Log-linear Convergence and Optimal Bounds for the $(1+1)$-ES

Mohamed Jebalia, Anne Auger, Pierre Liardet

Mohamed Jebalia, Anne Auger, and Pierre Liardet

Log-linear Convergence and Optimal Bounds for the $(1+1)$-ES

Proceedings of Evolution Artificielle 2007, pp 207-218.

Errata :

Proof of Lemma 1: The surface area of the d-dimensional unit ball should read $S_d = 2\pi^{d/2}/\Gamma(\frac{d}{2})$.
Proof of Proposition 2, last equation: The right hand-side of the first ligne should be $\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(\ln^- \left(\left\| \frac{\mathrm{X}_m}{\|\mathrm{X}_m\|} + \sigma x \right\| \right) \right) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma)$.

Log-linear Convergence and Optimal Bounds for the (1 + 1)-ES

Mohamed Jebalia[1], Anne Auger[1], and Pierre Liardet[2]

[1] TAO Team, INRIA Futurs
Université Paris Sud, LRI
91405 Orsay cedex, France
{mohamed.jebalia,anne.auger}@lri.fr
[2] Université de Provence
UMR-CNRS 6632, 39 rue F. Joliot-Curie
13453 Marseille cedex 13, France
liardet@cmi.univ-mrs.fr

Abstract. The (1 + 1)-ES is modeled by a general stochastic process whose asymptotic behavior is investigated. Under general assumptions, it is shown that the convergence of the related algorithm is sub-log-linear, bounded below by an explicit log-linear rate. For the specific case of spherical functions and scale-invariant algorithm, it is proved using the Law of Large Numbers for orthogonal variables, that the linear convergence holds almost surely and that the best convergence rate is reached. Experimental simulations illustrate the theoretical results.

1 Introduction

Evolutionary algorithms (EAs) are bio-inspired stochastic search algorithms that iteratively apply operators of variation and selection to a population of candidate solutions. Among EAs, adaptive Evolution Strategies (ESs) are recognized as state of the art algorithms when dealing with continuous optimization problems. Adaptive ESs sequentially adapt the parameters of the search distribution, usually a multivariate normal distribution, based on the history of the search. Several adaptation schemes have been introduced in the past. The one-fifth success rule [1, 2] considers the adaptation of one parameter, referred as the step-size, based on the success probability. The most advanced adaptation scheme, the Covariance Matrix Adaptation (CMA), adapts the full covariance matrix of the multivariate normal distribution [3].

The first theoretical works carried out in the context of Evolution Strategies focused on the so-called progress rate defined as a one-step expected progress towards the optimum [1, 4]. The progress rate approach consists in looking for step-sizes maximizing the expected progress. This amounts to investigating an artificial step-size adaptation scheme called scale-invariant, in which, at each iteration, the step-size is proportional to the distance to the optimum. The results derived in the context of the progress rate theory hold asymptotically in the

dimension of the search space and the techniques used do not allow to obtain finite dimension estimations.

Finite dimension results were obtained in the context of 'comma' strategies on the class of the so-called sphere functions, mapping $\mathbb{R}^d$ into $\mathbb{R}$ (d being the dimension of the search space) and defined as

$$f(x) = \mathrm{g}(\|x\|^2)\,, \tag{1}$$

where $\mathrm{g} : [0, +\infty[\mapsto \mathbb{R}$ is an increasing function and $\|.\|$ denotes the usual euclidian norm on $\mathbb{R}^d$. On this class of functions, scale-invariant ESs [5] and self-adaptive ESs (which use a real adaptation rule) [5, 6] do converge (or diverge) with order one, or log-linearly[1].

In this paper, finite dimension results are investigated and the focus is on the simplest ES, namely the (1+1)-ES. Section 2 introduces the mathematical model associated to the algorithm in a general framework and provides preliminary results. In Section 3, a sharp lower bound of the log-convergence rate is proved. In Section 4, it is shown that this lower bound is reached for a scaled-invariant algorithm on the class of sphere functions. The proof of convergence on the class of sphere functions uses the Law of Large Numbers for orthogonal random variables. A central limit theorem is also derived from this analysis. In Section 5 our results are discussed and related to previous works. Some numerical experiments illustrating the theoretical results are presented.

2 Mathematical model for the (1 + 1)-ES

Let $\mathbb{R}^d$ be equipped with the Borel σ-algebra and the Lebesgue measure. In the sequel we always assume that $(\mathcal{N}_n)_n$ denotes a sequence of random vectors (r.vec.) independent and identically distributed (i.i.d.), defined on a suitable probability space (Ω, P), with common law the multivariate isotropic normal distribution on $\mathbb{R}^d$ denoted by $\mathcal{N}(0, I_d)$ [2]. Let $(\sigma_n)_n$ be a given sequence of positive random variables (r.var.). We also assume that for each index n, σ_n is defined on Ω and is independent of $\mathcal{N}_n$; further we will also require that the sequences $(\sigma_n)_n$ and $(\mathcal{N}_n)_n$ are mutually independent. Finally, let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function (which is always assumed to be Lebesgue measurable) and let $\delta_n : \mathbb{R}^d \times \Omega \to \{0, 1\}$ $(n \geq 0)$ be the measurable function defined by $\delta_n(x, \omega) := \mathbf{1}_{\{f(x+\sigma_n(\omega)\mathcal{N}_n(\omega)) \leq f(x)\}}$. In this paper, $(1 + 1)$-ES algorithms are modeled by the $\mathbb{R}^d$-valued random process $(X_n)_{n\geq 0}$ defined on Ω by the recurrence relation

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \delta_n(\mathrm{X}_n, I_\Omega)\sigma_n\mathcal{N}_n\,, \tag{2}$$

where I_Ω is the identity function $\omega \mapsto \omega$ on Ω and X_0 is given.

[1] We say that the sequence $(\mathrm{X}_n)_n$ converges log-linearly to zero (resp. diverges log-linearly) if there exists $c < 0$ (resp. $c > 0$) such that $\lim_n \frac{1}{n} \ln \|\mathrm{X}_n\| = c$.

[2] $\mathcal{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix the identity I_d.

The classical terminology used for algorithms defined by (2) stresses the parallel with the biology: the iteration index n is referred as generation, the random vector X_n is called the parent, the perturbed random vector $\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma_n \mathcal{N}_n$ is the n-th offspring. The scalar r.var. σ_n is called step-size. The r.var. δ_n translates the plus selection "+" in the $(1+1)$-ES: the offspring is accepted if and only if its fitness value is smaller than the fitness of the parent. Several heuristics have been introduced for the adaptation of the step-size σ_n, the most popular being the one-fifth success rule [1, 2].

Notations and preliminary results

For a real valued function $x \mapsto h(x)$ we introduce its positive part $h^+(x) := \max\{0, h(x)\}$ and negative part $h^- = (-h)^+$. In other words $h = h^+ - h^-$ and $|h| = h^+ + h^-$. In the sequel, we denote by e_1 a unitary vector in $\mathbb{R}^d$. The following technical lemmas will be useful in the sequel.

Lemma 1. *Let $\mathcal{N}$ be a r.vec. of distribution $\mathcal{N}(0, I_d)$. The map $F : [0, \infty] \to [0, +\infty]$ defined by $F(+\infty) := 0$ and*

$$F(\sigma) := E\left[\ln^-\left(\|e_1 + \sigma\mathcal{N}\|\right)\right] = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \qquad (3)$$

otherwise, is continuous on $[0, +\infty]$ (endowed with the usual compact topology), finite valued and strictly positive on $]0, \infty[$.

Proof. The integral (3) always exists but could be infinite. In any case, $F(\sigma)$ is independent of the choice of e_1 due to the invariance of $\mathcal{N}$ under rotations. For convenience we choose $e_1 = (1, 0, \ldots, 0)$ so that $\ln^-(\|e_1 + \sigma x\|) = 0$ if $x = (x_1, \ldots, x_d)$ with $x_1 \geq 0$. Let $f_1 : \mathbb{R}^d \times [0, \infty] \to [0, +\infty]$ be defined by

$$f_1(x, \sigma) = \ln^-(\|e_1 + \sigma x\|^2) e^{-\frac{\|x\|^2}{2}}$$

for $x \neq (-1/\sigma, 0, \ldots, 0)$ and $f_1((-1/\sigma, 0, \ldots, 0), \sigma) = +\infty$ (with $\sigma > 0$) and finally $f_1(x, +\infty) = 0$ $(= \lim_{\sigma \to +\infty} f_1(x, \sigma))$. Notice that $f_1(x, \sigma) = 0$ if $x_1 \geq 0$ and readily $f_1((x_1, x_2, \ldots, x_d), \sigma) = f_1((x_1, \epsilon_2 x_2, \ldots, \epsilon_d x_d), \sigma)$ for any $(\epsilon_2, \ldots, \epsilon_d)$ in $\{-1, +1\}^{d-1}$ so that we can restrict the integration giving $F(\sigma)$ to the domain $\mathcal{D} :=]-\infty, 0[\times]0, \infty[^{d-1}$, more precisely one has

$$F(\sigma) = \frac{1}{4}\Big(\frac{2}{\pi}\Big)^{d/2} \int_{\mathcal{D}} f_1(x, \sigma) dx \qquad (4)$$

with in addition f_1 is finite everywhere in $\mathcal{D}$. From the definition of $F(+\infty)$ and f_1 one has $\frac{1}{4}(2/\pi)^{d/2} \int_{\mathcal{D}} f_1(x, +\infty) dx = 0 = F(+\infty)$ so that (4) holds also for $\sigma = +\infty$. Now, for any real number $\sigma > 0$ fixed, the inequality $f_1(x, \sigma) > 0$ holds on $B_\sigma := \{x \in \mathcal{D};\ \|e_1 + \sigma x\| < 1\}$ which is a nonempty open set, therefore $F(\sigma) > 0$. In addition, $f_1(x, 0) = 0$ for all x and so, $F(0) = 0$. Passing to spherical coordinates (with $d \geq 2$)we obtain after partial integration

$$\int_{\mathcal{D}} f_1(x) dx = 2c_d \int_0^{+\infty} \int_0^{\pi/2} \ln^-(|\sigma r - e^{i\theta_1}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}\theta_1 dr\ d\theta_1$$

where

$$c_d = \int_0^{\pi/2} \cdots \int_0^{\pi/2} \sin^{d-3}(\theta_2) \ldots \sin(\theta_{d-2}) d\theta_2 \ldots d\theta_{d-1}$$

for $d \geq 3$ and $c_2 = 1$. With the classical Wallis integral $W_{d-2} = \int_0^{\pi/2} \sin^{d-2}\theta \; d\theta$ and the surface area of the d-dimensional unit ball $S_d = 2\pi^{d/2}/\Gamma(\frac{n}{2})$ we have $S_d = 2^d c_d W_{d-2}$ and after collecting the above results we get

$$F(\sigma) = \Big(\frac{1}{2\pi}\Big)^{d/2} \frac{1}{W_{d-2}\Gamma(\frac{d}{2})} \int_0^{+\infty} \int_0^{\pi/2} \ln^-(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) \, dr \, d\theta \, .$$

The integrand $g : (r, \theta, \sigma) \mapsto \ln^-(|\sigma r - e^{i\theta}|) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta)$ defined on the set $]0, +\infty[\times[0, \pi/2] \times [0, \infty]$ (with $g(r, \theta, +\infty) = 0$) is continuous. In fact, the continuity is clear at each point (r, θ, σ) with $\sigma \neq +\infty$ and for the points $(r, \theta, +\infty)$, one has $g(\rho, \alpha, \sigma) = 0$ on $]r/2, +\infty[\times[0, \pi/2]\times]\frac{4}{r}, +\infty]$. Moreover, g is dominated by $g_1 : (r, \theta) \mapsto \ln^-(\sin\theta) r^{d-1} e^{-r^2/2}$ *i.e.*, $g(r, \theta, \sigma) \leq g_1(r, \theta)$ for all (r, θ, σ) in $]0, +\infty[\times[0, \pi/2] \times [0, +\infty]$. Since g_1 is integrable, the continuity of F on $[0, +\infty]$ follows from the Lebesgue dominated convergence theorem. For the remaining case $d = 1$ the conclusions of the lemma follow easily from (4) that gives $F(\sigma) = \frac{1}{2\sqrt{2\pi}} \int_0^{\infty} \ln^-(|1 - \sigma r|) e^{-\frac{r^2}{2}} dr$. □

Corollary 1. *The supremum $\tau := \sup F([0, +\infty])$ is reached and $\sigma_F := \min F^{-1}(\tau)$ exists. Moreover $0 < \sigma_F < +\infty$ and $0 < \tau < +\infty$.*

Proof. This corollary is a straightforward consequence of the continuity of F according to Lemma 1 which implies that $F^{-1}(\tau)$ is nonempty and compact. □

Lemma 2. *Let X denote a r.vec. in $\mathbb{R}^d$ such that $\|X\|^{-1}$ is finite almost surely. Let σ be a non negative random variable and let $\mathcal{N}$ be a random vector in $\mathbb{R}^d$ with distribution $\mathcal{N}(0, I_d)$ and independent of $\sigma\|X\|^{-1}$. Assume that*

$$E\Big(\ln\Big(1 + r\frac{\sigma}{\|X\|}\Big)\Big) \in O(e^{cr})$$

with a constant $c \geq 0$, then the expectation of $\ln^+(\|X\|^{-1}\|X + \sigma\mathcal{N}\|)$ is finite.

Proof. Obviously $E(\ln^+(\|X\|^{-1}\|X + \sigma\mathcal{N}\|)) \leq E(\ln(1 + \frac{\sigma}{\|X\|}\|\mathcal{N}\|))$. Using the independency of $\sigma\|X\|$ and $\mathcal{N}$, and passing to the spherical coordinates, one gets

$$\begin{aligned}
E\big(\ln\big(1 + \frac{\sigma}{\|X\|}\|\mathcal{N}\|\big)\big) &\leq E\big(\int_{\mathbb{R}^d} \ln(1 + \frac{\sigma}{\|X\|}\|x\|) e^{-\frac{\|x\|^2}{2}} dx\big) \\
&= S_d E\big(\int_0^{+\infty} \ln(1 + r\frac{\sigma}{\|X\|}) r^{d-1} e^{-\frac{r^2}{2}} dr\big) \\
&= S_d \int_0^{+\infty} E(\ln(1 + r\frac{\sigma}{\|X\|})) r^{d-1} e^{-\frac{r^2}{2}} dr \\
&\ll \int_0^{+\infty} r^{d-1} e^{cr - \frac{r^2}{2}} dr < +\infty \, .
\end{aligned}$$

□

Remark 1. The assumption $E(\ln(1 + r\frac{\sigma}{\|X\|})) \in O(e^{cr})$ (with $c = 0$) is verified if there exists $\alpha > 0$ such that the expectation of the r.var. $(\sigma/\|X\|)^\alpha$ is finite.

3 Lower bounds for the (1 + 1)-ES

In this section, we consider a general measurable objective function $f : \mathbb{R}^d \to \mathbb{R}$. We prove that the $(1 + 1)$-ES defined by (2) for minimizing f, under suitable assumptions, satisfies for all x^* in $\mathbb{R}^d$ and all indices $n \geq 0$:

$$-\infty < E(\ln \|\mathrm{X}_n - x^*\|) - \tau \leq E(\ln \|\mathrm{X}_{n+1} - x^*\|) < +\infty \tag{5}$$

where τ is defined in Corollary 1.

If x^* is a limit point of (X_n) (that could be a local optimum of f), (5) means that the expected log-distance to x^* cannot decrease more than τ, in other words, $-\tau$ is a lower bound for the convergence rate of $(1 + 1)$-ES. The proof of this result uses the following easy Lemma whose proof is left to the reader.

Lemma 3. *Let Z and V be r.vec. and let Θ be any r.var. valued in $\{0, 1\}$. Assume that the r.var. $\ln(\|Z\|)$ is finite almost surely. Then the following inequalities*

$$\begin{aligned} \ln(\|Z\|) - \ln^-(\|Z\|^{-1}\|Z + V\|) &\leq \ln(\|Z + \Theta V\|) \\ &\leq \ln(\|Z\|) + \ln^+(\|Z\|^{-1}\|Z + V\|) \end{aligned} \tag{6}$$

hold almost surely. □

We are ready to prove the following general theorem.

Theorem 1 (Lower bounds for the $(1+1)$-ES). *Let $(\mathrm{X}_n)_n$ be the sequence of random vectors verifying (2) with a given objective function f as above. Assume that for each step $n = 0, 1, 2, \ldots$ the random vector $\mathcal{N}_n$ is independent of both the random variable σ_n and the random vector X_n. Let x^* be any vector in $\mathbb{R}^d$ and suppose that $E(\big|\ln(\|\mathrm{X}_0 - x^*\|)\big|) < +\infty$ and for all $n \geq 0$,*

$$E\Big(\ln(1 + r\frac{\sigma_n}{\|X_n - x^*\|})\Big) \in O(e^{c_n r})$$

with a constant $c_n \geq 0$. Then

$$E\left(|\ln\left(\|\mathrm{X}_n - x^*\|\right)|\right) < +\infty \ ,$$

and

$$E(\ln(\|\mathrm{X}_n - x^*\|)) - \tau \leq E(\ln(\|\mathrm{X}_{n+1} - x^*\|))\,, \tag{7}$$

for all $n \geq 0$, where τ is defined in Corollary 1. In particular, the convergence of the $(1 + 1)$-ES is at most linear, in the sense that

$$\inf_{n\in\mathbb{N}} \frac{1}{n} E\big(\ln\big(\|\mathrm{X}_n - x^*\|/\|\mathrm{X}_0 - x^*\|\big)\big) \geq -\tau\,. \tag{8}$$

Proof. Set $Z_n = \mathrm{X}_n - x^*$, $\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma_n \mathcal{N}_n$ and $\tilde{\mathrm{Z}}_n = \tilde{\mathrm{X}}_n - x^*$. We prove the integrability of $\ln(\|Z_n\|)$ by induction. By assumption $E\big(\ln(\|Z_0\|)\big)$ is finite. Suppose that $E\big(\ln\|Z_n\|\big)$ is finite, then $0 < \|Z_n\| < +\infty$ almost surely, hence $\ln\big(\|Z_{n+1}\|\big)$ is also finite almost surely. We claim that $E\big(\ln(\|Z_{n+1}\|)\big)$ is finite. By applying Lemma 3 we get (6) and derive

$$\ln^+(\|Z_{n+1}\|) \leq \ln^+(\|Z_n\|) + \ln^+\big(\|Z_n\|^{-1}(\|Z_n + \sigma_n \mathcal{N}_n\|)\big)\,. \tag{9}$$

By Lemma 2 the expectation of $\ln^+\big(\|Z_n\|^{-1}(\|Z_n + \sigma_n\mathcal{N}_n\|)\big)$ is finite and using (9) we conclude that $E\big(\ln^+(\|Z_{n+1}\|)\big) < +\infty$. It remains to show that $E\big(\ln^-(\|Z_{n+1}\|)\big)$ is also finite. Using the first inequality in (6) we obtain

$$\ln^-(\|Z_{n+1}\|) \leq -\ln(\|Z_n\|) + \ln^-\Big(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\Big) + \ln^+(\|Z_{n+1}\|)\,. \tag{10}$$

For each $n \geq 0$, let $\mathcal{F}_n$ denote the σ-algebra generated by the r.vec. X_n and the r.var. σ_n. Taking the conditional expectation we obtain

$$\begin{aligned}
&E[\ln^-(\|Z_{n+1}\|)\,|\,\mathcal{F}_n] \\
&\leq -\ln(\|Z_n\|) + E\Big[\ln^-\Big(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\Big)\,|\,\mathcal{F}_n\Big] + E\big[\ln^+\big(\|Z_{n+1}\|\big)\,|\,\mathcal{F}_n\big]\,.
\end{aligned}$$

Since the distribution $\mathcal{N}_n$ is invariant under rotation and independent of $\mathcal{F}_n$,

$$\begin{aligned}
E\big(\ln^-\Big(\Big\|\frac{Z_n}{\|Z_n\|} + \frac{\sigma_n}{\|Z_n\|}\mathcal{N}_n\Big\|\Big)\,|\,\mathcal{F}_n\big) &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \ln^-(\|e_1 + t_n x\|)e^{-\frac{\|x\|^2}{2}}dx \\
&= F(t_n)
\end{aligned}$$

where e_1 is any unit vector on $\mathbb{R}^d$, $t_n = \sigma_n/\|Z_n\|$ (and F is the map introduced in Lemma 1). Using Lemma 1, we get $E\big[\ln^-(\|Z_{n+1}\|)\,|\,\mathcal{F}_n\big] \leq -\ln(\|Z_n\|) + \tau + E\big[\ln^+(\|Z_{n+1}\|)\,|\,\mathcal{F}_n\big]$ (recall that $\tau = \max F([0,+\infty])$). Passing to the expectation we get

$$E\big[\ln^-(\|Z_{n+1}\|)\big] \leq -E\big[\ln(\|Z_n\|)\big] + \tau + E\big[\ln^+(\|Z_{n+1}\|)\big] < +\infty\,.$$

Hence $E[|\ln(\|Z_{n+1}\|)|]$ is finite for all $n \geq 0$. Moreover, we also get

$$E(\ln\|Z_{n+1}\|) \geq E(\ln\|Z_n\|) - \tau$$

and after summing such inequalities we obtain

$$E\left(\ln(\|Z_n\|/\|Z_0\|)\right) \geq -\tau n$$

and (8) follows. □

When x^* is a local minimum of the objective function, $E(\ln\|\mathrm{X}_n - x^*\|) - E(\ln\|\mathrm{X}_{n+1} - x^*\|)$ represents the expected log-distance reduction towards x^* at the n-th step of iteration, called *log-progress* in [7]. Theorem 1 shows that the log-progress is bounded above by $\tau = F(\sigma_F)$.

4 Spherical functions and the scale-invariant algorithm

In this section we prove that the lower bound $-\tau$ obtained in Theorem 1 is reached for spherical objective functions when $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ $(n \geq 0)$. Recall that sphere objective functions are defined by $f(x) = g(\|x\|^2)$ where g is any increasing map, so that the acceptance condition $f(\mathrm{X}_{n+1}) \leq f(\mathrm{X}_n)$ is equivalent to $\|\mathrm{X}_{n+1}\| \leq \|\mathrm{X}_n\|$. It follows that $(\|\mathrm{X}_n\|)_{n\geq 0}$ is a non-increasing sequence of positive random variables (finite almost surely), hence converges pointwise almost surely. For spherical functions, Lemma 3 becomes:

Lemma 4. *Let X and W be any random vectors and let $\Theta = \mathbf{1}_{\{f(X+W)\leq f(X)\}}$ and assume that the random variable $\ln(\|X\|)$ is finite almost surely. Then the equality*

$$\ln(\|X + \Theta W\|) - \ln(\|X\|) = -\ln^+(\|X\|^{-1}\|X + W\|) \tag{11}$$

holds almost surely.

Proof. The equality (11) emphasizes the fact that $\|X+\Theta\| \leq \|X\|$ with equality on the event $\{\Theta = 0\}$ $(= \{\|X + W\| > \|X\|\})$. □

Proposition 1. *Let $(\mathrm{X}_n)_n$ be the sequence of random vectors valued in $\mathbb{R}^d$ satisfying the recurrence relation (2) involving spherical function $f(x) = g(\|x\|^2)$ where $g : [0, \infty[\to \mathbb{R}$ is an increasing map. Assume that $E(\ln(\|X_0\|)$ is finite and that, at each step n, the random vector $\mathcal{N}_n$ is independent of both the random variable σ_n and the random vector X_n. Then $E(\ln(\|X_n\|)$ is finite for all indices n, the inequalities*

$$E(\ln(\|X_n\|) - \tau \leq E(\ln(\|X_{n+1}\|)$$

hold, where τ is defined above in Corollary 1, and

$$\ln(\|X_n\|) - \ln(\|X_{n+1}\|) = \ln^-(\|X_n\|^{-1}\|X_n + \sigma_n\mathcal{N}_n\|) < +\infty \ a.s. \tag{12}$$

Proof. By construction $\|X_{n+1}\| \leq \|X_n\| \leq \|X_0\|$ so that $E(\ln^+(\|X_{n+1}\|)) \leq E(\ln^+(\|X_0\|)) < +\infty$. Now assume that $\ln(\|X_n\|)$ is integrable, hence $0 < \|X_n\| < +\infty$ a.s. and so, by Lemma 4, to obtain the inequalities and equality asserted in the proposition it is enough to prove that $E(\ln^-(\|X_n\|^{-1}\|X_n + \sigma_n\mathcal{N}_n\|)) \leq \tau$. But similarly to the end part of the proof of Theorem 1 we have $E(\ln^-(\|X_n\|^{-1}\|X_n + \sigma_n\mathcal{N}_n\|)) = E(F(\sigma_n/\|X_n\|)) \leq \tau$. □

Now we pay attention to the particular case where $\sigma_n = \sigma\|\mathrm{X}_n\|$ with $\sigma > 0$ fixed. The resulting $(1 + 1)$-ES is said to be *scale-invariant*, and is modeled by the d-dimensional random process

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \delta_n(\mathrm{X}_n, I_\Omega)\sigma\|\mathrm{X}_n\|\mathcal{N}_n \qquad (n \geq 0)\,. \tag{13}$$

For convenience of the reader we collect the hypothesis that govern the scale-invariant random process (13):

(HSI) *The sequence of random vectors $(\mathcal{N}_n)_n$ in $\mathbb{R}^d$ is i.i.d. with common law $\mathcal{N}(0, I_d)$, is independent of the initial random vector X_0 and $\ln(\|X_0\|)$ has a finite expectation.*

Notice that Assumption (HSI) implies in particular that for $m \geq n \geq 0$, $\mathcal{N}_m$ is independent of X_n and by Proposition 1, $\ln(\|X_n\|)$ has a finite expectation. The update rule (13) is not so realistic because in practice, at each step n, the distance of X_n to the optimum is unknown. Nevertheless, we will show that the stochastic process defined by (13) converges log-linearly for sphere functions and that for $\sigma = \sigma_F$ the convergence rate in log is equal to $-F(\sigma_F)$ $(= -\tau)$. In other words, the choice $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ correspond to the adaptation scheme that gives the optimal convergence rate for isotropic Evolution Strategies.

It is usual for studying stochastic search algorithms to consider log-linear convergence of X_n by investigating the stability of $\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|)$. This idea was introduced in the context of ESs by Bienvenüe and François [5] and exploited in [6]. The process X_n given by (13) has a remarkable property expressed in terms of orthogonality of the random sequences $Y_n = \ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right) - F(\sigma)$:

Proposition 2. *Consider the random variables*

$$Y_n := \ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right) - F(\sigma)$$

where F is defined by (4) and let $\sigma > 0$. Under the hypothesis (HSI) *the followings hold:*

1. *For $n \geq 0$, $E(Y_n) = 0$ and $E(|Y_n|^2) < +\infty$.*
2. *Let $(Y'_n)_{n\geq 0}$ be the sequence of random variables*

$$Y'_n := \ln^-(\|e_1 + \sigma\mathcal{N}_n\|) - F(\sigma).$$

   The random variables Y_n $(n \geq 0)$ are identically distributed and for every $n \geq 0$, Y_n and Y'_n follow the same distribution.
3. *The sequence of random variables $(Y_n)_{n\geq 0}$ is orthogonal, i.e. for all indices i, j, with $i \neq j$ one has $E(Y_i) = 0$, $E(Y_i^2) < +\infty$ and $E(Y_iY_j) = 0$.*

Proof. The isotropy of the standard d-dimensional normal distribution gives

$$\begin{aligned} E\Big(\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right) \mid \mathrm{X}_n\Big) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \\ &= F(\sigma) \end{aligned}$$

hence $E\left[\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n\right\|\right)\right] = E\left[F(\sigma)\right]$ and so, $E(Y_n) = 0$. Let $F_2 : [0,\infty] \to [0,+\infty[$ be defined by $F_2(\infty) = 0$ and, for $t \in [0,+\infty[$,

$$F_2(t) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[\ln^-(\|e_1 + tx\|)\right]^2 e^{-\frac{\|x\|^2}{2}} dx\,. \tag{14}$$

Similarly to the proof of Lemma 1, we prove that F_2 is continuous, hence bounded. Now, from the definitions of F and F_2 one has

$$E(|Y_n|^2) = F_2(\sigma) - (F(\sigma))^2 < +\infty\,. \tag{15}$$

This ends the proof of the first point.

The random vectors Y_n and Y'_n have the same distribution if their characteristic functions are identical. But successively

$$\begin{aligned} E(e^{itY_n} \,|\, \mathrm{X}_n) &= e^{-itF(\sigma)} E\Big(e^{it\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|\right)} \,|\, \mathrm{X}_n\Big) \\ &= \frac{e^{-itF(\sigma)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it\ln^-\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma x\right\|\right)} e^{-\|x\|^2/2} dx \\ &= \frac{e^{-itF(\sigma)}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it\ln^-\left(\left\|e_1+\sigma x\right\|\right)} e^{-\|x\|^2/2} dx \\ &= E(e^{itY'_n})\,. \end{aligned}$$

Therefore $E(e^{itY_n}) = E(E(e^{itY_n} \,|\, \mathrm{X}_n)) = E(e^{itY'_n})$. To finish the proof we show the orthogonality property of the Y_n ($n \geq 0$). Let n and m be indices such that $n < m$. The random vector Y_n is $\sigma(\mathrm{X}_n, \mathcal{N}_n)$-measurable, so that

$$E(Y_m Y_n \,|\, \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n) = Y_n E(Y_m | \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n)\,.$$

Using the independency of $\mathcal{N}_m$ with the random vectors. X_n, $\mathcal{N}_n$ and X_m, we get

$$\begin{aligned} E(Y_m | \mathrm{X}_n, \mathrm{X}_m, \mathcal{N}_n) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \Big(\ln^-\left(\Big\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma x\Big\|\right)\Big) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma) \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \Big(\ln^-(\|e_1 + \sigma x\|)\Big) e^{-\frac{\|x\|^2}{2}} dx - F(\sigma) = 0\,, \end{aligned}$$

that implies $E(Y_m Y_n) = 0$. □

With the above notations define the random vectors $S_n = Y_0 + \cdots + Y_n$ and $S'_n = Y'_0 + \cdots + Y'_n$. Under the hypothesis (HSI), the characteristic function of S_n can be written as $E(itS_n) = E(E(itS_n \,|\, X_0, \mathcal{N}_0, \ldots, \mathcal{N}_{n-1}))$ and so, $E(itS_n) = E(itS'_n) = (E(itY'_0))^{n+1}$. But the random vectors Y_n are i.i.d. with expectation 0 and variance $F_2(\sigma) - F(\sigma)^2$ (see (15)). As a consequence, the central limit theorem holds for both $(Y_n)_n$ and $(Y'_n)_n$:

Theorem 2. *Under the hypothesis* (HSI) *one has*

$$\lim_{n\to+\infty} P\left(\frac{\ln(\|X_n\|) - \ln(\|X_0\|) + F(\sigma)n}{\sqrt{(F_2(\sigma) - F(\sigma)^2)n}} \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{u^2}{2}} du\,.$$

The pointwise stability of $\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|)$ is obtained by applying the following Law of Large Numbers (LLN) for orthogonal random variables (see [10, p. 458] where a more general statement is given).

Theorem 3 (LLN for Orthogonal Random Variables). *Let $(Y_n)_{n\geq 0}$ be a sequence of identically distributed real random variables with finite variance and orthogonal, i.e., for all indices i, j, with $i \neq j$ one has $E(Y_i) = 0$, $E(Y_i^2) < +\infty$ and $E(Y_i Y_j) = 0$. Then*

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} Y_k = 0 \quad a.s.$$

We are now ready to prove the following main result

Theorem 4. *Let $\sigma > 0$ and let $(\mathrm{X}_n)_n$ be the sequence of random vectors satisfying the recurrence relation (13) with $f(x) = g(\|x\|^2)$ where g is an increasing map. Assume that the hypothesis* (HSI) *holds. Then $(\mathrm{X}_n)_n$ converges log-linearly to the minimum, in the sense that*

$$\lim_n \frac{1}{n} \ln \Big(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|} \Big) = -F(\sigma) (< 0) \quad \text{a.s.} \tag{16}$$

where F is defined by (4). The optimal convergence rate is obtained for $\sigma = \sigma_F := \min F^{-1}(\max F)$ (see Corollary 1).

Proof. In case $\sigma_n = \sigma \|\mathrm{X}_n\|$ for all indices n the equality (12) becomes

$$\ln \|\mathrm{X}_{n+1}\| - \ln \|\mathrm{X}_n\| = - \ln^- \Big(\Big\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \Big\| \Big) .$$

and after summing the equations for $k = 0, \ldots, n-1$, we obtain

$$\frac{1}{n} \left(\ln \|\mathrm{X}_n\| - \ln \|\mathrm{X}_0\| \right) = -\frac{1}{n} \sum_{k=0}^{n-1} \ln^- \Big(\Big\| \frac{X_k}{\|X_k\|} + \sigma \mathcal{N}_k \Big\| \Big) .$$

Proposition 2 and Theorem 3 end the proof. □

5 Discussion and conclusion

Theorems 1 and 4 show that optimal bounds for the convergence rate of an isotropic $(1+1)$-ES with multivariate normal distribution are reached for the scale-invariant algorithm with $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ for the sphere function, where σ_F maximizes

$$F(\sigma) = E(\ln^- \|e_1 + \sigma \mathcal{N}\|) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^-(\|e_1 + \sigma x\|) e^{-\frac{\|x\|^2}{2}} dx \ .$$

From (12) and from the isotropy of the multivariate normal distribution $\mathcal{N}$, it follows that finding σ maximizing F amounts to finding σ maximizing the log-progress $E(\ln \|\mathrm{X}_n\|) - E(\ln \|\mathrm{X}_{n+1}\|)$.

Most of the works based on the progress rate, consist in finding σ maximizing estimations of the expected progress $E(\|\mathrm{X}_n\|) - E(\|\mathrm{X}_{n+1}\|)$ (when d goes to infinity) [1, 4]. Note that the definition of progress in those works does not consider

$\ln \|\mathrm{X}_n\|$ and so is different from the one underlying our study. Assuming that both definitions matches[3], our results give an interpretation of this approach in terms of lower bounds for convergence of ESs.

The lower bounds derived in this paper are tight. Consequently they can be used in practice to assess the performances of a given step-size adaptation strategy comparing the convergence rate achieved by the strategy with the optimal one, given by the scale-invariant algorithm.

The numerical estimation of the optimal convergence rate $-\tau$ can be achieved with a Monte Carlo integration: for different σ, $F(\sigma)$ equals the expectation $E(\ln^- \|e_1 + \sigma\mathcal{N}\|)$. This expectation can be estimated by summing independent samplings of the random variable $\ln^- \|e_1 + \sigma\mathcal{N}\|$. This is illustrated in Fig 1.

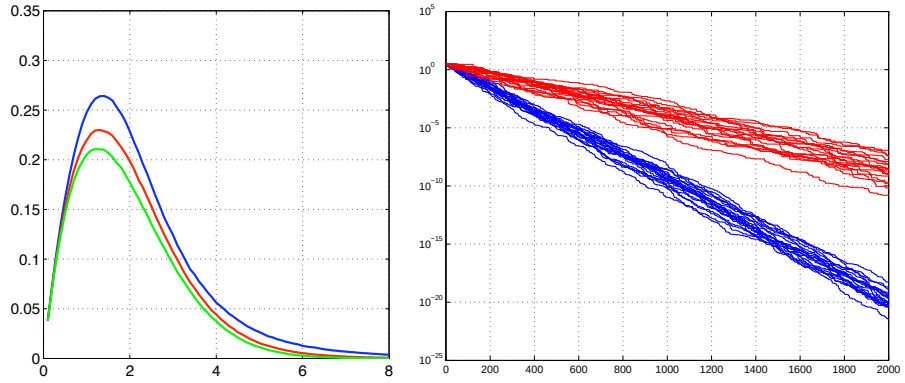

Fig. 1. Left: Plot of the function $\sigma \mapsto dF(\sigma/d)$ (Eq. (4)) versus σ for $d = 5$ (resp. 10, 30) and $0 \leq \sigma \leq 8$. The upper curve corresponds to $d = 5$, the middle one to $d = 10$ and the lower one to $d = 30$. Note that the function F defined in (4) implicitly depends on d. Using the more explicit notation F_d instead of F, the plots represent actually $\sigma \mapsto dF_d(\sigma/d)$. For $d = 10$, we see that σ_F maximizing F (defined in Corollary 1) approximately equals 0.13. The plots were obtained doing Monte Carlo estimations of F using 10^6 samples.
Right: Twenty realizations of the scale-invariant algorithm on the sphere function for $d = 10$. The y-axis shows the distance to the optimum (in log-scale) and the x-axis the number of iterations n. The twenty curves below correspond to the optimal algorithm, *ie.* $\sigma_n = \sigma_F \|\mathrm{X}_n\|$ for all n where σ_F equals to 0.13 (value maximizing the curve of F on the left for $d = 10$). The twenty curves above correspond to 20 realizations of the scale-invariant algorithm for $\sigma_n = 0.3\|\mathrm{X}_n\|$. Observed, the log-linear convergence as well as the optimality of the scale-invariant algorithm for $\sigma = \sigma_F$.

The analysis of the log-linear convergence carried out in this paper relies on the application of the Strong Law of Large Numbers for orthogonal random

[3] This will be true asymptotically in the dimension d, though we do not prove it rigorously in this paper.

variables. This study uses deeply the invariance under rotations of the standard d-dimensional multivariate normal distribution and does not cover directly the usual case of stable Markov chains that will be investigated in future works.

Acknowledgments

The authors thank the referees for their constructive remarks on the previous version that lead to this new version and are very grateful to Nicolas Monmarché for his encouragements. This work receives partial supports from the ANR/RNTL project Optimisation Multidisciplinaire (OMD) and from the ACI CHROMALGEMA.

References

1. I. Rechenberg, "Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution", Fromman-Hozlboog Verlag, Stuttgart (1973).
2. S. Kern, S, Müller, N. Hansen, D. Büche, J. Ocenasek, P. Koumoutsakos, "Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review", Natural Computing **3** (2004) 77–112.
3. N. Hansen, A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies", Evolutionary Computation **9** (2001) 159–195.
4. H.G. Beyer, "The Theory of Evolution Strategies", Springer (2001).
5. A. Bienvenüe, O. François, "Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties". Theoretical Computer Science **306** (2003) 269–289.
6. A. Auger, "Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains", Theoretical Computer Science **334** (2005) 35–69.
7. A. Auger, N. Hansen, "Reconsidering the progress rate theory for evolution strategies in finite dimensions", In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006), (2006) 445–452.
8. O. Teytaud, S. Gelly, S., "General lower bounds for evolutionary algorithms", In Ninth International Conference on Parallel Problem Solving from Nature PPSN IX, (2006) (LNCS 4193/2006) 21–31.
9. J. Jägersküpper, "Lower bounds for hit-and-run direct search", In Proceedings of Stochastic Algorithms: Foundations and Applications (SAGA 2007), (LNCS 4665/2007) 118–129.
10. M. Loève, "Probability Theory" (3rd Edition), Van Nostrand (New York, 1963).