Dimension-Independent Convergence Rate for Non-isotropic $(1, \lambda) - ES$

Anne Auger[1,2], Claude Le Bris[1,3], and Marc Schoenauer[2]

[1] CERMICS – ENPC
Cité Descartes, 77455 Marne-La-Vallée, France
{auger,lebris}@cermics.enpc.fr
[2] INRIA Rocquencourt, Projet Fractales
BP 105, 78153 LE CHESNAY Cedex, France
marc.schoenauer@inria.fr
[3] INRIA Rocquencourt, Projet MIC MAC
BP 105, 78153 LE CHESNAY Cedex, France

Abstract. Based on the theory of non-negative super martingales, convergence results are proven for adaptive $(1, \lambda) - ES$ (i.e. 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. Rigorous numerical investigations on some 1-dimensional functions validate the theoretical results. Trends for future research are indicated.

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 $(\mu \stackrel{+}{,} \lambda) - ES$ 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 function slightly more general than quadratic functions, the so-called $(Q - K) -$ *strongly convex* functions [9]. These latter results deal with the so-called *adaptive* version of evolution strategies, in which the step-size is computed

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2723, pp. 512–524, 2003.

at each iteration according to some measures 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 $(1, \lambda) - \sigma - SA - ES$ 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 papers 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 probably be improved in the near future. An important point about these latter two results is that they use the theory of super-martingales [12], a somewhat more sophisticated technique than all previously cited works (with the remarkable exception of [8]).

The same super martingale technique will be used in this paper, 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 will also be studied: as in previous relevant work, some geometrical upper-bounds will be derived, and their sharpness will be tested through numerical experiments.

The paper 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 adaptive ES (see formula (1) below). Section 3 gives the convergence results and the main ideas of the proofs (due to size limitation, the complete proofs cannot be given here, see [2] for all the details). First, the one-dimensional case is thoroughly studied: in the case of the sphere function analytical results are obtained for the sphere function, 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. This is done 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 next 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 (w.r.t. their minimum) nor convex). The article closes with some discussion and trends for future work.

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 f be a real-valued function defined on $\mathbb{R}$ to be minimized. The general adaptive $(1, \lambda)$-Evolution Strategy algorithm we will consider henceforth is of the form:

$$\begin{cases} X^0 \in \mathbb{R}, \\ X^{n+1} = \arg\min\{f(X^n + \sigma H(X^n) N_i^n), i \in [1, \lambda]\}, \end{cases} \tag{1}$$

where X_n is the random variable modeling the parent at the generation n, $(N_i^n)_{i=1,\dots,\lambda}$ are independent standard normal random variables, $H(x)$ is real-valued function (for conciseness, only two cases will be considered in the following: $H(x) = |x|$ or $H(x) = |f'(x)|$, but other cases, such as $H(x) = |f(x) - f^*|$ can be treated by the same technique, see [1])), and 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 sought.

Section 3 gives answers to both questions, first for the sphere function (section 3.1), as exact convergence rates can be easily computed, and then for twice continuously differentiable functions with particular properties 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 – Again

The sphere function ($f(X) = |X|^2$) has always been 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, many things get simpler, and most quantities of interest can be computed analytically.

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

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 $H(x) = |x|$ satisfies,*

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

where $Y(\lambda)$, the random variable 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].

Convergence in L^p. The following theorem is an immediate consequence of Lemma 1:

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

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

Hence, the algorithm converges or diverges in L^p norm geometrically. Moreover, there exists a value $\sigma_c(\lambda, p)$ such that X^n converges in L^p norm iff $\sigma \in]0, \sigma_c(\lambda, p)]$. 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,

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

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 the value of X^n and is hence for instance equal to the value for $X^n = 1$: $(\forall n > 0), \varphi_p(X^n, \sigma, \lambda) = \mathbb{E}(|1 + \sigma Y(\lambda)|^p - 1)$.

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).

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. *Assume that* $\mathbb{E}(\ln(|1 + \sigma Y(\lambda)|)) < \infty$. *Then, for the sphere function, the random variable* X^n *defined by (1) with* $H(x) = |x|$ *satisfies,*

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

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 $H(x) = |x|$ (or $H(x) = |x - x^*|$ for some minimizer x^* of f) is the case with constant (normalized) step-size, as defined for instance in [3]. Though this algorithm has not a practical interest because it supposes that a minimum is already known, it will allow us to develop the technique of analysis to be later applied to the more interesting case $H(x) = |f'(x)|$.

The first step of this analysis consists in finding a value σ_c such that $f(X^n)$ is a super martingale for $\sigma \in]0, \sigma_c[$. The convergence of the processes $f(X^n)$ and X^n will immediately follow (see [12]). For this purpose, we state some assumptions on f:

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 $x \neq 0$, $|\frac{f'(x)}{x}| \geq \alpha > 0$*

Remark 2. All our proofs (see [2]) still go through 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 3. Assumption (H1) above implies that f is monotonously decreasing on $\mathbb{R}^-$ and increasing on $\mathbb{R}^+$.

In the sequel, $\mathcal{F}_n$ denotes the filtration adapted to the process $f(X^n)$.

Lemma 2. *Assume $\lambda \leq 2$. Let g be defined by*

$$g(\sigma,\lambda,\alpha,M) = \mathbb{E}\left(\min_{1\leq i\leq\lambda}\left(\alpha N^i + \sigma\frac{M}{2}(N^i)^2\right)\right), \tag{8}$$

and let $\sigma_c(\lambda,\alpha,M)$ be the solution of

$$g(\sigma_c(\lambda,\alpha,M),\lambda,\alpha,M) = 0 \tag{9}$$

Assume f satisfies Assumption (H1), $f(X^n)$ is a $\mathcal{F}_n$-super martingale[2] *for $0 \leq \sigma \leq \sigma_c(\lambda,\alpha,M)$.*

Remark 4. The value $\sigma_c(\lambda,\alpha,M)$ defined by equation (9) always exists and is unique for $\lambda \geq 2$, α and M given, because $g(\sigma,\lambda,\alpha,M)$ is a strictly increasing and continuous function w.r.t. σ, and satisfies $g(0,\lambda,\alpha,M) < 0$ and $\lim_{\sigma\to+\infty} g(\sigma,\lambda,\alpha,M) = +\infty$.

Key point of the proof. The demonstration of this result relies on the following inequality, based on Taylor formula:

$$\mathbb{E}(f(X^{n+1})|\mathcal{F}_n) \leq f(X^n) + \sigma|X^n|^2 g(\sigma,\lambda,\alpha,M) \;\; a.s. \tag{10}$$

Convergence result. From this Lemma, the theory of non-negative super martingale [12] gives the following theorem.

Theorem 3. *Assume $\lambda \geq 2$, assume f satisfies Assumption (H1), and $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$ with $\sigma_c(\lambda,\alpha,M)$ defined by equation (9), 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.*

Convergence Speed

Theorem 4. *Assume $\lambda \geq 2$, assume f satisfies Assumptions (H1), and that $\sigma \in]0,\sigma_c(\lambda,\alpha,M)[$, with $\sigma_c(\lambda,\alpha,M)$ defined by (9), 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 $1+\sigma C g(\sigma,\lambda,\alpha,M)$.

[2] Z^n is a super martingale if it satisfies $\mathbb{E}(Z^{n+1}|\mathcal{F}_n) \leq Z^n$

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

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

Contrary to the previous section, unimodality is not mandatory in the present section to obtain the convergence result *per se*. But, some local convexity is needed to derive the convergence rate. We consider the following assumptions,

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

Lemma 3. *Assume $\lambda \geq 2$. Let h be defined by*

$$h(\sigma, \lambda, M) = \mathbb{E}\left(\min_{1 \leq i \leq \lambda}\left(N^i + \sigma \frac{M}{2}(N^i)^2\right)\right) \tag{11}$$

and let $\sigma_c^{'}(\lambda, M)$ be the solution of

$$h(\sigma_c^{'}(\lambda, M), \lambda, M) = 0 \tag{12}$$

Then, if f satisfies Assumption (H2), $f(X^n)$ is a $\mathcal{F}_n$-super martingale for $0 \leq \sigma \leq \sigma_c^{'}(\lambda, M)$.

Remark 6. The proof of the existence of $\sigma_c^{'}(\lambda, M)$ is exactly the same as in Remark 4.

Key point of the proof. Once again, the demonstration of the above result relies on the following inequality.

$$\mathbb{E}(f(X^{n+1})|\mathcal{F}_n) \leq f(X^n) + \sigma |f'(X^n)|^2 h(\sigma, \lambda, M) \ \ a.s. \tag{13}$$

Convergence result. A straightforward corollary of this Lemma is that $f(X^n)$ converges almost surely. The following theorem then gives the convergence of $f'(X^n)$.

Theorem 5. *Assume f satisfies Assumption (H2). Assume $\lambda \geq 2$ and $\sigma \in]0, \sigma_c^{'}(\lambda, M)[$. 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.*

Remark 7. If we moreover suppose that f is unimodal and that the only minimum is 0, then the algorithm converges globally: $f(X^n)$ converges to 0 a.s.

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 8. Example of non-trivial functions satisfying both Assumptions (H2) and (H3) will be given in the numerical experiments (see section 5).

Theorem 6. *Assume* $\lambda \geq 2$. *Assume* f *satisfies Assumptions (H2)(H3) and that* $\sigma \in]0, \sigma_c^{'}(\lambda, M)[$. *Then* $f(X^n)$ *converges geometrically to 0 at the rate* $(1 + \sigma Ch(\sigma, \lambda, M))$ *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* $\sigma = \sigma_s^{'}(\lambda, M)$ *where* $\sigma_s^{'}(\lambda, M)$ *minimizes* $1 + \sigma Ch(\sigma, \lambda, M)$.

On the optimality of the general estimates when applied to the sphere function. Going back to the sphere function, the values in Assumption (H1),(H2) and (H3) are $M = 2$, $\alpha = 2$ and $C = 4$, and straightforward calculus gives $\mathbb{E}(|1 + \sigma Y(\lambda)|^2) = \mathbb{E}(min_{1 \leq i \leq \lambda}(1 + \sigma N_i)^2) = 1 + \sigma g(\sigma, \lambda, 2, 2) = 1 + 2\sigma h(\frac{\sigma}{2}, \lambda, 2)$. It is thus easy to show that the critical values given in Theorems 3, 4, 5 and 6 are the optimal values given by equations (5) and (6).

3.4 Results in Higher Dimensions

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

$$\begin{cases} X^0 \in \mathbb{R}^d, \\ X^{n+1} = \arg\min\{f(X^n + \sigma(H_k(X^n)N_k^{n,i})_{k \in [1,d]}), i \in [1, \lambda]\}, \end{cases} \tag{14}$$

where X_n is the random variable modeling the parent at the generation n, $(N_k^{n,i}), k \in [1, d], i \in [1, \lambda]$ 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 $H_k(x) = \frac{\partial f(x)}{\partial x_k}$ will be considered here. The situation is then similar to that studied in section 3.3. Assumption (H2)(ii) then becomes, $||D^2 f||_d = \sup_{x \in \mathbb{R}^d} \frac{\|D^2 fx\|_d}{\|x\|_d} \leq M$. Similar derivations allow one to prove the following equation which is the equivalent of equation (13),

$$f(X^{n+1}) \leq f(X^n) + \sigma \sum_{k=1}^{d} (\frac{\partial f(X^n)}{\partial x_k})^2 \left[N_k^{n,i} + \frac{M}{2}\sigma(N_k^{n,i})^2\right] \ a.s.$$

from which derives exactly the same result than that of Lemma 3. In particular, the critical value σ_c, below which convergence takes place, is again defined by equations (11) and (12). 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 will discuss the results of previous section in the light of past related work from the literature.

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 Rudolph's work, either using also super martingale [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. [3]

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 a 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 almost useless.

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

[3] In this line, we would like to mention that there seems to be a lot of room for improvement in the proofs we present here (see [1]). Assumptions of regularity and convexity are likely to be relaxed. We are currently working on such extensions. Definite conclusions are however yet to be obtained

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.

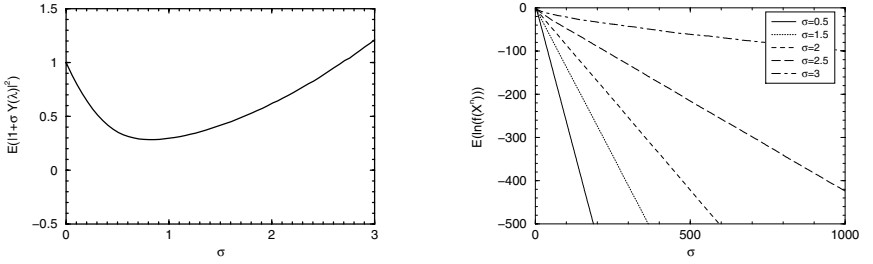
5 Numerical Experiments

All numerical experiments presented in the sequel are based on the Monte Carlo approximation of the expectation of a random variable. The expectation $\mathbb{E}(Z)$ of a random variable Z is approximated by $\frac{1}{K}\sum_{k=1}^{K} Z_k$, where Z_k are K independent 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,

$$\mathbb{E}(Z) \in [\frac{1}{K}\sum_{k=1}^{K} Z_k - \sqrt{VarZ}\frac{1.96}{\sqrt{K}}, \frac{1}{K}\sum_{k=1}^{K} Z_k + \sqrt{VarZ}\frac{1.96}{\sqrt{K}}] \tag{15}$$

5.1 Computation of the Constants

The Monte-Carlo method described above has been used to compute approximate values of the constants σ_c and σ_s from section 3.



(a) (b)

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

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 $\sigma_s(\lambda, 2) \approx 0.8$. Note that this method allows us to plot the progress rate (3.1) 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 σ_c, σ_s, $\sigma_c^{'}$ and $\sigma_s^{'}$ for some functions that are not quadratic, in order to test their optimality (whereas these constants

are known to be optimal in the case of quadratic functions, see section 3.3, where optimal means here that theses constants are the limit values between convergence and divergence.)

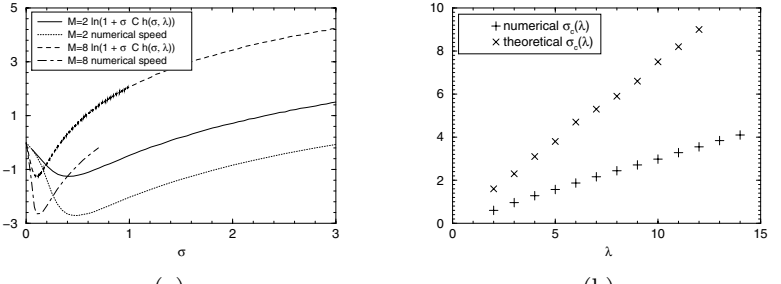
First, we need to circumvent a difficulty. Indeed, when evaluating $\mathbb{E}(f(X^n))$ with the Monte Carlo method, the relative error given by the Central Limit Theorem ($\sqrt{Var(f(X^n))}\frac{1.96}{\sqrt{K}\mathbb{E}(f(X^n))}$) grows geometrically with the number of generations n (the exact computation 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.

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

The functions f_M, defined by equation (16) 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)).

$$f_M(x) = \frac{M}{2} \begin{cases} x^2 & if x < 0 \\ x \arctan(x) & if x > 0 \end{cases} \tag{16}$$

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



(a) (b)

Fig. 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 2 (a) plots the slopes of those linear functions (determined using linear regressions), and the theoretical values $\sigma Ch(\sigma, \lambda, \alpha, M)$, 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

Convergence results and geometrical convergence rates for adaptive $(1, \lambda) - ES$ have been proved for a sub-class of C^2 functions. The optimality of the critical value for the step size and the resulting convergence rate have been proved for the sphere function and numerical experiments have demonstrated their validity for more general functions. The extension of the results to the d-dimensional case with a *non-isotropic* ES algorithm (14) leads to a critical value of the step-size and a convergence rate that are independent of the dimension, improving over previous work. On-going work is concerned with relaxing the regularity and convexity assumptions: it should be possible to nevertheless obtain. similar results for convergence and convergence rates. In addition on can envision the extension to a more practically useful algorithm, where the step-size is adapted proportionally to $|f(x) - f^*|$ (where f^* is the value at the global optimum). However, the d-dimension case of this latter algorithm will probably lead to dimension-dependent convergence rate. Finally, similar analysis should be possible for self-adaptive $(1, \lambda) - ES$, but probably requiring regularity assumptions on the objective function.

References

1. A.Auger. *ES, théorie et applications au contrôle en chimie.* PhD thesis, Université Paris 6, in preparation.
2. A. Auger, C. Le Bris, and M. Schoenauer. Rigorous analysis of some simple adaptative es. *Technical Report INRIA*, http://cermics.enpc.fr/∼auger/.
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 pitfalls. *Submitted*, 2001. 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.