GLOBAL OPTIMIZATION REQUIRES GLOBAL INFORMATION

by

William Baritompa and Chris Stephens
Department of Mathematics & Statistics
University of Canterbury
Christchurch, New Zealand

No. 147 September, 1996

Abstract

There are many global optimization algorithms which do not use global information. We broaden previous results, showing limitations on such algorithms, even if allowed to run forever. We show deterministic algorithms must sample a dense set to find the global optimum value and can never be guaranteed to converge only to global optimizers. Further, analogous results show introducing a stochastic element does not overcome these limitations. An example is simulated annealing in practice. Our results show there are functions for which the probability of success is arbitrarily small.

Key Words. Global optimization, convergence, stochastic algorithms, deterministic algorithms

Global Optimization Requires Global Information

William Baritompa Chris Stephens

September 30, 1996

Abstract

There are many global optimization algorithms which do not use global information. We broaden previous results, showing limitations on such algorithms, even if allowed to run forever. We show deterministic algorithms must sample a dense set to find the global optimum value and can never be guaranteed to converge only to global optimizers. Further, analogous results show introducing a stochastic element does not overcome these limitations. An example is simulated annealing in practice. Our results show there are functions for which the probability of success is arbitrarily small.

Key Words. Global optimization, convergence, stochastic algorithms, deterministic algorithms

1 Introduction

There are many global optimization algorithms which use global information. For instance, the algorithms of Piyavskii-Shubert [1, 2] and its variants [3], Mladineo [4], Wood [5], Brent [6], and Breiman & Cutler [7], interval methods [8, 9]), and "standard" simulated annealing ([10] for discrete setting, [11, 12] for continuous). All of these algorithms share the properties that they require global information in the form of a parameter (e.g. Lipschitz constant, bound on second derivate, functional form, or a cooling schedule) and they are guaranteed to find the global optimum. One criticism of these algorithms is that this information, being of a global nature, is hard to obtain (or simply may not be available).

Thus there is a desire to design algorithms which avoid the need for global information. A number of algorithms have been proposed with this in mind. For instance, the DIRECT algorithm of Jones et al. [13], Strongin's algorithm [14], algorithms of Gergel ([15]), pages 3–4) and Sergeyev [15], and simulated annealing in practice. While performing well in empirical tests, all of these algorithms, and indeed all algorithms which avoid global information, have inherent theoretical limitations.

Inherent limitations of algorithms which stop after a finite stage are well known. Solis & Wets [16] point out that "the search for a good[sic]

stopping criterion seems doomed to fail", because as noted by Dixon "even with [the domain] compact convex and f twice differentiable, at each step of the algorithm there will remain an unsampled square region of nonzero measure v (volume) on which f can be redefined (by means of spline fits) so that the unsampled region now contain the global minimum." Thus, after the run it possible that the algorithm failed. Our results strengthen this by implying the existence of functions, a priori, for which the probability of success is arbitrarily small.

Some limitations for algorithms which are allow to run forever are also known. Hansen et al. [17] found a class of functions which Strongin's algorithm fails to converge. It is well known that all deterministic algorithms which use only function values at sample points converge to the global optimum on all continuous functions if and only if it searches a dense set (Törn and Žilinskas [18] provide a proof for this).

In this paper we extend the above results. We show that the result reported in Törn and Žilinskas holds for algorithms which can run forever and use other local information, such as derivatives, in addition to function values. We describe other classes of functions for which these algorithms fail. We show for convergence to the localization of the global optimum, the algorithm always fails, regardless of whether or not it searches a dense set.

Introducing a stochastic element to algorithms is often seen as a way to overcome these limitations, (so that no function can be found that will definitely fail). However, we show that there are analogous results for stochastic algorithms. For instance, for simulated annealing in practice there are functions for which the probability of success is arbitrarily small.

2 Definitions and Notation

Our results require very few conditions on the objective function. Here we consider functions $f : D \to \mathbb{R}$, with $D \subseteq \mathbb{R}^n$ a compact set with no isolated points. Note, functions from standard domains such as the closure of a bounded open subset, or a feasible region in $\mathbb{R}^n$ satisfying "reasonable" constraints, are included.

For the results in this paper, the class of functions for which the algorithm is designed must contain sufficiently many functions. Intuitively we require conditions that allow arbitrarily large modifications of a function on arbitrarily small neighborhoods without affecting the function elsewhere. Formally we have the following.

Definition 2.1 *A non-empty class of functions $\mathcal{F}$ is* sufficiently rich, *if it consists of continuous functions and $\forall y \in \mathbb{R}$, $\forall x \in D$, $\forall f \in \mathcal{F}$ and $\forall N \subseteq D$ an open set containing x, there exists $g \in \mathcal{F}$ such that $g(x) = y$ and $g_{|D\setminus N} = f_{|D\setminus N}$.*

Commonly used examples are continuous functions, C^n, C^∞, continuous functions with a unique global optimum, Lipschitz continuous functions, and functions with Lipschitz continuous derivatives. Many non-standard classes of functions satisfy our definition. For example, continuous functions with continuous first partial derivative and having multiple global optima.

We now provide a formal definition of local information. Let $\mathcal{X}_{finite}$ be the set of all finite sequences in D.

Definition 2.2 Local information *for a family $\mathcal{F}$ is a function, LI, defined on $\mathcal{F} \times \mathcal{X}_{finite}$ satisfying $\forall f, g \in \mathcal{F}$, $\forall X \in \mathcal{X}_{finite}$, $\forall N$ open in D containing X, if $f_{|N} = g_{|N}$ then $LI(f, X) = LI(g, X)$.*

We have intentionally left the range of a local information function unspecified as there are many diverse examples. Local information includes any information depending on function values and any "limiting" information at a finite number of sample points. Examples of such limiting information include (partial and directional) derivatives, but also includes less common "limiting" information, for instance, the limiting fractal dimension at a point. Also, any formula depending on these examples (indeed, on any local information) is itself local information. Thus, the maximum sample point, the maximum slope between sample points and the interpolating polynomial through sample points are local information.

Examples of non-local information are the Lipschitz constant, the constant of uniform continuity, bounds on higher derivatives, the level set associated with a function value, the measure of the basin of attraction, the "depth" of the function, the number of local optima, the functional form and the global optimum itself. If the family of functions has a distribution associated with it, the induced distribution for any of the above examples is also non-local information.

In this paper, we call non-local information *global information*, since we concerned with showing that this is a necessary condition for guaranteeing convergence. Note, some non-local information, such as the Lipschitz constant of the function on a proper subset of the domain, is not global in the sense of being information about the function over the whole domain. Such information still may not guarantee convergence of an algorithm that uses it. More refined versions of our results could characterize which types of global information are sufficient.

We define formally the algorithms under consideration. They produce sample sequences $(x_0, x_1, x_2, \dots)$. We denote by X_k the partial sequence $(x_0, \dots, x_k)$.

Definition 2.3 *A* deterministic sequential sampling algorithm on a class of functions $\mathcal{F}$ *is an algorithm for which there is a local information function LI such that for all $f \in \mathcal{F}$, when running on f, x_{k+1} depends only on $LI(f, X_k)$.*

Note, in a deterministic sequential sampling algorithm x_{k+1} is, itself, local information.

Definition 2.4 *A* stochastic sequential sampling algorithm on a class of functions $\mathcal{F}$ *is an algorithm for which there is a local information function* LI *such that for all* $f \in \mathcal{F}$, *when running on* f, x_{k+1} *depends on* $LI(f, X_k)$ *and* ω_{k+1}, *an instance of a random variable.*

The sample sequence produced by a stochastic sequential algorithm when run on the function f is denoted by X_f. Note that X_f is a random variable. Let ω be $(\omega_1, \omega_2, \dots)$. We denote an instance of X_f by $X_f(\omega)$.

X_f is an infinite sequence as we are considering algorithms which "run forever". Denote the closure of X_f by $\bar{X}_f$ and the sub-sequential limit points by X'_f. Note, X'_f is never empty, as D is compact and $\bar{X}_f = X_f \cup X'_f$.

Often algorithms in literature are justified but showing that they "converge" in the limit. Using the above ideas, we now define this formally. Let X^*_f be the set of global optimizers, $\{x^* \in D | f(x^*) \leq f(x), \forall x \in D\}$.

Definition 2.5 *An algorithm is said to* see the global optimum of f *if* $\bar{X}_f \cap X^*_f \neq \emptyset$.

This form of convergence is typical of algorithms for which the emphasis is on finding the global optimal value. For instance, the DIRECT algorithm uses this type of convergence.

Definition 2.6 *An algorithm is said to* localize the global optimizers *if* $X'_f = X^*_f$ *(or weaker* $\emptyset \neq X'_f \subseteq X^*_f$*).*

This convergence is more typical of algorithms which emphasize finding the location of (some of) the global optimizers in the limit. For instance, Sergeyev's algorithm and simulated annealing use this type of convergence.

3 Main Results

Both definitions of convergence seem easy to satisfy because we allow the algorithm to run forever. However, our results show algorithms which do not use global information have inherent limitations.

Practical realizations of the algorithm terminate with only an initial finite segment of the sample sequence produced, (ideally with some indication of error). Clearly, this compounds the limitations. The results in this paper can be extended to terminating sequential sampling algorithms by repeating the "optimizers found".

3.1 Deterministic Case

We now give the theorems for the the two forms of convergence defined above. These theorems are (almost) special cases of the stochastic results. Since in this case X_f is no longer a random variable, but a determined sequence, the proofs are simple and quite intuitive.

Theorem 3.1 *Any deterministic sequential sampling algorithm on a sufficiently rich class of functions $\mathcal{F}$ sees the global optimum of every function $g \in \mathcal{F}$ if and only if $\bar{X}_f = D$ for every function $f \in \mathcal{F}$.*

Proof: It follows immediately from the definitions that sampling a dense set implies seeing the global optimum (of the same function in fact). For the converse, suppose that there exists a function $f \in \mathcal{F}$ such that $\bar{X}_f \neq D$ and $\bar{X}_f \cap X_f^* \neq \emptyset$. If there is no such f we are done. Since $\bar{X}_f \neq D$, there exist $x_0 \in D \setminus \bar{X}_f$. Take a neighborhood about this point whose closure is disjoint from $\bar{X}_f$. Find another function g agreeing with our original function outside this neighborhood and taking a value smaller than the global optimum of f at x_0. Since

$$\forall n, LI(f, (x_1, x_2, \cdots, x_n)) = LI(g, (x_1, x_2, \cdots, x_n)),$$

running the algorithm on g gives sample sequence $X_g = X_f$ and so fails to see the global optimum of g. □

Since localizing implies seeing, it follows immediately that an algorithm localizing for every g implies $\bar{X}_f = D$ for every f. However, we have the following stronger result.

Theorem 3.2 *For any deterministic sequential sampling algorithm on a sufficiently rich class of functions $\mathcal{F}$, there exists a function in $\mathcal{F}$ for which the algorithm fails to localize the global optimizers.*

Proof: Let f be a function for which $D \setminus X_f^*$ is uncountable, (the existence of such a function is assured by the conditions on $\mathcal{F}$ and D ([19], Theorem 2-80, page 88).) If X_f' is not contained in X_f^* then we are finished, so assuming that $X_f' \subseteq X_f^*$ gives $\bar{X}_f \setminus X_f^* = X_f \setminus X_f^*$ which contains at most countably many points. As $D \setminus X_f^*$ is uncountable, $\bar{X}_f \neq D$. Theorem 3.1 gives a g such that $\bar{X}_g \cap X_g^* = \emptyset$ so $X_g' \not\subseteq X_g^*$. □

3.2 Stochastic Case

The essence of the proofs in the stochastic case relies upon the same ideas as the deterministic case, however quite a few technical difficulties had to be overcome because X_f is a random variable.

We start with a lemma which shows that a bound on the probability of a point being seen extends to a neighbourhood.

Lemma 3.1 *Let X be any random sequence of points in D. If for some $x_0 \in D$ and probability p,*

$$P(x_0 \in X') < p$$

then there exists $y_0 \in D$ and a neighborhood N of y_0 in D, such that

$$P(\bar{X} \cap \bar{N} \neq \emptyset) < p.$$

Furthermore, if M is any closed subset of D and $x_0 \notin M$ then it is possible to choose y_0 and N such that $\bar{N}$ and M are disjoint.

Proof: Consider the non-negative real valued random variable

$$R = \inf\{\|x - x_0\| : x \in X, x \neq x_0\}.$$

Note, whenever X is constantly x_0, $R = 0$. Clearly, $R = 0$ implies $x_0 \in X'$ so $P(R = 0) \leq P(x_0 \in X') < p$. By right-hand continuity of the cumulative distribution function, there exists $\delta > 0$ such that $P(R \leq \delta) < p$.

Therefore,

$$P(\bar{X} \cap B \neq \emptyset) \leq P(R \leq \delta) < p$$

where $B = \{x : \|x - x_0\| < \delta, x \neq x_0\}$ is the punctured open ball of radius δ centered at x_0.

Finally, since D has no isolated points and $x_0 \notin M$ there exists $y_0 \in D \cap B \setminus M$. Let N be a neighborhood of y_0 whose closure is contained within $B \setminus M$. Then

$$P(\bar{X} \cap \bar{N} \neq \emptyset) \leq P(\bar{X} \cap B \neq \emptyset) < p.$$

and furthermore, $\bar{N}$ is disjoint from M. □

We now give the stochastic analog for Theorem 3.1.

Theorem 3.3 *For any probability p and any stochastic sequential sampling algorithm,*

$$P(\textit{algorithm sees the global optimum of } g) \geq p, \forall g) \in \mathcal{F}$$

if and only if

$$P(x \in \bar{X}_f) \geq p, \forall x \in D, f \in \mathcal{F}.$$

Proof: If the probability for each point in the domain being in $\bar{X}_f$ is greater than or equal to p, it follows immediately that the global optimum points (of the same function) have probability greater than or equal to p of being seen. For the converse, suppose that there exists $f \in \mathcal{F}$ and $x_0 \in D$ such that $P(x_0 \in \bar{X}_f) < p$. Clearly, $x_0 \in X'_f$ implies $x_0 \in \bar{X}_f$, so $P(x_0 \in X'_f) < p$ and Lemma 3.1 gives a non-empty neighborhood $N \subseteq D$ such that

$$P(\bar{X}_f \cap \bar{N} \neq \emptyset) < p.$$

Because $\mathcal{F}$ is sufficiently rich there exists $g \in \mathcal{F}$ such that

$$f_{|D\setminus N} = g_{|D\setminus N} \tag{1}$$

and

$$\min_{x\in D} f(x) > \min_{x\in D} g(x). \tag{2}$$

Since $\forall k x_{k+1}$ depends only on $LI(f, X_k)$ and ω_{k+1}, it follows from (1) that if $\bar{X}_f(\omega) \cap \bar{N} = \emptyset$ then $X_f(\omega) = X_g(\omega)$ and so $\bar{X}_f(\omega) = \bar{X}_g(\omega)$. Therefore,

$$P\left(\bar{X}_g \cap \bar{N} \neq \emptyset\right) < p.$$

As the global optimum of g is contained within N, P(algorithms sees the global optimum of g) $< p$. □

As in Section 3.1, we immediately get that P(algorithm localizes the global optimum of g) $\geq p, \forall g$ implies $P(x \in \bar{X}_f) \geq p, \forall x$ and f.

In the deterministic case, Theorem 3.2 showed attempts to localize are guaranteed to fail. For stochastic algorithms the existence of functions with *zero* probability of localizing cannot be guaranteed. However, we do have the following result, analogous to Theorem 3.2 and stronger than the above.

Theorem 3.4 *For any stochastic sequential sampling algorithm and any $\epsilon > 0$ there exists a function $f \in \mathcal{F}$ such that*

$$P(\textit{algorithm localizes the global optimum of } f) < \epsilon.$$

Proof: Suppose, to the contrary, that for all $f \in \mathcal{F}$ and for some fixed $\epsilon > 0$,

$$P(X'_f \subseteq X^*_f) \geq \epsilon. \tag{3}$$

We obtain a contradiction by showing this allows the construction of a function $g \in \mathcal{F}$ and a subset M such that $P(X'_g \subseteq M) > 1$!

Let $n > 2$ be an integer such that $1/n < \epsilon$ and let $x_1, \ldots, x_{n+2}$ be $n+2$ distinct points in D. We construct g in the following iterative manner.

For $i \in \{1, \ldots, n+2\}$, we produce $f_i \in \mathcal{F}$ and open set N_i with $\bar{N}_i$ disjoint from $\{x_{i+1}, \ldots, x_{n+2}\}$ such that

$$P(X'_{f_i} \subseteq M_i) > i(n-1)/n^2 \tag{4}$$

where

$$M_i = \bigcup_{k=1}^{i} N_k$$

The function $g = f_{n+2} \in \mathcal{F}$ and $M = M_{n+2}$ give the desired contradiction.

The specific details are:

For $i = 1$, let N_1 be a neighborhood of x_1 whose closure is disjoint from $\{x_2, \ldots, x_{n+2}\}$. Since $\mathcal{F}$ is sufficiently rich there exists $f_1 \in \mathcal{F}$ such that $X^*_{f_1} \subseteq N_1$. Then from (3) we have (4) in the case of $i = 1$,

$$P(X'_{f_1} \subseteq N_1) \geq \epsilon > (n-1)/n^2.$$

For $i \leq n+1$, assume we have the required functions and sets. Since x_{i+1} is disjoint from M_i,

$$P(x_{i+1} \in X'_{f_i} | X'_{f_i} \subseteq M_i) = 0 < 1/(ni - i).$$

By Lemma 3.1, there exists $y_{i+1} \in D$ and a neighborhood N_{i+1} of y_{i+1}, whose closure is disjoint from $\bar{M}_i \cup \{x_{i+2}, \ldots, x_n\}$, such that,

$$P(\bar{X}_{f_i} \cap \bar{N}_{i+1} \neq \emptyset | X'_{f_i} \subseteq M_i) < 1/(ni - i).$$

or

$$P(\bar{X}_{f_i} \cap \bar{N}_{i+1} = \emptyset | X'_{f_i} \subseteq M_i) > (ni - i - 1)/(ni - i) \tag{5}$$

Since $\mathcal{F}$ is sufficiently rich, there exists $f_{i+1} \in \mathcal{F}$ such that

$$f_{i+1|D\setminus N_{i+1}} = f_{i|D\setminus N_{i+1}}$$

and

$$X^*_{f_{i+1}} \subseteq N_{i+1}.$$

As in the proof to Theorem 3.3, if $\bar{X}_{f_i}(\omega) \cap \bar{N}_{i+1} = \emptyset$ then $X'_{f_{i+1}}(\omega) = X'_{f_i}(\omega)$. So that

$$\begin{aligned} P(X'_{f_{i+1}} \subseteq M_i) &\geq P(X'_{f_i} \subseteq M_i \text{ and } \bar{X}_{f_i} \cap \bar{N}_{i+1} = \emptyset) \\ &= P(X'_{f_i} \subseteq M_i) \cdot P(\bar{X}_{f_i} \cap \bar{N}_{i+1} = \emptyset | X'_{f_i} \subseteq M_i) \end{aligned}$$

From (4) (for i) and (5) we get

$$P(X'_{f_{i+1}} \subseteq M_i) > (ni - i - 1)/n^2. \tag{6}$$

Observe $[(X'_{f_{i+1}} \subseteq M_i)$ or $(X'_{f_{i+1}} \subseteq N_{i+1})]$ implies $X'_{f_{i+1}} \subseteq M_i \cup N_{i+1}$, the events $(X'_{f_{i+1}} \subseteq M_i)$ and $(X'_{f_{i+1}} \subseteq N_{i+1})$ are mutually exclusive (as N_{i+1} is disjoint from M_i) and $X^*_{f_{i+1}} \subseteq N_{i+1}$. Therefore (3) and (6) give

$$\begin{aligned} P(X'_{f_{i+1}} \subseteq M_{i+1}) &= P(X'_{f_{i+1}} \subseteq M_i \cup N_{i+1}) \\ &\geq P(X'_{f_{i+1}} \subseteq M_i) + P(X'_{f_{i+1}} \subseteq N_{i+1}) \\ &\geq P(X'_{f_{i+1}} \subseteq M_i) + P(X'_{f_{i+1}} \subseteq X^*_{f_{i+1}}) \\ &> (ni - i - 1)/n^2 + 1/n \\ &= (i+1)(n-1)/n^2. \end{aligned}$$

Thus (4) holds for $i + 1$.

4 Examples

The function g constructed where the algorithm fails, may have quite high global constants associated with it, e.g. for Lipschitz continuous functions it may have a high Lipschitz constant. This is precisely the main point. If an overall limit for the Lipschitz constant in the class were known (even probabilistically) and used as a parameter to the algorithm, such examples could be prevented.

For every finitely terminating sequential sampling algorithm on a sufficiently rich class, there is, a priori, a function for which the probability of seeing the global optimum is arbitrarily small. The key here is that a finite terminating algorithm that sees the global optimum can be modified, by repeating the best point, to one which localizes. Theorem 3.4 applies.

An algorithm using "hidden" local information on which the next sample point depends is, strictly speaking, not a sequential sampling algorithm. That is, it uses "global" information. For example, using the results of an internal finite local search or reporting only record values, give such algorithms. Of course, these algorithms correspond to algorithms which do not hide any local information, and suffer the same limitations.

Global information may in fact be disguised. For example, in Sergeyev's paper there are parameters r and ξ involved in the algorithm. Careful reading of a convergence result for this algorithm says for every function there is an r^* dependent on ξ, past which convergence is always guaranteed. However the proof shows $r\xi$ must exceed a multiple of the overall Lipschitz constant involved. So $r\xi$ is in fact a global constant for the given function. Similar comments hold for Gergel's algorithm.

Simulated annealing provides another example where global information is disguised. "Standard" simulated annealing localizes if the cooling schedule is slow enough. Hajek ([20] shows (in the deterministic setting) a necessary and sufficient condition on the cooling schedule depends on the depth of the lowest local minimum. This is clearly a global parameter. In the continuous case where gradients are used, our results show the cooling schedule must depend on global properties. So, attempts to find a suitable (or optimize an existing) cooling schedule by pre-sampling or adjusting the cooling schedule on the run using sample points are doomed to failure. Theorem 3.4 shows that there are always functions in sufficiently rich classes for which the probability of success of such a scheme is arbitrarily small.

Other algorithms do not have hidden global information. The DIRECT algorithm of Jones et al. is guaranteed to see the global optimum and uses only local information. This is a result of looking in a dense set of the domain. This is really where the local information *becomes* global. In practice, however, one stops the algorithm after a finite number of steps, and there will always exist a function for which the global optimum is in an unsampled region.

5 Extensions

The proofs of Theorems 3.1 and 3.3 need only that the global optimum can be extracted from the information obtained at the sample points. A more general result is thus available for other definitions of convergence. For instance, Wood's algorithm "brackets" X^* at each iteration and is proven to converge in the sense that the infinite intersection of the brackets is equal to X^*. Wood's algorithm requires global information to do this, but conceivably another algorithm using only local information could attempt to approximate this approach. Such an algorithm would only converge on all functions in a sufficiently rich class if it sampled a dense set.

In practice one weakens the requirements for X^* to allow for approximate answers. For instance, the set X^*_ϵ of points whose value is within ϵ of the global optimum is often sought and our results apply.

Alternatively, rather than search for points whose value in within an acceptable tolerance of the global optimum, one may be interested in, say, the best 5% percent of the domain. Formally, for $0 \leq \alpha \leq 1$, and α-best point is a point in $X^*_\alpha = \{x : \text{the relative volume of } \{x' | f(x') < f(x)\} \leq \alpha\}\}$. Results for convergence to X^*_α require modifying "sufficiently rich", from meaning functions can be changed at arbitrary points in a given open set, to allowing changes at all points in an arbitrary set of relative volume α contained in a given open set. Analogous to Theorems 3.1 and 3.2, a deterministic algorithm sees an α-best point for every g if and only if the relative volume of $\bar{X}_f$ is greater than or equal to $1 - \alpha$ for every f, and it alway fails to localize every function (for $\alpha < 1/2$). For the stochastic case, the analog to Theorem 3.3 is $P(\text{algorithm sees an } \alpha\text{-best point of } g) \geq p, \forall g$ if and only if $P(\bar{X}_f \cap A \neq \emptyset) \geq p), \forall f, \forall A$ a set of relative volume α. Note, the latter condition is met by choosing at least $N = \ln(1-p)/\ln(1-\alpha)$ points uniformly in the domain. By repeating the best of these points, we establish, in contrast to Theorem 3.4, that there exists an algorithm which localizes a α-best point of every f with arbitrarily high probability. Thus convergence, in terms of seeing and localizing to the α-best points is realizable in practice. We conjecture that N is also a lower bound on the number of points required for stochastic sequential sampling algorithms. Only algorithms which use global information could improve upon this.

All our results assume the algorithms sample only from the "feasible" set. Often functions are defined on a larger domain but one is interested in the global optimum when constraints are satisfied. Algorithms such as relaxed dual and penalty methods use infeasible sample points. Our results can easily be extended to handle this by appropriate reformulation.

Finally, the results in this paper apply to algorithms attempting to find global information other than (approximate) global optimum. All that is necessary is the appropriate definition of sufficiently rich class. The class needs to have functions in it agreeing closely to others associated with different global information.

6 Summary and Conclusions

Limitations on specific types of algorithms have been noted before. We have looked at a more general class of (finitely terminating and infinite) algorithms, both deterministic and stochastic, that might utilize function and derivative values (indeed, any type of limiting information) but not global information. All of these algorithms have theoretical limitations.

Theorem 3.3 means such algorithms will succeed frequently on all functions if and only if all points in the domain are frequently seen. That is, the algorithms must use brute force. Theorem 3.4 shows that attempts to localize the global optima on all functions with such algorithms is doomed to failure. In the real-world, algorithms must stop after a finite time, and there exists functions for which deterministic algorithms fail or stochastic algorithms fail with arbitrarily high probability. So, if no global information about problem is utilized, the real-world function may be one on which the algorithm (likely) fails. We cannot have justified confidence in the results.

However, many of these algorithms do have heuristic justification. They are often designed for certain real-world problems and perform well when tested on these and similar problems. Indeed, using these global optimization heuristics are often far more practical than running general algorithms until the mathematically proven stopping criteria are satisfied. These real-world and test functions must have nicer characteristics, than "randomly chosen" functions from a formal class.

By our results, this "niceness" must be a global characteristic. This illustrates the need to quantify this "niceness" into useful global parameters. Success of such an undertaking would result in algorithms with the practical usefulness of current heuristics with the addition of mathematically justified confidence in the results.

Acknowledgments

We would like to thanks Don Jones for his useful comments and references and the anonymous referees for some of the examples.

References

[1] S. A. Piyavskii. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12:57–67, 1972.

[2] B. O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9:379–388, 1972.

[3] P. Hansen, B. Jaumard, and S-H. Lu. Global optimization of univariate Lipschitz functions: II. New algorithms and computational comparison. *Mathematical Programming*, 55:273–292, 1992.

[4] R. H. Mladineo. An algorithm for finding the global maximum of multimodal, multivariate functions. *Mathematical programming*, pages 188–200, 1986.

[5] G. R. Wood. Multidimensional bisection and global optimization. *Computers and Mathematics with Applications*, pages 161–172, 1991.

[6] R. P. Brent. *Algorithms for minimization without derivatives.* Prentice-Hall, Inc., 1973.

[7] L. Breiman and A. Cutler. A deterministic algorithm for global optimization. *Mathematical Programming*, 58:179–199, 1993.

[8] H. Ratcheck and J. Ronke. *New computer methods for global optimization.* Ellis Horwood Limited, Chichester, England, 1988.

[9] Eldon Hansen. *Global optimization using interval analysis*, volume 165 of *Pure and Applied Mathematics.* Marcel Dekker, Inc., 270 Madison Avenue, New York, New York 10016, 1992.

[10] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:621–680, 1993.

[11] F. Aluffi-Pentini, V. Parisi, and F. Zirilli. Global optimization and stochastic differential equations. *Journal of Optimization Theory and Applications*, 47(1):1–16, September 1985.

[12] Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal of Control and Optimization*, 24(5):1031–1043, September 1986.

[13] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of optimization theory and application*, pages 157–181, 1993.

[14] R. G. Strongin. On the convergence of an algorithm for finding a global extremum. *Engineering Cybernetics*, pages 549–555, 1973.

[15] Yaroslav D. Sergeyev. A global optimization algorithm using derivatives and local tuning. Technical Report 1, Istituto per la sistemistica e L'Informatica, 1994.

[16] F. J. Solis and R. J-B. Wets. Minimization by random search techniques. *Mathematics of Operation Research*, 6:19–30, 1981.

[17] P. Hansen, Jaumard B., and Lu S-H. On using estimates of Lipschitz constants in global optimization. *Journal of Optimization Theory and Applications*, 75:195–200, 1992.

[18] Aimo Törn and Antanas Žilinskas. *Global Optimization*, volume 350 of *Lecture Notes in Computer Science.* Springer-Verlag, Berlin Heidelberg, 1989.

[19] John G. Hocking and Gail S. Young. *Topology.* Addison-Wesley Publishing Company, Inc., 1961.

[20] Bruce Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, May 1988.