Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Theoretical Computer Science 306 (2003) 269–289

Theoretical
Computer Science

www.elsevier.com/locate/tcs

Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties

Alexis Bienvenüe[a,*], Olivier François[b]

[a] *LMC, BP 53, 38041, Grenoble cedex 9, France*
[b] *TIMC-TIMB, Campus-Santé, F-38706, La Tronche cedex, France*

Received 20 December 2001; received in revised form 4 April 2003; accepted 15 April 2003
Communicated by G. Rozenberg

Abstract

This paper presents simple proofs for the global convergence of evolution strategies in spherical problems. We investigate convergence properties for both adaptive and self-adaptive strategies. Regarding adaptive strategies, the convergence rates are computed explicitly and compared with the results obtained in the so-called "rate-of-progress" theory. Regarding self-adaptive strategies, the computation is conditional to the knowledge of a specific induced Markov chain. An explicit example of chaotic behavior illustrates the complexity in dealing with such chains. In addition to these proofs, this work outlines a number of difficulties in dealing with evolution strategies.

Keywords: Evolution strategies; Global convergence; Markov chains

1. Introduction

For almost three decades, the theory of evolution strategies (ES) has focussed on convergence toward optima of simple objective functions [1,2]. Of these solvable models, the sphere problem is among the most frequently studied [13,5]. Despite the simplicity of this model, the analytical treatment of convergence issues however proves to be especially difficult.

In lieu of convergence assessments, the performances of evolution strategies are usually measured through a quantity called the *rate-of-progress* that evaluates the average

* Corresponding author.
E-mail address: alexis.bienvenue@imag.fr (A. Bienvenüe).

0304-3975/03/$ - see front matter
doi:10.1016/S0304-3975(03)00284-6

improvement after a single step of the algorithm. While interesting observations can be made from this measure, the question of whether the strategies converge or not remains open. For instance, practitioners often refer to [13, Fig. 5.9] for choosing good parameter settings in $(1,\lambda)$-ES for which a single parent is replaced by one of its λ offspring. The figure actually describes a relationship between the rate of progress and the universal step length (standard deviation). Although the figure suggests that the strategies should diverge from the optimum for large step lengths nothing establishes the result formally.

Several ways of tackling the convergence issue have been proposed in the literature. Rudolph used a mathematical tool called *martingale theory* that ensures the convergence of $(1+\lambda)$-ES [10,11]. Yin et al. have identified $(1,\lambda)$-ES as being relevant to the theory of *stochastic approximation* [14,15] . However, this theory has the drawback of relying on heavy stability hypotheses (e.g., Theorem 4.1 [15]), that can hardly be checked in practice, and few concrete results can be established following this approach.

So far, the most complete convergence theory has been achieved by Beyer [3–5] whose work concerns one-step measures as well as global convergence results. Nonetheless, Beyer's approach makes use of several approximations that are non-rigorous from the mathematical viewpoint. A first approximation called the *first order approximation* amounts to considering averaged instead of random behaviors. In a second kind of approximation, fluctuations are modeled as Gaussian noises which can be a crude approximation of the actual behavior. While partly confirmed by numerical experiments of the standard evolution strategies (i.e., Gaussian mutations, log-normal self-adaptation), these techniques can hardly be applied to other contexts.

In the next section (except for Section 3), we consider the problem of minimizing the sphere function

$$f(x) = \|x\|^2, \quad x \in \mathbb{R}^d, \ d \geqslant 1,$$

where $\|.\|$ denotes the d-dimensional Euclidean norm

$$\|x\|^2 = \sum_{i=1}^{d} x_i^2.$$

Following Eiben et al. [6], $(1,\lambda)$-ES can be classified according to the existing approaches for step length control: adaptive or self-adaptive. The purpose of this note is twofold. First, it presents simple global convergence proofs for adaptive and self-adaptive evolution strategies on the spherical problem (Sections 2 and 4). Second, it points out some intrinsic difficulties that seem to be associated with global convergence issues, and that make ES algorithms highly complex systems. Some of these difficulties are of geometrical nature (Sections 2–4), some other are strongly related to nonlinearities inside the dynamics (Section 5). For instance, a common pitfall in the evolutionary computation community consists of believing that having downhill oriented progress vectors is a sufficient condition for global convergence (see the remark after Theorem 3.1 in [15]). Section 2 shows that this can be false, and an additional difficulty comes in convex problems where progress vectors may not be oriented in the correct direction (Section 3). In Section 5, we provide an example of chaotic be-

havior underlying self-adaptive ES, and we compute the rate of convergence exactly. In this example, the ratio between the distance from the optimum and the current step length wanders as a sequence of uniform random variables, which is again a surprising observation which respect to the existing literature [4,5].

2. The $(1, \lambda)$-evolution strategy

Let λ be a fixed integer. The $(1,\lambda)$-ES strategy defines a Markovian dynamics for which a basic step is usually described as follows [13,5]. Let $x \in \mathbb{R}^d$ $(d \geqslant 1)$ be any initial arbitrary solution, and σ a positive constant called the *step length*. Let $\xi^1, \ldots, \xi^\lambda$ be λ random centered variables sampled from a symmetrical distribution of finite mean. Here, symmetry means that the distribution is invariant by any d-dimensional orthogonal transformation. A typical choice is the multivariate Gaussian distribution of covariance matrix identity I_d. The ξ^ℓ's are assumed to be independent of each other and from the past. The update rule consists of computing a new solution y as follows:

$$y = \arg\min\{f(x+\sigma\xi^1), \ldots, f(x+\sigma\xi^\lambda)\}, \quad \sigma > 0.$$

The distribution of the ξ^ℓ's is usually called the *offspring distribution*, and the integer λ corresponds to the number of offspring. The $(1,\lambda)$-ES dynamics in d dimensions is associated with a Markov transition kernel on $\mathbb{R}^d$ that we denote by $p_\sigma(y|x)$.

Historically, the first attempt to build a convergence theory of ES was directed toward the analysis of a single step of the algorithm, i.e., the solutions produced from the transition kernel $p_\sigma(y|x)$ [13]. This static approach is known as the *rate-of-progress theory*. It is based on a specific normalization rule in which the step length is divided by the Euclidean norm of the current solution

$$\sigma^* = \frac{\sigma}{r},$$

where

$$r = \|x\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

In the rate-of-progress theory, σ^* is kept constant, and is called the *universal step length*. The goal of this theory is seeking which parameter settings maximize a static quantity defined as the *progress rate* (see [13])

$$\phi(x) = \int \frac{\|y\| - \|x\|}{\|x\|}\, p_\sigma(y|x)\, \mathrm{d}y.$$

In this section, we shall consider the adaptive dynamics defined as follows. Let $X_0 = x$ be the initial solution, we take

$$X_{n+1} = \arg\min\{f(X_n+\sigma_n\xi^1), \ldots, f(X_n+\sigma_n\xi^\lambda)\}, \quad n \geqslant 0, \tag{1}$$

where

$$\sigma_n = \sigma^* r_n \tag{2}$$

and

$$r_n = \|X_n\|.$$

Although the above algorithm is seldom used by practitioners in solving real-world problems, it has the properties required in order to compare global convergence results with those obtained from the static approach. The definition of this adaptive strategy translates the hypothesis that σ_n/r_n should be constant. In contrast with previous works, we shall seek optimal universal step lengths on the basis of the convergence of the dynamics.

Let us start with a set of elementary remarks and further definitions. The transition kernel of the adaptive $(1,\lambda)$-ES is equal to

$$\hat{p}(y|x) = p_{\sigma^*\|x\|}(y|x),$$

where $\sigma = \sigma^*\|x\|$. Denote by $Y_x^{(\sigma)}$ (resp. $\hat{Y}_x$) and call *progress vector* associated to a step length σ (resp. universal step length σ^*), a random variable of probability distribution density $p_\sigma(x+\cdot|x)$ (resp. $\hat{p}(x+\cdot|x)$). The *average progress vectors* are quantities defined as

$$\zeta_\sigma(x) = E[Y_x^\sigma] = \int y\, p_\sigma(x+y|x)\,\mathrm{d}y, \tag{3}$$

and

$$\hat{\zeta}(x) = E[\hat{Y}_x] = \int y\, \hat{p}(x+y|x)\,\mathrm{d}y. \tag{4}$$

Note that the progress vectors possess an interesting rescaling property. Consider any $\alpha > 0$. Then we have

$$\alpha\hat{Y}_x = \alpha Y_x^{\sigma^*\|x\|} = Y_{\alpha x}^{\alpha\sigma^*\|x\|} = \hat{Y}_{\alpha x}, \tag{5}$$

where the identities hold in distribution, i.e., the progress vector starting from αx has the same distribution as α times the progress vector starting from x. As a consequence, the adaptive dynamics can be rescaled so that they restart from $x = e_1$ (or $x = -e_1$) at each generation, where $e_1 = (1,0,\ldots,0)$ is the unit vector in d dimensions.

First, we consider the case of one dimension ($d = 1$). Because, the distribution of $\hat{Y}_1$ will play a crucial role in computing a global convergence rate for (X_n), we give here its explicit description.

Lemma 2.1. *Let $d = 1$ and $\hat{Y}_1$ be the progress vector obtained by starting from $x = 1$ (with step length σ^*). Let g_{λ,σ^*} denote the probability density function of $\hat{Y}_1$. Then we have*

$$g_{\lambda,\sigma^*}(t) = \frac{\lambda}{\sigma^*}\, p_\xi\left(\frac{t}{\sigma^*}\right)\left[F_\xi\left(\frac{-|t+1|-1}{\sigma^*}\right) + 1 - F_\xi\left(\frac{|t+1|-1}{\sigma^*}\right)\right]^{\lambda-1},$$
$$t \in \mathbb{R},$$

where p_ξ and F_ξ are respectively the density and the cumulative distribution function of the offspring distribution.

Proof. Let $q_{\lambda,\sigma,x}$ be the density of $Y_x^{(\sigma)}$. Because $Y_x^{(\sigma)}$ equals t when one of the λ offspring equals $x+t$ and $f(x+t)$ is lower than all other offspring values, we have

$$q_{\lambda,\sigma,x}(t) = \frac{\lambda}{\sigma}\, p_\xi(t/\sigma)\Pr[f(x+\sigma\xi) > f(x+t)]^{\lambda-1}, \tag{6}$$

where ξ is sampled from the offspring distribution p_ξ. Rewriting this equation in the particular case $x=1$ and $\sigma=\sigma^*$ leads to the result. □

In one dimension, the adaptive $(1,\lambda)$-ES converges or diverges at a linear rate that can be computed as follows.

Theorem 2.1. *Let $d=1$ and (X_n) be the Markov chain associated to the adaptive $(1,\lambda)$-ES defined by Eqs. (1), (2) with universal step length σ^*, and $r_n=|X_n|$. Assume that $\hat{\zeta}(1)$ is finite. Then,*

$$\frac{1}{n}\ln r_n \to \kappa(\sigma^*,\lambda) \quad \text{as } n\to\infty,$$

where

$$\kappa(\sigma^*,\lambda) = \int_{-\infty}^{+\infty} \ln(|1+y|)\, g_{\lambda,\sigma^*}(y)\,\mathrm{d}y.$$

Proof. Assume that $r_0>0$. According to Eq. (5), the dynamics of (r_n) can be rescaled in the following way:

$$r_{n+1} = |X_n + \hat{Y}_{X_n}| = r_n|1+\operatorname{sign}(X_n)\hat{Y}_1|.$$

Using the symmetry of the offspring distribution, we obtain that

$$\ln r_{n+1} = \ln r_0 + \sum_{k=1}^{n} \ln\,|1+Y_k|,$$

where the Y_k's are i.i.d. random variables sampled from the probability density function g_{λ,σ^*}. The result follows from the strong law of large numbers. □

The function $\kappa(\sigma^*,\lambda)$ that appears in Theorem 2.1 can hardly be computed explicitly even for small values of λ. Fig. 2 displays the values of $\kappa(\sigma^*,3)$ obtained numerically in the case of Gaussian mutations. With $\lambda=3$ the optimal step length σ^*_{opt} is around 0.94 and the rate of convergence is greater than 1.1. Besides, the dynamics diverge for $\sigma^*>\sigma^*_c=4.73$. The shape of the curve is quite similar to Schwefel's curves (Fig. 5.9. [13]) that display the relationships between the rates of progress and the universal step length.

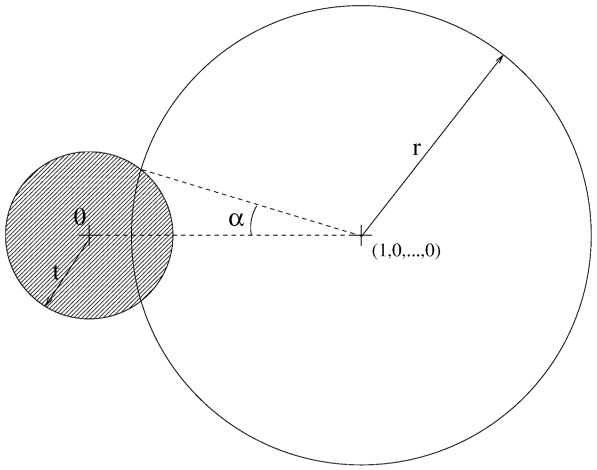

Fig. 1. The geometrical definition of $\alpha(r,t)$. $A(r,t)$ is the relative surface area of the sphere of radius r which does not intersect the sphere of radius t.

Let us now give a generalization of this result in d dimensions. The extension of Theorem 2.1 to $d \geqslant 2$ involves the computation of a new constant $\kappa(d,\sigma,\lambda)$. This constant will be defined as a double integral whatever the dimension. Denote by $\hat{Y}_1$ the progress vector after starting from e_1 with step length σ^*. We set

$$\hat{r}_1 = \|e_1 + \hat{Y}_1\|.$$

When all offspring norms are greater than t, we have $\hat{r}_1 > t$. This implies that

$$\Pr[\hat{r}_1 > t] = \Pr[\|e_1 + \sigma^* \xi\| > t]^{\lambda}. \tag{7}$$

Because spherical problems are invariant by orthogonal transformations, we have

$$\mathrm{P}[\|e_1 + \sigma^* \xi\| > t] = \int_0^{+\infty} A(r,t)\, p_{\xi}(r/\sigma^*)\, \mathrm{d}(r/\sigma^*), \tag{8}$$

where $A(r,t)$ is the relative surface area of the sphere of radius r centered at e_1 which does not intersect the sphere of radius t centered at O (see Fig. 1), and p_{ξ} is the probability density function of ξ.

If $r < 1-t$ or $r > 1+t$, then we take $A(r,t) = 1$. Otherwise, $A(r,t)$ is mathematically defined as

$$A(r,t) = \frac{2}{d\sqrt{\pi}} \frac{\Gamma(d/2+1)}{\Gamma(d/2-\frac{1}{2})} \int_0^{\alpha(r,t)} \sin^{d-2} u \, \mathrm{d}u.$$

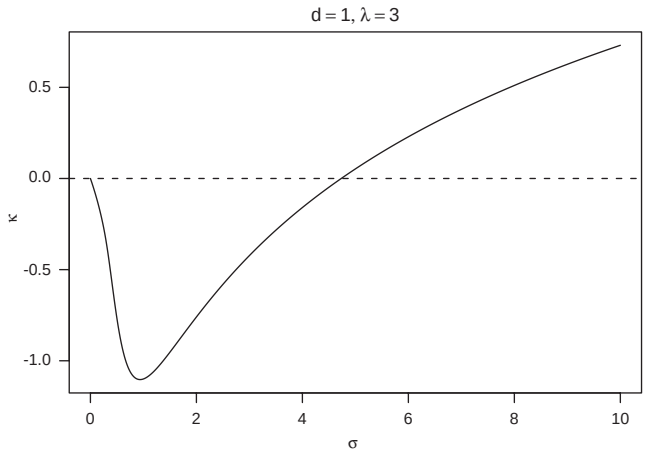

Fig. 2. The values of $\kappa(\sigma,3)$ obtained from three offspring with optimal step length $\sigma^*_{\mathrm{opt}} \approx 0.94$ in one dimension. The dynamics diverge for $\sigma^* > \sigma^*_c \approx 4.73$.

After calculation, we have

$$A(r,t) = \begin{cases} \dfrac{1}{\pi}\left[\alpha(r,t) - \cos\alpha(r,t) \displaystyle\sum_{j=0}^{(d-4)/2} \frac{2^{2j}(j!)^2}{(2j+1)!} \sin^{2j+1}\alpha(r,t)\right] & \text{for } n = 2k, \\ \dfrac{1}{2}\left[1 - \cos\alpha(r,t) \displaystyle\sum_{j=0}^{(d-3)/2} \frac{(2j)!}{2^{2j}(j!)^2} \sin^{2j}\alpha(r,t)\right] & \text{for } n = 2k+1, \end{cases}$$

where $\alpha(r,t) \in [0,\pi]$ is the half-angle defined as

$$\cos\alpha(r,t) = \frac{r^2 + 1 - t^2}{2r}.$$

Putting Eqs. (7) and (8) together, we find the probability density function f_R of $\hat{r}_1$

$$f_R(t) = \frac{\lambda}{\sigma^{*\lambda}} \int_0^{+\infty} \frac{\partial A(r,t)}{\partial t}\, p_\xi\left(\frac{r}{\sigma^*}\right) \mathrm{d}r \left(\int_0^{+\infty} A(r,t)\, p_\xi\left(\frac{r}{\sigma^*}\right) \mathrm{d}r\right)^{\lambda-1}. \qquad (9)$$

Finally, a proof similar to Theorem 2.1 leads to the following result.

Theorem 2.2. *Let (X_n) be the Markov chain defined by the adaptive $(1,\lambda)$-ES in Eqs. (1), (2), and $r_n = \|X_n\|$. Assume that $\hat{\zeta}(e_1)$ is finite. Then, we have*

$$\frac{1}{n} \ln r_n \to \kappa(d,\sigma^*,\lambda) \quad \text{as } n \to \infty,$$

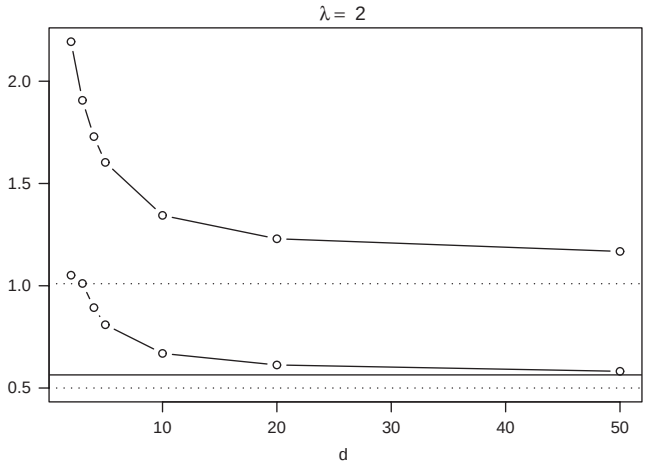

Fig. 3. The critical normalized step length $d\sigma_c^*(d)$ (up) and the optimal normalized step length $d\sigma_{\text{opt}}^*(d)$ (down) as functions of the dimension d for $(1,2)$-adaptive ES. The dotted lines correspond to the (constant) values obtained with respect to Schwefel's rate-of-progress theory. Schwefel's constant is 0.5. Beyer's constant is $c_{1,\lambda} = 0.5642$.

where

$$\kappa(d,\sigma^*,\lambda) = \int_0^{+\infty} \ln t f_R(t)\,\mathrm{d}t$$

and $f_R(t)$ *is given in Eq.* (9).

The improper integral $\kappa(d,\sigma^*,\lambda)$ cannot be expressed into a closed formula. When ξ is distributed according to the standard multivariate Gaussian distribution, the probability distribution of

$$\frac{1}{\sigma^*}\|e_1 + \sigma^*\xi\| = \left\|\frac{e_1}{\sigma^*} + \xi\right\|$$

is the shifted chi-square distribution with d degrees of freedom and location parameter $\mu = 1/2(\sigma^*)^2$. Its probability density function can be formulated as

$$f(t) = e^{-\mu} \sum_{n\geqslant 0} \frac{\mu^n}{n!} f_{\chi 2(d+2n)}(t), \quad t \geqslant 0.$$

In this situation, we obtain a simpler formula

$$f_R(t) = \frac{\lambda}{\sigma^{*\lambda}} f\left(\frac{t}{\sigma^*}\right)\left(\int_t^{+\infty} f\left(\frac{u}{\sigma^*}\right)\mathrm{d}u\right)^{\lambda-1}, \quad t \geqslant 0.$$

Based on numerical resolution, Figs. 3–5 allow us to make precise comparisons with the curves obtained from the rate-of-progress theory in the Gaussian framework. In Fig. 3, the number of offspring was taken equal to $\lambda = 2$, whereas it was equal to

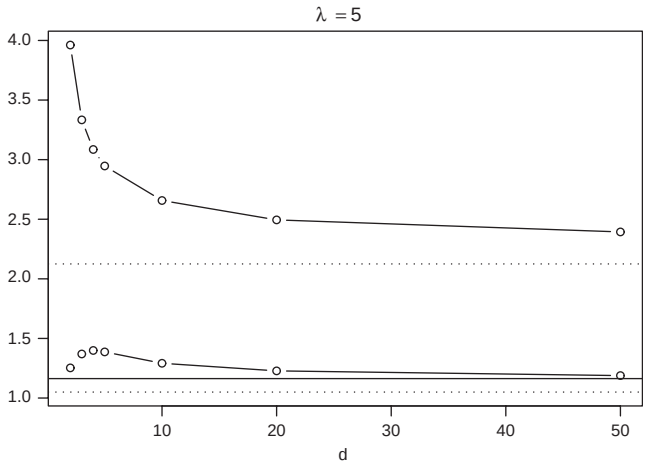

Fig. 4. The critical normalized step length $d\sigma_c^*(d)$ (up) and the optimal normalized step length $d\sigma_{\text{opt}}^*(d)$ (down) as functions of the dimension d for $(1,5)$-adaptive ES. The dotted lines correspond to the (constant) values obtained with respect to Schwefel's rate-of-progress theory. Schwefel's constant is 1.05. Beyer's constant is $c_{1,\lambda} = 1.1630$.

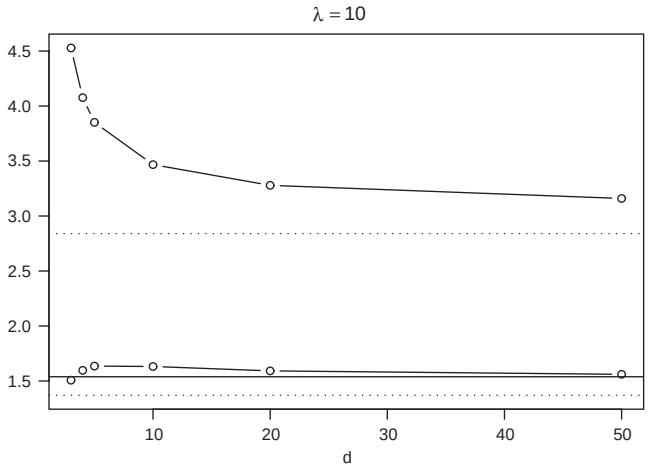

Fig. 5. The critical normalized step length $d\sigma_c^*(d)$ (up) and the optimal normalized step length $d\sigma_{\text{opt}}^*(d)$ (down) as functions of the dimension d for $(1,10)$-adaptive ES. The dotted lines correspond to the (constant) values obtained with respect to Schwefel's rate-of-progress theory. Schwefel's constant is 1.37. Beyer's constant is $c_{1,\lambda} = 1.5388$.

$\lambda = 5, 10$ in Figs. 4 and 5. Two curves are displayed in each figure. The upper curves describe the relationship between critical values σ_c^* for which $\kappa(d, \sigma_c^*, \lambda) = 0$ and the problem dimension d. The lower curves concern optimal values σ_{opt}^*. The dotted lines recall the values obtained from [13] and the straight lines recall those of [5]. The

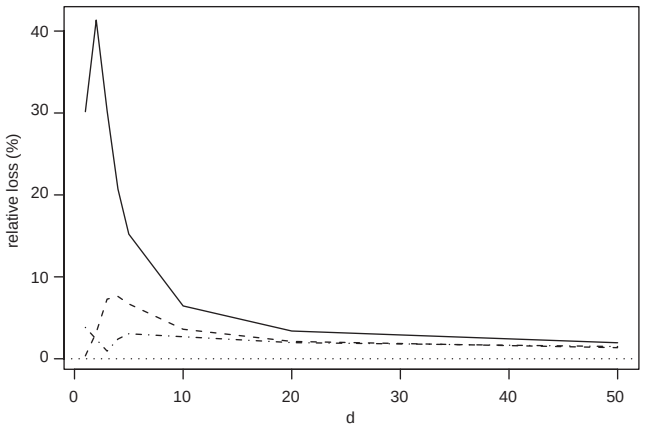

Fig. 6. Relative speed loss when using the optimal σ^*'s from Schwefel's rate-of-progress theory as a function of the dimension, for $\lambda = 2, 5, 10$ (from top to bottom).

second constant corresponds to $c_{1,\lambda}$, i.e. the values computed when d goes to infinity [5]. They are equal to $c_{1,2} = 0.5642$, $c_{1,5} = 1.1630$, and $c_{1,10} = 1.5388$. These values provide a good fit to the limit of $d\sigma^*_{\text{opt}}$, and this leads us to conjecture that

$$\sigma^*_{\text{opt}} \sim \frac{c_{1,\lambda}}{d}, \quad \text{as } d \to \infty.$$

Another remarkable fact is that Schwefel's critical and optimal values always underestimate the true critical and optimal step length, significantly for small dimensions. Because of this property, selecting optimal parameters from the rate-of-progress approach always warrant that global convergence holds. Nevertheless, our work shows that such choices can be improved. Relative losses in performances using Schwefel's constants can be found in Fig. 6. For large dimensions, the relative losses are however lower than 2–3% and Schwefel's approximations can be considered accurate.

3. Remarks on convex problems

Because the function $f(x) = x^2$ is symmetric, the average progress vector $\hat{\zeta}(x)$ is always oriented in the downhill direction (see Lemma 3.1). Believing that this ensures convergence is a common pitfall in evolutionary computation. For instance, Yin et al. [15] used this argument to justify the global convergence of more general adaptive strategies. Theorem 2.1 shows that the algorithm may be divergent even when average progress vector is oriented properly (because κ can take positive values).

This section details another remark about convex problems which have been considered as natural extensions to spherical problems in [10]. The convergence properties of adaptive $(1, \lambda)$-ES obtained in Theorem 2.1 are more a consequence of symmetry than a consequence of convexity. For convex problems, the progress vector is *not*

always oriented in the downhill direction, as will be shown by Theorem 3.1, which also provides a counterexample to Proposition 3.1 in [15].

In this section, the adaptive algorithm using the step length $\sigma = \sigma^* |x|$ will be replaced, as in [15], by

$$\sigma = H(f_x(x)), \tag{10}$$

where f_x denotes the gradient of f, and $H : \mathbb{R} \to \mathbb{R}_+$ is a real-valued function for which $H(x) = 0$ implies $x = 0$. This choice generalizes $\sigma = \sigma^* |x|$ to non-spherical problems. In this section, notations $\hat{p}$, $\hat{\zeta}$ and $\hat{Y}$ will refer to the adaptation scheme defined by (10) instead of (2). Let us start with a simple lemma.

Lemma 3.1. *Let x be fixed. If f is such that, for all $t > 0$, $f(x+t) > f(x-t)$, then $\hat{\zeta}(x) < 0$.*

If f is symmetric ($f(x) = f(-x)$), then $\hat{\zeta}(x)$ is always downhill oriented.

Proof. This is a simple consequence of (6). Indeed, for $t > 0$, we have

$$\Pr[f(x+\sigma Z) > f(x+t)] < \Pr[f(x+\sigma Z) > f(x-t)]. \qquad \square$$

Next, set $H(x) = |x|$, and consider the function f defined by

$$f(x) = \begin{cases} hx^2 & \text{if } x < 0, \\ hax^2 & \text{otherwise,} \end{cases} \tag{11}$$

where $a > 0$ and $h > 0$. With this particular function, the rescaling property (5) can be rewritten in the following way (the additional superscripts h and a are used to emphasize the dependence on the new parameters): For $\alpha > 0$ and $h' > 0$, let

$$Y_x^{(\sigma,h,a)} = Y_x^{(\sigma,h',a)}, \tag{12}$$

$$Y_x^{(\sigma,h,a)} = \alpha Y_{x/\alpha}^{(\sigma/\alpha,\alpha^2 h,a)}. \tag{13}$$

In distribution, we have

$$\hat{Y}_{\alpha x} = \alpha \hat{Y}_x. \tag{14}$$

Theorem 3.1. *Let $\hat{\zeta}(x)$ be the average progress vector for adaptive $(1,\lambda)$-ES, as in Eq. (4). For all $a > 0$, there exists a value $h_1(a)$ such that, for all $h > h_1(a)$, the sign of $\hat{\zeta}(x)$ is given by Table 1*:

Table 1

	$x < 0$	$x > 0$
$a < 1$	+	+
$a = 1$	+	−
$a > 1$	−	−

Proof. The cases $a=1$, $a<1, x<0$ and $a>1, x>0$ follow from Lemma 3.1. We inspect the case $a>1$, $x<0$. The remaining case can be treated similarly.

Thanks to Eq. (14), $\hat{\zeta}(x)$ has the sign of $\hat{\zeta}(-1)$. Using (12) and (13), we have

$$\begin{aligned}\hat{\zeta}(-1) &= E[Y_{-1}^{(2h,h,a)}] \\ &= E[Y_{-1/h}^{(2,h^3,a)}] = E[Y_{-1/h}^{(2,1,a)}].\end{aligned}$$

Hence, $\hat{\zeta}(-1)$ goes to $E[Y_0^{(2,1,a)}]$ as h goes to infinity. So we need to prove that $E[Y_0^{(2,1,a)}]$ is negative to complete the proof. Thanks to Eq. (6), we can write $E[Y_0^{(2,1,a)}] = I_+ + I_-$, where

$$I_+ = \lambda \int_0^{+\infty} t(1 - F_\xi(t/\sigma) + F_\xi(-t\sqrt{a}/\sigma))^{\lambda-1} p_\xi(t/\sigma)/\sigma \, \mathrm{d}t,$$

and

$$\begin{aligned}I_- &= \lambda \int_{-\infty}^{0} t(F_\xi(t/\sigma) + 1 - F_\xi(-t/\sigma\sqrt{a}))^{\lambda-1} p_\xi(t/\sigma)/\sigma \, \mathrm{d}t \\ &= -\lambda \int_0^{+\infty} t(1 - F_\xi(t/\sigma) + F_\xi(-t/\sigma\sqrt{a}))^{\lambda-1} p_\xi(t/\sigma)/\sigma \, \mathrm{d}t.\end{aligned}$$

But, for $t>0$ and $a>1$, $-t/\sigma\sqrt{a} > -t\sqrt{a}/\sigma$, so that $E[Y_0^{(2,1,a)}] = I_+ + I_- < 0$ for all $\sigma>0$. □

From the table of Theorem 3.1, we can conclude that the average progress vector is not always downhill oriented in the adaptive $(1,\lambda)$-ES. To see why Theorem 3.1 provides a counterexample to Proposition 3.1 of [15], f can be transformed into a twice differentiable function with bounded derivatives by modification in a small neighborhood of zero. The modified function matches with the hypothesis of [15], but their conclusions are erroneous.

Regarding the particular landscape f, the behavior of the adaptive $(1,\lambda)$-ES can be studied again.

Theorem 3.2. *Consider the Markov chain (X_n) corresponding to the adaptive $(1,\lambda)$-ES algorithm and the objective function f given in Eq. (11). Let $r_n = |X_n|$. We have*

$$\frac{1}{n} \ln r_n \to K_{h,a} \quad \text{as } n \to \infty,$$

where the constant $K_{h,a}$ can be computed in terms of the offspring distribution in the following way:

$$K_{h,a} = \sum_{\varepsilon \in \{\pm 1\}} \frac{p_{-\varepsilon,\varepsilon}}{p_{\varepsilon,-\varepsilon} + p_{-\varepsilon,\varepsilon}} E[\ln |1 + \varepsilon \hat{Y}_\varepsilon|],$$

and

$$p_{\varepsilon,\varepsilon'} = \Pr[\varepsilon\varepsilon'(1 + \varepsilon \hat{Y}_\varepsilon) > 0].$$

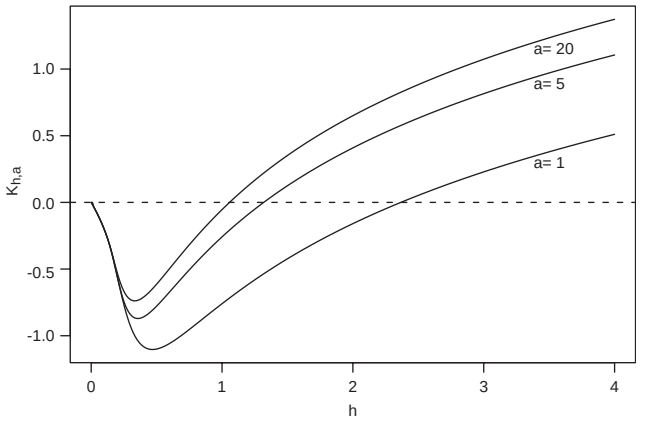

Fig. 7. The values of $K_{h,a}$ for the adaptive $(1,3)$-ES and the non-symmetric objective function defined by (11), as a function of h, for $a = 1, 5, 20$.

Proof. From Eq. (14), one step of the adaptive $(1,\lambda)$-ES is given by

$$x \mapsto x + |x|\hat{Y}_{\mathrm{sign}(x)} = x[1 + \mathrm{sign}(x)\hat{Y}_{\mathrm{sign}(x)}].$$

Let $S_n = \mathrm{sign}(X_n)$. The process $(S_n)_n$ is a Markov chain with transition probabilities $p_{\varepsilon,\varepsilon'}$, with stationary distribution π given by

$$\pi(\varepsilon) = \frac{p_{-\varepsilon,\varepsilon}}{p_{-\varepsilon,\varepsilon} + p_{\varepsilon,-\varepsilon}}.$$

In addition, the process $(S_n, |X_{n+1}/X_n|)_n$ is a process defined on the Markov chain (S_n), so that the process $(\ln|X_{n+1}/X_n|)_n$ satisfies the law of large numbers (see e.g. [7]). □

The value of $K_{h,a}$ can be obtained numerically. When $K_{h,a}$ is negative (resp. positive), the algorithm converges (resp. diverges) linearly. Fig. 7 gives the value of $K_{h,a}$ as a function of h, for several values of a, and for a standard Gaussian offspring distribution. Remark that the asymmetry of f is parameterized by a. The constant $K_{h,a}$ increases as a increases: Symmetry is a factor that improves convergence speed.

4. The self-adaptive $(1, \lambda)$-evolution strategy

In self-adaptive $(1,\lambda)$-ES, the process of evolution is exploited to determine which changes are the most advantageous with respect to the fitness of individuals. A major difference with adaptive evolution strategies is that the step length is then evolved by the evolutionary algorithm rather than exogenously defined.

The self-adaptive ES can be defined as follows. Let $X_0 \in \mathbb{R}^d$ $(d \geqslant 1)$ be any initial arbitrary solution. Let X_n be the solution obtained after n steps and σ_n be the associated

step length. The solution at time $n+1$ is computed from a sample $\xi^1,\dots,\xi^\lambda$ of λ independent random centered variables taken from a symmetrical d-dimensional probability distribution, and a sample $\eta^1,\dots,\eta^\lambda$ of λ independent nonnegative one-dimensional random variables. More specifically, we have

$$X_{n+1} = \arg\min\{f(X_n + \sigma_n\eta^1\xi^1),\dots,f(X_n + \sigma_n\eta^\lambda\xi^\lambda)\}. \tag{15}$$

In addition, the step lengths σ_n are updated according to the multiplication by the variable $\eta^\star$ that leads to the best increment at each iteration. The symbol $\star$ indicates which label corresponds to the selected offspring

$$\begin{aligned}\sigma_{n+1} &= \sigma_n\eta^\star,\\ X_{n+1} &= X_n + \sigma_n\eta^\star\xi^\star.\end{aligned} \tag{16}$$

In typical choices, the ξ^ℓ's are multivariate Gaussian random variables with diagonal covariance matrices, and the η^ℓ's are usually sampled according to standard lognormal distributions.

Denote by X_z^1, $\eta_z^{1\star}$, $\xi_z^{1\star}$ the variables obtained after a single step of the dynamics described by Eq. (16). The subscript z indicates that we start from $X_n = ze_1$ where e_1 is the unit vector $e_1 = (1,0,\dots,0)$, and z is a nonnegative number. The superscript 1 means that we take $\sigma_n = 1$. Hence, we have

$$X_z^1 = ze_1 + \eta_z^{1\star}\xi_z^{1\star}.$$

In addition, we consider the Euclidean norm of X_z^1 obtained as follows:

$$r_z^1 = \|X_z^1\|.$$

The following lemma establishes a useful result: The Euclidean norm $r_n = \|X_n\|$ can be rescaled so that the normalized variables $Z_n = r_n/\sigma_n$ possess their own autonomous Markovian dynamics.

Lemma 4.1. *Let (X_n) be defined as in Eq. (15) and*

$$Z_n = r_n/\sigma_n$$

with $r_n = \|X_n\|$. Then (Z_n) is an homogeneous Markov chain. Starting from $Z_0 = z$, a single step of this chain yields a random variable whose distribution is the same as

$$Z_1 = \frac{r_z^1}{\eta_z^{1\star}}.$$

Proof. In order to emphasize the rescaling, Eqs. (15) and (16) can be rewritten as follows:

$$Z_{n+1} = \frac{\|X_{n+1}\|}{\sigma_{n+1}} = \frac{\|X_n/\sigma_n + \eta^\star\xi^\star\|}{\eta^\star}, \tag{17}$$

where

$$\bigstar = \underset{1\leqslant \ell \leqslant \lambda}{\arg\min} \left\| \frac{X_n}{\sigma_n} + \eta^{\ell}\xi^{\ell} \right\|. \tag{18}$$

According to the symmetry of the offspring distribution, this is equivalent to

$$Z_{n+1} = \frac{\|Z_n + \eta^{\bigstar}\xi^{\bigstar}\|}{\eta^{\bigstar}}, \tag{19}$$

and

$$\bigstar = \underset{1\leqslant \ell \leqslant \lambda}{\arg\min} \left\| Z_n + \eta^{\ell}\xi^{\ell} \right\|. \tag{20}$$

We see from Eq. (19) that Z_{n+1} depends on the past through Z_n only, and this property is shared by $\eta^{\bigstar}$ and $\xi^{\bigstar}$ as well. This proves that (Z_n) is an homogeneous Markov chain. Starting from $Z_0 = z$, we see that

$$Z_1 = \frac{\|ze_1 + \eta_z^{1\bigstar}\xi_z^{1\bigstar}\|}{\eta^{1\bigstar}} = \frac{r_z^{1\bigstar}}{\eta^{1\bigstar}}. \qquad \square \tag{21}$$

Comments. Lemma 4.1 actually states that the Z evolution can be decoupled from the r evolution. Hence it provides a general proof of Beyer's result that the *σ evolution can be decoupled from the r evolution* [4]. In addition, we have shown that this property is a natural consequence of the symmetry in the model, and is independent of the nature of mutations (ξ and η are arbitrary variables). It has also been shown within the framework of progress rate theory, that the evolution of normalized mutation strength can be described by a Chapman–Kolmogorov equation. For all measurable set B, the Chapman–Kolmogorov equations can be written as

$$\mathrm{P}_{Z_{n+1}}(B) = \int_B \int_0^{\infty} \mathrm{P}_{Z_n}(\mathrm{d}z)K(z,\mathrm{d}t),$$

where the transition kernel can be computed as

$$K(z,\mathrm{d}t) = \mathrm{P}(r_z^1/\eta_z^{1\bigstar} \in \mathrm{d}t).$$

When studying the r_n evolution, the Z_n's play the role of hidden variables. The dynamics of Z_n are independent of r_n but in return the Z_n's act as latent variables in the r_n evolution. The model is therefore relevant to the theory of Hidden Markov chains (e.g., [9]). When the dynamics of (Z_n) are sufficiently mixing, a unique stationary probability distribution exists, and can be found as the solution to the following integral equation:

$$\mu(\mathrm{d}t) = \mu K(\mathrm{d}t) = \int_0^{\infty} \mu(\mathrm{d}z)K(z,\mathrm{d}t). \tag{22}$$

From an algorithmic point of view, the Markov chain (Z_n) must enjoy good stability properties. Irreducibility ensures that every set A will be visited by the chain but

this property is too weak to guarantee that Z_n will enter A often enough. Here, we assume that (Z_n) is Harris-recurrent [8]. Harris-recurrence is a stronger concept than irreducibility. Let us recall this concept. Let N_A be the number of passages in A. The set A is *Harris-recurrent* if

$$\mathrm{P}_z(N_A = \infty) = 1, \quad z \in A.$$

The chain (Z_n) is Harris-recurrent if there exists a measure ψ such that (Z_n) is ψ-irreducible and for every set A with $\psi(A) > 0$, A is Harris-recurrent. We have the following result.

Theorem 4.1. *Let (X_n) be defined as in Eq. (15) and $Z_n = r_n/\sigma_n$ with $r_n = \|X_n\|$. Assume that the Markov chain (Z_n) is Harris-recurrent. Then (X_n) either converges or diverges linearly. The rate of convergence is given by*

$$\frac{1}{n} \ln r_n \to \kappa = \int_0^\infty E[\ln(r_z^1/z)]\, \mathrm{d}\mu(z), \quad \text{as } n \to \infty,$$

where μ is the solution of the integral Eq. (22).

Proof. Let

$$Y_z^1 = \eta_z^{1\star} \zeta_z^{1\star} = X_z^1 - z e_1.$$

Assume that $r_0 > 0$. We have

$$\ln r_{n+1} = \ln r_0 + \sum_{k=1}^{n} \ln \left\| e_1 + \frac{Y_{Z_k}^1}{Z_k} \right\|.$$

By the law of large numbers holds (see [8,9]), we have

$$\frac{1}{n} \ln r_n \to \int_0^\infty E[\ln \|e_1 + Y_z^1/z\|]\, \mathrm{d}\mu(z), \quad \text{as } n \to \infty. \qquad \square$$

Theorem 4.1 actually points out that global convergence (or divergence) of the self-adaptive $(1, \lambda)$-ES can be decided from the inspection of a much simpler chain than the original one. The theorem shows that the Harris-recurrence of (Z_n) is a crucial step in establishing global convergence of ES. Several technical hypotheses allow proving the Harris-recurrence property [8,9]. The most popular criterion consists of showing that the only bounded harmonic functions are constant. Nevertheless, even with the simplest assumptions about ξ and η, checking this criterion remains difficult for self-adaptive $(1, \lambda)$-ES's.

Again, convergence occurs at a linear rate. The rate of convergence can be expressed as a two-dimensional integral as follows:

$$\int_0^\infty E[\ln(r_z^1/z)]\, \mathrm{d}\mu(z) = \int_0^\infty \int_{-\infty}^\infty \ln |z'/z|\; p_1(z'|z)\, \mathrm{d}z' \mathrm{d}\mu(z). \tag{23}$$

Computing expression (23) remains difficult unless the stationary distribution μ is known, and the next section will show that this distribution may take unexpected

shapes. Nevertheless, if we assume an infinite number of offspring ($\lambda = \infty$), the single step dynamics becomes biased toward local average improvement

$$E[\|ze_1 + \eta_z^{1\star}\xi_z^{1\star}\|] < z \quad \text{for all } z > 0.$$

In this situation, we have

$$\ln E[\|ze_1 + Y_z^1\|\,] < \ln z.$$

By Jensen's inequality, we obtain that

$$E[\ln \|ze_1 + Y_z^1\|] < \ln E[\|ze_1 + Y_z^1\|],$$

and $\kappa < 0$, i.e., the self-adaptive ES converges. Remark that checking the above condition does not involve any self-adaptation property because it assumes that $\sigma = 1$.

In view of further progress in building a rigorous theory, mathematical efforts should be directed toward the understanding of the recurrence and the stability of (Z_n). A step in this direction was made by Beyer [4] who proposed conditions for the recurrence of $1/Z_n = \sigma_n/r_n$ bearing on the first momentum of the self-adaptation distribution.

We ran numerical simulations of the one-dimensional dynamics in which we took ξ to be Gaussian and η sampled according to the Gamma distribution. Using the shape parameter $\alpha = 1$ and the scale $\beta = 0.6$ in the self-adaptation distribution, we observed that the simulations diverge. On the other hand, they converge when $\alpha = 10$ and $\beta = 0.06$ (we used $\lambda = 10$ offspring in this experiment). Since the expectation of the Gamma distribution is $\alpha\beta$, a condition bearing on the first momentum cannot be able to discriminate between convergence and divergence, and more complex conditions must be investigated.

5. Example of chaos under self-adaptive-ES

In this section, we consider a strongly simplified model of self-adaptive ES and show that its dynamics underly a chaotic behavior. Some intrinsic difficulties with these algorithms are related to nonlinearities inside the dynamics, and make self-adaptive ES belong to the class of highly complex systems.

In our model, both η and ξ's are Bernoulli random variables. More specifically, $\xi^1, \ldots, \xi^\lambda$ are created according to the following model. For $\ell = 1, \ldots, \lambda$,

$$\xi^\ell = \begin{cases} -1 \text{ with probability } 1/2, \\ +1 \text{ with probability } 1/2. \end{cases} \tag{24}$$

The λ random variables $\eta^1, \ldots, \eta^\lambda$ are generated as follows. For $\ell = 1, \ldots, \lambda$,

$$\eta^\ell = \begin{cases} 1/2 \text{ with probability } 1/2, \\ 2 \quad \text{with probability } 1/2. \end{cases} \tag{25}$$

This model consists of an obvious discrete version of the standard self-adaptive $(1, \lambda)$-ES in which mutations are Gaussian and the self-adaptive rule uses lognormal random

variables. Here Gaussian variables are merely replaced with symmetrical Bernoulli random variables.

As in Section 4, we consider the normalized states

$$Z_n = \frac{|X_n|}{\sigma_n}, \quad n \geqslant 0.$$

Given $Z_n = z$, the variable Z_{n+1} can take one of the four following values:

$$|z/2 - 1|, \quad |2z - 1|, \quad |2z + 1|, \quad |z/2 + 1|.$$

Define the following probabilities

$$\begin{aligned} p_1(\lambda) &= 1 - (\tfrac{3}{4})^{\lambda}, \\ p_2(\lambda) &= (\tfrac{3}{4})^{\lambda}(1 - (\tfrac{2}{3})^{\lambda}), \\ p_3(\lambda) &= (\tfrac{1}{2})^{\lambda}(1 - (\tfrac{1}{2})^{\lambda}), \\ p_4(\lambda) &= (\tfrac{1}{2})^{\lambda}. \end{aligned}$$

According to the value of z, the $p_i(\lambda)$'s describe the chances that Z_{n+1} equals $|z/2-1|$ or $|2z-1|$ or $|2z+1|$ or $|z/2+1|$ up to a permutation of these four values. This is routine to check that the most probable transitions can be described as follows:

If $Z_n = z \leqslant 1/2$, then $Z_{n+1} = 1 - 2z$ with probability $p_1(\lambda)$,

If $Z_n = z \in]1/2, 5/4]$, then $Z_{n+1} = 2z - 1$ with probability $p_1(\lambda)$,

If $Z_n = z \in]5/4, 2]$, then $Z_{n+1} = 1 - z/2$ with probability $p_1(\lambda)$,

If $Z_n = z > 2$, then $Z_{n+1} = z/2 - 1$ with probability $p_1(\lambda)$.

Since the model is not amenable to an exact analysis, our focus is on the limiting dynamics obtained by taking an infinite number of offspring, i.e., $\lambda = \infty$. In this limit, the dynamics become deterministic

$$Z_{n+1} = T(Z_n),$$

where T is obtained from the above description and is given in Fig. 8. After a finite number of steps, the deterministic transformation T yields a sequence in $(0, 1)$ that (theoretically[1]) has the same behavior as a sequence of random numbers distributed according to the uniform distribution [12]. For this model, the stationary distribution of Z_n can therefore be determined exactly. The next result gives the rate of convergence explicitly.

Theorem 5.1. *Consider the dynamics defined by Eqs.* (24), (25) *and* $\lambda = \infty$ (i.e. $p_1(\lambda) = 1$). *Then, we have*

$$\frac{1}{n} \ln r_n \to -\ln 2 \quad a.s.$$

Proof. Let T be the deterministic transformation on the interval $[0, 1]$ such that

$$Z_{n+1} = T(Z_n).$$

[1] Given the finite representation of floating point numbers in the computer, simulated sequences might converge to a fixed value.

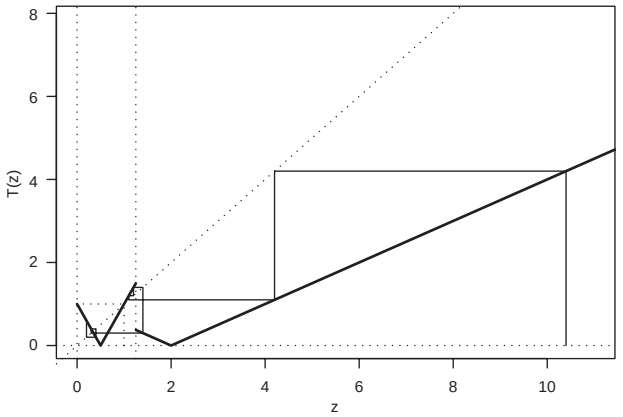

Fig. 8. The graph of T (boldface) and a sequence of numbers obtained from the iterations of $z_{n+1} = T(z_n)$.

By the ergodic theorem [12] and Theorem 4.1, $\ln r_n/n$ converges to

$$K = \int_0^1 E[\ln(r_{n+1}/r_n) \,|\, Z_n = z] \, \mathrm{d}z.$$

Given $Z_n = z$ in $(0,1)$,

$$\frac{r_{n+1}}{r_n} = \left| \frac{X_{n+1}}{\sigma_{n+1}} \frac{\sigma_{n+1}}{\sigma_n z} \right| = \left| \frac{1-2z}{2z} \right|.$$

Then, we have

$$K = \int_0^1 \ln \left| \frac{2z-1}{2z} \right| \mathrm{d}z = -\ln 2. \qquad \square$$

Comments. In other existing examples [5], the stationary distributions of Z_n are approximated as peaky distributions of shape close to the lognormal or the Gamma densities. Because our model is a discrete version of the standard self-adaptive ES, the fact that Z_n converges to the uniform distribution on $[0,1]$ was rather unexpected. This points out the issue of the robustness of the ES dynamics with respect to the hypothesis of which distributions are used for sampling the mutations.

6. Discussion

This paper has presented elementary proofs for the global convergence of adaptive and self-adaptive evolution strategies in simplified frameworks.

Regarding adaptive strategies, our work outlines that selecting parameters according to the optimality of the rate of progress (the traditional approach) is a circumspect

and risk less approach. The true optimal parameters are actually underestimated in this traditional approach and global convergence is always guaranteed. Nevertheless, our results enable quantifying the relative loss in convergence speed when using these parameters. Computing the exact rate of convergence of $(1,\lambda)$-ES is neither more difficult nor more computationally intensive than computing rates of progress, and the benefit is obvious. In view of further works, note that global convergence can be obtained for other strategies in a similar way. Recombinant strategies or noisy sphere problems are indeed amenable to the same kind of analysis.

Several authors have attempted to extent convergence results to problems different from the sphere. We have underlined a number of potential pitfalls in doing so. In convex problems for instance, the "average progress vector" $\zeta(x)$ is not necessarily oriented in the downhill direction (a condition for convergence). This fact emphasizes the difficulty in defining classes of problems for which the strategies work well.

Mathematical analyses of self-adaptive strategies are even more difficult as they involve Markov chains whose behavior can be complex. Proving linear convergence and estimating convergence rates can nevertheless be done through the examination of a simpler induced Markov chain. We have constructed an example for which the behavior of this simpler Markov chain can be studied exactly. This example illustrates the complexity in dealing with self-adaptation where chaotic behaviors may underpin the dynamics. In the future, a challenging issue will consist of exhibiting recurrence and stability conditions for this induced chain.

Acknowledgements

This work is supported by the AlPB project. We are grateful to the anonymous referee for her/his useful comments on the first version of the manuscript.

References

[1] T. Bäck, Evolutionary Algorithms in Theory and Practice, Oxford University Press, New York, 1996.
[2] T. Bäck, H.P. Schwefel, An overview of evolutionary algorithms for parameters optimization, Evol. Comput. 1 (1993) 1–24.
[3] H.-G. Beyer, Toward a theory of evolution strategies: some asymptotical results from the $(1,+\lambda)$-theory, Evol. Comput. 1 (2) (1993) 165–188.
[4] H.-G. Beyer, Toward a theory of evolution strategies: self-adaptation, Evol. Comput. 3 (3) (1996) 311–347.
[5] H.-G. Beyer, The Theory of Evolution Strategies, Natural Computing Series, Springer, Heidelberg, 2001.
[6] A.E. Eiben, R. Hinterding, Z. Michialewicz, Parameter control in evolutionary algorithms, IEEE Trans. Evol. Comput. 3 (1999) 124–141.
[7] J. Janssen, Les processus $(J-X)$, Cahiers Centre Études Rech. Opér. 11 (1969) 181–214.
[8] S. Meyn, R. Tweedie, Markov Chains and Stability Analysis, Springer, New York, 1994.
[9] C.P. Robert, G. Casella, Monte-Carlo Markov Chains, Springer, New York, 1999.
[10] G. Rudolph, Convergence Properties of Evolutionary Algorithms, Kovac, Hamburg, 1997.
[11] G. Rudolph, Finite Markov chain results in evolutionary computation: a tour d'horizon, Fund. Inform. 35 (1998) 1–22.
[12] D. Ruelle, Chaotic Evolution and Strange Attractors, Cambridge University Press, Cambridge, 1987.

[13] H.P. Schwefel, Evolution and Optimum Seeking, Wiley, New York, 1995.
[14] G. Yin, G. Rudolph, H.-P. Schwefel, Establishing connections between evolutionary algorithms and stochastic approximation, Informatica 6 (1) (1995) 93–116.
[15] G. Yin, G. Rudolph, H.-P. Schwefel, Analyzing $(1, \lambda)$ evolution strategy via stochastic approximation methods, Evol. Comput. 3 (4) (1996) 473–489.