known in advance which of the variables are important and which are not. If the (uniform) sample points are projected into an arbitrary subspace, they still follow a uniform distribution over this subspace. However, if the grid points are projected into an arbitrary subspace, they may very well form groups of mutually close points, that cover the subspace in an unsatisfactory manner.

Nonetheless, Pure Random Search can hardly be taken seriously as a computational proposal. Several extensions of this method have been proposed that also start from a uniform sample over S (hence, Proposition 1 can be applied), but that at the same time involve local searches from some or all points in the sample. The simplest way to make use of a *local search procedure* P occurs in a folklore method known as Multistart.

Multistart

Step 1. Draw a point from a uniform distribution over S.

Step 2. Apply P to the new sample point.

Step 3. A termination criterion indicates whether to stop or to return to Step 1. The local minimum with smallest function value found is the candidate value for y_*.

Although this method is obviously more attractive than Pure Random Search, several inefficiencies still remain. However, let us first consider the question of a proper stopping rule for this method. Our treatment will be brief, since the details of our approach are reported elsewhere; it was initiated in Zielinski (1981) and extended in Boender (1984). It is based on a *Bayesian* estimate of the *number of local minima* W and the *relative size* of each *region of attraction* $\Theta_l = m(R_{x^*})/m(S)$, $l = 1, \ldots, W$, where R_{x^*} is the *region of attraction* of the local minimum x^*, i.e., the set of points in S starting from which P will converge to x^*. If the values of these parameters would be given, then is is possible to determine the possibility that W different local minima are found during N local searches (Boender, 1984).

This probability can be used in a *Bayesian* approach in which the unknowns $W, \Theta_1, \ldots, \Theta_W$ are assumed to be themselves random variables for which a *prior distribution* can be specified. Given the outcome of an application of Multistart, Bayes's rule is used to compute the *posterior distribution*, which incorporates both the prior beliefs and the sample information.

After lengthy calculations, surprisingly simple expressions emerge for the posterior distribution and posterior expectation of several interesting parameters, some of which are stated in the next theorem.

Proposition 2 (Boender, 1984). *If w different local minima have been found as the result of N local searches started in uniformly distributed points, if we assume a priori for the number of local minima $\underset{\sim}{W}$ that each integer of $[1, \infty)$ is equally probable, and if we assume that given $\underset{\sim}{W} = W$ the relative sizes of the regions of attraction $\underset{\sim}{\Theta}_1, \ldots, \underset{\sim}{\Theta}_W$ follow a uniform distribution on the $(W-1)$-dimensional unit simplex, then*

(i) *the posterior probability that there are K local minima is equal to*

$$\frac{(K-1)!K!(N-1)!(N-2)!}{(N+K-1)!(K-w)!w!(w-1)!(N-w-2)!}; \tag{10}$$

(ii) *the posterior expectation of the number of local minima is*

$$\frac{w(N-1)}{N-w-2} \tag{11}$$

provided that $N > w+2$);

(iii) *the posterior expected relative size of the non-observed regions of attraction is*

$$\frac{w(w+1)}{N(N-1)}. \quad \square \tag{12}$$

This theoretical framework is quite an attractive one, the more so since it can be easily extended to yield *optimal Bayesian stopping rules*. As in the previous section, such rules incorporate assumption about the costs and potential benefits of further experiments and weigh these against each other probabilistically to calculate the optimal stopping point. Several loss structures and corresponding stopping rules are described in Boender and Rinnooy Kan (1985).

Two observations conclude this Bayesian analysis. First, note that the posterior distributions and expectations do not depend on the dimension of the problem. The number of local searches that has to be performed (and hence the computational effort) only depends on the number of minima located. Second, the a priori assumption that every number of local minima is equally probable may appear to be very pessimistic. However, it appears that the prior assumption is rapidly dominated by the influence of sample information.

It is appropriate here to mention Betro (1981) and Snijman and Fatti (1985) in which related Bayesian analyses are described.

Nonetheless, in spite of the scope that Multistart offers for analysis, the procedure is still lacking in efficiency. The main reason for this is that it will inevitably cause each local minimum to be found several times. To avoid all these time consuming local searches, P should ideally be invoked no more than once in every region of attraction.

A first attempt to modify Multistart in this way can be found in Hartman (1973). In this method, a local search is started only when a point is drawn whose function value is less than the smallest local minimum value found so far. It should be obvious that under this rule the global minimum may not be found even if a point is sampled in R_{x^*}. A far more successful adaptation of Multistart is provided by the *clustering methods*, which form the main subject of this paper.

The basic idea behind the clustering methods is to start from a uniform sample from S, to create groups of mutually close points that correspond to relevant regions of attraction, and to start P once in every such region. Two ways to create such groups from the initial sample have been proposed. The first, called *reduction* (Becker and Lago, 1970) removes a certain fraction of the sample points with the highest function values. The second, called *concentration* (Törn, 1978), transforms the sample by allowing one or at most a few *steepest descent* steps from every point.

Clustering methods may fail in two different ways. Firstly, the resulting groups of points, or *clusters*, may contain several regions of attraction, so that the global minimum can be missed. Secondly, one region of attraction may be divided over several clusters, in which case the corresponding minimum will be located more than once.

These failures are especially likely to occur for methods based on concentration, where the sample is transformed in an unpredictable way. Better results are possible in the case of sample reduction, in which case some properties are known of the distribution of the resulting points. This makes it possible to identify the clusters more easily and to avoid the two above-mentioned possible failures to a certain extent. The methods based on reduction will be dealt with in the next section.

3. Clustering methods

In this section we aim for solution methods for this global optimization problem that satisfy the following—partially conflicting—demands. On the one hand the method must be *asymptotically correct*, i.e. if the method would be continued sufficiently long, then the smallest function value found during this process must converge to the global minimum value. (If the method is stochastic, convergence in a probabilistic sense will be required.) On the other hand, the method must be *efficient*: if the method is stopped after a reasonable amount of time, it should have produced results which compare favourably to the results obtainable by other methods in the same time period.

For reasons given in the previous section, the methods we will consider are variants of Multistart. The methods are iterative, and fit in the following framework.

Global framework

Step 1 (Global phase). N points are drawn from a uniform distribution over S. The function is evaluated in these points, and the points are added to the (initially empty) sample.

Step 2 (Local phase). A procedure selects a (possibly empty) subset of the enlarged sample, and a local search procedure P is applied to each of the elements of this subset. The stationary points, which are found during these local searches and which were not detected previously, are added to an (initially empty) set X^*.

Step 3. A stopping rule decides whether to return to return to Step 1 or to stop. If the method is stopped, then the element of X^* with smallest function value is the candidate solution.

In theory, N may be chosen to be any positive integer, including 1. However, in practice it may be more efficient to choose a somewhat larger value (e.g. 50 or 100).

In this section, we assume that the local search procedure P in Step 2 is *strictly descent*: starting from any point $x \in S$, it generates a sequence of points x_k, with

$$x_{k+1} = x_k + \alpha_k p_k \quad (\|p_k\| = 1, \alpha_k > 0), \tag{13}$$

which converges to a stationary point $\bar{x}$, such that moreover

$$f(x_k + \beta p_k) \leq f(x_k + \alpha p_k) \tag{14}$$

for all k and all α, β satisfying $0 \leq \alpha < \beta \leq \alpha_k$. Thus, there exists a path from x to $\bar{x}$ along which the function is nonincreasing.

We shall also assume that this path is completely contained in S. As a result, we can now derive some important properties of the regions of attraction of P.

Let

$$L(y) = \{x \in S \,|\, f(x) \leq y\} \tag{15}$$

be the (y-) *level set* of f, and let $L_x(y)$ (with $y \geq f(x)$) denote the (connected) *component* of $L(y)$ containing x. It is not difficult to see that both these sets contain all their accumulation points and are hence closed.

Lemma 3. *If, for any $x \in S$ and $y \geq f(x)$, a procedure which is strictly descent is started from a point in $L_x(y)$, then the sequence x_k generated by this procedure will converge to a stationary point $\bar{x}$ in $L_x(y)$.*

Proof. Since all points of the sequence x_k are located in S and since S is convex, the interval $[x_k, x_{k+1}]$ is completely contained in S, for every k. Suppose that there exists a k such that $x_k \in L_x(y)$ and $x_{k+1} \notin L_x(y)$. It follows that there exists an element $\bar{x}$ on the line segment $[x_k, x_{k+1}]$ with $f(\bar{x}) > y$. However, if $f(\bar{x}) > y$, then $f(\bar{x}) > f(x_k)$ which contradicts (14). We conclude that $x_k \in L_x(y)$ for every k. The sequence x_k must converge to a stationary point $\bar{x}$. Since $x_k \in L_x(y)$ for every k and $L_x(y)$ is closed, the result is immediate. □

For some local minimum x^*, let $\bar{y} \in \mathbb{R}$ be the smallest y for which $L_{x^*}(y)$ contains a stationary point other than x^*. (If there is no such y then $\bar{y}$ is the maximum of f over S). We define the *basin* B_{x^*} of x^* as the component of the interior of $L_{x^*}(y)$ containing x^*.

Hence, for any $x \in B_{x^*}$, $L_{x^*}(f(x))$ contains x^* as its only stationary point (as it is for B_{x^*}), so that the next theorem follows immediately from Lemma 3.

Theorem 4 (Dixon et al., 1975). *If a procedure that is strictly descent is started from any point x in B_{x^*}, then the sequence x_k generated by this procedure converges to x^*,* i.e. $B_{x^*} \subset R_{x^*}$ *for strictly descent procedures.* □

We note that, although condition (14) cannot be verified computationally a slight variation of it is more tractable. Let us define a procedure satisfying (13) to be *ε-descent* if the sequence converges to a stationary point; if $f(x_{k+1}) \leq f(x_k)$ for all k and moreover

$$f(x_k + i\varepsilon p_k) \leq f(x_k + (i-1)\varepsilon p_k) \quad \left(i = 1, 2, \ldots, \left[\frac{\alpha_k}{\varepsilon}\right]\right). \tag{16}$$

Results similar to Lemma 3 and Theorem 4 hold for ε-descent procedures.

Lemma 5. *If, for some $x \in S$ and $y \geq f(x)$, there is no point x_1 with $x_1 \in L(y)$, $x_1 \notin L_x(y)$ which is within ε-distance of an element $L_x(y)$ (i.e. $L_x(y)$ is not too close to another component of $L(y)$), then an ε-descent procedure started from a point in $L_x(y)$ will converge to a stationary point in $L_x(y)$.*

Proof. The proof is similar to the proof of Lemma 3, however, if $f(x_k) \in L_x(y)$ and $f(x_{k+1}) \notin L_x(y)$, then the line segment $[x_k, x_{k+1}]$ must contain an interval of length ε on which the function values exceed y, which contradicts (16). □

For any local minimum x^*, we define

$$B^{\varepsilon}_{x^*} = \{x \in B_{x^*} \mid \exists x_1 \text{ with } x_1 \notin B_{x^*} \text{ and } \|x - x_1\| \leq \varepsilon\},$$

$$y_\varepsilon = \inf_{x \in B^{\varepsilon}_{x^*}} f(x),$$

$$B_{x^*}(\varepsilon) = \{x \in B_{x^*} \mid f(x) < y_\varepsilon\},$$

i.e. $B_{x^*}(\varepsilon)$ is the subset of B_{x^*} which consist of points that have a smaller function value than any point in B_{x^*} that is within ε-distance of the boundary of B_{x^*}.

Theorem 6. *If a procedure that is ε-descent is started from any point x in $B_{x^*}(\varepsilon)$ then the sequence x_k generated by this procedure converges to x^*.*

Proof. If $B_{x^*}(\varepsilon)$ is empty, the theorem is clearly true. If $B_{x^*}(\varepsilon)$ is not empty, then it contains the local minimum x^*. For any k, suppose that $x_k \in B_{x^*}(\varepsilon)$ and $x_{k+1} \notin B_{x^*}(\varepsilon)$. If $x_{k+1} \in B_{x^*}$, this would imply $f(x_{k+1}) > f(x_k)$. However, $x_{k+1} \notin B_{x^*}$ will (as in the proof of Lemma 5) lead to a contradiction of (16). Hence, $x_k \in B_{x^*}(\varepsilon)$ for all k, which implies that the procedure will converge to x^*. □

As a final assumption, let us suppose that $\bar{x}$, the ultimate point of convergence for P, is actually a local minimum of f. This assumption seems to be an innocent one from an empirical point of view (Wolfe, 1969, 1971); if P should get stuck in a saddlepoint, we could always leave it after a suitable perturbation.

Let us now return to the global framework mentioned above and assume that P is strictly descent. If the subset selected in Step 2 equals the set of points which are added to the sample in Step 1, then the global framework reduces to Multistart. It is not efficient, however, to apply P to every sample point. Obviously, it would be preferable to apply P to a sample point if and only if this point is located in a region of attraction belonging to a minimum that has not yet been found. The set of minima found would then equal the set of minima found by Multistart, but it would probably be obtained at less costs. The purpose of our research has been to develop a method in which P is started exactly once in every region of attraction in which points have been sampled.

We will first examine the case in which P is only applied to sample points with relatively small function value. More precisely, we will consider procedures that start by temporarily removing a prespecified fraction $1-\gamma$ of the sample points $(0<\gamma<1)$, whose function values are relatively high. The remaining points form the *reduced sample.* If $y_k^{(i)}$ is the i-th smallest function value in a sample of size kN (obtained after k iterations of the global framework), then all elements of the reduced sample are elements of $L(y_k^{(\gamma kN)})=\{x\in S \,|\, f(x)\leqslant y_k^{(\gamma kN)}\}$. (Let us note here that, to facilitate the notation, we shall ignore various necessary integer round-ups and round-downs as in the case of γkN; they do not affect the analysis at all.) Again it is not very efficient to actually apply P to every reduced sample point, i.e. every point in $L(y_k^{(\gamma kN)})$. Instead, we will seek for methods in which P is started exactly once in every region of attraction which contains a reduced sample point. (Note that the probability that a region of attraction contains a reduced sample point depends on γ. In general very little can be said about an optimal choice of this parameter. Typically a value between 0, 1 and 0, 2 is chosen.)

We now examine the consequences of this approach for the stopping rule in Step 3, which decides whether or not the search for the global minimum has been sufficiently thorough. Recall that the Bayesian stopping rules described in Section 2 only depend on the number of points sampled and the number of minima that are obtained by starting local searches at these points. Therefore, they are not only applicable to Multistart, but to every method which, given a sample, results in the same set of minima as Multistart. In particular, these stopping rules are applicable to the methods in which exactly one local search is started in every region of attraction in which points have been sampled. Note that for these latter methods, the number N of proposition 2 still equals the number of sample points, but no longer equals the number of local searches.

For methods in which P is applied exactly once in every region of attraction containing at least one reduced sample point, the situation is more complicated. To analyze this situation, for any γ with $0<\gamma<1$, let $y_\gamma\in\mathbb{R}$ be such that

$$\phi(y_\gamma)=\frac{m[\{x\in S \,|\, f(x)\leqslant y_\gamma\})}{m(S)}=\gamma, \tag{17}$$

i.e., y_γ is the *γ-quantile* of f. Since ϕ is a monotonically increasing continuous function, there exists a unique value y_γ satisfying (17). If we would apply P to every sample point in $L(y_\gamma)$, then the Bayesian analysis, as described in Section 2, can still be applied. We can simply ignore the sample points whose function value exceeds y_γ, and apply the Bayesian analysis to the remaining points. Since the remaining points are still distributed according to a uniform distribution over $L(y_\gamma)$, the analysis can be adapted in a trivial way. (Note that the number N in proposition 2 now equals the number of sample point multiplied by γ.)

However, since we do not know y_γ in advance we cannot apply P to the sample points in $L(y_\gamma)$. Instead, we aim for methods in which P is applied to points in the level set $L(y_k^{(\gamma kN)})$, such that all minima whose regions of attraction contain a reduced sample point are found. Hence, the level above which the sample points are ignored depends on the sample. Therefore, the cell probabilities are no longer constant over time and the Bayesian analysis is formally no longer applicable. However, it is known that $y_k^{(\gamma kN)}$ does converge to y_γ with probability 1 (Bahadur, 1966). Hence, we may apply the adapted stopping rules as though $y_k^{(\gamma kN)}$ does not vary with k.

Unfortunately, we will not succeed entirely in our search for a method in which P is started exactly once in every region of attraction which contains a sample point, respectively a reduced sample point. In particular, we will not be able to exclude the possibility that P is not applied to a (reduced) sample point, although it would have led to a local minimum which has not yet been found. However, we apply the stopping rules as though this possibility does not exist, and will justify our use of these rules by showing that the probability that an error of the above type is made goes to 0 when the sample size increases.

Now, given kN sample points that have been drawn from a uniform distribution over S and given a set of stationary points X^*, we must determine a subset of the sample points to which P will be applied. To do so, we will use the reduced sample to estimate the components of $L(y_k^{(\gamma kN)})$. A local search is then started once in each component that does not contain an element of X^*. The rationale of this approach is that if P is applied to an element of a component of $L(y_k^{(\gamma kN)})$, then P is known to converge to a local minimum in that component (see Lemma 3).

How can the components of $L(y_k^{(\gamma kN)})$ be identified? Intuitively, as a result of the removal of a fraction of the points with higher function values, groups of points that are relatively close to each other are created, each of which corresponds to such a component. The natural way to identify these groups (and through them, the components) is to make use of *cluster analysis*. However, there are several reasons not be use the ordinary clustering methods. The main reason is that we have more information about the problem than just the location and the function value of the reduced sample points. This extra information includes the fact that the reduced sample points are known to be a subset of a uniform sample and the fact that the groups searched for correspond to the components of a level set of a continuously differentiable function. Since this information cannot be translated into measurable characteristics of the reduced sample points, it must be ignored or used in a different way.

The methods which we will describe shortly can be viewed as standard clustering techniques which have been adapted to our specific problem, and all fit in the following framework. The clusters are created one by one, and each cluster is initiated by a *seed point*. Selected points of the reduced sample are added to the cluster until a *termination criterion* is satisfied. Under conditions to be specified, the local search procedure is started from a point in the cluster.

In the next three subsections, we will describe three methods that fit in this framework, but differ in the rule by which they select the points that are added to the cluster and in the corresponding termination criterion. This difference is mainly due to the different ways in which the methods exploit the fact that the reduced sample is known to be a subset of the original sample of uniformly distributed points. The first two of these methods already appeared in an earlier article (Boender et al., 1982) and hence will be described only briefly. The analysis of their properties, however, is new.

3.1. Density clustering

Analogously to Törn (1976), we will let the clusters in this approach correspond to the reduced sample points in a subset T_i, $i = 0, 1, 2, \ldots$, of S of stepwise increasing volume, where T_0 equals the seed point of the cluster and $T_{i+1} \supset T_i$, $i = 1, 2, \ldots$. A cluster is terminated if in a step no points are added to the cluster. However, we will adjust Törn's method in three ways; the choice of the seed points, the shape of the sets T_i and the increase in volume of these sets in each step.

We first turn to the choice of the seed points. As we will see later, it is advantageous to choose a local minimum as the seed point. Therefore, the local minima in X^* are first used as seed points. If all local minima known have been used as a seed point already, and there are still reduced sample points that have to be clustered, then a local search is started in the unclustered reduced sample point $\bar{x}$ with smallest function value. If the resulting local minimum x^* was already known, then $\bar{x}$ is assigned to the cluster that was initiated by x^*, and again a local search is started from the unclustered reduced sample point with smallest function value. If the resulting local minimum was not yet known, then it is chosen as the next seed point.

Let us now consider the shape of the sets T_i. Recall, that the cluster is initiated by a local minimum x^* and that it should correspond to $L_{x^*}(y_k^{(\gamma k N)})$. This suggests to let T_i correspond to $L_{x^*}(y)$ for stepwise increasing values of y. The actual sets $L_{x^*}(y)$ may be hard to construct, but since f is twice continuously differentiable, we can approximate these sets by the level sets $\tilde{L}(y)$ around x^* that are defined by the second order approximation $\tilde{f}$ of f around x^*:

$$\tilde{f}(x) = f(x^*) + \tfrac{1}{2}(x - x^*)^{\mathrm{T}} H(x^*)(x - x^*). \tag{18}$$

Hence, in step i we let T_i be the set $\{x \in S \,|\, (x - x^*)^{\mathrm{T}} H(x^*)(x - x^*) \leq r_i^2\}$, for some r_i to be determined below, with $r_{i+1} > r_i$, $i = 1, 2, \ldots$. (An approximation of $H(x^*)$ may be obtained, for example, as a byproduct of a quasi-Newton local search procedure.)

Finally, we derive the rate at which r_i should increase with i so as to ensure proper termination of a cluster. The probability that the cluster is terminated in step

i, equals the probability that the set $A_i = \{x \in S \mid x \in T_i, x \notin T_{i-1}\}$ does not contain any reduced sample points. To determine this probability for the case that there are still unclustered reduced sample points in $L_{x^*}(y_k^{(\gamma kN)})$, i.e. the probability of erroneous termination, we assume that the sets $L_{x^*}(y)$, with $f(x^*) \leq y \leq y_k^{(\gamma kN)}$ can be properly approximated by ellipsoids, so that $T_i \subset L_{x^*}(y_k^{(\gamma kN)})$. Given this assumption, the probability of erroneous termination in step i, say α_k, equals the probability that none of the kN original sample points is located in A_i. Using (8) it follows that

$$\alpha_k = (1 - m(A_i)/m(S))^{kN}. \tag{19}$$

Let us choose $m(A_i)$, and hence r_i, such that the probability α_k that the cluster is terminated incorrectly in step i, decreases with increasing k. To achieve this $kN \cdot m(A_i)$ must increase with k. However, we also must avoid unnecessary local searches, which implies that $m(A_i)$ should be as small as possible. Therefore, for some $\sigma > 0$, we choose $m(A_i) = (m(S)\sigma \log kN)/kN$, so that

$$\alpha_k = \left(1 - \frac{\sigma \log kN}{kN}\right)^{kN}. \tag{20}$$

It is not hard to verify that, for some constants $c_1, c_2 > 0$, we have that

$$c_1 k^{-\sigma} \leq \left(1 - \frac{\sigma \log k}{k}\right)^k \leq c_2 k^{-\sigma} \tag{21}$$

for all k. Hence, if we terminate the cluster in step i if no unclustered reduced sample point exists in T_i with (Boender et al., 1982)

$$r_i = \pi^{-1/2}\left(i\Gamma\left(1+\frac{n}{2}\right)(\det H(x^*))^{1/2} m(S)\frac{\sigma \log kN}{kN}\right)^{1/n}, \tag{22}$$

then the probability that the cluster is terminated incorrectly in step i, decreases polynomially fast with increasing k.

A stepwise description of this method follows. Let w be the number of local minima x^* with $f(x^*) \leq y_k^{(\gamma kN)}$, which are known at the start of the procedure.

Density Clustering

Step 1 (Determine reduced sample). Determine the reduced sample by taking the γkN points with the smallest function values. Set $j := 1$.

Step 2 (Determine seed points). Set $i := 1$. If all reduced sample points have been assigned to a cluster, stop.

If $j \leq w$, then choose the j-th local minimum in X^* as the next seed point. If $j > w$, then apply P to the unclustered reduced sample point $\bar{x}$ with the smallest function value. If the resulting local minimum x^* is an element of X^*, then assign $\bar{x}$ to the cluster initiated by x^* and repeat step 2. If $x^* \notin X^*$, then add x^* to X^*, set $w := w + 1$ and let x^* be the next seed point.

Step 3 (Form cluster). Add all unclustered reduced sample points which are within distance r_i of the seed point x^* to the cluster initiated by x^*. If no point has been added to the cluster for this specific value of r_i, then set $j := j + 1$ and go to Step 2, else set $i := i + 1$ and repeat Step 3.

Unfortunately, if the set $L_{x^*}(y_k^{(\gamma kN)})$ differs substantially from an ellipsoid, then this influences both the probability that the cluster is terminated incorrectly and the probability that the cluster is expanded incorrectly, in an unpredictable way. To arrive at a satisfactory clustering method, it is necessary that the shape of the resulting clusters is not fixed. Intuitively the shape of the clusters should converge to the shape of the actual sets $L_{x^*}(y_k^{(\gamma kN)})$ with increasing k. A method which satisfies this property is presented in the next subsection.

3.2. *Single linkage clustering*

In the adapted Single Linkage method, the clusters are formed sequentially, and each cluster is again initiated by a seed point. After a cluster C is initiated, we find an unclustered point x such that

$$d(x, C) = \min_{x_1 \in C} \|x - x_1\| \tag{23}$$

is minimal. We add x to C and repeat until $d(x, C)$ exceeds the critical distance r_k.

Early implementations of Single Linkage and Density Clustering were the subject of limited computational experiments. These experiments showed that Single Linkage indeed approximates the sets $L_{x^*}(y_k^{(\gamma kN)})$ more accurately than Density Clustering. However, to prove rigorously the superiority of Single Linkage it turns out that we must slightly adjust the rule according to which the seed points are selected. The reason is that it will turn out to be difficult to analyze Single Linkage in the regions near the boundary of S and in neighbourhoods of the elements of X^*. Therefore, we will define the procedure so as never to start a local search in these regions. This may imply that no local search is started from any point in a certain cluster. This, however, is not a serious drawback if we first redefine S in a slightly different way. For some $\tau > 0$, we let Q_τ be the set of points in S that are within distance τ of a point on the boundary of S, and we let S_τ be the set of points in S which are not within distance τ of a point on the boundary of S, so that $S_\tau = S \backslash Q_\tau$. We assume that all local minima of f occur in the interior of S_τ.

We will also have to give special treatment to the neighbourhoods of the elements of X^*. For some fixed and small υ, let X_υ^* be the set $\{x \in S \mid \|x - \bar{x}\| < \upsilon$, for any $\bar{x} \in X^*\}$. Recall that we assumed that a positive constant ε can be specified, such that the distance between any two stationary points exceeds ε. Hence, we can choose υ such that the distance between any two stationary points exceeds 2υ. We will now give a stepwise description of the adjusted Single Linkage procedure.

Single Linkage

Step 1 (Determine reduced sample). Determine the reduced sample by taking the γkN sample points with the smallest function values. Let X^1 be the set of minima in X^*, let w be the number of elements of X^1, and set $j := 1$.

Step 2 (Determine seed points). If all reduced sample points have been assigned to a cluster, stop.

If $j \leqslant w$, then choose the j-th local minimum in X^1 as the next seed point; go to Step 3.

Otherwise, determine the point $\bar{x}$ which has the smallest function value among the unclustered reduced sample points; $\bar{x}$ is the next seed point. If $\bar{x} \notin Q_\tau^*$ and if $\bar{x} \notin X_\upsilon^*$, then apply P to $\bar{x}$ to find a local minimum x^*; add new stationary points encountered during this search (possibly including x^*) to X^*, and adjust X^1 and w if necessary.

Step 3 (Form cluster). Initiate a cluster by the seed point which is determined in Step 2. Add reduced sample points which are within distance r_k of a point already in the cluster to the cluster, until no more such points exist. Set $j := j+1$, and go to Step 2.

Let us now analyze Single Linkage, and determine an appropriate value for the critical distance r_k. This critical distance will be chosen to depend on kN only so as to minimize the probabilities of two possible failures of the method: the probability that a local search is started, although the resulting minimum is known already, and the probability that no local search is started in a component of $L(y_k^{(\gamma kN)})$ which contains reduced sample points.

Let us first consider the probability that P is applied incorrectly to some reduced sample point. For a suitable choice of r_k, we will prove that the probability that a local search is started, let alone started incorrectly, tends to 0 with increasing k. For this purpose we divide S into three subsets. Let Y_υ be the set of elements in S that are within distance υ of a stationary point of f. (Note that $X_\upsilon^* \subset Y_\upsilon$, since the definition of X_υ^* only involves stationary points already detected.) We already defined Q_τ to be the set of elements in S that are within distance τ of a point on the boundary of S. Finally we let $M_{\tau,\upsilon}$ consist of the elements in S that do not belong to Q_τ or Y_υ. Note that we defined Q_τ and Y_υ as open sets, so that $M_{\tau,\upsilon}$ is closed and therefore compact. We will start our analysis by considering the elements of $M_{\tau,\upsilon}$; the large majority of the reduced sample points belongs to this set. We wish to determine the probability that P is applied to a reduced sample point $\underline{x}$ with $\underline{x} = a \in M_{\tau,\upsilon}$. Let $B_{a,r}$ be the set $\{x \in S \mid \|x-a\| \leqslant r\}$. Suppose that B_{a,r_k} contains a sample point z with $f(z) < f(a)$. Clearly, z then belongs to the reduced sample, and if z is assigned to a cluster then a will be assigned to that cluster too. Moreover, it is easy to check that we will not apply the local search procedure to a before z has been assigned to a cluster. Thus, the probability that a local search is started in a reduced sample point $\underline{x} = a \in M_{\tau,\upsilon}$, is certainly smaller than the probability that there is no sample point $\underline{z}$ in B_{a,r_k} with $f(\underline{z}) < f(a)$. To calculate this latter probability, we need the following lemma.

Lemma 7. *For any $\tau > 0$ and $\upsilon > 0$, let a be an element of $M_{\tau,\upsilon}$, let $B_{a,r} = \{x \in S \mid \|x-a\| \leqslant r\}$, and let $A_{a,r} = \{x \in S \mid \|x-a\| \leqslant r$ and $f(x) < f(a)\}$. Then, uniformly in a,*

$$\lim_{r \downarrow 0} \frac{m(A_{a,r})}{m(B_{a,r})} \geqslant \tfrac{1}{2}. \tag{24}$$

Proof. Consider the set

$$D_{a,r}=\{x\in S\,|\,\|x-a\|\leqslant r \text{ and } g(a)^{\mathrm{T}}(x-a)+\tfrac{1}{2}cr^2<0\}, \tag{25}$$

where c is a positive constant which is greater than the supremum over S of the eigenvalues of $H(x)$. From the Taylor expansion of f around a, we know that for all $x\in S$, with $\|x-a\|\leqslant r$, there exists a $\theta, 0\leqslant\theta\leqslant 1$, such that

$$\begin{aligned} f(x)-f(a) &= g(a)^{\mathrm{T}}(x-a)+\tfrac{1}{2}(x-a)^{\mathrm{T}}H(a+\theta(x-a))(x-a) \\ &\leqslant g(a)^{\mathrm{T}}(x-a)+\tfrac{1}{2}c(x-a)^{\mathrm{T}}(x-a) \\ &\leqslant g(a)^{\mathrm{T}}(x-a)+\tfrac{1}{2}cr^2. \end{aligned} \tag{26}$$

Hence, if $x\in D_{a,r}$, then $x\in A_{a,r}$. Thus, we have proved that $D_{a,r}\subset A_{a,r}$.

Now consider an orthogonal matrix U (so that $U^{\mathrm{T}}U=UU^{\mathrm{T}}=I$), for which

$$U^{\mathrm{T}}e_1=\frac{1}{\|g(a)\|}g(a), \tag{27}$$

where e_1^{T} is the n-dimensional vector $(1,0,\ldots,0)$. Obviously, such a matrix U always exists, because condition (27) only fixes the first row of U to be equal to $g(a)/\|g(a)\|$, the norm of which is 1.

We can now rewrite the set $D_{a,r}$ as follows

$$\begin{aligned} D_{a,r} &= \{x\in S\,|\,(x-a)^{\mathrm{T}}U^{\mathrm{T}}U(x-a)\leqslant r^2 \text{ and } g(a)^{\mathrm{T}}U^{\mathrm{T}}U(x-a)+\tfrac{1}{2}r^2c<0\} \\ &= \{x\in S\,|\,(U(x-a))^{\mathrm{T}}U(x-a)\leqslant r^2 \text{ and } (Ug(a))^{\mathrm{T}}U(x-a)+\tfrac{1}{2}r^2c<0\} \\ &= \{x\in S\,|\,(U(x-a))^{\mathrm{T}}U(x-a)\leqslant r^2 \text{ and } \|g(a)\|e_1^{\mathrm{T}}U(x-a)+\tfrac{1}{2}r^2c<0\}. \end{aligned} \tag{28}$$

Hence, the matrix U defines a 1-1 correspondence between the elements of $D_{a,r}$ and the elements of

$$G_{a,r}=\{z\in\mathbb{R}^n\,|\,\|z\|\leqslant r \text{ and } \|g(a)\|e_1^{\mathrm{T}}z+\tfrac{1}{2}r^2c<0\}. \tag{29}$$

Note that for r sufficiently small, all points $x\in\mathbb{R}^n$ satisfying $\|x-a\|\leqslant r$ are contained in S, because it follows from the definition of $M_{\tau,\upsilon}$ that this is certainly true if $r<\tau$. The transformation defined by U does not change the distances between points and the angles between vectors, because, for every $x_1, x_2\in\mathbb{R}_n$,

$$x_1^{\mathrm{T}}x_2=x_1^{\mathrm{T}}U^{\mathrm{T}}Ux_2=(Ux_1)^{\mathrm{T}}Ux_2=z_1^{\mathrm{T}}z_2.$$

Moreover, $m(D_{a,r})=m(G_{a,r})$, since the determinant of U is 1. Since, for $r<\tau$, $B_{a,r}=\{x\in\mathbb{R}^n\,|\,\|x-a\|\leqslant r\}$, we obtain that $m(B_{a,r})=r^n\pi^{n/2}/\Gamma(1+(n/2))$. Since f is continuously differentiable, and $M_{\tau,\upsilon}$ is compact, the minimum of $\|g(x)\|$ over $M_{\tau,\upsilon}$ exists. Let p be this minimum; p cannot be zero because $M_{\tau,\upsilon}$ does not contain any stationary point, so that $p>0$. Hence, if the first coordinate of z, say $z^{(1)}$, is smaller than $-\tfrac{1}{2}r^2c/p$, and if $\|z\|\leqslant r$, then $z\in G_{a,r}$. Since the intersection of $\{z\in\mathbb{R}^n\,|\,\|z\|\leqslant r\}$

with the hyperplane $z^{(1)}=0$ is an $(n-1)$ dimensional hyperball with measure $\pi^{(n-1)/2}r^{n-1}/\Gamma(1+(n-1)/2)$, it follows that

$$m(G_{a,r}) \geqslant \tfrac{1}{2}\cdot\frac{r^n\pi^{n/2}}{\Gamma\left(1+\frac{n}{2}\right)}-\frac{r^{n-1}\pi^{(n-1)/2}}{\Gamma\left(1+\frac{n-1}{2}\right)}\cdot\frac{cr^2}{2p} \tag{30}$$

Thus,

$$\lim_{r\downarrow 0}\frac{m(A_{a,r})}{m(B_{a,r})} \geqslant \lim_{r\downarrow 0}\frac{m(D_{a,r})}{m(B_{a,r})} = \lim_{r\downarrow 0}\frac{m(G_{a,r})}{m(B_{a,r})}$$

$$\geqslant \lim_{r\downarrow 0}\tfrac{1}{2}-\frac{cr^{n+1}\pi^{(n-1)/2}}{2p\Gamma\left(1+\frac{n-1}{2}\right)}\cdot\frac{\Gamma\left(1+\frac{n}{2}\right)}{r^n\pi^{n/2}}=\tfrac{1}{2}. \tag{31}$$

Since the above reasoning is independent of the choice of $a\in M_{\tau,\upsilon}$ the result is now immediate. □

Actually, $|m(A_{a,r})/m(B_{a,r})-\frac{1}{2}|$ can be shown to be $O(r/\|g(a)\|)$, so that the limit considered in (24) is precisely equal to $\frac{1}{2}$.

Lemma 7 is valid for any positive τ and positive υ, so that we can choose these numbers as small as we like. Note, that if $\upsilon=0$, but a is not a stationary point, then the limit in (24) still equals $\frac{1}{2}$. However, the convergence is not uniform in a, because $\|g(a)\|$ can become arbitrarily small.

Let us now return to the probability that, for some reduced sample point $\underline{x}$, with $\underline{x}=a\in M_{\tau,\upsilon}$, there exists a sample point $\underline{z}$ in B_{a,r_k} with $f(\underline{z})<f(a)$, which bounds the probability that a local search is started in a. To calculate this probability, let us first consider the simpler case where $\underline{x}$ is an arbitrary sample point, with $\underline{x}=a\in M_{\tau,\upsilon}$. The remaining $kN-1$ sample points are still distributed according to a uniform distribution over S and hence, the probability that none of these $kN-1$ uniform points is in A_{a,r_k}, i.e. is within distance r_k of a and has a smaller function value than a, equals (cf. (8))

$$(1-m(A_{a,r_k})/m(S))^{kN-1}. \tag{32}$$

Moreover, provided that r_k tends to 0 with increasing k, we know from Lemma 7 that, for any β with $0<\beta<\frac{1}{2}$, there exists a k_0 such that, for $k>k_0$,

$$\frac{m(A_{a,r_k})}{m(B_{a,r_k})}\geqslant\beta. \tag{33}$$

Hence, for any sample point $\underline{x}=a\in M_{\tau,\upsilon}$, the probability that there is no sample point $\underline{z}$ in B_{a,r_k} with $f(\underline{z})<f(a)$ is smaller than

$$(1-\beta m(B_{a,r_k})/m(S))^{kN-1} \tag{34}$$

(for sufficiently large k).

Analogously to Subsection 3.1, we can choose r_k in such a way that the probability (34) is constant or decreases with k. For instance, for some $\sigma > 0$, we can choose

$$r_k = \pi^{-1/2}\left(\Gamma\left(1+\frac{n}{2}\right) m(S) \frac{\sigma \log kN}{kN}\right)^{1/n}, \tag{35}$$

so that, for k large enough, $m(A_{a,r_k}) \geq (\beta\sigma \log kN)/kN$. Hence, for this specific choice of r_k we proved that the probability that for some sample point $\underline{x} = a \in M_{\tau,\upsilon}$, there is no sample point $\underline{z}$ in B_{a,r_k} with $f(\underline{z}) < f(a)$ is $\mathrm{O}(k^{-\beta\sigma})$ (note that we can omit N in all $\mathrm{O}(\cdot)$ terms since N is a constant). Since the number of sample points in iteration k is kN, we may conclude that the probabilty that there exists a sample point in $M_{\tau,\upsilon}$ which has no other sample point within distance r_k with smaller function value is $kN\ \mathrm{O}(k^{-\beta\sigma})$ or $\mathrm{O}(k^{1\ \beta\sigma})$. It follows that the probability that there exists a reduced sample point in $M_{\tau,\upsilon}$ which has no other sample point within distance r_k with smaller function value must also be $\mathrm{O}(k^{1-\beta\sigma})$. Hence, we proved that, for any $\beta < \frac{1}{2}$, the probability that a local search is started from any element of $M_{\tau,\upsilon}$ is $\mathrm{O}(k^{1-\beta\sigma})$. Obviously, if $\sigma > 2$, then we can choose $1/\sigma < \beta < \frac{1}{2}$, so that the probability that a local search is started from any element of $M_{\tau,\upsilon}$ in iteration k tends to 0 with increasing k.

Moreover, if we let $\underline{\xi}_k$ be the number of local searches started from points in $M_{\tau,\upsilon}$ in iteration k, and if we choose $\sigma > 4$, then it is easy to show that

$$\sum_{k=1}^{\infty} \Pr[\underline{\xi}_k > 0] < \infty. \tag{36}$$

Hence, it follows from the Borel-Cantelli Lemma that even if the sampling and clustering continues for ever, then the total number of local searches ever started in $M_{\tau,\upsilon}$ is finite with probability 1.

We now turn to the probability that a local search is started in Q_τ and Y_υ. It follows from the description of Single Linkage that no local search will ever be started in Q_τ. To analyze the situation in Y_υ we need one more assumption: we assume that if we apply P to a point which is within distance υ of a stationary point $\bar{x}$, then we will recognize $\bar{x}$ as such and add it to X^* (if necessary). Because we can choose υ as small as we want, this assumption is reasonable. Hence, we start P at most once in the neighbourhood of any stationary point. Since the number of stationary points is finite, we may conclude that the probability that P is applied to a point in Y_υ tends to 0 with increasing k.

Thus, we proved the following theorem.

Theorem 8. *If the critical distance r_k of Single Linkage is determined by* (35) *with $\sigma > 2$, then the probability that a local search is started by Single Linkage in iteration k tends to* 0 *with increasing k. If $\sigma > 4$, then, even if the sampling continues for ever, the total number of local searches ever started by Single Linkage is finite with probability* 1. □

We will now consider the second possible failure of Single Linkage, i.e. the possibility that no local minimum is found in a component of $L(y_k^{(\gamma kN)})$, although this component contains a sample point. We shall prove that with a probability increasing to 1, such a failure will not occur. In analyzing this probability, we again encounter the difficulty that the components of $L(y_k^{(\gamma kN)})$ depend on the specific value of the random variable $\underline{y}_k^{(\gamma kN)}$. As before, we therefore focus on the components of $L(y_\gamma)$.

To examine the probability that no local minimum is found in a component of $L(y_\gamma)$, say $L_a(y_\gamma)$, although a sample point is contained in $L_a(y_\gamma)$ we will first prove some general results. Roughly speaking, these results will show that if $x_1 \in L_a(y_\gamma)$ and $x_2 \in L(y_\gamma)\backslash L_a(y_\gamma)$, then the components of $L(y_k^{(\gamma kN)})$ containing x_1 and x_2 respectively are sufficiently far apart. The fact that those components have been defined to be closed sets will play an important role in the proofs. For any $y \in \mathbb{R}$ and any $a \in L(y)$, let $V_a(y)$ be the set of elements in $L(y)$ which are not contained in $L_a(y)$. Furthermore, let the distance between two subsets E_1 and E_2 of the $\mathbb{R}^n$, say $d(E_1, E_2)$, be defined as the infimum of the distances between any element of E_1 and any element of E_2.

Lemma 9. *For all* $y \in \mathbb{R}$, *there exists a* $\delta > 0$ *such that, for all* $a \in L(y)$, $d(L_a(y), V_a(y)) \geq \delta$.

Proof. Since every component of $L(y)$ contains a stationary point (see Lemma 3), $L(y)$ only consists of a finite number of components. Hence, $V_a(y)$ is the union of a finite number of components and is therefore closed. It follows from the definition of $L_a(y)$ and $V_a(y)$ that $L_a(y) \cap V_a(y) = \emptyset$. If $V_a(y) = \emptyset$, then the theorem is trivially true, so that we may assume that $V_a(y) \neq \emptyset$.

The remaining part of the proof is by contradiction; suppose that $d(L_a(y), V_a(y)) = 0$. Then there exist a sequence α_i in $L_a(y)$ and a sequence β_i in $V_a(y)$ with $\|\alpha_i - \beta_i\| < 1/i$, $i = 1, 2, \ldots$. Since both $L_a(y)$ and $V_a(y)$ are bounded (by S), both sequences contain a convergent subsequence, $\alpha_{i(j)}$ and $\beta_{i(j)}$, such that $\|\alpha_{i(j)} - \beta_{i(j)}\| \leq 1/i(j)$, for every positive integer j, and

$$\lim_{j\to\infty} \alpha_{i(j)} = \alpha, \lim_{j\to\infty} \beta_{i(j)} = \beta. \tag{37}$$

Since both $L_a(y)$ and $V_a(y)$ are closed, we have that $\alpha \in L_a(y)$ and $\beta \in V_a(y)$ and $\|\alpha - \beta\| = 0$. This, however, contradicts $L_a(y) \cap V_a(y) = \emptyset$. Thus, there exist a δ_a such that $d(L_a(y), V_a(y)) > \delta_a$. Obviously, δ_a is equal for all a that belong to the same component of $L(y)$. Since $L(y)$ only consists of a finite number of components, we can choose δ independent of a, which completes the proof. □

Lemma 10. *There exists an* $\varepsilon > 0$ *and a* $\delta > 0$ *such that for any* $y \leq y_\gamma + \varepsilon$, *for any* $a \in L(y_\gamma)$ *and for any minimum* $x^* \in L(y)$ *which does not belong to* $L_a(y_\gamma)$, *we have that* $d(L_a(y), L_{x^*}(y)) \geq \delta$.

Proof. If $y = y_\gamma$, then the reslt follows immediately from Lemma 9, since $L_{x^*}(y) \subset V_a(y)$.

Now suppose that $y < y_\gamma$. It follows from the definition of a component that $L_a(y) \subset L_a(y_\gamma)$ and that $L_{x^*}(y) \subset L_{x^*}(y_\gamma)$. Hence, because of Lemma 9 there exists a $\delta_1 > 0$ such that

$$d(L_a(y), L_{x^*}(y)) \geq d(L_a(y_\gamma), L_{x^*}(y_\gamma)) \geq d(L_a(y_\gamma), V_a(y_\gamma)) \geq \delta_1 . \tag{38}$$

The interesting case arises when $y > y_\gamma$. Since f only has a finite number of stationary points, there exists an ε_1 such that $L(y_\gamma + \varepsilon_1)$ contains no more stationary points than $L(y_\gamma)$. Hence, we may assume that $x^* \in L(y_\gamma)$. Suppose that a and x^* belong to the same component of $L(y_\gamma + \varepsilon)$ for any $\varepsilon > 0$. Then, they also belong to the same component of $A_i = \{x \in S \,|\, f(x) < y_\gamma + 1/i\}$ for any positive integer i. It is easy to prove that A_i and its components are open. Hence, if a and x^* belong to the same component of A_i, then there exists a path joining a and x^*. Since a and x^* both belong to $L(y_\gamma)$ and there exist a δ_1, such that $d(L_a(y_\gamma), L_{x^*}(y_\gamma)) \geq \delta_1$, we have that, for every i, there must exist an $\alpha_i \in A_i$ for which

$$\begin{aligned} &y_\gamma < f(\alpha_i) < y_\gamma + \frac{1}{i}, \\ &d(\alpha_i, V_a(y_\gamma)) > \tfrac{1}{3}\delta_1 , \\ &d(\alpha_i, L_a(y_\gamma)) > \tfrac{1}{3}\delta_1 . \end{aligned} \tag{39}$$

The sequence α_i contains a convergent subsequence $\alpha_{i(j)}$ such that $\lim_{j \to \infty} \alpha_{i(j)} = \alpha$ and $f(\alpha) = y_\gamma$. This, however, contradicts (39). We may conclude that there exists an ε such that a and x^* do not belong to the same component of $L(y_\gamma + \varepsilon)$. Since there are only a finite number of components and a finite number of minima, we can choose ε independent of a and x^*. By Lemma 9 it now follows that there exists a δ_2 (independent of a and x^*) such that $d(L_a(y_\gamma + \varepsilon), L_{x^*}(y_\gamma + \varepsilon)) \geq \delta_2$. Hence, if $y_\gamma < y \leq y_\gamma + \varepsilon$, then, for all $a \in L(y_\gamma)$and $x^* \in L(y)$, $x^* \notin L_a(y_\gamma)$, we have proven that $d(L_a(y), L_{x^*}(y)) \geq \delta_2$. By choosing $\delta = \min\{\delta_1, \delta_2\}$ the result is now immediate. □

We now return to the probability that no local minimum is found by Single Linkage in a component $L_a(y_\gamma)$ although there is a sample point in $L_a(y_\gamma)$. We shall show first that $L_a(y_\gamma)$ must contain a local minimum which is in a sense conveniently located. The possibility that $L_a(y_\gamma)$ contains a number of local minima, only one of which may be discovered, creates some extra difficulties in the reasoning that follows below.

First, since $L_a(y_\gamma)$ is compact, there must exist a point $e \in S$ which is the global minimum of f over $L_a(y_\gamma)$, i.e. $f(e) < f(x)$ for all $x \in L_a(y_\gamma)$. (We assume that the global minimum of f over $L_a(y_\gamma)$ is unique; this, however, is not essential.) Since P cannot leave a component of a level set in which it is started (Lemma 3), and since e has the smallest function value in $L_a(y_\gamma)$, it follows that if P is started in e

then it stops in e as well. Hence, e is a local minimum of f over S, so that e is in the interior of S_τ.

If $f(e)=y_\gamma$, then the sample point in $L_a(y_\gamma)$ must equal the local minimum e, so that a local minimum in $L_a(y_\gamma)$ has been found.

To analyze the usual situation that $f(e)<y_\gamma$, it is convenient to prove the following theorem, which provides useful information about the location of the local minimum e.

Lemma 11. *If, for any component $L_a(y_\gamma)$ of $L(y_\gamma)$, e is the unique global minimum of f over $L_a(y_\gamma)$, and if $f(e)<y_\gamma$, then there exists a neighbourhood E of e satisfying:*

$$E\subset L_a(y_\gamma); \tag{40}$$

$$\textit{if } x_1\in E \textit{ and if } x_2\in L_a(y_\gamma)\backslash E, \textit{ then } f(x_1)<f(x_2); \tag{41}$$

$$\textit{if } \bar{x} \textit{ is any stationary point other than } e, \textit{ and if } x\in E, \textit{ then } \|x-\bar{x}\|>\upsilon; \tag{42}$$

$$E\cap Q_\tau=\emptyset; \tag{43}$$

$$m(E)>0 \tag{44}$$

Proof. Let Y_υ^e be the set of points which are within distance υ of any stationary point other than e, and let $Z_1=\{x\in L_a(y_\gamma)\,|\,x\in Y_\upsilon^e\cup Q_\tau\}$. If $Z_1=\emptyset$, then we define $\bar{y}$ to be y_γ, else $\bar{y}$ is the infimum of f over Z_1. Now let E be the set $\{x\in L_a(y_\gamma)\,|\,f(x)<\bar{y}\}$. E obviously satisfies (40)-(44). Since we assumed that all local minima are in S_τ, and that a positive constant ε exists such that the distance of any two stationary points exceeds ε, it follows that $\bar{y}>f(e)$ (choose $\upsilon<\varepsilon$). Hence, it follows from the continuity of f that $m(E)>0$. □

It follows from (44) and (8) that the probability that E contains a sample point tends to 1 with increasing k. Suppose that $y_\gamma^{(\gamma kN)}\leq y_\gamma+\varepsilon$, for the ε mentioned in Lemma 10, and suppose that E contains a reduced sample point x (the probability that both events occur simultaneously tends to 1 with increasing k). Let $\bar{x}$ be the seed point of the cluster to which x is assigned. There are four possibilities:

$$\bar{x}\in E,\quad \bar{x}\in L_a(y_\gamma)\backslash E,\quad \bar{x}\in L(y_\gamma)\backslash L_a(y_\gamma)\quad\text{or}\quad \bar{x}\notin L(y_\gamma).$$

(i) If $\bar{x}\in E$, then it follows from (42) and (43) that either the local minimum e has been located already, or P is applied to $\bar{x}$ to find a minimum in $L_a(y_\gamma)$.

(ii) If $\bar{x}\in L_a(y_\gamma)\backslash E$, then $f(\bar{x})>f(x)$ by (41). It follows from the description of Single Linkage that a point x cannot be assigned to a seed point $\bar{x}$ which is not a minimum, if $f(\bar{x})>f(x)$. Hence, $\bar{x}$ must be a local minimum in $L_a(y_\gamma)$.

(iii) Suppose that $\bar{x}\notin L(y_\gamma)$. Since all seed points are in $L(y_\gamma+\varepsilon)$. and since there is no local minimum x^* with $y_\gamma<f(x^*)\leq y_\gamma+\varepsilon$ (see the proof of Lemma 10), it follows that $\bar{x}$ is not a minimum. However, x cannot be assigned to a seed point $\bar{x}$ which is not a minimum if $f(\bar{x})>f(x)$. Hence $\bar{x}$ cannot be outside $L(y_\gamma)$, and this case cannot occur.

(iv) Suppose that $\bar{x} \in L(y_\gamma) \backslash L_a(y_\gamma)$. Obviously, the component $L_{\bar{x}}(y_\gamma)$ contains a minimum x^*, since if P is applied to $\bar{x}$, it converges to a minimum in $L_{\bar{x}}(y_\gamma)$ by assumption. Hence, it follows from Lemma 10 that there is a $\delta > 0$, such that there is no point in $L_a(y_\gamma)$ within distance δ of any point in $L_{\bar{x}}(y_\gamma)$. It follows that if the critical distance r_k of Single Linkage is smaller than δ, then x cannot be assigned to the cluster initiated by $\bar{x}$.

Thus, if r_k tends to 0 with increasing k, and if $\underline{\xi}$ is the index of the iteration in which a local minimum is found in a component $L_a(y_\gamma)$ which contains a sample point, then we have shown that the probability that $\underline{\xi}$ is less than k tends to 1 with increasing k. Hence, for every $\varepsilon > 0$ there exists a k_0 such that $\Pr[\underline{\xi} < k_0] > 1 - \varepsilon$, and we may conclude that $\underline{\xi}$ is finite with probability 1. Hence, we have proved the following theorem.

Theorem 12. *If the critical distance r_k of Single Linkage tends to 0 with increasing k, then, in every component of $L(y_\gamma)$ in which a point has been sampled, a local minimum will be found by Single Linkage within a finite number of iterations with probability* 1. □

Note that we can omit the provision that the component $L_a(y_\gamma)$ of $L(y_\gamma)$ must contain a sample point, if we assume that the measure of $L_a(y_\gamma)$ is positive. If no local minimum x^* exists with $f(x^*) = y_\gamma$, then this latter assumption is satisfied for every component of $L(y_\gamma)$.

3.3. Mode Analysis

Density clustering and Single Linkage are based on very simple properties of the uniform distribution. In Density Clustering, a cluster is expanded if a certain region contains a reduced sample point and in Single Linkage a reduced sample point is assigned to a cluster if it is within the critical distance from a point which has already been assigned to the cluster. In principle it should be possible to design superior methods by using the information of more than two sample points simultaneously. The *mode analysis* approach to clustering (Wishart, 1969) is an example of an approach where several points are used simultaneously to determine the regions in which there is a high density of points to be clustered. However, the method proposed in Wishart (1969) is not suitable for our purpose since it ignores much of the information available, like the fact that the reduced sample points are a subset of the uniform sample. The technique proposed in Spircu (1979) (based on Parzen, 1962) allows one to estimate the distribution from which the reduced sample points are drawn, ignoring however that this distribution changes over time through its dependence on $L(y_k^{(\gamma kN)})$. This difficulty can be overcome (Ruygrok, 1982), but the resulting method is cumbersome and inferior to the much simpler one presented below.

We shall describe a method in which S is partitioned into small hypercubes or *cells*. We say that a cell A is *full* if it contains more than

$$\frac{1}{2}\frac{m(A)kN}{m(S)} \tag{45}$$

reduced sample points (i.e. more than half the expected number of sample points in A). If a cell is not full it is *empty*. We say that two cells are *neighbours*, or *neighbouring* cells, if they contain elements which are arbitrarily close to each other. We shall let a cluster correspond to a connected subset of S which corresponds to a number of full cells. These clusters can be found by applying a Single Linkage type algorithm to the full cells, such that if two cells are neighbours, then they are assigned to the same cluster.

A stepwise description follows below. (The sets Q_τ, Y_v and X_v^* are needed and defined for the same reasons as in the previous subsection.)

Mode Analysis

Step 1 (Determine reduced sample). Determine the reduced sample by taking the γkN sample points with the smallest function values.

Step 2 (Define cells). Divide S into v cells.

Step 3 (Determine full cells). For each cell, determine the number of reduced sample points in the cell. If this number exceeds (45) then the cell is full, else it is empty.

Step 4 (Determine seed cell). If all full cells have been assigned to a cluster, stop. If an unclustered full cell exists which contains a minimum which is in X^*, then this cell is the new seed cell; go to Step 5.

Determine the point $\bar{x}$ which has the smallest function value among the reduced sample points which are in unclustered full cells. The cell which contains $\bar{x}$ is the new seed cell.

If $\bar{x} \in S_\tau$ and if $\bar{x} \notin X_v^*$, then apply P to $\bar{x}$ to find a local minimum x^*; add new stationary points encountered during this search (possibly including x^*) to X^*.

Step 5 (Form cluster). A cluster is initiated by the seed cell which is determined in Step 4.

Full cells which are a neighbour of a cell already in the cluster are assigned to the cluster, until there are no more such cells. Go to Step 4.

Since the properties of Mode Analysis do not really depend on $\sqrt[n]{v}$ being integer, we will, for the sake of analysis, assume that S is a hypercube and that $\sqrt[n]{v}$ is an integer so that S can be divided in v equal hypercubes. For some $\sigma > 0$, we choose v to be equal to $kN/(\sigma \log kN)$ so that each cell has measure $(m(S)\sigma \log kN)/kN$.

Intuitively speaking, we can say that Mode Analysis and Single Linkage will result in similar clusters and similar sets X^*, since the measure of the points within the critical distance (35) of a given point equals the measure of a cell. However, Mode Analysis seems somewhat less dependent on the particular irregularities of the sample, because it considers a number of sample points at the same time ($\sigma \log kN$ in expectation).

It is not possible to prove the superiority of either of the two methods rigorously. For instance, consider a one-dimensional function, such that $L(y_k^{(\gamma kN)})$ consists of two components. Let the distance between both components be d and let r_k equal $(m(S)\sigma \log kN)/kN$, i.e. the critical distance of Single Linkage and the cell size of Mode Analysis. Let us assume that the probability that a cell is full is high if the fraction of the cell that intersects with $L(y_k^{(\gamma kN)})$ exceeds $\frac{1}{2}$, and that a cell is probably empty if this fraction is smaller than $\frac{1}{2}$. (We shall see later that this is true if σ is large enough.) Obviously, if $d > 2r_k$, then both methods will recognize both components as such, and if $d < \frac{1}{2}r_k$, then it is likely that both methods will fail to detect both components. If $r_k < d < 2r_k$, then Single Linkage always detects both components, whereas, with small probability, Mode Analysis may fail. However, if $\frac{1}{2}r_k < d < r_k$, then the probability that Single Linkage will assign all reduced sample points to the same cluster is considerable, since this will happen if two reduced sample points exist in the different components which are within distance r_k of each other. For Mode Analysis this probability is smaller, since it is possible that the region between both components covers an important part of one of the cells, in which case this cell is probably empty.

Clearly, the above arguments loose their relevance if k tends to infinity, since the distance between the components does not tend to zero with increasing k and r_k. For k large enough it turns out that Mode Analysis and Single Linkage are very similar. Actually, it is possible to adjust the analysis of Single Linkage such that it can be applied to Mode Analysis to yield similar results. The most important adjustment of the analysis is needed in Lemma 7, for which we will now give the appropriate extension.

Lemma 13. *Let S be partitioned into equal hypercubes with edgelength r_1. For any $\tau > 0$ and $\upsilon > 0$, let a be an element of $M_{\tau,\upsilon}$, and let C_1 be the cell which contains a. Then there exists a cell C_2 (depending on r_1) which is a neighbour of C_1 such that, uniformly in a,*

$$\lim_{r_1 \downarrow 0} \frac{m(\{x \in C_2 \mid f(x) < f(a)\})}{m(C_2)} = 1. \tag{46}$$

Proof. We will proof this theorem by adjusting the proof of Lemma 7; notations used in this latter proof have the same meaning here. We can choose r and r_1, such that $r = 2r_1\sqrt{n}$. Hence, we may assume that C_1 and all its neighbours are contained in $B_{a,r}$. We know that there exists a transformation $z = U(x-a)$ which maps $B_{a,r}$ into $\{z \in \mathbb{R}^n \mid \|z\| \leq r\}$, such that the orthogonal matrix U satisfies (27) and such that the image of the hypercube C_1 is a hypercube with edgelength r_1 containing the origin.

We first prove that there exists a cell C_2 which is a neighbour of C_1 of which the image (under the above transformation) is completely contained in $A^- = \{z \in \mathbb{R}^n \mid z^{(1)} \leq 0\}$. For this purpose, note that each hypercube has a vertex of which the first coordinate is smaller than or equal to the first coordinate of any other member of the hypercube. Hence, C_1 has a vertex, say a_1, which has the property that, for all $x \in C_1$,

$$e_1^{\mathrm{T}} U(a_1 - a) \leqslant e_1^{\mathrm{T}} U(x - a). \tag{47}$$

Since $a \in C_1$ and $e_1^{\mathrm{T}} U(a - a) = 0$, it follows that the image of a_1, $U(a_1 - a)$ is in A^-. Obviously, each vertex (in $M_{T,v}$) is shared by 2^n cells, and each of these cells can be characterized by the fact whether or not the i-th coordinate of the elements in the cell is greater than the i-th coordinate of this vertex ($i = 1, 2, \ldots, n$). More formally, a cell which has a_1 as a vertex consists of elements x that can be written as

$$x = a_1 + \sum_{i=1}^{n} \lambda_i e_i, \tag{48}$$

where e_i is the i-th unit vector, and either $-r_1 \leqslant \lambda_i \leqslant 0$ or $0 \leqslant \lambda_i \leqslant r_1$ for every $i = 1, 2, \ldots, n$. Now consider a cell, say C_2, whose elements satisfy (48) where

$$\begin{aligned} &\lambda_i \leqslant 0 \quad \text{if } e_1^{\mathrm{T}} U e_i \geqslant 0, \\ &\lambda_i \geqslant 0 \quad \text{if } e_1^{\mathrm{T}} U e_i < 0. \end{aligned} \tag{49}$$

Hence,

$$\begin{aligned} e_1^{\mathrm{T}} U(x - a) &= e_1^{\mathrm{T}} U \left((a_1 - a) + \sum_{i=1}^{n} \lambda_i e_i \right) \\ &= e_1^{\mathrm{T}} U(a_1 - a) + \sum_{i=1}^{n} \lambda_i e_1^{\mathrm{T}} U e_i \\ &\leqslant 0. \end{aligned} \tag{50}$$

It follows that C_2 is completely contained in A^-. It is easy to show that C_2 is not C_1 since, for all $x \in C_2$ we have that

$$e_1^{\mathrm{T}} U(x - a) \leqslant e_1^{\mathrm{T}} U(a_1 - a), \tag{51}$$

while for the elements of C_1 the reverse is true. Thus, we have proved that C_2 is a neighbour of C_1 which is completely contained in A^-.

We know from the proof of Lemma 7 that there exist positive constants c and p, such that the elements whose images are in

$$\left\{ z \in \mathbb{R}^n \,\middle|\, \|z\| \leqslant r \text{ and } z^{(1)} \leqslant -\tfrac{1}{2} \frac{r^2 c}{p} \right\} \tag{52}$$

have a function value smaller than $f(a)$. Since the image of C_2 is a hypercube with edgelength r_1, which is completely contained in $\{z \in \mathbb{R}^n \mid \|z\| \leqslant r \text{ and } z^{(1)} \leqslant 0\}$, simple calculations yield

$$m(\{x \in C_2 | f(x) < f(a)\}) \geqslant r_1^n - (r_1\sqrt{n})^{n-1} \cdot \frac{r^2 c}{2p}. \tag{53}$$

Thus

$$\lim_{r \downarrow 0} \frac{m(\{x \in C_2 | f(x) < f(a)\})}{m(C_2)} \geqslant \lim_{r_1 \downarrow 0} \frac{r_1^n - r_1^{n-1} \cdot \dfrac{r^2 c n^{(n-1)/2}}{2p}}{r^n}$$

$$= \lim_{r_1 \downarrow 0} 1 - \frac{2r_1 c n^{(n+1)/2}}{p} = 1. \tag{54}$$

(Recall that $r = 2r_1\sqrt{n}$.) Since the above arguments are independent of a, the result is now immediate. □

To determine the probability that a local search is started by Mode Analysis, we divide S in the three sets Q_τ, Y_υ and $M_{\tau,\upsilon}$ again. As in Single Linkage, no local search is every started in Q_τ, and, from our earlier assumption, for any stationary point, only one local search can be started within distance υ of this point.

Now let us consider the probability that P is applied to a reduced sample point $\underline{x} = a \in M_{\tau,\upsilon}$. Obviously, Mode Analysis will not start P from a point a if the cell containing a has a neighbour which is full and contains a sample point with a function value smaller than $f(a)$. To determine the probability of the above event, let us first consider the simpler case where a is an arbitrary sample point in $M_{\tau,\upsilon}$.

Let C_1 be the cell in which a is located. Obviously, the remaining $kN - 1$ sample points are still distributed according to a uniform distribution over S. From Lemma 13 we know, that for any $\frac{1}{2} < \beta < 1$, there exists a k_0 such that if $k > k_0$, then there exists a neighbour C_2 of C_1 with

$$\frac{m(\{x \in C_2 | f(x) < f(a)\})}{m(C_2)} \geqslant \beta. \tag{55}$$

Since $m(C_2) = (m(S)\sigma \log kN)/kN$, the probability that less than $\frac{1}{2}\sigma \log kN$ of the $kN - 1$ sample points are in $\{x \in C_2 | f(x) < f(a))\}$ is smaller than

$$\sum_{i=0}^{\frac{1}{2}\sigma \log kN - 1} \binom{kN-1}{i} \left(\frac{\beta\sigma \log kN}{kN}\right)^i \left(1 - \frac{\beta\sigma \log kN}{kN}\right)^{kN-i-1} \tag{56}$$

Obviously, for k large enough (56) is smaller than

$$\sum_{i=1}^{\frac{1}{2}\sigma \log kN} 2\binom{kN}{i} \left(\frac{\beta\sigma \log kN}{kN}\right)^i \left(1 - \frac{\beta\sigma \log kN}{kN}\right)^{kN-i} \tag{57}$$

Using Chernoff's inequality (Erdös and Spencer, 1974), it follows that we can choose β such that (57) is $O(k^{-\sigma/10})$. Hence, for an arbitrary sample point a in $M_{\tau,\upsilon}$, we proved that the probability that the cell containing a has no neighbour with more than $\frac{1}{2}\sigma \log kN$ sample points with a function value smaller than $f(a)$ is $O(k^{-\sigma/10})$. (Of course, this is not the sharpest possible bound, but it suffices for our purpose.) Since the number of sample points in iteration k is kN, the probability that there exists a sample point a in $M_{\tau,\upsilon}$, such that the cell containing a has no neighbour with more than $\frac{1}{2}\sigma \log kN$ sample points with a function value smaller than $f(a)$ is $O(k^{1-\sigma/10})$. It is not difficult to verify that the same statements are true with respect to the reduced sample. Hence, if $\sigma > 10$, then the probability that a local search is started by Mode Analysis in iteration k tends to 0 with increasing k. As in Single Linkage, we can use the Borel-Cantelli Lemma to prove the following analagon of Theorem 8.

Theorem 14. *If the number of cells in Mode Analysis is $kN/(\sigma \log kN)$ with $\sigma > 10$, then the probability that a local search is started by Mode Analysis in iteration k tends to* 0 *with increasing k. If $\sigma > 20$, then, even if the sampling continues for ever, the total number of local searches started by Mode Analysis is finite with probability* 1. □

Let us now consider the second possible failure of Mode Analysis, i.e. the possibility that no local minimum is found in a component $L_a(y_\gamma)$ although a sample point exists in $L_a(y_\gamma)$. The analysis of this possibility is similar to the analysis of the corresponding possibility for Single Linkage. Ignoring details, we just state the final result.

Theorem 15. *If the number of cells in Mode Analysis is $kN/(\sigma \log kN)$ with $\sigma > 0$, then, in every component $L(y_\gamma)$ in which a point has been sampled, a local minimum will be found by Mode Analysis within a finite number of iterations with probability* 1. □

4. Concluding remarks

The methods that have been described in the three previous subsections share one major deficiency. Although we know that a region of attraction cannot intersect with two different components of a level set, it is possible that a component of $L(y_k^{(\gamma kN)})$ contains more than one region of attraction. Since only one local search is started in each cluster, it is therefore possible that a local minimum may not be found although its region of attraction contains a reduced sample point. In this section we briefly consider three possible remedies to overcome this problem.

A first remedy is to replace every reduced sample point x by another point which is the result of a *steepest descent step* started in x, i.e. a one-dimensional search from

x in the direction of the negative gradient in x. The clustering procedure can then be applied to the resulting *transformed sample* as though it was the reduced sample (Boender et al., 1982). From a theoretical point of view, however, the transformed sample has the disadvantage that its elements are no longer a subset of the original uniform sample. Thus, the analysis of the clustering methods that has been described in the previous subsections is no longer valid.

A second remedy is based on the observations that the negative gradient at a point which belongs to the region of attraction of a minimum x^*, will generally have a component in the direction of x^*. The methods described in the foregoing subsections can be improved by inspecting the gradient at a point before assigning it to a cluster. More precisely, if a cluster is initiated by a seed point $\bar{x}$ (usually a local minimum), we then approximate the derivative of f in x in the direction of $\bar{x}$ by

$$\frac{f(x+h(\bar{x}-x))-f(x)}{h\|\bar{x}-x\|} \tag{58}$$

for small h, and we reject x for the cluster if this value is positive. In the case of Mode Analysis, we could inspect the gradient at the reduced sample point with the smallest function value in a cell before assigning the cell to the cluster. Although this *gradient criterium* can be incorrect and may affect the methods in an unpredictable way, it turns out to be very useful from a computational point of view (see Boender et al., 1982).

A third possible remedy affects the methods even more deeply. Since we are only interested in the global minimum, we could reduce the sample even further. If, for some $\gamma \in (0, 1)$, we would (re)define the reduced sample to contain the $(kN)^{\gamma}$ sample points with the smallest function values, we could prove much stronger (asymptotic) results. However, the statements are valid only if all reduced sample points are arbitrarily close to a global minimum. Obviously, at that moment the problem has been solved a long time ago already. Moreover, the analysis would not yield any insight into the way in which the resulting methods would function before we arrive in the asymptotic case.

We conclude that the idea of sample reduction and clustering gives rise to interesting methods, but does not solve the problem satisfactorily from a theoretical point of view. In particular we cannot always avoid that a cluster contains several regions of attraction, so that a (local) minimum may still be missed. In Part II of this paper we will deal with this problem in a more fundamental way. The results obtained in Part I will turn out to play an important role in Part II as well.

Acknowledgements

The research of the first author was partially supported by a NATO Scientist Fellowship. The research of the second author was partially supported by the

Netherlands Foundation for Mathematics SMC with financial aid from the Netherlands Organization for Advancement of Pure Research (ZWO).

References

R.S. Anderssen, "Global optimization," in: R.S. Anderssen, L.S. Jennings and D.M. Ryan, eds., *Optimization* (University of Queensland Press, 1972) pp. 1-15.

R.S. Anderssen and P. Bloomfield, "Properties of the random search in global optimization," *Journal of Optimization Theory and Applications* 16 (1975) 383-398.

F. Archetti and B. Betro, "A priori analysis of deterministic strategies for global optimization," in: Dixon and Szegö (1978a) pp. 31-48.

F. Archetti and B. Betro, "On the effectiveness of uniform random sampling in global optimization problems," Technical Report, University of Pisa (Pisa, Italy, 1978b).

F. Archetti and F. Frontini, "The application of a global optimization method to some technological problems" (1978), in: Dixon and Szegö (1978a) pp. 179-188.

R.R. Bahadur, "A note on quantiles in large samples," *Annals of Mathematical Statistics* 37 (1966) 577-580.

R.W. Becker and G.V. Lago, "A global optimization algorithm," in: *Proceedings of the 8th Allerton Conference on Circuits and Systems Theory* (1970).

B. Betro, "Bayesian testing of nonparametric hypotheses and its application to global optimization," Technical Report, CNR-IAMI (Italy, 1981).

C.G.E. Boender, A.H.G. Rinnooy Kan, L. Stougie and G.T. Timmer, "Global optimization: A stochastic approach," in: F. Archetti and M. Cugiani, eds., *Numerical Techniques for Stochastic Systems* (North-Holland, Amsterdam, 1980) pp. 387-394.

C.G.E. Boender, A.H.G. Rinnooy Kan, L. Stougie and G.T. Timmer, "A stochastic method for global optimization," *Mathematical Programming* 22 (1982) 125-140.

C.G.E. Boender and A.H.G. Rinnooy Kan, "A Bayesian analysis of the number of cells of a multinomial distribution," *The Statistician* 32 (1983) 240-248.

C.G.E. Boender, "The generalized multinomial distribution: A Bayesian analysis and applications," Ph.D. Dissertation, Erasmus Universiteit Rotterdam (Centrum voor Wiskunde en Informatica, Amsterdam, 1984).

C.G.E. Boender and A.H.G. Rinnooy Kan, "Bayesian Stopping rules for a class of stochastic global optimization methods," Technical Report, Econometric Institute, Erasmus University Rotterdam (1985).

C.G.E. Boender, A.H.G. Rinnooy Kan and G.T. Timmer, "A stochastic approach to global optimization," in: K. Schittkowski, ed., *Computational Mathematical Programming* (NATO ASI Series, Vol. F15, Springer-Verlag, Berlin, 1985) pp. 291-308.

S.H. Brooks, "A discussion of random methods for seeking maxima," *Operations Research* 6 (1958) 244-251.

K.L. Chung, *A Course in Probability Theory* (Academic Press, London, 1974).

L. Devroye, "Progressive global random search of continuous functions," *Mathematical Programming* 15 (1978) 330-342.

L.C.W. Dixon and G.P. Szegö, eds., *Towards Global Optimization* (North-Holland, Amsterdam, 1975).

L.C.W. Dixon, J. Gomulka and G.P. Szegö, "Towards global optimization," in: Dixon and Szegö (1975) pp. 29-54.

L.C.W. Dixon and G.P. Szegö, eds., *Towards Global Optimization* 2 (North-Holland, Amsterdam, 1978a).

L.C.W. Dixon and G.P. Szegö, "The global optimization problem" (1978b), in: Dixon and Szegö (1978a) pp. 1-15.

P. Erdös and J. Spencer, *Probabilistic Methods in Combinatorics* (Academic Press, London, 1979).

J.K. Hartman, "Some experiments in global optimization," *Naval Research Logistics Quarterly* 20 (1973) 569-576.

V.V. Ivanov, "On optimal algorithms of minimization in the class of functions with the Lipschitz condition," *Information Processing* 2 (1972) 1324-1327.

E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics* 33 (1962) 1065-1076.

A.H.G. Rinnooy Kan and G.T. Timmer, "Stochastic methods for global optimization," *American Journal of Mathematical and Management Sciences* 4 (1984) 7-40.

A.H.G. Rinnooy Kan and G.T. Timmer, "Stochastic global optimization methods. Part II: Multi level methods," *Mathematical Programming* 38 (1987) 57-78 (this issue).

R.Y. Rubinstein, *Simulation and the Monte Carlo Method* (John Wiley & Sons, New York, 1981).

A.J. Ruygrok, "Mode Analysis in globaal optimaliseren," Master Thesis, Erasmus University Rotterdam (in Dutch) (1982).

J.A. Snijman and L.P. Fatti, "A multistart global minimization algorithm with dynamic search trajectories," Technical Report, University of Pretoria (Republic of South Africa, 1985).

I.M. Sobol, "On an estimate of the accuracy of a simple multidimensional search," *Soviet Math. Dokl.* 26 (1982) 398-401.

F.J. Solis and R.J.E. Wets, "Minimization by random search techniques," *Mathematics of Operations Research* 6 (1981) 19-30.

L. Spircu, "Cluster analysis in global optimization," *Economic Computation and Economic Cybernetic Studies and Research* 13 (1979) 43-50.

A.G. Sukharev, "Optimal strategies of the search for an extremum," *Computational Mathematics and Mathematical Physics* 11 (1971) 119-137.

A.A. Törn, "Cluster analysis using seed points and density determined hyperspheres with an application to global optimization," in: *Proceeding of the Third International Conference on Pattern Recognition, Coronado, California* (1976) *pp.* 394-398.

A.A. Törn, "A search clustering approach to global optimization" (1978), in: Dixon and Szegö (1978a) pp. 49-62.

D. Wishart, "Mode Analysis: A generalization of nearest neighbour which reduces chaining effects," in: A.J. Cole, ed., *Numerical Taxonomy* (Academic Press, New York, 1969).

P. Wolfe, "Convergence conditions for ascent methods," *Siam Review* 11 (1969) 226-235.

P. Wolfe, "Convergence conditions for ascent methods II: some corrections," *Siam Review* 13 (1971) 185-188.

R. Zielinski, "A stochastic estimate of the structure of multi-extremal problems," *Mathematical Programming* 21 (1981) 348-356.