Stochastic global optimization

Anatoly Zhigljavsky,
School of Mathematics, Cardiff University, Cardiff, U.K.

Stochastic global optimization methods are methods for solving a global optimization problem incorporating probabilistic (stochastic) elements, either in the problem data (the objective function, the constraints, etc.), or in the algorithm itself, or in both.

Global optimization is a very important part of applied mathematics and computer science. The importance of global optimization is primarily related to the applied areas such as engineering, computational chemistry, finance and medicine amongst many other fields. For the state of the art in the theory and methodology of global optimization we refer to the 'Journal of Global Optimization' and two volumes of the 'Handbook of Global Optimization' [1,2]. If the objective function is given as a 'black box' computer code, the optimization problem is especially difficult. Stochastic approaches can often deal with problems of this kind much easier and more efficiently than the deterministic algorithms.

The problem of global minimization. Consider a general minimization problem $f(x) \to \min_{x \in X}$ with objective function $f(\cdot)$ and feasible region X. Let x^* be a global minimizer of $f(\cdot)$; that is, x^* is a point in X such that $f(x^*) = f_*$ where $f_* = \min_{x \in X} f(x)$. Global optimization problems are usually formulated so that the structure of the feasible region X is relatively simple; this can be done on the expense of increased complexity of the objective function.

A global minimization algorithm is a rule for constructing a sequence of points $x_1, x_2, \ldots$ in X such that the sequence of record values $y_{on} = \min_{i=1\ldots n} f(x_i)$ approaches the minimum f_* as n increases. In addition to approximating the minimal value f_*, one often needs to approximate at least one of the minimizers x_*.

Heuristics. Many stochastic optimization algorithms where randomness is involved have been proposed heuristically. Some of these algorithms are based on analogies with natural processes; the well-known examples are evolutionary algorithms [3] and simulated annealing [4]. Heuristic global optimization algorithms are very popular in applications, especially in discrete optimization problems. Unfortunately, there is a large gap between practical efficiency of stochastic global optimization algorithms and their theoretical rigor.

Stochastic assumptions about the objective function. In deterministic global optimization, Lipschitz-type conditions on the objective function are heavily exploited. Much research have been done in stochastic global optimization where stochastic assumptions about the objective function are used in a manner similar to how the Lipschitz condition is used in deterministic algorithms. A typical example of a stochastic assumption of this kind is the postulation that $f(\cdot)$ is a realization of a certain stochastic process. This part of stochastic optimization is well described in [5], Chapter 4 and will not be pursued in this article.

Global random search (GRS). The main research in stochastic global optimization deals with the so-called 'global random search' (GRS) algorithms which involve random decisions in the process of choosing the observation points. A general GRS algorithm assumes that a sequence of random points $x_1, x_2, \ldots, x_n$ is generated where for each $j \geqslant 1$ the point x_j has some probability distribution P_j. For each $j \geqslant 2$, the distribution P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on the results of the objective function evaluations at these points (the function evaluations may not be noise-free). The number of points n, $1 \leqslant n \leq \infty$ (the stopping rule) can be either deterministic or random and may depend on the results of function evaluation at the points $x_1, \ldots, x_n$.

Three important classes of GRS algorithms. In the algorithm which is often called 'pure random search' (PRS) all the distributions P_j are the same (that is, $P_j = P$ for all j) and the points x_j are independent. In Markovian algorithms the distribution P_j depends only on the previous point x_{j-1} and $f(x_{j-1})$, the objective function value at x_{j-1}. In the so-called population-based algorithms the distributions P_j are updated only after a certain number of points with previous distribution have been generated.

Attractive features of GRS. GRS algorithms are very popular in both theory and practice. Their popularity is owed to several attractive features that many global random search algorithms share: (a) the structure of GRS algorithms is usually simple; (b) these algorithms are often rather insensitive to the irregularity of the objective function behaviour, to the shape of the feasible region, to the presence of noise in the objective function evaluations, and even to the growth of dimensionality; (c) it is very easy to construct GRS algorithms guaranteeing theoretical convergence.

Drawbacks of GRS. Firstly, the practical efficiency of the algorithms often depends on a number of parameters, but the problem of the choice of these parameters frequently has little relevance to the theoretical results concerning the convergence of the algorithms. Secondly, for many global random search algorithms an analysis on good parameter values is lacking or just impossible. Thirdly, the convergence rate can be painfully slow, see discussion below. Improving the convergence rate (or efficiency of the algorithms) is a problem that much research in the theory of global random search is devoted to.

Main principles of GRS. A very large number of specific global random search algorithms exist, but only a few main principles form their basis. These principles can be summarized as follows: (i) random sampling of points at which $f(\cdot)$ is evaluated, (ii) random covering of the space, (iii) combination with local optimization techniques, (iv) the use of different heuristics including cluster-analysis techniques to avoid clumping of points around a particular local minima, (v) Markovian construction of algorithms, (vi) more frequent selection of new trial points in the vicinity of 'good' previous points, (vii) use of statistical inference, and (viii) decrease of randomness in the selection rules for the trial points. In constructing a particular global random search method, one usually incorporates several of these principles, see [5] where all these principles are carefully considered.

Convergence of GRS. To establish the convergence of a particular GRS algorithm, the classical Borel-Cantelli theorem is usually used. The corresponding result can be formulated as follows, see [5], Theorem 2.1. Assume that $X \subseteq \mathbb{R}^d$ with $0 < \mathrm{vol}(X) < \infty$ and $\sum_{j=1}^{\infty} \inf P_j(B(x,\varepsilon)) = \infty$ for all $x \in X$ and $\varepsilon > 0$, where $B(x,\varepsilon) = \{y \in X : ||y-x||_2 \leq \varepsilon\}$ and the infimum is taken over all possible locations of previous points $x_1, \ldots, x_{j-1}$ and the results of the objective function evaluations at these points. Then with probability one, the sequence of points $x_1, x_2, \ldots$ falls infinitely often into any fixed neighbourhood of any global minimizer.

In practice, a very popular rule for selecting the sequence of probability measures P_j is $P_j = \alpha_j P_0 + (1-\alpha_j)Q_j$, where $0 \leq \alpha_j \leqslant 1$, P_0 is the uniform distribution on X and Q_j is an arbitrary probability measure on X. In this case, the corresponding GRS algorithm converges if $\sum_{j=1}^{\infty} \alpha_j = \infty$.

Rate of convergence of PRS. Assume $X \subseteq \mathbb{R}^d$ with $\mathrm{vol}(X) = 1$ and the points $x_1, x_2, \ldots, x_n$ are independent and have uniform distribution on X (that is, GRS algorithm is PRS). The rate of convergence of PRS to the minimizer x_* is the fastest possible (for the worst continuous objective function) among all GRS algorithms. To guarantee that PRS reaches the ε-neighbourhood $B(x_*, \varepsilon)$ of a point x_* with probability at least $1-\gamma$, we need to perform at least $n_* = \lceil -\log(\gamma)\cdot\Gamma\left(\frac{d}{2}+1\right)/(\pi^{\frac{d}{2}}\varepsilon^d)\rceil$ iterations, where $\Gamma(\cdot)$ is the Gamma-function. This may be a very large number even for reasonable values of d, ε and γ. For example, if $d = 10$ and $\varepsilon = \gamma = 0.1$ then $n_* \simeq 0.9\cdot 10^{10}$. See Sect. 2.2.2 in [5] for an extensive discussion on convergence and convergence rates of PRS and other GRS algorithms.

Markovian GRS algorithms. In a Markovian GRS algorithm, the distribution P_j depends only on the previous point x_{j-1} and its function value $f(x_{j-1})$; that is, the sequence of points $x_1, x_2, \ldots$ constitutes a Markov chain. The most known Markovian GRS algorithms are the simulated annealing methods [4]. If a particular simulated annealing method creates a time-homogeneous Markov chain then the corresponding stationary distribution of this Markov chain is called Gibbs distribution. Parameters of the simulated annealing can be chosen so that the related Gibbs distribution is concentrated in a narrow neighbourhood of the global minimizer x_*. The convergence to the Gibbs distribution can be very slow resulting in a slow convergence of the corresponding simulated annealing algorithm. The convergence of all Markovian GRS algorithms is generally slow as the information about the objective function obtained during the search process is used ineffectively.

Population-based methods. Population-based methods are very popular in practice [3]. These methods generalize the Markovian GRS algorithms in the following way: rather than to allow the distribution P_j of the next point x_j to depend on the previous point x_{j-1}, it is now the distribution of a population of points (descendants, or children) depends on the previous population of points (parents) and the objective function values at these points. There are many heuristic arguments associated with these methods [3]. There are also various probabilistic models of the population-based algorithms [6].

Statistical inference in GRS. The use of statistical procedures can significantly accelerate the convergence of GRS algorithms. Statistical procedures can be especially useful for defining the stopping rules and the population sizes in the population-based algorithms. These statistical procedures are based on the use of the asymptotic theory of extreme order statistics and the related theory of record moments. As an example, consider PRS and the corresponding sample $S = \{f(x_j), j = 1, \ldots, n\}$. This is an independent sample of values from the distribution with c.d.f. $F(t) = \int_{f(x)\leq t} P(dx)$ and the support $[f_*, f^*]$, where $f^* = \sup_{x\in X} f(x)$. It can be shown that under mild conditions on f and P, this distribution belongs to the domain of attraction of the Weibull distribution, one of the extreme value distributions. Based on this fact, one can construct efficient statistical procedures for f_* using several minimal order statistics from the sample S.

For the theory, methodology and the use of probabilistic models and statistical inference in GRS, we refer to [5] and [6].

References

[1] R.Horst R. and P.Pardalos P., eds. (1995) Handbook of global optimization, Kluwer Acad. Publ., Dordrecht.

[2] Pardalos P. and Romeijn E., eds. (2002) Handbook of global optimization, Vol. 2, Kluwer Acad. Publ., Dordrecht.

[3] Glover F. and Kochenberger G.A. (2003) Handbook on metaheuristics, Kluwer Acad. Publ., Dordrecht.

[4] Van Laarhoven P.J.M. and Aarts E.H.L. (1987) Simulated annealing: theory and applications, D. Reidel Publishing Co., Dordrecht.

[5] Zhigljavsky A. and Zilinskas A. (2008) Stochastic global optimization, Springer, N.Y.

[6] Zhigljavsky A. (1991) Theory of global random search, Kluwer Acad. Publ., Dordrecht.