Mirrored Sampling and Sequential Selection for Evolution Strategies

Dimo Brockhoff, Anne Auger, Nikolaus Hansen, Dirk Arnold, Tim Hohm

Mirrored Sampling and Sequential Selection for Evolution Strategies

Dimo Brockhoff[1], Anne Auger[1], Nikolaus Hansen[1], Dirk V. Arnold[2], and Tim Hohm[3]

[1] TAO Team, INRIA Saclay, LRI Paris Sud University, 91405 Orsay Cedex, France
firstname.lastname@inria.fr
[2] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1W5
dirk@cs.dal.ca
[3] Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland
tim.hohm@unil.ch

Abstract. This paper reveals the surprising result that a single-parent non-elitist evolution strategy (ES) can be locally faster than the (1+1)-ES. The result is brought about by *mirrored sampling* and *sequential selection*. With mirrored sampling, two offspring are generated symmetrically or *mirrored* with respect to their parent. In sequential selection, the offspring are evaluated sequentially and the iteration is concluded as soon as one offspring is better than the current parent. Both concepts complement each other well. We derive exact convergence rates of the $(1, \lambda)$-ES with mirrored sampling and/or sequential selection on the sphere model. The log-linear convergence of the ES is preserved. Both methods lead to an improvement and in combination the (1,4)-ES becomes about 10% faster than the (1+1)-ES. Naively implemented into the CMA-ES with recombination, mirrored sampling leads to a bias on the step-size. However, the (1,4)-CMA-ES with mirrored sampling and sequential selection is unbiased and appears to be faster, more robust, and as local as the (1+1)-CMA-ES.

1 Introduction

Evolution strategies (ESs) are robust stochastic search algorithms designed to minimize objective functions f that map a continuous search space $\mathbb{R}^d$ into $\mathbb{R}$. The $(1, \lambda)$-ES is a non-elitist and rather local search algorithm where 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 $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^i$, i.e., the solution with the lowest objective function value, is *selected* to become the next parent $\boldsymbol{X}_{k+1}$. The elitist version of this algorithm, the $(1 + \lambda)$-ES, selects $\boldsymbol{X}_{k+1}$ as the best among the λ offspring *and* the parent $\boldsymbol{X}_k$.

The (1+1)-ES is arguably the most local, and the locally fastest, variant of an evolution strategy. In a local search scenario, the (1+1)-CMA-ES outperforms its non-elitist counterparts typically by a factor of about 1.5 [10]. Also in the BBOB-2009 benchmarking exercise[4], the (1+1)-CMA-ES, restarted many times, performed surprisingly well on two highly multi-modal functions with weak overall structure (f_{21} and f_{22}).

[4] http://coco.gforge.inria.fr/doku.php?id=bbob-2009

However, we regard elitist selection generally as less robust, as for instance witnessed by its poor performance on the BBOB-2009 noisy testbed [5] (a single outlier fitness measurement can survive for an arbitrarily long time) or its failure on the attractive sector function f_6. Therefore, we pursue the objective to construct local non-elitist ESs with a convergence speed competitive to the (1+1)-ES and without the disadvantages of elitist selection. This is achieved by derandomization of random samples and a greedy acceptance mechanism in the $(1, \lambda)$-ES with (very) small λ.

Derandomization of random numbers has been previously introduced as antithetic variables for isotropic samples [11] and for the CMA-ES by replacing the sequence of uniform random numbers used for sampling a multivariate normal distribution by scrambling-Halton and Sobol sequences [3, ref. [27]]. However, both approaches can introduce a bias on the step-size update as we will discuss later.

Objectives of this paper. In this paper we present the concepts of mirrored (derandomized, antithetic) sampling and sequential selection within evolution strategies. We derive theoretical results on their convergence rates. We discuss their implementation into CMA-ES, in particular with respect to the question of an unbiased step-size, and present some empirical performance results.

2 Mirrored Sampling and Sequential Selection

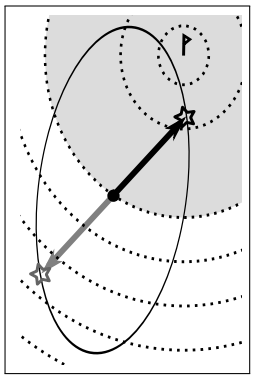
In this section, we present the concepts of mirrored samples and sequential selection, which we have recently benchmarked in the special case of the (1,2)- and the (1,4)-CMA-ES [3, ref. [3–10]]. Here, we describe both concepts for the $(1 \overset{+}{,} \lambda)$-ES.

Mirrored sampling uses a single random vector instantiation to create two offspring, one by adding and the other by subtracting the vector. In **Fig. 1**, the $(1, \lambda_\mathrm{m})$-ES is given, but mirrored sampling is entirely independent of the selection scheme.

We denote by $\boldsymbol{X}_k$ the parent at iteration k and consider the $(1 \overset{+}{,} \lambda_\mathrm{m})$-ES with even λ. In each iteration k, we sample $\lambda/2$ random vectors $(\boldsymbol{\mathcal{N}}_k^{2i-1})_{1 \le i \le \lambda/2}$. A given vector $\boldsymbol{\mathcal{N}}_k^{2i-1}$ is used for two offspring that equal $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^{2i-1}$ and $\boldsymbol{X}_k - \boldsymbol{\mathcal{N}}_k^{2i-1}$. They are thus *mirrored* or *symmetric* with respect to the parent $\boldsymbol{X}_k$. For odd λ, every other iteration, the first offspring uses the mirrored last vector from the previous iteration, see j in Fig. 1. Consequently, in the (1+1$_\mathrm{m}$)-ES, a mirrored sample is used if and only if the iteration index is even. Note that in the $(1 \overset{+}{,} \lambda_\mathrm{m})$, two mirrored offspring are entirely dependent and, in a sense, complementary, similarly to antithetic variables for Monte-Carlo numerical integration [3, ref. [14]].

Mirrored sampling has also been used in an attempt to increase the robustness of Evolutionary Gradient Search (EGS) [1]. In contrast to its use here, its utility in EGS lies in the ability to compute a stochastic gradient approximation by means of finite differences that do not involve the (possibly noisy) fitness value of a single parental solution. With a large sample size, the use of mirrored samples also increases the rate of convergence of EGS on the sphere model.

Sequential selection. Evaluating a sampled solution and its mirrored counterpart can result in unnecessary function evaluations: on unimodal objective functions with convex sub-level sets, $\{x \mid f(x) \le c\}$ for $c \in \mathbb{R}$, such as the sphere function, $f(x) = \|x\|^2$,



given: $\boldsymbol{X}_k \in \mathbb{R}^d, j \in \mathbb{N}, \lambda \in \mathbb{N}^+, f : \mathbb{R}^d \to \mathbb{R}$
$i \leftarrow 0$
while $i < \lambda$ **do**
  $i \leftarrow i + 1, j \leftarrow j + 1$
  if *mirrored sampling* and $j \equiv 0 \pmod 2$ **then**
    $\boldsymbol{X}_k^i = \boldsymbol{X}_k - \boldsymbol{\mathcal{N}}_k^{i-1}$ use previous sample
  else
    $\boldsymbol{X}_k^i = \boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^i$
  if *sequential selection* and $f(\boldsymbol{X}_k^i) < f(\boldsymbol{X}_k)$ **then**
    $j \leftarrow 0$ start with a new sample in the next iteration
    break;
end while
return $\boldsymbol{X}_{k+1} = \mathrm{argmin}\{f(\boldsymbol{X}_k^1), \ldots, f(\boldsymbol{X}_k^i)\}$

Fig. 1. Left: If for a unimodal function with convex sub-level sets, a sampled solution is better than its parent (dark arrow into shaded region of better objective function values), the mirrored one (gray) is always worse. **Right**: Pseudocode for one iteration step of mirrored sampling and sequential selection, returning the new parent $\boldsymbol{X}_{k+1}$. $\boldsymbol{\mathcal{N}}_{k+1}^0 = \boldsymbol{\mathcal{N}}_k^\lambda$ and before the first iteration, j is even. The pseudocode captures all combinations with/without mirrored sampling and/or sequential selection. The last line depicts comma-selection but can be replaced by plus selection

the mirrored solution $\boldsymbol{X}_k - \boldsymbol{\mathcal{N}}$ must be worse than the parent $\boldsymbol{X}_k$, if $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}$ was better than $\boldsymbol{X}_k$, see Fig. 1. Sequential selection, originally introduced to save such unnecessary function evaluations, is however *independent of mirrored sampling*: 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. If the first $\lambda - 1$ offspring are worse than the parent, the original selection scheme is applied.

Sequential selection applied to $(1+\lambda)$-selection coincides with (1+1)-selection: in both cases any offspring is accepted if and only if it is better than the parent[5]. The $(1, \lambda)$-ES with sequential selection is denoted as $(1, \lambda^{\mathrm{s}})$-ES and shown in Fig. 1. Note that an alternative view of the (1,λ^{s})-ES is as (1+1)-ES that periodically replaces the parent if no improvement is found after λ candidate samples.

Combining mirrored sampling and sequential selection. As the concepts of mirrored sampling and sequential selection are independent, they can be applied simultaneously. With plus selection we obtain the (1+1$^{\mathrm{s}}_{\mathrm{m}}$)-ES, independently of λ. Compared to the (1+1$_{\mathrm{m}}$)-ES, the (1+1$^{\mathrm{s}}_{\mathrm{m}}$)-ES does not use the mirrored vector after a success. With comma selection, the resulting algorithm is denoted by $(1, \lambda^{\mathrm{s}}_{\mathrm{m}})$-ES and shown in Fig. 1. In order to profit most profoundly from the interplay of mirrored sampling and sequential selection—namely from the increased likelihood that the mirrored solution is good, if the unmirrored solution was poor—we intertwine newly sampled solutions and their mirrored versions, i.e., we evaluate the offspring in the order $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^1$, $\boldsymbol{X}_k - \boldsymbol{\mathcal{N}}_k^1$, $\boldsymbol{X}_k + \boldsymbol{\mathcal{N}}_k^3$, $\boldsymbol{X}_k - \boldsymbol{\mathcal{N}}_k^3$, ...

[5] However, the iteration counters differ and other parts of the algorithm might essentially depend on λ or the iteration counter.

3 Convergence Rates on the Sphere and Lower Bounds

In this section, we investigate theoretically the gain we can expect from mirrored samples and sequential selection on spherical functions. We are interested in convergence rates for isotropic $(1, \lambda)$-ESs with adaptive step-size where an offspring i at iteration k equals $\boldsymbol{X}_k + \sigma_k \boldsymbol{\mathcal{N}}^i$ with $\sigma_k > 0$ being the step-size. Here, $(\boldsymbol{\mathcal{N}}^i)_{1 \leq i \leq \lambda}$ will denote i.i.d. random vectors following a multivariate normal distribution with identity covariance matrix. Though (independently) sampled anew each iteration, we drop the dependency on k in the notation.

The dynamics and thus the convergence rate of a step-size adaptive ES obviously depends on the step-size rule. We will study here an (artificial) step-size setting that we call *scale-invariant step-size*, where σ_k is proportional to the distance to the optimum assumed w.l.o.g. in 0, that is $\sigma_k = \sigma \|\boldsymbol{X}_k\|$ for $\sigma > 0$. We will also explain how convergence rates with scale-invariant step-size on spherical functions relate to optimal bounds for convergence rates of general adaptive step-size ESs.

Preliminaries. The fastest convergence that can be achieved by step-size adaptive ESs is linear convergence, where the logarithm of the distance to the optimum decreases to $-\infty$ linearly like the number of function evaluations increases [3, ref. [13]]. An example of linear convergence is illustrated in **Fig. 2** for three different instances of the (1,2)- and (1,2$_\mathrm{m}$)-ESs. We now establish a formal definition of linear convergence taking into account that different numbers of evaluations are performed per iteration. Let T_k be the number of function evaluations performed until iteration k. Almost sure (a.s.) linear convergence takes place if there exists a constant $c \neq 0$, such that

$$\frac{1}{T_k} \ln \frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \to c \ \ a.s.^{6} \tag{1}$$

The convergence rate c is the slope of the curves in Fig. 2. The $(1 \overset{+}{,} \lambda)$- and $(1 \overset{+}{,} \lambda_\mathrm{m})$-ES perform λ evaluations per iteration and therefore $T_k = \lambda k$. In the sequel $\mathcal{M}$ denotes the set of functions $g : \mathbb{R} \mapsto \mathbb{R}$ that are strictly increasing.

How do we prove linear convergence for scale-invariant step-size? We explain now the main idea behind the proofs that we cannot present in detail due to space limitations but which can be found in [3]. Assume that the number of offspring per iteration is fixed to λ such that $T_k = \lambda k$. The first step of the proofs expresses the left-hand side (LHS) of (1) as a sum of k terms exploiting standard properties of the logarithm function:

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

We then exploit the isotropy of the sphere function, the isotropy of the 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 and identically distributed. A law of large numbers (LLN)[7] therefore implies that the right-hand side (RHS) of (2) converges when

[6] Literally, *convergence* of $\boldsymbol{X}_k$ takes place only if $c < 0$.

[7] This also requires verifying some technical conditions, such that the expectation and the variance of $\ln(\|\boldsymbol{X}_{i+1}\|/\|\boldsymbol{X}_i\|)$ are finite.

k goes to infinity to $E[\ln(\|\boldsymbol{X}_{i+1}\|/\|\boldsymbol{X}_i\|)]$ almost surely. For more details see [3, ref. [13]].

Convergence rate for the $(1,\lambda)$***-ES.*** Linear convergence for the $(1,\lambda)$-ES with scale-invariant step-size has been shown for instance in [4]. We restate the result while denoting the first coordinate of a vector $\boldsymbol{Z}$ by $[\boldsymbol{Z}]_1$.

Theorem 1. *For a $(1,\lambda)$-ES with scale-invariant step-size ($\sigma_k = \sigma\|\boldsymbol{X}_k\| > 0$) on the class of spherical functions $g(\|\boldsymbol{x}\|)$, $g \in \mathcal{M}$, linear convergence holds with*

$$\frac{1}{\lambda}\frac{1}{k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \xrightarrow[k\to\infty]{} \frac{1}{2}\frac{1}{\lambda}E\left[\ln\left(1+\sigma\min_{1\le i\le\lambda}\left(2[\boldsymbol{\mathcal{N}}^i]_1+\sigma\|\boldsymbol{\mathcal{N}}^i\|^2\right)\right)\right] a.s., \qquad (3)$$

where $(\boldsymbol{\mathcal{N}}^i)_{1\le i\le\lambda}$ are λ independent random vectors.

The proof follows the sketch presented above. Exploiting the isotropy of the sphere and the scale-invariant step-size rule, we find that the random variable $\|\boldsymbol{X}_{i+1}\|^2/\|\boldsymbol{X}_i\|^2$, for all i, is distributed as the random variable $Z_{(1,\lambda)} = 1+\sigma\min_{1\le i\le\lambda}(2[\boldsymbol{\mathcal{N}}^i]_1 + \sigma\|\boldsymbol{\mathcal{N}}^i\|^2)$. Applying the LLN to (2), we prove the linear convergence with convergence rate $\frac{1}{2}\frac{1}{\lambda}E[\ln(Z_{(1,\lambda)})]$.

Convergence rate for the $(1,\lambda_{\mathrm{m}})$***-ES.*** In a similar manner we derive the linear convergence for the $(1,\lambda)$-ES with mirrored samples.

Theorem 2. *For a $(1,\lambda_{\mathrm{m}})$-ES with even λ and scale-invariant step-size ($\sigma_k = \sigma\|\boldsymbol{X}_k\| > 0$) on the class of spherical functions $g(\|x\|)$, for $g \in \mathcal{M}$, linear convergence holds and*

$$\frac{1}{\lambda}\frac{1}{k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \xrightarrow[k\to\infty]{} \frac{1}{2}\frac{1}{\lambda}E\left[\ln\left(1+\sigma\min_{1\le i\le\lambda/2}\left(-2|[\boldsymbol{\mathcal{N}}^i]_1|+\sigma\|\boldsymbol{\mathcal{N}}^i\|^2\right)\right)\right] a.s. \quad (4)$$

where $(\boldsymbol{\mathcal{N}}^i)_{1\le i\le\lambda/2}$ are $\lambda/2$ independent random vectors.

The difference to the previous proof lies in the expression of the random variable $\|\boldsymbol{X}_{i+1}\|^2/\|\boldsymbol{X}_i\|^2$ equal to $Z_{(1,\lambda_{\mathrm{m}})} = 1+\sigma\min_{1\le i\le\lambda/2}\left(-2|[\boldsymbol{\mathcal{N}}^i]_1|+\sigma\|\boldsymbol{\mathcal{N}}^i\|^2\right)$ in distribution.

Convergence rate for the $(1,2^{\mathrm{s}})$***-ES.*** To tackle the convergence of algorithms with sequential selection, we need to handle the fact that T_k, the number of offspring evaluated until iteration k, is a random variable, because the number of offspring per iteration is itself not a constant but a random variable in this case. This difficulty can be solved for λ even as we illustrate for $\lambda = 2$.

Theorem 3. *For a $(1,2^{\mathrm{s}})$-ES with scale-invariant step-size ($\sigma_k = \sigma\|\boldsymbol{X}_k\| > 0$) on the class of spherical functions $g(\|\boldsymbol{x}\|)$, for $g \in \mathcal{M}$, linear convergence holds and*

$$\frac{1}{T_k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \xrightarrow[k\to\infty]{} \frac{1}{2}\frac{E\left[\ln\left(1+\sigma\left(Y_1 1_{\{Y_1<0\}}+\min(Y_1,Y_2)1_{\{Y_1\ge 0\}}\right)\right)\right]}{2-p_s(\sigma)} a.s. \qquad (5)$$

where T_k is the random variable for the number of function evaluations until iteration k, $Y_1 = 2[\boldsymbol{\mathcal{N}}^1]_1+\sigma\|\boldsymbol{\mathcal{N}}^1\|^2$, $Y_2 = 2[\boldsymbol{\mathcal{N}}^2]_1+\sigma\|\boldsymbol{\mathcal{N}}^2\|^2$ with $\boldsymbol{\mathcal{N}}^1$, $\boldsymbol{\mathcal{N}}^2$ being two independent random vectors and $p_s(\sigma) = \Pr(2[\boldsymbol{\mathcal{N}}^1]_1+\sigma\|\boldsymbol{\mathcal{N}}^1\|^2 < 0)$ corresponds to the probability that the first offspring is better than its parent.

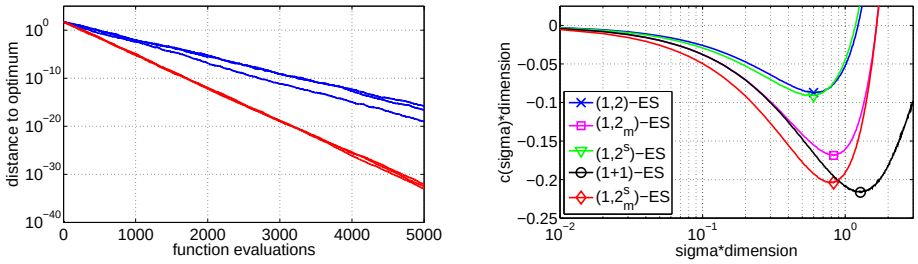


Fig. 2. Left: Evolution of distance to the optimum versus number of function evaluations for the (1,2)-ES (3 upper curves) and (1,2$_\mathrm{m}$)-ES (3 lower curves) with scale-invariant step-sizes ($d = 20$, $\sigma = 0.6/d$) on $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$; **Right**: Convergence rate $c(\sigma)$ multiplied by the dimension d versus $\sigma \cdot d$ for different algorithms with scale-invariant step-size in dimension $d = 20$. The estimated best convergence rate for each algorithm is depicted by a marker

The first step of the proof expresses the LHS of (5) as $A_k = k/T_k$ times $B_k = \frac{1}{k}\ln(\|\boldsymbol{X}_k\|/\|\boldsymbol{X}_0\|)$. Then we handle both terms separately. For B_k, we proceed as before and obtain convergence towards $\frac{1}{2}E[\ln Z_{(1,2^\mathrm{s})}]$ with $Z_{(1,2^\mathrm{s})} = 1 + \sigma\left(Y_1 1_{\{Y_1<0\}} + \min(Y_1, Y_2)\, 1_{\{Y_1\geq 0\}}\right)$. 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 $1/A_k = \frac{1}{k}\sum_{i=1}^{k}\Lambda_i$. Using the isotropy of the sphere function and the multivariate normal distribution and exploiting the scale-invariance of the step-size, we prove that Λ_i are identically distributed and independent. We can again apply the LLN and prove that $1/A_k$ converges almost surely to $E(\Lambda_1)$. Moreover, we prove that $E(\Lambda_1) = 2 - p_s(\sigma)$.

Convergence rate for the $(1, 2^\mathrm{s}_\mathrm{m})$***-ES.*** To establish the results for the (1,2)-ES with mirrored samples and sequential selection, we proceed exactly as in Theorem 3. Note that similar results can be derived for the (1,4)-ES with sequential selection [3].

Theorem 4. *For a* $(1, 2^\mathrm{s}_\mathrm{m})$*-ES with scale-invariant step-size* ($\sigma_k = \sigma\|\boldsymbol{X}_k\| > 0$) *on the sphere function* $g(\|\boldsymbol{x}\|)$*, for* $g \in \mathcal{M}$*, linear convergence holds and*

$$\frac{1}{T_k}\ln\frac{\|\boldsymbol{X}_k\|}{\|\boldsymbol{X}_0\|} \xrightarrow[k\to\infty]{} \frac{1}{2}\frac{1}{2-p_s(\sigma)} \times E\left[\ln\left(1 - 2\sigma|[\boldsymbol{\mathcal{N}}]_1| + \sigma^2\|\boldsymbol{\mathcal{N}}\|^2\right)\right] a.s. \quad (6)$$

where T_k *is the random variable for the number of function evaluations until iteration* k*,* $\boldsymbol{\mathcal{N}}$ *is a random vector following a multivariate normal distribution, and* $p_s(\sigma) = \Pr(2[\boldsymbol{\mathcal{N}}]_1 + \sigma\|\boldsymbol{\mathcal{N}}\|^2 < 0)$ *is the probability that the first offspring is successful.*

Link between convergence rates on the sphere and lower bounds for convergence. The convergence rates in (3), (4), (5) and (6) depend on σ. The RHS of Fig. 2 illustrates the dependence on σ for $\lambda = 2$. For the $(1, \lambda)$- and the $(1, \lambda_\mathrm{m})$-ES, the minimal values in σ of the RHS of (3) and (4) correspond to the fastest convergence rate that can be achieved on any function with any step-size adaptation technique. The proof is similar to the one presented in [3, ref. [13]] for the (1+1)-ES. For the $(1, \lambda^\mathrm{s})$-ES and $(1, \lambda^\mathrm{s}_\mathrm{m})$-ES, our result might be less general, but the minimal values in σ of the RHS of (5) and (6) are at least the fastest convergence rates that can be achieved on spherical functions with any step-size adaptation technique. We refer to [3] for details of the proofs.

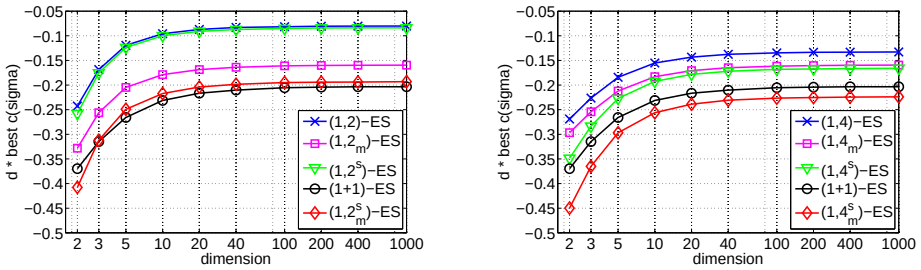


Fig. 3. Estimated optimal convergence rates on the sphere function for several algorithms with scale-invariant-constant step-size depending on the dimension d.

Numerical simulation of convergence rates. To evaluate the improvements that can be brought about 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 10^6 times the random variables inside the expectation and averaged to obtain an estimate of the convergence rate for different σ. Here, σ has been chosen such that $0.01 \leq \sigma \cdot d \leq 3$ 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 right-hand plot of Fig. 2 shows resulting convergence rate estimates versus σ in dimension 20. The step-sizes for the best measured convergence rates for the (1,2)-ESs are smaller than for the (1+1)-ES. The same is true for the (1,4)-ESs (not shown).

Fig. 3 presents the estimated best convergence rates for several algorithms for different dimensions. The strongest effect is observed from mirrored sampling in the (1,2)-ES. Only in dimension 2, the improvement is smaller than a factor of 1.5. Sequential selection alone offers little benefit for the (1,2)-ES, but the effect from mirrored sampling and sequential selection is clearly overadditive and the $(1,2^{\mathrm{s}}_{\mathrm{m}})$-ES almost achieves the progress rate of the (1+1)-ES. In the (1,4)-ES, the impact of mirrored sampling or sequential selection is similar and less than a factor of 1.5. Their combined effect is close to additive and the $(1,4^{\mathrm{s}}_{\mathrm{m}})$-ES becomes significantly faster than the (1+1)-ES.

4 Application to the CMA-ES Algorithm

We implemented *mirrored sampling* and *sequential selection* into the well-known *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES), where in addition to the step-size, the covariance matrix of the multivariate normal distribution is adapted [3, ref. [16,17,21,23]]. The additional implementational and numerical effort for the method is negligible and even fewer random numbers need to be sampled with mirrored vectors. For parent number $\mu = 1$, the implementation is straightforward in both cases. Taking $\mu > 1$ with sequential selection, the decision for when to conclude the iteration is not entirely obvious and we stick to $\mu = 1$ for sequential selection.

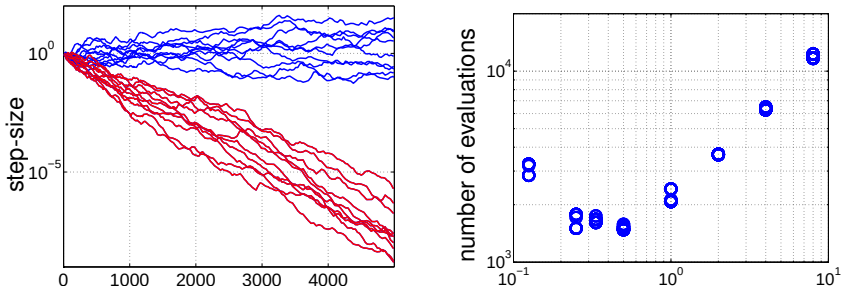


Fig. 4. Left: Step-size σ versus number of function evaluations of 20 runs on a purely random fitness function in dimension 10. The upper ten graphs show the $(5/5_{\mathrm{w}}, 10)$-CMA-ES revealing a random walk on $\log(\sigma)$. The lower ten graphs show the $(5/5_{\mathrm{w}}, 10_{\mathrm{m}})$-CMA-ES and reveal a strong bias of σ due to the recombination of mirrored vectors. **Right**: Number of function evaluations to reach function value 10^{-9} on the 20-D sphere function, versus multiplier of the default damping parameter d_σ for the $(1, 2^{\mathrm{s}}_{\mathrm{m}})$-CMA-ES starting from search point all-ones with $\sigma = 1$. Shown are three runs per d_σ-value. For smaller values of the multiplier the algorithm fails

Mirrored sampling with recombination. Taking $\mu > 1$ seems to have, a priori, no impact on the implementation of mirrored samples. Unfortunately, for $\mu > 1$, mirrored sampling introduces a strong bias on the step-size and the covariance matrix update in the $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES under neutral selection (i.e., "pure random" selection). This effect is shown in **Fig. 4**, left. The bias is due to the recombination of mirrored offspring and systematically reduces the sampling variance. The bias can facilitate premature convergence for example in a noisy selection situation and is therefore considered as undesirable [6]. On the other hand, the bias can help to focus the convergence to a single optimum in a multi-modal or rugged search landscape. We have experimented with several ways to remove the bias, but leave the question of "which way is the best" open to future work. In the following also for mirrored sampling, only $\mu = 1$ is used.

Parameter setting. We modified the damping parameter for the **step-size** to $d_\sigma = 0.3 + 2\mu_{\mathrm{w}}/\lambda + c_\sigma$. Here, $1 \le \mu_{\mathrm{w}} \le \mu$ is the effective selection mass determined by the recombination weights and therefore $\mu_{\mathrm{w}} = \mu = 1$ in our case and usually $c_\sigma \ll 1$ [7]. For a given μ_{w}, the modification introduces a dependency of d_σ on λ. The setting was found by performing experiments on the sphere function, where the convergence rate is a unimodal function of d_σ. The default d_σ was chosen, such that in all cases (a) decreasing d_σ from the default value by a factor of two led to a better performance than increasing it by a factor of two, (b) decreasing d_σ by a factor of three never led to an observed failure (this is not always achieved for $\lambda = 2$ without mirroring), and (c) the performance with d_σ was at most two times slower than the optimal performance in the tuning graph. An example of a tuning graph for the $(1, 2^{\mathrm{s}}_{\mathrm{m}})$-CMA-ES is shown in Fig. 4, right. The graph meets the specifications (a)–(c), but ideally d_σ could have been chosen almost two times smaller in this case. For λ as large as 1000 and dimension up to 5, even smaller values for d_σ are useful, but not exploited in the given default value.

For $\mu_{\mathrm{w}}/\lambda = 0.35$ and $\mu_{\mathrm{w}} \le d + 2$, where d is the dimension, the former default setting of d_σ is recovered. For a smaller ratio of μ_{w}/λ or for $\mu_{\mathrm{w}} > d+2$, the new setting allows faster changes of σ and might be harmful in a noisy or too rugged landscape. In

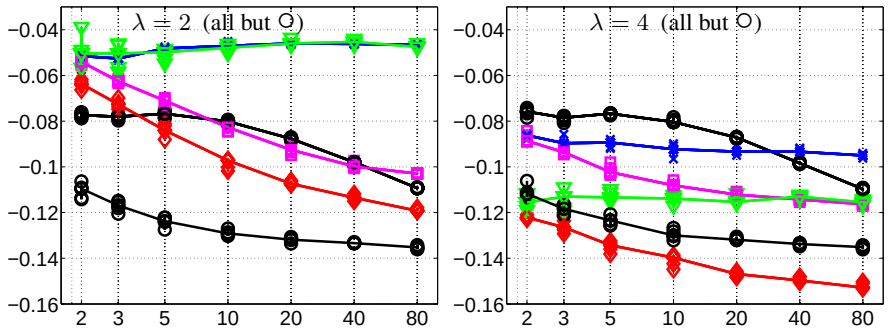


Fig. 5. Serial convergence rates $d\ln(\|\boldsymbol{X}_k\|/\|\boldsymbol{X}_0\|)/T_k$ versus dimension d of the CMA-ES on the sphere function with $\|\boldsymbol{X}_0\| = 1$, initial step-size $1/d$ and $\|\boldsymbol{X}_k\| \approx 10^{-150}$. ○: default (1+1)-CMA-ES (lower graph) and (6/6$_\mathrm{w}$,12)-CMA-ES; ×: (1,λ)-CMA-ES; ▽: sequential (1,λ^s)-CMA-ES; □: mirrored (1,λ_m)-CMA-ES; ◇: mirrored and sequential (1,$\lambda^\mathrm{s}_\mathrm{m}$)-CMA-ES. For each setting, five runs are shown and lines connect the median. Lower values are better

order to prevent a detrimental increment of the step-size for very large values of μ_w, the step-size multiplier is clamped from above at $\exp(1)$.

The learning rate for the **covariance matrix** in the CMA was originally designed for values of $\lambda \geq 5$. We rectified the learning rate of the rank-one update for small values of λ: the multiplier 2 is replaced by $\min(2, \lambda/3)$, resulting in $c_1 = \min(2, \lambda/3)/((d + 1.3)^2 + \mu_\mathrm{w})$. Similar as for the damping factor d_σ, the new value was guided by the specifications (a)–(c) from above when replacing d_σ with $1/c_1$ and optimizing the sphere function with a non-spherical initial covariance matrix and (d) the condition number of the final covariance matrix is smaller than ten. The learning rate for the rank-μ update of the covariance matrix is unchanged and zero for $\mu = 1$ [3, ref. [17,20]].

Convergence speed on the sphere. Similar to Fig. 3, we show in **Fig. 5** the convergence speed of various CMA-ES variants on the sphere function. We used `cmaes.m`, version `3.41.beta`, from `http://www.lri.fr/~hansen/cmaes_inmatlab.html` for implementing mirrored sampling and sequential selection. The resulting code is available at `http://coco.gforge.inria.fr/doku.php?id=bbob-2010-results`. In Fig. 3, the variance of the sample distribution was chosen optimal. In the CMA-ES, the covariance matrix is adapted and either *cumulative step-size adaptation* or the *1/5th success rule* is used for step-size control, in the non-elitist and the elitist variant respectively. While the overall convergence speed in moderate or large dimension is roughly two times slower than in Fig. 3, the ordering of the different variants essentially remains the same. The new sampling and selection schemes lead to a significant speedup. In low dimension, the convergence rate remains far from optimal, in accordance with observations in [2].

Experiments with BBOB-2010. The (1,2)- and the (1,4)-CMA-ES with mirrored sampling and/or sequential selection have been extensively empirically studied on 54 noisy [9] and noiseless [8] functions in the companion papers [3, ref. [3–10]]. Mirrored sampling improves the performance (number of function evaluations to reach a target function value) consistently on many functions by about a factor of two in the (1,2)-CMA-ES and by a much smaller but non-negligible factor in the (1,4)-CMA-ES. The larger factor for $\lambda = 2$ mainly reflects the comparatively poor performance of the baseline

(1,2)-selection. On the attractive sector function f_6, the performance gain is more than a factor of three even for the (1,4)-CMA-ES in dimension 20. Additional sequential selection improves the performance again on many functions, typically by 10–30% for both values of λ. Even for the (1,4)-ES, the effect of mirrored sampling is still slightly more pronounced than that of sequential selection. Overall, the $(1, 4^{\mathrm{s}}_{\mathrm{m}})$-CMA-ES is consistently faster than the (1,$2^{\mathrm{s}}_{\mathrm{m}}$)-CMA-ES. On the noisy functions, the picture is qualitatively the same. Surprisingly, the differences are not less pronounced. Even sequential selection never impairs the performance significantly. In conclusion from this rather huge benchmarking exercise, the $(1, 4^{\mathrm{s}}_{\mathrm{m}})$-CMA-ES becomes the candidate of choice to replace the (1+1)-CMA-ES as *the* fast and robust local search ES.

Acknowledgments. This work receives support by the French national research agency (ANR) within the SYSCOMM project ANR-08-SYSC-017 and within the COSINUS project ANR-08-COSI-007-12.

References

1. Arnold, D.V., Salomon, R.: Evolutionary gradient search revisited. IEEE Transactions on Evolutionary Computation 11(4), 480–495 (2007)
2. Arnold, D.V., Van Wart, D.C.S.: Cumulative step length adaptation for evolution strategies using negative recombination weights. In: Giacobini, M., et al. (eds.) EvoWorkshops. LNCS, vol. 4974, pp. 545–554. Springer (2008)
3. Auger, A., Brockhoff, D., Hansen, N.: Mirrored sampling and sequential selection for evolution strategies. Research Report RR-7249, INRIA Saclay—Île-de-France (June 2010)
4. Auger, A., Hansen, N.: Reconsidering the progress rate theory for evolution strategies in finite dimensions. In: Keijzer, et al. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006). pp. 445–452. ACM Press (2006)
5. Auger, A., Hansen, N.: Benchmarking the (1+1)-CMA-ES on the BBOB-2009 noisy testbed. In: Rothlauf, F., et al. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2009), Companion Material. pp. 2467–2472. ACM Press (2009)
6. Hansen, N.: An analysis of mutative σ-self-adaptation on linear fitness functions. Evolutionary Computation 14(3), 255–275 (2006)
7. Hansen, N.: The CMA evolution strategy: a comparing review. In: Lozano, J., Larranaga, P., Inza, I., Bengoetxea, E. (eds.) Towards a new evolutionary computation. Advances on estimation of distribution algorithms, pp. 75–102. Springer (2006)
8. Hansen, N., Finck, S., Ros, R., Auger, A.: Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Tech. Rep. RR-6829, INRIA (2009), `http://coco.gforge.inria.fr/bbob2010-downloads`, updated February 2010
9. Hansen, N., Finck, S., Ros, R., Auger, A.: Real-parameter black-box optimization benchmarking 2009: Noisy functions definitions. Tech. Rep. RR-6869, INRIA (2009), `http://coco.gforge.inria.fr/bbob2010-downloads`, updated February 2010
10. Igel, C., Suttorp, T., Hansen, N.: A computational efficient covariance matrix update and a (1+1)-CMA for evolution strategies. In: Keijzer, et al. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006). pp. 453–460. ACM Press (2006)
11. Teytaud, O., Gelly, S., Mary, J.: On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy. In: Runarsson, T., et al. (eds.) Conference on Parallel Problem Solving from Nature (PPSN IX). LNCS, vol. 4193, pp. 32–41. Springer (2006)