



**HAL**  
open science

# Mirrored Sampling and Sequential Selection for Evolution Strategies

Anne Auger, Dimo Brockhoff, Nikolaus Hansen

► **To cite this version:**

Anne Auger, Dimo Brockhoff, Nikolaus Hansen. Mirrored Sampling and Sequential Selection for Evolution Strategies. [Research Report] RR-7249, 2010. inria-00472650v1

**HAL Id: inria-00472650**

**<https://inria.hal.science/inria-00472650v1>**

Submitted on 15 May 2010 (v1), last revised 17 Jun 2010 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Mirrored Sampling and Sequential Selection for Evolution Strategies*

Anne Auger, Dimo Brockhoff, Nikolaus Hansen

**N° 7249**

April 2010

Domaine 1



*R*apport  
de recherche



## Mirrored Sampling and Sequential Selection for Evolution Strategies

Anne Auger\*, Dimo Brockhoff†, Nikolaus Hansen‡

Domaine : Mathématiques appliquées, calcul et simulation  
Équipes-Projets Thème Apprentissage et Optimisation (TAO)

Rapport de recherche n° 7249 — April 2010 — 21 pages

**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 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 they can sometimes even double the convergence rate. 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.

**Key-words:** continuous optimization, evolution strategies, derandomization, CMA-ES, theory

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.

\* INRIA Saclay—Île-de-France, projet TAO, Bat 490, Université Paris-Sud, 91405 Orsay Cedex, France, [anne.auger@inria.fr](mailto:anne.auger@inria.fr), <http://www.lri.fr/~auger/>

† INRIA Saclay—Île-de-France, projet TAO, Bat 490, Université Paris-Sud, 91405 Orsay Cedex, France, [dimo.brockhoff@inria.fr](mailto:dimo.brockhoff@inria.fr), <http://www.lri.fr/~brockho/>

‡ INRIA Saclay—Île-de-France, projet TAO, Bat 490, Université Paris-Sud, 91405 Orsay Cedex, France, [nikolaus.hansen@inria.fr](mailto:nikolaus.hansen@inria.fr), <http://www.lri.fr/~hansen/>



## 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  $\lambda$  candidate solutions, the offspring, are created from a single parent,  $\mathbf{X}_k \in \mathbb{R}^d$ . The  $\lambda$  offspring are generated by adding  $\lambda$  independent random vectors  $(\mathcal{N}_k^i)_{1 \leq i \leq \lambda}$  to  $\mathbf{X}_k$ . Then, the *best* of the  $\lambda$  offspring  $\mathbf{X}_k + \mathcal{N}_k^i$ , i.e., the solution with the lowest objective function value, is *selected* to become the next parent  $\mathbf{X}_{k+1}$ . The elitist version of this algorithm, the  $(1 + \lambda)$ -ES, selects  $\mathbf{X}_{k+1}$  as the best among the  $\lambda$  offspring *and* the parent  $\mathbf{X}_k$ .

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 1.5 [Igel et al., 2006]. Also in the BBOB-2009 benchmarking exercise<sup>1</sup>, the  $(1+1)$ -CMA-ES, restarted many times, performed surprisingly well on two highly multi-modal functions with weak overall structure ( $f_{21}$  and  $f_{22}$ ). However, we regard elitist selection generally as less robust, as for instance witnessed by its poor performance on the BBOB-2009 noisy testbed [Auger and Hansen, 2009] (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  $\lambda$ .

Derandomization of random numbers has been previously introduced for the CMA-ES [Teytaud and Gelly, 2007] by replacing the sequence of uniform random numbers used for sampling a multivariate normal distribution by scrambling-Halton and Sobol sequences. However, such an approach 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) 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 [Auger et al., 2010e,f,a,b,g,h,c,d]. Here, we describe both concepts for the  $(1 \ddagger \lambda)$ -ES.

**Mirrored sampling** uses a single random vector instantiation to create two offspring, one by adding and the other by subtracting the vector. We introduce mirrored sampling for the  $(1 \ddagger \lambda)$ -ES. In **Fig. 1**, the  $(1, \lambda_m)$ -ES is given, but mirrored sampling is entirely independent of the chosen selection scheme.

We denote by  $\mathbf{X}_k$  the parent at iteration  $k$  and consider the  $(1 \ddagger \lambda_m)$ -ES with even  $\lambda$ . In each iteration  $k$ , we sample  $\lambda/2$  random vectors  $(\mathcal{N}_k^{2i-1})_{1 \leq i \leq \lambda/2}$ . A given

<sup>1</sup><http://coco.gforge.inria.fr/doku.php?id=bbob-2009>

vector  $\mathcal{N}_k^{2i-1}$  is used for two offspring that equal  $\mathbf{X}_k + \mathcal{N}_k^{2i-1}$  and  $\mathbf{X}_k - \mathcal{N}_k^{2i-1}$ . They are thus *mirrored* or *symmetric* with respect to the parent  $\mathbf{X}_k$ . For odd  $\lambda$ , every other iteration, the first offspring uses the mirrored last vector from the last iteration, see  $j$  in Fig. 1. Consequently, in the  $(1+1_m)$ -ES, a mirrored sample is used if and only if the iteration index is even. Note that in the  $(1 \ddagger \lambda_m)$ , two mirrored offspring are entirely dependent and, in a sense, complementary, similarly to antithetic variables for Monte-Carlo numerical integration [Hammersley and Handscomb, 1967].

Mirrored sampling has also been used in an attempt to increase the robustness of Evolutionary Gradient Search (EGS) [Arnold and Salomon, 2007]. 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,  $\{x \mid f(x) \leq c\}$  for  $c \in \mathbb{R}$ , such as the sphere function,  $f(x) = \|x\|^2$ , the mirrored solution  $\mathbf{X}_k - \mathcal{N}$  must be worse than the parent  $\mathbf{X}_k$ , if  $\mathbf{X}_k + \mathcal{N}$  was better than  $\mathbf{X}_k$ , 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 immediately concluded, 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.

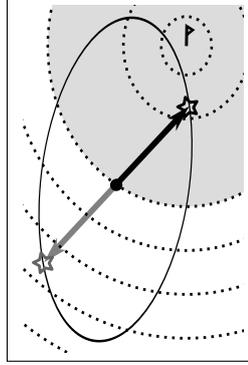
Sequential selection applied to  $(1+\lambda)$ -selection coincides with  $(1+1)$ -selection: in both cases each offspring is accepted if and only if it is better than the parent<sup>2</sup>. The  $(1, \lambda)$ -ES with sequential selection is denoted as  $(1, \lambda^s)$ -ES and shown in Fig. 1. Note that an alternative view of the  $(1, \lambda^s)$ -ES is as  $(1+1)$ -ES that periodically replaces the parent if no improvement is found after  $\lambda$  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  $(1+1_m^s)$ -ES, independently of  $\lambda$ . Compared to the  $(1+1_m)$ -ES, the  $(1+1_m^s)$ -ES does not use the mirrored vector after a success. With comma selection, the resulting algorithm is denoted by  $(1, \lambda_m^s)$ -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  $\mathbf{X}_k + \mathcal{N}_k^1$ ,  $\mathbf{X}_k - \mathcal{N}_k^1$ ,  $\mathbf{X}_k + \mathcal{N}_k^3$ ,  $\mathbf{X}_k - \mathcal{N}_k^3$ ,  $\dots$

### 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  $\mathbf{X}_k + \sigma_k \mathcal{N}^i$  with  $\sigma_k > 0$  being the step-size. In contrast to the previous

<sup>2</sup>However, the iteration counters differ and other parts of the algorithm might essentially depend on  $\lambda$  or the iteration counter.



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

Figure 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  $\mathbf{X}_{k+1}$ .  $\mathcal{N}_{k+1}^0 = \mathcal{N}_k^\lambda$  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

section, here  $(\mathcal{N}_k^i)_{1 \leq i \leq \lambda}$  will denote i.i.d. random vectors following a multivariate normal distribution.

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  $\sigma_k$  is proportional to the distance to the optimum assumed w.l.o.g. in 0, that is  $\sigma_k = \sigma \|\mathbf{X}_k\|$  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 [Auger and Hansen, 2010]. An example of linear convergence is illustrated in Fig. 2 for three different instances of the (1,2)- and (1,2<sub>m</sub>)-ESs. Since we are interested in comparing the speed of different strategies that do not use the same number of function evaluations per iteration and that even might not have a fixed number of function evaluations per iteration, we need to come up with a formal definition of linear convergence taking into account the different number of evaluations performed per iteration. Formally, 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

$$\frac{1}{T_k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} \rightarrow c \text{ a.s.}^3 \quad (1)$$

The convergence rate  $c$  is the slope of the curves in Fig. 2. The  $(1 \dagger \lambda)$ - and  $(1 \dagger \lambda_m)$ -ES perform  $\lambda$  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.

<sup>3</sup>Literally, *convergence* of  $\mathbf{X}_k$  takes place only if  $c < 0$ .

**How do we prove linear convergence for scale-invariant step-size?** We explain now the main idea behind the proofs that we detail later in Section 5. Assume that the number of offspring per iteration is fixed to  $\lambda$  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:

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

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  $\ln(\|\mathbf{X}_{i+1}\|/\|\mathbf{X}_i\|)$  are independent identically distributed. A law of large numbers (LLN)<sup>4</sup> therefore implies that the right-hand side (RHS) of (2) converges when  $k$  goes to infinity to  $E[\ln(\|\mathbf{X}_{i+1}\|/\|\mathbf{X}_i\|)]$  almost surely.

**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 [Auger and Hansen, 2006]. We restate the result while denoting the first coordinate of a vector  $\mathbf{Z}$  by  $[\mathbf{Z}]_1$ .

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

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

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

*Proof: see page 12*

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  $\|\mathbf{X}_{i+1}\|^2/\|\mathbf{X}_i\|^2$ , for all  $i$ , is distributed as the random variable  $\mathbf{Z}_{(1, \lambda)} = 1 + \sigma \min_{1 \leq i \leq \lambda} (2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2)$ . Applying the LLN to (2) we prove the linear convergence with convergence rate  $\frac{1}{2} \frac{1}{\lambda} E[\ln(\mathbf{Z}_{(1, \lambda)})]$ .

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

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

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

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

*Proof: see page 15*

The difference with the previous proof lies in the expression of the random variable  $\|\mathbf{X}_{i+1}\|^2/\|\mathbf{X}_i\|^2$  equal to  $\mathbf{Z}_{(1, \lambda_m)} = 1 + \sigma \min_{1 \leq i \leq \lambda/2} (-2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2)$  in distribution.

<sup>4</sup>Verifying some technical conditions such that the expectation and the variance of  $\ln(\|\mathbf{X}_{i+1}\|/\|\mathbf{X}_i\|)$  are finite.

**Convergence rate for the  $(1, 2^s)$ -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  $\lambda$  even as we illustrate for  $\lambda = 2$ .

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

$$\frac{1}{T_k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} \xrightarrow{k \rightarrow \infty} \frac{1}{2} \frac{E[\ln(1 + \sigma(Y_1 1_{\{Y_1 < 0\}} + \min(Y_1, Y_2) 1_{\{Y_1 \geq 0\}}))]}{2 - p_s(\sigma)} \quad a.s. \quad (5)$$

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

*Proof: see page 16*

The first step of the proof expresses the LHS of (5) as  $A_k = k/T_k$  times  $B_k = \frac{1}{k} \ln(\|\mathbf{X}_k\|/\|\mathbf{X}_0\|)$ . Then we handle both terms independently. For  $B_k$ , we proceed as before and obtain convergence towards  $\frac{1}{2} E[\ln \mathbf{Z}_{(1, 2^s)}]$  with  $\mathbf{Z}_{(1, 2^s)} = 1 + \sigma(Y_1 1_{\{Y_1 < 0\}} + \min(Y_1, Y_2) 1_{\{Y_1 \geq 0\}})$ . For the term  $A_k$ , we denote by  $\Lambda_i$  the number of offspring evaluated at iteration  $i$ . Then,  $T_k = \Lambda_1 + \dots + \Lambda_k$  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  $\Lambda_i$  are identically distributed and independent. We can again apply the LLN and prove that  $1/A_k$  converges almost surely to  $1/E(\Lambda_1)$ . Moreover, we prove that  $E(\Lambda_1) = 2 - p_s(\sigma)$ .

**Convergence rate for the  $(1, 2_m^s)$ -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 as we will see below.

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

$$\frac{1}{T_k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} \xrightarrow{t \rightarrow \infty} \frac{1}{2} \frac{1}{2 - p_s(\sigma)} \times E[\ln(1 + 2\sigma([\mathcal{N}]_1 + |[\mathcal{N}]_1|) 1_{\{2[\mathcal{N}]_1 + \sigma^2 \|\mathcal{N}\|^2 < 0\}} - 2\sigma([\mathcal{N}]_1 + \sigma^2 \|\mathcal{N}\|^2))] \quad a.s. \quad (6)$$

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

*Proof: see page 19*

**Convergence rate for the  $(1, 4^s)$ -ES with sequential selection.** The convergence rate for the case of  $\lambda = 4$  can be shown in a similar way than above:

**Theorem 5.** For a  $(1, 4^s)$ -ES with scale-invariant step-size, i.e.  $\sigma_k = \sigma \|\mathbf{X}_k\|$  on the sphere function  $g(\|x\|)$ , for  $g \in \mathcal{M}$ , log-linear convergence holds almost surely

$$\frac{1}{T} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} \xrightarrow[k \rightarrow \infty]{} \frac{1}{2} \frac{1}{E(\Lambda^s)} E \left[ \ln \left( 1 + \sigma \left( Y_1 1_{\{Y_1 \leq 0\}} + Y_2 1_{\{Y_1 \geq 0, Y_2 \leq 0\}} + Y_3 1_{\{Y_1 \geq 0, Y_2 \geq 0, Y_3 \leq 0\}} + \left( \min_{1 \leq i \leq 4} (Y_i) \right) 1_{\{Y_i \geq 0, i=1,2,3\}} \right) \right) \right], \quad (7)$$

where  $T$  is the random variable for number of function evaluations till iteration  $k$ ,  $Y^i = 2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2$ , with  $(\mathcal{N}^i)_{1 \leq i \leq 4}$  being four independent random vectors following a multivariate normal distribution and  $E(\Lambda^s)$  corresponds to expected number of offspring evaluated per iteration.

Denoting  $p_s(\sigma) = \Pr(2[\mathcal{N}]_1 + \sigma \|\mathcal{N}\|^2 \leq 0)$  the probability of success of one offspring, we have

$$E(\Lambda^s) = p_s(\sigma) + 2(1 - p_s(\sigma))p_s(\sigma) + 3(1 - p_s(\sigma))^2 p_s(\sigma) + 4(1 - p_s(\sigma))^3$$

*Proof: see page 19*

#### Convergence rate for the $(1, 4_m^s)$ -ES with sequential selection.

**Theorem 6.** For a  $(1, 4_m^s)$ -ES with scale-invariant step-size, i.e.  $\sigma_k = \sigma \|\mathbf{X}_k\|$  on the sphere function  $g(\|x\|)$ , for  $g \in \mathcal{M}$ , log-linear convergence holds almost surely

$$\frac{1}{T} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} \xrightarrow[t \rightarrow \infty]{} \frac{1}{2} \frac{1}{E(\Lambda_m^s)} E \left[ \ln \left( 1 + \sigma \left( Y_1 1_{\{Y_1 \leq 0\}} + \tilde{Y}_1 1_{\{Y_1 \geq 0, \tilde{Y}_1 \leq 0\}} + Y_3 1_{\{Y_1 \geq 0, \tilde{Y}_1 \geq 0, Y_3 \leq 0\}} + \min(Y_1, \tilde{Y}_1, Y_3, \tilde{Y}_3) 1_{\{Y_1 \geq 0, \tilde{Y}_1 \geq 0, Y_3 \geq 0\}} \right) \right) \right], \quad (8)$$

where  $T$  is the random variable for number of function evaluations till iteration  $k$ ,  $Y^i = 2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2$ ,  $\tilde{Y}^i = 2[\mathcal{N}^i]_1 - \sigma \|\mathcal{N}^i\|^2$ ,  $i = 1, 3$  and with  $(\mathcal{N}^i)_{i=1,3}$  being two independent random vectors following a multivariate normal distribution and  $E(\Lambda_m^s)$  corresponds to expected number of offspring evaluated per iteration.

*Proof: see page 19*

#### Link between convergence rates on the sphere and lower bounds for convergence.

The convergence rates in (3), (4), (5) and (6) depend on  $\sigma$ . The RHS of Fig. 2 illustrates the dependence on  $\sigma$  for  $\lambda = 2$ . For the  $(1, \lambda)$ - and the  $(1, \lambda_m)$ -ES, the minimal values in  $\sigma$  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 [Auger and Hansen, 2010] for the  $(1+1)$ -ES. For the  $(1, \lambda^s)$ -ES and  $(1, \lambda_m^s)$ -ES, our result might be less general though, but the minimal values in  $\sigma$  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.

**Numerical simulation of convergence rates.** To conclude on the improvements that can be brought by mirrored samples and sequential selection, we now need to compare the different convergence rates. However, those convergence rates are expressed only implicitly as the expectation of some random variables. We therefore simulate the

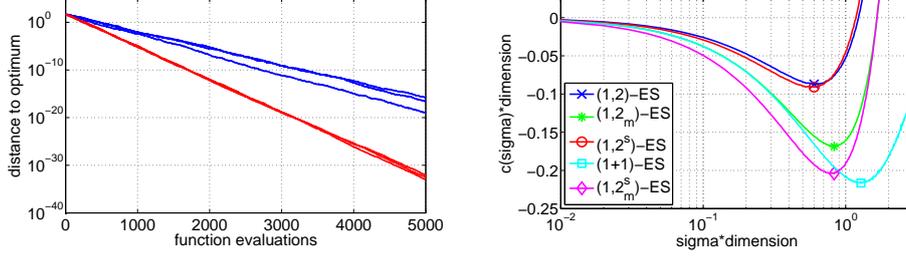


Figure 2: **Left:** Evolution of distance to the optimum versus number of function evaluations for the (1,2)-ES (3 upper curves) and (1,2<sub>m</sub>)-ES (3 lower curves) with scale-invariant step-sizes ( $d = 20$ ,  $\sigma = 0.6/d$ ) on  $f(x) = \|x\|^2$ ; **Right:** Convergence rate  $c(\sigma)$  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

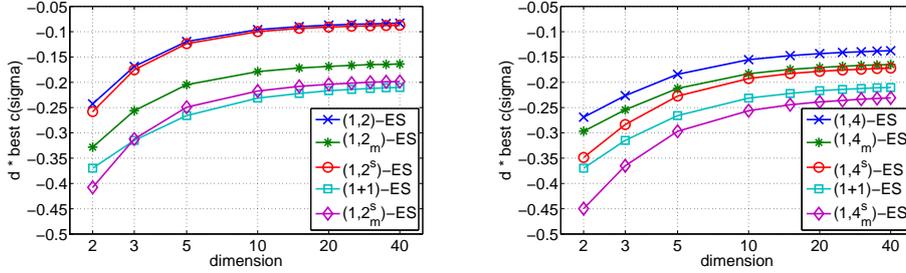


Figure 3: Estimated optimal convergence rates on the sphere function for several algorithms with scale-invariant-constant step-size depending on the dimension  $d$ . For estimating the convergence rates of Sec. 3,  $10^6$  generations/iterations have been sampled

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  $\sigma$ . Here,  $\sigma$  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  $\sigma$  in dimension 20. The step-size for the best measured convergence rate for the (1,2)-ESs is 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 on 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<sub>m</sub><sup>s</sup>)-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<sub>m</sub><sup>s</sup>)-ES becomes significantly faster than the (1+1)-ES.

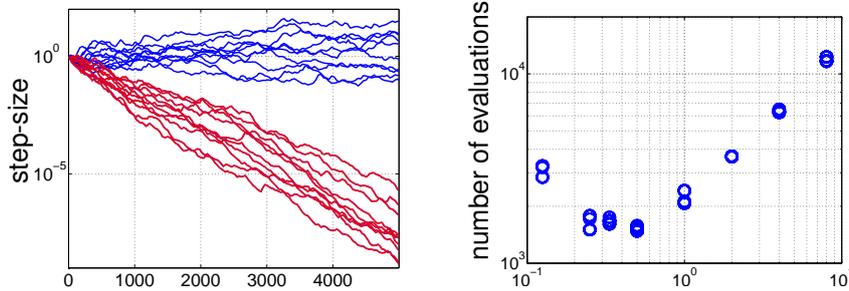


Figure 4: **Left:** Step-size  $\sigma$  versus number of function evaluations of 20 runs on a purely random fitness function in dimension 10. The upper ten graphs show the  $(5/5_w, 10)$ -CMA-ES revealing a random walk on  $\log(\sigma)$ . The lower ten graphs show the  $(5/5_w, 10_m)$ -CMA-ES and reveal a strong bias of  $\sigma$  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  $d_\sigma$  for the  $(1, 2_m^s)$ -CMA-ES starting from search point all-ones with  $\sigma = 1$ . Shown are three runs per  $d_\sigma$ -value. For smaller values of the multiplier the algorithm fails

## 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 [Hansen and Ostermeier, 2001, Hansen et al., 2003, Hansen, 2006b, Hansen et al., 2009c]. 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.

**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  $(\mu/\mu_w, \lambda)$ -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 in an ambiguous selection situation and is therefore considered as undesirable [Hansen, 2006a]. 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,  $\mu = 1$  is used.

**Parameter setting.** We modified the damping parameter for the **step-size** to  $d_\sigma = 0.3 + 2\mu_w/\lambda + c_\sigma$ . Here,  $1 \leq \mu_w \leq \mu$  is the effective selection mass determined by the recombination weights and therefore  $\mu_w = \mu = 1$  in our case and usually  $c_\sigma \ll 1$  [Hansen, 2006b]. For a given  $\mu_w$ , the modification introduces a dependency of  $d_\sigma$  on  $\lambda$ . The setting was found by performing experiments on the sphere function, where the performance is a unimodal function of  $d_\sigma$ . The default  $d_\sigma$  was chosen, such that in all cases (a) decreasing  $d_\sigma$  from the default value by a factor of two leads

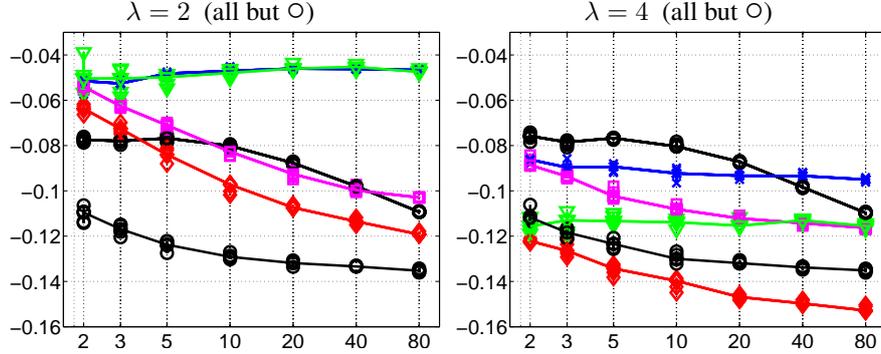


Figure 5: Serial convergence rates  $d \ln(\|\mathbf{X}_k\|/\|\mathbf{X}_0\|)/T_k$  versus dimension  $d$  of the CMA-ES on the sphere function with  $\|\mathbf{X}_0\| = 1$ , initial step-size  $1/d$  and  $\|\mathbf{X}_k\| \approx 10^{-150}$ .  $\circ$ : default (1+1)-CMA-ES (lower graph) and (6/6<sub>w</sub>,12)-CMA-ES;  $\times$ : (1,λ)-CMA-ES;  $\nabla$ : sequential (1,λ<sup>s</sup>)-CMA-ES;  $\square$ : mirrored (1,λ<sub>m</sub>)-CMA-ES;  $\diamond$ : mirrored and sequential (1,λ<sub>m</sub><sup>s</sup>)-CMA-ES. For each setting, five runs are shown and lines connect the median. Lower values are better

to a better performance than increasing it by a factor of two, (b) decreasing  $d_\sigma$  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  $d_\sigma$  is at most two times slower than the optimal performance in the tuning graph. An example of a tuning graph for the (1, 2<sub>m</sub><sup>s</sup>)-CMA-ES is shown in **Fig. 4**, right. The graph meets the specifications (a)–(c), but ideally  $d_\sigma$  could have been chosen almost two times larger in this case. For  $\lambda$  as large as 1000 and dimension up to 5, even smaller values for  $d_\sigma$  are useful, but not covered in the given default value.

For  $\mu_w/\lambda = 0.35$  and  $\mu_w \leq d + 2$ , where  $d$  is the dimension, the former default setting of  $d_\sigma$  is recovered. For a smaller ratio of  $\mu_w/\lambda$  or for  $\mu_w > d + 2$ , the new setting allows faster changes of  $\sigma$  and might then be harmful in a noisy or too rugged landscape. In order to prevent a detrimental increment of the step-size for very large values of  $\mu_w$ , 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  $\lambda$ : the multiplier 2 is replaced by  $\min(2, \lambda/3)$ , resulting in  $c_1 = \min(2, \lambda/3)/((d+1.3)^2 + \mu_w)$ . Similar as for the damping factor  $d_\sigma$ , the new value was guided by the specifications (a)–(c) from above when replacing  $d_\sigma$  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- $\mu$  update of the covariance matrix is unchanged and zero for  $\mu = 1$  [Hansen and Kern, 2004, Hansen, 2009].

**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](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 [Arnold and Wart, 2008].

**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 and noiseless functions [Hansen et al., 2009a,b] in the companion papers [Auger et al., 2010e,f,a,b,g,h,c,d]. 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 section 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  $\lambda$ . Even for the (1,4)-ES, the effect of mirrored sampling is still slightly more pronounced than that of sequential selection. Overall, the  $(1, 4_m^s)$ -CMA-ES is consistently faster than the  $(1, 2_m^s)$ -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  $(1, 4_m^s)$ -CMA-ES becomes the candidate of choice to replace the (1+1)-CMA-ES as *the* fast and robust local search ES.

## 5 Appendix

### 5.1 Proof of Theorem 1

*Proof.* In a  $(1, \lambda)$ -ES with scale-invariant step-size, the best of  $\lambda$  independent offspring is selected such that  $\|\mathbf{X}_{k+1}\|$  satisfies

$$\|\mathbf{X}_{k+1}\| = \min_{1 \leq i \leq \lambda} \{ \|\mathbf{X}_k + \sigma \|\mathbf{X}_k\| \mathcal{N}_k^i \|\} .$$

We can factor out  $\|\mathbf{X}_k\|$  and we obtain

$$\|\mathbf{X}_{k+1}\| = \|\mathbf{X}_k\| \min_{1 \leq i \leq \lambda} \{ \|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i \|\} .$$

Let us define the random variable  $Y_k$  as

$$Y_k = \min_{1 \leq i \leq \lambda} \{ \|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i \|\} ,$$

such that  $\|\mathbf{X}_{k+1}\| = \|\mathbf{X}_k\| Y_k$ . Taking the logarithm we get

$$\ln \|\mathbf{X}_{k+1}\| = \ln \|\mathbf{X}_k\| + \ln Y_k .$$

Summing up the previous equation for  $k = 0$  to  $K - 1$  and dividing by  $\lambda K$  we obtain after renaming the summation index  $i$  and the integer  $K, k$ :

$$\frac{1}{\lambda} \frac{1}{k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} = \frac{1}{\lambda} \frac{1}{k} \sum_{i=0}^{k-1} \ln Y_i \quad (9)$$

From Lemma 1,  $(Y_k)_k$  are independent and identically distributed as  $Y = \min_{1 \leq i \leq \lambda} \{\|e_1 + \sigma \mathcal{N}^i\|\}$  such that by the LLN<sup>5</sup>, the right hand side of (9) converges to  $\frac{1}{\lambda} E[\ln Y]$ . Moreover by Lemma 3, the expectation of  $\ln Y$  satisfies

$$E[\ln Y] = \frac{1}{2} E \left[ \ln \left( 1 + \sigma \min_{1 \leq i \leq \lambda} (2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2) \right) \right]. \quad (10)$$

Putting together (9) and (10) we obtain (3).  $\square$

**Lemma 1.** *Let  $Y_k = \min_{1 \leq i \leq \lambda} \{\|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|\}$ . Then  $(Y_k)_{k \in \mathbb{N}}$  are independent and identically distributed as  $Y = \min_{1 \leq i \leq \lambda} \{\|e_1 + \sigma \mathcal{N}^i\|\}$  where  $\mathcal{N}^i$  for  $i = 1, \dots, \lambda$  are  $\lambda$  independent vectors following a multivariate normal distribution with covariance matrix identity.*

*Proof.* (i) By using Lemma 2, we first prove that  $Y_k$  are identically distributed according to  $Y$  and thus for all  $t \in \mathbb{R}$  and for all  $k$ ,

$$E[e^{itY_k}] = E[e^{itY}]. \quad (11)$$

Let  $\xi_k = \sigma(\mathbf{X}_0, \mathcal{N}_0^1, \dots, \mathbf{X}_{k-1}, \mathcal{N}_{k-1}^1, \dots, \mathcal{N}_{k-1}^\lambda, \mathbf{X}_k)$  be the smallest  $\sigma$ -algebra such that  $(\mathbf{X}_0, \mathcal{N}_0^1, \dots, \mathbf{X}_{k-1}, \mathcal{N}_{k-1}^1, \dots, \mathcal{N}_{k-1}^\lambda, \mathbf{X}_k)$  is measurable, then  $E[e^{itY_k}] = E[E[e^{itY_k} | \xi_k]] = E[E[e^{it \min_{1 \leq i \leq \lambda} \{\|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|\}} | \xi_k]]$ . By independence of  $(\mathcal{N}_k^i)$  for  $1 \leq i \leq \lambda$  to the past and thus to  $\xi_k$  we have that

$$E[e^{it \min_{1 \leq i \leq \lambda} \{\|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|\}} | \xi_k] = h_t^{(1, \lambda)}(\mathbf{X}_k / \|\mathbf{X}_k\|)$$

where  $h_t^{(1, \lambda)}$  is defined in Lemma 2. Thus by Lemma 2  $E[e^{itY_k} | \xi_k] = h_t^{(1, \lambda)}(e_1) = E[e^{itY}]$ , thus (11) holds.

(ii) For showing the independence of  $(Y_k)_{k \in \mathbb{N}}$ , we will prove that for all  $k$ , for all  $t_0 \in \mathbb{R}, \dots, t_k \in \mathbb{R}$ ,  $E[e^{it_0 Y_0} \dots e^{it_k Y_k}] = E[e^{it_0 Y_0}] \dots E[e^{it_k Y_k}]$ , however thanks to (11) we only need to prove that  $E[e^{it_0 Y_0} \dots e^{it_k Y_k}] = E[e^{it_0 Y}] \dots E[e^{it_k Y}]$ . We will proceed by induction and suppose that for all  $t_0 \in \mathbb{R}, \dots, t_{k-1} \in \mathbb{R}$

$$E[e^{it_0 Y_0} \dots e^{it_{k-1} Y_{k-1}}] = E[e^{it_0 Y}] \dots E[e^{it_{k-1} Y}]$$

and prove that for all  $t_0 \in \mathbb{R}, \dots, t_k \in \mathbb{R}$

$$E[e^{it_0 Y_0} \dots e^{it_k Y_k}] = E[e^{it_0 Y}] \dots E[e^{it_k Y}]. \quad (12)$$

We can rewrite the LHS of (12) as

$$E[e^{it_0 Y_0} \dots e^{it_k Y_k}] = E[E[e^{it_0 Y_0} \dots e^{it_k Y_k} | \xi_k]].$$

Since  $e^{it_0 Y_0} \dots e^{it_{k-1} Y_{k-1}}$  is bounded and  $\xi_k$ -measurable

$$E[e^{it_0 Y_0} \dots e^{it_k Y_k} | \xi_k] = e^{it_0 Y_0} \dots e^{it_{k-1} Y_{k-1}} E[e^{it_k Y_k} | \xi_k].$$

By definition of  $Y_k$ ,  $E[e^{it_k Y_k} | \xi_k] = E[e^{it_k \min_{1 \leq i \leq \lambda} \{\|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|\}} | \xi_k]$  and by independence of  $(\mathcal{N}_k^i)_{1 \leq i \leq \lambda}$  to the past and thus to  $\xi_k$  we have that

$$E[e^{it_k \min_{1 \leq i \leq \lambda} \{\|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|\}} | \xi_k] = h_t^{(1, \lambda)}(\mathbf{X}_k / \|\mathbf{X}_k\|)$$

<sup>5</sup>We should also prove that  $Y$  is integrable, since this step is slightly technical we refer to [Jebalia et al., 2008] for the details.

where  $h_t^{(1,\lambda)}$  is defined in Lemma 2. Since the norm of the vector  $\mathbf{X}_k/\|\mathbf{X}_k\|$  is one, we know from Lemma 2 that  $E[e^{it_k Y_k}] = E[e^{it_k Y}]$  where  $Y = \min_{1 \leq i \leq \lambda} \{\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|\}$  and thus

$$E[e^{it_0 Y_0} \dots e^{it_k Y_k} | \xi_k] = e^{it_0 Y_0} \dots e^{it_{k-1} Y_{k-1}} E[e^{it_k Y}]$$

Taking the expectation of the previous equation we obtain

$$E[e^{it_0 Y_0} \dots e^{it_k Y_k}] = E[e^{it_0 Y_0} \dots e^{it_{k-1} Y_{k-1}}] E[e^{it_k Y}]$$

and thus by the induction assumption we obtain (12) and thus the independence of  $(Y_k)_{k \in \mathbb{N}}$ .  $\square$

**Lemma 2.** *Let  $t \in \mathbb{R}$  and  $h_t^{(1,\lambda)}$  be the mapping from  $\mathbb{R}^d$  to  $\mathbb{R}$  defined for all  $x \in \mathbb{R}^d$  as  $h_t^{(1,\lambda)}(x) = E \left[ e^{it \min_{1 \leq i \leq \lambda} \|x + \sigma \mathcal{N}^i\|} \right]$  where  $\mathcal{N}^i$  for  $i = 1, \dots, \lambda$  are  $\lambda$  independent random vectors following a multivariate normal distribution with covariance matrix identity. Then for all vectors  $u_1$  and  $u_2$  with  $\|u_1\| = \|u_2\| = 1$ ,  $h_t^{(1,\lambda)}(u_1) = h_t^{(1,\lambda)}(u_2)$  and thus without loss of generality for all  $u$ ,  $\|u\| = 1$ ,*

$$h_t^{(1,\lambda)}(u) = E \left[ e^{itY} \right]$$

with  $Y = \min_{1 \leq i \leq \lambda} \|\mathbf{e}_1 + \sigma \mathcal{N}^i\|$  where  $\mathbf{e}_1$  is the vector  $(1, 0, \dots, 0)$ .

*Proof.* This result is a consequence of the fact that the standard  $d$ -dimensional normal distribution is spherical. Let  $x \in \mathbb{R}^d$  and  $\|x\| = 1$ . Let  $O$  be an orthogonal matrix such that  $Ox = \mathbf{e}_1$ . Since  $O$  is an orthogonal matrix,  $\|Oy\| = \|y\|$  for all  $y \in \mathbb{R}^d$  and therefore for all  $i$   $\|x + \sigma \mathcal{N}^i\| = \|O(x + \sigma \mathcal{N}^i)\|$  almost surely. Besides,  $\|O(x + \sigma \mathcal{N}^i)\| = \|\mathbf{e}_1 + \sigma O\mathcal{N}^i\|$ . However, since  $\mathcal{N}^i$  is spherical,  $O\mathcal{N}^i$  has the same law as  $\mathcal{N}^i$  and thus  $\|\mathbf{e}_1 + \sigma O\mathcal{N}^i\|$  and  $\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|$  have the same distribution and thus  $\|x + \sigma \mathcal{N}^i\| (= \|\mathbf{e}_1 + \sigma O\mathcal{N}^i\|)$  and  $\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|$  have the same distribution. Therefore  $\min_{1 \leq i \leq \lambda} \|x + \sigma \mathcal{N}^i\|$  and  $\min_{1 \leq i \leq \lambda} \|\mathbf{e}_1 + \sigma \mathcal{N}^i\|$  have the same distribution and thus admit the same characteristic function, i.e.

$$h_t^{(1,\lambda)}(x) = E \left[ e^{itY} \right] .$$

$\square$

**Lemma 3.** *Let  $Y = \min_{1 \leq i \leq \lambda} \{\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|\}$  where  $\mathcal{N}^i$  for  $i = 1, \dots, \lambda$  are  $\lambda$  independent multivariate normal distribution with covariance matrix identity, then*

$$E[\ln Y] = \frac{1}{2} E \left[ \ln \left( 1 + \sigma \min_{1 \leq i \leq \lambda} (2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2) \right) \right] .$$

*Proof.* First we write  $\ln Y = \frac{1}{2} \ln Y^2$ . Moreover since the square function is increasing we have that  $Y^2 = \min_{1 \leq i \leq \lambda} \{\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|^2\}$ . Let us now develop  $\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|^2$ , we have

$$\|\mathbf{e}_1 + \sigma \mathcal{N}^i\|^2 = 1 + 2\sigma[\mathcal{N}^i]_1 + \sigma^2 \|\mathcal{N}^i\|^2 .$$

Moreover

$$\min_{1 \leq i \leq \lambda} 1 + 2\sigma[\mathcal{N}^i]_1 + \sigma^2 \|\mathcal{N}^i\|^2 = 1 + \min_{1 \leq i \leq \lambda} (2\sigma[\mathcal{N}^i]_1 + \sigma^2 \|\mathcal{N}^i\|^2) .$$

such that

$$\ln Y^2 = \ln \left[ 1 + \min_{1 \leq i \leq \lambda} (2\sigma[\mathcal{N}^i]_1 + \sigma^2 \|\mathcal{N}^i\|^2) \right]$$

and thus

$$\ln Y = \frac{1}{2} \ln[1 + \min_{1 \leq i \leq \lambda} (2\sigma[\mathcal{N}^i]_1 + \sigma^2 \|\mathcal{N}^i\|^2)]$$

Hence by taking the expectation of the previous equation, we obtain the result.  $\square$

## 5.2 Proof of Theorem 2

*Proof.* For a  $(1, \lambda_m)$ -ES, we have

$$\|\mathbf{X}_{k+1}\|^2 = \min_{1 \leq i \leq \lambda/2} \{ \|\mathbf{X}_k + \sigma \|\mathbf{X}_k\| \mathcal{N}_k^i\|^2, \|\mathbf{X}_k - \sigma \|\mathbf{X}_k\| \mathcal{N}_k^i\|^2 \} .$$

We can factor out  $\|\mathbf{X}_k\|$  and we obtain

$$\|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k\|^2 \min_{1 \leq i \leq \lambda/2} \{ \|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{X}_k / \|\mathbf{X}_k\| - \sigma \mathcal{N}_k^i\|^2 \} .$$

Let us define the random variable  $Y_k$  as

$$Y_k = \min_{1 \leq i \leq \lambda/2} \{ \|\mathbf{X}_k / \|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{X}_k / \|\mathbf{X}_k\| - \sigma \mathcal{N}_k^i\|^2 \},$$

such that  $\|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k\|^2 Y_k$ . Taking the logarithm we get

$$\ln \|\mathbf{X}_{k+1}\|^2 = \ln \|\mathbf{X}_k\|^2 + \ln Y_k .$$

Summing up the previous equation for  $k = 0$  to  $K - 1$  and dividing by  $\lambda K$  we obtain after renaming the summation index  $i$  and the integer  $K, k$ :

$$\frac{1}{\lambda} \frac{1}{k} \ln \frac{\|\mathbf{X}_k\|^2}{\|\mathbf{X}_0\|^2} = \frac{1}{\lambda} \frac{1}{k} \sum_{i=0}^{k-1} \ln Y_i ,$$

dividing both sides by 2 we obtain

$$\frac{1}{\lambda} \frac{1}{k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} = \frac{1}{2} \frac{1}{\lambda} \frac{1}{k} \sum_{i=0}^{k-1} \ln Y_i . \quad (13)$$

From Lemma 4,  $(Y_k)_k$  are independent and identically distributed as

$$Y_{(1, \lambda_m)} = \min_{1 \leq i \leq \lambda/2} \{ \|\mathbf{e}_1 + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}_k^i\|^2 \}$$

such that by the LLN<sup>6</sup>, the right hand side of (13) converges to  $\frac{1}{2} \frac{1}{\lambda} E[\ln Y_{(1, \lambda_m)}]$ . Moreover by Lemma 6, the expectation of  $\ln Y_{(1, \lambda_m)}$  satisfies

$$E[\ln Y_{(1, \lambda_m)}] = \frac{1}{2} E \left[ \ln \left( 1 + \sigma \min_{1 \leq i \leq \lambda/2} (-2[\mathcal{N}^i]_1 + \sigma \|\mathcal{N}^i\|^2) \right) \right] . \quad (14)$$

Putting together (13) and (10) we obtain (4).  $\square$

<sup>6</sup>We should also prove that  $Y$  is integrable, since this step is slightly technical we refer to [Jebalia et al., 2008] for the details.

**Lemma 4.** Let  $Y_k$  be the random variable defined as  $Y_k = \min_{1 \leq i \leq \lambda/2} \{\|\mathbf{X}_k/\|\mathbf{X}_k\| + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{X}_k/\|\mathbf{X}_k\| - \sigma \mathcal{N}_k^i\|^2\}$ . Then  $(Y_k)_{k \in \mathbb{N}}$  are independent and identically distributed as

$$Y_{(1, \lambda_m)} = \min_{1 \leq i \leq \lambda/2} \{\|\mathbf{e}_1 + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}_k^i\|^2\}$$

*Proof.* The proof follows the exact same line as the proof of Lemma 4.  $\square$

**Lemma 5.** Let  $a \in \mathbb{R}^d$  and  $b \in \mathbb{R}^d$ , the following holds

$$\min\{\|a + b\|^2, \|a - b\|^2\} = \|a\|^2 + \|b\|^2 - 2|a^T b|$$

*Proof.* The following holds

$$\begin{aligned} \min\{\|a + b\|^2, \|a - b\|^2\} &= \min\{\|a\|^2 + \|b\|^2 + 2a^T b, \|a\|^2 + \|b\|^2 - 2a^T b\} \\ &= \|a\|^2 + \|b\|^2 - 2|a^T b| \end{aligned}$$

$\square$

**Lemma 6.** Let  $Y_{(1, \lambda_m)}$  be the random variable defined as  $\min_{1 \leq i \leq \lambda/2} \{\|\mathbf{e}_1 + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}_k^i\|^2\}$ , then

$$E[\ln Y_{(1, \lambda_m)}] = E \left[ \ln \left( 1 + \sigma \min_{1 \leq i \leq \lambda/2} (-2|\mathcal{N}_k^i|_1 + \sigma \|\mathcal{N}_k^i\|^2) \right) \right]. \quad (15)$$

*Proof.* Let  $i$  an integer satisfying  $1 \leq i \leq \lambda$ , we can apply Lemma 5 to the term  $\min\{\|\mathbf{e}_1 + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}_k^i\|^2\}$  and we obtain

$$\min\{\|\mathbf{e}_1 + \sigma \mathcal{N}_k^i\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}_k^i\|^2\} = 1 - 2\sigma|\mathcal{N}_k^i|_1 + \sigma^2\|\mathcal{N}_k^i\|^2 \quad (16)$$

and thus

$$Y_{(1, \lambda_m)} = \min_{1 \leq i \leq \lambda/2} (1 - 2\sigma|\mathcal{N}_k^i|_1 + \sigma^2\|\mathcal{N}_k^i\|^2)$$

after simplification we obtain

$$Y_{(1, \lambda_m)} = 1 + \sigma \min_{1 \leq i \leq \lambda/2} (-2|\mathcal{N}_k^i|_1 + \sigma \|\mathcal{N}_k^i\|^2)$$

Taking the logarithm and the expectation of the previous equation we obtain (15).  $\square$

### 5.3 Proof of Theorem 3

*Proof.* Starting from the LHS of (5) we write

$$\frac{1}{T_k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} = \frac{k}{T_k} \frac{1}{k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|}.$$

Let  $A_k = k/T_k$  and  $B_k = \frac{1}{k} \ln(\|\mathbf{X}_k\|/\|\mathbf{X}_0\|)$ . We will handle both terms independently. We first start to handle the term  $B_k$ . Because of the sequential selection, the offspring  $\mathbf{X}_k + \sigma \|\mathbf{X}_k\| \mathcal{N}_k^1$  will be accepted if it is better than its parent  $\mathbf{X}_k$ , i.e., if  $\|\mathbf{X}_k + \sigma \|\mathbf{X}_k\| \mathcal{N}_k^1\| \leq \|\mathbf{X}_k\|$  or equivalently if  $\|\mathbf{X}_k + \sigma \|\mathbf{X}_k\| \mathcal{N}_k^1\|^2 \leq \|\mathbf{X}_k\|^2$  that can be simplified into

$$\|\mathbf{X}_k/\|\mathbf{X}_k\| + \sigma \mathcal{N}_k^1\|^2 \leq 1.$$

Let us denote  $W_k^1 = \|\mathbf{X}_k/\|\mathbf{X}_k\| + \sigma\mathcal{N}_k^1\|^2$ , the offspring  $\mathbf{X}_k/\|\mathbf{X}_k\| + \sigma\mathcal{N}_k^1$  is accepted if  $1_{\{W_k^1 \leq 1\}}$ . Therefore the update equation for  $\|\mathbf{X}_k\|^2$  reads

$$\|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k + \sigma\|\mathbf{X}_k\|\mathcal{N}_k^1\|^2 1_{\{W_k^1 < 1\}} + \left( \min_{i=1,2} \{\|\mathbf{X}_k + \sigma\|\mathbf{X}_k\|\mathcal{N}_k^i\|^2\} \right) 1_{\{W_k^1 \geq 1\}}$$

i.e., the first offspring is accepted if it is better than  $\mathbf{X}_k$  otherwise the best among the two offspring is accepted. We can now factor out  $\|\mathbf{X}_k\|$  in the previous equation.

$$\|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k\|^2 \left[ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma\mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \left( \min_{i=1,2} \left\{ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma\mathcal{N}_k^i \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}} \right] . \quad (17)$$

Let  $Z_k$  be the random variable defined as

$$Z_k = \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma\mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \left( \min_{i=1,2} \left\{ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma\mathcal{N}_k^i \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}} .$$

We can write that  $\|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k\|^2 Z_k$ . Taking the logarithm, summing up, dividing by  $k$  and then by 2 as in the proof of Theorem 2, we obtain

$$B_k = \frac{1}{k} \ln \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_0\|} = \frac{1}{2} \frac{1}{k} \sum_{i=0}^k \ln Z_i .$$

From Lemma 7, the random variables  $(Z_i)$  are independent and identically distributed such that we obtain by the LLN<sup>7</sup>

$$B_k \xrightarrow[k \rightarrow \infty]{} \frac{1}{2} E[\ln Z]$$

where  $Z$  is defined in Lemma 7. Moreover, from Lemma 9,

$$E[\ln Z] = E[\ln(1 + \sigma(Y_1 1_{\{Y_1 < 0\}} + \min(Y_1, Y_2) 1_{\{Y_1 \geq 0\}}))] . \quad (18)$$

We are now going to prove that  $1/A_k = T_k/k$  converges to  $2 - p_s(\sigma)$ . Let  $\Lambda_i$  be the number of offspring evaluated at iteration  $i$ . Then,  $T_k = \Lambda_1 + \dots + \Lambda_k$  and thus

$$1/A_k = \frac{1}{k} \sum_{i=1}^k \Lambda_i .$$

From Lemma 8, the random variables  $\Lambda_i$  are independent and identically distributed as  $\Lambda$  and  $\Lambda$  is integrable. Thus applying the LLN,  $1/A_k$  converges to  $E[\Lambda]$ . Moreover from Lemma 8,  $E[\Lambda] = 2 - p_s(\sigma)$  and thus

$$1/A_k \xrightarrow[k \rightarrow \infty]{} 2 - p_s(\sigma) \quad (19)$$

Putting together (18) and (19) we obtain (5).  $\square$

<sup>7</sup>We should also prove that  $Y$  is integrable, since this step is slightly technical we refer to [Jebalia et al., 2008] for the details.

**Lemma 7.** Let  $(Z_k)_{k \in \mathbb{N}}$  be the sequence of random variables defined as

$$Z_k = \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \left( \min_{i=1,2} \left\{ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^i \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}} .$$

Then  $(Z_k)$  are independent and identically distributed as

$$Z = \left\| \mathbf{e}_1 + \sigma \mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \left( \min_{i=1,2} \left\{ \left\| \mathbf{e}_1 + \sigma \mathcal{N}_k^i \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}} .$$

*Proof.* The proof is similar to the one of Lemma 1.  $\square$

**Lemma 8.** Let  $\Lambda_k$  be the number of offspring evaluated at iteration  $k$  of a  $(1, 2^s)$ -ES. Then the random variables  $(\Lambda_k)_{k \in \mathbb{N}}$  are independent and identically distributed as  $\Lambda$  where

$$\Lambda = 1 \times 1_{\{2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 < 0\}} + 2 \times 1_{\{2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 \geq 0\}} .$$

Moreover  $E[\Lambda] < \infty$  and  $E[\Lambda] = 2 - p_s(\sigma)$ .

*Proof.* We are using the definition of the random variable  $W_k^1$  that has been introduced in the proof of Theorem 3. The random variable  $\Lambda_k$  satisfies  $\Lambda_k = 1 \times 1_{\{W_k^1 \leq 1\}} + 2 \times 1_{\{W_k^1 > 1\}}$ , i.e. one offspring is evaluated if the first offspring is better than  $\mathbf{X}_k$  (that is  $1_{\{W_k^1 \leq 1\}}$ ) otherwise two offspring are evaluated. The random variables  $W_k^1$  are independent and identically distributed with common distribution  $\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2$ : the proof is similar to the one of Lemma 1. Therefore,  $\Lambda_k$  is distributed as

$$\Lambda = 1 \times 1_{\{\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2 < 1\}} + 2 \times 1_{\{\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2 \geq 1\}}$$

We can simplify  $1_{\{\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2 < 1\}}$  into  $1_{\{2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 < 0\}}$  developing the term  $\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2$  and thus

$$\Lambda = 1 \times 1_{\{2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 < 0\}} + 2 \times 1_{\{2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 \geq 0\}} .$$

Since  $\Lambda$  is positive and almost surely bounded by 2 it is integrable. Let denote  $p_s(\sigma)$  the probability of success of the first offspring, then since  $W_k^1$  is distributed as  $\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2$ ,  $p_s(\sigma)$  satisfies  $p_s(\sigma) = \Pr(\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2 < 1)$  and after simplification also equals  $p_s(\sigma) = \Pr(2[\mathcal{N}^1] + \sigma \|\mathcal{N}^1\|^2 < 0)$ . We have thus that  $E[\Lambda_k] = p_s(\sigma) + 2(1 - p_s(\sigma))$  or  $E[\Lambda] = 2 - p_s(\sigma)$ .  $\square$

**Lemma 9.** The random variable  $Z$  defined in Lemma 7 is distributed as the random variable

$$1 + \sigma (Y_1 1_{\{Y_1 < 0\}} + \min(Y_1, Y_2) 1_{\{Y_1 \geq 0\}})$$

where  $Y_1 = 2[\mathcal{N}^1]_1 + \sigma \|\mathcal{N}^1\|^2$ ,  $Y_2 = 2[\mathcal{N}^2]_1 + \sigma \|\mathcal{N}^2\|^2$  with  $\mathcal{N}^1, \mathcal{N}^2$  being two independent random vectors following a standard multivariate normal distribution.

*Proof.* As in the proof of Lemma 8, we have that  $W_k^1$  is distributed as  $\|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2$ . Moreover simplifying  $\|\mathbf{e}_1 + \sigma \mathcal{N}_k^1\|^2$  into  $1 + 2\sigma[\mathcal{N}^1]_1 + \sigma^2 \|\mathcal{N}^1\|^2$ , we obtain the result.  $\square$

## 5.4 Proof of Theorem 4

*Proof.* The proof is similar to the proof of Theorem 3. What changes is the update equation for  $\|\mathbf{X}_k\|^2$ . Because of the mirroring we here have

$$\begin{aligned} \|\mathbf{X}_{k+1}\|^2 = \|\mathbf{X}_k\|^2 & \left[ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \right. \\ & \left. \left( \min \left\{ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^1 \right\|^2, \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} - \sigma \mathcal{N}_k^1 \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}} \right]. \end{aligned} \quad (20)$$

Defining

$$\begin{aligned} Z_k = & \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^1 \right\|^2 1_{\{W_k^1 < 1\}} + \\ & \left( \min \left\{ \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} + \sigma \mathcal{N}_k^1 \right\|^2, \left\| \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|} - \sigma \mathcal{N}_k^1 \right\|^2 \right\} \right) 1_{\{W_k^1 \geq 1\}}, \end{aligned} \quad (21)$$

we obtain that  $Z_k$  are independent and identically distributed as

$$\begin{aligned} Z = & \|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2 1_{\{2[\mathcal{N}^1]_1 + \sigma \|\mathcal{N}^1\|^2 < 0\}} + \\ & \left( \min \left\{ \|\mathbf{e}_1 + \sigma \mathcal{N}^1\|^2, \|\mathbf{e}_1 - \sigma \mathcal{N}^1\|^2 \right\} \right) 1_{\{2[\mathcal{N}^1]_1 + \sigma \|\mathcal{N}^1\|^2 \geq 0\}}, \end{aligned} \quad (22)$$

that can be written in a simpler way as

$$Z = 1 + 2\sigma([\mathcal{N}^1]_1 + |\mathcal{N}^1|) 1_{\{2[\mathcal{N}^1]_1 + \sigma^2 \|\mathcal{N}^1\|^2 < 0\}} - 2\sigma|\mathcal{N}^1| + \sigma^2 \|\mathcal{N}^1\|^2.$$

Following the same argument as in the proof of Theorem 3, we thus obtain (6).  $\square$

## 5.5 Proof of Theorem 5

*Proof.* The proof is a straightforward generalization of the proof of Theorem 3.  $\square$

## 5.6 Proof of Theorem 6

*Proof.* The proof is a straightforward generalization of the proof of Theorem 3.  $\square$

## References

- Dirk V. Arnold and Ralf Salomon. Evolutionary gradient search revisited. *IEEE Transactions on Evolutionary Computation*, 11(4):480–495, 2007.
- Dirk V. Arnold and D. C. Scott Van Wart. Cumulative step length adaptation for evolution strategies using negative recombination weights. In Mario Giacobini et al., editors, *EvoWorkshops*, volume 4974 of *LNCS*, pages 545–554. Springer, 2008.
- Anne Auger and Nikolaus Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

- Anne Auger and Nikolaus Hansen. Benchmarking the (1+1)-CMA-ES on the BBOB-2009 noisy testbed. In Rothlauf [2009], pages 2467–2472.
- Anne Auger and Nikolaus Hansen. Theory of evolution strategies: a new perspective. In Anne Auger and Benjamin Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing, 2010. accepted.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Mirrored variants of the (1,2)-CMA-ES compared on the noiseless BBOB-2010 testbed. In *GECCO (Companion)*, 2010a.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Mirrored variants of the (1,2)-CMA-ES compared on the noisy BBOB-2010 testbed. In *GECCO (Companion)*, 2010b.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Mirrored variants of the (1,4)-CMA-ES compared on the noiseless BBOB-2010 testbed. In *GECCO (Companion)*, 2010c.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Mirrored variants of the (1,4)-CMA-ES compared on the noisy BBOB-2010 testbed. In *GECCO (Companion)*, 2010d.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Investigating the impact of sequential selection in the (1,2)-CMA-ES on the noiseless BBOB-2010 testbed. In *GECCO (Companion)*, 2010e.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Investigating the impact of sequential selection in the (1,2)-CMA-ES on the noisy BBOB-2010 testbed. In *GECCO (Companion)*, 2010f.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Investigating the impact of sequential selection in the (1,4)-CMA-ES on the noiseless BBOB-2010 testbed. In *GECCO (Companion)*, 2010g.
- Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. Investigating the impact of sequential selection in the (1,4)-CMA-ES on the noisy BBOB-2010 testbed. In *GECCO (Companion)*, 2010h.
- John Michael Hammersley and David Christopher Handscomb. *Monte Carlo methods*. Methuen’s monographs on applied probability and statistics. Methuen, 1967.
- Nikolaus Hansen. An analysis of mutative  $\sigma$ -self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006a.
- Nikolaus Hansen. The CMA evolution strategy: a comparing review. In J.A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006b.
- Nikolaus Hansen. Benchmarking a BI-population CMA-ES on the BBOB-2009 function testbed. In Rothlauf [2009], pages 2389–2396.

- Nikolaus Hansen and Stefan Kern. Evaluating the CMA evolution strategy on multimodal test functions. In X. Yao et al., editors, *Parallel Problem Solving from Nature PPSN VIII*, volume 3242 of *LNCS*, pages 282–291. Springer, 2004.
- Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
- Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation. *Evolutionary Computation*, 11(1):1–18, 2003.
- Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009a. URL <http://coco.gforge.inria.fr/bbob2010-downloads>. Updated February 2010.
- Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. Real-parameter black-box optimization benchmarking 2009: Noisy functions definitions. Technical Report RR-6869, INRIA, 2009b. URL <http://coco.gforge.inria.fr/bbob2010-downloads>. Updated February 2010.
- Nikolaus Hansen, Andre Niederberger, Lino Guzzella, and Petros Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009c.
- Christian Igel, Thorsten Suttrop, and Nikolaus Hansen. A computational efficient covariance matrix update and a (1+1)-CMA for evolution strategies. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 453–460. ACM Press, 2006.
- M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the (1+1)-ES. In N. Monmarché and al., editors, *Proceedings of Evolution Artificielle (EA'07)*, volume 4926 of *LNCS*, pages 207–218. Springer, 2008.
- Franz Rothlauf, editor. *Genetic and Evolutionary Computation Conference, GECCO 2009, Proceedings, Montreal, Québec, Canada, July 8-12, 2009, Companion Material*, 2009. ACM.
- Olivier Teytaud and Sylvain Gelly. DCMA, yet another derandomization in covariance-matrix-adaptation. In D. Thierens et al., editors, *Genetic and Evolutionary Computation Conference (GECCO)*, pages 955–922. ACM Press, 2007.



---

Centre de recherche INRIA Saclay – Île-de-France  
Parc Orsay Université - ZAC des Vignes  
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex  
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier  
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq  
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique  
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex  
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex  
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex  
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

---

Éditeur  
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)  
<http://www.inria.fr>  
ISSN 0249-6399