A CONDITIONAL GAUSSIAN MARTINGALE ALGORITHM FOR GLOBAL OPTIMIZATION

MANUEL L. ESQUÍVEL

Abstract. A stochastic algorithm for determination of a global minimum of a real valued continuous function over a compact interval of the real line satisfying a mild convexity condition is presented, an implementation is used to annalize an example and the convergence of this algorithm is proved. The algorithm may be thought to belong to the *random search* class but although we use Gaussian distributions at each repetition, the mean is changed at each repetition to be the intermediate minimum found at the preceeding repetition and the standard deviation is halved from one repetition to the next. The convergence proof is quite simple relying on the fact that the sequence of intermediate random minima is shown to be an uniformly integrable conditional Gaussian martingale.

1. Introduction

Quite some attention has been recently devoted to stochastic algorithms as the set of more than 300 references in the textbook [Spall 03] testifies. Global optimization methods using randomized search strategies are object of a thorough synthetic presentation in [Zabinsky 03] which also presents applications of these methods to engineering problems. Negative results as in [Stephens et al. 98] show that overconfidence on the effectiveness of stochastic methods is not desirable but, nevertheless, it is natural to speculate that an adequate randomized algorithm can perform better than a deterministic one in global optimization, at least in most of the situations. Theoretical results as in [Yin 99] and [Shi et al. 99] indicate that stochastic algorithms may be thought to be as reliable as deterministic ones and efforts in order to find better performing algorithms continue to be pursued as in [Raphael et al. 03]. In this paper we present a new simple convergent stochastic algorithm with good implementation properties. The main feature of this algorithm allows to recover some interesting properties of other stochastic algorithms such as the clustering and adaptiveness properties simultaneously with the property of continuing to search the whole domain at each repetition, which is a feature of simulated annealing.

2. The martingale algorithm

The algorithm presented may be included, on a first approximation, in the class of random research methods as this class *consists of algorithms which generate a sequence of points in the feasible region following some prescribed probability distribution or sequence of probability distributions*, according to [Horst et al. 95, p. 835]. The main idea of the method studied here is to change, at each new repetition, the location and dispersion parameters of the probability Gaussian distribution in order to concentrate the points from which the new intermediate minimum will be selected in the region where there is a greater chance of

Date: 8th September 2004.

1991 *Mathematics Subject Classification.* 68W20, 90C15.

Key words and phrases. Global optimization, random search algorithms, algorithm convergence.

This work was done with partial support from the FCT (Fundação para a Ciência e Tecnologia), program POCTI (Portugal/FEDER-EU).

finding a global minimum not precluding, however, a new intermediate minimum to be found outside this region. We use a sequence of Gaussian distributions taking at each repetition the mean equal to the intermediate minimum found in the preceeding repetition and the standard deviation equal to half of the standard deviation taken in the preceeding repetition. Let us now briefly describe the algorithm postponing to the following section the convergence result. We want to find a global minimum of a real function f defined over a compact interval C having length c. From now on, $\mathcal{U}(C)$ will denote the uniform distribution over C and $\mathcal{N}(m, \sigma)$ denotes the Gaussian distribution with mean m and standard deviation σ. Using the presentation protocol suggested in [Pintér 96] the schematic algorithm follows.

Step 0 Set $j = 0$;
Step 1 Generate $x_1^0, x_2^0, \ldots, x_N^0$ from the uniform distribution over the domain C.
Step 2 Choose $y_0 = x_{i_0}^0$ such that $f(x_{i_0}^0) = \min\{f(x_i^0) : 1 \leq i \leq N\}$. Increment j.
Step 3 Generate $x_1^j, x_2^j, \ldots, x_N^j$ from the normal distribution $\mathcal{N}(y_{j-1}, c/2^j)$ with mean y_{j-1} and standard deviation $c/2^j$, c being the length of C.
Step 4 Choose $y_j = x_{i_0}^j$ such that $f(x_{i_0}^j) = \min\{f(x_i^j) : 1 \leq i \leq N\}$.
Step 5 Perform a stopping test and stop, or increment j and return to Step 3.

Observe that steps 1 and 2 are useful in order to choose a starting point for the algorithm in C. The repetitions of steps 3, 4 and 5, provide a sort of clustering of the random test points around the intermediate minimum found at the preceeding repetition.

This algorithm's core is easily implemented in a programming language allowing symbolic computation. For algorithm implementation purposes we may extend f to the whole real line by defining $f(x) = +\infty$ for $x \notin C$. The crude implementation that follows was done in Mathematica 5 and does not covers the stopping rule.

1. `{ptmin, ptmax, cont, rep}`
   The variables `ptmin, ptmax, cont, rep` denote the left and right extremum of the compact set C, the number of points that will be generated at random with different distributions and the number of repetitions, respectively.
2.
```
alea1 = Table[Random[Real, ptmin, ptmax], {i, 1, cont}]
T1 = Table[alea1[[i]], f[alea1[[i]]], {i, 1, cont}]
M1 = Select[Column[T1, 1], f[#] <= Min[Column[T1, 2]] &]
eMes = M1
```
   The variable `M1` denotes the point correspondent to y_0 in Step 2 of the algorithm above.
3.
```
For[i = 1, i ≤ rep, i++,
alea = Table[Random[NormalDistribution[eMes[[1]], (ptmax - ptmin)/2^i]], {j,
1, cont}];
Tes = Table[alea[[j]], f[alea[[j]]], {j, 1, cont}];
eMes = Select[Column[Tes, 1], f[#] <= Min[Column[Tes, 2]] &]]
eMes[[1]]
```
   The variable `eMes[[1]]` denotes the intermediate minimum correspondent to y_{rep} in Step 4 of the algorithm above.

Using this implementation some preliminary tests were performed. For instance, 1.30084 the global minimum of $f(x) = x^4 - 3x^2 + x$ in $[-2, 2]$ was found in 0.05 seconds in a Macintosh PowerBook G4, using samples of only 10 points for each of 12 repetitions.

3. On the Algorithm Convergence

The algorithm introduced converges to a global minimum under the hypothesis of continuity of the function f defined on a compact set and if there is a convex level set near the minimum value of f. The sequences of intermediate random minima, each one defined for an arbitrary number of trial points, are shown to be martingales having nice convergence properties.

Theorem 3.1. *Let f be a real valued continuous function defined over the reals and C, a compact interval in $\mathbb{R}$, a neighborhood of z the global minimum of f. Set $m = \min_{x\in C} f(x)$ and suppose that for some $\epsilon > m$, we have that $Z(\epsilon) = \{x \in C : f(x) < \epsilon\}$ is an interval. Consider $N \in \mathbb{N} - \{0\}$ fixed and define almost surely and recursively the sequence $(Y_j^N)_{j\in\mathbb{N}}$ by:*

$$Y_0^N := \left\{ X_{i_0}^0 : f(X_{i_0}^0) = \min_{1\leq i\leq N} \{f(X_i^0) : X_1^0, \ldots, X_N^0 \in \mathcal{U}(C) \textit{ i.i.d. } \} \right\} ;$$

and for all $j \geq 1$

$$Y_j^N := \left\{ X_{i_0}^j : f(X_{i_0}^j) = \min_{1\leq i\leq N} \{f(X_i^j) : X_1^j, \ldots, X_N^j \in \mathcal{N}(y, \frac{c}{2^j}) \textit{ if } Y_{j-1}^N(\omega) = y, \textit{ i.i.d.}\} \right\} .$$

Then, for all $N \geq 1$, $\lim_{j\to+\infty} Y_j^N = y^N \in \mathbb{R}$ almost surely and, $\lim_{j\to+\infty} y^N =: z$ defined to be a point of C such that $f(z) = min_{x\in C} f(x)$.

Proof. The proof goes along the following lines. We show that $(Y_j^N)_{j\in\mathbb{N}}$ is an uniformly integrable martingale which converges almost surely and in L^1 to a constant y^N. The sequence $(y^N)_{N\geq 1}$ may be supposed to converge to $y \in C$ and then, we will show that if $f(y) > f(z)$, and $Z(\epsilon)$ is a sufficiently small interval around z we should have with arbitrary high probability for N large enough, for all $i \in \{1, \ldots, N\}$ and for all j large enough that $X_i^j \notin Z(\epsilon)$; this is impossible due to the laws of these random variables. For simplicity we will denote Y_j for Y_j^N whenever no confusion is expected.

Let us first establish convergence. For every $j \geq 1$ and $(y_1, \ldots, y_{j-1}) \in \mathbb{R}^{j-1}$ we have by construction that:

$$\mathbb{E}[Y_j \mid Y_1 = y_1, \ldots, Y_{j-1} = y_{j-1}] = \mathbb{E}[Y_j \mid Y_{j-1} = y_{j-1}] = y_{j-1} .$$

As a consequence

$$\mathbb{E}[Y_j \mid (Y_1, \ldots, Y_{j-1})] = \mathbb{E}[Y_j \mid Y_{j-1}] = Y_{j-1} ,$$

thus showing that $(Y_j)_{j\in\mathbb{N}}$ is a martingale with respect to the natural filtration $(\mathcal{F}_j)_{j\in\mathbb{N}}$ with $\mathcal{F}_j = \sigma(\{Y_i : i \leq j\})$. Denote by y the common value of $\mathbb{E}[Y_j]$ and observe that as the variance of this random variable is, by construction, given by $\mathbb{V}[Y_j] = (c/2^j)^2$, we have $\mathbb{E}[Y_j^2] = \mathbb{V}[Y_j] + y^2 = c^2/2^{2j} + y^2$. Being so, $\sup_{j\geq 1} \mathbb{E}[Y_j^2] \leq c^2/4 + y^2$ and we conclude that $(Y_j)_{j\in\mathbb{N}}$ is an uniformly integrable martingale.

Then, there exist an integrable random variable Y such that the martingale converges almost surely and in L^1 to Y and moreover for all $j \in \mathbb{N}$ we have $Y_j = \mathbb{E}[Y \mid \mathcal{F}_j]$. As $\mathbb{E}[Y_j] = \mathbb{E}[Y]$ and we have by Markov inequality that

$$\forall \epsilon > 0 \;\; \mathbb{P}\left[|Y_j - \mathbb{E}[Y]| > \epsilon\right] \leq \frac{\mathbb{V}[Y_j]}{\epsilon^2} = \frac{c^2}{2^{2j}\epsilon^2} ,$$

thus showing that $(Y_j)_{j\in\mathbb{N}}$ converges in probability to $\mathbb{E}[Y]$. As a consequence, there exists a subsequence of $(Y_j)_{j\in\mathbb{N}}$ converging almost surely to $\mathbb{E}[Y]$ and finally, as a result of the martingale almost surely convergence result, we may conclude that $Y = \mathbb{E}[Y] = y$.

Now remember that y depends on N. We then have a sequence $(y^N)_{N\geq 1}$ of points in C which is a compact set. We may suppose that there exists $y \in C$ such that $\lim_{N\to+\infty} y^N = y$, possibly by resorting to a subsequence.

Suppose now that $f(y) > f(z)$. We will derive a contradiction. By the hypotheses assumed to hold on f, take ϵ such that $f(z) < \epsilon < f(y)$ and suppose that $Z(\epsilon)$ is a level set of f verifying:

$$\forall x \in Z(\epsilon) :=]z-\epsilon, z+\epsilon[\ ,\ \forall u \in C - Z(\epsilon)\ \ f(x) < f(u)\ . \tag{3.1}$$

As $C - Z(\epsilon)$ is a neighborhood of y, consider also ϵ_1 such that with $Y(\epsilon_1) :=]y-\epsilon_1, y+\epsilon_1[$ we have $Z(\epsilon) \cap Y(\epsilon_1) = \emptyset$. Again, by Markov inequality, for any $j \geq 0$ and any $N \geq 1$:

$$\mathbb{P}\left[\bigcup_{l \geq j} \{\left|Y_l^N - y^N\right| > \epsilon_1/2\}\right] \leq \frac{4c^2}{\epsilon_1^2} \sum_{l \geq j} \frac{1}{2^{2l}} \xrightarrow[j_0 \to +\infty]{} 0\ .$$

Now, choose $1 > \delta > 0$ and $j_0 = j_0(\delta) \geq 1$ such that for $j \geq j_0$ and with $\Omega_j^N := \bigcap_{l \geq j} \{\left|Y_l^N - y^N\right| \leq \epsilon_1/2\}$ we have $\mathbb{P}[\Omega_j^N] \geq 1 - \delta$. Note that j_0 does not depend on N.

Let us make explicit a fundamental observation. Suppose given a $N \geq 1$ and that for any i, j, ω we have $X_i^j(\omega) \in Z(\epsilon)$. Then $f(Y_j(\omega)) \leq f(X_i^j(\omega))$ thus implying $Y_j(\omega) \in Z(\epsilon)$. Set $N_0 \geq 1$ such that for all $N \geq N_0$ we have $\left|y^N - y\right| < \epsilon_1/2$,

As a consequence, for any $N \geq N_0$, $i \in \{1, \ldots, N\}$, $j \geq j_0$ and $\omega \in \Omega_j^N$, we will have $X_i^j(\omega) \notin Z(\epsilon)$.

We will now consider $j_1 \geq j_0$ such that $\sqrt{2}c/2^{j_1} < d(Z(\epsilon), Y(\epsilon_1))$, the distance in $\mathbb{R}$ between $Z(\epsilon)$ and $Y(\epsilon_1)$ and define for any $N \geq N_0$, $\Omega^N := \Omega_{j_1}^N$. This implies that for all $j \geq j_1$ and $\omega \in \Omega^N$ the function $\phi(x) = \exp(-2^{2j}(Y_j^N(\omega) - x)^2/2c^2)$ is convex over $Z(\epsilon)$. We will need this fact below.

Choose an $N \geq N_0$ and consider for every $j \geq 0$ the vector of random variables $\mathcal{Y}_j^N := (Y_j^N, Y_{j+1}^N \ldots)$. We have the following chain of equalities and upper bounds.

(3.2)

$$1 - \delta \leq \mathbb{P}[\Omega^N] = \mathbb{E}[\mathbb{E}[\mathbb{I}_{\Omega^N} \mid \mathcal{Y}_{j_1}^N]] =_{(a)} \mathbb{E}\left[\mathbb{E}[\bigcap_{1 \leq i \leq N, j \geq j_1} \mathbb{I}_{\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N} \mid \mathcal{Y}_{j_1}^N]\right] =_{(b)}$$

$$= \mathbb{E}\left[\prod_{1 \leq i \leq N, j \geq j_1} \mathbb{E}[\mathbb{I}_{\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N} \mid Y_{j-1}^N]\right] =_{(c)} \prod_{j \geq j_1} \prod_{1 \leq i \leq N} \mathbb{E}\left[\mathbb{E}[\mathbb{I}_{\{X_1^j \notin Z(\epsilon)\} \cap \Omega^N} \mid Y_{j-1}^N]\right] =_{(d)}$$

$$= \prod_{j \geq j_1} \left(1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{-\frac{2^{2j}(y^N - x)^2}{2c^2}} dx\right)^N \leq_{(e)} \exp\left(-\frac{N}{c\sqrt{2\pi}} \sum_{j \geq j_1} 2^{2j} e^{-\frac{2^{2j}(y^N - x_j)^2}{2c^2}}\right)\ ,$$

where the equalities and inequalities may be justified as follows.

(a) As a consequence of the fundamental observation above.

(b) As the events $\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N$ for $1 \leq i \leq N, j \geq j_1$ are, by construction, conditionally independent given $\mathcal{Y}_{j_1}^N$ and then as we have $\mathbb{E}[\mathbb{I}_{\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N} \mid \mathcal{Y}_{j-1}^N] = \mathbb{E}[\mathbb{I}_{\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N} \mid Y_{j-1}^N]$ for all $1 \leq i \leq N, j \geq j_1$.

(c) As for $j \geq j_1$ the events $\{X_i^j \notin Z(\epsilon)\} \cap \Omega^N$ for $1 \leq i \leq N$ are, by construction, conditionally independent given $Y_{j_1}^N$ and, as for $j \geq j_1$ the random variables X_i^j for $1 \leq i \leq N$ are identically distributed.

(d) As we have, by construction,

$$\mathbb{P}[X_1^j \notin Z(\epsilon) \mid Y_{j-1}^N = u] = 1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{\frac{-2^{2j}(u-x)^2}{2c^2}} dx\ ,$$

we also have almost surely for $\omega \in \Omega^N$:

$$\mathbb{E}[\mathbb{1}_{\{X_1^j \notin Z(\epsilon)\} \cap \Omega^N} \mid Y_{j-1}^N](\omega) = 1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{\frac{-2^{2j}(Y_{j-1}^N(\omega)-x)^2}{2c^2}} dx \ .$$

Finally, as the function ϕ defined above is convex, as $1 > \mathbb{P}[\Omega^N] \geq 1 - \delta > 0$ and as $\mathbb{E}[Y_j^N] = y^N$ then, by Jensen's inequality and Fubini theorem we have:

$$\mathbb{E}[\mathbb{1}_{\{X_1^j \notin Z(\epsilon)\} \cap \Omega^N}] = \mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\{X_1^j \notin Z(\epsilon)\} \cap \Omega^N} \mid Y_{j-1}^N] \mathbb{1}_{\Omega^N}\right] =$$
$$= \mathbb{E}\left[\left(1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{\frac{-2^{2j}(Y_{j-1}^N(\omega)-x)^2}{2c^2}} dx\right) \mathbb{1}_{\Omega^N}\right] \leq$$
$$\leq \mathbb{P}[\Omega^N] \left(1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{\frac{-2^{2j}(y^N-x)^2}{2c^2}} dx\right) \leq \left(1 - \frac{2^j}{c\sqrt{2\pi}} \int_{Z(\epsilon)} e^{\frac{-2^{2j}(y^N-x)^2}{2c^2}} dx\right) .$$

(e) By choosing $x_j^N \in Z(\epsilon)$ such that $\int_{Z(\epsilon)} e^{\frac{-2^{2j}(y^N-x)^2}{2c^2}} dx = 2\epsilon \ e^{\frac{-2^{2j}(y^N-x_j^N)^2}{2c^2}}$ and by the fact that for $N \geq 1$ and $0 \leq x \leq 1$ we have $(1-x)^N \leq e^{-Nx}$.

We may now conclude. As $x_j^N \in Z(\epsilon)$ and $y^N \in Y(\epsilon_1)$, we have $(y^N - x_j^N)^2 \geq d(Z(\epsilon), Y(\epsilon_1))^2$ and so the series on the most right-hand side of formula of 3.2 converges. Letting N go to $+\infty$, as j_1 does not depend on N, we will have a contradiction and finally, we must have $f(z) = f(y)$. □

Remark 1. From the preceeding proof it is clear that the same ideas may be applied to a multidimensional situation in order to get an analogue of this theorem. Nevertheless, as the present proof relies heavily on the fact that there exists a level set near the minimum which is an interval, some other line of attack has to be chosen for the multidimensional case.

References

[Horst et al. 95] R. Horst, P. M. Pardalos (editors) *Handbook of Global Optimization.* Kluwer Academic Publishers 1995.

[Pintér 96] J. D. Pintér, *Global optimization in action. Continuous and Lipschitz optimization: algorithms, implementations and applications.*, Nonconvex Optimization and Its Applications. 6. Dordrecht: Kluwer Academic Publishers. xxvii 1996.

[Raphael et al. 03] B. Raphael, I.F.C. Smith, A direct stochastic algorithm for global search. *Appl. Math. Comput.* 146, No.2-3, 729-758 (2003).

[Shi et al. 99] D. Shi, J. Peng, A new theoretical framework for analyzing stochastic global optimization algorithms. *J. Shanghai Univ.* 3, No.3, 175-180 (1999).

[Spall 03] J. C. Spall, , *Introduction to stochastic search and optimization. Estimation, simulation, and control.*, Wiley-Interscience Series in Discrete Mathematics and Optimization. Hoboken, NJ: Wiley. xx, 2003.

[Stephens et al. 98] C.P. Stephens, W. Baritompa, Global optimization requires global information. *J. Optimization Theory Appl.* 96, No.3, 575–588 (1998).

[Yin 99] G. Yin, Rates of convergence for a class of global stochastic optimization algorithms. *SIAM J. Optim.* 10, No.10, 99–120 (1999).

[Zabinsky 03] Z. B. Zabinsky, , *Stochastic adaptive search for global optimization.* Nonconvex Optimization and Its Applications 72. Boston, MA: Kluwer Academic Publishers. xviii, 2003.

Departamento de Matemática, FCT/UNL, Quinta da Torre, 2829-516 Caparica, Portugal and CMAF/UL.

E-mail address: `mle@fct.unl.pt`

URL: `http://ferrari.dmat.fct.unl.pt/personal/mle`