Analyzing the Impact of Mirrored Sampling and Sequential Selection in Elitist Evolution Strategies

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

Dimo Brockhoff
System Modeling and Optimization Team
Laboratoire d'Informatique
École Polytechnique
91128 Palaiseau Cedex, France
brockho@lix.polytechnique.fr

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

ABSTRACT

This paper presents a refined single parent evolution strategy that is derandomized with mirrored sampling and/or uses sequential selection. The paper analyzes some of the elitist variants of this algorithm. We prove, on spherical functions with finite dimension, linear convergence of different strategies with scale-invariant step-size and provide expressions for the convergence rates as the expectation of some known random variables. In addition, we derive explicit asymptotic formulae for the convergence rate when the dimension of the search space goes to infinity. Convergence rates on the sphere reveal lower bounds for the convergence rate of the respective step-size adaptive strategies. We prove the surprising result that the (1+2)-ES with mirrored sampling converges at the same rate as the (1+1)-ES without and show that the tight lower bound for the (1+λ)-ES with mirrored sampling and sequential selection improves by 16% over the (1+1)-ES reaching an asymptotic value of about -0.235.

Categories and Subject Descriptors

G.1.6 [**Numerical Analysis**]: Optimization—*global optimization, unconstrained optimization*; F.2.1 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems

General Terms

Algorithms, Theory

Keywords

Evolution Strategies, Mirroring, Sequential Selection, Plus-Selection

1. INTRODUCTION

Evolution Strategies (ESs) are robust search algorithms designed to minimize objective functions f that map a continuous search space $\mathbb{R}^d$ into $\mathbb{R}$. In a $(1 \dot{+} \lambda)$-ES, λ candidate solutions, the offspring, are created from a single parent, $\boldsymbol{X}_k \in \mathbb{R}^d$. The λ offspring are generated by adding λ *independent* random vectors $(\boldsymbol{\mathcal{N}}_k^i)_{1\leq i\leq\lambda}$ to $\boldsymbol{X}_k$. Then, the *best* of the λ offspring $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^i$ in case of comma selection or of the λ offspring plus parent in case of plus selection is *selected* to become the next parent $\boldsymbol{X}_{k+1}$. The (1+1)-ES is arguably the most local, and the locally fastest, variant of an evolution strategy.

Derandomization of random numbers is a general technique where the independent samples are replaced by dependent ones with the objective of accelerating algorithm convergence. Derandomization by means of antithetic variables for isotropic samples was first introduced within general ESs in [27]. Mirrored samples, as used in this paper, are a special case, where the number of independent events is reduced by a factor of two only. In [26], the sequence of uniform random numbers used for sampling a multivariate normal distribution was replaced by scrambling-Halton and Sobol sequences. These sequences achieved consistent improvements of CMA-ES (covariance matrix adaptation evolution strategy) mainly on unimodal test functions, typically with $\leq 30\%$ speed-up and most pronounced in dimension 2. The improvements are however difficult to attribute to a cause for at least two reasons. First, in CMA-ES with $\mu > 1$, quasi-random numbers possibly introduce a (strong) bias on the step-size. For mirrored samples and Sobol sequences, we have verified this bias empirically (shown for mirrored sampling in [17]). The bias can improve convergence rates,[1] but violates the demand on a stochastic search algorithm to be unbiased [19, 22]. Second, random rotations of the quasi-random vector sets in [26] lead to a significant loss of the advantage. The investigated functions were however unrotated. This makes the identity as initial covariance matrix, represented in the given coordinate system and in connection with the quasi-random numbers, presumably a choice that is unintentionally biased towards the function testbed.

Consequently, it remains to be investigated to what extend the improvements can be attributed to a bias on the variance of the sum of selected vectors (leading to the bias

[1]For mirrored sampling this most probably happens if random vectors with different lengths are realized, which is the case in particular in small dimensions.

on the step-size), to a coordinate system dependency, or to the quasi-random structure itself.

Our own experiments with derandomizations beyond mirroring, similar to those in [27], revealed the most pronounced effects (unsurprisingly) by mirroring and in small populations. We have not considered algorithms that are—by themselves or in combination with CMA-ES—biased or not rotationally invariant.

Mirrored sampling is a derandomization technique similar to antithetic variables that was recently introduced within $(1+\lambda)$ and $(1,\lambda)$-ESs [17]. In addition, mirrored sampling has been coupled with *sequential selection*, a modification of the $(1,\lambda)$ and $(1+\lambda)$ selection schemes where the offspring are evaluated sequentially and the iteration is concluded as soon as one offspring is better than its current parent [17].

Sequential selection and mirrored sampling have been implemented within the CMA-ES and extensively empirically studied on 54 noiseless [20] and noisy [21] functions in a series of papers [4–10, 12–15]. In summary, the variants with mirrored mutation and sequential selection improved their baseline algorithms (without these two ideas) on almost all functions for almost all target values where the combination of the two concepts was never statistically significantly worse than the standard algorithms. In particular for the elitist (1+1)-CMA-ES, additional mirrored mutation and sequential selection improved the performance by about 17% on the non-separable ellipsoid function, by about 20% on the ellipsoid, the discus, and the sum of different powers functions, and by 12% on the sphere function while no statistically significant worsening of the performance was reported [6].

So far, theoretical investigations of mirrored sampling and sequential selection is restricted to comma selection [17]. Convergence rates of the scale-invariant step-size $(1,\lambda)$-ES with mirrored sampling and sequential selection on spherical functions have been derived and lower bounds for the convergence of the different strategies were compared. Those results hold for finite dimensions of the search space. In this paper, we aim at generalizing those theoretical results to plus selection: we extend finite dimension convergence proofs to plus selection and complement those results with asymptotic estimates of the convergence rates when the dimension goes to infinity.

The paper is structured as follows. In Section 2, we describe the $(1\,\dot{+}\,\lambda)$-ES with mirrored sampling and sequential selection and derive general properties. In Section 3, we derive the linear convergence of the $(1+\lambda)$-ES with mirrored sampling and sequential selection with scale-invariant step-size on spherical functions. We express the convergence rate in terms of the expectation of a random variable. In addition, we establish that the (1+1)-ES and the $(1+2_{\mathrm{m}})$-ES with mirrored sampling exhibit the same convergence rate. In Section 5, we derive some simple expressions for the asymptotic normalized convergence rate of the different algorithms, where asymptotic refers to the dimension tending to infinity. In Section 6, we numerically simulate the convergence rates for different dimensions and appraise quantitatively the improvements brought by mirrored sampling and sequential selection.

Notations: In this paper, a multivariate normal distribution with mean vector zero and covariance matrix identity will be called *standard* multivariate normal distribution. The first vector $(1,0,\ldots,0)$ of the canonical basis will be denoted $\boldsymbol{e}_1$.

Algorithm 1 Pseudocode for the $(1+\lambda)$-ES and the $(1,\lambda)$-ES with all combinations with/without mirrored sampling and/or sequential selection. $\boldsymbol{X}_k \in \mathbb{R}^d$ denotes the current search point and σ_k the current step-size at iteration k. $(\boldsymbol{\mathcal{N}}_m)_{m\in\mathbb{N}}$ is a sequence of random vectors. In this paper, *skip mirror* is true whenever sequential selection is true.

given: $f : \mathbb{R}^d \to \mathbb{R}$, $\boldsymbol{X}_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, $\lambda \in \mathbb{N}^+$, $(\boldsymbol{\mathcal{N}}_m)_{m\in\mathbb{N}}$

```
m ← 0    number of random samples used
j ← 0    use previous sample if j is even
k ← 0    iteration counter for notational consistency
while stopping criterion not fulfilled do
  i ← 0    offspring counter
  while i < λ do
    i ← i + 1, j ← j + 1
    if mirrored sampling and j ≡ 0 (mod 2) then
      X_k^i = X_k − σ_k N_m    use previous sample
    else
      m ← m + 1
      X_k^i = X_k + σ_k N_m
    if f(X_k^i) ≤ f(X_k) then
      if skip mirror then
        j ← 0    continue with a fresh sample
      if sequential selection then
        break
  end while
  if plus selection then
    X_{k+1} = argmin{f(X_k), f(X_k^1), ..., f(X_k^i)}
  else
    X_{k+1} = argmin{f(X_k^1), ..., f(X_k^i)}
  σ_{k+1} = update(σ_k)
  k ← k + 1    iteration counter
end while
```

2. $(1\,\dot{+}\,\lambda)$-ES WITH MIRRORED SAMPLING AND SEQUENTIAL SELECTION

In this section, we introduce the $(1\,\dot{+}\,\lambda)$-ES with mirrored sampling and sequential selection and derive general theoretical results on those algorithms.

2.1 Algorithm Description

Mirrored mutations and sequential selection have been introduced in [17] and are two independent ideas for improving simple local search strategies such as $(1\,\dot{+}\,\lambda)$-ESs. Algorithm 1 shows the pseudocode of a combination of both concepts within the $(1+\lambda)$-ES and the $(1,\lambda)$-ES. Note that we describe the algorithms without specifying which sampling distribution is used—though most of the time, for Evolution Strategies, multivariate normal distributions are used. However, since we will derive some results that are independent of the choice of the sampling distribution, we keep the description general and indicate when a standard multivariate normal distribution is required.

Mirrored sampling: The idea behind mirrored sampling is to derandomize the generation of new sample points. Instead of using two independent random vectors to create two offspring, with mirrored sampling only a single random vector instantiation is used to create two offspring: one by adding and the other by subtracting the vector from the current search point. The two instantiations are called *mirrored* or *symmetric* with respect to the parent $\boldsymbol{X}_k$ at itera-

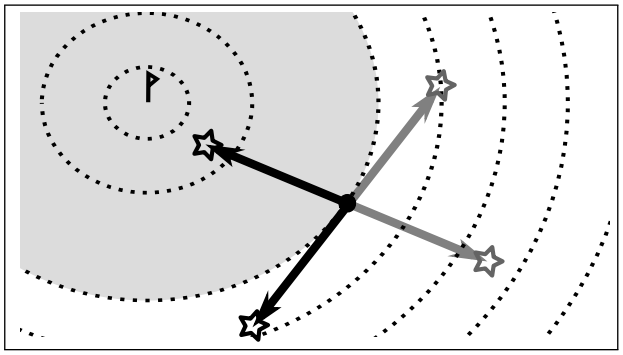

Figure 1: Mirrored sampling on an objective function with convex sub-level sets. The shaded region represents the set of solutions with a better objective function value than the parent solution (black dot). *Not both mirrored offspring can be better than the parent solution at the same time*: shown are two examples of an offspring (black) and its mirrored version (gray) where either one or both offspring are worse than the parent.

tion k if they take place in the same iteration. For odd λ, every other iteration, the first offspring uses the mirrored last vector from the previous iteration, see j in Algorithm 1. Consequently, in the (1+1$_\mathrm{m}$)-ES, a mirrored sample is used if and only if the iteration index is even (given *skip mirror* in Algorithm 1 is false).

When evaluating a sampled solution and its mirrored counterpart, sometimes unnecessary function evaluations are performed: for example, on unimodal objective functions with convex sub-level sets, $\{x \mid f(x) \leq c\}$ for $c \in \mathbb{R}$, such as the sphere function, $f(x) = \|x\|^2$, the mirrored solution $\boldsymbol{X}_k - \boldsymbol{\mathcal{N}}$ is always worse than the parent $\boldsymbol{X}_k$ if $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}$ was better than $\boldsymbol{X}_k$, see Fig. 1 and Proposition 2 below. Setting *skip mirror* to true in Algorithm 1 prevents these mirrored samples from being realized.[2]

Note that in the (1+λ_m)-ES, two mirrored offspring are entirely dependent and, in a sense, complementary, similar to antithetic variables for Monte-Carlo numerical integration [18]. Mirrored sampling is also similar to using a symmetric difference quotient instead of the standard one-sided difference quotient. The technique has been applied to Evolutionary Gradient Search (EGS) with good success [1].

Sequential selection: In sequential selection, the offspring are evaluated one by one, compared to their parent, and the iteration is concluded immediately if one offspring is better than its parent. Sequential selection has been introduced in the context of comma selection, where it aims to combine the robustness advantage of comma selection with the speed advantage of elitist plus selection [17]. Sequential selection and mirrored sampling are independent of each other and can be employed separately within ESs, see Algorithm 1. We will see that sequential selection, in the elitist context, essentially comes down to the (1+1)-ES. The (1+λ)-ES variant employing sequential selection is denoted by (1+λ^s)-ES.

Combining mirrored sampling and sequential selection: Sequential selection has been combined with mirrored sampling in the $(1, \lambda)$-ES and, although not explicitly mentioned, *skip mirror* was in this case always applied [17]. Both concepts complement each other well for the $(1, \lambda)$-ES. Sequential selection tends to reduce the realized population size to a minimum and mirrored sampling improves the performance in particular for very small population sizes. In this paper as well, *skip mirror* is set to true when sequential selection and mirrored sampling are combined. With sequential selection, it is important that independent and mirrored offspring are evaluated alternately in order to profit immediately from the increased probability of the mirrored offspring being better after the independently drawn offspring was worse than the parent [17].

The (1+λ)-ES with mirrored sampling and sequential selection is denoted as (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES where the superscript refers to sequential selection and the subscript to mirrored sampling with *skip mirror* set to true. All results in this paper refer to strategies where *skip mirror* is true when sequential selection is applied.

[2] In the (1+1$_\mathrm{m}$)-ES, these unnecessary mirrored solutions fall closely together with the previous parental solution.

2.2 General Properties of Mirrored Sampling and Sequential Selection

In this section, we derive general results on evolution strategies using mirrored sampling and sequential selection. Let us first recognize that the (1+λ^s)-ES is essentially a (1+1)-ES with smaller iteration counter. In both strategies, the parent is updated if and only if the currently sampled offspring is better (but see also Remark 1). Now, we also establish for mirrored sampling that (1+1$_\mathrm{ms}$)-ES and (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES all evaluate the same points, provided they use the same (constant or scale-invariant) step-size and the same random instance for generating the offspring.

PROPOSITION 1. *The (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES is for any $\lambda \geq 1$ the same algorithm—with possibly different iteration counter, given the same random vectors and the same step-sizes are used (for example the step-size σ_k is either constant or scale-invariant, i.e., $\sigma_k = \sigma\|X_k\|$ with σ constant).*

PROOF. We prove that the state of the algorithm (apart from the iteration counter) does not depend on λ. Independently of λ, because of sequential selection applied in combination with plus selection, any new evaluated offspring is sampled from the *best ever* evaluated point so far. Since step-size only depends on the parent or is constant, the same offspring will be sampled provided the random samples used are also independent of λ. However, the samples used are taken one by one from $(\boldsymbol{\mathcal{N}}_m, -\boldsymbol{\mathcal{N}}_m)_{m\in\mathbb{N}}$ where because *skip mirror* is true, some mirrored vectors $-\boldsymbol{\mathcal{N}}$ are skipped. But the decision of whether or not to skip the mirroring of a sample is also independent of λ since it only depends on a comparison between the single last offspring and the parent. □

Due to this result, the notations (1+1$_\mathrm{ms}$)-ES, (1+2$^\mathrm{s}_\mathrm{ms}$)-ES and (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES refer, in this paper, all to the same strategy. However, this might not be the case in practice.

REMARK 1. *In practice, the behavior of ESs with sequential selection depends on λ, because the step-size is typically updated at the end of each iteration and therefore more often with small λ.*

REMARK 2. *For $\mu = 1$, sequential selection and/or mirroring have been combined with CMA-ES and extensively studied with plus and comma selection and different step-size rules [4–10, 12–15].*

We derive now some results on objective functions with convex sub-level sets. We first establish that on objective functions with convex sub-level sets two mirrored offspring cannot be both better than their parent (see also Fig. 1).

PROPOSITION 2. *Let f be an objective function with convex sub-level sets, then two mirrored offspring cannot be simultaneously strictly better than their parent.*

PROOF. Considering the convex sub-level set, given by the parent solution, and the tangent hyperplane in this solution, the two mirrored offspring can never lie on the same side of the tangent hyperplane, see Fig. 1. At the same time, due to the convexity of the sub-level set and the definition of the tangent hyperplane, all solutions that are better than the parent solution lie on the same side of the tangent hyperplane such that not both mirrored offspring can have better objective function values at the same time. □

A consequence of Proposition 2 is that on objective functions with convex sub-level sets, sequential selection applied with two mirrored offspring has no effect on the sequence of *accepted* solutions. In this case, sequential selection combined with skip mirror only reduces the number of evaluated solutions.

COROLLARY 3 (IDENTICAL TRACE FOR $\lambda = 2$). *On objective functions with convex sub-level sets, the $(1+2_{\mathrm{m}})$-ES and the $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES deliver the same sequence of parental solutions (given they use the same random vectors and step-sizes). The same holds for the $(1, 2_{\mathrm{m}})$-ES and the $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$-ES.*

PROOF. We consider the iteration step k and assume that $m = k$ at the beginning of the inner while loop. According to Proposition 2, it can never happen that both offspring are better than the parent and only two remaining cases need to be investigated. (i) In case of both offspring being worse than the parent, both plus-selection algorithms will keep the parent whereas the better of the two offspring is taken as the new parent by both comma-strategy algorithms. (ii) In case that one of the two offspring is not worse than the parent, the other must be worse and all algorithms will take the better offspring as their new parent solution—there is only a difference between the algorithms if the first offspring is not worse than the parent. Only in this case, the algorithms with sequential selection will directly accept the first offspring as next parent while the other variants evaluate unnecessarily the second (worse) offspring as well. Since either both offspring are evaluated or the sample associated to the non-evaluated offspring is skipped, in the next iteration, a fresh sample $\boldsymbol{\mathcal{N}}_{m+1}$ will be used for the first offspring thus $m = k+1$ at the beginning of the next inner while loop. □

Because sequential selection evaluates sometimes only one solution per iteration, the corollary implies that on functions with convex sub-level sets, the $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES (or $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$-ES) will converge faster than the $(1+2_{\mathrm{m}})$-ES (or $(1, 2_{\mathrm{m}})$-ES) in case of convergence and diverge faster in case of divergence.

We can additionally establish that for all strategies with two offspring and sequential selection, the number of offspring evaluated per iteration is the same, independent of mirroring and elitism:

LEMMA 4. *Assume the $(1+2^{\mathrm{s}})$, $(1, 2^{\mathrm{s}})$, $(1+2^{\mathrm{s}}_{\mathrm{ms}})$, $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$-ESs start at iteration k from the same parent $\boldsymbol{X}_k$, sample the same first offspring $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}$, and optimize the same objective function. Then the number of evaluated offspring at iteration k will be the same for all strategies.*

PROOF. In all the cases, the number of evaluated offspring will be 1 if $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}$ is not worse than $\boldsymbol{X}_k$ and 2 otherwise. □

Finally, we find that for λ to infinity, comma strategies using sequential selection without or with mirroring converge to the (1+1)-ES or the $(1+1_{\mathrm{ms}})$-ES, respectively:

LEMMA 5 (EQUIVALENCE OF $(1, \infty^{\mathrm{s}})$-ES AND (1+1)-ES). *Using scale-invariant or constant step-size, the $(1, \infty^{\mathrm{s}})$-ES is equivalent to the (1+1)-ES and the $(1, \infty^{\mathrm{s}}_{\mathrm{ms}})$-ES is equivalent with the $(1+1_{\mathrm{ms}})$-ES.*

PROOF. The proof follows directly from the algorithm descriptions, similar to the proof of Proposition 1. □

3. LINEAR CONVERGENCE AND LOWER BOUNDS

Evolution Strategies are rank-based search algorithms and as such cannot exhibit a faster convergence than linear [25]. We here define linear convergence as the logarithm of the distance to the optimum decreasing linearly with the increasing number of function evaluations. An example of linear convergence is illustrated in Fig. 2 for three different instances of the (1+1)-ES with scale-invariant step-size. Formally, for the (1+1)-ES, let $\boldsymbol{X}_k$ be the estimate of the solution at iteration k. Almost sure (a.s.) linear convergence takes place if there exists a constant $c \neq 0$, such that

$$\frac{1}{k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \to c \ \ a.s. \tag{1}$$

Literally, *convergence* of $\boldsymbol{X}_k$ takes place only if $c < 0$ and for $c > 0$ the term *divergence* is more appropriate. If the above expression goes to zero, the strategy might still converge sub-linearly [16]. In this paper, we analyze algorithms that do not have a constant number of function evaluations per iteration and we will use the following generalization of (1) that accounts for the actual number of function evaluations: let T_k be the number of function evaluations performed until iteration k. Almost sure linear convergence takes place if there exists a constant $c \neq 0$, such that

$$\frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \to c \ \ a.s. \tag{2}$$

For the (1+1)-ES both equations are equivalent and for the $(1 \dagger \lambda)$-ES we have $T_k = k\lambda$. The constant c is called the *convergence rate* and it corresponds to the slope of the curves in Fig. 2. The dynamics and thus the convergence rate of a step-size adaptive ES obviously depends on the step-size rule. The fastest convergence rates for adaptive step-size ESs are in general reached for a specific step-size rule in the so-called *scale-invariant step-size* ES where the step-size σ_k at time k is proportional to the distance to the optimum. Assuming the optimum w.l.o.g. in 0, the scale-invariant step-size is $\sigma_k = \sigma\|\boldsymbol{X}_k\|$ for $\sigma > 0$ on spherical functions $g(\|x\|)$ for $g \in \mathcal{M}$ where $\mathcal{M}$ denotes the set of functions $g : \mathbb{R} \mapsto \mathbb{R}$ that are strictly increasing [23]. ESs with scale-invariant step-sizes are artificial algorithms as they use the distance to the optimum to adapt the step-size. However, they are interesting to study as (1) they are a realistic approximation of step-size adaptive isotropic ESs where $\|\boldsymbol{X}_k\|/\sigma_k$ is

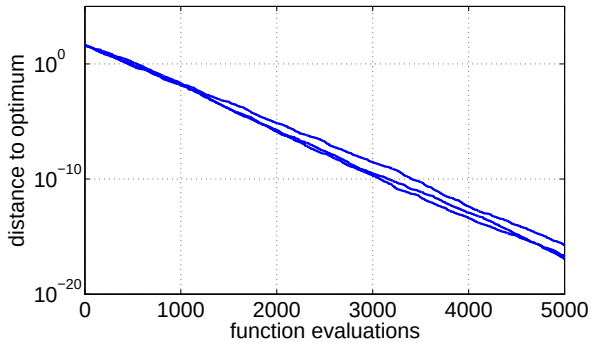

Figure 2: Distance to optimum versus number of function evaluations for three different instances of a (1+1)-ES minimizing the sphere function $g(\|x\|), g \in \mathcal{M}$ and with scale-invariant step-size $\sigma_k = \sigma\|\boldsymbol{X}_k\|$ for $d = 20$ and $\sigma = 0.6/d$. Linear decrease is observed, the convergence rate corresponds to the slope of the curves.

usually a stable Markov Chain, here modeled as a constant, and (2) they achieve, for the right choice of the constant, optimal convergence rates. In addition, the simplification of $\|\boldsymbol{X}_k\|/\sigma_k$ being a constant induces in general much simpler theoretical analysis. We now state formally the linear convergence of a (1+1)-ES with scale-invariant step-size and give an implicit expression for the convergence rate:

THEOREM 1 (LINEAR CONVERGENCE OF (1+1)-ES [23]). *The (1+1)-ES with scale-invariant step-size ($\sigma_k = \sigma\|\boldsymbol{X}_k\|$) converges linearly on the class of spherical functions $g(\|x\|)$, $g \in \mathcal{M}$, and*

$$\lim_{k\to\infty} \frac{1}{k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \mathrm{CR}_{(1+1)}(\sigma), \tag{3}$$

with

$$\mathrm{CR}_{(1+1)}(\sigma) = -\frac{1}{2} E\big[\ln^- \big(1 + \underbrace{2\sigma[\boldsymbol{\mathcal{N}}]_1}_{\text{gain if negative}} + \underbrace{\sigma^2\|\boldsymbol{\mathcal{N}}\|^2}_{\text{loss}}\big)\big] \ ,$$

where $\ln^-$ is the negative part of the function $\ln$, i.e., $\ln^-(x) = -\min(\ln(x), 0)$, $\boldsymbol{\mathcal{N}}$ is a standard multivariate normal distribution and $[\boldsymbol{\mathcal{N}}]_1$ is the projection of $\boldsymbol{\mathcal{N}}$ onto the first coordinate $\boldsymbol{e}_1$.

The function $\sigma \mapsto \mathrm{CR}_{(1+1)}(\sigma)$ gives for each $\sigma > 0$ the convergence rate of the (1+1)-ES with step-size $\sigma_k = \sigma\|\boldsymbol{X}_k\|$. The function has been studied in [23] and is plotted in Fig 4 (left) for different dimensions. The minimum of $\sigma \mapsto \mathrm{CR}_{(1+1)}(\sigma)$ gives for a given dimension the lower bound for the convergence rate of (1+1)-ES with offspring sampled with a standard multivariate normal distribution and *any step-size adaptation mechanism* on *any objective function* as formally stated now:

THEOREM 2 (LOWER BOUND FOR (1+1)-ES [23]). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a measurable objective function and $x^* \in \mathbb{R}^d$. Assume that at each iteration k, the standard multivariate normal distribution used to sample the offspring is independent of σ_k and $\boldsymbol{X}_k$ and that $E[|\ln\|\boldsymbol{X}_0 - x^*\|\|] < \infty$, then the convergence of the step-size adaptive (1+1)-ES is at most linear and*

$$\inf_{k\in\mathbb{N}} E[\ln\|\boldsymbol{X}_k - x^*\|/\|\boldsymbol{X}_0 - x^*\|] \geq \inf_{\sigma} \mathrm{CR}_{(1+1)}(\sigma) \ .$$

Objective for the rest of the paper: In the rest of the paper, we investigate the linear convergence of mirrored and sequential variants of the (1+λ)-ES with scale-invariant step-size. As for the (1+1)-ES, the minimum of the convergence rate in σ will represent lower bounds for the convergence rate of step-size adaptive methods with a standard multivariate normal sampling on any objective function. Before tackling the linear convergence of the different variants, we explain the main proof idea behind the linear convergence proofs.

How to prove linear convergence of scale-invariant step-size ESs? We sketch the proof idea in the case of the (1+1)-ES and we will explain in the core of the paper how to generalize this proof in particular for the case of a non-constant number of evaluation per iteration. The first step of the proof expresses the left-hand side (LHS) of (1) as a sum of k terms exploiting standard properties of the logarithm function:

$$\frac{1}{k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{1}{k} \sum_{i=0}^{k-1} \ln \frac{\|\boldsymbol{X}_{i+1}\|}{\|\boldsymbol{X}_i\|} \ . \tag{4}$$

We then exploit the isotropy of the sphere function, the isotropy of the standard multivariate normal distribution and the scale-invariant step-size rule to prove that all terms $\ln(\|\boldsymbol{X}_{i+1}\|/\|\boldsymbol{X}_i\|)$ are independent identically distributed (i.i.d.). A law of large numbers[3] (LLN) therefore implies that the right-hand side (RHS) of (4) converges when k goes to infinity to $E[\ln(\|\boldsymbol{X}_{i+1}\|/\|\boldsymbol{X}_i\|)]$ almost surely.

4. CONVERGENCE RATE ON SPHERICAL FUNCTIONS IN FINITE DIMENSION

In this section, we analyze the linear convergence of the (1+2$_\mathrm{m}$)-ES and the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES for a fixed dimension d of the search space. Before to establish the main results, we derive some technical results and introduce some useful definitions.

4.1 Preliminary Results and Definitions

We establish first a lemma that simplifies the writing of the acceptance event of mirrored offspring.

LEMMA 6. *Let $\boldsymbol{X}_{\boldsymbol{e}_1} = \boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}$ and $\boldsymbol{X}^\mathrm{m}_{\boldsymbol{e}_1} = \boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}$ be two mirrored offspring sampled from the parent $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$. On spherical functions, the acceptance event $\{\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\| \leq 1\}$ can be written as $\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}$. Similarly, the acceptance event of $\boldsymbol{X}^\mathrm{m}_{\boldsymbol{e}_1}$ satisfies $\{\|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\| \leq 1\} = \{-2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}$.*

PROOF. We remark first that $\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\| \leq 1$ is equivalent to $\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\|^2 \leq 1$. We now develop $\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\|^2$ as $1 + 2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2$ and we immediately obtain that $1 + 2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2 \leq 1$ is equivalent to $2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2 \leq 0$. We proceed similarly for the acceptance event of $\boldsymbol{X}^\mathrm{m}_{\boldsymbol{e}_1}$. □

In the sequel, we will need to use the indicator function of the acceptance events of mirrored offspring sampled from $\boldsymbol{e}_1$. For that reason we define the random variables W_1 and W_1^m in the following way:

DEFINITION 1. *Let $W_1 = 2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$ and $W_1^\mathrm{m} = -2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$.*

[3] Verifying some technical conditions such that the expectation and the variance of $\ln(\|\boldsymbol{X}_{i+1}\|/\|\boldsymbol{X}_i\|)$ are finite.

We can now express the indicator of the acceptance event of $\boldsymbol{X}_{\boldsymbol{e}_1}$ as

$$1_{\{\boldsymbol{X}_{\boldsymbol{e}_1} \text{ is better than } \boldsymbol{e}_1\}} = 1_{\{W_1 \leq 0\}} \ , \tag{5}$$

and the indicator of the acceptance of $\boldsymbol{X}^{\mathrm{m}}_{\boldsymbol{e}_1}$ as

$$1_{\{\boldsymbol{X}^{\mathrm{m}}_{\boldsymbol{e}_1} \text{ is better than } \boldsymbol{e}_1\}} = 1_{\{W_1^{\mathrm{m}} \leq 0\}} \ . \tag{6}$$

Using the expression of W_1 and a straightforward derivation, we find the following alternative expression for the convergence rate of the (1+1)-ES:

$$\mathrm{CR}_{(1+1)}(\sigma) = \frac{1}{2} E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right] \ . \tag{7}$$

We now establish two technical lemmas that will be useful to prove the equality of the convergence rate of the (1+1)-ES and the (1+2$_{\mathrm{m}}$)-ES.

LEMMA 7. *Let $\boldsymbol{\mathcal{N}}$ be a standard multivariate normal distribution, the following equality holds*

$$\begin{aligned} &E\left[\ln(1 + (2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2) 1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}}\right] = \\ &E\left[\ln(1 + (-2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2) 1_{\{-2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}}\right] \end{aligned} \tag{8}$$

or, using the notations W_1 and W_1^{m}

$$E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right] = E\left[\ln(1 + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}})\right] \tag{9}$$

PROOF. Since $\boldsymbol{\mathcal{N}}$ is a standard multivariate normal distribution, $-\boldsymbol{\mathcal{N}}$ follows the same distribution as $\boldsymbol{\mathcal{N}}$ and thus (8) follows. □

LEMMA 8. *The following equation holds*

$$\begin{aligned} &E\left[\ln\left(1 + \sigma W_1 1_{\{W_1 \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}}\right)\right] \\ &\qquad = 2E\left[\ln\left(1 + \sigma W_1 1_{\{W_1 \leq 0\}}\right)\right] \end{aligned} \tag{10}$$

where $W_1 = 2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$ and $W_1^{\mathrm{m}} = -2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$ with $\boldsymbol{\mathcal{N}}$ a random vector following a standard multivariate normal distribution.

PROOF. According to Proposition 2, two mirrored offspring cannot be simultaneously better than their parent on the sphere function. Since $\{W_1 \leq 0\}$ and $\{W_1^{\mathrm{m}} \leq 0\}$ are the acceptance events of mirrored offspring started from $\boldsymbol{e}_1$ on the sphere function ((5) and (6)), we know that they are incompatible such that $1_{\{W_1 \leq 0\}}$ and $1_{\{W_1^{\mathrm{m}} \leq 0\}}$ are not simultaneously equal to 1. Consequently

$$\begin{aligned} &\ln\left(1 + \sigma W_1 1_{\{W_1 \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}}\right) = \\ &\qquad \ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}}) + \ln(1 + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}}) \ . \end{aligned}$$

Using the linearity of the expectation, we obtain that

$$\begin{aligned} &E\left[\ln\left(1 + \sigma W_1 1_{\{W_1 \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}}\right)\right] = \\ &E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right] + E\left[\ln(1 + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}})\right] \ . \end{aligned}$$

We now use Lemma 7 and obtain that the RHS of the last equation equals $2E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right]$. Hence the result. □

4.2 Convergence Rate for the $(1 + 2_{\mathrm{m}})$-ES

In this section, we prove the linear convergence of the (1+2$_{\mathrm{m}}$)-ES with scale-invariant step-size and prove the surprising result that the convergence rate of the (1+2$_{\mathrm{m}}$)-ES equals the convergence rate of the (1+1)-ES. In a (1+2$_{\mathrm{m}}$)-ES with scale-invariant step-size, two mirrored offspring $\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}$ and $\boldsymbol{X}_k - \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}$ are sampled from the parent $\boldsymbol{X}_k$ where $\boldsymbol{\mathcal{N}}$ is a standard multivariate normal distribution independent of $\boldsymbol{X}_k$ and of the past (we omit the dependence in k for the sampled vectors for the sake of readability). Since on the sphere function, the offspring cannot be simultaneously better than their parent (see Proposition 2), the update equation for $\|\boldsymbol{X}_k\|$ reads:

$$\begin{aligned} \|\boldsymbol{X}_{k+1}\| = &\|\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| \times 1_{\{\|\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| \leq \|\boldsymbol{X}_k\|\}} + \\ &\|\boldsymbol{X}_k - \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| \times 1_{\{\|\boldsymbol{X}_k - \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| \leq \|\boldsymbol{X}_k\|\}} + \\ &\|\boldsymbol{X}_k\| \times 1_{\{\|\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| > \|\boldsymbol{X}_k\|, \|\boldsymbol{X}_k - \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}\| > \|\boldsymbol{X}_k\|\}} \ . \end{aligned} \tag{11}$$

Before to prove the linear convergence of the (1+2$_{\mathrm{m}}$)-ES with scale-invariant step-size, we need to establish the following lemma:

LEMMA 9. *Let Z_k be the sequence of random variables*

$$\begin{aligned} Z_k = \frac{1}{2} \ln \big[&\|\boldsymbol{Y}_k + \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{Y}_k + \sigma\boldsymbol{\mathcal{N}}\| \leq 1\}} + \\ &\|\boldsymbol{Y}_k - \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{Y}_k - \sigma\boldsymbol{\mathcal{N}}\| \leq 1\}} + 1_{\{\|\boldsymbol{Y}_k + \sigma\boldsymbol{\mathcal{N}}\| > 1, \|\boldsymbol{Y}_k - \sigma\boldsymbol{\mathcal{N}}\| > 1\}}\big] \end{aligned}$$

where $\boldsymbol{Y}_k = \boldsymbol{X}_k / \|\boldsymbol{X}_k\|$ with $\boldsymbol{X}_k$ defined with (11). Then Z_k are independent identically distributed as

$$Z^{(1+2_{\mathrm{m}})} = \frac{1}{2} \ln\left[1 + \sigma W_1 1_{\{W_1 \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}}\right] \ .$$

Moreover $E[|Z^{(1+2_{\mathrm{m}})}|] < \infty$.

PROOF. Because of the isotropy of the distribution of $\boldsymbol{\mathcal{N}}$ and of the sphere function, in distribution Z_k equals

$$\begin{aligned} Z_k \stackrel{d}{=} \frac{1}{2} \ln \big[&\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\| \leq 1\}} + \\ &\|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\| > 1\}} + 1_{\{\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\| > 1, \|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\| > 1\}}\big] \end{aligned} \tag{12}$$

where we have replaced $\boldsymbol{Y}_k$ by $\boldsymbol{e}_1$. The independence of Z_k comes from the fact that $\boldsymbol{\mathcal{N}}$ is independent of $\boldsymbol{Y}_k$ and from the isotropy of the sphere. The detailed proof of those two points is rather technical and we refer to [11, Lemma 1 and Lemma 2] to see how to have a fully formal proof. We are now going to simplify the following term

$$\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{W_1 \leq 0\}} + \|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{W_1^{\mathrm{m}} \leq 0\}} + 1_{\{W_1 > 0, W_1^{\mathrm{m}} > 0\}}$$

that comes into play in the RHS of (12). Developing $\|\boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}\|^2$ as $1 + 2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2$ and $\|\boldsymbol{e}_1 - \sigma\boldsymbol{\mathcal{N}}\|^2$ as $1 - 2\sigma[\boldsymbol{\mathcal{N}}]_1 + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2$, we can simplify the previous equation into

$$\begin{aligned} &1_{\{W_1 \leq 0\}} + \sigma W_1 1_{\{W_1 \leq 0\}} + 1_{\{W_1^{\mathrm{m}} \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}} \\ &\qquad + 1_{\{W_1 > 0, W_1^{\mathrm{m}} > 0\}} \ . \end{aligned}$$

Since $1_{\{W_1 \leq 0\}} + 1_{\{W_1^{\mathrm{m}} \leq 0\}} + 1_{\{W_1 > 0, W_1^{\mathrm{m}} > 0\}} = 1$ we can simplify the previous equation into

$$1 + \sigma W_1 1_{\{W_1 \leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}} \ .$$

Injecting this in (12), we obtain the result. The proof of the fact that $E[|Z^{(1+2_{\mathrm{m}})}|] < \infty$ comes from the proof of the integrability of $\ln[1 + W_1 1_{\{W_1 \leq 0\}}]$ that has been shown in detail in [23]. □

We are now ready to prove the linear convergence of the (1+2$_{\mathrm{m}}$)-ES and express its convergence rate as the expectation

of the random variable $Z^{(1+2_\mathrm{m})}$ introduced in the previous lemma divided by 2.

THEOREM 3. *For the* $(1+2_\mathrm{m})$*-ES with scale-invariant step-size on the class of spherical functions* $g(\|x\|)$, $g \in \mathcal{M}$, *linear convergence holds and*

$$\lim_{k\to\infty} \frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \mathrm{CR}_{(1+2_\mathrm{m})}(\sigma) \tag{13}$$

where

$$\begin{aligned}\mathrm{CR}_{(1+2_\mathrm{m})}(\sigma) &= \frac{1}{2} E[Z^{(1+2_\mathrm{m})}] \\ &= \frac{1}{4} E\left[\ln\left(1 + \sigma W_1 1_{\{W_1\le 0\}} + \sigma W_1^\mathrm{m} 1_{\{W_1^\mathrm{m}\le 0\}}\right)\right] \end{aligned} \tag{14}$$

where $W_1 = 2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$ *and* $W_1^\mathrm{m} = -2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2$ *with* $\boldsymbol{\mathcal{N}}$ *a random vector following a standard multivariate normal distribution.*

PROOF. We start from (11), square it, normalize the equation by $\|\boldsymbol{X}_k\|$ and take the logarithm. We obtain

$$\begin{aligned}\frac{1}{2}\ln\frac{\|\boldsymbol{X}_{k+1}\|^2}{\|\boldsymbol{X}_k\|^2} = \frac{1}{2}\ln\big[&\|\boldsymbol{Y}_k + \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{Y}_k+\sigma\boldsymbol{\mathcal{N}}\|\le 1\}} + \\ &\|\boldsymbol{Y}_k - \sigma\boldsymbol{\mathcal{N}}\|^2 1_{\{\|\boldsymbol{Y}_k-\sigma\boldsymbol{\mathcal{N}}\|\le 1\}} + 1_{\{\|\boldsymbol{Y}_k+\sigma\boldsymbol{\mathcal{N}}\|>1, \|\boldsymbol{Y}_k-\sigma\boldsymbol{\mathcal{N}}\|>1\}}\big]\end{aligned}$$

where $\boldsymbol{Y}_k = \boldsymbol{X}_k/\|\boldsymbol{X}_k\|$. According to Lemma 9, by isotropy of the standard multivariate normal distribution, the random variables in the RHS of the previous equation are independent and identically distributed as

$$Z^{(1+2_\mathrm{m})} = \frac{1}{2}\ln\left[1 + \sigma W_1 1_{\{W_1\le 0\}} + \sigma W_1^\mathrm{m} 1_{\{W_1^\mathrm{m}\le 0\}}\right]$$

In addition, by Lemma 9, $E[|Z^{(1+2_\mathrm{m})}|] < \infty$, and we can thus apply the LLN for independent random variables to

$$\frac{1}{T_k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{1}{2k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{1}{4k}\sum_{i=0}^{k-1}\ln\frac{\|\boldsymbol{X}_{i+1}\|^2}{\|\boldsymbol{X}_i\|^2}$$

and we obtain (14). □

Putting together (7), Lemma 8 and the expression of the convergence rate of the $(1+2_\mathrm{m})$-ES found in the previous theorem, we immediately obtain that the (1+1)-ES and the $(1+2_\mathrm{m})$-ES converge at the same rate. This result is stated in the following corollary.

COROLLARY 10. *On the class of spherical functions, the* (1+1)*-ES and* $(1+2_\mathrm{m})$*-ES with scale-invariant step-size converge at the same convergence rate, i.e.*

$$\mathrm{CR}_{(1+1)}(\sigma) = \mathrm{CR}_{(1+2_\mathrm{m})}(\sigma) \text{ for all } \sigma \ .$$

We close this section with a geometrically based argumentation for the corollary. Consider the tangent hyperplane at the parent location that divides the space into two half spaces and only one of the half spaces contains better solutions. The (1+1)-ES samples isotropically into both half spaces integrating over the entire space. The $(1+2_\mathrm{m})$-ES samples one offspring into one half space and the second one into the other. Together, the offspring integrate over exactly the same region as the single offspring in the (1+1)-ES. The worse offspring is never successful, while the better offspring realizes twice the expected improvement of the offspring in the (1+1)-ES.

4.3 Convergence Rate for the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES

In this section, we analyze the convergence rate of the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES. According to Proposition 1, for all λ, the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES with scale-invariant step-size evaluate the same points in the search space provided they use the same independent random sequence $(\boldsymbol{\mathcal{N}}_m)_{m\in\mathbb{N}}$. Therefore, also the convergence rate of the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES is independent of λ. Note that this is true because we investigate the convergence rate defined as log-progress *per function evaluation* and not per iteration. Though we could think that the easiest algorithm to analyze is the $(1+1_\mathrm{ms})$-ES, we investigate the $(1+2^\mathrm{s}_\mathrm{ms})$ for which iterations are independent—contrary to the $(1+1_\mathrm{ms})$—allowing thus to apply *directly* the LLN for *independent* random variables.

THEOREM 4. *For the* $(1+\lambda^\mathrm{s}_\mathrm{ms})$*-ES with scale-invariant step-size on the class of spherical functions* $g(\|x\|)$, $g \in \mathcal{M}$, *linear convergence holds and for all* λ

$$\lim_{k\to\infty}\frac{1}{T_k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{2}{2-p_s(\sigma)}\,\mathrm{CR}_{(1+1)}(\sigma) \tag{15}$$

where $p_s(\sigma) = \Pr(2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \le 0)$ *is the probability that the offspring* $\boldsymbol{X}_{\boldsymbol{e}_1} = \boldsymbol{e}_1 + \sigma\boldsymbol{\mathcal{N}}$ *is better than its parent* $\boldsymbol{e}_1$ *where* $\boldsymbol{\mathcal{N}}$ *is a standard multivariate normal distribution.*

PROOF. We have seen in Proposition 1 that the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES with scale-invariant step-size evaluates the same points for all λ. Therefore for all λ, the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ESs with scale-invariant step-size have the same convergence rate. Let us analyze the $(1+2^\mathrm{s}_\mathrm{ms})$-ES. Let us write $\frac{1}{T_k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|}$ as $A_k B_k$ with $A_k = k/T_k$ and $B_k = \frac{1}{k}\ln(\|\boldsymbol{X}_k\|/\|\boldsymbol{X}_0\|)$. We are going to handle both terms separately. For B_k, we exploit Corollary 3 where we have seen that, starting from the same parent, the $(1+2_\mathrm{m})$-ES and $(1+2^\mathrm{s}_\mathrm{ms})$-ES have the same parent for the next iteration for objective functions with convex sub-level sets. Thus the sequence of parents $\boldsymbol{X}_k$ is the same for a $(1+2_\mathrm{m})$-ES and a $(1+2^\mathrm{s}_\mathrm{ms})$-ES and thus the expected relative improvement *per iteration* will be the same for both algorithms. By Corollary 10, we have that B_k goes to $2\,\mathrm{CR}_{(1+1)}(\sigma)$ (the factor 2 comes from the normalization by evaluations for the convergence rate of the $(1+2_\mathrm{m})$-ES). For the term A_k, we denote by Λ_i the number of offspring evaluated at iteration i. Then, $T_k = \Lambda_1 + \ldots + \Lambda_k$ and $1/A_k = \frac{1}{k}\sum_{i=1}^k \Lambda_i$. Similarly to [11, Lemma 8], the Λ_k are independent and identically distributed. In addition, according to Lemma 4, the number of evaluated offspring for the $(1+2^\mathrm{s}_\mathrm{ms})$ is the same as for the $(1, 2^\mathrm{s})$, we can therefore use the result shown in [6, Lemma 8] and obtain that $1/A_k$ converges almost surely to $2 - p_s(\sigma)$.

Therefore A_k times B_k converges to

$$\frac{2}{2-p_s(\sigma)}\,\mathrm{CR}_{(1+1)}(\sigma) \ . \quad \square$$

We see in (15) that the convergence rate of the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES is expressed as the product of the convergence rate of the (1+1)-ES times $2/(2-p_s(\sigma))$. The term $2/(2-p_s(\sigma))$—which is always larger or equal one—is the gain brought by sequential selection. Indeed, as sketched in the proof of the theorem, the gain brought by sequential selection in strategies with two offspring (with mirrored or non-mirrored sampling) always equals $2/(2-p_s(\sigma))$.

We give a useful expression for the success probability $p_s(\sigma)$ for a single offspring on the sphere function.

LEMMA 11. *For all $\sigma > 0$, we have*

$$p_s(\sigma) = \Pr\left([\boldsymbol{\mathcal{N}}]_1 \leq -\frac{d}{2}\sigma \underbrace{\frac{\|\boldsymbol{\mathcal{N}}\|^2}{d}}_{\text{close to } 1}\right) \tag{16}$$

PROOF. The lemma follows from the definition of $p_s(\sigma) = \Pr\left(2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\right)$ □

The expression suggests that $\sigma \propto 1/d$ achieves a fairly d-independent success probability. A typical, close to optimal value is $\sigma \approx 1.2/d$ with $p_s \approx 1/4$ and $2/(2-p_s) \approx 1.16$.

Finally, we can give the upper bound for the speed-up brought by sequential selection, when $\lambda = 2$.

COROLLARY 12 (SPEED-UP FOR $\lambda = 2$). *The upper bound for the speed-up brought by sequential selection for $\lambda = 2$ is given by*

$$\frac{2}{2-p_s} < \frac{4}{3} = 1.333\ldots \tag{17}$$

PROOF. From Lemma 11 we find for $\sigma > 0$ that $p_s < \Pr([\boldsymbol{\mathcal{N}}]_1 \leq 0) = 1/2$ which implies (17). For $\sigma = 0$ we have no speed-up. □

This upper bound holds equally well for savings by sequential selection whether or not *skip mirror* is applied.

5. ASYMPTOTIC CONVERGENCE RATES

So far, we have proven the linear convergence of some scale-invariant step-size ESs *for a fixed dimension of the search space*. In this section, we want to study how the finite dimension convergence rates derived previously behave when the dimension goes to infinity. We have observed that the convergence rate of the $(1+1_{\text{ms}})$-ES is a function of the convergence rate of the (1+1)-ES and of the probability of success p_s. We therefore study those two quantities asymptotically in order to obtain the asymptotic behavior of $\text{CR}_{(1+1_{\text{ms}})}$. Both asymptotic estimates were already (less rigorously) derived in [24].

5.1 Asymptotic Probability of Success

We first derive the limit of the probability of success $p_s(\sigma/d)$ when d goes to infinity.

LEMMA 13. *For all $\sigma > 0$*

$$\begin{aligned}\lim_{d\to\infty} p_s\left(\frac{\sigma}{d}\right) &= \Pr([\boldsymbol{\mathcal{N}}]_1 \leq -\sigma/2) \\ &= \Phi\left(-\frac{\sigma}{2}\right)\end{aligned} \tag{18}$$

where Φ is the cumulative distribution of a standard normal distribution, i.e. $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt$ or, with the error function erf, $\Phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$.

PROOF. We start from the expression of $p_s(\sigma/d)$:

$$p_s(\sigma/d) = \Pr\left(2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\right) \tag{19}$$

$$= E\left[1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}}\right] \tag{20}$$

From the LLN, we know that

$$\lim_{d\to\infty} \frac{1}{d}\|\boldsymbol{\mathcal{N}}\|^2 = \lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \mathcal{N}_i^2 = 1$$

almost surely, where $\mathcal{N}_i$ are i.i.d. standard normal distributions that are the coordinates of the vector $\boldsymbol{\mathcal{N}}$. Thus

$$2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2 \xrightarrow[d\to\infty]{} 2[\boldsymbol{\mathcal{N}}]_1 + \sigma$$

and therefore we have that

$$1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}} \xrightarrow[d\to\infty]{} 1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0\}} a.s.$$

Since $1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2\}} \leq 1$, we can apply the Lebesgue dominated convergence theorem that implies that

$$E\left[1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2 \leq 0\}}\right] \xrightarrow[d\to\infty]{} E\left[1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0\}}\right] \ .$$

We can rewrite the RHS of the last equation as

$$E\left[1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0\}}\right] = \Pr(2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0) = \Pr([\boldsymbol{\mathcal{N}}]_1 \leq -\sigma/2) \ .$$

Moreover, $\Pr([\boldsymbol{\mathcal{N}}]_1 \leq -\sigma/2) = \Phi(-\sigma/2)$. □

5.2 Asymptotic Convergence Rate of the (1+1)-ES

We will derive now the asymptotic convergence rate of the (1+1)-ES with scale-invariant step-size and find that it coincides with the negative of the well-known progress rate of the (1+1)-ES [24]. We first need to derive the following technical lemma:

LEMMA 14. *Let $\mathcal{N}$ be a standard normal distribution, the following equation holds*

$$E[\mathcal{N} 1_{\{\mathcal{N} \leq -\sigma/2\}}] = -\frac{1}{\sqrt{2\pi}} \exp(-\frac{\sigma^2}{8}) \ , \tag{21}$$

for all $\sigma > 0$.

PROOF. In a first step we write the LHS of (21) using the density of a normal distribution

$$E[\mathcal{N} 1_{\{\mathcal{N} \leq -\sigma/2\}}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\sigma/2} x \exp(-\frac{x^2}{2}) dx \ . \tag{22}$$

By integrating the RHS of (22) we obtain the result. □

We are now ready to derive the limit of the convergence rate of the (1+1)-ES.

THEOREM 5. *Let $\sigma > 0$, the convergence rate of the (1+1)-ES with scale-invariant step-size on spherical functions satisfies at the limit*

$$\lim_{d\to\infty} d \times \text{CR}_{(1+1)}\left(\frac{\sigma}{d}\right) = \frac{-\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\sigma^2}{8}\right) + \frac{\sigma^2}{2}\Phi\left(-\frac{\sigma}{2}\right) \tag{23}$$

where Φ is the cumulative distribution of a normal distribution.

PROOF. We are going to investigate the almost sure limit of the random variable inside the RHS of

$$CR_{(1+1)}(\sigma/d) = \frac{1}{2} E\left[\ln(1 + \frac{\sigma}{d}\min(2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2, 0))\right] \ . \tag{24}$$

The following equation holds almost surely

$$\begin{aligned}\lim_{d\to\infty} d \times \frac{1}{2}\ln(1 + \sigma/d\min(2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2, 0)) \\ \xrightarrow[d\to\infty]{} \frac{1}{2}\sigma\min(2[\boldsymbol{\mathcal{N}}]_1 + \sigma, 0) \ .\end{aligned} \tag{25}$$

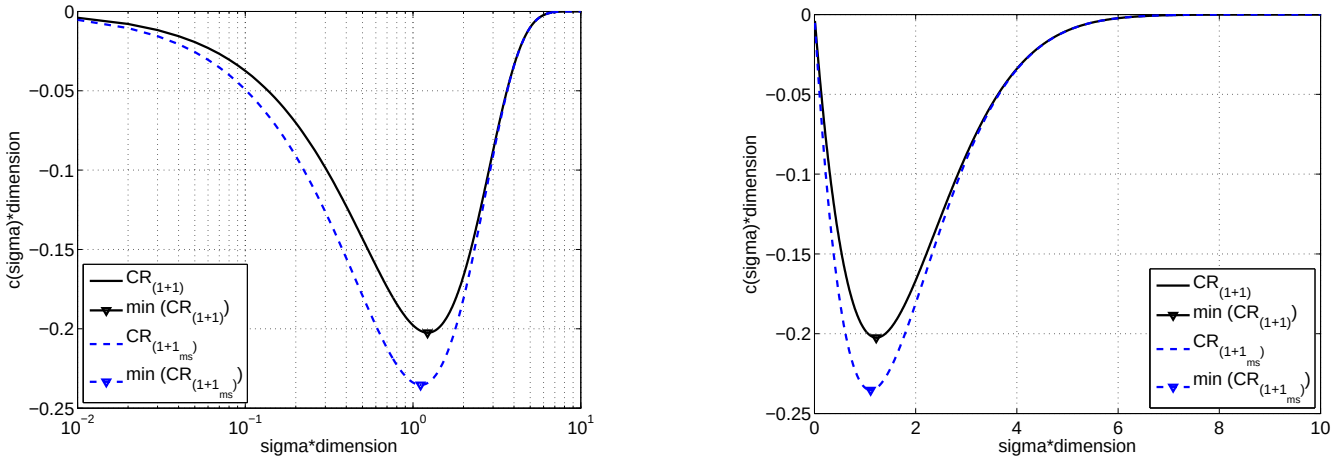

Figure 3: Theoretical limit results of the convergence rate for the (1+1)-ES (solid line) and for the $(1+\lambda^{s}_{ms})$-ES (any $\lambda \geq 1$, dashed line) if d goes to infinity. Left: versus $\sigma \cdot d$ in log scale; right: versus $\sigma \cdot d$ in linear scale.

Assuming the uniform integrability of

$$d \times \frac{1}{2} \ln(1 + \sigma/d \min(2[\boldsymbol{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}\|^2, 0)) \ ,$$

we deduce that

$$\lim_{d\to\infty} d \times \mathrm{CR}_{(1+1)}(\sigma/d) = \frac{\sigma}{2} E[\min(2[\boldsymbol{\mathcal{N}}]_1 + \sigma, 0)] \ .$$

Moreover,

$$\begin{aligned} E[\min(2[\boldsymbol{\mathcal{N}}]_1 + \sigma, 0)] &= E[(2[\boldsymbol{\mathcal{N}}]_1 + \sigma) 1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0\}}] \\ &= 2E[([\boldsymbol{\mathcal{N}}]_1) 1_{\{2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0\}}] \\ &\quad + \sigma \Pr(2[\boldsymbol{\mathcal{N}}]_1 + \sigma \leq 0) \\ &= 2E[([\boldsymbol{\mathcal{N}}]_1) 1_{\{[\boldsymbol{\mathcal{N}}]_1 \leq -\sigma/2\}}] \\ &\quad + \sigma \Pr([\boldsymbol{\mathcal{N}}]_1 \leq -\sigma/2) \ . \end{aligned}$$

Thus,

$$\lim_{d\to\infty} d \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{d}\right) = \sigma E[[\boldsymbol{\mathcal{N}}]_1 1_{\{[\boldsymbol{\mathcal{N}}]_1 \leq -\frac{\sigma}{2}\}}] + \frac{\sigma^2}{2}\Phi\left(\frac{-\sigma}{2}\right) \ .$$

Using now Lemma 14, we obtain the result. □

This limit of the normalized convergence rate of the (1+1)-ES found in the previous theorem equals to the negated progress rate of the (1+1)-ES on the sphere function [24].

5.3 Deriving the Asymptotic Convergence Rate of the $(1+\lambda^{s}_{ms})$-ES

We can now combine Lemma 13 and Theorem 5 to derive the asymptotic convergence rate of the $(1+\lambda^{s}_{ms})$-ES with scale-invariant step-size. Note again that the $(1+\lambda^{s}_{ms})$-ES is here the same as the $(1+1_{ms})$-ES.

THEOREM 6. *Let $\sigma > 0$, the convergence rate of the $(1+1_{ms})$-ES with scale-invariant step-size on spherical functions satisfies*

$$\lim_{d\to\infty} d \times \mathrm{CR}_{(1+1_{ms})}\left(\frac{\sigma}{d}\right) = \frac{2}{2 - \Phi(-\sigma/2)} \times \left(\frac{-\sigma}{\sqrt{2\pi}} \exp(-\frac{\sigma^2}{8}) + \frac{\sigma^2}{2}\Phi\left(-\frac{\sigma}{2}\right)\right) \ . \quad (26)$$

PROOF. Since the convergence rate of the $(1+1_{ms})$-ES equals the convergence rate of the (1+1)-ES times $2/(2-p_s)$ we have that the limit for d to infinity satisfies

$$\begin{aligned} &\lim_{d\to\infty} d \times \mathrm{CR}_{(1+1_{ms})}\left(\frac{\sigma}{d}\right) \\ &\quad = \lim_{d\to\infty}\left(\frac{2}{2-p_s}\right) \times \lim_{d\to\infty}\left(d \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{d}\right)\right) \ . \end{aligned}$$

Using Lemma 13 for the limit of $2/(2-p_s)$ and Proposition 5, we obtain the result. □

Figure 3 represents the limit of the normalized convergence rates of the (1+1)-ES and the $(1+1_{ms})$-ES. The minimal value of the convergence rate of the (1+1)-ES and $(1+1_{ms})$-ES respectively equal approximately -0.202 and -0.235. Mirrored sampling and sequential selection speed up the fastest single-parent evolution strategy asymptotically by a factor of about 1.16.

6. NUMERICAL SIMULATION OF CONVERGENCE RATES

To conclude on the improvements that can be brought by mirrored samples and sequential selection, we now compare the different convergence rates. However, those convergence rates are expressed only implicitly as the expectation of some random variables. We therefore simulate the convergence rate with a Monte-Carlo technique. For each convergence rate expression, we have simulated 10^7 times the random variables inside the expectation and averaged to obtain an estimate of the expectation and therefore of the convergence rate for different σ. Here, σ has been chosen such that $0.01 \leq \sigma \cdot d \leq 10$ and with steps of 0.01 in $\sigma \cdot d$. The minimum of the measured convergence rates over $\sigma \cdot d$ is used as estimate of the *best* convergence rate for each algorithm and dimension—resulting in a slightly (systematically) smaller value than the true one, due to taking the minimal value from several random estimates.

The plots of **Fig. 4** show the resulting convergence rate estimates versus σ in several dimensions. Overall, mirroring and/or sequential selection do not essentially change the

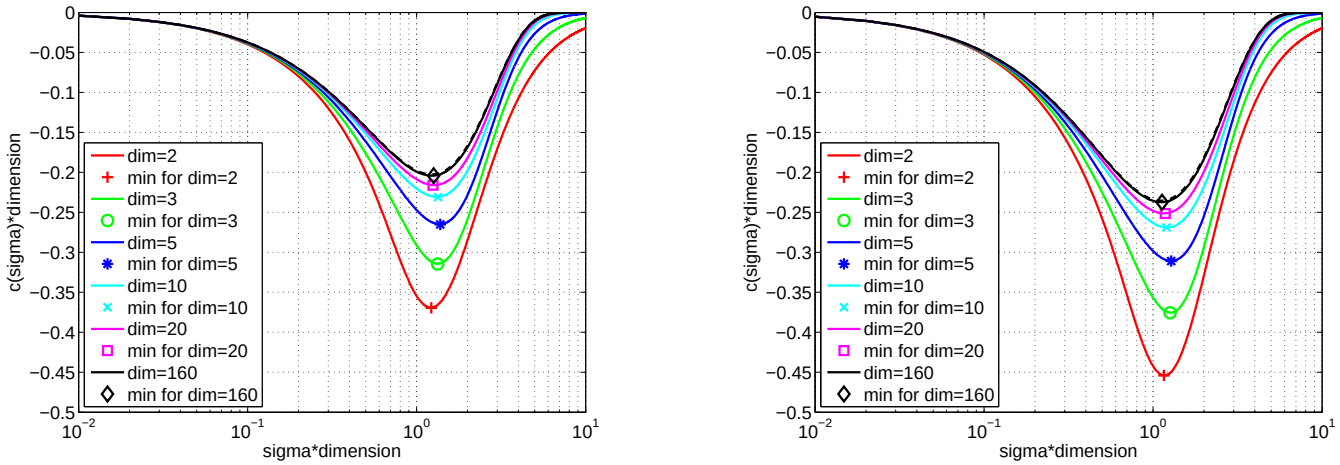

Figure 4: Convergence rate $c(\sigma)$ for different dimensions d and the (1+1)-ES and (1+2$_\mathrm{m}$)-ES (both have the same convergence rate, left figure), and the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES (the same for all $\lambda \geq 1$, right figure), all with scale-invariant step-size. The dashed (uppermost) line shows the limit result for d to infinity.

picture. The (1+1)-ES realizes the largest optimal step-size of all variants, also compared with comma selection (not shown). **Figure 5** shows the relative improvement. For small step-sizes the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES is up to about 33% faster (compare (17)). For large step-sizes, both, (1 + 1)- and (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES, show very similar convergence rates. For close to optimal step-sizes (somewhat above one), the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES is about 15% to 20% faster.

Figure 6 presents the estimated best convergence rates for several algorithms versus dimension. Here, the $(1, 4^\mathrm{s}_\mathrm{ms})$-ES is shown additionally.[4] The convergence rate of the $(1, \lambda^\mathrm{s}_\mathrm{ms})$-ES is monotonically increasing in λ (not shown) and in the limit for $\lambda \to \infty$, the $(1, \lambda^\mathrm{s}_\mathrm{ms})$-ES coincides with the (1+1$_\mathrm{ms}$)-ES. In small dimension, already for $\lambda = 4$ the convergence rate is very close to the convergence rate of the (1+1$_\mathrm{ms}$)-ES. In all cases, the convergence rate of the $(1, 4^\mathrm{s}_\mathrm{ms})$-ES is closer to the (1+1$_\mathrm{ms}$)-ES than to the (1+1)-ES. The difference between the original (1+1)-ES and the (1+1$_\mathrm{ms}$)-ES is roughly between 15 and 20%. In dimension 320, the values are very close to the limit value.

7. DISCUSSION

In this paper we have analyzed the (1+λ)-ES with mirrored sampling and/or sequential selection. With sequential (plus) selection, the parameter λ loses most of its meaning. Given that the step-size (and all other variation parameters) are updated in an identical way, the (1+λ^s)-ES, where $^\mathrm{s}$ denotes sequential selection, and also the $(1, \infty^\mathrm{s})$-ES depict the same strategy for all $\lambda \geq 1$. The same holds analogously for the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ESs, where the subscript $_\mathrm{ms}$ denotes *m*irrored sampling with *s*kip mirror applied (on success).

We have obtained tight lower bounds for the convergence rate of the (1+2$_\mathrm{m}$)-ES and of the (1+1$_\mathrm{ms}$)-ES that coincides with the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ESs for any $\lambda \geq 1$. These bounds are also the convergence rate with scale-invariant optimal step-size on the sphere function. The (1+2$_\mathrm{m}$)-ES has the same convergence rate as the (1+1)-ES, asymptotically with the

[4]Note that in previous publications such as [4, 5, 17], the slightly different notation $(1, 4^\mathrm{s}_\mathrm{m})$-ES was used for the same algorithm.

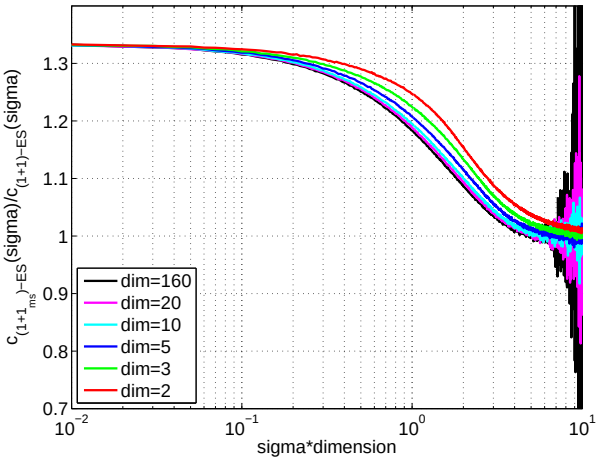

Figure 5: Relative improvement in the convergence rates $c(\sigma)$ of the (1+1$_\mathrm{ms}$)-ES over the (1+1)-ES plotted versus σ times dimension for scale-invariant step sizes. Smaller dimensions show the (slightly) larger improvements. The huge fluctuations to the right are due to the small success probability for large step-sizes and therefore a large variance when measuring a few events.

dimension to ∞ being $\geq -0.202\ldots$ The asymptotic convergence rate of the (1+1$_\mathrm{ms}$)-ES is $\geq -0.235\ldots$ and the relation

$$\mathrm{CR}_{(1+\lambda^\mathrm{s}_\mathrm{ms})}(\sigma) = \mathrm{CR}_{(1+1_\mathrm{ms})}(\sigma) \tag{27}$$

$$= \frac{2}{2 - p_s(\sigma)} \mathrm{CR}_{(1+2_\mathrm{m})}(\sigma) \tag{28}$$

$$= \frac{2}{2 - p_s(\sigma)} \mathrm{CR}_{(1+1)}(\sigma) \tag{29}$$

holds, where $p_s(\sigma)$ is the probability that an offspring, sampled isotropically with step-size σ, is better than its parent.

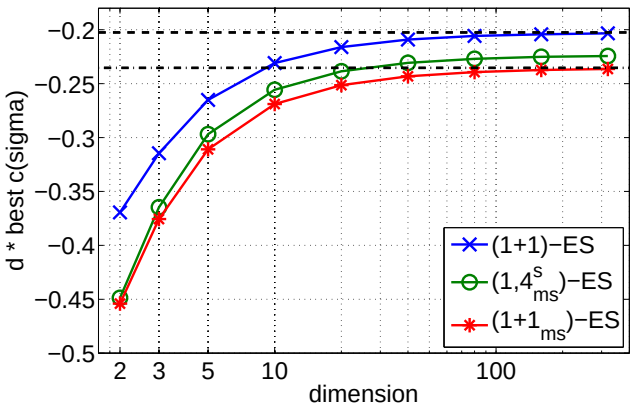

Figure 6: Estimated optimal convergence rates, in parts extracted from Fig. 4, multiplied by the dimension and plotted versus dimension for the algorithms $(1+1)$-ES (equivalent to $(1+2_{\mathrm{m}})$-ES), $(1+1_{\mathrm{ms}})$-ES, and $(1, 4^{\mathrm{s}}_{\mathrm{ms}})$-ES with scale-invariant step-size. In addition, the theoretical limit results for d to infinity are shown for the $(1+1)$-ES (dashed) and for the $(1+1_{\mathrm{ms}})$-ES (dotted-dashed).

The factor $2/(2 - p_s(\sigma)) < 4/3$ is the improvement brought by sequential selection for $\lambda = 2$, with plus as well as comma selection.

As to our knowledge, the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is now the single-parent evolution strategy with the fastest known convergence rate, more than 15% faster than the (1+1)-ES, but no more than 5% faster than the $(1, 4^{\mathrm{s}}_{\mathrm{ms}})$-ES. Only strategies with weighted recombination can exhibit even faster convergence rates (also denoted as serial efficiencies), namely ≥ -0.25 when positive recombination weights are used [2].

The convergence rates derived assume that the step-size equals a constant times the distance to the optimum. This assumption simplifies the linear convergence derivation as the law of large numbers for independent random variables can then be used. For real adaptation schemes however, the analysis on spherical functions is in general more complicated, as $\sigma_k/\|\boldsymbol{X}_k\|$ is not a constant but a Markov chain. Law of large numbers for Markov chains can be used to prove linear convergence, the difficult task being to prove that $\sigma_k/\|\boldsymbol{X}_k\|$ is stable enough to satisfy a LLN [3, 16].

References

[1] D. V. Arnold and R. Salomon. Evolutionary gradient search revisited. *IEEE Transactions on Evolutionary Computation*, 11(4):480–495, 2007.

[2] D.V. Arnold. Optimal weighted recombination. *Foundations of Genetic Algorithms*, pages 215–237, 2005.

[3] A. Auger. Convergence results for the $(1, \lambda)$-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science*, 334(1–3):35–69, 2005.

[4] A. Auger, D. Brockhoff, and N. Hansen. Benchmarking the (1,4)-CMA-ES With Mirrored Sampling and Sequential Selection on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1617–1624. ACM, 2010.

[5] A. Auger, D. Brockhoff, and N. Hansen. Benchmarking the (1,4)-CMA-ES With Mirrored Sampling and Sequential Selection on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1625–1632. ACM, 2010.

[6] A. Auger, D. Brockhoff, and N. Hansen. Comparing the (1+1)-CMA-ES with a Mirrored (1+2)-CMA-ES with Sequential Selection on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1543–1550. ACM, 2010.

[7] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,2)-CMA-ES on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1591–1596. ACM, 2010.

[8] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,2)-CMA-ES on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1605–1610. ACM, 2010.

[9] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,4)-CMA-ES on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1597–1604. ACM, 2010.

[10] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,4)-CMA-ES on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1611–1616. ACM, 2010.

[11] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Sampling and Sequential Selection for Evolution Strategies. Rapport de Recherche RR-7249, INRIA Saclay—Île-de-France, June 2010.

[12] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,2)-CMA-ES Compared on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1551–1558. ACM, 2010.

[13] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,2)-CMA-ES Compared on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1575–1582. ACM, 2010.

[14] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,4)-CMA-ES Compared on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1559–1566. ACM, 2010.

[15] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,4)-CMA-ES Compared on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1583–1590. ACM, 2010.

[16] A. Auger and N. Hansen. Theory of evolution strategies: a new perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing, 2010. In press.

[17] D. Brockhoff, A. Auger, N. Hansen, D. V. Arnold, and T. Hohm. Mirrored Sampling and Sequential Selection for Evolution Strategies. In R. Schaefer et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN XI)*, pages 11–21. Springer, 2010.

[18] J. M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Methuen's monographs on applied probability and statistics. Methuen, 1967.

[19] N. Hansen. An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.

[20] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009. Updated February 2010.

[21] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noisy functions definitions. Technical Report RR-6869, INRIA, 2009. Updated February 2010.

[22] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[23] 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.

[24] I. Rechenberg. Optimierung technischer Systeme nach Prinzipien der biologischen Evolution Dr.-Ing. Dissertation. Technical report, Verlag Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1973.

[25] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In *Conference on Parallel Problem Solving from Nature (PPSN 2006)*, volume 4193, pages 21–31. Springer, 2006.

[26] O. Teytaud and S. Gelly. DCMA, yet another derandomization in covariance-matrix-adaptation. In D. Thierens et al., editors, *Genetic and Evolutionary Computation Conference (GECCO)*, pages 955–922. ACM Press, 2007.

[27] O. Teytaud, S. Gelly, and J. Mary. On the Ultimate Convergence Rates for Isotropic Algorithms and the Best Choices Among Various Forms of Isotropy . In T. P. Runarsson et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN IX)*, volume 4193 of *LNCS*, pages 32–41. Springer, 2006.