Log-linear Convergence of the Scale-invariant $(\mu/\mu\_w, \lambda)$-ES and Optimal μ for Intermediate Recombination for Large Population Sizes

Mohamed Jebalia, Anne Auger

Log-linear Convergence of the Scale-invariant $(\mu/\mu_w, \lambda)$-ES and Optimal μ for Intermediate Recombination for Large Population Sizes

Mohamed Jebalia and Anne Auger

TAO Team - INRIA Saclay-Ile-de-France, LRI, Paris-Sud University,
91405 Orsay Cedex, France
firstname.lastname@inria.fr

Abstract. Evolution Strategies (ESs) are population-based methods well suited for parallelization. In this paper, we study the convergence of the $(\mu/\mu_w, \lambda)$-ES, an ES with weighted recombination, and derive its optimal convergence rate and optimal μ especially for large population sizes. First, we theoretically prove the log-linear convergence of the algorithm using a scale-invariant adaptation rule for the step-size and minimizing spherical objective functions and identify its convergence rate as the expectation of an underlying random variable. Then, using Monte-Carlo computations of the convergence rate in the case of equal weights, we derive optimal values for μ that we compare with previously proposed rules. Our numerical computations show also a dependency of the optimal convergence rate in $\ln(\lambda)$ in agreement with previous theoretical results.

1 Introduction

Evolution Strategies (ESs) are robust stochastic search methods [2, 3] for solving continuous optimization problems where the goal is to minimize[1] a real valued objective function f defined on an open subset of $\mathbb{R}^d$. At each iteration of an ES, new solutions are in general generated by adding Gaussian perturbations (mutations) to some (optionally recombined) current ones. These Gaussian mutations are parameterized by the step-size giving the general scale of the search, and the covariance matrix giving the principal directions of the Gaussian distribution. In state-of-the art ESs, these parameters are adapted at each iteration [1–4]. We focus on isotropic ESs where the step-size is adapted and the covariance matrix is kept equal to the identity matrix I_d and therefore the search distribution is spherical. Adaptation in ESs allows them to have a log-linear behavior (convergence or divergence) when minimizing spherical objective functions [10, 13, 5, 7]. Log-linear convergence (resp. divergence) means that there exists a constant value $c < 0$ called convergence rate (resp. $c > 0$) such that the distance to the optimum, d_n, at an iteration n satisfies $\lim_n \frac{1}{n} \ln(d_n) = c$. Spherical objective functions are defined as

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

[1] Without loss of generality, the minimization of a real value function f is equivalent to the maximization of $-f$.

where $g : [0, \infty[\mapsto \mathbb{R}$ is a strictly increasing function, $x \in \mathbb{R}^d$ and $\|.\|$ denotes the Euclidean norm on $\mathbb{R}^d$. Log-linear behavior holds also when minimizng spherical functions perturbed by noise [11].

In this paper, we investigate ESs with weighted recombination, denoted $(\mu/\mu_w, \lambda)$-ES, and used in the state-of-the-art ES, the Covariance Matrix Adaptation-ES (CMA-ES) [4]. The $(\mu/\mu_w, \lambda)$-ES is an ES which evolves a single solution. Let $\mathbf{X}_n$ be the solution (the parent) at iteration n, λ new solutions $\mathbf{Y}_n^i$ (offspring) are then generated using independent Gaussian samplings of mean $\mathbf{X}_n$. Then, the offspring are evaluated, the μ best offspring $(\mathbf{Y}_n^{i:\lambda})_{1 \leq i \leq \mu}$ are selected and the new solution $\mathbf{X}_{n+1}$ is obtained by recombining these selected offspring using recombination weights denoted $\left(w^i\right)_{1 \leq i \leq \mu}$, i.e., $\mathbf{X}_{n+1} = \sum_{i=1}^{\mu} w^i \mathbf{Y}_n^{i:\lambda}$ [2]. We will specifically study the $(\mu/\mu_w, \lambda)$-ES with large (offspring) population size λ compared to the search space dimension d, i.e., $\lambda \gg d$. This is motivated by the increasing possibilities of parallelization with the raise of the number of parallel machines, supercomputers and grids. ESs are population-based methods and then are well suited for parallelization which consists in distributing the number of evaluations λ on the processes available. The performance of the $(\mu/\mu_w, \lambda)$-ES as a function of λ has been theoretically investigated [14, 16]. Under the approximation $d \to +\infty$, the study in [14] investigated the $(\mu/\mu_w, \lambda)$-ES minimizing any spherical function and using an artificial step-size adaptation rule termed scale-invariant which sets the step-size at each iteration proportionally to the distance of the current solution to the optimum. The progress rate φ which measures the one-step expected progress to the optimum verifies $\varphi = O\left(\mu \ln\left(\frac{\lambda}{\mu}\right)\right)$ [14]. This suggests that, if μ is chosen proportional to λ, the progress rate of the $(\mu/\mu_w, \lambda)$-ES can be linear in μ and in λ. The study in [16] is based on a theoretical computations of lower bounds for the convergence ratio which measures the convergence rate in probability of wide classes of ESs. It shows that the convergence ratio of the $(\mu/\mu_w, \lambda)$-ES varies at best linearly with $\ln(\lambda)$ for sufficiently large λ when minimizing any spherical function [16]. This suggests that the bound found in [14] is not tight for finite dimensions.

A natural question arising when using recombination is how to choose the number of offspring μ to be recombined. Studies based on computations of the progress rate when the search space dimension goes to infinity suggest to use $\mu = \lfloor \frac{\lambda}{4} \rfloor$ [14] or $\mu = \lfloor \frac{\lambda}{2} \rfloor$ [6][3] for two different choices of the (positive) weights $\left(w^i\right)_{1 \leq i \leq \mu}$. CMA-ES which has been designed to work well on small population sizes uses $\mu = \lfloor \frac{\lambda}{2} \rfloor$ as a default parameter. However, when using a large population size λ, the convergence rate of some real-world algorithms tested in [15, 8] using the rules $\mu = \lfloor \frac{\lambda}{4} \rfloor$ or $\mu = \lfloor \frac{\lambda}{2} \rfloor$ as recommended in [14, 6] is worse than the theoretical prediction of [16]. This is due to the fact that the rules used in these tests for choosing μ, are recommended by the studies performed under the approximation $(d \to +\infty)$ [14, 6] and thus under the assumption $\lambda \ll d$. For some values of λ and d such that $\lambda \gg d$, Beyer [17] computed, using some approximations permitted by the assumption $(d \to +\infty)$, optimal choices for μ when minimizing spherical functions. However, no explicit rule for the choice of μ has been

[2] If $\mu = 1$, only the best offspring is taken and then the $(\mu/\mu_w, \lambda)$-ES is simply the $(1, \lambda)$-ES.

[3] The rule proposed in [6] where negative weights are allowed is rather $\mu = \lambda$, but the study implies that if the weights can be only positive the rule becomes $\mu = \lfloor \frac{\lambda}{2} \rfloor$.

proposed when $\lambda \gg d$. Performing experiments with $\lambda \gg d$ on a $(\mu/\mu_w, \lambda)$-ES using equal weights, the so-called self-adaptation rule for the step-size and two variants for the covariance matrix adaptation, Teytaud [9] proposed to choose μ equal to $\min\{d, \lfloor \frac{\lambda}{4} \rfloor\}$.

Since it is in general difficult to appraise whether the effect observed when changing the setting of one parameter on a real algorithm is coming from the fact that the setting of an other parameter may subsequently becomes sub-optimal, we want here to identify independently of any real step-size or covariance matrix update rule the optimal setting for μ especially for lagre λ. This optimal setting can be used to identify a rule for choosing best optimal values μ in real-world algorithms like CMA-ES. We want also to verify whether an optimal choice for μ allows to have a dependency of the convergence rate in $\ln(\lambda)$ and thus reach the lower bounds predicted by [16]. In order to do so, we perform in this paper a theoretical and numerical investigation of the convergence and the optimal choice for μ relative to the isotropic $(\mu/\mu_w, \lambda)$-ES. We focus on large population sizes. The objective functions investigated are the spherical functions allowing ESs which do not use recombination to reach optimal convergence rates [5, 7]. In Section 2, we present the mathematical formulation of the algorithm. In Section 3, we identify the optimal step-size adaptation rule of the algorithm when minimizing spherical functions. In Section 4, we theoretically prove the log-linear convergence of the algorithm using the scale-invariant adaptation rule and identify its convergence rate. In Section 5, using Monte-Carlo computations of the convergence rate, optimal μ values and optimal convergence rates are derived for some dimensions and in the specific case of equal weights $(w^i)_{1 \leq i \leq \mu}$. A new rule for choosing μ is proposed based on our results. Throughout the paper, we explain only the basic ideas of the proofs because of space limitation. For complete proofs, we refer to [12].

2 Mathematical Formulation of the Isotropic $(\mu/\mu_w, \lambda)$ Evolution Strategy Minimizing Spherical Functions

Throughout the remainder of this paper, we suppose that μ and λ are two positive integers such that $1 \leq \mu \leq \lambda$, and that the recombination weights $(w^i)_{1 \leq i \leq \mu}$ are positive constants summing to one, i.e., $\sum_{i=1}^{\mu} w^i = 1$. In this section we will introduce the mathematical formulation of the isotropic $(\mu/\mu_w, \lambda)$-ES for minimizing a spherical function (1). Let $\mathbf{X}_0 \in \mathbb{R}^d$ be the first solution randomly chosen using a law absolutely continuous with respect to the Lebesgue measure. Let σ_0 be a strictly positive variable (possibly) randomly chosen. Let $(\mathbf{N}_n^i)_{i \in [1,\lambda], n \in \mathbb{Z}^+}$, be a sequence of random vectors defined on a probability space $(\Omega, \mathcal{A}, P)$, independent and identically distributed (i.i.d.) with common law the isotropic multivariate normal distribution on $\mathbb{R}^d$ with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix identity I_d, which we will simply denote $\mathbf{N}$. We assume that the sequence $(\mathbf{N}_n^i)_{i \in [1,\lambda], n \in \mathbb{Z}^+}$ is independent of $\mathbf{X}_0$. Let σ_n be the step-size mutation at iteration n such that for all $(i, n) \in [1, \lambda] \times \mathbb{Z}^+$, σ_n and $\mathbf{N}_n^i$ are independent. An offspring $\mathbf{Y}_n^i$ where $i = 1, \ldots, \lambda$ writes as $\mathbf{Y}_n^i := \mathbf{X}_n + \sigma_n \mathbf{N}_n^i$, and its objective function value is $g(\|\mathbf{Y}_n^i\|)$ in our case of minimization of spherical functions. Let $\mathbf{N}_n^{i:\lambda}(\mathbf{X}_n, \sigma_n)$ $(1 \leq i \leq \mu)$ denotes the mutation vector relative to the i^{th} best offspring according to its fitness value. As the function g is increasing, the vectors

$\mathbf{N}_n^{i:\lambda}(\mathbf{X}_n, \sigma_n)$ (where, for all i in $\{1, \ldots, \mu\}$, the indices $i:\lambda$ are in $\{1, \ldots, \lambda\}$) verify:

$$\begin{aligned} &\left\|\mathbf{X}_n + \sigma_n \mathbf{N}_n^{1:\lambda}(\mathbf{X}_n, \sigma_n)\right\| \quad \leq \ldots \leq \left\|\mathbf{X}_n + \sigma_n \mathbf{N}_n^{\mu:\lambda}(\mathbf{X}_n, \sigma_n)\right\| \text{ and} \\ &\left\|\mathbf{X}_n + \sigma_n \mathbf{N}_n^{\mu:\lambda}(\mathbf{X}_n, \sigma_n)\right\| \leq \left\|\mathbf{X}_n + \sigma_n \mathbf{N}_n^{j}\right\| \forall j \in \{1, \ldots, \lambda\} \backslash \{1:\lambda, \ldots, \mu:\lambda\}\,. \end{aligned} \tag{2}$$

Using the fact that $\sum_{i=1}^{\mu} w^i = 1$, the new parent $\mathbf{X}_{n+1} = \sum_{i=1}^{\mu} w^i \mathbf{Y}_n^{i:\lambda}$ can be rewritten as:

$$\mathbf{X}_{n+1} = \mathbf{X}_n + \sigma_n \sum_{i=1}^{\mu} w^i \mathbf{N}_n^{i:\lambda}(\mathbf{X}_n, \sigma_n)\,. \tag{3}$$

In the specific case where the scale-invariant rule is used for the adaptation of $(\sigma_n)_{n\in\mathbb{Z}^+}$, i.e., $\sigma_n = \sigma\|\mathbf{X}_n\|$ (with $\sigma > 0$), the previous equation becomes:

$$\mathbf{X}_{n+1} = \mathbf{X}_n + \sigma\|\mathbf{X}_n\| \sum_{i=1}^{\mu} w^i \mathbf{N}_n^{i:\lambda}(\mathbf{X}_n, \sigma\|\mathbf{X}_n\|)\,. \tag{4}$$

Finally, σ_n is updated, i.e., σ_{n+1} is computed independently of $\mathbf{N}_{n+1}^i$ for all $i \in [1, \lambda]$. Throughout the remainder of this paper, we will denote in a general context where $u \in \mathbb{R}^d$, $s \in \mathbb{R}$ and $(\mathbf{N}_n^i)_{i\in[1,\lambda],n\in\mathbb{Z}^+}$ is a sequence of random vectors (i.i.d.) with common law $\mathbf{N}$ and such that for all $(i, n) \in [1, \lambda] \times \mathbb{Z}^+$, $\mathbf{N}_n^i$ is independent of u and s, $\mathbf{N}_n^{i:\lambda}(u, s)$ the random vector which verifies (2) where $\mathbf{X}_n$ and σ_n are respectively replaced by u and s. For $n = 0$ and $i \in \{1, \ldots, \mu\}$, the notation $\mathbf{N}_0^{i:\lambda}(u, s)$ will be replaced by the notation $\mathbf{N}^{i:\lambda}(u, s)$.

3 Optimal Step-size Adaptation Rule When Minimizing Spherical Functions

The (log-linear) convergence rate of the isotropic scale-invariant $(\mu/\mu_w, \lambda)$-ES minimizing any spherical function and satisfying the recurrence relation (4) is, as will be shown in Section 4, the function V that we will introduce in the following definition.

Definition 1. *Let* e_1 *denotes the unit vector* $(1, 0, \ldots, 0) \in \mathbb{R}^d$*. For* $\sigma \geq 0$*, let* $\mathrm{Z}(\sigma)$ *be the random variable defined as* $\mathrm{Z}(\sigma) := \left\|\mathrm{e}_1 + \sigma \sum_{i=1}^{\mu} w^i \mathbf{N}^{i:\lambda}(\mathrm{e}_1, \sigma)\right\|$ *where the random variables* $\mathbf{N}^{i:\lambda}(\mathrm{e}_1, \sigma)$ *are obtained similarly to (2) but with* $n = 0$ *and* $(\mathbf{X}_n, \sigma_n)$ *replaced by* (e_1, σ)*. We introduce the function* V *as the function mapping* $[0, +\infty[$ *into* $\mathbb{R}$ *as follows:*

$$\mathrm{V}(\sigma) := E\left[\ln \mathrm{Z}(\sigma)\right] = E\left[\ln\left\|\mathrm{e}_1 + \sigma \sum_{i=1}^{\mu} w^i \mathbf{N}^{i:\lambda}(\mathrm{e}_1, \sigma)\right\|\right]\,. \tag{5}$$

Fig. 1 (left) represents numerical computations of the function V in some specific settings. In the following proposition, we show that V is well defined and we study its properties. Note that in the following, the notation V will be sometimes replaced by V_μ when we need to stress the dependence of V on μ.

Proposition 1. *The function* V *introduced in (5) has the following properties:*

(i) V *is well defined for* $d \geq 1$*, and continuous for* $d \geq 2$*, on* $[0, +\infty[$.
(ii) *For* $d \geq 2$, $\lim_{\sigma \to +\infty} \mathrm{V}(\sigma) = +\infty$.
(iii) *If* $\mu \leq \frac{\lambda}{2}$*, for* $d \geq 2$, $\exists\ \bar{\sigma} > 0$ *such that* $\mathrm{V}(\bar{\sigma}) < 0$.
(iv) *If* $\mu \leq \frac{\lambda}{2}$*, for* $d \geq 2$, $\exists\ \sigma_{opt} > 0$ *such that* $\inf_{\{\sigma \geq 0\}} \mathrm{V}(\sigma) = \mathrm{V}(\sigma_{opt}) < 0$.
(v) *For* $d \geq 2$ *and* $\lambda \geq 2$*, if* $\mu \leq \lambda/2$, $\exists\ (\sigma_{opt}, \mu_{opt})$ *such that* $\mathrm{V}_{\mu_{opt}}(\sigma_{opt}) = \inf_{\{\sigma \geq 0, \mu \leq \lambda/2\}} \mathrm{V}_{\mu}(\sigma) < 0$.

Summary of the proof A basic step in the proof of (i) and (ii) is to write V as the sum of $\mathrm{V}^+(\sigma) := E\left[\ln^+ \mathrm{Z}(\sigma)\right]$ and $\mathrm{V}^-(\sigma) := E\left[\ln^- \mathrm{Z}(\sigma)\right]$. Then, for (i), integrands in these quantities are upper bounded by quantities which do not depend on σ and the result follows by the Lebesgue dominated convergence theorem for continuity. For (ii), we show that $\mathrm{V}(\sigma)$ is lower bounded by an expectation of a given random variable which depends on σ. We show using the Monotone convergence theorem that this lower bound converges to infinity when σ goes to infinity and then the result follows. For proving (iii), we prove before, using the concept of uniform integrability of a family of random variables that $d\ \mathrm{V}\left(\frac{\sigma^*}{d}\right)$ ($\sigma^* > 0$ fixed) converges to a certain limit depending on σ^* when d goes to $+\infty$. Using the fact that this limit can be negative for a given σ^* we prove our claim. (iv) is proven using (i), (ii) and (iii) and the intermediate value theorem. (v) follows easily from (iv).

An important point that we can see from this proposition is that, given $\lambda \geq 2$ and $d \geq 2$, and under the condition $\mu \leq \lambda/2$, μ and σ can be chosen such that the relative convergence rate V is optimal (v). We conducted numerical computations of V in the case where $d = 10$, $\lambda = 10$ and equal weights $(w^i)_{1 \leq i \leq \mu}$. The cases with $\mu = 1, 2$ and 5 are represented in Fig. 1 (left). It can be seen that the curves are in conformity with (i), (iii), (iv) and (v) of Proposition 1. In particular, for each μ, there exists a σ_{opt} realizing the minimum of V and we can see that the optimal μ (among the represented μ values 1,2 and 5) is 2. In the following theorem, we will see that the optimal value of V is also the optimal convergence rate in expectation that can be reached by the $(\mu/\mu_w, \lambda)$-ES minimizing a spherical function and using any step-size adaptation rule $(\sigma_n)_{n \geq 0}$, or more precisely, the smallest value of $\frac{1}{n} E\left[\ln \frac{\|\mathbf{X}_n\|}{\|\mathbf{X}_0\|}\right]$ that can be reached by the sequence $(\mathbf{X}_n)_{n \geq 0}$ satisfying the recurrence relation (3). This optimal value corresponds also to the smallest value of $\frac{1}{n} E\left[\ln \frac{\|\mathbf{X}_n\|}{\|\mathbf{X}_0\|}\right]$ that can be reached by the isotropic scale-invariant $(\mu/\mu_w, \lambda)$-ES minimizing a spherical function, i.e., where $(\mathbf{X}_n)_{n \geq 0}$ satisfies the recurrence relation (4) with $\sigma = \sigma_{opt}$.

Theorem 1. *Let* $(\mathbf{X}_n)_{n \geq 0}$ *be the sequence of random vectors satisfying the recurrence relation (3) and relative to the* $(\mu/\mu_w, \lambda)$*-ES minimizing any spherical function (1). Then, for* $\lambda \geq 2$ *and* $d \geq 2$*, we have*

$$\frac{1}{n} E\left[\ln \frac{\|\mathbf{X}_n\|}{\|\mathbf{X}_0\|}\right] \geq \mathrm{V}(\sigma_{opt})\,, \tag{6}$$

where σ_{opt} *is given in Proposition 1 as* $\sigma_{opt} = \mathit{argmin}_{\{\sigma > 0\}} \mathrm{V}(\sigma)$ *and* $\mathrm{V}(\sigma_{opt})$ *corresponds to* $\frac{1}{n} E\left[\ln\left(\frac{\|\mathbf{X}_n\|}{\|\mathbf{X}_0\|}\right)\right]$ *for a* $(\mu/\mu_w, \lambda)$*-ES using the specific scale-invariant adaptation rule with* $\sigma_n = \sigma_{opt}\|\mathbf{X}_n\|$ *and minimizing any spherical function (1).*

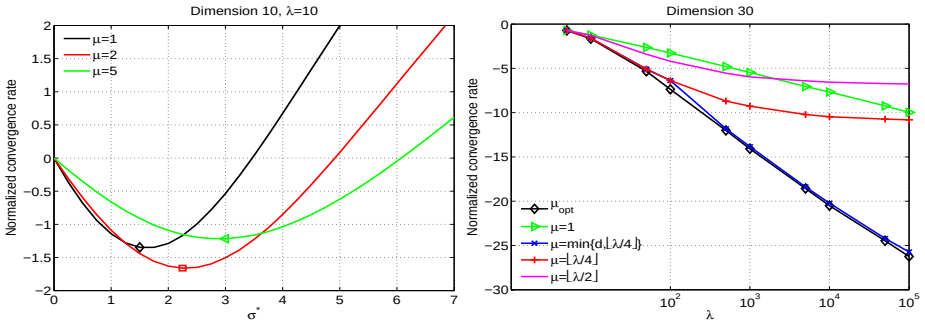

Fig. 1. Left: Plots of the normalized convergence rate $d \times \mathrm{V}_\mu(\frac{\sigma^*}{d})$ where V_μ $(= \mathrm{V})$ is defined in (5) as a function of $\sigma^* > 0$ with $d = 10$, $\lambda = 10$, $w^i = \frac{1}{\mu}$, $\forall\, i = 1, \ldots, \mu$ and $\mu \in \{1, 2, 5\}$. The plots were obtained doing Monte-Carlo estimations of V using 10^6 samples. **Right:** Optimal convergence rate $(d \times \mathrm{V}_\mu(\frac{\sigma^*_{\mathrm{opt}}}{d}))$ associated to different choices of μ as a function of λ for dimension 30 and μ_{opt} realizing the minimum of $(\sigma^*, \mu) \mapsto \mathrm{V}_\mu(\frac{\sigma^*}{d})$.

Summary of the proof The first step for proving the theorem is to remark that:

$$E\left[\ln \frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|}\right] = E\left[E\left[\ln \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \frac{\sigma_k}{\|\mathbf{X}_k\|} \sum_{i=1}^{\mu} w^i \mathbf{N}_k^{i:\lambda}\left(\frac{\mathbf{X}_k}{\|\mathbf{X}_k\|}, \frac{\sigma_k}{\|\mathbf{X}_k\|}\right)\right\| \Big| \left(\mathbf{X}_k, \sigma_k\right)\right]\right].$$

By the isotropy of the norm function and of the multivariate normal distribution, the term $\frac{\mathbf{X}_k}{\|\mathbf{X}_k\|}$ in the previous equation can be replaced by e_1. Then $E\left[\ln \frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|}\right] = E\left[\mathrm{V}\left(\frac{\sigma_k}{\|\mathbf{X}_k\|}\right)\right]$ where $E\left[\mathrm{V}\left(\frac{\sigma_k}{\|\mathbf{X}_k\|}\right)\right]$ is, by Proposition 1, lower bounded by $\mathrm{V}(\sigma_{opt})$. The result follows from summing such inequalities from $k = 0$ to $k = n - 1$.

This theorem states that the artificial scale-invariant adaptation rule with the specific setting $\sigma_n = \sigma_{opt}\|\mathbf{X}_n\|$ is the rule which allows to obtain the best convergence rate of the $(\mu/\mu_w, \lambda)$-ES when minimizing spherical functions. The relative convergence rate is then a tight lower bound that can be reached in this context. Then, for our study on minimization of spherical functions, we will use the $(\mu/\mu_w, \lambda)$-ES with the artificial scale-invariant adaptation rule, i.e., with $\sigma_n = \sigma\|\mathbf{X}_n\|$ where σ is a strictly positive constant. In the specific case where σ equals σ_{opt}, the convergence rate is optimal.

4 Log-Linear Behavior of the Scale-invariant $(\mu/\mu_w, \lambda)$-ES Minimizing Spherical Functions

Log-linear convergence of ESs can be in general shown using the application of different Law of Large Numbers (LLN) such as LLN for independent or orthogonal random variables or LLN for Markov chains. Log-linear behavior has been shown for ESs which do not use recombination [10, 5, 13, 7]. The key idea of the proof is stated in the following proposition.

Proposition 2. *Let $\sigma \geq 0$ and let $(\mathbf{X}_n)_n$ be the sequence of random vectors satisfying the recurrence relation (4). We introduce the sequence of random variables $(\mathrm{Z}_n)_{n\in\mathbb{Z}^+}$ by $\mathrm{Z}_n := \left\| \frac{\mathbf{X}_n}{\|\mathbf{X}_n\|} + \sigma \sum_{i=1}^{\mu} w^i \mathbf{N}_n^{i:\lambda}\left(\frac{\mathbf{X}_n}{\|\mathbf{X}_n\|}, \sigma\right)\right\|$ where $\mathbf{N}_n^{i:\lambda}\left(\frac{\mathbf{X}_n}{\|\mathbf{X}_n\|}, \sigma\right)$ are obtained similarly to (2) but with replacing $(\mathbf{X}_n, \sigma_n)$ by $\left(\frac{\mathbf{X}_n}{\|\mathbf{X}_n\|}, \sigma\right)$. Then for $n \geq 0$, we have*

$$\frac{1}{n} \ln \frac{\|\mathbf{X}_n\|}{\|\mathbf{X}_0\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \mathrm{Z}_k \ a.s. \tag{7}$$

Using the isotropy of the norm function and of the multivariate normal distribution, the terms $\ln \mathrm{Z}_k$ appearing in the right hand side of the previous equation are independent identically distributed with a common expectation $\mathrm{V}(\sigma)$ which we have proved to be finite in Proposition 1. The following theorem is then obtained by the application of the LLN for independent identically distributed random variables with a finite expectation to the right hand side of the previous equation.

Theorem 2 (Log-linear Behavior of the Scale-invariant $(\mu/\mu_w, \lambda)$-ES). *The scale-invariant $(\mu/\mu_w, \lambda)$-ES defined in (4) and minimizing any spherical function (1) converges (or diverges) log-linearly in the sense that for $\sigma > 0$ the sequence $(\mathbf{X}_n)_n$ of random vectors given by the recurrence relation (4) verifies*

$$\lim_{n\to+\infty} \frac{1}{n} \ln \|\mathbf{X}_n\| = \mathrm{V}(\sigma) \tag{8}$$

almost surely, where V refers to the quantity defined in (5).

Theorem 2 establishes that, provided that V is non zero, the convergence of the scale-invariant $(\mu/\mu, \lambda)$-ES minimizing any spherical objective function given in (1) is log-linear. This theorem also provides the convergence (or divergence) rate $\mathrm{V}(\sigma)$ of the sequence $(\ln(\|\mathbf{X}_n\|))_n$: If $\mathrm{V}(\sigma) < 0$, the distance to the optimum, $(\|\mathbf{X}_n\|)_{n\geq 0}$, converges log-linearly to zero and if $\mathrm{V}(\sigma) > 0$, the algorithm diverges log-linearly. From Proposition 1, we know that, for all $d \geq 2$, for all $\lambda \geq 2$ and all $\mu \geq 1$ with the condition $\mu \leq \lambda/2$, there exists $\sigma > 0$ such that $\mathrm{V}(\sigma) < 0$ and therefore the algorithm converges. Moreover, by the same proposition, we know that for any $d, \lambda \geq 2$ there is an optimal choice of (σ, μ) such that the optimal convergence rate is reached.

A practical interest of this result is that if someone chooses the optimal value of μ and is able to tune the adaptation rule of his algorithm such that the quantity $\frac{\sigma_n}{\|\mathbf{X}_n\|}$ is (after an adaptation time) stable around the optimal value for σ, a convergence rate close to the optimal convergence rate can be obtained at least for spherical functions. This can be useful especially for choosing μ when the population size λ is large.

The goal is then to compute those optimal values (i.e., μ_{opt} and σ_{opt}) depending on λ and d. Fortunately, another important point of Theorem 2 is that the convergence rate is expressed in terms of the expectation of a given random variable (see Definition 1). Therefore, the convergence rate V can be numerically computed using Monte-Carlo simulations. Numerical computations allowing to derive optimal convergence rate values and relative optimal values of μ will be investigated in the following section.

5 Numerical Experiments

In this section, we numerically compute, for a fixed dimension and λ, values of μ leading to optimal convergence rates. We compare the convergence rate associated to those optimal μ with the ones obtained with previous choices of μ (proportional to $\lambda/2, \ldots$). We also investigate how the optimal convergence rate depends on the population size λ in particular for $\lambda \gg d$. The context of our numerical study is the specific $(\mu/\mu_w, \lambda)$-ES with intermediate recombination, i.e., with equal weights $w^i = \frac{1}{\mu}, \ \ (i = 1, \ldots, \mu)$ which is simply denoted $(\mu/\mu, \lambda)$-ES.

Since V is expressed in terms of expectation of a random variable, we can perform a Monte-Carlo simulation of the normalized convergence rate $d \times \mathrm{V}_\mu \left(\frac{\sigma^*}{d} \right)$ where $\sigma^* > 0$ is called normalized step-size. The values computed are then relative to the scale-invariant $(\mu/\mu, \lambda)$-ES with $\sigma_n = \frac{\sigma^*}{d} \|\mathbf{X}_n\|$ and minimizing a spherical function. Our experimental procedure relies on finding the minimal value of $(\sigma^*, \mu) \mapsto \mathrm{V}_\mu \left(\frac{\sigma^*}{d} \right)$ for μ in a range $\mu_{\mathbf{range}}$ and for values of σ^* taken in a range $\sigma_{\mathbf{range}}$. As a first experiment, we took $\mu_{\mathbf{range}} = \{2^k; k \in \mathbb{Z}^+ \text{and } 2^k \leq \frac{\lambda}{2}\}$ and $\sigma_{\mathbf{range}} = \ln(\mu + 1) * \ln(\lambda) * [0 : 0.1 : 3]$. We experimented discrete values of λ from $\lambda = 5$ to $\lambda = 10^5$ with a number of Monte-Carlo samplings decreasing as a function of λ from 10^4 to 500. These first computations show that for the values of λ and d tested, the approximation

$$\min_{\{\sigma^* \in \sigma_{\mathbf{range}}\}} d \times \mathrm{V}_\mu \left(\frac{\sigma^*}{d} \right) \simeq a(\lambda, d) \ln^2(\mu) + b(\lambda, d) \ln(\mu) + c(\lambda, d) \qquad (9)$$

is reliable (for $\mu > 1$) and we determined numerically the coefficients $a(\lambda, d), b(\lambda, d)$ and $c(\lambda, d)$. Using these quadratic approximations, we performed a second serie of tests where the values of μ were taken around the optimal value of the polynomial approximation, $\sigma_{\mathbf{range}} = m * \ln(\mu + 1) * \ln(\lambda) * [0 : 0.1 : 3]$ (with $m \leq \frac{2}{3}$) and using more Monte-Carlo samplings.

In Fig. 2 (left), we plotted the normalized optimal convergence rate values and the normalized convergence rates relative to the rule $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$ from [9] as a function of λ and for different dimensions. It can be seen that the optimal convergence rate is, for λ sufficiently large, linear as a function of $\ln(\lambda)$. This result is in agreement with the results in [16]. This figure shows also that the rule $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$ provides convergence rates very close to optimal ones. The curves in Fig. 2 (left) are smooth. However, to obtain the exact optimal values of μ (denoted μ_{opt}), we would need a very large number of Monte-Carlo samplings and (in parallel) a very small discretisation in σ^* that is not affordable. Therefore, we plotted in Fig. 2 (right), the ranges of μ values giving the optimal convergence rate up to a precision of 0.2, as a function of λ and for dimensions $d = 2, 10, 30$ and 100. Those ranges are called 0.2-confidence intervals in μ in the sequel. In the same graph, we plotted values of μ computed as the argmin of the polynomial approximation (9) that we denote μ_{th}. It can be seen that μ_{th} values are in the 0.2-confidence interval in μ. However, the values $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$ for $\lambda = 10^4$ and $d \in \{10, 30, 100\}$, are not in the 0.2-confidence interval in μ. In Figure 1, we compare, for $d = 30$, optimal convergence rates for different choices of μ, namely $\mu = 1$, $\lfloor \frac{\lambda}{4} \rfloor$ ([14]) and $\lfloor \frac{\lambda}{2} \rfloor$ ([6]), $\min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$ ([9]) and the optimal rule (i.e., μ_{opt} values). We

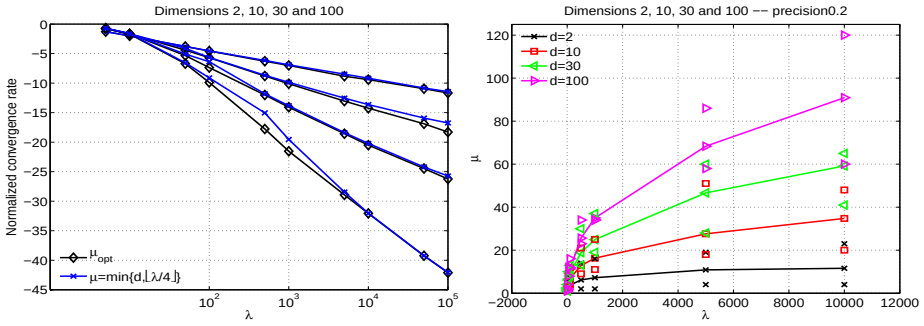

Fig. 2. Left: Plots of the normalized optimal convergence rate $d \times \mathrm{V}_{\mu_{opt}} \left(\frac{\sigma^*_{opt}}{d} \right)$ where $V_\mu (= \mathrm{V})$ is defined in (5) and convergence rate relative to the rule $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$, as a function of λ (log-scale for λ) for dimensions 2, 10, 30 and 100 (from top to bottom). **Right:** Plots of the values μ_{th} (solid lines with markers) giving the optimal μ relative to the quadratic approximation (9) together with extremity of range of μ values (shown with markers) giving convergence rates up to a precision of 0.2 from the optimal numerical value. The dimensions represented are 2, 10, 30 and 100 (from bottom to top).

observe that for μ equal $\lfloor \frac{\lambda}{4} \rfloor$ and $\lfloor \frac{\lambda}{2} \rfloor$, the convergence rate does not scale linearly in $\ln(\lambda)$ and is thus sub-optimal. For $\mu = 1$ and $\min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$, the scaling is linear in $\ln(\lambda)$ and close from the optimal convergence rate for $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$.

Fig. 2 (right) suggests also that the values of μ_{th} vary as a function of $\ln(\lambda)$. Further investigations show that for λ large $\ln(\mu_{th}) = \alpha(d) \ln^2(\ln(\lambda)) + \beta(d)$ where $\alpha(d), \beta(d) > 0$ are some constants that have to be tuned for each dimension (see [12]).

6 Conclusion

In this paper, we have developed a complementary theoretical/numerical approach in order to investigate the isotropic $(\mu/\mu_w, \lambda)$-ES minimizing spherical functions. First, we have shown the log-linear convergence of this algorithm (provided good choice of parameters) with a scale-invariant adaptation rule for the step-size and we have expressed the convergence rate as the expectation of a given random variable. Second, thanks to the expression of the convergence rate, we have numerically computed, using Monte-Carlo simulations, optimal values for the choice of μ and $\frac{\sigma_n}{d_n}$ and their relative optimal convergence rates. We have investigated in particular large values of λ. Our results suggest that the optimal μ is monotonously increasing in λ as opposed to the rule $\mu = \min\{\lfloor \frac{\lambda}{4} \rfloor, d\}$ proposed in [9] but that however this latter rule gives a convergence rate close to the optimal one. We have confirmed as well that for the rules $\mu = \lfloor \frac{\lambda}{4} \rfloor$ and $\lfloor \frac{\lambda}{2} \rfloor$, the convergence rate does not scale linearly in $\ln(\lambda)$ and is thus sub-optimal.

Acknowledgments The authors would like to thank Nikolaus Hansen for his advises on how to approach the problem tackled in the paper. This work received support by the French national research agency (ANR) within the COSINUS project ANR-08-COSI-007-12.

References

1. M. Schumer and K. Steiglitz. Adaptive step size random search. *Automatic Control, IEEE Transactions on*, 13:270–276, 1968.
2. I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.
3. H.-P. Schwefel. Collective phenomena in evolutionary systems. In P. Checkland and I. Kiss, editors, *Problems of Constancy and Change-The Complementarity of Systems Approaches to Complexity, Proc. of 31st Annual Meeting Int'l Soc. for General System Research*,2:1025–1033, Budapest, 1987.
4. Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
5. A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.
6. D.V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms 8*, pages 215–237. Springer Verlag, 2005.
7. M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the (1+1)-ES. In N. Monmarché and al., editors, *Proceedings of Evolution Artificielle (EA'07)*, volume 4926 of *LNCS*, pages 207–218. Springer, 2008.
8. F. Teytaud and O. Teytaud. On the parallel speed-up of Estimation of Multivariate Normal Algorithm and Evolution Strategies. In *Proceedings of EvoStar 2009*, pages 655–664. 2009.
9. F. Teytaud. A new selection ratio for large population sizes. In *Proceedings of EvoStar. 2010.*
10. A. Bienvenüe and O. François. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Th. Comp. Sc.*, 306(1-3):269-289, 2003.
11. M. Jebalia, A. Auger and N. Hansen. Log-linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments. *Algorithmica*, (to appear), 2010.
12. M. Jebalia and A. Auger. Log-linear Convergence of the Scale-invariant $(\mu/\mu_w, \lambda)$-ES and Optimal μ for Intermediate Recombination for Large Population Sizes. Research Report n=°7275, INRIA, 2010.
13. Anne Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 334(1–3):35–69, 2005.
14. H.-G. Beyer. *The Theory of Evolution Strategies*. Nat. Comp. Series. Springer-Verlag, 2001.
15. H.-G. Beyer and B. Sendhoff. Covariance Matrix Adaptation revisited - The CMSA Evolution Strategy. In Günter Rudolph and al., editors, *Proceedings of Parallel Problem Solving from Nature (PPSN X)*, volume 5199 of *(LNCS)*, pages 123–132. Springer Verlag, 2008.
16. O. Teytaud and H. Fournier. Lower Bounds for Evolution Strategies Using VC-dimension. In Günter Rudolph and al., editors, *Proceedings of PPSN X*, pages 102–111. Springer, 2008.
17. H.-G. Beyer. *Toward a Theory of Evolution Strategies: On the Benefits of Sex - the (μ/μ,λ) Theory*. *Evolutionary Computation*, 3(1):81–111, 1995.