Learning Rate Adaptation by Line Search in Evolution Strategies with Recombination

Armand Gissler
forename.lastname@polytechnique.edu
Inria and CMAP, Ecole Polytechnique, IP Paris, CNRS
Palaiseau, France

Anne Auger
forename.lastname@inria.fr
Inria and CMAP, Ecole Polytechnique, IP Paris, CNRS
Palaiseau, France

Nikolaus Hansen
forename.lastname@inria.fr
Inria and CMAP, Ecole Polytechnique, IP Paris, CNRS
Palaiseau, France

ABSTRACT

In this paper, we investigate the effect of a learning rate for the mean in Evolution Strategies with recombination. We study the effect of a half-line search after the mean shift direction is established, hence the learning rate value is conditioned to the direction. We prove convergence and study convergence rates in different dimensions and for different population sizes on the sphere function with the step-size proportional to the distance to the optimum.

We empirically find that a perfect half-line search increases the maximal convergence rate on the sphere function by up to about 70%, assuming the line search imposes no additional costs. The speedup becomes less pronounced with increasing dimension. The line search reduces—however does not eliminate—the dependency of the convergence rate on the step-size. The optimal step-size assumes considerably smaller values with line search, which is consistent with previous results for different learning rate settings. The step-size difference is more pronounced in larger dimension and with larger population size, thereby diminishing an important advantage of a large population.

CCS CONCEPTS

• **Theory of computation → Bio-inspired optimization**; **Non-convex optimization**.

KEYWORDS

Evolution Strategy, line search, convergence rate

ACM Reference Format:
Armand Gissler, Anne Auger, and Nikolaus Hansen. 2022. Learning Rate Adaptation by Line Search in Evolution Strategies with Recombination. In *Genetic and Evolutionary Computation Conference (GECCO '22), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512290.3528760

1 INTRODUCTION

Evolution Strategies (ES) are stochastic numerical optimization algorithms that aim at optimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ (we assume in this paper an unconstraint case). They are typically used for difficult problems that can be non-convex, non-differentiable, ill-conditionned, multi-modal, or noisy.

In a $(\mu/\mu, \lambda)$-ES, at each iteration $t+1$ given a current incumbent solution $X_t \in \mathbb{R}^n$ and step-size $\sigma_t > 0$, we generate λ independent samples following a multivariate normal distribution $\mathcal{N}(X_t, \sigma_t^2 I_n)$ with mean X_t and covariance matrix $\sigma_t^2 I_n$ (where I_n denotes the identity matrix of size $n \times n$). Specifically, we consider λ i.i.d. random variables $U_{t+1}^i \sim \mathcal{N}(0, I_n)$, $i = 1, \ldots, \lambda$, and define the permutation of λ elements φ such that

$$f(X_t + \sigma_t U_{t+1}^{\varphi(1)}) \leqslant f(X_t + \sigma_t U_{t+1}^{\varphi(2)}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\varphi(\lambda)}) \quad . \quad (1)$$

We use the $\mu = \lfloor \lambda/2 \rfloor$ best samples to update the mean

$$X_{t+1} = X_t + \kappa_{t+1} \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{\varphi(i)} =: X_t + \kappa_{t+1} \sigma_t S_{U_{t+1}}^{\varphi} \quad (2)$$

where $\kappa_{t+1} > 0$ is the learning rate at iteration $t+1$ and $w_1, \ldots, w_\mu$ are positive recombination weights summing to one. In our simulations, we set the weights proportional to $w_i \propto \log((\lambda+1)/2) - \log(i)$ where $\mu = \lfloor \lambda/2 \rfloor$ [11]. Adaptation of the step-size σ_t is crucial to achieve linear convergence. We define its update in an abstract way as $\sigma_{t+1} = \bar{\sigma}(X_{t+1}, (U_{t+1}^{\varphi(i)})_{i=1,\ldots,\lambda}, \sigma_t)$.

While the learning rate for the mean, κ, is typically set to 1, its influence has been studied in previous works. First, Rechenberg [14] introduced the mutation enhancement factor κ, realizing that larger mutations due to a larger step-size ("mutate big") help to differentiate solutions under noise but need to be dialed back ("inherit small") when the new incumbent is computed (in this case, κ appears in Eq. (1) instead of Eq. (2), which is just another formulation of an equivalent algorithm). Progress rates for the $(1, \lambda)$-ES minimizing a noisy sphere function have also been analyzed for learning rate values ≤ 1 in [9]. More recently, the quality gain on different convex-quadratic functions of a $(\mu/\mu, \lambda)$-ES has been derived and numerically investigated as a function of the learning rate [2].

Figure 1 shows convergence rates with optimal σ simulated on the sphere function versus κ from Eq. (2) for different dimensions and different population sizes (here $\kappa_0 = \kappa_{t+1} = \kappa$ is constant). For decreasing $\kappa < 1$, the convergence rate decreases and approaches zero when κ approaches zero in all cases. The decline starts earlier in smaller dimension or with larger population size. Because $\kappa \times \sigma$ must be roughly constant to achieve maximal convergence rates, small learning rates κ imply large step-sizes σ and the latter disturb the ranking in Eq. (1). For $\sigma \to \infty$, the ranking becomes dominated by the lengths of U_i instead of their directions. The optimal κ is generally between one and two in dimension up to five and it becomes difficult to determine for larger dimension because the

GECCO '22, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9237-2/22/07…$15.00
https://doi.org/10.1145/3512290.3528760

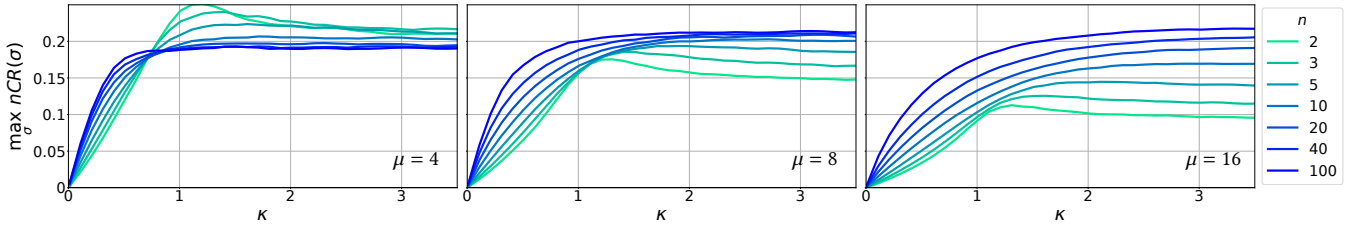

Figure 1: Optimal convergence rate, see Eq. (5), versus the learning rate on the sphere function for different dimensions n and parent numbers μ.

Algorithm 1 $(\mu/\mu, \lambda)$-ES

Require: $X_0 \in \mathbb{R}^n$, $\sigma_0 > 0$, $\kappa_0 > 0$

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: Sample independently $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$
3: $\varphi \leftarrow \arg\operatorname{sort}\{f(X_t + \sigma_t U_{t+1}^i) : i = 1, \ldots, \lambda\}$
4: $\kappa_{t+1} \leftarrow \bar{\kappa}(X_t, \sigma_t \sum_{i=1}^\mu w_i U_{t+1}^{\varphi(i)}, \kappa_t)$
5: $X_{t+1} \leftarrow X_t + \kappa_{t+1}\sigma_t \sum_{i=1}^\mu w_i U_{t+1}^{\varphi(i)}$
6: $\sigma_{t+1} \leftarrow \bar{\sigma}(X_{t+1}, (U_{t+1}^{\varphi(i)})_{i=1,\ldots,\lambda}, \sigma_t)$

graphs are very flat. Overall, the convergence rate gain compared to $\kappa = 1$ does not exceed 50%. The natural question arises, how much these moderate gains improve when κ_{t+1} is chosen *depending* on, that is, *conditional to* the mean shift direction $S_{U_{t+1}}^\varphi$ by executing a half-line search.

In the previously mentioned works, the choice of κ_{t+1} in Eq. (2) was independent of the mean shift direction, $S_{U_{t+1}}^\varphi$. However, in the context of randomized direct search akin to the (1+1)-ES, the effect of a line search along $S_{U_{t+1}}^\varphi$ has already been investigated [12, 15]. In this paper, we integrate this idea into the $(\mu/\mu, \lambda)$-ES. We prove the convergence of the $(\mu/\mu, \lambda)$-ES with half-line search when $\sigma_t = \alpha\|X_t\|$ and specifically address the question by how much a cost-free half-line search can improve the optimal convergence rate. We denote

$$(x, v, \kappa) \mapsto \bar{\kappa}(x, v, \kappa) \tag{3}$$

an abstract learning rate update such that $\kappa_{t+1} = \bar{\kappa}(X_t, \sigma_t S_{U_{t+1}}^\varphi, \kappa_t)$. The adaptive learning rate algorithm that we study is summarized in Algorithm 1.

Linear convergence of Evolution Strategies is empirically observed on many functions and was proven for different variants of step-size adaptive $(1+1)$-ES on the sphere function $f(x) = \|x\|^2$ [12] and on positively homogeneous functions [7] and on wider classes of functions that include smooth strongly convex functions [1]. Analysis for step-size adaptive $(\mu/\mu, \lambda)$-ES are more recent and hold on composites of smooth scaling-invariant functions with strictly increasing functions [16]. For a stochastic algorithm, linear convergence can be defined as the existence of a convergence rate $\mathrm{CR} > 0$ such that

$$\mathrm{CR} = -\frac{1}{\Gamma} \lim_{t\to\infty} \frac{1}{t} \ln \frac{\|X_t - x^*\|}{\|X_0 - x^*\|} \tag{4}$$

$$= -\frac{1}{\Gamma} \lim_{t\to\infty} \mathbb{E} \ln \frac{\|X_{t+1} - x^*\|}{\|X_t - x^*\|} \tag{5}$$

where Γ is the cost per iteration (i.e., the number of calls to f at each step t).

A useful and widely used conceptual algorithm we also adopt here sets the step-size proportional to the distance to the optimum, x^*, i.e., $\sigma_t = \alpha\|X_t - x^*\|$ for a given $\alpha > 0$. The $(\mu/\mu, \lambda)$-ES with this step-size and with fixed learning rate equal to 1 converges linearly, notably on the sphere function [13] and we can equivalently define the (normalized) convergence rate CR as

$$-\frac{1}{\Gamma} \mathbb{E} \ln \frac{\|X_{t+1} - x^*\|}{\|X_t - x^*\|} = -\frac{1}{\Gamma} \mathbb{E}\left[\ln \|X_1 - x^*\| \mid X_0 = x^* + e_1\right]. \tag{6}$$

The term within the limit in Eq. (5) is sometimes referred to as log-progress [5, 6]. *Progress rate theory* refers to a series of work consisting in estimating the progress rate (or equivalently the convergence rate) of the algorithm with step-size proportional to the optimum as a function of the different parameters of the algorithm. It typically assumes that the dimension goes to infinity. Estimates of the convergence rate depending on relevant parameters of the algorithms are useful to understand the influence and relevance of these parameters. The optimal convergence rate, $\max_\alpha \mathrm{CR}(\alpha)$, also bounds the possible convergence rate with any adaptive σ_t [13]. In contrast to many studies on progress rates, in this paper we provide a rigorous mathematical derivation of convergence rates together with asymptotic estimates when the dimension goes to infinity.

The rest of the paper is organized as follows. Section 2 outlines sufficient conditions for linear convergence of the $(\mu/\mu, \lambda)$-ES with adaptive learning rate. Section 3 shows that the $(\mu/\mu, \lambda)$-ES with optimal adaptive learning rate satisfies these conditions. We give expressions of the convergence rates and numerical estimations to see the potential benefits from a line search. In Section 4, we study one particular line search and Section 5 concludes the paper. Supplementary proofs are provided in [10].

Notations. Let $\mathbb{N}$ and $\mathbb{R}_+$ be respectively the set of non-negative natural and real numbers, $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and $n \in \mathbb{N}^*$. We consider the objective function $f\colon \mathbb{R}^n \to \mathbb{R}$. Let $\mu, \lambda \in \mathbb{N}^*$ and $\mu \leqslant \lambda$. Let $\mathbf{w} = (w_1, \ldots, w_\mu)$ be a vector of μ strictly positive weights summing to one and $\mu_\mathrm{w} = (\sum_{i=1}^\mu w_i^2)^{-1}$. Given $U = (U^1, \ldots, U^\lambda) \in \mathbb{R}^{n\times\lambda}$ and a permutation φ, we define $S_U^\varphi = \sum_{i=1}^\mu w_i U^{\varphi(i)}$ and analogously S_V^φ and $S_\mathcal{N}^\varphi$.

For any vector $x \in \mathbb{R}^n$, $[x]_k$ denotes the k-th coordinate of x. We denote I_n the identity matrix of size $n \times n$ and $\mathcal{N}(0, I_n)$ the standard multivariate normal distribution. We refer to a half-line search also as line search.

2 LINEAR CONVERGENCE OF THE $(\mu/\mu, \lambda)$-ES WITH ADAPTIVE LEARNING RATE

In this section, we generalize results of linear convergence for the $(\mu/\mu, \lambda)$-ES with step-size proportional to the distance to the optimum and fixed learning rate $\kappa = 1$ to the case with adaptive learning rate. Our analysis encompasses a constant learning rate $\kappa \neq 1$ and a learning rate obtained via a *perfect line search* over $\kappa \geq 0$. We optimize the sphere function and assume that the optimum is in zero w.l.o.g. (the studied algorithms are invariant under search space translations). We also generalize proofs of asymptotic estimates of the convergence rate when the dimension goes to infinity.

Given a starting point $x \in \mathbb{R}^n$ and a direction $v \in \mathbb{R}^n$, *perfect line search* yields the learning rate $\kappa \geqslant 0$ which minimizes $\kappa \mapsto f(x+\kappa v)$. Hence, in Algorithm 1

$$\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}_{\text{PLS}}(x, v) := \arg\min_{\kappa \geqslant 0} f(x + \kappa v) \quad . \tag{7}$$

We introduce a few assumptions on the function $\bar{\kappa}$. First, we assume that $\bar{\kappa}$ is scaling-invariant.

(A1) Scaling-invariance: the function $\bar{\kappa}$ satisfies $\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}(x/r, v/r, \kappa_{\text{init}})$ for all $r > 0$ and inital mean $x \in \mathbb{R}^n$ and direction $v \in \mathbb{R}^n$.

This assumption is trivially satisfied with a constant learning rate on any function.

We remind that a function $f : \mathbb{R}^n \to \mathbb{R}$ is scaling-invariant [8, 17] with respect to $x^* \in \mathbb{R}^n$ if and only if for all $x, y \in \mathbb{R}^n$ and $r > 0$,

$$f(rx + x^*) \leqslant f(ry + x^*) \Leftrightarrow f(x + x^*) \leqslant f(y + x^*) \quad .$$

For a learning rate from perfect line search as in Eq. (7), assumption (A1) is satisfied when f is a scaling-invariant function with a finite arg min, for example the sphere function or any convex-quadratic function. This is formalized in the next lemma.

Lemma 2.1. *Consider the learning rate update $\bar{\kappa}_{\text{PLS}}$ associated to the perfect line search optimizing a lower-bounded continuous scaling-invariant function (w.r.t. 0). Then $\bar{\kappa}_{\text{PLS}}$ satisfies (A1). In particular, (A1) is satisfied when minimizing the sphere function.*

Proof. Since we assume that the scaling-invariant function is lower-bounded and continuous, the arg min of the perfect line search exists. Let $x, v \in \mathbb{R}^n$ and $r > 0$. Then, exploiting the definition of scaling-invariance, we have

$$\begin{aligned}\bar{\kappa}_{\text{PLS}}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v) &= \arg\min_{\kappa \geqslant 0} \frac{1}{r} f(x + \kappa v) \\ &= \arg\min_{\kappa \geqslant 0} f\left(\frac{x}{r} + \kappa \frac{v}{r}\right) = \bar{\kappa}_{\text{PLS}}\left(\frac{x}{r}, \frac{v}{r}\right) \quad .\end{aligned}$$

Thus, (A1) holds. □

Second, we look at the mean update of the $(\mu/\mu, \lambda)$-ES. Starting from any random variable X, optimizing the sphere function, we require the following.

(A2) Isotropy on the sphere: The distribution of the norm of the updated mean starting from $X/\|X\|$ equals the distribution starting from e_1, i.e.

$$\left\| \frac{X}{\|X\|} + \alpha\bar{\kappa}\left(\frac{X}{\|X\|}, \alpha S_U^{\varphi}, \kappa\right) S_U^{\varphi} \right\| \overset{d}{=} \left\| e_1 + \alpha\bar{\kappa}\left(e_1, \alpha S_V^{\varphi}, \kappa\right) S_V^{\varphi} \right\| \tag{8}$$

where the law of X has a positive pdf, $U^1, \ldots, U^\lambda, V^1, \ldots, V^\lambda$ are independent standard Gaussians vectors, and the U^i are ordered by φ w.r.t $\|X + \alpha\|X\|U^i\|$ and the V^i w.r.t $\|e_1 + \alpha V^i\|$.

We establish now that invariance of $\bar{\kappa}$ to rotation with respect to the first two arguments implies condition (A2).

Lemma 2.2. *If the learning rate update $\bar{\kappa}$ is rotation-invariant, i.e., $\bar{\kappa}(Rx, Rv, \kappa) = \bar{\kappa}(x, v, \kappa)$ for any rotation matrix R, then (A2) holds.*

Proof. Consider R_X the rotation which maps $\frac{X}{\|X\|}$ to e_1. If $\bar{\kappa}$ is rotation-invariant, then

$$\bar{\kappa}\left(\frac{X}{\|X\|}, \alpha \sum_{i=1}^{\mu} w_i U^{\varphi(i)}, \kappa\right) = \bar{\kappa}\left(e_1, \alpha \sum_{i=1}^{\mu} w_i (R_X U)^{\varphi(i)}, \kappa\right)$$

applying rotation-invariance to R_X. Then the $(R_X U)^{\varphi(i)}$ are ordered with respect to $\|e_1 + \alpha R_X U^i\|$, hence (A2) holds. □

For a constant learning rate, the property (A2) is known [13] and is key for deriving the expression of the convergence rate as the expected log progress starting from e_1 with step-size α.

The almost sure asymptotic linear convergence with a constant learning rate is derived using the strong law of large numbers (LLN) applied to the i.i.d. random variables $\|X_{t+1}\| \,/\, \|X_t\|$ [13]. We remind that the strong LLN applied to a sequence of i.i.d. random variables $\{Y_t; t \geq 0\}$ with common law Y states that $\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} g(Y_k) = \mathbb{E}[g(Y)]$ for any continuous function g such that $\mathbb{E}|g(Y)| < \infty$.

Assuming (A1) and (A2) and a learning rate update that is independent of the previous learning rate, we can generalize the property that the random variables $\{\|X_{t+1}\| \,/\, \|X_t\|, t \geq 0\}$ are i.i.d. to the case with varying κ as proven in the next lemma.

Lemma 2.3. *Assume that (A1) and (A2) hold and that $\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}(x, v)$. Denote $Z_t = \|X_{t+1}\|/\|X_t\|$. Then the $(Z_t)_{t\in\mathbb{N}}$ are i.i.d. with the same law as $\|e_1 + \alpha\bar{\kappa}(e_1, \alpha S_V^{\varphi}) S_V^{\varphi}\|$, where $S_V^{\varphi} = \sum_{i=1}^{\mu} w_i V^{\varphi(i)}$ and $V^{\varphi(i)}$ are λ i.i.d. standard n-dimensional Gaussian vectors ordered w.r.t. $\|e_1 + \alpha V^i\|$.*

The proof is given in [10].

We can directly use the previous lemma to prove the asymptotic linear convergence of a $(\mu/\mu, \lambda)$-ES with adaptive learning rate such that $\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}(x, v)$. Indeed, since

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|X_{k+1}\|}{\|X_k\|}$$

and the random variables $\{Z_k = \frac{\|X_{k+1}\|}{\|X_k\|}, k \geq 0\}$ are i.i.d. with distribution $\left\| e_1 + \alpha\bar{\kappa}\left(e_1, \alpha S_V^{\varphi}\right) S_V^{\varphi} \right\|$, we deduce from the strong LLN that almost surely

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \mathbb{E} \ln \left\| e_1 + \alpha\bar{\kappa}\left(e_1, \alpha S_V^{\varphi}\right) S_V^{\varphi} \right\| \quad , \tag{9}$$

given the RHS of Eq. (9) is integrable. This is known when the learning rate is constant [13], and can be easily generalized for optimal line search, for details see [13, Proof of Proposition 1 (i)].

Overall the following linear convergence result holds under (A1) and (A2) and the learning rate update satisfying $\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}(x, v)$.

Theorem 2.4. *Consider a $(\mu/\mu, \lambda)$-ES with adaptive learning rate (Algorithm 1) optimizing the sphere function with optimum in zero and with $\sigma_t = \alpha\|X_t\|$. Assume that the learning rate update satisfies (A1) and (A2). Additionally, assume that $\bar{\kappa}$ does not depend on the parameter κ_{init} (i.e. $\bar{\kappa}(x, v, \kappa_{\text{init}}) = \bar{\kappa}(x, v)$), and that the random variable $\ln \|e_1 + \alpha\bar{\kappa}(e_1, \alpha S_U^\varphi) S_U^\varphi\|$ is integrable where $S_U^\varphi = \sum_{i=1}^{\mu} w_i U^{\varphi(i)}$, $U^i \sim \mathcal{N}(0, I_n)$ i.i.d., and index $\varphi(i)$ sorts the U^i w.r.t. $\|e_1 + \alpha U^i\|$. Let C be the constant cost in terms of number of function evaluations for updating the learning rate. Denote $\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa})$ as*

$$\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}) = -\frac{\mathbb{E} \ln \left\| e_1 + \alpha\kappa S_U^\varphi \right\|}{\lambda + C} \quad , \tag{10}$$

$$= -\frac{1}{2} \frac{\mathbb{E}\left[\ln\left(1 + 2\alpha\kappa [S_U^\varphi]_1 + \|\alpha\kappa S_U^\varphi\|^2\right)\right]}{\lambda + C} \tag{11}$$

where $\kappa = \bar{\kappa}(e_1, \alpha S_U^\varphi)$. Then, $\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa})$ is the convergence (or divergence) rate of the algorithm in the sense that the normalized expected log progress

$$\frac{1}{\lambda + C} \mathbb{E} \ln \frac{\|X_{t+1}\|}{\|X_t\|} = -\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}) \tag{12}$$

is the same for all $t = 1, 2, \ldots$ and almost surely

$$\lim_{t \to \infty} \frac{1}{t \times (\lambda + C)} \ln \frac{\|X_t\|}{\|X_0\|} = -\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}) \quad . \tag{13}$$

Proof. The almost sure convergence in Eq. (13) derives from the strong LLN as in Eq. (9). Eq. (12) is deduced from taking the expectation of $\ln \frac{\|X_{t+1}\|}{\|X_t\|}$, which has a known law given in Lemma 2.3, where we know that $\left(\frac{\|X_{t+1}\|}{\|X_t\|}\right)_{t \in \mathbb{N}}$ is i.i.d. Finally the developed form of Eq. (11) is obtained by using the identity

$$\ln \|e_1 + v\| = \frac{1}{2} \ln \left(1 + 2[v]_1 + \|v\|^2\right)$$

applied to $v = \alpha\bar{\kappa}(e_1, \alpha S_U^\varphi) S_U^\varphi$. □

In order to drop the assumption that $\bar{\kappa}$ is independent of the parameter κ_{init} in Theorem 2.4, we consider the process $(X_t/\|X_t\|, \kappa_t)$ which is a homogenous Markov chain when assuming that (i) the step-size is proportional to the distance to the optimum and (ii) the learning rate update function satisfies a scaling-invariance property.

Lemma 2.5. *Consider the $(\mu/\mu, \lambda)$-ES defined in Algorithm 1 with adaptive learning rate $\kappa_{t+1} = \bar{\kappa}(X_t, \sigma_t S_{U_{t+1}}^\varphi, \kappa_t)$ optimizing the sphere function (with optimum in 0) with step-size $\sigma_t = \alpha\|X_t\|$. Assume (A1) holds. Define $(Z_t, \kappa_t) := (X_t/\|X_t\|, \kappa_t)$. Then (Z_t, κ_t) is a homogenous Markov chain which satisfies*

$$(Z_{t+1}, \kappa_{t+1}) = \left(\frac{Z_t + \alpha\kappa_{t+1} S_{U_{t+1}}^\varphi}{\left\| Z_t + \alpha\kappa_{t+1} S_{U_{t+1}}^\varphi \right\|}, \bar{\kappa}\left(Z_t, \alpha S_{U_{t+1}}^\varphi, \kappa_t\right) \right) \quad , \tag{14}$$

where φ sorts the U_{t+1}^i w.r.t. $\|Z_t + \alpha U_{t+1}^i\|$.

Proof. Let $t \in \mathbb{N}$. Note that φ sorts the U_{t+1}^i w.r.t. $\|X_t + \sigma_t U_{t+1}^i\| = \|X_t + \alpha\|X_t\| U_{t+1}^i\|$ so equivalently w.r.t. $\|Z_t + \alpha U_{t+1}^i\|$. Then, $\kappa_{t+1} = \bar{\kappa}(X_t, \alpha\|X_t\| S_{U_{t+1}}^\varphi, \kappa_t) = \bar{\kappa}(Z_t, \alpha S_{U_{t+1}}^\varphi, \kappa_t)$ by (A1), and $X_{t+1} = X_t + \kappa_{t+1}\, \alpha\|X_t\| S_{U_{t+1}}^\varphi = \|X_t\|(Z_t + \alpha\kappa_{t+1} S_{U_{t+1}}^\varphi)$, thus $Z_{t+1} = (Z_t + \alpha\kappa_{t+1} S_U^\varphi) \,/\, \|Z_t + \alpha\kappa_{t+1} S_U^\varphi\|$. □

We pose the assumption that the chain $(X_t/\|X_t\|, \kappa_t)$ is ergodic.

(A3) The Markov chain $(X_t/\|X_t\|, \kappa_t)$ is ergodic (i.e. irreducible and positive recurrent) with invariant probability measure π. Also assume that

$$\mathbb{E}_{(Z,\kappa)\sim\pi}[|\ln \|Z + \alpha\bar{\kappa}(Z, \alpha S_U^\varphi, \kappa) S_U^\varphi\||] < +\infty \quad . \tag{15}$$

Remark that proving the ergodicity of the chain can be cumbersome. Ergodicity allows to use the LLN for Markov chains to conclude linear convergence and get an expression of the convergence rate.

Theorem 2.6. *Consider a $(\mu/\mu, \lambda)$-ES with adaptive learning rate (Algorithm 1) optimizing the sphere function with step-size proportional to the distance to the optimum with multiplicative factor α. Assume that the learning rate update satisfies (A1), (A2) and (A3). Let C denote the constant cost of the line search corresponding to the update function $\bar{\kappa}$. The same conclusions as in Theorem 2.4 hold where the convergence rate is expressed as*

$$\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}) = -\frac{\mathbb{E}_{\kappa\sim\pi_\kappa} \ln \left\| e_1 + \alpha\bar{\kappa}\left(e_1, \alpha S_U^\varphi, \kappa\right) S_U^\varphi \right\|}{\lambda + C} \quad , \tag{16}$$

where φ sorts the U^i w.r.t. $\|e_1 + \alpha U^i\|$ and where the U^i are i.i.d. r.v. of distribution $\mathcal{N}(0, I_n)$, and π_κ is the marginal distribution of π w.r.t. the last variable.

Proof. We have

$$\frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|X_{k+1}\|}{\|X_k\|}$$

$$= \sum_{k=0}^{t-1} \ln \left\| \frac{X_k}{\|X_k\|} + \alpha\bar{\kappa}(X_k/\|X_k\|, \alpha S_{U_{k+1}}^\varphi, \kappa_k) S_{U_{k+1}}^\varphi \right\|.$$

As (A3) holds, we can apply the ergodic theorem to get that

$$\lim_{t \to \infty} \frac{1}{t} \ln \frac{\|X_t\|}{\|X_0\|} = \mathbb{E}_{(Z,\kappa)\sim\pi} \ln \left\| Z + \alpha\bar{\kappa}(Z, \alpha S_U^\varphi, \kappa) S_U^\varphi \right\| \quad ,$$

then we get the desired result by (A2). □

2.1 Asymptotic Limit of Convergence Rates

We derive in the next theorem the asymptotic convergence rate when the dimension tends to infinity. To do so, we need an asymptotic assumption on the update function $\bar{\kappa}$.

(A4) Let $u^1, \ldots, u^\mu \in \mathbb{R}^{\mathbb{N}}$ be μ infinite dimensional random vectors such that the $u_k^i, i = 1, \ldots, \mu; k \in \mathbb{N}^*$, are i.i.d. r.v. of distribution $\mathcal{N}(0, 1)$. For any $v \in \mathbb{R}^{\mathbb{N}}$, denote $[v]_{\leqslant n}$ the vector of $\mathbb{R}^n$ such that $[[v]_{\leqslant n}]_i = [v]_i$ for $i = 1, \ldots, n$. Then, there exists a function $\bar{\kappa}^\infty \colon \mathbb{R}^\mu \times \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{n \to \infty} \bar{\kappa}\left(e_1, \frac{\alpha}{n} \sum_{i=1}^{\mu} w_i [u^i]_{\leqslant n}, \cdot\right) = \bar{\kappa}^\infty(([u^1]_1, \ldots, [u^\mu]_1), \cdot) \quad . \tag{17}$$

This assumption is trivially satisfied for a constant learning rate $\bar{\kappa} = \kappa_0$ with $\bar{\kappa}^\infty = \kappa_0$. We will prove in Lemma 3.2 that (A4) is satisfied for perfect line search.

Theorem 2.7. *Suppose (A1),(A2) and (A4) hold. Furthermore, assume that the assumptions in Theorem 2.4 hold. Additionally, assume that $\bar{\kappa}$ is upper-bounded or equal to $\bar{\kappa}_{\text{PLS}}$. Then, the convergence rate*

of the $(\mu/\mu,\lambda)$-ES with adaptive learning rate satisfies the following limit when $n \to \infty$

$$\lim_{n\to\infty} n\mathrm{CR}\left(\frac{\alpha}{n}, \mathbf{w}, \bar{\kappa}\right) = -\frac{\mathbb{E}\left[2\alpha\kappa^{\infty} S_{\mathcal{N}}^{\varphi} + (\alpha\kappa^{\infty})^2/\mu_{\mathrm{w}}\right]}{2(\lambda + C)} \tag{18}$$

where $\kappa^{\infty} := \bar{\kappa}^{\infty}((\mathcal{N}^{\varphi(i)})_i, 1)$ and the $\mathcal{N}^{\varphi(i)}$ are the order statistics of λ i.i.d. standard normal distributions.

Proof. Let $\mathcal{S}_\lambda$ be the set of permutations of $\{1,\dots,\lambda\}$. Let $U^1,\dots,U^\lambda$ be λ independent standard multivariate normal distributions. Then, we have that

$$\ln\left(1 + 2\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\varphi}\right)[S_U^{\varphi}]_1 + \left\|\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\varphi}\right)S_U^{\varphi}\right\|^2\right)$$
$$= \sum_{\phi\in\mathcal{S}_\lambda} \ln\left(1 + 2\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\phi}\right)[S_U^{\phi}]_1 + \left\|\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\phi}\right)S_U^{\phi}\right\|^2\right)\delta_n^{\phi}$$

where $\delta_n^{\phi} = \mathbf{1}_{\{h_{\alpha/n}(U^{\phi(1)})<\cdots<h_{\alpha/n}(U^{\phi(\lambda)})\}}$, and, for any $i = 1,\dots,\lambda$ $h_{\alpha/n}(U^i) = 2U_1^i + \frac{\alpha}{n}\|U^i\|^2$ such that, almost surely, we have that $\lim_{n\to\infty} h_{\alpha/n}(U^i) = 2U^i + \alpha$. Therefore, almost surely, for any permutation $\phi \in \mathcal{S}_\lambda$:

$$\mathbf{1}_{\{h_{\alpha/n}(U^{\phi(1)})<\cdots<h_{\alpha/n}(U^{\phi(\lambda)})\}} \xrightarrow[n\to\infty]{} \mathbf{1}_{\{[U^{\phi(1)}]_1<\cdots<[U^{\phi(\lambda)}]_1\}} \ .$$

Moreover, by the law of Large Numbers we have that $\left(\|S_U^{\phi}\|^2/n\right)^{-1}$ converges almost surely to $\left(\sum_{i=1}^{\mu} w_i^2\right)^{-1} = \mu_{\mathrm{w}}$, so we have almost surely the following limit

$$\lim_{n\to\infty} n\ln\left(1 + 2\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\phi}\right)[S_U^{\phi}]_1 + \left\|\frac{\alpha}{n}\bar{\kappa}\left(e_1, \frac{\alpha}{n}S_U^{\phi}\right)S_U^{\phi}\right\|^2\right)$$
$$= 2\alpha\bar{\kappa}^{\infty}(([U^{\phi(i)}]_1)_i)[S_U^{\phi}]_1 + \alpha^2(\bar{\kappa}^{\infty})^2(([U^{\phi(i)}]_1)_i)(\mu_{\mathrm{w}})^{-1} \ .$$

The assumption that $\bar{\kappa}$ is upper-bounded or equal to $\bar{\kappa}_{\mathrm{PLS}}$ implies the uniform integrability of the LHS of the above equation, we refer to [10, Lemma A.1] for a full proof.

Thus, by the dominated convergence theorem

$$\lim_{n\to\infty} n\,\mathrm{CR}(\alpha/n, \mathbf{w}, \bar{\kappa}) = \frac{1}{2(\lambda + C)}$$
$$\times\,\mathbb{E}\left[2\alpha\bar{\kappa}^{\infty}((\mathcal{N}^{\varphi(i)})_i)S_{\mathcal{N}}^{\varphi} + \alpha^2\bar{\kappa}^{\infty}((\mathcal{N}^{\varphi(i)})_i)^2(\mu_{\mathrm{w}})^{-1}\right] \ .$$

□

With Theorems 2.4 and 2.7, we recover the convergence rates (and the asymptotic limit of the convergence rates) of the $(\mu/\mu,\lambda)$-ES with constant learning rate. This result however can be established more easily [13].

Now, we are interested in update functions $\bar{\kappa}$ that depend on the previous line search result. We assume ergodicity via (A3) and moreover convergence (in law) of the invariant measure:

(A5) Let $\pi_\kappa = \int_{\mathbb{R}^n} \pi(dx,\cdot)$ be the marginal distribution of the probability measure π defined in (A3) w.r.t. the second variable. Assume then that π_κ converges (in law) to π_κ^{∞} when n tends to ∞.

Theorem 2.8. *Suppose (A1-5) hold. Then, the following limit holds when $n \to \infty$,*

$$\lim_{n\to\infty} n\mathrm{CR}\left(\frac{\alpha}{n}, \mathbf{w}, \bar{\kappa}\right) = -\frac{\mathbb{E}\left[2\alpha\kappa^{\infty} S_{\mathcal{N}}^{\varphi} + \alpha^2[\kappa^{\infty}]^2/\mu_{\mathrm{w}}\right]}{2(\lambda + C)} \tag{19}$$

where $\kappa^{\infty} = \bar{\kappa}^{\infty}((\mathcal{N}^{\varphi(i)})_i, \kappa)$, the $\mathcal{N}^{\varphi(i)}$ are λ ordered i.i.d. standard Gaussian r.v., and κ is a r.v. of distribution π_κ^{∞}, independent of the $\mathcal{N}^i$.

Proof. The proof is very similar to the proof of Theorem 2.7, yet we don't use the dominated convergence theorem, but the convergence in law assumed in (A5). □

3 PERFECT LINE SEARCH

We study in this section the $(\mu/\mu,\lambda)$-ES where the learning rate is adapted with a perfect line search as defined in Eq. (7). While this is not a realistic practical method, the convergence rate with perfect line search bounds the possible benefit from a half-line search on the sphere function.

Given a starting point x and a search direction v, we can compute the optimal learning rate on the sphere function, that is, the arg min of the map $\kappa \in \mathbb{R}_+ \mapsto \|x + \kappa v\|^2$ which corresponds to the learning rate associated with perfect line search.

Lemma 3.1. *For any $x \in \mathbb{R}^n$ and $v \in \mathbb{R}^n \setminus \{0\}$,*

$$\bar{\kappa}_{\mathrm{PLS}}(x, v) = \max\left(0, -\frac{\langle x, v\rangle}{\langle v, v\rangle}\right) = \max\left(0, -\frac{\|x\|}{\|v\|}\cos(x, v)\right) \ . \tag{20}$$

Proof. We denote $L\colon \kappa \in \mathbb{R}_+ \mapsto \|x + \kappa v\|^2$. Hence, the derivative of L in $\kappa \geqslant 0$ $L'(\kappa) = 2\langle v, x + \kappa v\rangle$ is positive if and only if $\kappa > -\langle x, v\rangle/\langle v, v\rangle$. Hence the minimum of L is reached for $\kappa = \max(0, -\langle x, v\rangle/\langle v, v\rangle)$. □

The learning rate update function $\bar{\kappa}_{\mathrm{PLS}}$ satisfies Assumptions (A1), (A2) and (A4).

Lemma 3.2. *Suppose that the learning rate is adapted via perfect line search, then (A1), (A2) and (A4) hold. Moreover, the infinite dimension limit for $\bar{\kappa}_{\mathrm{PLS}}$ as described in (A4) is*

$$\bar{\kappa}^{\infty}(u^1,\dots,u^\lambda) = \frac{\mu_{\mathrm{w}}}{\alpha}\mathbf{1}_{\sum_{i=1}^{\mu} w_i u^i < 0}\sum_{i=1}^{\mu} -w_i u^i \ . \tag{21}$$

The proof is given in [10]. Next, we provide the convergence rate of the ES with perfect line search on the sphere function.

Theorem 3.3. *We denote $\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}_{\mathrm{PLS}})$ the convergence rate of the $(\mu/\mu,\lambda)$-ES from Algorithm 1 with weights $\mathbf{w}$, step-size $\sigma_t = \alpha\|X_t\|$, $\alpha > 0$, and the learning rate from perfect line search. Then,*

$$\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}_{\mathrm{PLS}}) = -\frac{1}{2\lambda}\mathbb{E}\left[\mathbf{1}_{[S_U^{\varphi}]_1<0}\ln\left(1 - \frac{\left([S_U^{\varphi}]_1\right)^2}{\|S_U^{\varphi}\|^2}\right)\right] \ , \tag{22}$$

where the U^i are i.i.d. r.v. of distribution $\mathcal{N}(0, I_n)$, and φ sorts the U^i w.r.t. $\|e_1 + \alpha U^i\|$.

Proof. By Theorem 2.4 and Lemmas 2.1, 3.2, $\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}_{\mathrm{PLS}}) = -\mathbb{E}\left[\ln\left(1 + 2\alpha\bar{\kappa}\left(e_1, \alpha S_U^{\varphi}\right)[S_U^{\varphi}]_1 + \alpha^2\bar{\kappa}\left(e_1, \alpha S_U^{\varphi}\right)^2 \|S_U^{\varphi}\|^2\right)\right] / (2\lambda)$. By Lemma 3.1, we know that

$$\bar{\kappa}_{\mathrm{PLS}}(e_1, \alpha S_U^{\varphi}) = \max\left(0, -\frac{\langle S_U^{\varphi}, e_1\rangle}{\langle S_U^{\varphi}, S_U^{\varphi}\rangle}\right) = -\mathbf{1}_{[S_U^{\varphi}]_1<0}\frac{[S_U^{\varphi}]_1}{\alpha\|S_U^{\varphi}\|^2} \quad .$$

This yields the desired result. □

The convergence rate for perfect line search in Eq. (22) is, trivially, always positive. The only dependency of Eq. (22) on α or σ_t comes from the φ-sorting of U^i (ranking selection). On the sphere function with isotropic sampling, the dependency on $\alpha > 0$ disappears when the U^i are sampled with constant length[1] or for dimension to infinity.

In the next theorem, we derive the limit of the convergence rate obtained in Theorem 3.3 when the dimension goes to infinity.

Theorem 3.4. *When the dimension n goes to infinity, we have the following limit for the convergence rate with perfect line search:*

$$\lim_{n\to\infty} n\mathrm{CR}\left(\frac{\alpha}{n}, \mathbf{w}, \bar{\kappa}_{\mathrm{PLS}}\right) = \frac{\mu_{\mathrm{w}}\mathbb{E}\left[{S_{\mathcal{N}}^{\varphi}}^2 \mathbf{1}_{S_{\mathcal{N}}^{\varphi}<0}\right]}{2\lambda} \tag{23}$$

where φ sorts the λ independent r.v. $\mathcal{N}^i \sim \mathcal{N}(0,1)$.

Proof. This follows from Theorem 2.7 and Lemma 3.2 by plugging Eq. (21) in Eq. (18). □

The limit in Eq. (23) does not anymore depend on the parameter $\alpha > 0$, as the selection in Theorem 3.4 only depends on order statistics of the (one-dimensional) Gaussian variables $\mathcal{N}^i$ which are independent of α.

3.1 Numerical Results

Figure 2 shows numerical estimations of the convergence rates on the sphere function with perfect line search (dashed lines) as obtained in Eq. (22), and without line search and $\kappa_t = 1$ (solid lines) plotted versus $n \times \alpha$ for different dimensions n and for $\mu = 4$, 8 and 16. The estimations are obtained by Monte-Carlo estimations of the expectation with 10^5 samples. Black lines show the limit of the convergence rates for $n \to \infty$ using Eq. (23).

Perfect and cost-free line search invariably increases the convergence rate—the κ from Eq. (7) always improves X_{t+1} compared to any initial κ. The dependency of the convergence rate on σ is considerably less pronounced with line search and even vanishes for $n \to \infty$. The product $\kappa\sigma$ (not shown) remains roughly constant for smaller than optimal values of σ and for $\sigma \to 0$ [2] and decreases to zero only for $\sigma \to \infty$.

Figure 3 shows empirical distributions of the learning rate from Eq. (20) for the best step-sizes in Figure 2 marked with crosses. For step-sizes which are optimal with $\kappa = 1$ (solid lines), all distributions stay relatively close to $\kappa = 1$, as expected, but with a modest bias to larger values, increasing with increasing μ. For step-sizes which are optimal with perfect line search, we observe larger values of κ in particular with larger dimension and larger population size. The probability that the optimal κ is negative is always smaller than 2.1% ($\mu = 4, n = 2$) and 3.4% ($\mu = 4, n = 5$) respectively with and without line search[2].

Figure 4 compares the maximal convergence rates (left subfigure) as marked by crosses in Figure 3 and the respective step-sizes (middle) by plotting the ratios with perfect line search versus $\kappa = 1$. Perfect line search increases the convergence rate by at most a factor of 1.72 (72%) as observed with $\mu = 16$ in dimension two. For larger dimension, in particular when $n > \mu$, the gain is decreasing and gets asymptotically 19.5, 8.0, 3.7% for $\mu = 4, 8, 16$, respectively.

As observed before, the optimal step-size is invariably smaller with κ adaptation than with $\kappa = 1$. The effect is more pronounced with increasing dimension and larger population size (Figure 4, middle). For $n = 20$ and $\mu = 16$, the optimal step-size is more than three times smaller.

Algorithm 2 Dichotomic line search

Require: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_{\mathrm{init}} > 0$, $\varepsilon > 0$, $\beta \in (0,1)$
Result: κ
$\kappa^0 \leftarrow \frac{\kappa_{\mathrm{init}}}{2}$, $\kappa^1 \leftarrow 2 \times \kappa_{\mathrm{init}}$
while $\kappa^1 - \kappa^0 \geqslant \varepsilon \times \kappa_{\mathrm{init}}$ **do**
  if $f(X + \kappa^0 v) < f(X + \kappa^1 v)$ **then**
    $\kappa^1 \leftarrow \beta\kappa^1 + (1-\beta)\kappa^0$
  else
    $\kappa^0 \leftarrow \beta\kappa^0 + (1-\beta)\kappa^1$
$\kappa \leftarrow \mathrm{argmin}\left\{f(X + \kappa v); \kappa = \kappa^0, \kappa^1\right\}$

4 DICHOTOMIC LINE SEARCH

We introduce a comparison-based line search algorithm (tailored to a symmetrical function like the sphere function) to which we apply the theoretical results of Section 2.

Algorithm 2 presents a dichotomic line search for the learning rate $\kappa \in [\kappa_{\mathrm{init}}/2, 2\kappa_{\mathrm{init}}]$. After initializing κ^0, κ^1 as $\kappa_{\mathrm{init}}/2, 2\kappa_{\mathrm{init}}$, we update in each iteration the worse of the two κ values towards the better by the factor $1-\beta$. We stop when both values are close enough and return the better. We denote by $\bar{\kappa}_{\mathrm{DLS}}^{\varepsilon,\beta}$ the resulting learning rate update $\bar{\kappa}(x, v, \kappa_{\mathrm{init}}) = \bar{\kappa}_{\mathrm{DLS}}^{\varepsilon,\beta}(x, v, \kappa_{\mathrm{init}})$ with parameters x, v, ε, β, and $\kappa_{\mathrm{init}} = 1$. The cost of this update is constant, because the search stops after a deterministic number of iterations which depends only on β and ε as shown in the next lemma.

Lemma 4.1. *The dichotomic line search, Algorithm 2, with parameters $X, v, \kappa_{\mathrm{init}}, \varepsilon, \beta$ requires $C(\beta, \varepsilon) = \max\{0, \lceil \ln(2\varepsilon/3)/\ln\beta \rceil\} + 2$ evaluations of f.*

Proof. Consider the line search obtained with $\bar{\kappa}_{\mathrm{DLS}}^{\varepsilon,\beta}(e_1 v, \kappa_{\mathrm{init}})$. We denote $(\kappa^{i,0}, \kappa^{i,1})_{i=0,\dots}$ the value of κ^1 and κ^0 over the iterations of this line search. Then $\kappa^{0,1} - \kappa^{0,0} = 3/2 \times \kappa_{\mathrm{init}}$, and for any $i = 0, 1, \dots, \kappa^{i+1,1} - \kappa^{i+1,0} = \beta \times (\kappa^{i,1} - \kappa^{i,0})$. Hence, $\kappa^{i,1} - \kappa^{i,0} = 3/2\beta^i\kappa_{\mathrm{init}}$

[1] Indeed, given an initial mean x, step-size $\alpha\|x\|$ and random vectors $(U^i)_i$, the ranking of candidate solutions is given by ranking $\left\|x + \alpha\|x\|U^i\right\|^2 = \|x\|^2 + \alpha^2\|x\|^2\|U^i\|^2 + 2\alpha\|x\|\langle x, U^i\rangle$. Hence when $\|U^i\|$ is a constant, the ranking will only be determined by the projection of the U^i on x and is thus independent of α.

[2] Therefore, a full line search would only marginally increase the convergence rates. For the (1+1)-ES, a full line search increases the convergence rate by a factor of 2 compared to a half-line search, similarly to introducing negative recombination weights [3].

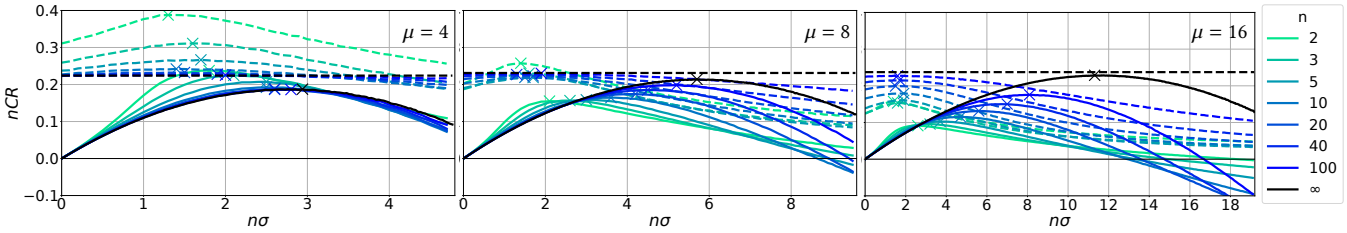

Figure 2: Convergence rate versus step-size on the sphere function without line search (solid lines) and with perfect line search (dashed lines) in different dimensions. Larger values are better, crosses indicate the maximum. The perfect line search is assumed to impose no additional costs, hence it increases the convergence rate for any given n, μ, σ.

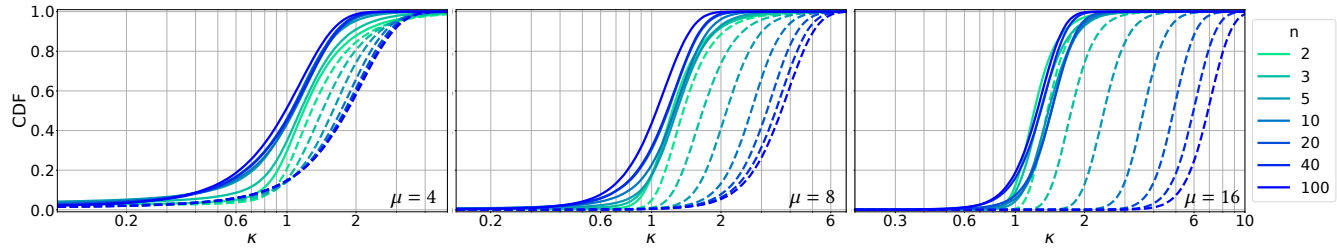

Figure 3: Empirical CDF of the optimal learning rate κ, see Eq. (7), on the sphere function with optimal step-size and perfect line search (dashed) and optimal step-size without line search (solid). The optimal learning rate is proportional to the positive part of the cosine of the angle between the gradient and the sampled direction, see Eq. (20).

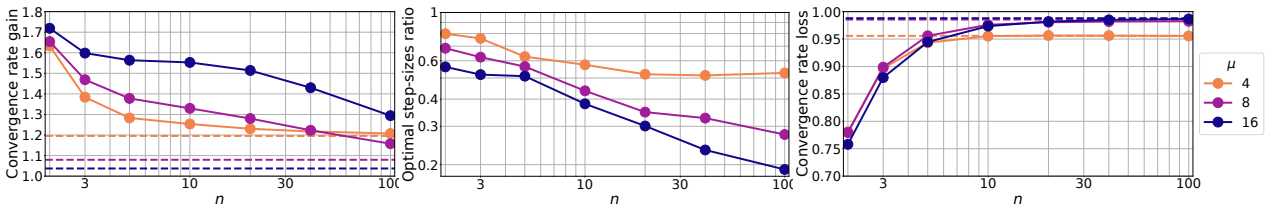

Figure 4: Left: Ratio of the maximal convergence rates from Figure 2 with perfect line search and without, plotted versus dimension n for different values of μ; dashed lines give the asymptotic limit for $n \to \infty$. Middle: respective (optimal) step-size ratios. Right: Ratio of the maximal convergence rates on the sphere function with dichotomic line search from Algorithm 2 versus perfect line search assuming zero cost in both cases.

is smaller than $\varepsilon\kappa_{\text{init}}$ if and only if $i \geqslant \lceil \ln(2\varepsilon/3)/\ln\beta \rceil$. Thus, exactly $\max\{0, \lceil \ln(2\varepsilon/3)/\ln\beta \rceil\}$ iterations are performed before the stopping criterion in Algorithm 2, line 3 is achieved. □

The update function $\bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}}$ satisfies the scaling invariance property (A1), as well as the isotropy on the sphere (A2), and the infinite dimension convergence (A4) as proven in the next lemma.

Lemma 4.2. *On the sphere function, the update function $\bar{\kappa} = \bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}}$ satisfies the assumptions of scaling-invariance (A1), rotation-invariance hence isotropy (A2), and convergence when the dimension tends to ∞ (A4). Moreover, given μ real numbers $u^1, \ldots, u^\mu$ and a non-negative real number κ, Algorithm 2 returns for any given ε, β the infinite-dimension limit of $\bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}}$ as defined in (A4), $\bar{\kappa}^{\infty}((u^i)_i, \kappa)$, when initialized with parameters $X = \alpha^{-1}\mu_{\text{w}} \sum_{i=1}^{\mu} w_i u^i$, $v = 1$, $\kappa_{\text{init}} = \kappa$.*

The proof is given in [10].

To apply Theorem 2.6 and conclude linear convergence of the algorithm, we need to prove that the Markov chain $(X_t/\|X_t\|, \kappa_t)_t$ is ergodic according to (A3). We suspect that ergodicity is difficult to prove and thus leave it as an assumption in the next theorem where we conclude linear convergence and give the expression of the convergence rate depending on the stationary distribution of the Markov chain $(X_t/\|X_t\|, \kappa_t)_t$.

Theorem 4.3. *Let $\beta \in (0,1)$ and $\varepsilon > 0$. We consider the dichotomic line search from Algorithm 2 and the resulting update function $\bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}}$ for κ optimizing the sphere function with step-size $\sigma_t = \alpha\|X_t\|$.*

We assume that (A3) holds, with the marginal distribution of the invariant probability measure π w.r.t. the second variable denoted π_κ.

The $(\mu/\mu, \lambda)$-ES with update function for the learning rate $\bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}}$ converges linearly. We denote $\text{CR}(\alpha, \mathbf{w}, \bar{\kappa}^{\varepsilon,\beta}_{\text{DLS}})$ the convergence rate of $(\mu/\mu, \lambda)$-ES with weights $\mathbf{w}$, and a step-size proportional to the

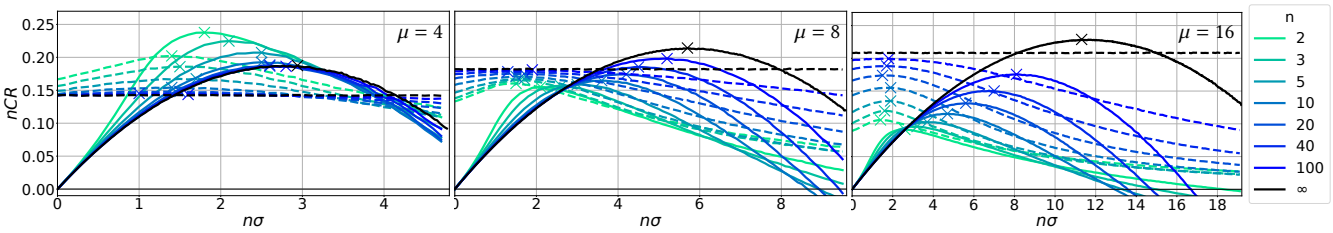

Figure 5: Convergence rate versus step-size on the sphere function without line search (solid lines) and with dichotomic line search (dashed lines) in different dimensions. Larger values are better. Dichotomic line search improves the speed per iteration but imposes additional costs, which correspond to the number of functions calls (here 4).

distance to the optimum up to a multiplicative factor α, with the learning rate from an adaptive line search as given by $\bar{\kappa} = \bar{\kappa}^{\varepsilon,\beta}_{\mathrm{DLS}}$. We denote the constant cost of such a line search by $C(\beta, \varepsilon)$. Then,

$$\mathrm{CR}(\alpha, \mathbf{w}, \bar{\kappa}^{\varepsilon,\beta}_{\mathrm{DLS}}) = -\frac{\mathbb{E}_{\hat{\kappa}\sim\pi_\kappa} \ln \left\| e_1 + \bar{\kappa}^{\varepsilon,\beta}_{\mathrm{DLS}} \left(e_1, \alpha S^{\varphi}_U, \hat{\kappa} \right) \alpha S^{\varphi}_U \right\|^2}{2(\lambda + C(\beta, \varepsilon))}. \quad (24)$$

This theorem results from Theorem 2.6.

We derive now the convergence rate of ES with dichotomic line search with memory when the dimension goes to ∞.

Theorem 4.4 (Asymptotic convergence rate for dichotomic line search). *Let $\alpha > 0$ and $\varepsilon > 0$, $\beta \in (0, 1)$. Consider $\tilde{\kappa}^{\infty}(\hat{\kappa}, \mathbf{N})$ the result of Algorithm 2 with parameters $X = \alpha^{-1}\mu_{\mathrm{w}} \sum_{i=1}^{\mu} w_i \mathcal{N}^{\varphi(i)} \in \mathbb{R}^1, v = 1 \in \mathbb{R}^1, \hat{\kappa}, \varepsilon, \beta$, where $\mathbf{N} = (\mathcal{N}^{\varphi(1)}, \ldots, \mathcal{N}^{\varphi(\lambda)})$ are the order statistics of λ i.i.d. standard Gaussian r.v. $\mathcal{N}^i$.*

We assume that (A3) and (A5) hold for the $(\mu/\mu, \lambda)$-ES with learning rate update function $\bar{\kappa}^{\varepsilon,\beta}_{\mathrm{DLS}}$. Then, the infinite-dimensional limit of the convergence rate given by Theorem 4.3 is

$$\lim_{n\to\infty} n\mathrm{CR}\left(\frac{\alpha}{n}, \mathbf{w}, \bar{\kappa}^{\varepsilon,\beta}_{\mathrm{DLS}}\right) = -\frac{\mathbb{E}\left[\tilde{\kappa}^{\infty}(\hat{\kappa}, \mathbf{N})\, \alpha S^{\varphi}_{\mathcal{N}} + \frac{(\tilde{\kappa}^{\infty}(\hat{\kappa}, \mathbf{N}))^2}{2\mu_{\mathrm{w}}}\right]}{\lambda + C(\beta, \varepsilon)} \quad (25)$$

where $\hat{\kappa}$ is a r.v. of distribution π^{∞}_{κ}, independent of the $\mathcal{N}^i$.

This follows from Theorem 2.8.

4.1 Numerical Results

Figure 5 compares, analogous to Figure 2, numerical estimations for the convergence rates on the sphere function without line search (solid lines) with adaptive dichotomic line search (dashed lines) as in Eq. (24) for finite and in Eq. (25) for infinite dimension, where $\beta = 0.5$ and $\varepsilon = 0.2$, plotted versus σ. To estimate the convergence rate with line search, we discard the first 10^3 of 10^5 samples in order to diminish the effect of κ initialization.

Qualitatively, the dichotomic line search implementation shows similar behavior as perfect line search in Figure 2. The dependency of the convergence rate on σ is attenuated and the rate remains largely unaffected for $\sigma \to 0$. For extreme values of σ, line search always improves the convergence rate compared to $\kappa = 1$. Also the optimal step-size is similar to the one with perfect line search.

Quantitatively, dichotomic line search improves the maximal convergence rate with $\kappa = 1$ only for the largest population size, $\mu = 16$, in all finite dimensions by about 25% and for $\mu = 8$ only in smaller dimensions and only by a smaller margin.

The loss of convergence speed of the implemented line search compared to perfect line search has two reasons: (i) the search costs increase from λ to $\lambda + 4$, which is an increase by 50%, 25%, and 12.5% for $\mu = 4, 8, 16$, respectively. (ii) the resulting κ remain suboptimal. Figure 4, right, shows the loss from the suboptimal κ. The loss is 20–25% in dimension two and decreases quickly to <5% in dimension ten.

5 SUMMARY AND DISCUSSION

Based on the idea of rescaled mutations or, equivalently, a learning rate on the mean in Evolution Strategies (ES), we introduce a line search in the $(\mu/\mu, \lambda)$-ES. A line search identifies the learning rate for the mean update *conditioned* to its direction. Conditioning the learning rate is particularly effective in small dimensions but has overall surprisingly little effects on the convergence speed on the sphere function.

We generalize convergence proofs and estimates of the convergence rates in finite and infinite dimension, previously obtained for learning rate one, to different and adaptive (varying) learning rates. One central assumption is that the step-size is proportional to the distance to the optimum. We analyze in detail the case of a perfect half-line search.

We observe convergence rate improvements by up to 70% for a cost-free line search compared to 50% with the optimal constant learning rate and about 25% with our implementation of a simple line search. These advantages are similar to the 56% gain with mirrored sampling [4], but here reduced in practice by the additional costs from the line search evaluations which also impede parallelization.

The optimal step-size in ES with multi-recombination increases with the parent population size. Learning rate adaptation removes this dependency in our experiments and the optimal step-size is invariably smaller. As a larger step-size is one import advantage of choosing larger population sizes[3], this effect seems to seriously diminish the applicability of learning rate adaptation in ES with (intermediate) multi-recombination in practice.

[3] A larger step-size causes a more global search behavior and is an important aspect for the improved performance on multimodal functions and, arguably, on noisy functions. For $\mu \ll n$, the step-size is proportional to μ [3, 14]. In Figure 2, we observe only a factor of $\mu^{0.75}$ for μ between 4 and 16 in dimension 100.

REFERENCES

[1] Youhei Akimoto, Anne Auger, Tobias Glasmachers, and Daiki Morinaga. Global linear convergence of evolution strategies on more than smooth strongly convex functions. *SIAM Journal on Optimization*, 2022 (accepted).

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. Quality gain analysis of the weighted recombination evolution strategy on general convex quadratic functions. *Theoretical Computer Science*, 832:42–67, 2020.

[3] Dirk V Arnold. Weighted multirecombination evolution strategies. *Theoretical computer science*, 361(1):18–37, 2006.

[4] Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Mirrored sampling in evolution strategies with weighted recombination. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 861–868, 2011.

[5] Anne Auger and Nikolaus Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In *Proceedings of the 8th annual conference on genetic and evolutionary computation GECCO*, pages 445–452. ACM, 2006.

[6] Anne Auger and Nikolaus Hansen. Theory of evolution strategies: a new perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, chapter 10, pages 289–325. World Scientific Publishing, 2011.

[7] Anne Auger and Nikolaus Hansen. Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the (1+1) ES with generalized one-fifth success rule. *arXiv:1310.8397 [cs.NA]*, 2013.

[8] Anne Auger and Nikolaus Hansen. Linear convergence of comparison-based step-size adaptive randomized search via stability of markov chains. *SIAM Journal on Optimization*, 26(3):1589–1624, 2016.

[9] Hans-Georg Beyer. Mutate large, but inherit small! On the analysis of rescaled mutations in $(\tilde{1}, \tilde{\lambda})$-ES with noisy fitness data. In *International Conference on Parallel Problem Solving from Nature*, pages 109–118. Springer, 1998.

[10] Armand Gissler, Anne Auger, and Nikolaus Hansen. Supplementary material for Learning rate adaptation by line search in evolution strategies with recombination. *hal-03626292*, 2022.

[11] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

[12] Jens Jägersküpper. Lower bounds for hit-and-run direct search. In Juraj Hromkovič, Richard Královič, Marc Nunkesser, and Peter Widmayer, editors, *Stochastic Algorithms: Foundations and Applications*, volume 4665, pages 118–129, Berlin, Heidelberg, 2007. Springer.

[13] Mohamed Jebalia and Anne Auger. Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-ES and optimal μ for intermediate recombination for large population sizes. In *International Conference on Parallel Problem Solving from Nature*, pages 52–62. Springer, 2010.

[14] Ingo Rechenberg. *Evolutionsstrategie'94*, volume 1581. Frommann–Holzboog–Verlag, Stuttgart (Germany), 1994.

[15] Sebastian U Stich, Christian L Muller, and Bernd Gartner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013.

[16] Cheikh Touré, Anne Auger, and Nikolaus Hansen. Global linear convergence of evolution strategies with recombination on scaling-invariant functions. *Journal of Global Optimization*, 2022 (under revision).

[17] Cheikh Touré, Armand Gissler, Anne Auger, and Nikolaus Hansen. Scaling-invariant functions versus positively homogeneous functions. *Journal of Optimization Theory and Applications*, 191:363–383, 2021.