Rigorous analysis of some simple adaptive ES

Anne Auger, Claude Le Bris, Marc Schoenauer

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Rigorous analysis of some simple adaptive ES

Anne Auger — Claude Le Bris — Marc Schoenauer

N° 4914

Août 2003

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4914--FR+ENG

INRIA
ROCQUENCOURT

Rigorous analysis of some simple adaptive ES

Anne Auger*† ‡ , Claude Le Bris† , Marc Schoenauer* §

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Fractales

Rapport de recherche n° 4914 — Août 2003 — 21 pages

Abstract: Based on the theory of non-negative supermartingales, convergence results are proven for adaptive $(1, \lambda) - ES$ with Gaussian mutations, and geometrical convergence rates are derived. In the d-dimensional case ($d > 1$), the algorithm studied here uses a different step-size update in each direction. However, the critical value for the step-size, and the resulting convergence rate do not depend on the dimension. Those results are discussed with respect to previous works. Thorough numerical investigations on some 1-dimensional functions validate the theoretical results.

Key-words: adaptive evolutionary algorithms, evolution strategies, martingales, convergence analysis, rate of convergence

* Projet Fractales, INRIA Rocquencourt, BP 105, 78153 LE CHESNAY Cedex, France
† CERMICS – ENPC, Cité Descartes, 77455 Marne-La-Vallée, France,{auger}{lebris}@cermics.enpc.fr
‡ Projet MIC MAC, INRIA Rocquencourt, BP 105, 78153 LE CHESNAY Cedex, France
§ Marc.Schoenauer@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Analyse de Stratégies d'Evolutions adaptatives

Résumé : Ce document présente l'étude mathématique d'une classe d'algorithmes d'optimisation stochastiques spécifiquement dédiés à l'optimisation réelle: les stratègies d'évolution. En utilisant la théorie des martingales, on prouve la convergence de $(1, \lambda) - ES$ adaptatifs dans le cas de fonctions à minimiser régulières. On montre aussi que les vitesses de convergence sont géométriques. Les résultats théoriques sont ensuite comparés et validés avec les expérimentations numériques de ces algorithmes.

Mots-clés : algorithmes évolutionnaires adaptatifs, stratégie d'évolution, martingales, convergence, vitesse de convergence

1 Introduction

Since their invention in the mid-sixties (see the seminal books by Rechenberg [7] and Schwefel [10]), Evolution Strategies have been thoroughly studied from the theoretical point of view.

Early studies on two very particular functions (the *sphere* and the *corridor*) have concerned the progress rate of the $(1+1)-ES$, and have lead, by extrapolation to any function, to the famous *one-fifth* rule. The huge body of work by Beyer, including many articles, and somehow summarized in his book [3], has pursued along similar lines, studying more general algorithm, from the full $(\mu \dagger \lambda) - ES$ to the $(\mu/\mu, \lambda) - ES$ with recombination and the $(1, \lambda) - \sigma - SA - ES$ with self-adaptation. However, though giving important insights about the way ES actually work, the study of local progress measures, such as the progress rate, does not lead to global convergence results of the algorithm.

Some global convergence results, together with the associated (geometrical) convergence rates have been obtained for convex functions [8, 13], and for a class of functions slightly more general than quadratic functions, the so-called $(Q-K)-$ *strongly convex* functions [9]. These latter results deal with the *adaptive* version of evolution strategies, in which the step-size is computed at each iteration according to some measure on the current population (the terminology used here is taken from [6]) – namely the norm of the gradient of the fitness function.

Note that the results in [13] have been criticized in [4], in which an analytical approach is provided in the case of the sphere function when the step-size is the norm of the parent itself. In that case, the strong law of large number gives an almost sure convergence.

The state-of-the-art in practical ES, however, recommends using *self-adaptive* ES, in which the step-size is adjusted by the evolution itself at the individual level. Whereas of course the results by Beyer on the $(1, \lambda)$-σ-SA-ES do address self-adaptive ES [3], only recently some global convergence results regarding self-adaptive ES-like algorithms were published [5, 11]. However, the algorithms studied in those works do not consider the standard normal mutation, but rather use a simplified mutation operator: only a finite number of variation of the step-size are allowed in [5], while [11] considers a uniform mutation. Moreover, these works only consider the simple and symmetrical function $f(x) = |x|$. Finally, [5] does not give any estimation of the convergence rate, and the proof in [11] relies on a numerical estimate of some inequality - though this might be improved in the near future. An important point about these two latter results is that they use the theory of super-martingales [12], a more sophisticated technique than all previously cited works (with the remarquable exception of [8]).

The same supermartingale technique is also used in the present article, to analyze some adaptive ES with Gaussian mutation, in which the step-size is adapted either using the distance to the global optimum or using gradient information about the fitness function, but in a different way than in [13, 9]. Moreover, the speed of convergence is also studied: as in previous relevant works, some geometrical upper-bounds are derived, and their sharpness are tested through numerical experiments.

The article is organized as follows. Next section formally describes the adaptive ES under study. We configure ES with an adaptivity that evolves more deterministically than in standard self adaptative ES (see formula (1) below). Section 3 gives the convergence results and speed of convergence results. First, the one-dimensional case is thoroughly studied: in the case of the sphere function analytical results are obtained, before two different ways of adapting the step-size are studied in turn for a more general class of functions. It is indeed to be noted that our proofs and techniques are not restricted to the specific cases we deal with here. Next, the optimality of the critical value of the step size and convergence rate obtained is proved for the sphere function. The case of larger dimension is finally presented. The originality is that we derive estimates of the convergence rate that do not depend on the dimension. The analysis is carried out on a specific algorithm where the step-size is adapted independently in each dimension.

In section 4, our results are thoroughly discussed, in the light of previous works on adaptive algorithm (already cited in the Introduction). Section 5 gives experimental evidences (in one dimension only) that demonstrate the validity of the critical value of the step size and of the convergence rate, for more general functions (such as functions that are neither symmetric nor convex). The article ends with some discussion and trends for future work.

Let us close this introduction by mentioning that most of the results proven here have been announced in [2].

2 Notations and algorithm

For the sake of simplicity, the results will first be presented in dimension 1. The case of higher dimensions will be introduced in Section 3.4. Let the function f to be minimized be a function defined on $\mathbb{R}$. The general adaptive $(1, \lambda)$-Evolution Strategy algorithm we will consider henceforth is of the form:

$$\left\{ \begin{array}{l} X^0 \in \mathbb{R}, \\ X^{n+1} = \arg\min\{f(X^n + \sigma H(X^n)\mathcal{N}_i^n), i \in [1, \lambda]\}, \end{array} \right. \tag{1}$$

where X_n is the random variable modeling the parent at the generation n, $(\mathcal{N}_i^n)_{i=1,\ldots,\lambda}$ are independent standard normal random variables. For conciseness, we shall only consider in the following two cases function H: $H(x) = |x|$ or $H(x) = |f'(x)|$. However, other cases, such as $H(x) = |f(x) - f^*|$ can be treated by the same technique, and will be detailed in see [1]. Parameter σ is a positive real parameter, often referred to as the *step-size* (or normalized step-size e.g. in [10, 3], in the case where $H(x) = |x|$).

This paper is concerned with studying the behavior of algorithm (1), or, more precisely, with addressing the issue of the range of values for σ for which the algorithm converges[1]. Moreover, whenever convergence takes place, bounds for the convergence rate will also be determined.

[1] Both *almost sure* convergence and convergence in L^p (w.r.t the norm $\mathbb{E}(|X|^p)^{\frac{1}{p}}$) will be examined.

Section 3 addresses the two aspects, first for the sphere function (Section 3.1 collects results from the literature, especially from [4]), since exact convergence rates can then be easily computed, and next for twice continuously differentiable functions (with some particular properties, see Assumption (H1)(H2)(H3) below) in the case $H(x) = |x|$ (Section 3.2) and $H(x) = |f'(x)|$ (Section 3.3).

3 Convergence results the $(1, \lambda)$-ES.

3.1 The sphere function

The sphere function ($f(X) = |X|^2$) is the preferred test function of authors studying the theory of Evolution Strategies [7, 10, 8, 3, 4, 5, 11]. Indeed, when f is the sphere function, things get simpler, and many interesting quantities can be computed analytically.

For instance, it is clear that both cases $H(x) = |x|$ and $H(x) = |f'(x)|$ behave identically (up to a factor 2). But another important simplification concerns the algorithm itself:

Lemma 1. *For the sphere function, the random variable X^n defined by (1) with $H(x) = |x|$ is such that*

$$X^{n+1} = X^n(1 + \sigma Y(\lambda)) \tag{2}$$

where the random variable $Y(\lambda)$ defined by

$$1 + \sigma Y(\lambda) = \arg\min\{(1 + \sigma N_1^n)^2, ..., (1 + \sigma N_\lambda^n)^2\} \tag{3}$$

does not depend on σ.

A detailed proof, with the exact distribution of $Y(\lambda)$ can be found in [4].

3.1.1 Convergence in L^p

The following theorem is an immediate consequence of Lemma 1:

Theorem 1. (Convergence in L^p) *For the sphere function, the random variable X^n defined by (1) with $H(x) = |x|$ satisfies*

$$\mathbb{E}(|X^n|^p) = \mathbb{E}(|X_0|^p)\mathbb{E}(|1 + \sigma Y(\lambda)|^p)^n. \tag{4}$$

Hence, the algorithm converges or diverges in L^p geometrically. Moreover, there exists a value $\sigma_c(\lambda, p)$ such that X^n converges in L^p if and only if $\sigma \in]0, \sigma_c(\lambda, p)]$. This value is defined by

$$\sigma_c(\lambda, p) = \inf\{\sigma \text{ such that } \mathbb{E}(|1 + \sigma Y(\lambda)|^p) \geq 1\}. \tag{5}$$

Remark 1. *It can be proved that $\mathbb{E}(|1 + \sigma Y(\lambda)|^p)$ has a unique minimum w.r.t. σ, which gives the best convergence rate. This minimum $\sigma_s(\lambda, p)$ is thus defined by*

$$\sigma_s(\lambda, p) = argmin\{\mathbb{E}(|1 + \sigma Y(\lambda)|^p), \sigma \in]0, \sigma_c(\lambda, p)[\}. \tag{6}$$

3.1.2 An alternative view on the progress rate

Interestingly, this result meets early studies of ES [7, 10, 3] that did look at the *progress rate* φ_p, defined by:

$$\varphi_p(X^n, \sigma, \lambda) = \mathbb{E}\left(\frac{|X^{n+1}|^p - |X^n|^p}{|X^n|^p} | X^n \right). \tag{7}$$

The progress rate measures the expectation of change from one iteration of the algorithm to the next one, conditionally to the current parent X^n (Note that this conditional dependency is often left implicit in the cited works. Those early works determine, for a given λ, the optimal step size σ which minimizes the progress rate. In general, this quantity depends on the current point X^n and will not be very useful to study the dynamics of the algorithm.

However, in the case of the sphere function, things are different. A direct consequence of Lemma 1 is that for the sphere function with $H(x) = |x|$, the progress rate does not depend on n, and hence is for instance equal to the value for $X^n = 1$:

$$\forall n > 0, \;\; \varphi_p(X^n, \sigma, \lambda) = \mathbb{E}(|1 + \sigma Y(\lambda)|^p - 1). \tag{8}$$

Hence, minimizing the progress rate as in [10, 3] thus amounts to finding the value of σ such that $\mathbb{E}(|1 + \sigma Y(\lambda)|^p)$ is minimal – and this is exactly the value given by equation (6).

3.1.3 Convergence almost surely

For the almost sure convergence, Lemma 1 and the strong law of large numbers gives the following result (see [4] for more details),

Theorem 2. (Convergence almost surely, [4]) *Assume that* $\mathbb{E}(\ln(|1 + \sigma Y(\lambda)|)) < \infty$. *Then, for the sphere function, the random variable* X^n *defined by (1) with* $H(x) = |x|$ *is such that*

$$\frac{1}{n} \ln(|X^n|) \xrightarrow[n \to \infty]{} \mathbb{E}(\ln(|1 + \sigma Y(\lambda)|)) \quad \textit{almost} \quad \textit{surely}.$$

Thus the critical value $\sigma_c(\lambda, as)$ is here defined as $\sup\{\sigma \backslash \mathbb{E}(\ln(|1 + \sigma Y(\lambda)|) < 1\}$.

The following two sections will prove similar results for more general functions, for each of the cases $H(x) = |x|$ and $H(x) = |f'(x)|$.

3.2 Convergence of the $(1, \lambda)$-ES with $H(x) = |x|$

The case where $H(x) = |x - x^*|$ for some minimizer $x*$ of f is the case with constant (normalized) step-size, as defined for instance in [3]. The algorithm under study here therefore reads from (1)

$$\begin{cases} X^0 \in \mathbb{R}, \\ X^{n+1} = \arg\min\{f(X^n + \sigma |X^n - x^*| \mathcal{N}_i^n), i \in [1, \lambda]\}, \end{cases} \tag{9}$$

We are aware that this algorithm may be considered to have a poor practical interest because it supposes that a minimizer $x*$ is already known. However, from the methodological viewpoint, its consideration will allow us to develop the technique of analysis to be applied later on to some more interesting cases: first here to the case $H(x) = |f'(x)|$, and second, in a forthcoming publication [1], to the case $H(x) = |f(x) - \inf f|$.

3.2.1 Convergence

The first step of the analysis consists in finding a value $\sigma_c(\alpha, \lambda)$ such that, for $\sigma \in]0, \sigma_c(\alpha, \lambda)[$, $f(X^n)$ is a supermartingale. The convergence of the processes $f(X^n)$ and X^n will then immediately follow.

For this purpose, we need some assumptions on f, that we now state:

Assumptions (H1)

(i) The function f has a unique global minimizer x^*. Without loss of generality, we assume that $x^* = 0$ and $f(0) = 0$, and therefore $\forall x \in \mathbb{R}, f(x) > 0$.

(ii) The function f is twice continuously differentiable.

(iii) There exists M finite such that, for all $x \in \mathbb{R}$, $|f''(x)| \leq M$.

(iv) There exists $\alpha > 0$ such that, for all $x \neq 0$, $|\frac{f'(x)}{x}| \geq \alpha > 0$

The key assumptions for our proof are (i) and (iv), where the latter one basically amounts to saying that f exhibits some strict coercivity at the vicinity of its minimizer. In addition to these crucial assumptions we assume technical properties ((ii)-(iii)) for simplicity. The latter ones may be shown to be unneccessary, at the price of tedious technical details we do not want to enter here. Remark 2 below pursues in this line.

In the same vein, a straightforward extension of Assumptions (H1) consists in supposing the three last properties hold only in a neighborhood of zero. Alternatively, one may make use of the trick indicated in Remark 3 below.

Remark 2. *The C^2 regularity assumption (H1)(ii) can be weakened into a $C^{1+\alpha}$ regularity with $0 < \alpha \leq 1$. The constant M of Assumption (H1)(iii) could then be replaced by the constant $M_\alpha = sup_{x,y \in \mathbb{R}} \frac{|f'(x) - f'(y)|}{|x-y|^\alpha}$ induced by the $C^{1+\alpha}$ regularity.*

Remark 3. *It is to be noted that all our proofs still hold when the process X^n is replaced by $\inf(\sup(X^n, -A), A)$ in equation (1) for some large A. Such a modification is an easy trick to render Assumptions (H1) easier to fulfill.*

Remark 4. *Assumption (H1) above clearly implies that f is monotonously decreasing on $\mathbb{R}^-$ and increasing on $\mathbb{R}^+$. Considering the extension described in Remark 3, it may hold only on $[-A, 0]$ and $[0, A]$ respectively.*

Remark 5. *Actually, the need for Assumption (H1) comes from the fact that the choice $H(x) = |x - x^*|$ is to some extent not a good choice. As we will see below when considering e.g. the case $H(x) = |f'(x)|$, we can then prove convergence with weaker assumptions on f, and it is only when convergence rates are to be evaluated that stronger assumptions come into play.*

We now introduce the following function

$$g(\sigma, \lambda, \alpha, M) = \mathbb{E}\left(\min_{1\leq i\leq\lambda}\left(\alpha N^i + \sigma\frac{M}{2}(N^i)^2\right)\right), \tag{10}$$

and define $\sigma_c(\lambda, \alpha, M)$ as the solution to

$$g(\sigma_c(\lambda, \alpha, M), \lambda, \alpha, M) = 0 \tag{11}$$

The existence and uniqueness of $\sigma_c(\lambda, \alpha, M)$ is stated in Lemma 2 below. The convergence of $f(X^n)$ for $\sigma < \sigma_c(\lambda, \alpha, M)$ is contained in

Theorem 3. *If f satisfies Assumption (H1), and $\sigma \in]0, \sigma_c(\lambda, \alpha, M)[$ with $\sigma_c(\lambda, \alpha, M)$ defined by equation (11), then, when n goes to $+\infty$, $f(X^n)$ converges to 0, both almost surely and in L^1, and X^n converges to 0 both almost surely and in L^2.*

The proof of this theorem relies on Lemma 2 and Lemma 3. We first state these two Lemma and their proofs before completing the demonstration of Theorem 3.

Lemma 2. *Fix α as in Assumption (H1)(iv). Let $\lambda \geq 2$ be an integer. Let g be defined by (10). Then, $g(\sigma, \alpha, \lambda, M)$ is a strictly increasing and continuous function of its argument σ. It satisfies $g(0, \alpha, \lambda, M) < 0$ and $\lim_{\sigma\to+\infty} g(\sigma, \alpha, \lambda, M) = +\infty$. Consequently there exists a unique $\sigma_c(\lambda, \alpha, M)$ such that $g(\sigma_c(\alpha, \lambda), \alpha, \lambda, M) = 0$ and therefore $g(\sigma_c(\alpha, \lambda), \alpha, \lambda, M) < 0$ when $\sigma \in [0, \sigma_c(\lambda, \alpha, M)[$.*

Proof. Let σ_1 and σ_2 such that $\sigma_1 < \sigma_2$, then

$$\min_{1\leq i\leq\lambda} \alpha\mathcal{N}^i + \sigma_1\frac{M}{2}(\mathcal{N}^i)^2 < \min_{1\leq i\leq\lambda} \alpha\mathcal{N}^i + \sigma_2\frac{M}{2}(\mathcal{N}^i)^2 \text{ almost surely}$$

And, thus,

$$g(\alpha, \sigma_1, \lambda) = \mathbb{E}(\min_{1\leq i\leq\lambda} \alpha\mathcal{N}^i + \sigma_1\frac{M}{2}(\mathcal{N}^i)^2) < \mathbb{E}(\min_{1\leq i\leq\lambda} \alpha\mathcal{N}^i + \sigma_2\frac{M}{2}(\mathcal{N}^i)^2) = g(\alpha, \sigma_2, \lambda).$$

The continuity of $g(, \sigma, \lambda, \alpha, M)$ can be proved by extracting an increasing subsequence σ_n such that $\sigma_n \to \sigma$ and by using the monotone convergence theorem. The fact that $g(0, \lambda\alpha, M) = \mathbb{E}(\min_{1\leq i\leq\lambda} \mathcal{N}^i) < 0$ for $\lambda \geq 2$ is a result of order statistics. Finally, $\lim_{\sigma\to\infty} g(\sigma, \lambda, \alpha, M) = +\infty$ is a consequence of the monotone convergence theorem.

□

The property of supermartingale of the process $f(X^n)$ is now stated in Lemma 3. In the sequel, $\mathcal{F}_n$ denotes the filtration adapted to the process $f(X^n)$. Let us recall that $f(X^n)$ is said to be a $\mathcal{F}_n$-supermartingale if $f(X^n) \in L^1$ and satisfies $\mathbb{E}(f(X^{n+1})|\mathcal{F}_n) \leq f(X^n)$ almost surely.

Lemma 3. *Assume f satisfies (H1). Then*

$$\mathbb{E}(f(X^{n+1})|\mathcal{F}_n) \leq f(X^n) + \sigma|X^n|^2 g(\sigma, \lambda, \alpha, M). \tag{12}$$

It follows that $f(X^n)$ is a $\mathcal{F}_n$-supermartingale for $0 \leq \sigma \leq \sigma_c(\lambda, \alpha, M)$.

Proof. From the hypotheses $f'' \leq M$, $f(0) = 0$ and $f'(0) = 0$, we deduce that $\forall x,\ f(x) \leq \frac{M}{2}x^2$. In order to show that $f(X^n) \in L^1$, it suffices to show that $X^n \in L^2$. This can be proven by recurrence, using the inequality

$$|X^{n+1}| \leq |X^n| + \sigma|X^n| max_{1\leq i\leq\lambda}(|\mathcal{N}^i|),$$

which comes from the monotonicity of f mentioned in Remark 4, the independence between $|X^n|$ and the fact that $max_{1\leq i\leq\lambda}(|\mathcal{N}^i|) \in L^2$.
Then, by definition, we know

$$f(X^{n+1}) \leq f(X^n + \sigma|X^n|\mathcal{N}_i^n) \quad \forall 1 \leq i \leq \lambda$$

Using, then the Taylor Lagrange formula, and (H1)-(iii), we obtain,

$$f(X^{n+1}) \leq f(X^n) + \sigma|X^n|\mathcal{N}_i^n f'(X^n) + \frac{\sigma^2}{2}|X^n|^2 M(\mathcal{N}_i^n)^2 \quad \forall 1 \leq i \leq \lambda$$

thus, using Assumption (H1)(iv),

$$f(X^{n+1}) \leq f(X^n) + \sigma|X^n||f'(X^n)|(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2\alpha}(\mathcal{N}_i^n)^2) \quad \forall 1 \leq i \leq \lambda,$$

whence

$$f(X^{n+1}) \leq f(X^n) + \sigma|X^n||f'(X^n)| \min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2\alpha}(\mathcal{N}_i^n)^2).$$

Let $\mathcal{H}_n$ be the filtration adpated to the process X^n, we have, $\mathcal{F}_n \subset \mathcal{H}_n$. We first prove that $f(X^n)$ is a $\mathcal{H}_n$-supermartingale and will deduce next that $f(X^n)$ is a $\mathcal{F}_n$-supermartingale. We take the conditional espectation of the previous inequality,

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma|X^n||f'(X^n)|\mathbb{E}\left(\min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2\alpha}(\mathcal{N}_i^n)^2)|\mathcal{H}_n\right)$$

It can be easily proved, using the symetry of the distribution of $\mathcal{N}_i^n$ that,

$$\mathbb{E}(\min_{1\leq i\leq\lambda}(-\mathcal{N}_i^n + \frac{M\sigma}{2\alpha}(\mathcal{N}_i^n)^2)) = \mathbb{E}(\min_{1\leq i\leq\lambda}(+\mathcal{N}_i^n + \frac{M\sigma}{2\alpha}(\mathcal{N}_i^n)^2))$$

Hence,

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma|X^n|\frac{|f'(X^n)|}{\alpha}g(\sigma,\lambda,\alpha,M)$$

Since for $0 < \sigma \leq \sigma_c(\lambda,\alpha,M)$, $g(\sigma,\lambda,\alpha,M) \leq 0$ we obtain

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma|X^n|^2 g(\sigma,\lambda,\alpha,M),$$

i.e. (12). The property, $\mathbb{E}(\mathbb{E}(.|\mathcal{H}_n)|\mathcal{F}_n)) = \mathbb{E}(.|\mathcal{F}_n)$ (see [12]), yields the fact that for $0 < \sigma \leq \sigma_c(\lambda,\alpha,M)$, $f(X^n)$ is a $\mathcal{F}_n$-supermartingale.

□

We are now in position to prove Theorem 3.
proof[of Theorem 3] We start from inequality (12) of Lemma 3, and we take its expectation, which is legitimate since we have seen in the proof of Lemma 3 that $f(X^n)$ and $|X^n|^2$ are both in L^1. We have

$$\mathbb{E}(f(X^{n+1})) \leq \mathbb{E}(f(X^n)) + \sigma\mathbb{E}(|X^n|^2)g(\sigma,\lambda,\alpha,M) \tag{13}$$

From Lemma 2, we know that, for $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$, $g(\sigma,\lambda,\alpha,M) < 0$, thus, for $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$, $\mathbb{E}(f(X^n))$ is a decreasing positive sequence that consequently converges. We rewrite (13) as

$$\mathbb{E}(|X^n|^2)(-\sigma g(\sigma,\lambda,\alpha,M)) \leq \mathbb{E}(f(X^n)) - \mathbb{E}(f(X^{n+1})). \tag{14}$$

For every $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$, the right-hand side of (14) converges to zero, and the left-hand side is positive. This implies that $\mathbb{E}(|X^n|^2)$ converges to 0. We can thus extract a subsequence $|X^{\gamma(n)}|^2$ of $|X^n|^2$ which converges almost surely to zero.As f is continuous, $f(X^{\gamma(n)}) \xrightarrow[n\to\infty]{\text{a.s}} f(0) = 0$. From Lemma 3, we know that, for $\sigma \in]0,\sigma_c(\alpha,\lambda)[$, the process $f(X^n)$ is a positive supermartingale. This implies that $f(X^n)$ converges almost surely (see [12]). From the uniqueness of the limit for $f(X^n)$, we conclude that $f(X^n)$ converges almost surely to 0. For the convergence of $f(X^n)$ in L^1, we deduce from Assumption (H1)(iii) that, $\forall x,\ f(x) \leq \frac{M}{2}x^2$ and use the convergence of $\mathbb{E}(|X^n|^2)$. It follows that $\mathbb{E}(f(X^n))$ converges to zero.

□

3.2.2 Convergence speed

Theorem 4. *Assume f satisfies Assumptions (H1), and that $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$, with $\sigma_c(\lambda,\alpha,M)$ defined by (11), then $f(X^n)$ converges geometrically to 0 in the following senses:*

(i) **(Convergence a.s.)***:* $\dfrac{f(X^n)}{(1+\sigma C g(\sigma,\lambda,\alpha,M))^n}$ *converges to some random variable Y,*

(ii) **(Convergence in L^1)***:* $\mathbb{E}(f(X^n)) \leq (1+\sigma C g(\sigma,\lambda,\alpha,M))^n \mathbb{E}(f(X^0))$,

where $C = \frac{2}{M}$ *and* M *is defined by (H1)(iii). In addition, the best convergence rate is reached for* $\sigma = \sigma_s(\lambda, \alpha, M)$ *where* $\sigma_s(\lambda, \alpha, M)$ *is the unique value of* σ *that minimizes* $1 + \sigma C g(\sigma, \lambda, \alpha, M)$.

Proof. We choose $\sigma < \sigma_c(\alpha, \lambda)$ and we again start from inequality (12)

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma |X^n|^2 g(\sigma, \lambda, \alpha, M)$$

As mentioned above, Assumption (H1) in particular implies that $f(x) \leq \frac{M}{2} x^2$ for all x, and therefore we obtain

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n)(1 + \sigma \frac{2}{M} g(\sigma, \lambda, \alpha, M))$$

For σ such that $0 < 1 + \sigma \frac{2}{M} g(\sigma, \lambda, \alpha, M) < 1$, $f(X^n)/(1 + \sigma \frac{2}{M} g(\sigma, \lambda, \alpha, M))^n$ is a positive super martingale which converges almost surely. Taking the expectation leads to the geometric convergence of $\mathbb{E}(f(X^n))$ to zero at the rate announced in the Theorem.

□

3.3 Convergence of the $(1, \lambda)$-ES with $H(x) = |f'(x)|$

We now deal with the algorithm

$$\left\{ \begin{array}{l} X^0 \in \mathbb{R}, \\ X^{n+1} = \arg\min\{f(X^n + \sigma |f'(X^n)| \mathcal{N}_i^n), i \in [1, \lambda]\}, \end{array} \right. \tag{15}$$

The general outline of the demonstration in this case is the same as in the previous section: First, find a value $\sigma_c(\alpha, \lambda)$ such that $f(X^n)$ is a supermartingale for $\sigma \in]0, \sigma_c(\alpha, \lambda)[$. Then, derive the convergence and the speed of convergence of $f(X^n)$.

Contrary to the previous section, convexity is not mandatory in the present section to obtain the convergence result *per se*. However, some local convexity is indeed needed to derive the convergence rate. In particular, it is simply assumed in the begining of this section that

Assumption (H2)

(i) The function f *is bounded from below (say by zero) and is twice continuously differentiable.*

(ii) There exists M *finite such that, for all* x,

$$|f''(x)| \leq M.$$

Remark 6. *Once again, using the truncation trick mentioned in Remark 3 weakens this assumption which is then satisfied for every* C^2 *function.*

3.3.1 Convergence of $f'(X^n)$

In order to lighten the notation, we henceforth denote by

$$h(\sigma,\lambda) = g(1,\sigma,\lambda) = \mathbb{E}\left(\min_{1\leq i\leq\lambda}\left(\mathcal{N}^i + \sigma\frac{M}{2}(\mathcal{N}^i)^2\right)\right) \tag{16}$$

and define, $\sigma_c^{'}(\lambda)$ as the solution of,

$$g(1,\sigma_c^{'}(\lambda),\lambda) = 0 \tag{17}$$

The proof of existence and uniqueness of this constant is exactly the same as the proof of Lemma 2 in the previous section. The convergence of the process $f'(X^n)$ to zero is stated in the following theorem.

Theorem 5. *Assume f satisfies Assumption (H2). Assume $\lambda \geq 2$ and $\sigma \in]0,\sigma_c^{'}(\lambda)[$. Then $f'(X^n)$ converges to 0 in L^2. If we moreover assume that $f(X^n)$ is bounded then $f'(X^n)$ converges almost surely.*

The proof of this theorem relies on two lemmas, that we state and prove before completing the proof of the Theorem itself. In the sequel $\mathcal{F}_n$ again denotes the filtration adapted to the process $f(X^n)$.

Lemma 4. *Let $\sigma_c^{'}(\lambda)$ defined by (17). Then, $f(X^n)$ is a $\mathcal{F}_n$-supermartingale for $0 \leq \sigma \leq \sigma_c^{'}(\lambda)$.*

Proof. The proof follows the same lines as that of Lemma 3. We first remark that,

$$f(X^{n+1}) \leq f(X^n + \sigma|f'(X^n)|\mathcal{N}_i^n) \quad \forall 1 \leq i \leq \lambda$$

Using successively the Taylor Lagrange formula, the property (H2)(ii), and taking the minimum over i of the right hand side, we obtain,

$$f(X^{n+1}) \leq f(X^n) + \sigma|f'(X^n)|^2 \min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2}(\mathcal{N}_i^n)^2) \tag{18}$$

Let $\mathcal{H}_n$ be again the filtration adpated to the process X^n. Taking the conditional espectation of the previous inequality, we have

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma|f'(X^n)|^2\mathbb{E}\left(\min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2}(\mathcal{N}_i^n)^2)|\mathcal{H}_n\right)$$

and therefore

$$\mathbb{E}(f(X^{n+1})|\mathcal{H}_n) \leq f(X^n) + \sigma|f'(X^n)|^2 h(\sigma,\lambda) \tag{19}$$

For $0 < \sigma \leq \sigma_c^{'}(\alpha,\lambda)$, this implies (taking the expectation of both sides) that, by induction $f(X^n) \in L^1$, and foremost that $f(X^n)$ is a $\mathcal{F}_n$-supermartingale.

□

The next lemma, which indeed contains a standard result of the theory of random processes, is given for consistency.

Lemma 5. *Let $\mathcal{F}_n$ be an increasing filtration adapted to a process X^n. We assume that, $X^n \xrightarrow[n\to\infty]{a.s} 0$ and that X^n is bounded.*
Then,

$$\mathbb{E}(X^n|\mathcal{F}_n) \xrightarrow[n\to\infty]{a.s} 0$$

Proof. Fix an integer n_0. We have

$$|\mathbb{E}(X^n|\mathcal{F}_n)| \leq \mathbb{E}(sup_{p\geq n_0}|X_p||\mathcal{F}_n) \text{for all } \; n \geq n_0$$

Hence,

$$\limsup\left|\mathbb{E}(X^n|\mathcal{F}_n)\right| \leq \mathbb{E}(sup_{p\geq n_0}|X_p||\mathcal{F}_\infty)$$

With the dominated convergence theorem,

$$\mathbb{E}(sup_{p\geq n_0}|X_p||\mathcal{F}_\infty) \xrightarrow{n_0\to\infty} 0,$$

thus we have

$$\mathbb{E}(X^n|\mathcal{F}_n) \xrightarrow[n\to\infty]{\text{p.s}} 0.$$

□

We now give the demonstration of Theorem 5.

Proof. For the first point of the theorem, we take the expectation of the equation (19),

$$\mathbb{E}(f(X^{n+1})) \leq \mathbb{E}(f(X^n)) + \sigma\mathbb{E}(|f'(X^n)|^2)h(\sigma,\lambda)$$

Thus, for $0 \leq \sigma \leq \sigma_c^{'}(\alpha,\lambda)$, $\mathbb{E}(f(X^n))$ is a positive decreasing sequence which consequently converges. In addition,

$$-\sigma\mathbb{E}(|f'(X^n)|^2)h(\sigma,\lambda) \leq \mathbb{E}(f(X^n)) - \mathbb{E}(f(X^{n+1}))$$

implies that, for $0 < \sigma < \sigma_c^{'}(\alpha,\lambda)$,

$$\mathbb{E}(|f'(X^n)|^2) \to 0. \tag{20}$$

In order to now get the almost sure convergence, we come back to inequality (19) and Lemma 5. We know according to (19) that

$$|f'(X^n)|^2 \leq \frac{\mathbb{E}(f(X^n) - f(X^{n+1})|\mathcal{F}_n)}{-\sigma h(\sigma,\lambda)}$$

Thanks to Lemma 4, $f(X^n)$ is a positive supermartingale. It converges almost surely. Hence $f(X^n) - f(X^{n+1})$ converges almost surely to 0. As we also assume that $f(X^n)$ is bounded, we deduce from Lemma 5 that $\mathbb{E}(f(X^n) - f(X^{n+1})|\mathcal{F}_n)$ converges almost surely to 0. Thus $|f'(X^n)|^2$ converges almost surely to 0.

□

3.3.2 Convergence speed

An additional hypothesis, somewhat connected to convexity, is now needed to estimate the convergence speed. Before we state it, we set by convention that

$$\inf_{\mathbb{R}} f = 0,$$

otherwise $f(x)$ should be replaced by $f(x) - \inf_{\mathbb{R}} f$ in the assumption below.

Assumption (H3)

- *There exists $C > 0$ such that*

$$\inf_{\mathbb{R}} \frac{|f'(x)|^2}{f(x)} \geq C.$$

Remark 7. *It is important to note that Assumption (H3) implies that the set of global minimizers x^* of f is indeed an interval, and that any critical point is such a minimizer.*

Theorem 6. *Assume f satisfies Assumptions (H2)(H3) and that $\sigma \in]0, \sigma_c'(\lambda)[$. Then $f(X^n)$ converges geometrically to 0 at the rate $(1 + \sigma Ch(\sigma, \lambda))$ both almost surely and L^1 (in the sense of Theorem 4, and with the constant C defined in (H3)). The best convergence rate is reached for $\sigma = \sigma_s'(\lambda)$ where $\sigma_s'(\lambda)$ minimizes $1 + \sigma Ch(\sigma, \lambda)$.*

Proof. We choose $\sigma \in]0, \sigma_c'(\lambda)[$ and come back to (19) to write, using Assumption (H3),

$$\mathbb{E}(f(X^{n+1})|\mathcal{F}_n) \leq f(X^n)\,(1 + \sigma Ch(\sigma, \lambda)) \tag{21}$$

For σ such that $0 < 1 + \sigma Ch(\sigma, \lambda) < 1$, $f(X^n)/(1 + \sigma Ch(\sigma, \lambda))$ is a positive super martingale which thus converges almost surely. If we take the expectation of equation (21), we obtain that $\mathbb{E}(f(X^n))$ geometrically converges to zero at the rate $(1 + \sigma Ch(\sigma, \lambda))^n$.

Remark 8. On the optimality of the general estimates when applied to the sphere function *It is enlightening to compare the results obtained in the present section and the one of the previous one when the function under consideration is the sphere function $f(x) = x^2$. The algorithm (15) then reads*

$$\begin{cases} X^0 \in \mathbb{R}, \\ X^{n+1} = \arg\min\{f(X^n + 2\sigma|X^n|\mathcal{N}_i^n), i \in [1, \lambda]\}, \end{cases} \tag{22}$$

and is exactly that of the previous section when $H(x) = |x|$ and σ is replaced by 2σ. In this particular case, Theorem 6 claims that the rate of convergence is $1+4\sigma g(1,\sigma,\lambda)$ since $C = 4$ in Assumption (H3). On the other hand, Theorem 4 claims that the rate is $1+\sigma\frac{2}{M}g(\alpha,\sigma,\lambda)$ and therefore here, where $\alpha = 2$, $M = 2$, $\sigma = 2\sigma$, that it reads $1 + 2\sigma g(2, 2\sigma, \lambda) = 1 + 4\sigma g(1,\sigma,\lambda)$. It follows that the critical values $\sigma_c(\alpha,\lambda)$ and $\sigma_s(\alpha,\lambda)$ given in Section 3.1 are the same.

3.4 The case of higher dimensions

The algorithm defined in (1) must be slightly modified when going to dimension $d > 1$. The general form of the *non-isotropic* ES algorithm considered here is:

$$\begin{cases} X^0 \in \mathbb{R}^d, \\ X^{n+1} = \arg\min\{f(X^n + \sigma(H_k(X^n)\mathcal{N}_k^{n,i})_{k\in[1,d]}), i \in [1,\lambda]\}, \end{cases} \tag{23}$$

where X_n is the random variable modeling the parent at the generation n, $(\mathcal{N}_k^{n,i}), k \in [1,d], i \in [1,\lambda]$ are independent standard normal random variables, and $H_k(x), k \in [1,d]$ are d real-valued functions. Different step-sizes are here applied to the different directions, similarly with what can be done as far as self-adaptation is concerned [10].

Only the case of practical interest where $H_k(x) = \dfrac{\partial f(x)}{\partial x_k}$ will be considered here. The situation is then similar to that studied in Section 3.3, a similar result to that of Lemma 4 can be proven. Indeed, let us briefly sketch the modified argument. We first replace assumption (H2)(ii) by

There exists M finite such that $D^2 f$, the Hessian matrix of f, satisfies

$$D^2 f \leq M\,\mathbf{Id},$$

in the sense of symmetric matrices.

and assumption (H3) by

There exists $C > 0$ such that

$$\|\nabla f\|^2 \geq Cf.$$

We next write that, for all $1 \leq i \leq \lambda$,

$$\begin{aligned} f(X^{n+1}) &\leq f(X^n + \sigma(\frac{\partial f(X^n)}{\partial x_k}\mathcal{N}_k^{n,i})) \\ &= f(X^n) + \sigma\sum_{k=1}^{d}(\frac{\partial f(X^n)}{\partial x_k})^2\,\mathcal{N}_k^{n,i} \\ &\quad + \frac{1}{2}D^2 f(Z^n).(\sigma(\frac{\partial f(X^n)}{\partial x_k}\mathcal{N}_k^{n,i})_k, \sigma(\frac{\partial f(X^n)}{\partial x_k}\mathcal{N}_k^{n,i})_k) \\ &\leq f(X^n) + \sigma\sum_{k=1}^{d}(\frac{\partial f(X^n)}{\partial x_k})^2\,[\mathcal{N}_k^{n,i} + \frac{M}{2}\sigma(\mathcal{N}_k^{n,i})^2] \end{aligned}$$

Then, it is straightforward to see that all our arguments go through *mutatis mutandis*, yielding the same conclusions and the same criteria of convergence.

In particular, the critical value σ_c, below which convergence takes place, is again defined by equations (16) and (17). The more remarkable fact here is that this critical value (and hence the convergence rate that comes with it) does not depend on the dimension.

4 Discussion

This section discusses the results of previous sections in the light of past related work from the litterature.

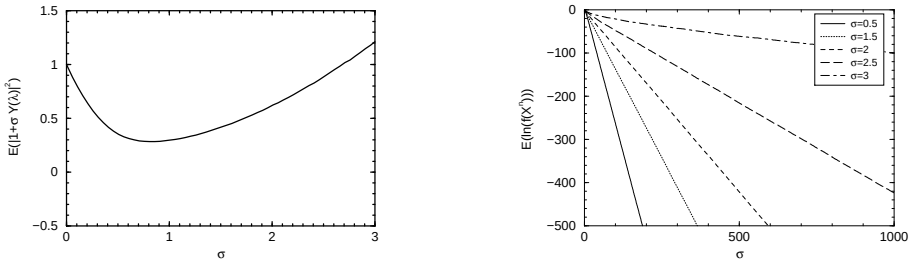
First, it should be clear that only works proposing global convergence results are relevant for comparison here, as opposed to all work studying local convergence (see Section 3.1 for a link with those works).

The work whose results are most similar to the ones presented here are by far Rudoplh's work, either using also supermartingale [8], or somehow simplified and based on order statistics [9]. There are however quite a few differences.

First, Rudolph's results are based on some strong convexity of function f – but it is fair to say that on the other hand, he only needs f to be differentiable once – whereas convexity is not required here for the convergence result, and, as expected, only weak convexity is necessary to obtain the geometrical converge rate.

Second, whereas Rudolph chooses all offspring uniformly on some hypersphere (or radius σ), the algorithm considered here uses the "true" Gaussian mutation. A common argument is that both mutations behave similarly in high dimension. However, when it comes to theoretical results, such consideration is of no help. Indeed, the method used by Rudolph based on order statistics [9] can also be applied with Gaussian mutation, and gives the same kind of convergence result: there exists a critical value σ_c such that whenever σ lies in $]0, \sigma_c[$ the algorithm converges. Unfortunately, this constant σ_c is then defined as $2\frac{\mathbb{E}(N^{\lambda:\lambda})}{M\mathbb{E}((N^{\lambda:\lambda})^2)}$, where $N^{\lambda:\lambda}$ is the λ^{th} order statistics for standard normal random variables. The problem is that this quantity is a very poor upper bound: for instance, it decreases for large values of λ, making the result less relevant.

A noticeable difference with Rudolph's algorithm in [9] lies in the case where the dimension is greater than 1: the offspring of parent X^n in Rudolph's algorithms are chosen using $H(x) = \sigma||\nabla f(x)||N$ (notation of equation (1)), for some vector of standard normal random variables N. The approach proposed here is different (see Section 3.4), and the results are indeed far more appealing: the upper-bound geometrical rate obtained by Rudolph goes to 1 when the dimension goes to ∞ (despite the fact that he does not use Gaussian mutation), while the one proposed here does not depend on the dimension. However, the gap between the two approaches remains open, as it has not been possible up to now to analyze the algorithm 1 with Rudolph's H function.



(a) (b)

Figure 1: (a) $\mathbb{E}(|1+\sigma Y(4)|^2)$ with respect to σ – see equation (3). (b) $\mathbb{E}(ln(f(X^n)))$ with respect to the number of generations n for different values of σ.

5 Numerical experiments

All numerical experiments presented in the sequel are based on the Monte Carlo approximation of the expectation of a random variable. Define

$$\mathbb{E}(Z) \approx \frac{1}{K}\sum_{k=1}^{K} Z_k,$$

where Z_k are K independant random variables with the same law than Z. Then, for instance, from the central limit theorem, for large values of K ($K = 1500$ in all numerical experiments presented here), with probability 0.95,

$$\mathbb{E}(Z) \in [\frac{1}{K}\sum_{k=1}^{K} Z_k - \sqrt{VarZ}\frac{1.96}{\sqrt{K}}, \frac{1}{K}\sum_{k=1}^{K} Z_k + \sqrt{VarZ}\frac{1.96}{\sqrt{K}}] \tag{24}$$

5.1 Computation of the constants

The Monte-Carlo method described above has been used to compute approximate values of the constants $\sigma_c(\alpha,\lambda)$ and $\sigma_s(\alpha,\lambda)$ from Section 3.

A first example is given by the plot of $\mathbb{E}(|1+\sigma Y(\lambda)|^2)$ against σ for the sphere function on Figure 1 (a), for $\lambda = 4$. The limit value of σ for which $\mathbb{E}(|1+\sigma Y(\lambda)|^2) \leq 1$ is $\sigma_c(\lambda,2) = 2.7$, and the corresponding minimal value for $\mathbb{E}(|1+\sigma Y(\lambda)|^2)$ is $\sigma_s(\alpha,\lambda,2) \approx 0.8$. Note that this method allows us to plot the progress rate (8) for any dimension d, as in [10, 3], without any assumption regarding $d \to +\infty$.

5.2 Optimality of the constants

The idea here is to compare the constants $\sigma_c(\alpha,\lambda)$, $\sigma_s(\alpha,\lambda)$, $\sigma_c^{'}(\lambda)$ and $\sigma_s^{'}(\lambda)$ for some functions that are not quadratic, in order to test their optimality.

First, we need to circumvent a difficulty. Indeed, when evaluating $\mathbb{E}(f(X^n))$ with the Monte Carlo method, the relative error given by the Central Limit Theorem ($\frac{\sqrt{Var(f(X^n))}}{\sqrt{K}\mathbb{E}(f(X^n))}\frac{1.96}{}$) grows geometricaly with the number of generations n (the exact computation can be made easily on the sphere function). On the other hand, that of evaluating $\mathbb{E}(\ln(f(X^n)))$ decreases in $\frac{1}{\sqrt{n}}$. Hence, all numerical tests have been performed on the process $ln(f(X^n))$. This fact in turn requires to come back to the convergence analysis. Indeed, it turns out that the arguments used to treat the minimization of f also hold for the minimization of $ln(f)$. Of course, since the a.s. convergence of $f(X^n)$ implies that of $ln(f(X^n))$, we know sufficient conditions for such a convergence. But, more than that, $ln(f(X^n)$ converges in the same fashion and under the same conditions as $f(X^n)$ with an arithmetic rate replacing the geometric rate of Theorems 4 and 6. Let us briefly indicate here the modifications of the arguments needed to establish these facts. We argue e.g. in the case $H(x) = |f'(x)|$. We start from (18), namely

$$f(X^{n+1}) \leq f(X^n) + \sigma|f'(X^n)|^2 \min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2}(\mathcal{N}_i^n)^2),$$

take the logarithm of both sides, and use $ln(1+u) \leq u$ to obtain

$$ln(f(X^{n+1})) \leq ln(f(X^n)) + \sigma\frac{|f'(X^n)|^2}{f(X^n)} \min_{1\leq i\leq\lambda}(sgn(f'(X^n))\mathcal{N}_i^n + \frac{M\sigma}{2}(\mathcal{N}_i^n)^2).$$

Then, we pass to the conditional expectation, and next use Assumption (H3) to write

$$\mathbb{E}(ln(f(X^{n+1}))|\mathcal{H}_n) \leq ln(f(X^n)) + \sigma Ch(\sigma,\lambda).$$

This shows that $ln(f(X^n))$ (and in fact $-n\sigma Ch(\sigma,\lambda) + ln(f(X^n))$) is a supermartingale. A consequence is also the bound from above

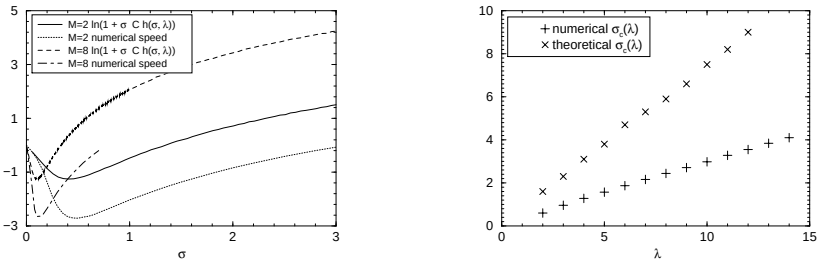
$$\mathbb{E}(ln(f(X^n)) \leq \mathbb{E}(ln(f(X^0)) + n\sigma Ch(\sigma,\lambda),$$

which yields the convergence of $\mathbb{E}(ln(f(X^n))$ to $-\infty$. Likewise, all results follow for the function $ln(f)$ as they used to for f, with the same critical (resp. optimal) values for the convergence rates. These rates are of course arithmetic instead of geometric. Anyway, this equivalence will enable us to test the optimality of our analysis.

Only numerical results concerning the case $H(x) = |f'(x)|$ will be shown here.

The functions f_M, defined by equation (25) below, are examples among the class of non symetrical functions satisfying both Assumptions (H2) and (H3) that will be used for all experiments (where $M > 0$ is the value used in Assumption (H2)-(ii)).

$$f_M(x) \quad = \frac{M}{2} \quad \begin{cases} x^2 & if\, x < 0 \\ x\arctan(x) & if\, x > 0 \end{cases} \tag{25}$$



(a) (b)

Figure 2: (a) Theoretical and numerical speeds of convergence for functions f_2 and f_8. (b) Numerical $\sigma^{'}_{cnum}(\lambda)$ and theoretical $\sigma^{'}_{c}(\lambda)$.

Figure 1 (b) plots $\frac{1}{K}\sum_{k=1}^{K} ln(f_2(X_k^n))$ against the number of generations for different values of σ. The relative error

$$\frac{\mathbb{E}(ln(f_2(X^n))) - \frac{1}{K}\sum_{k=1}^{K} ln(f_2(X_k^n))}{\frac{1}{K}\sum_{k=1}^{K} ln(f_2(X_k^n))}$$

given by equation (24), is here bounded by 0.01. This corroborates the linear rate of convergence predicted by our theoretical study.

Figure 2 (a) plots the slopes of those linear functions (determined using linear regressions), and the theoretical values $\sigma Ch(\sigma, \lambda)$, for $\lambda = 4$ and for both functions f_2 and f_8. Both curves have the same shapes. Moreover, on these functions, the theoretical bounds indeed underestimate the threshold, as expected.

Studying only fonction f_2, the intersection between the theoretical curve and the x-axis gives a numerical approximation $\sigma^{'}_{c}(4) \approx 1.4$ of the theoretical value $\sigma_c(4)$ – and in the sequel, $\sigma^{'}_{cnum}(4)$ will denote the intersection between the experimental curve and the x-axis. From Figure 2 (a), it comes that $\sigma^{'}_{c_{num}}(4) \approx 3.1$. Defining similarly $\sigma^{'}_{s_{num}}(4)$ as the critical point of the the numerical curve, it may also be noted on the same Figure that $\sigma^{'}_{s}(4) \leq \sigma^{'}_{s_{num}}(4)$.

It may be observed from the same Figure 2 (a) that both theoretical and numerical curves present the same scaling transformation when M is increased – even though the theoretical bound still seems pessimistic. Last, Figure 2 (b) shows, for function f_2, the numerical $\sigma^{'}_{cnum}(\lambda)$ and theoretical $\sigma^{'}_{c}(\lambda)$ for $\lambda = 2, ..., 13$. Both are linear increasing functions in λ.

6 Conclusions and perspectives

We have concentrated in this article on some convergence results and geometrical convergence rates for adaptive $(1, \lambda) - ES$ in the specific cases of a sub-class of C^2 functions. Our

results have been extended to the d-dimensional case with a *non-isotropic* ES algorithm. In the specific cases we have dealt with, they are quite satisfactory, as shown by the comparison to experiments, for our analysis is sharp enough to reproduce most of the *qualitative* behaviours that are observed in the practice (convergence speed, behaviour with respect to the population size, the dimension of the ambient space). Of course, on the *quantitative* level, there is room for improvement. We now intend to pursue our efforts in two different directions. First, we aim at improving our results on adaptive algorithms. This direction is twofold: relaxing the regularity and convexity assumptions for the same adaptive algorithms, and considering other adaptive algorithms, in particular one more practically useful algorithm, where the step-size is adapted proportionaly to $|f(x) - \inf f|$. The second direction deals with self-adaptive $(1, \lambda) - ES$.

References

[1] A.Auger. ES, théorie et applications au contrôle en chimie. *Thèse de l'Université Paris 6*, (in preparation).

[2] A. Auger, C. Le Bris, and M. Schoenauer. Dimension-independent convergence rate for non-isotropic $(1, \lambda)$-ES. In Erick Cantù-Paz et al., editor, *Lecture Notes in Computer Science*, volume 2723, pages 512–524. Springer Verlag, 2003.

[3] H.-G. Beyer. *The Theory of Evolution Strategies*. Springer, Heidelberg, 2001.

[4] A. Bienvenüe and O. François. Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties. *Theoretical Computer Science A*, to appear. http://www-lmc.imag.fr/lmc-sms/Alexis.Bienvenue/.

[5] J.M DeLaurentis, L. A. Ferguson, and W.E. Hart. On the convergence properties of a simple self-adaptive evolutionary algorithm. In W.B. Langdon & al., editor, *Proceedings of the Genetic and Evolutionary Conference*, pages 229–237. Morgan Kaufmann, 2002.

[6] A. E. Eiben, R. Hinterding, and Z. Michalewicz. Parameter control in Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124, 1999.

[7] I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.

[8] G. Rudolph. Convergence of non-elitist strategies. In Z. Michalewicz, J. D. Schaffer, H.-P. Schwefel, D. B. Fogel, and H. Kitano, editors, *Proceedings of the First IEEE International Conference on Evolutionary Computation*, pages 63–66. IEEE Press, 1994.

[9] G. Rudolph. Convergence rates of evolutionary algorithms for a class of convex objective functions. *Control and Cybernetics*, 26(3):375–390, 1997.

[10] H.-P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[11] M.A. Semenov. Convergence velocity of an evolutionary algorithm with self-adaptation. In W.B. Langdon & al., editor, *Proceedings of the Genetic and Evolutionary Conference*, pages 210–213. Morgan Kaufmann, 2002.

[12] D. Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, 2000.

[13] G. Yin, G. Rudolph, and H.-P Schwefel. Analysing $(1, \lambda)$ evolution strategy via stochastic approximation methods. *Evolutionary Computation*, 3(4):473–489, 1996.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399