Mirrored Sampling in Evolution Strategies With Weighted Recombination

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

Dimo Brockhoff
Sysmo Team
LIX, Ecole Polytechnique
91128 Palaiseau Cedex, France
brockho@lix.polytechnique.fr

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

ABSTRACT

This paper introduces mirrored sampling into evolution strategies (ESs) with weighted multi-recombination. Two further heuristics are introduced: *pairwise selection* selects at most one of two mirrored vectors in order to avoid a bias due to recombination. *Selective mirroring* only mirrors the worst solutions of the population. Convergence rates on the sphere function are derived that also yield upper bounds for the convergence rate on any spherical function. The optimal fraction of offspring to be mirrored is regardless of pairwise selection one without selective mirroring and about 19% with selective mirroring, where the convergence rate reaches a value of 0.390. This is an improvement of 56% compared to the best known convergence rate of 0.25 with positive recombination weights.

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

1. INTRODUCTION

Derandomization of random numbers is a general technique where independent samples are replaced by dependent ones. Recent studies showed, for example, how derandomization via *mirrored sampling* can improve $(1, \lambda)$- and $(1 + \lambda)$-ESs [6, 3]. Instead of generating λ independent and identically distributed (i.i.d.) search points in iteration k as $\boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ where $\boldsymbol{X}_k$ is the current search point, σ_k the current step-size, and $\boldsymbol{\mathcal{N}}^i$ a random sample from a multivariate normal distribution, the $(1 \dot{+} \lambda)$-ES with mirrored sampling always pairs samples one by one and produces $\lambda/2$ independent search points $\boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ and $\lambda/2$ dependent ones as $\boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}^i$ $(1 \leq i \leq \lambda/2)$. In the end, the best out of these λ search points is used as next search point $\boldsymbol{X}_{k+1}$ in the $(1, \lambda)$-ES (and in the $(1+\lambda)$-ES the best out of the λ new *and* the old $\boldsymbol{X}_k$). Several ES variants using mirrored mutations showed noticeable improvements over their unmirrored counterparts—not only in theoretical investigations on simple test functions such as the sphere function, but also in exhaustive experiments within the COCO/BBOB framework [6, 3]. Up to now, the results were restricted to single-parent $(1 \dot{+} \lambda)$-ESs though the idea is, in principle, applicable in a straight-forward manner to population-based ESs such as the $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES where the μ best out of the λ offspring are used to compute the new search point $\boldsymbol{X}_{k+1}$ via weighted recombination. However, the direct application of mirrored mutations in population-based ESs, as for example proposed in a more general way by Teytaud et al. [12], results in an undesired bias on the step-size, as was argued already in [6].

The purpose of this paper is to introduce mirrored mutations into ESs with weighted recombination *without* introducing a bias on the length of the recombined step. The main idea hereby is *pairwise selection* that allows only the better solution of a mirrored/unmirrored solution pair to possibly contribute to the weighted recombination. In detail, the contributions of this paper are

- the introduction of several ES variants that combine mirrored mutations and weighted recombination without a bias on the recombined step,
- a theoretical investigation of the algorithms' convergence rates (in finite and infinite dimension) on spherical functions,
- the computation of optimal recombination weights, and
- an experimental comparison of convergence rates with only positive recombination weights, in particular evaluating the impact of mirrored mutations.

The paper is organized as follows. After introducing the baseline $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES, Sec. 2 explains in detail how mirrored mutations can be introduced in this algorithm. Section 3 theoretically investigates the convergence rate of three variants in finite and infinite dimension. Section 4 presents a comparison of the algorithms based on the numerical estimation of their convergence rates on the sphere function. Section 5 summarizes and concludes the paper.

Notations. For a (random) vector $\boldsymbol{x} \in \mathbb{R}^n$, $[\boldsymbol{x}]_1$ will denote its first coordinate. The vector $(1, 0, \ldots, 0)$ will be denoted $\boldsymbol{e}_1$. A random vector following a multivariate normal distribution with mean vector zero and covariance matrix identity will be called *standard* multivariate normal distribution.

Algorithm 1 $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m})$-ES with random or selective mirroring with or without resampled mutation lengths for the mirrored vectors

1: **given**: $f : \mathbb{R}^d \to \mathbb{R}$, $\boldsymbol{X}_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, $\lambda_\mathrm{iid} \in \mathbb{N}^+$, $\lambda_\mathrm{m} \in \{0, \ldots, \lambda_\mathrm{iid}\}$, $\mu \leq \lambda_\mathrm{iid}$, $(\boldsymbol{\mathcal{N}}^r)_{r\in\mathbb{N}}$, weights $\boldsymbol{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^\mu |w_i| = 1$ and $|\{w_i \geq 0\}| \geq \lambda_\mathrm{m}$
2: $r \leftarrow 0$ *(index of current random sample)*
3: $k \leftarrow 0$ *(iteration counter for notational consistency)*
4: **while** stopping criterion not fulfilled **do**
5: ** generate λ_iid offspring independently **/*
6: $i \leftarrow 0$ *(offspring counter)*
7: **while** $i < \lambda_\mathrm{iid}$ **do**
8: $i \leftarrow i + 1$
9: $r \leftarrow r + 1$
10: $\boldsymbol{X}_k^i = \boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^r$
11: **if** selective mirroring **then**
12: $\boldsymbol{X}_k^1, \ldots, \boldsymbol{X}_k^{\lambda_\mathrm{iid}} = \mathrm{argsort}(f(\boldsymbol{X}_k^1), \ldots, f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}}))$
13: */* mirror λ_m offspring */*
14: **while** $i < \lambda_\mathrm{iid} + \lambda_\mathrm{m}$ **do**
15: $i \leftarrow i + 1$
16: */* dependent sample (with new length $\|\boldsymbol{\mathcal{N}}^r\|$) */*
17: **if** resample lengths **then**
18: $r \leftarrow r + 1$;
19: $\boldsymbol{X}_k^i = \boldsymbol{X}_k - \frac{\sigma_k \|\boldsymbol{\mathcal{N}}^r\|}{\|\boldsymbol{X}_k^{i-\lambda_\mathrm{m}} - \boldsymbol{X}_k\|} (\boldsymbol{X}_k^{i-\lambda_\mathrm{m}} - \boldsymbol{X}_k)$
20: **else**
21: $\boldsymbol{X}_k^i = \boldsymbol{X}_k - (\boldsymbol{X}_k^{i-\lambda_\mathrm{m}} - \boldsymbol{X}_k)$
22: */* weighted recombination */*
23: $\boldsymbol{X}_k^1, \ldots, \boldsymbol{X}_k^{\lambda_\mathrm{iid}} =$
24: $\mathrm{argsort}(f(\boldsymbol{X}_k^1), \ldots, f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}-\lambda_\mathrm{m}}),$
25: $\min\{f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}-\lambda_\mathrm{m}+1}), f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}+1})\}, \ldots,$
26: $\min\{f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}}), f(\boldsymbol{X}_k^{\lambda_\mathrm{iid}+\lambda_\mathrm{m}})\})$
27: $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \sum_{i=1}^\mu w_i (\boldsymbol{X}_k^i - \boldsymbol{X}_k)$
28: $\sigma_{k+1} = \mathrm{update}(\sigma_k, \boldsymbol{X}_k^1, \ldots, \boldsymbol{X}_k^{\lambda_\mathrm{iid}}, \boldsymbol{X}_k)$
29: $k \leftarrow k + 1$ *(iteration counter)*

2. MIRRORING AND WEIGHTED RECOMBINATION

2.1 The Standard $(\mu/\mu_\mathrm{w}, \lambda)$-ES

As our baseline algorithm, we briefly recapitulate the standard $(\mu/\mu_\mathrm{w}, \lambda)$-ES with weighted recombination and show its pseudo code in Algorithm 1, where $\lambda_\mathrm{m} = 0$ and therefore $\lambda = \lambda_\mathrm{iid}$. Weighted recombination generalizes intermediate multi-recombination (where all weights are equal), has been studied in [10, 7, 2], and must be nowadays regarded as standard in ESs. Given a starting point $\boldsymbol{X}_0 \in \mathbb{R}^d$, an initial step-size $\sigma_0 > 0$, a population size $\lambda \in \mathbb{N}^+$, and weights $\boldsymbol{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^\mu |w_i| = 1$ for a chosen $1 \leq \mu \leq \lambda$, the $(\mu/\mu_\mathrm{w}, \lambda)$-ES generates at iteration k λ independent search points from a multivariate normal distribution with mean $\boldsymbol{X}_k$ and variance σ_k^2 and recombines the best μ of them in terms of a weighted sum to become the new mean $\boldsymbol{X}_{k+1}$ of the next iteration (in case of negative weights, the steps must be recombined, see line 27 in Algorithm 1).

Typically, μ is chosen as $\lfloor \lambda/2 \rfloor$ and $w_i = \ln(\frac{\lambda+1}{2}) - \ln(i) > 0$ in the scope of the CMA-ES [7]. As update rule for the step-size σ_k in the $(\mu/\mu_\mathrm{w}, \lambda)$-ES, several techniques such as self-adaptation [11] or cumulative step-size adaptation [9] are available. Of particular theoretical interest is the scale-invariant (constant) step-size $\sigma_k = \sigma \|\boldsymbol{X}_k\|$ which depends on the distance to the optimum assumed WLOG in zero and which allows to prove bounds on the convergence rate of evolution strategies with any adaptive step-size update, see Sec. 3.

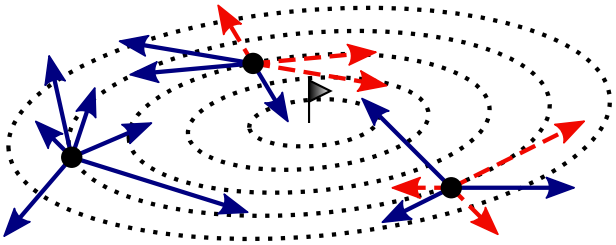

Figure 1: Illustration of i.i.d. mutations (left) and mirrored mutations (middle) and mirrored mutations with resampled lengths (right). Dashed arrows depict the mirrored samples. Dotted lines connect points with equal function value.

2.2 The Mirroring Idea

Derandomized mutations [12] and more recently mirrored mutations [6, 3] have been proposed to replace the independent mutations in evolution strategies by dependent ones in order to reduce the probability of "unlucky events"—resulting in an increase in the convergence speed of the algorithms. Instead of sampling all λ offspring i.i.d., an algorithm with mirrored mutations samples only $\lceil \lambda/2 \rceil$ i.i.d. offspring as $\boldsymbol{X}_k^i = \boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ $(1 \leq i \leq \lceil \lambda/2 \rceil)$ and up to $\lfloor \lambda/2 \rfloor$ further offspring depending on the already drawn samples as $\boldsymbol{X}_k^i = \boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}^{i-\lceil \lambda/2 \rceil}$ (for $\lceil \lambda/2 \rceil + 1 \leq i \leq \lambda$), see Fig. 1, left versus middle.

In evolution strategies with weighted recombination and cumulative step-size adaption, mirrored mutations cause a bias towards smaller step-sizes [6, Fig. 4], see Fig 2. The bias can cause premature convergence of the algorithm.[1] The reason for the bias is that if both samples $\boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ and $\boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}^i$ are considered within weighted recombination, they partly cancel each other out and the realized shift of $\boldsymbol{X}_k$ will be smaller than with independent mutations. Consequently, derandomized step-size control like cumulative step-size adaptation [9] will cause the step-size to shrink.

In this paper, we therefore introduce *pairwise selection* which prevents this bias: unmirrored and mirrored offspring are paired two-by-two and only the better among the unmirrored sample $\boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ and its mirrored counterpart $\boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}^i$ is used within the weighted recombination but never both. Here, we introduce a new notation: the number of independent offspring per iteration is denoted by λ_iid and the number of mirrored offspring per iteration is denoted by λ_m, where each iteration $\lambda = \lambda_\mathrm{iid} + \lambda_\mathrm{m}$ solutions are evaluated on f. As a result $0 \leq \lambda_\mathrm{m} \leq \lambda_\mathrm{iid}$ which results in the standard $(\mu/\mu_\mathrm{w}, \lambda)$-ES in case $\lambda_\mathrm{m} = 0$. We denote the new algorithm $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid} + \lambda_\mathrm{m})$-ES.

Note that the idea of sequential mirroring of [6, 3], i.e., stopping the generation of new offspring as soon as a better solution than the parent is found, is not applied here. With

[1] Mutative self-adaptation has no such bias, but suffers in combination with weighted recombination from a far too small control target step-size and can achieve close to optimal step-sizes only with a peculiar parameter tuning.

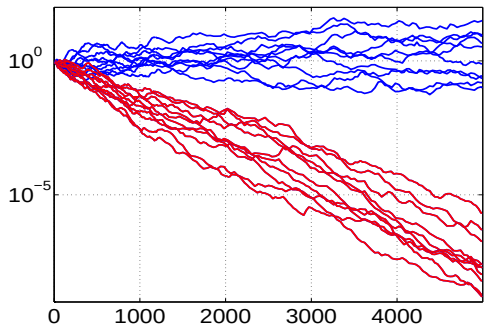

Figure 2: Illustration of the bias towards smaller step-sizes under random selection introduced by recombination of mirrored vectors in the CMA-ES. Shown is the step-size σ versus the number of function evaluations of 20 runs on a purely random fitness function in dimension 10. The upper ten graphs show the $(5/5_\mathrm{w}, 10)$-CMA-ES revealing a random walk on $\log(\sigma)$. The lower ten graphs show the $(5/5_\mathrm{w}, 5+5_\mathrm{m})$-CMA-ES without pairwise selection of mirrored samples.

recombination, the meaning of a comparison with "the parent" is not unique and additional algorithm design decisions were necessary[2]. Instead, *selective mirroring* is introduced.

2.3 Random Versus Selective Mirroring

We consider two variants of the $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m})$-ES[3] that differ in the choice of mirrored offspring. The $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m}^\mathrm{rand})$-ES, where λ_m randomly chosen offspring are mirrored, and the $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m}^\mathrm{sel})$-ES with *selective mirroring*, where only the λ_m worst offspring are selected for mirroring.

The reason behind the latter variant of selecting the worst offspring for mirroring is the following: in particular on fitness functions with convex sublevel sets[4] we do not expect the best of λ_iid offspring to improve by mirroring. For an offspring that is better than the current search point $\boldsymbol{X}_k$, the mirroring would always result in a worse solution since never both an independently drawn solution and its mirrored counterpart can be better than the parent in case of convex sublevel sets [3, Proposition 2].

Regarding the comparison of random and selective mirroring, two questions arise: (i) how much faster can an ES become with selective mirroring and (ii) what is the optimal choice for the number λ_m of mirrored offspring. Both questions will be answered in the following by theoretical investigations of the algorithms' convergence rates.

[2]The super-parent and distribution mean $\boldsymbol{X}_k$, resulting from the weighted recombination, is not directly comparable to the offspring, because depending on d, μ and λ with a large probability all i.i.d. sampled offspring might be worse. However, a feasible heuristic could be to compare with the best offspring from the last iteration.

[3]Adaptive variants with a variable number of mirrored offspring that depends on the observed fitness function values have also been considered but are not included here.

[4]The sublevel set S_l contains all points in search space with a fitness value of at most l: $S_l = \{x \in \mathbb{R}^d \mid f(x) \leq l\}$.

2.4 Resampled Mirrored Vector Lengths

Within the $(\mu/\mu_\mathrm{w}, \lambda)$-ES, solutions that happen to originate from a comparatively long step tend to be worse than average. Therefore, the solutions chosen by *selective mirroring* are biased towards long mutation steps and their mirrors tend to be bad solely because they originate from a long mirrored step (still they tend to be better than the original λ_iid samples). Hence, we consider a variant of mirroring where the lengths of the mirrored vectors $\sigma_k \boldsymbol{\mathcal{N}}^r$ are i.i.d. resampled, i.e., where $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \sigma_k \boldsymbol{\mathcal{N}}^r$ is replaced by $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \sigma_k \frac{\|\boldsymbol{\mathcal{N}}^{r+1}\|}{\|\boldsymbol{\mathcal{N}}^r\|} \boldsymbol{\mathcal{N}}^r$ with $\|\boldsymbol{\mathcal{N}}^{r+1}\|$ the newly sampled length of the mirrored vector, cp. Fig. 1, right.

We refer to this last variant as $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m}^\mathrm{sel})$-ES with *resampled (mutation) lengths*. Algorithm 1 shows the pseudo code of all variants with random/selective mirroring and with/without resampled lengths of the mirrored offspring. Theoretical results in the next section will not only show how much improvement in the convergence rate can be gained by the resampled lengths but also that the variants with and without resampled lengths are the same if the dimension goes to infinity.

2.5 Algorithm Parameters

The algorithms we have described involve several parameters, the number λ_iid of independent samples, the number λ_m of mirrored offspring, the number μ of offspring to be recombined, the weights $\boldsymbol{w}$ for recombination, and the step-size σ_k. The convergence rates depend on the choice of these parameters. In the sequel, we investigate upper bounds for the convergence rate on spherical functions by investigating the optimal choice of the different parameters, given $\sum |w_i| = 1$.

3. CONVERGENCE RATE UPPER BOUNDS ON SPHERICAL FUNCTIONS

In order to find optimal settings for the different parameters, we investigate convergence rates on spherical functions having WLOG the optimum in zero, i.e., $g(\|x\|)$, $g \in \mathcal{M}$ where $\mathcal{M}$ denotes the set of functions $g : \mathbb{R} \mapsto \mathbb{R}$ that are strictly increasing. Convergence rates depend on the step-size adaptation chosen. We study the case of the scale-invariant constant step-size where $\sigma_k = \sigma \|\boldsymbol{X}_k\|$, that we refer to as scale-invariant ES (however, as most ESs are scale invariant, the name is somewhat abusive). For an optimal choice of the constant σ, the scale-invariant ES gives the upper bound for convergence rates achievable by any strategy with step-size adaptation on spherical functions (see below). This case is of great relevance, because also practical step-size control mechanisms can achieve close-to-optimal step-sizes on spherical functions. For different algorithm variants with scale-invariant step-size, we prove linear convergence in expectation in the following sense: there exists a $\mathrm{CR} \in \mathbb{R}$ such that for all $k, k_0 \in \mathbb{N}$ with $k > k_0$

$$\frac{1}{\Lambda} \frac{1}{k-k_0} E\left[-\ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_{k_0}\|}\right] = \mathrm{CR} \ , \qquad (1)$$

where Λ is the number of evaluations per iteration introduced to define the convergence rate per function evaluation. The constant CR[5] is the convergence rate of the algorithm

[5]Convergence takes place if and only if $\mathrm{CR} > 0$, and for non-elitist ESs only numerical integration of CR (expressed

and depends on the dimension d and the parameters σ, λ_{iid}, λ_{m}, μ and $\boldsymbol{w}$, see Sections 2.1 and 2.5 and Algorithm 1. The convergence rate defined in (1) is compatible with almost sure convergence [4]. Hence, we will prove that with scale-invariant step-size, almost surely

$$-\frac{1}{\Lambda}\frac{1}{k-k_0}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_{k_0}\|} \xrightarrow[k\to\infty]{} \text{CR} \ . \tag{2}$$

More loosely we may say that $\|\boldsymbol{X}_{k+1}\| \approx \exp(-\Lambda\,\text{CR})\|\boldsymbol{X}_k\|$.

3.1 The $(\mu/\mu_{\text{w}},\lambda)$-ES

To serve as a baseline algorithm for a later comparison with algorithms using mirrored mutations, we first investigate the convergence rate of the scale-invariant version of the standard $(\mu/\mu_{\text{w}},\lambda)$-ES (see Algorithm 1 with $\lambda_{\text{m}}=0$).

3.1.1 Finite Dimension Results

At each step, λ independent vectors following a standard multivariate normal distribution $\boldsymbol{\mathcal{N}}^i$ are sampled to create the offspring $\boldsymbol{X}_k^i = \boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i$. The offspring are ranked according to their fitness function value. We denote $(\boldsymbol{Z}_{1:\lambda},\ldots,\boldsymbol{Z}_{\lambda:\lambda})$ the sorted vector of multivariate normal distributions, such that the best offspring equals $\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{Z}_{1:\lambda}$, the second best $\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{Z}_{2:\lambda}$, etc. The distribution of $(\boldsymbol{Z}_{1:\lambda},\ldots,\boldsymbol{Z}_{\lambda:\lambda})$ depends *a priori* on $\boldsymbol{X}_k$. However, in the scale-invariant step-size case on spherical functions the distribution is independent of $\boldsymbol{X}_k$ and is determined by ranking of $\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\|$ for $i=1\ldots\lambda$, that can be simplified to ranking $2[\boldsymbol{\mathcal{N}}^i]_1+\sigma\|\boldsymbol{\mathcal{N}}^i\|^2$. These results are stated in the following lemma.

LEMMA 1. *For the scale-invariant $(\mu/\mu_{\text{w}},\lambda)$-ES minimizing spherical functions, the probability distribution of the vector $(\boldsymbol{Z}_{1:\lambda},\ldots\boldsymbol{Z}_{\lambda:\lambda})$ is independent of $\boldsymbol{X}_k$ and equals*

$$(\boldsymbol{Z}_{1:\lambda},\ldots,\boldsymbol{Z}_{\lambda:\lambda}) = \text{argsort}\{h_\sigma(\boldsymbol{\mathcal{N}}^1),\ldots,h_\sigma(\boldsymbol{\mathcal{N}}^\lambda)\} \tag{3}$$

where $h_\sigma(\boldsymbol{x}) = 2[\boldsymbol{x}]_1+\sigma\|\boldsymbol{x}\|^2$ and $(\boldsymbol{\mathcal{N}}^i)_{1\le i\le\lambda}$ are λ independent standard multivariate normal distribution.

PROOF. At iteration k, starting from $\boldsymbol{X}_k$, the distribution of the selected $\boldsymbol{\mathcal{N}}^i$ is determined by the ranking of

$$(\|\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i\|)_{1\le i\le\lambda} \ .$$

Normalizing by $\|\boldsymbol{X}_k\|$ will not change the ranking such that the distribution is determined by the ranking of

$$\left\|\frac{\boldsymbol{X}_k}{\|\boldsymbol{X}_k\|}+\sigma\boldsymbol{\mathcal{N}}^i\right\| \text{ for } 1\le i\le\lambda \ .$$

However, since the distribution of $\boldsymbol{\mathcal{N}}^i$ is spherical, the distribution of the selected $\boldsymbol{\mathcal{N}}^i$ will be the same if we start from any vector with unit norm (like $\frac{\boldsymbol{X}_k}{\|\boldsymbol{X}_k\|}$), so WLOG the distribution will be determined by ranking $\left\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\right\|$ for $1\le i\le\lambda$ or since composing by $g(x)=x^2$ will not change the ranking

$$\left\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\right\|^2 \text{ for } 1\le i\le\lambda \ .$$

We develop $\left\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\right\|^2$ and obtain $1+2\sigma[\boldsymbol{\mathcal{N}}^i]_1+\sigma^2\|\boldsymbol{\mathcal{N}}^i\|^2$. Ranking will not be affected if we subtract 1 and divide by σ such that the distribution of the selected $\boldsymbol{\mathcal{N}}^i$ is determined by the ranking with respect to $h_\sigma(\boldsymbol{\mathcal{N}}^i)$. □

In the $(\mu/\mu_{\text{w}},\lambda)$-ES, the μ best offspring $\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{Z}_{i:\lambda}$ for $i=1,\ldots,\mu$ are recombined into the new parent $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\sum_{i=1}^\mu w_i\boldsymbol{Z}_{i:\lambda}$ where $(w_1,\ldots,w_\mu)\in\mathbb{R}^\mu$ and $\sum_{i=1}^\mu|w_i|=1$. The next theorem gives the expression of the convergence rate associated to the $(\mu/\mu_{\text{w}},\lambda)$-ES with scale-invariant step-size as a function of σ and $\boldsymbol{w}=(w_1,\ldots,w_\mu)$.

THEOREM 1. *For the $(\mu/\mu_{\text{w}},\lambda)$-ES with scale-invariant step-size on $g(\|x\|)$, $g\in\mathcal{M}$, (1) and (2) hold and the convergence rate equals*

$$\text{CR}(\sigma,\boldsymbol{w}) = -\frac{1}{2\lambda}E\ln\left(1+2\sum_{i=1}^\mu\sigma w_i\left[\boldsymbol{Z}_{i:\lambda}\right]_1+\left\|\sum_{i=1}^\mu\sigma w_i\boldsymbol{Z}_{i:\lambda}\right\|^2\right) \tag{4}$$

where $w_i\in\mathbb{R}$ and $\sum_{i=1}^\mu|w_i|=1$.

PROOF. We start from

$$\|\boldsymbol{X}_{k+1}\| = \left\|\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\sum_{i=1}^\mu w_i\boldsymbol{Z}_{i:\lambda}\right\|$$

that we normalize by $\|\boldsymbol{X}_k\|$ and take the logarithm

$$\ln\frac{\|\boldsymbol{X}_{k+1}\|}{\|\boldsymbol{X}_k\|} = \ln\left\|\boldsymbol{X}_k/\|\boldsymbol{X}_k\|+\sigma\sum_{i=1}^\mu w_i\boldsymbol{Z}_{i:\lambda}\right\| \ . \tag{5}$$

Using the isotropy of the sphere function and of the multivariate normal distribution, together with the previous lemma, we find that the random variables in the RHS of the previous equation are i.i.d. distributed as $\ln\|\boldsymbol{e}_1+\sigma\sum_{i=1}^\mu w_i\boldsymbol{Z}_{i:\lambda}\|$. Applying the Law of Large Numbers to

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

we find thus that

$$\frac{1}{\lambda k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} = \frac{1}{\lambda}E[\ln\|\boldsymbol{e}_1+\sum_{i=1}^\mu\sigma w_i\boldsymbol{Z}_{i:\lambda}\|]$$

We develop the convergence in the RHS of the previous equation using the identity

$$\ln\|\boldsymbol{e}_1+u\| = \frac{1}{2}\ln\left[1+2u_1+\|u\|^2\right], \text{ for } u\in\mathbb{R}^n \tag{6}$$

that can be obtained in a straightforward way by writing $\ln\|\boldsymbol{e}_1+u\|$ as $\frac{1}{2}\ln\|\boldsymbol{e}_1+u\|^2$ and developing the norm. We then obtain that (2) holds with the convergence rate CR as given in the theorem. To obtain the convergence in expectation as defined in (1), we take the expectation in (5). For a more detailed argumentation why the expectation exists and for the independence of the random variables $\ln\|\boldsymbol{X}_k\|/\|\boldsymbol{X}_k\|$, we refer to [8]. □

The convergence rate of the $(\mu/\mu_{\text{w}},\lambda)$-ES is a function of σ and the weights. The optimal convergence rate computes with $\text{CR}(\sigma,\boldsymbol{w})$ from (4) and the constraint $\sum|w_i|=1$ to

$$\text{CR}^{\text{opt}}_{(\mu/\mu_{\text{w}},\lambda)} = \max_{\boldsymbol{y}\in\mathbb{R}^\mu}\text{CR}_{(\mu/\mu_{\text{w}},\lambda)}\left(\sum_{i=1}^d|y_i|,\frac{\boldsymbol{y}}{\sum_{i=1}^d|y_i|}\right) \ . \tag{7}$$

as an expectation of some continuous random variables and thus as an integral) can reveal the sign of CR. In contrast to our results, "classical" progress-rate derivations [1, 2, 5] only approximate the actual strategy behavior and the result of the approximation can be comparatively poor for small values of d/λ. Consequently, the classical progress-rate φ might be negative even when de facto convergence takes place, or vice versa [4].

Optimal weights can be obtained as $w_i^{\text{opt}} = y_i^{\text{opt}} / \sum_{i=1}^{\mu} |y_i^{\text{opt}}|$ with $\boldsymbol{y}^{\text{opt}} = \text{argmax}_{\boldsymbol{y} \in \mathbb{R}^\mu} \text{CR}_{(\mu/\mu_\text{w},\lambda)}(\sum_{i=1}^{d} |y_i|, \boldsymbol{y}/\sum_{i=1}^{d} |y_i|)$ and the optimal step-size equals $\sum_{i=1}^{\mu} |y_i^{\text{opt}}|$.

The convergence rate of the $(\mu/\mu_\text{w}, \lambda)$-ES with scale-invariant constant step-size gives the upper bound for the convergence rate of any step-size adaptive $(\mu/\mu_\text{w}, \lambda)$-ES with isotropic mutations on spherical functions [8, Theorem 1].

3.1.2 Asymptotic Results

In the following, we investigate the limit of the convergence rate when the dimension of the search space goes to infinity.

THEOREM 2. *The convergence rate of the $(\mu/\mu_\text{w}, \lambda)$-ES on the class of spherical functions $g(\|x\|)$, $g \in \mathcal{M}$, given scale-invariant step-size and $\sum_{i=1}^{\mu} |w_i| = 1$, satisfies*

$$\lim_{d\to\infty} d \,\text{CR}\left(\frac{\sigma}{d}, \boldsymbol{w}\right) = -\frac{1}{\lambda}\left(\frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 + \sigma \sum_{i=1}^{\mu} w_i E(\mathcal{N}_{i:\lambda})\right)$$

where $\mathcal{N}_{i:\lambda}$ is the ith order statistic of λ independent normal distributions with mean 0 and variance 1, i.e., the ith smallest of λ independent variables $\mathcal{N}_i \sim N(0,1)$.

PROOF. Let $(\boldsymbol{\mathcal{N}}^i)_{1\le i\le\lambda}$ be λ independent standard multivariate normal distributions. With the set of all permutations of $\{1,\ldots,\lambda\}$ denoted by $\mathcal{P}(\lambda)$, we have that

$$\frac{1}{2\lambda}\ln\left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i[\boldsymbol{Z}_{1:i}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \boldsymbol{Z}_{1:i}\|^2}{d}\right] = \frac{1}{2\lambda}\sum_{\pi\in\mathcal{P}(\lambda)} \ln\left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \boldsymbol{\mathcal{N}}^{\pi(i)}\|^2}{d}\right] \cdot 1_{\{h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(1)}) \le \ldots \le h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(\lambda)})\}} \quad . \quad (8)$$

For any permutation $\pi \in \mathcal{P}(\lambda)$ and any $1 \le i \le \lambda$

$$h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(i)}) = 2[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma}{d}\|\boldsymbol{\mathcal{N}}^{\pi(i)}\|^2$$

such that $\lim_{d\to\infty} h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(i)}) = 2[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \sigma$. Therefore,

$$1_{\{h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(1)}) \le \ldots \le h_{\sigma/d}(\boldsymbol{\mathcal{N}}^{\pi(\lambda)})\}} \xrightarrow[d\to\infty]{} 1_{\{[\boldsymbol{\mathcal{N}}^{\pi(1)}]_1 \le \ldots [\boldsymbol{\mathcal{N}}^{\pi(\lambda)}]_1\}} \quad . \quad (9)$$

In addition, since every component of the vector $\sum_{i=1}^{\mu} w_i \boldsymbol{\mathcal{N}}^{\pi(i)}$ follows a standard normal distribution with mean zero and variance $\sum_{i=1}^{\mu} w_i^2$, we have by the Law of Large Numbers that $\|\sum_{i=1}^{\mu} w_i \boldsymbol{\mathcal{N}}^{\pi(i)}\|^2/d$ converges to $\sum_{i=1}^{\mu} w_i^2$ and thus

$$\frac{d}{2}\ln\left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \boldsymbol{\mathcal{N}}^{\pi(i)}\|^2}{d}\right] \xrightarrow[d\to\infty]{} \sigma\sum_{i=1}^{\mu} w_i[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 \quad . \quad (10)$$

Injecting the limits from (9) and (10) into (8), we obtain

$$\frac{d}{2\lambda}\ln\left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i[\boldsymbol{Z}_{1:i}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \boldsymbol{Z}_{1:i}\|^2}{d}\right] \xrightarrow[d\to\infty]{} \sum_{\pi\in\mathcal{P}(\lambda)} \frac{1}{\lambda}\left(\sigma\sum_{i=1}^{\mu} w_i[\boldsymbol{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2\right) \cdot 1_{\{[\boldsymbol{\mathcal{N}}^{\pi(1)}]_1 \le \ldots [\boldsymbol{\mathcal{N}}^{\pi(\lambda)}]_1\}} . \quad (11)$$

In the RHS of the previous equation, we recognize the distribution of order statistics of standard normal distributions. Thus,

$$\frac{d}{2\lambda}\ln\left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i[\boldsymbol{Z}_{1:i}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \boldsymbol{Z}_{1:i}\|^2}{d}\right] \xrightarrow[d\to\infty]{} \frac{1}{\lambda}\left(\sigma\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda} + \frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2\right) \quad . \quad (12)$$

To find the announced result, we need to obtain the limit in expectation. To do so we need to verify that the random variables are uniformly integrable. For this quite technical step we refer to [8]. □

The asymptotic "classical" progress rate φ of the $(\mu/\mu_\text{w}, \lambda)$-ES is derived from an approximation of $1 - \|\boldsymbol{X}_{k+1}\|/\|\boldsymbol{X}_k\|$ and coincides with the limit from the previous theorem

$$\varphi_{(\mu/\mu_\text{w},\lambda)}(\sigma, \boldsymbol{w}) = \lambda \lim_{d\to\infty} d \,\text{CR}\left(\frac{\sigma}{d}, w\right) \quad .$$

As for the finite dimensional case, we consider the variable $\boldsymbol{y} = \sigma \cdot \boldsymbol{w} \in \mathbb{R}^\mu$ with $\sigma = \sum_{i=1}^{d} |y_i|$ and compute the optimal asymptotic convergence rate that is reached for $y_i^{\text{opt}} = -E(\mathcal{N}_{i:\lambda})$ to

$$\text{CR}^{\text{opt},\infty}_{(\mu/\mu_\text{w},\lambda)} = \max_{\boldsymbol{y}\in\mathbb{R}^\mu} \lim_{d\to\infty} d \,\text{CR}\left(\frac{\sigma}{d}, \frac{\boldsymbol{y}}{\sigma}\right) = \frac{1}{2\lambda}\sum_{i=1}^{\mu} E(\mathcal{N}_{i:\lambda})^2 \quad (13)$$

as already found in [2]. Optimal weights are proportional to y_i^{opt}, thus $w_i^{\text{opt},\infty} = -E(\mathcal{N}_{i:\lambda})/\sum_{i=1}^{\mu} |E(\mathcal{N}_{i:\lambda})|$. Whether or not *negative* weights are allowed does not effect the optimal *positive* weight values, aside from the normalization factor.

3.2 The $(\mu/\mu_\text{w}, \lambda)$-ES With Random and Selective Mirroring

Following the same approach as in the previous section, we analyze the convergence rates of the different mirroring variants first for finite dimension and then asymptotically in the dimension. We define as $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\lambda_\text{iid}})$, the vector of ordered steps to be recombined; namely for a given ES variant, the best point to be recombined (for which the highest weight will be given) is $\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{Z}_1$, the second best $\boldsymbol{X}_k + \sigma\|\boldsymbol{X}_k\|\boldsymbol{Z}_2$, ... Among the different algorithm variants, the distribution of the vector of ordered steps changes. In the sequel, we express this distribution for the ES variants with random mirroring, selective mirroring and selective mirroring with resampled length.

Selected vector for random mirroring. In random mirroring, we mirror λ_m arbitrary vectors among the λ_iid independent ones. Without loss of generality, we can mirror the λ_m last vectors. For the mirrored pairs, only the best of the two vectors is recombined. The distribution of the resulting vector of ordered steps is expressed in the following lemma:

LEMMA 2. *In the $(\mu/\mu_\text{w}, \lambda_\text{iid}+\lambda_\text{m}^\text{rand})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\lambda_\text{iid}}) = \text{argsort}\{h_\sigma(\boldsymbol{\mathcal{N}}^1), \ldots, h_\sigma(\boldsymbol{\mathcal{N}}^{\lambda_\text{iid}-\lambda_\text{m}}), \min\{h_\sigma(\boldsymbol{\mathcal{N}}^{\lambda_\text{iid}-\lambda_\text{m}+1}), h_\sigma(-\boldsymbol{\mathcal{N}}^{\lambda_\text{iid}-\lambda_\text{m}+1})\}, \ldots, \min\{h_\sigma(\boldsymbol{\mathcal{N}}^{\lambda_\text{iid}}), h_\sigma(-\boldsymbol{\mathcal{N}}^{\lambda_\text{iid}})\}\} \quad (14)$$

where $h_\sigma(\boldsymbol{x}) = 2[x]_1 + \sigma\|x\|^2$.

PROOF. Let $(\boldsymbol{\mathcal{N}}^i)_{1\leq i\leq\lambda}$ be λ independent standard multivariate normal distributions. At iteration k starting from $\boldsymbol{X}_k$, we rank the individuals $\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i$ for $1\leq i\leq\lambda_{\text{iid}}-\lambda_{\text{m}}$ and the best of the mirrored/unmirrored pairs for the λ_{m} last individuals, i.e., we rank $\|\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i\|$ for $1\leq i\leq\lambda_{\text{iid}}-\lambda_{\text{m}}$ with $\min\{\|\boldsymbol{X}_k+\sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i\|,\|\boldsymbol{X}_k-\sigma\|\boldsymbol{X}_k\|\boldsymbol{\mathcal{N}}^i\|\}$ for $i=\lambda_{\text{iid}}-\lambda_{\text{m}}+1,\ldots,\lambda_{\text{iid}}$. As in Lemma 1, we find that the ranking does not change if we normalize by $\|\boldsymbol{X}_k\|$ and if we start from $\boldsymbol{e}_1$ such that the distribution is determined by the ranking of $\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\|$ for $1\leq i\leq\lambda_{\text{iid}}-\lambda_{\text{m}}$ and $\min\{\|\boldsymbol{e}_1+\sigma\boldsymbol{\mathcal{N}}^i\|,\|\boldsymbol{e}_1-\sigma\boldsymbol{\mathcal{N}}^i\|\}$ for $i=\lambda_{\text{iid}}-\lambda_{\text{m}}+1,\ldots,\lambda_{\text{iid}}$. As in Lemma 1, we square the terms and develop them to find that the distribution is determined by the ranking according to h_σ as given in (14). □

Selected vector for selective mirroring. In selective mirroring, where we mirror the worse offspring, we need to sort the λ_{iid} offspring to determine which offspring to mirror. Let $\mathcal{Y}$ be defined as

$$\mathcal{Y}:=(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{\lambda_{\text{iid}}}):=\text{argsort}\{h_\sigma(\boldsymbol{\mathcal{N}}^1),\ldots,h_\sigma(\boldsymbol{\mathcal{N}}^{\lambda_{\text{iid}}})\}$$

where $h_\sigma(\boldsymbol{x})=2[x]_1+\sigma\|x\|^2$. Then for the worst λ_{m} vectors of $\mathcal{Y}$, we select the pair-wise best among offspring and mirrored one and we keep the other vectors unchanged:

$$\boldsymbol{Y}_i^{\text{sel}}=\boldsymbol{Y}_i, i=1,\ldots,\lambda_{\text{iid}}-\lambda_{\text{m}} \tag{15}$$

$$\boldsymbol{Y}_i^{\text{sel}}=\text{argmin}\{h_\sigma(\boldsymbol{Y}_i),h_\sigma(-\boldsymbol{Y}_i)\},\lambda_{\text{iid}}-\lambda_{\text{m}}+1\leq i\leq\lambda_{\text{iid}} \tag{16}$$

Finally, as expressed in the following lemma, the distribution of the λ_{iid} ordered steps to be recombined is the result of the sorting of the $\boldsymbol{Y}_i^{\text{sel}}$ vectors:

LEMMA 3. *In the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{\lambda_{\text{iid}}})=\text{argsort}\{h_\sigma(\boldsymbol{Y}_1^{\text{sel}}),\ldots,h_\sigma(\boldsymbol{Y}_{\lambda_{\text{iid}}}^{\text{sel}})\}\ , \tag{17}$$

where $\boldsymbol{Y}_i^{\text{sel}}$ is defined in (15) *and* (16).

PROOF. As in Lemma 1 and Lemma 2, the ranking can be done normalizing by $\boldsymbol{X}_k$ and starting from $\boldsymbol{e}_1$. Thus, it follows from the way we have defined $\boldsymbol{Y}_i^{\text{sel}}$ that the distribution of the vector of ordered steps is determined by (17). □

Selected vector for selective mirroring with resampled length. The selective mirroring with resampled length algorithm differs from the previous one for the mirroring step in that only the direction is kept and the length is independently resampled according to its original χ-distribution with d degrees of freedom. Assuming that sorting of the λ_{iid} offspring has been made according to $\mathcal{Y}$ as described above, the $\boldsymbol{Y}^{\text{sel}}$ vector is given by

$$\boldsymbol{Y}_i^{\text{sel}}=\boldsymbol{Y}_i \text{ for } i=1,\ldots,\lambda_{\text{iid}}-\lambda_{\text{m}} \tag{18}$$

and for $i=\lambda_{\text{iid}}-\lambda_{\text{m}}+1,\ldots,\lambda_{\text{iid}}$,

$$\boldsymbol{Y}_i^{\text{sel}}=\text{argmin}\left\{h_\sigma(\boldsymbol{Y}_i),h_\sigma\left(-\|\widetilde{\boldsymbol{\mathcal{N}}}^i\|\frac{\boldsymbol{Y}_i}{\|\boldsymbol{Y}_i\|}\right)\right\} \tag{19}$$

where $\widetilde{\boldsymbol{\mathcal{N}}}^i$ are independent vectors following a standard multivariate normal distribution. As for the previous algorithm, the distribution of the λ_{iid} ordered steps to be recombined is the result of the sorting of the $\boldsymbol{Y}_i^{\text{sel}}$ vectors:

LEMMA 4. *In the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{\lambda_{\text{iid}}})=\text{argsort}\{h_\sigma(\boldsymbol{Y}_1^{\text{sel}}),\ldots,h_\sigma(\boldsymbol{Y}_{\lambda_{\text{iid}}}^{\text{sel}})\}\ , \tag{20}$$

where $\boldsymbol{Y}_i^{\text{sel}}$ is defined in (18) *and* (19).

PROOF. As in Lemma 1 and Lemma 2, the ranking can be done normalizing by $\boldsymbol{X}_k$ and starting from $\boldsymbol{e}_1$. Thus, it follows from the way we have defined $\boldsymbol{Y}_i^{\text{sel}}$ that the distribution of the vector of ordered steps is determined by (20). □

Similarly as for the $(\mu/\mu_{\text{w}},\lambda)$-ES, we find that the convergence rate of the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{rand}})$-ES and the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES with and without resampled mutation lengths can be expressed in the following way

THEOREM 3. *The convergence rate of the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{rand}})$-ES and the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES with and without resampled mutation lengths equals*

$$\text{CR}(\sigma,\boldsymbol{w})=-\frac{E\ln\left[1+2\sum_{i=1}^{\mu}\sigma w_i\left[\boldsymbol{Z}_i\right]_1+\left\|\sum_{i=1}^{\mu}\sigma w_i\boldsymbol{Z}_i\right\|^2\right]}{2(\lambda_{\text{iid}}+\lambda_{\text{m}})}$$

where $w_i\in\mathbb{R}$ and $\sum_{i=1}^{\mu}|w_i|=1$ and the distributions of the random vector $(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{\lambda_{\text{iid}}})$ are defined in Lemma 2, 3 and 4 respectively.

PROOF. The proof is similar to the proof of Theorem 1 injecting the distribution of the random vectors $(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{\lambda_{\text{iid}}})$ for the different algorithms. □

As for the $(\mu/\mu_{\text{w}},\lambda)$-ES, optimal convergence rates are solutions of the maximization problem

$$\max_{\boldsymbol{y}\in\mathbb{R}^\mu}\text{CR}\left(\sum_{i=1}^{d}|y_i|,\frac{\boldsymbol{y}}{\sum_{i=1}^{d}|y_i|}\right) \tag{21}$$

with CR from Theorem 3.

3.2.1 Asymptotic Results

We investigate the limit of the convergence rate given in Theorem 3 when the dimension goes to infinity. For the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{rand}})$-ES, we define the random vector $(Z_1,\ldots,Z_{\lambda_{\text{iid}}})\in\mathbb{R}^{\lambda_{\text{iid}}}$ as

$$(Z_1,\ldots,Z_{\lambda_{\text{iid}}})=\text{argsort}\{\mathcal{N}^1,\ldots,\mathcal{N}^{\lambda_{\text{iid}}-\lambda_{\text{m}}},-|\mathcal{N}^{\lambda_{\text{iid}}-\lambda_{\text{m}}+1}|,\ldots,-|\mathcal{N}^{\lambda_{\text{iid}}}|\} \tag{22}$$

where the $\mathcal{N}^i$ are λ_{iid} independent standard normal distributions. For the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES with or without resampled lengths, we define the vector

$$(Z_1,\ldots,Z_{\lambda_{\text{iid}}})=\text{argsort}\{Y_1,\ldots,Y_{\lambda_{\text{iid}}-\lambda_{\text{m}}},-|Y_{\lambda_{\text{iid}}-\lambda_{\text{m}}+1}|,\ldots,-|Y_{\lambda_{\text{iid}}}|\} \tag{23}$$

where $(Y_1,\ldots,Y_{\lambda_{\text{iid}}})=\text{argsort}\{\mathcal{N}^1,\ldots,\mathcal{N}^{\lambda_{\text{iid}}}\}$. The asymptotic convergence rate for different variants is given in the following theorem.

THEOREM 4. *The convergence rate of the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{rand}})$-ES and the $(\mu/\mu_{\text{w}},\lambda_{\text{iid}}+\lambda_{\text{m}}^{\text{sel}})$-ES (with or without resampled lengths) with scale-invariant step-size and weights*

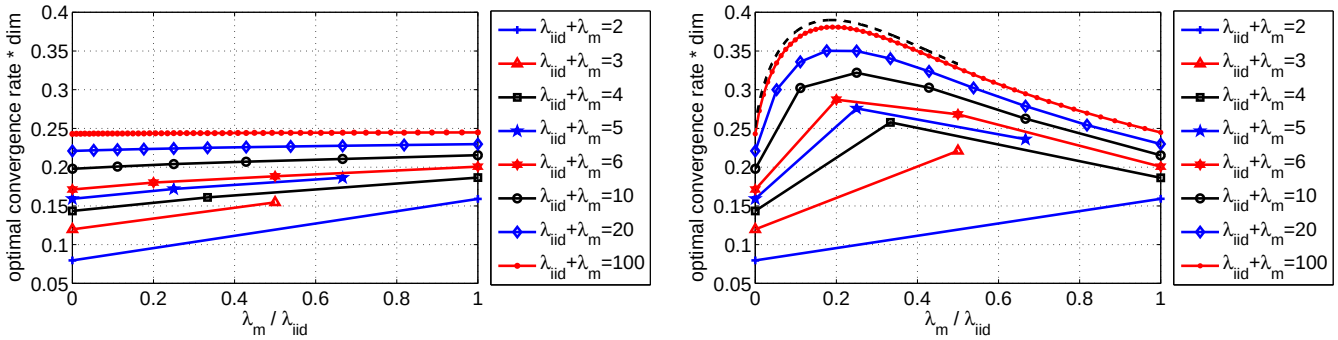

Figure 3: Optimal asymptotic convergence rates $\mathrm{CR}^{\mathrm{opt},\infty}(\lambda_{\mathrm{iid}}, \lambda_{\mathrm{m}})$, Equation (24), for the $(\mu/\mu_{\mathbf{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}})$-ES versus the ratio $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}}$ of mirrored and independent offspring for various $\lambda = \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}$. Left: $(\mu/\mu_{\mathbf{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{rand}})$-ES with random mirroring. Right: $(\mu/\mu_{\mathbf{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{sel}})$-ES with selective mirroring, mirroring the worst λ_{m} from λ_{iid} independent offspring. In addition, the righthand plot shows the theoretical result for $\lambda_{\mathrm{iid}} \to \infty$ of Equation (25) as a dashed line.

$\boldsymbol{w} \in \mathbb{R}^{\mu}$ with $\sum_{i=1}^{\mu} |w_i| = 1$ on the class of spherical functions $g(\|x\|), g \in \mathcal{M}$ satisfies

$$\lim_{d\to\infty} d\,\mathrm{CR}\left(\frac{\sigma}{d}, w\right) = \frac{-1}{\lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}} \left(\frac{\sigma^2}{2} \sum_{i=1}^{\mu} w_i^2 + \sigma \sum_{i=1}^{\mu} w_i E(Z_i) \right)$$

where the distribution of Z_i is given in (22) for the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{rand}})$-ES and in (23) for the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{sel}})$-ES with or without resampled lengths for the mirroring.

Proof. The proof follows the same lines as the proof for the $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES (Theorem 2). The limit is the same for selective mirroring with or without resampled length because asymptotically $\|\boldsymbol{\mathcal{N}}\|/\|\boldsymbol{\mathcal{N}}'\|$ goes to one when d goes to infinity for any two standard multivariate normal distribution $\boldsymbol{\mathcal{N}}$ and $\boldsymbol{\mathcal{N}}'$. □

Similarly to the $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES case, we find that the optimal convergence rate is given by

$$\mathrm{CR}^{\mathrm{opt},\infty}(\lambda_{\mathrm{iid}}, \lambda_{\mathrm{m}}) = \frac{1}{2(\lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}})} \sum_{i=1}^{\mu} E(Z_i)^2 \ , \qquad (24)$$

and the optimal weights equal $w_i^{\mathrm{opt}} = -E(Z_i)/\sum_{i=1}^{\mu} |E(Z_i)|$. We remark that the asymptotic convergence rate for the selective mirroring is the same with or without resampled lengths for the mirroring vectors. Thus, the resampling of lengths can only affect finite dimensional results.

We conclude this paragraph with a conjecture on an expression for the optimal asymptotic convergence rate of the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}})$-ES as a function of $\lambda_{\mathrm{iid}}/\lambda_{\mathrm{m}}$.

Conjecture 1. *The optimal d-asymptotic convergence rate of the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{sel}})$-ES with selective mirroring and positive recombination weights in the limit for $\lambda_{\mathrm{iid}} \to \infty$ and for $r = \lim_{\lambda_{\mathrm{iid}}\to\infty} \lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}} \le 1/2$ is given by*

$$\begin{aligned} \mathrm{CR}^{\mathrm{opt},\infty,\infty}_{\mathrm{sel\,mirr}}(r) &= \frac{1}{2}\frac{1}{1+r}\left(\frac{1}{2} + \int_{-\infty}^{G^{-1}(r)} x^2 g(x)dx\right) \qquad (25) \\ &= \frac{1}{2}\frac{1}{1+r}\left(\frac{1}{2} + r\underbrace{-\,G^{-1}(r)g(G^{-1}(r))}_{>0 \text{ for } 0<r<1/2}\right) \end{aligned}$$

where g and G are the pdf and cdf of the standard normal distribution respectively.

The $\mathrm{CR}^{\mathrm{opt},\infty,\infty}_{\mathrm{sel\,mirr}}$ is shown as the top (dashed) graph in Fig. 3, right. $\mathrm{CR}^{\mathrm{opt},\infty,\infty}_{\mathrm{sel\,mirr}}$ of 0 and 1/2 compute to 1/4 and 1/3 respectively and its unique maximum in $[0; 1/2]$ can be found for $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}} \to r = 0.188566 \pm 10^{-6}$ as 0.390015661 ± 10^{-9}.

4. SIMULATIONS

Due to their implicit nature, some of the above derived optimal convergence rates are difficult to compare directly. However, we can easily estimate the rates by means of Monte Carlo sampling allowing us to compare the performance of the proposed algorithms in infinite and finite dimension. Moreover, we find the optimal ratio of mirrored offspring and, theoretically and on spherical functions, the fastest algorithm. All convergence rates are estimated with only positive recombination weights and 10^6 samples are used for each combination of λ_{iid} and λ_{m} and for each algorithm. The MATLAB code is available at `http://canadafrance.gforge.inria.fr/mirroring/`.

Random and Selective Mirroring in Infinite Dimension. Figure 3 shows estimated optimal convergence rates versus the ratio of mirrored offspring. In all cases, the convergence rate monotonically improves with increasing number of offspring λ. For random mirroring (left subfigure), the convergence rate also increases monotonically with the number of mirrored offspring λ_{m}. The optimal ratio $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}}$ is therefore one. With an increasing number of offspring λ, the convergence rate however approaches 0.25 for any ratio $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}}$.

The results look quite different with *selective* mirroring (right subfigure): for $\lambda_{\mathrm{m}} \in \{0, \lambda_{\mathrm{iid}}\}$, selective mirroring cannot have any effect, but for any $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{iid}}$ the convergence rate is consistently better than with random mirroring and has a unique optimum slightly below $\lambda_{\mathrm{m}} = \lambda_{\mathrm{iid}}/5$.

Figure 4 shows the best convergence rates from Fig. 3 plotted versus $\lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}$ together with the corresponding optimal ratio $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}}$ for selective mirroring. For random mirroring, the known limit convergence rate for $\lambda \to \infty$ with $\lambda_{\mathrm{m}} \in \{0, \lambda_{\mathrm{iid}}\}$ is 0.25 (for $\lambda_{\mathrm{m}} = \lambda_{\mathrm{iid}}$ this follows immediately from the optimal value of 0.5 with negative recombination weights). For selective mirroring the limit is close to 0.39 with $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}} \approx 0.19$ (see Conjecture 1 above). Note that the unsmoothness of the ratio stems from discretization: not all values for the ratio of $\lambda_{\mathrm{m}}/\lambda_{\mathrm{iid}}$ are possible which in partic-

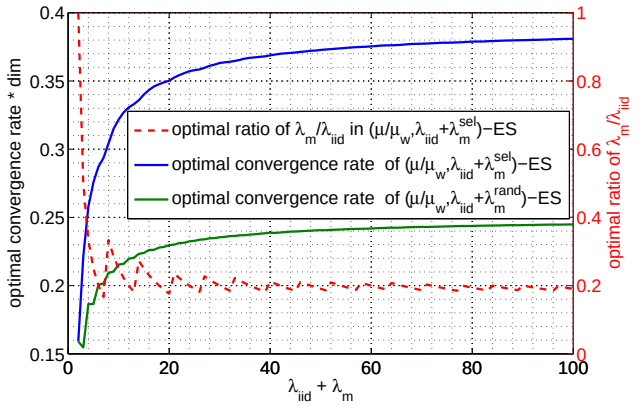

Figure 4: Extracted normalized optimal convergence rates (solid lines) for the $(\mu/\mu_{\mathbf{w}}, \lambda_{\mathrm{iid}}+\lambda_{\mathrm{m}}^{\mathrm{rand}})$-ES (bottom line) and $(\mu/\mu_{\mathbf{w}}, \lambda_{\mathrm{iid}}+\lambda_{\mathrm{m}}^{\mathrm{sel}})$-ES (top line) of Fig. 3 for different numbers of offspring $\lambda = \lambda_{\mathrm{iid}}+\lambda_{\mathrm{m}}$ together with the corresponding optimal ratio of mirrored and unmirrored offspring for the selective mirroring variant (dashed).

ular has an effect for small λ_{iid}.

5. SUMMARY AND CONCLUSION

We have introduced mirrored sampling in ESs with multi-recombination. Two important tricks are used: *selective mirroring* where only the worst λ_{m} offspring are mirrored and *pairwise selection* where at most one offspring from any mirrored couple is selected for recombination. Less importantly, the length of mirrored vectors might be resampled. Obtained theoretical results support the effectiveness of selective mirroring in particular: the new algorithm improves the known convergence rate record for ESs with positive recombination weights by 56% from 0.25 to 0.39. This is a huge improvement and the new $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}}+\lambda_{\mathrm{m}})$-ES, where $\lambda_{\mathrm{m}} \approx 0.19\lambda_{\mathrm{iid}}$, is also more than 60% faster than the fastest single-parent mirroring $(1{+}1_{\mathrm{ms}})$-ES and almost twice as fast as the regular (1+1)-ES in the asymptotic limit, cp. [3].

Only strategies with negative recombination weights are known to realize larger convergence rates, up to 0.5, cp. [2]. Negative weights however have the disadvantage that they use points for recombination that have never been—even remotely—evaluated, that is, they rely on quite specific properties of the fitness function. Compared to the strategy with optimal positive *and negative* recombination weights, the optimal $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}}+\lambda_{\mathrm{m}})$-ES loses out in two ways. About 19% additional offspring are evaluated—with negative weights they are simply used without being evaluated. These additional evaluations lead to a maximal loss of $1-1/1.19 = 16\%$ convergence speed. About $0.3\lambda_{\mathrm{iid}}$ offspring are entirely disregarded for recombination—they have small negative weights otherwise. From the overall loss of $(0.5-0.39)/0.5 = 22\%$ we can imply that the latter disregard contributes with a loss of 6% in convergence speed.

In preliminary experiments, not shown in this paper, mirrored sampling applied in CMA-ES, using the default recombination weights, improves the convergence speed in small populations, while its effect in large populations is almost negligible[6]. This is not surprising, as with $\lambda \gg d$ mirroring becomes much less effective, because offspring similar to the mirrored ones are already present in the population. Additionally, the reason to apply large populations is not to achieve faster convergence rates[7].

Considering that small populations are the default setting and that mirrored sampling is simple and has the potential to be cleverly exploited for the covariance matrix update, mirrored sampling might become a future standard method in practice.

6. REFERENCES

[1] D.V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms (FOGA 2005)*, pages 215–237. Springer Verlag, 2005.

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

[3] A. Auger, D. Brockhoff, and N. Hansen. Analyzing the impact of mirrored sampling and sequential selection in elitist evolution strategies. In *Foundations of Genetic Algorithms (FOGA 2011)*. ACM, 2011. to appear.

[4] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In *Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[5] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer-Verlag, 2001.

[6] D. Brockhoff, A. Auger, N. Hansen, D. V. Arnold, and T. Hohm. Mirrored sampling and sequential selection for evolution strategies. In *Parallel Problem Solving from Nature (PPSN XI)*, pages 11–21. Springer, 2010.

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

[8] M. Jebalia and A. Auger. Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-ES and optimal μ for intermediate recombination for large population sizes. Research Report RR-7275, INRIA, june 2010.

[9] A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaption based on non-local use of selection information. In *Conference on Problem Solving From Nature (PPSN III)*, pages 189–198, 1994.

[10] G. Rudolph. *Convergence Properties of Evolutionary algorithms*. Verlag Dr. Kovac, Hamburg, 1997.

[11] H.-P. Schwefel. *Evolution and Optimum Seeking*. Sixth-Generation Computer Technology Series. John Wiley & Sons, Inc., New York, 1995.

[12] 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 *Conference on Parallel Problem Solving from Nature (PPSN IX)*, pages 32–41. Springer, 2006.

[6] A strong adverse effect that was first observed on a single multimodal function was due to a too small evaluation budget and vanished under more appropriate experimental conditions.

[7] Note however that the effectiveness of mirroring does not imply that the ES operates only in a local neighborhood predominated by linear terms in the fitness function expansion.