Stationary probability density of stochastic search processes in global optimization

Arturo Berrones

Posgrado en Ingeniería de Sistemas, Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León AP 126, Cd. Universitaria, San Nicolás de los Garza, NL 66450, Mexico
E-mail: arturo@yalma.fime.uanl.mx

Received 18 October 2007
Accepted 7 December 2007
Published 18 January 2008

Abstract. A method for the construction of approximate analytical expressions for the stationary marginal densities of general stochastic search processes is proposed. By the marginal densities, regions of the search space that with high probability contain the global optima can be readily defined. The density estimation procedure involves a controlled number of linear operations, with a computational cost per iteration that grows linearly with problem size.

Keywords: heuristics, stochastic search

ArXiv ePrint: 0710.3561v1
1. Diffusion and global optimization

Stochastic strategies for optimization are essential to many of the heuristic techniques used to deal with complex, unstructured global optimization problems. Methods like simulated annealing [1]–[4] and evolutionary algorithms [5]–[7], have proven to be valuable tools, capable of give good quality solutions at a relatively small computational effort. In spite of their success, these approaches present a major drawback, namely the absence of valid bounds on the obtained solutions. A common feature of deterministic global optimization algorithms is the progressive reduction of the domain space until the global optimum has been found with arbitrary accuracy [8, 9]. An analogous property for stochastic algorithms has been largely lacking. In this contribution is introduced a method for the estimation of the asymptotic probability density of a general stochastic search process in global optimization problems. The convergence of the estimated density can be clearly assessed, and with the help of this density, reliable bounds for the location of the global optimum are derived. The procedure involves linear operations only, and a well defined number of evaluations of the given cost function. Although the method is in first instance defined for continuous problems, discrete cost functions can be treated by the addition of suitable constraints. The presented results indicate that by the proposed approach, regions of the search space can be discarded on a probabilistic basis. This property may be implemented in a variety of ways in order to improve existing or develop new optimization algorithms, and open the door for the construction of probabilistic optimality certificates in large scale non-linear optimization problems.

The roots of stochastic search methods can be traced back to the Metropolis algorithm [10], introduced in the early days of scientific computing to simulate the evolution of a physical system to thermal equilibrium. This process is the base of the simulated annealing technique [1], which makes use of the convergence to a global minimum in configurational energy observed in physical systems at thermal equilibrium as the temperature goes to zero. The method presented in this contribution is rooted in similar physical principles as those on which simulated annealing and related algorithms [1, 11, 2, 3] are based. However, in contrast with other approaches, the proposed method considers a density of points instead of Markov transitions of individual points. Moreover, the main goal of the proposed approach is not the convergence to global minima.
as a randomness parameter is reduced, but the approximation of the probability density after an infinitely long exploration time of the search space, keeping a fixed randomness.

Consider the minimization of a cost function of the form $V(x_1, x_2, \ldots, x_n, \ldots, x_N)$ with a search space defined over $L_{1,n} \leq x_n \leq L_{2,n}$. A stochastic search process for this problem is modeled by

$$\dot{x}_n = -\frac{\partial V}{\partial x_n} + \varepsilon(t),$$

(1)

where $\varepsilon(t)$ is an additive noise with zero mean. Equation (1), known as Langevin equation in the statistical physics literature [12,13], captures the basic properties of a general stochastic search strategy. Under an uncorrelated Gaussian noise with constant strength, equation (1) represents a search by diffusion, while a noise strength that is slowly varying in time gives a simulated annealing. Notice that when choosing an external noise of infinite amplitude, the dynamical influence of the cost function over the exploration process is lost, leading to a blind search. The model given by equation (1) can be also interpreted as an overdamped non-linear dynamical system composed by $N$ interacting particles.

The temporal evolution of the probability density of such a system in the presence of an additive Gaussian white noise, is described by a linear differential equation, the Fokker–Planck equation [12,13],

$$\dot{p} = \sum_{n=1}^{N} \frac{\partial}{\partial x_n} \left[ \frac{\partial V}{\partial x_n} p \right] + \sum_{n=1}^{N} \sum_{m=1}^{N} D_{n,m} \frac{\partial^2 p}{\partial x_n \partial x_m}$$

(2)

with diffusion matrix $D_{n,m} = D\delta_{n,m}$, where $D$ is a constant that is proportional to the noise strength. The direct use of equation (2) for optimization or deviate generation purposes would imply the calculation of high dimensional integrals. It results numerically much less demanding to perform the following one dimensional projection of equation (2). Consider the conditional density

$$p(x_n|x_{j\neq n}) = \frac{p(x_1, \ldots, x_N)}{\int p(x_{j\neq n})|x_n)p(x_n) \, dx_n},$$

(3)

which is related to the marginal $p(x_n)$ by

$$p(x_n) = \int p(x_n|x_{j\neq n})p(x_{j\neq n}) \, dx_{j\neq n}. $$

(4)

Under very general conditions (e.g., the absence of infinite cost values), the equation (2) has a stationary solution over a search space with reflecting boundaries [12,14]. Introducing equation (3) into (2), and considering the variables $\{x_{j\neq n} = x_j^*\}$ fixed, it turns out that the stationary conditional probability density satisfy the one dimensional Fokker–Planck equation

$$D \frac{\partial p(x_n|x_{j\neq n} = x_j^*)}{\partial x_n} + p(x_n|x_{j\neq n} = x_j^*) \frac{\partial V}{\partial x_n} = 0.$$ 

(5)

An important consequence of equation (5) is that the marginal $p(x_n)$ can be sampled by drawing points from the conditional $p(x_n|x_{j\neq n} = x_j^*)$ via a Gibbs sampling [15]. It is now shown how, due to the linearity of the Fokker–Planck equation, a particular form of Gibbs sampling can be constructed, such that it is not only possible to sample the marginal

doi:10.1088/1742-5468/2008/01/P01013
density, but to give an approximate analytical expression for it. From equation (5) follows a linear second order differential equation for the cumulative distribution \( y(x_{j\neq n} = x_j^*) \) in a family of well known constrained NP-hard problems.

\[
\frac{d^2 y}{dx_n^2} + \frac{1}{D} \frac{\partial V}{\partial x_n} \frac{dy}{dx_n} = 0, \\
y(L_{1,n}) = 0, \quad y(L_{2,n}) = 1.
\] (6)

Random deviates can be drawn from the density \( p(x_n|x_{j\neq n} = x_j^*) \) by the inversion method [16], based on the fact that \( y \) is an uniformly distributed random variable in the interval \( y \in [0, 1] \). Viewed as a function of the random variable \( x_n \), \( y(x_n|x_{j\neq n}) \) can be approximated through a linear combination of functions from a complete set that satisfy the boundary conditions in the interval of interest,

\[
\hat{y}(x_n|x_{j\neq n}) = \sum_{l=1}^{L} a_{n,l} \varphi_l(x_n).
\] (7)

Choosing for instance, a basis in which \( \varphi_l(0) = 0 \), the \( L \) coefficients are uniquely defined by the evaluation of equation (6) in \( L - 1 \) interior points. In this way, the approximation of \( y \) is performed by solving a set of \( L \) linear algebraic equations, involving \( L - 1 \) evaluations of the derivative of \( V \).

The proposed procedure is based on the iteration of the following steps:

1. Fix the variables \( x_{j\neq n} = x_j^* \) and approximate \( y(x_n|x_{j\neq n}) \) by the use of formulae (6) and (7).

2. By the use of \( \hat{y}(x_n|x_{j\neq n}) \) construct a lookup table in order to generate a deviate \( x_n^* \) drawn from the stationary distribution \( p(x_n|x_{j\neq n} = x_j^*) \).

3. Update \( x_n = x_n^* \) and repeat the procedure for a new variable \( x_{j\neq n} \).

The iteration of the three steps above give an algorithm for the estimation of the equilibrium distribution of the stochastic search process described by equation (1). A convergent representation for \( p(x_n) \) is obtained after taking the average of the coefficients \( a \)'s in the expansion (7) over the iterations. In order to see this, notice that from equation (4) follows that the marginal \( y(x_n) \) is given by the expected value of the conditional \( y(x_n|x_{j\neq n}) \) over the set \( \{x_{j\neq n}\} \),

\[
y(x_n) = E_{\{x_{j\neq n}\}}[y(x_n|x_{j\neq n})].
\] (8)

All the information on the set \( \{x_{j\neq n}\} \) is stored in the coefficients of the expansion (7). Therefore

\[
\langle \hat{y} \rangle = \sum_{l=1}^{L} \langle a_{n,l} \rangle \varphi_l(x_n) \rightarrow y(x_n),
\] (9)

where the brackets represent the average over the iterations of the density estimation procedure.

Previous preliminary applications of the density estimation method on the generation of suitable populations of initial points for optimization algorithms can be found in [17]. In the next section the capabilities of the proposed algorithm for the construction of reliable probabilistic bounds is tested on several benchmark unconstrained examples and in a family of well known constrained NP-hard problems.

doi:10.1088/1742-5468/2008/01/P01013
2. Examples

The fundamental parameters of the density estimation procedure, $L$ and $D$, have a clear meaning, which is very helpful for their selection. The diffusion constant ‘smooth’ the density. This is evident by taking the limit $D \to \infty$ in equation (6), which imply an uniform density in the domain. The number of base functions $L$, on the other hand, defines the algorithm’s capability to ‘learn’ more or less complicated density structures. Therefore, for a given $D$, the number $L$ should be at least large enough to assure that the estimation algorithm will generate valid distributions $y(x_n|x_{j\neq n})$. A valid distribution should be a monotone increasing continuous function that satisfy the boundary conditions. The parameter $L$ ultimately determines the computational cost of the procedure, because at each iteration a system of size $\propto L$ of linear algebraic equations must be solved $N$ times. Therefore, the user is able to control the computational cost through the interplay of the two basic parameters: for a larger $D$ a smoother density should be estimated, so a lesser $L$ can be used.

The density estimation algorithm is tested on the following benchmark unconstrained problems:

Schwefel:

$N = 6, \quad f = 418.9829N - \sum_{n=1}^{N} x_n \sin(\sqrt{|x_n|}),$

$-500 \leq x_n \leq 500,$

solution: $x^* = (420.9687, \ldots, 420.9687), \quad f(x^*) = 0.$

Levy No. 5:

$N = 2, \quad f = \sum_{i=1}^{5} i \cos((i-1)x_1 + i) \sum_{j=1}^{5} j \cos((j+1)x_2 + j)$

$+ (x_1 + 1.42513)^2 + (x_2 + 0.80032)^2,$

$-10 \leq x_n \leq 10,$

solution: $x^* = (-1.3068, -1.4248), \quad f(x^*) = -176.1375.$

Booth:

$N = 2, \quad f = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2,$

$-10 \leq x_n \leq 10,$

solution: $x^* = (1, 3), \quad f(x^*) = 0.$

Colville:

$N = 4, \quad f = 100(x_2 - x_1)^2 + (1 - x_1)^2 + 90(x_4 - x_3)^2 + (1 - x_3)^2$

$+ 10.1((x_2 - 1)^2 + (x_4 - 1)^2) + 19.8(x_2 - 1)(x_4 - 1),$  

$-10 \leq x_n \leq 10,$

solution: $x^* = (1, \ldots, 1), \quad f(x^*) = 0.$

doi:10.1088/1742-5468/2008/01/P01013
Stationary probability density of stochastic search processes in global optimization

Figure 1. Density estimation for the variable $x_1$ of the Schwefel function. Two different diffusion constants have been considered, taking $L = 100$ in both cases. In the two cases the density clearly represents the structure of the cost function. The density is sharply peaked around the optimal value for the lesser $D$.

Rosenbrock:

$N = 20, \quad f = \sum_{n=1}^{N} 100(x_{n+1}^2 - x_n^2)^2 + (x_n - 1)^2,$

$-10 \leq x_n \leq 10,$

solution: $x^* = (1, \ldots, 1), \quad f(x^*) = 0$.

For the experiments, the following specific form of the expansion (7) has been used,

$\hat{y} = \sum_{l=1}^{L} a_l \sin \left( (2l - 1) \frac{\pi}{L} (x_n - L_1) \right), $  \hspace{1cm} (10)

so the size of the algebraic system of equations is $L - 1$. The linear system has been solved by a LU decomposition routine [18]. The cost function gradients have been calculated numerically, with two function evaluations per derivative. In this way, the total number of cost function evaluations per iteration goes like $2(L - 1)N$. The monotonicity of $\hat{y}$ is approximately accomplished through a large enough $L$ under expansion (10). It should be remarked that the choice of the basis functions is rather arbitrary, and other complete sets may be considered (e.g., Hermite polynomials). The Fourier basis is used here only because of its relative simplicity.

In figure 1 two different pairs $L, D$ have been considered in the study of the Schwefel problem. This problem has a second best minimum at a relatively large distance of the global optimum. This is reflected on the estimated densities (which have been calculated analytically by taking derivatives on equation (10)), but at small $D$ a clear distinction between the two regions is made. The Schwefel problem is an example of a separable function, that is, a function given by a linear combination of terms, where each term involves a single variable. Separable
Figure 2. Probability densities associated to the Levy No. 5 problem, using the parameter values $L = 200$, $D = 70$ and a number of iterations $M = 300$. The densities maxima are at coordinates $(-1.3, -1.42)$.

problems generate an uncoupled dynamics of the stochastic search described by equation (1). Because of this fact, the estimation algorithm converges in only one iteration for separable problems. The Schwefel example also illustrates that the density estimation algorithm works well on functions that are not derivable in some points. This is a consequence of the finite number of gradient evaluations required by the procedure.

In contrast to the Schwefel function, the stochastic search process associated to the Levy No 5 problem represents a coupled non-linear dynamics. Despite that this problem has about 760 local minima [19], the estimation algorithm shows good convergence properties, as is illustrated in figure 2.

The two previous examples show how, once that the estimation algorithm as attained convergence, is possible to define a region of the search space in which with high probability the global optimum is located. This concept is more systematically studied through the introduction of normalized distances. The distance normalized with respect to the search space $L_{1,n} \leq x_n \leq L_{2,n}$ between two points $x$ and $x^*$ is

\[
doi:10.1088/1742-5468/2008/01/P01013
\]
Density estimation for the Booth problem. Semi-log scale has been used.

defined by

\[
\text{distance} = \sqrt{\frac{(x_1 - x_1^*)^2 + \cdots + (x_N - x_N^*)^2}{(L_{1,1} - L_{2,1})^2 + \cdots + (L_{1,N} - L_{2,N})^2}}. \tag{11}
\]

Two measures written in terms of normalized distances are presented in the examples of figures 3–5: (i) the distance between the global optimum and the point in which the density is maximum. (ii) The length of the 95% probability interval around the point of maximum probability.

Under general conditions a Gibbs sampling displays geometric convergence [20]. In the presented experiments the running time has been chosen such that the 95% probability interval differ in less than 0.01 between successive iterations. Additionally, several control runs from different and independent starting conditions have been performed,

doi:10.1088/1742-5468/2008/01/P01013
indicating convergence to the same corresponding region within the predefined accuracy. As expected for a Gibbs sampling, the density appears to contract to a region of space that is independent of the starting point [21]. The numerical realizations suggest that convergence is attained at a few hundreds of iterations for all of the examples, even for the 20-dimensional Rosenbrock problem, which as been reported to be difficult to solve by stochastic heuristics like genetic algorithms [22].

The numerical experiments show that the global optimum is contained in a region close to the point of maximum probability, and that this region gets more sharp as $D$ decreases. A straightforward application of this behavior would be, for instance, on simulated annealing-type algorithms. Starting with a large diffusion constant, search regions can be iteratively discarded. By the use of the density estimation method, a probability measure is associated with each region. In this way the user can define a certain level of precision in the search. Several statistical quantities can be readily
calculated like measures of confidence, for instance probability intervals or characteristic fluctuation sizes.

Because the estimation algorithm depends on linear operations only, additional non-linearities in the cost function can be treated with essentially the same efficiency, giving more freedom and flexibility in modeling. For instance, the application of the density estimation algorithm to constrained problems can be done in a very direct manner through the addition of suitable ‘energy barriers’ (or more precisely, ‘force barriers’). These barriers do not need to be very large. Their main purpose is not to define prohibited regions, but only regions with low probability. The original constrained problem is transformed to an unconstrained cost function with additional non-linearities. Of course, the design of adequate barriers may be a difficult problem-dependent task. However, at least for some problems the approach seems to be straightforward. This is illustrated on the classical NP-hard knapsack problem [23], which is defined as follows. Given a set of \( N \) items, each item having a profit \( q_n \) and a weight \( w_n \), select a subset of the items such that the total profit is maximized, while the total weight is below a capacity \( c \). The usual formulation of this problem is given in terms of binary variables,

\[
\begin{align*}
\max & \sum_{n=1}^{N} q_n x_n \\
\text{s.t.} & \sum_{n=1}^{N} w_n x_n \leq c, \\
x_n & \in \{0, 1\}.
\end{align*}
\] (12)

It is well known that many standard instances of the knapsack model can be efficiently solved by exact methods [23], which makes it an ideal example for experimentation with the density estimation algorithm. The knapsack problem can be formulated in terms of continuous variables by introducing a non-linear constraint,

\[
\begin{align*}
\min \quad & -\sum_{n=1}^{N} q_n x_n \\
\text{s.t.} & \sum_{n=1}^{N} w_n x_n \leq c, \\
& -x_n^2 + x_n \leq 0, \quad 0 \leq x_n \leq 1.
\end{align*}
\] (13)

The quadratic constraint is equivalent to the restriction to binary variables. The following transformation is proposed,

\[
\begin{align*}
\min \quad & -\sum_{n=1}^{N} q_n x_n + k_0 \sum_{n=1}^{N} \frac{1}{1 + \exp(-b_0[-x_n^2 + x_n])} + k_1 \frac{\exp\left( b_1 \sum_{n=1}^{N} w_n x_n - c \right)}{\exp\left( -b_2 \sum_{n=1}^{N} w_n x_n - c \right) + 1} - 1 \\
\text{s.t.} \quad & 0 \leq x_n \leq 1.
\end{align*}
\] (14)

For illustrative purposes, consider the instance \( q = (2, 3, 5), w = (3, 5, 7), c = 10 \) of the knapsack problem. By inspection, the solution is given by \( x = (1, 0, 1) \). In figure 6 typical densities produced by the estimation algorithm for this instance are shown. The selection...
Stationary probability density of stochastic search processes in global optimization

Figure 6. Probability densities associated to a three dimensional instance of a knapsack problem.

of the parameters has been done after the performance of short runs, measuring the effects of each of the non-linear terms. The parameters have been tuned such that: (a) the first non-linear term alone produce symmetric densities peaked in the neighborhood of \{0, 1\}. (b) The addition of the second non-linear term and the original cost function generate densities in which the configurations with maximum probability satisfy the \(\sum_{n=1}^{N} w_n x_n \leq c\) constraint. Notice that the configuration that corresponds to the global optimum has the maximum probability.

A larger example is presented in figure 7. The exact solution has been calculated with the branch and bound algorithm supplied in the GNU Linear Programming Kit [24]. The instance has been generated by taking \(w_n\) uniformly distributed in an interval \([1, R]\), and

\[
q_n = w_n + (\varepsilon_n - 1) \frac{R}{100} + \frac{R}{10},
\]

\(\varepsilon_n\) uniform deviate in \([-1, 1]\), (15)

which imply strong linear correlations between \(q_n\) and \(w_n\). Instances of this kind are relevant to real management problems in which the return of an investment is proportional to the sum invested within small variations [23]. It has been argued that these type of instances fall in a category which is close to the ‘worst case’ scenario for exact
Stationary probability density of stochastic search processes in global optimization

Figure 7. The density estimation of an instance of a knapsack problem with 30 variables.

Table 1. Density estimation on instances of the knapsack problem with 30 variables. The final two columns indicate the resulting 95% intervals and gaps to optimum like normalized distances/flips.

| $R$  | $c$  | $k_0$ | $k_1$ | $b_0$ | $b_1$ | $b_2$ | $M$  | $L$  | $D$  | Interval       | Gap   |
|------|------|-------|-------|-------|-------|-------|------|------|------|----------------|-------|
| 10   | 100  | 10    | 7.1   | 10    | 0.01  | $2b_1$| 300  | 100  | 1    | 0.425/5        | 0.185/1|
| 100  | 500  | 100   | 37.0  | 10    | 0.001 | $3b_1$| 300  | 100  | 15   | 0.432/6        | 0.363/4|
| 1000 | 3000 | 1000  | 315.0 | 10    | 0.0001| $3b_1$| 300  | 100  | 100  | 0.433/6        | 0.257/2|

algorithms [23]. Despite of that, the estimation method converge to densities in which the optimum is contained in a region with high probability. Moreover, from the definition of the normalized distance, follows that the closest integers to the elements of the vector that represent the point of maximum probability differ in $\sim 1$ positions from the exact solution.

Three independent realizations of the numerical experiment over instances generated by equation (15) have been performed, varying the orders of magnitude of $R$ and $c$. The results are summarized in table 1, indicating the number of $0 \leftrightarrow 1$ flips that the normalized distances imply.

It should be remarked that, although the instances of the knapsack model discussed in this section are quickly solvable by exact algorithms, no reference to the particular structure of the original problem has been used for the density estimation. In fact, the problem has been treated like a highly non-linear cost function of 30 variables.

3. Heuristics based on stationary density estimation

The potential benefits of the stationary density estimation algorithm as a tool for the construction of new heuristics for high dimensional global optimization problems is
illustrated through the following greedy random search procedure:

1. Run an iteration of the stationary density estimation algorithm.
2. An initial best point is given by the global maximum of the estimated density.
3. Define a population of \(N + 1\) points. One point is the current best solution and other is the current point with maximum probability density. The remaining \(N - 1\) points are randomly drawn from an uniform distribution centered around the best point. For each dimension, the corresponding uniform distribution has a length equal to the typical fluctuation size given by the estimated density.
4. Run a downhill simplex routine. The starting conditions are given by the simplex defined by the points generated at step (3) as vertices.
5. If from step (4) results a point which improves the best known objective value, then update the best point.
6. Run an iteration of the density estimation algorithm.
7. Go to step (3).

From an evolutionary perspective, the above procedure acts at two different levels. At a short time scale a finite population evolve locally in a very greedy fashion by means of the downhill simplex routine. On the other hand, at the end of each round of simplex steps the entire population is replaced, except for the best point. The large changes on the population composition are dictated by a long time scale dynamics, which is consistent with the learned information about the global cost landscape. This information is gained through the approximation of the long term statistical density of a diffusive search process.

The short time scale exploration of the solution space is dominated by the downhill simplex method, which is a deterministic search based on function evaluations of a simplex vertices [25]. In a \(N\) dimensional search space, a population of \(N + 1\) points defined by the corresponding \(N + 1\) vertices evolve under simple geometric transformations, namely reflection, expansion and contraction. At each iteration, a new trial point is generated by the image of the worst point in the simplex (reflection). If the new point is better than all other vertices, the simplex expands in its direction. If the trial point improves the worst point, a reflection from the new worst point is performed. A contraction step is made when the worst point is at least as good as the reflected point. In this way the simplex eventually surrounds a local minimum. In the experiments presented here, the implementation of the downhill simplex given by [18] has been used, with a fractional decrease of cost value of at least \(10 \times 10^{-4}\) as termination criteria. A maximum of 50,000 function evaluations in the downhill simplex routine is allowed.

From a given reference point \(\mathbf{x}^{(\text{best})}\) (step (3) of the greedy diffusive search described above), \(N - 1\) vertices of an initial simplex for each call to the downhill simplex routine are defined through the characteristic length scales \(\lambda_n\), as \(\mathbf{x}^{(i)} = (x_1^{(\text{best})} + \epsilon_1^{(i)}, \ldots, x_n^{(\text{best})} + \epsilon_n^{(i)}, \ldots, x_N^{(\text{best})} + \epsilon_N^{(i)})\) where \(i = 1, \ldots, N - 1\) is an index for the \(N - 1\) vertices constructed around the reference point and the \(\epsilon_n\)'s are independent random numbers uniformly distributed in the interval \([\frac{1}{2}\lambda_n, \frac{1}{2}\lambda_n]\). The estimated long term density provides a vertex (the point with maximum likelihood at the current stage) and other vertex is given by the current best point. The typical fluctuation sizes, which are given by the first two moments.
of the estimated density, provide a natural definition for the characteristic length scales,
\[ \lambda_n = \sqrt{\langle x_n^2 \rangle - \langle x_n \rangle^2}, \] (16)
where the brackets in this case indicate average with respect to the corresponding estimated marginal. Due to the simple form of the expansion (10), all the integrals over the variables domain that are needed for moment calculation are performed analytically. The resulting expressions are finite sums with $L$ terms.

For illustration purposes, consider the Rosenbrock test function. In figure 8 are presented some plots of the behavior of our greedy stochastic search for the Rosenbrock problem of $N = 20$ variables. The graphs represent the cost function values over successive iterations. Four samples from a total of 100 runs are plotted. The result of a version of the algorithm in which an uniform density over the search space is used instead of the estimated long term density is also plotted. The success in finding the optimum is defined by a gap size with the known global optimum lesser than 0.001. Each run consist of 100 iterations. Over the total number of runs of the greedy stochastic search, 90% have been successful, and the 100% of the runs outperform the search based on uniform distributions. An 80% of the runs have been successful in less than 48 iterations and 30% in less than 13 iterations. For the 13 iterations cases, an average number of 28 080 cost function evaluations was needed. It should be remarked that the 20-dimensional Rosenbrock test function has been reported to be extremely difficult to solve by randomized optimization algorithms. For instance, in an experiment similar to the one presented here, it has been reported in [26] a success rate of zero after 400 000 function evaluations for simulated annealing, cross-entropy and model reference Adaptive Search. On the other hand, in [22] is reported that genetic algorithms and scatter search methods are unable to succeed in the 20-dimensional Rosenbrock problem after 50 000 cost function evaluations.

A more exhaustive experimentation with possible heuristics based on the stationary density estimation algorithm is in progress.
4. Conclusion

The presented results strongly suggest that the proposed density estimation algorithm can be used to construct probabilistic bounds on the location of global optima for large classes of problems. The density estimation is performed in a well defined number of elementary operations. The developed theory and the numerical experiments indicate that any given desired precision on the bounds can be attained with some finite values of the basic parameters, performing a finite number of iterations. The total computational cost per iteration grows linearly with problem size. The algorithm estimates the marginal density of each separate variable, which makes it suitable for parallel implementation. These features make the proposed method a promising tool, opening the possibility of constructing probabilistic optimal certificates for large scale unstructured problems. Experimentation in this direction is in progress. On the other hand, the density estimation algorithm may be used to develop new heuristics or improve existing stochastic or deterministic algorithms.

Acknowledgment

This work was partially supported by the National Council of Science and Technology of Mexico under grant CONACYT J45702-A.

References

[1] Kirkpatrick S, Gelatt C D Jr and Vecchi M P, Optimization by simulated annealing, 1983 Science 220 671
[2] Gidas B, 1995 Topics in Contemporary Probability and its Applications (Prob. Stochastic Ser.) (Boca Raton, FL: CRC) pp 159–232
[3] Parpas P, Rustem B and Pistikopoulos E N, Linearly constrained global optimization and stochastic differential equations, 2006 J. Global Optim. 36 191
[4] Geman S and Hwang C R, Diffusions for global optimization, 1986 SIAM J. Control Optim. 24 1031
[5] Goldberg D, 1989 Genetic Algorithms in Search, Optimization and Machine Learning (Reading, MA: Addison-Wesley)
[6] Eiben A E and Smith J E, 2003 Introduction to Evolutionary Computing (Berlin: Springer)
[7] Pelikan M, Goldberg D E and Lobo F G, A survey of optimization by building and using probabilistic models, 2002 Comput. Optim. Appl. 21 5
[8] Floudas C A, 2000 Deterministic Global Optimization: Theory, Methods and Applications (Dordrecht: Kluwer)
[9] Horst R, Pardalos P M and Thoai N V, 1995 Introduction to Global Optimization (Dordrecht: Kluwer) pp 158–83
[10] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E, Equations of state calculations by fast computing machines, 1953 J. Chem. Phys. 21 1087
[11] Suykens J A K, Verhaege H and Vandewalle J, On-line learning Fokker–Planck machine, 1998 Neural Process. Lett. 7 81
[12] Risken H, 1984 The Fokker–Planck Equation (Berlin: Springer)
[13] Van Kampen N G, 1992 Stochastic Processes in Physics and Chemistry (Amsterdam: North-Holland)
[14] Grasman J and van Herwaarden O A, 1999 Asymptotic Methods for the Fokker–Planck Equation and the Exit Problem in Applications (Berlin: Springer)
[15] Geman S and Geman D, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, 1984 IEEE Trans. Pattern Anal. Mach. Intell. 6 721
[16] Devroye L, 1986 Non-Uniform Random Variate Generation (Berlin: Springer)
[17] Berroues A, 2007 IWANN 2007: Proc. of the 9th Int. Work-Conf. on Artificial Neural Networks (Springer Lecture Notes in Computer Science vol 4507) (Berlin: Springer) pp 1–8
[18] Press W, Teukolsky S, Vetterling W and Flannery B, 2005 Numerical Recipes in C++, the Art of Scientific Computing (Cambridge: Cambridge University Press)

doi:10.1088/1742-5468/2008/01/P01013
Stationary probability density of stochastic search processes in global optimization

[19] Parsopoulos K E and Vrahatis M N, *Recent approaches to global optimization problems through particle swarm optimization*, 2002 Nat. Comput. 1 235

[20] Roberts G O and Polson N G, *On the geometric convergence of the Gibbs sampler*, 1994 J. R. Stat. Soc. B 56 377

[21] Canty A, *Hypothesis tests of convergence in Markov chain Monte Carlo*, 1999 J. Comput. Graphical Stat. 8 93

[22] Laguna M and Martí R, *Experimental testing of advanced scatter search designs for global optimization of multimodal functions*, 2005 J. Global Optim. 33 235

[23] Pisinger D, *Where are the hard knapsack problems?*, 2005 Comput. Oper. Res. 32 2271

[24] http://www.gnu.org/software/glpk/

[25] Nelder J A and Mead R, *A simplex method for function minimization*, 1965 Comput. J. 7 308

[26] Hu J, Fu M and Marcus S, *A model reference adaptive search method for global optimization*, 2007 Oper. Res. 55 549

doi:10.1088/1742-5468/2008/01/P01013