Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

Mirrored Orthogonal Sampling for Covariance Matrix Adaptation Evolution Strategies

Hao Wang h.wang@liacs.leidenuniv.nl
LIACS, Leiden University, 2333 CA Leiden, The Netherlands

Michael Emmerich m.t.m.emmerich@liacs.leidenuniv.nl
LIACS, Leiden University, 2333 CA Leiden, The Netherlands

Thomas Bäck t.h.w.baeck@liacs.leidenuniv.nl
LIACS, Leiden University, 2333 CA Leiden, The Netherlands

Abstract

Generating more evenly distributed samples in high dimensional search spaces is the major purpose of the recently proposed *mirrored sampling* technique for evolution strategies. The diversity of the mutation samples is enlarged and the convergence rate is therefore improved by the mirrored sampling. Motivated by the mirrored sampling technique, this paper introduces a new derandomized sampling technique called *mirrored orthogonal sampling*. The performance of this new technique is both theoretically analyzed and empirically studied on the sphere function. In particular, the mirrored orthogonal sampling technique is applied to the well-known Covariance Matrix Adaptation Evolution Strategy (CMA-ES). The resulting algorithm is experimentally tested on the well-known Black-Box Optimization Benchmark (BBOB). By comparing the results from the benchmark, mirrored orthogonal sampling is found to outperform both the standard CMA-ES and its variant using mirrored sampling.

Keywords

Convergence of numerical methods, evolution strategies, mirrored orthogonal sampling.

1 Introduction

In Evolution Strategies (ESs), derandomized sampling aims at improving random sampling by generating "good" random mutation points from a certain distribution (usually Gaussian). Loosely speaking, the "goodness" of the mutation refers to its diversity, which measures how evenly the mutation points are arranged in the search space. Intuitively, the mutation sample having a high diversity could explore the search space more thoroughly. We will introduce the definition of diversity in Sec. 2.2.

Much effort has been devoted to the derandomized sampling and several methods are proposed. Niederreiter (1992) proposes to adopt *quasi-random* variables, which has already been applied to Genetic Algorithms (Kimura and Matsumura, 2005) and Evolution Strategies (Teytaud and Gelly, 2007). Despite their successes, these approaches are found to be more complicated to implement than the simple random sampling and may introduce undesired biases as will be shown later. A recent systematic overview of modern variants of evolution strategies can be found in Bäck et al. (2013).

The recently proposed *mirrored sampling* technique (Brockhoff et al., 2010) is a simple and effective derandomized sampling method for ES. Instead of generating i.i.d.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

Gaussian samples, the sample points are paired and symmetric to the current parent, where only one sample point of each pair is generated by the simple random sampling. In mirrored sampling, half of the sample points are independent and the other half are dependent. There is no additional computational cost needed for the mirrored sampling technique. It has been theoretically proven that the performance of $(1 \stackrel{+}{,} \lambda)$-ES can be improved by applying the mirrored sampling technique (Auger et al., 2011a).

The purpose of this paper is to elaborate and analyze an improvement of the mirrored sampling technique, which is called *mirrored orthogonal sampling*. Its basic idea has been briefly introduced in Wang et al. (2014) and it is based on the following intuition. In mirrored sampling, half of the samples are still obtained using the simple random sampling, which would suffer from the same problem as before, namely the sampling errors (see Sec. 2.2 for an in-depth discussion. Thus, the diversity of random samples could be further enhanced by improving the mirrored sampling technique. This is achieved by completely discarding the simple random sampling component in mirrored sampling and replacing it by the *orthogonal sampling*. The resulting technique is called *mirrored orthogonal sampling* here. This paper extends the our previous work (Wang et al., 2014) in the following aspects:

- The motivation, intuition and formalization of the mirrored orthogonal sampling are incorporated in detail.
- The procedure for tuning the strategy parameters for mirrored orthogonal sampling is discussed.
- The progress rate approach is applied to analyze and compare the simple random, mirrored and mirrored orthogonal sampling techniques.
- The benchmark results are presented and explained in detail.

Based on the convergence rate analysis approach in Brockhoff et al. (2010), we analyze the convergence rate of the isotropic $(1, \lambda)$-ES with mirrored orthogonal sampling on the sphere function. For the (μ, λ)-ES algorithm, a bias would occur during recombination, probably leading to premature convergence behavior. This bias is avoided by applying the *pairwise selection* technique, in which only the better point of a mirrored pair is allowed to participate in the weighted recombination. Mirrored orthogonal sampling is applied within a CMA-ES for the experimental validation.

This paper is organized as follows. Sec. 2 introduces the background of derandomized sampling as well as the motivation of our work. In Sec. 3.1, the new derandomized sampling approach is proposed and explained in detail. Sec. 3.2 concentrates on the implementation issues of our approach. Both the theoretical and empirical study of the convergence rates are presented in Sec. 4. The progress rate analysis is used to analyze the algorithm performance. In Sec. 5, the experimental results of mirrored orthogonal sampling are shown and compared to the other sampling methods. Finally, conclusions and possible directions for further research are given in Sec. 6. In this paper, we shall use n to denote the dimensionality of the search space, and λ to denote the population size of evolution strategy.

2 Background and related work

2.1 Evolution Strategy

In this paper, we are dealing with single-objective continuous optimization problems, namely the objective function of the form: $f : \mathbb{R}^n \to \mathbb{R}$. Without loss of generality, the

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

search space is assumed as the whole n-dimensional Euclidean space $\mathbb{R}^n$. Evolution Strategies (ESs) are population-based black-box optimization techniques, proposed to solve such problems efficiently (Schwefel, 1993; Bäck et al., 2013). The ES algorithm is built heavily on the so-called **mutation operator**, where random variables are used to perturb candidate solutions locally. Mostly commonly, the mutation operator is realized by taking a **simple random sample** (of size $\lambda > 1$) from the *multivariate Gaussian distribution*:

$$\mathbf{x}_i = \mathbf{m} + \sigma\boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{C}), \quad i = 1, 2, \ldots, \lambda, \tag{1}$$

where 1) $\mathbf{m}$ is the center of mass of the current population. 2) The covariance matrix $\mathbf{C}$ models the correlation among search variables. 3) The parameter $\sigma \in \mathbb{R}_{>0}$ is the so-called step-size, which scales the magnitude of the mutation. After obtaining the fitness value of $\{\mathbf{x}_i\}_{i=1}^{\lambda}$ on f, the only the μ best mutations ($\mu < \lambda$) are selected: $\mathbf{x}_{1:\lambda}, \mathbf{x}_{2:\lambda}, \ldots, \mathbf{x}_{\mu:\lambda}$. Note that $\mathbf{x}_{i:\lambda}$ stands for the mutation corresponds to the i-th ranked fitness value $f_{i:\lambda}$. Then, the center of mass $\mathbf{m}$ is re-calculated based on the so-called *weighted recombination* (Hansen and Ostermeier, 2001):

$$\mathbf{m} = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda},$$

where the weight is scaled down logarithmically with increasing ranks of the mutation. Typically, such an algorithm paradigm is termed $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES (Bäck et al., 2013), representing μ parents, λ mutations/offspring and μ_{w} mutations in the weighted recombination. Note that the covariance matrix is also adapted using the selected mutations. In the *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES) (Hansen, 2006), this is achieved by the maximum likelihood estimation. In addition, in CMA-ES the step-size σ is controlled based on the *Cumulative Step-size Adaptation* (CSA) mechanism (Hansen and Ostermeier, 2001), where the steps (displacement of $\mathbf{m}$ between iterations) are cumulated with exponential decay and the length of the cumulant is used to adjust the step-size. The changing rate of the step-size is regulated by the so-called *damping factor* d_σ. Please refer to Hansen (2006) for the detail. Note that the original damping factor is modified when applying the proposed sampling approach to CMA-ES (Sec. 3.4).

2.2 Sampling Error and Space Exploration

Given a mutation sample $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\lambda\} \subset \mathbb{R}^n$, the diversity $D(\mathcal{X})$ is defined as the minimal distance between sample points in $\mathcal{X}$. Intuitively, it can be measured by the lower bound of distance between points in $\mathcal{X}$:

$$D(\mathcal{X}) \coloneqq \inf \left\{ d(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \neq \mathbf{x}_j \in \mathcal{X} \right\},$$

where d is a distance metric in $\mathbb{R}^n$. The diversity should be maximized in order to obtain "good" mutation samples. Although many other criteria might apply (e.g. discrepancy) on the sample "goodness", we will just use the simple measure above. Generating n-dimensional random vectors from a multivariate Gaussian distribution is the key source of random variations in evolution strategies. The standard method to achieve this, *simple random sampling* (Eq. (1)), samples pseudo-random numbers directly from a certain distribution. However, it also suffers from the so-called *sampling error*, which describes the situation that the estimated properties (from a sample) differ largely from the property of the population. The sampling error is caused by unrepresentative or biased samples when the sample size is small.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

An example of biased samples is illustrated in Fig. 1, in which four i.i.d. mutation vectors are sampled from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{C})$ (the step-size is ignored in this case). The black solid ellipsoid represents the covariance matrix $\mathbf{C}$. The diversity of the mutation vectors is not satisfactory because the minimal distance between samples is relatively small. A strong sampling error incurs in this case because if the mean and covariance of the distribution are estimated from these four vectors, the results would deviate largely from $\mathbf{m}$ and $\mathbf{C}$.

Consequently, a large portion of the search space is not reached (at least half the space in this case). Moreover, if the objective function is locally convex near the optimum (as illustrated by the dashed ellipsoids). The probability that a new search point leads to an improvement can be very small, as shown by the area marked by vertical lines. Therefore, if the population size is small, a biased sample can take place such that none of the mutations leads to an improvement, hindering the progress of this generation. The sampling error has an even bigger side effect in modern evolution strategies (e.g., CMA-ES) because those algorithms tend to exploit small populations to speed up their convergence rate. To overcome this problem, it is proposed to apply derandomized sampling methods for a small population.

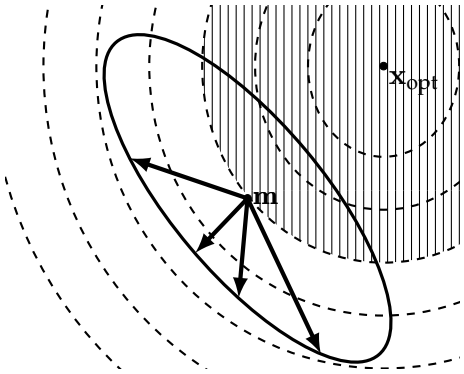


Figure 1: Example of a set of unsuccessful mutation samples. Four offspring are generated here while none of them is an improvement. This phenomenon reduces the convergence velocity of the algorithm.

2.3 Quasi-Random Sampling

There are some techniques proposed to reduce the sampling error as much as possible. The first method is called quasi-random sampling, which produces low-discrepancy sequences of samples (Dick and Pillichshammer, 2010). The discrepancy of a sequence is low if the proportion of points in the sequence falling into an arbitrary set is close to proportional to the size of this set. Low-discrepancy sequences are commonly used as a replacement of simple random samples from the uniform distribution. Intuitively, such sequences span the search space more "evenly" than the pseudo-random numbers. It is widely used in numerical approaches like the quasi-Monte-Carlo method (Niederreiter, 1992) to achieve a faster rate of convergence. Due to the advantages of quasi-random sampling, it is also applied in genetic algorithms (Kimura and Matsumura, 2005) and evolution strategies (Teytaud and Gelly, 2007). Specifically, it has already been applied to the well-known Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen and Ostermeier, 2001; Hansen et al., 2003). Teytaud and Gelly

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

(2007) propose to replace the simple random Gaussian samples by a low-discrepancy sequence in the mutation operator. The method for generating quasi-random samples according to the Gaussian distribution is also developed because the quasi-random samples are usually generated for a uniform distribution. It is also argued that the efficiency of CMA-ES is improved due to a better diversity of quasi-random samples. However, such an approach would cause a systematic bias on the step size adaptation (see Sec. 3.3).

2.4 Mirrored Sampling

The mirrored sampling technique (Brockhoff et al., 2010) is another method for obtaining "good" samples and it successfully accelerates the convergence of ESs (Auger et al., 2010). It is a simple and elegant idea in which a single random mutation is used to create two sample points: Instead of generating λ i.i.d. search points, only half of the mutation vectors are generated using simple random sampling, namely $\{\mathbf{z}_{2i-1}\}_{i=1}^{\lambda/2}$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})$. Each mutation vector $\mathbf{z}_{2i-1}$ is used to generate a pair of offspring, $\mathbf{x}_{2i-1} = \mathbf{m} + \mathbf{z}_{2i-1}$ and $\mathbf{x}_{2i} = \mathbf{m} - \mathbf{z}_{2i-1}$, which are symmetric about the center of mass $\mathbf{m}$ (parent point).

Algorithm 1 Mirrored sampling

procedure MIRRORED-SAMPLING($\mathbf{m}, \sigma, \mathbf{C}, \lambda$)
  $\mathbf{B}, \mathbf{D} \leftarrow$ EIGEN-DECOMPOSITION($\mathbf{C}$) ▷ $\mathbf{B}$: eigenvector and $\mathbf{D}$: eigenvalues
  if $\lambda \mod 2 \neq 0$ **and** $\mathbf{z}_{\text{last}}$ is set **then** ▷ if λ is an odd number
    $\mathbf{x}_\lambda \leftarrow \mathbf{m} - \sigma\mathbf{B}\mathbf{D}\mathbf{z}_{\text{last}}$ ▷ $\mathbf{z}_{\text{last}}$: the unused mutation from the last iteration
    $\lambda' \leftarrow \lambda - 1$
    Unset the static variable $\mathbf{z}_{\text{last}}$. ▷ Unset $\mathbf{z}_{\text{last}}$ once it is used
  else
    $\lambda' \leftarrow \lambda$
  end if
  for $i = 1 \rightarrow \lambda'$ **do**
    if $i \mod 2 = 0$ **then**
      $\mathbf{x}_i \leftarrow \mathbf{m} - \sigma\mathbf{B}\mathbf{D}\mathbf{z}_{i-1}$ ▷ The mirrored vector
    else
      $\mathbf{z}_i \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ The original vector
      $\mathbf{x}_i \leftarrow \mathbf{m} + \sigma\mathbf{B}\mathbf{D}\mathbf{z}_i$
    end if
  end for
  if $\lambda' \mod 2 \neq 0$ **then** ▷ Odd number of mutations are created
    Set the the static variable $\mathbf{z}_{\text{last}} \leftarrow \mathbf{z}_\lambda$ ▷ Save unused $\mathbf{z}_\lambda$ for the next iteration
  end if
  return $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\lambda\}$
end procedure

To make the following discussion clear, the mutation obtained directly from simple random sampling is called the original mutation. The mirrored sampling method is described in Alg. 1, acting as an alternative to the standard mutation operator (simple random sampling) in evolution strategies. For an odd λ, it begins with generating $\lceil \lambda/2 \rceil$ mutations in the first generation, corresponding to $\lceil \lambda/2 \rceil$ mirrored ones. To keep the population size alway as λ, all $\lceil \lambda/2 \rceil$ original mutations and $\lceil \lambda/2 \rceil - 1$ mirrored ones undergo the evaluation and selection procedure while the last mirrored mutation

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

is held out for the next iteration (Lines 18-21). In the next iteration, the hold-out mirrored mutation is used (Lines 3-9) and we only need to draw $\lfloor \lambda/2 \rfloor$ mutations. The following generations repeat this procedure. The static variable $\mathbf{z}_{\text{last}}$ in Alg. 1 stores the hold-out mutation. Here, the notation proposed in (Brockhoff et al., 2010) is used such that any ES algorithm using mirrored sampling is denoted as $(\mu \stackrel{+}{,} \lambda_{\text{m}})$-ES. Consequently, in the $(1 + 1_{\text{m}})$-ES, a mirrored mutation is used if and only if the iteration index is even. By using mirrored sampling, mutations in each mirrored pair are dependent and explore two anti-parallel directions such that the mirrored counterpart of an unsuccessful mutation has a certain chance to yield an improvement.

Note that the mirrored sampling method is very similar to the so-called *opposition-based learning* method (Rahnamayan et al., 2006), in which the candidate solution is mirrored with respect to the center of the smallest hyper-box covering the current population. This approach is implemented in the Differential Evolution (DE) algorithm to generate an opposite population occasionally, which improves the performance of DE (Rahnamayan et al., 2006).

2.5 Deterministic Orthogonal Sampling

Orthogonal sampling, which denotes a the sampling approach utilizing orthogonal search directions, is another solution to enhance the mutation diversity. This sampling scheme can be found in Coordinate Descent (Schwefel, 1993), Adaptive Coordinate Descent (ACiD) (Loshchilov et al., 2011) and Rosenbrock's Local Search (Rosenbrock, 1960). Intuitively, by taking samples on the orthogonal directions, the search space is covered more evenly.

Normally, in this approach, an orthogonal basis $\Xi = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n\}$ are maintained in each optimization iteration, which represents the possible search directions. In each iteration, a line search is conducted along a basis vector, which is achieved by sampling two trial points: one point is created by adding the basis vector to the current search point $\mathbf{m}$ while the other one is generated through mirroring. In the next iteration, a different basis vector in Ξ is picked for the exploration. The general framework of this method is summarized below:

1. Initialize the search point $\mathbf{m}$, a set of orthonormal basis vectors $\Xi = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n\}$ as the search directions and the step sizes $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ for each search direction.
2. If the termination condition is not satisfied, perform the following steps until (e) for each iteration. Let g be the iteration counter:
   (a) Choose base $\boldsymbol{\xi}_i$ as the exploration direction where $i = g \mod n$ and generate one trial point: $\mathbf{x}_1 = \mathbf{m} + \sigma_i \boldsymbol{\xi}_i$.
   (b) For Rosenbrock's local search, goto (c). For the other methods, use base $\boldsymbol{\xi}_i$ to generate the other trial point: $\mathbf{x}_2 = \mathbf{m} - \sigma_i \boldsymbol{\xi}_i$.
   (c) Evaluate the trial points $\mathbf{x}_1, \mathbf{x}_2$ (if $\mathbf{x}_2$ exists). Set the search point $\mathbf{m}$ to the one with the best fitness value.
   (d) Update the step size σ_i according to a deterministic or stochastic rule and increase the iteration counter g by one.
   (e) If $g \mod n = 0$, then update the basis Ξ according to the search points of the most recent n iterations.

When all the basis directions are tried, the orthogonal basis Ξ is either unchanged or updated based on the successful trials in the history. Note that the rules of the update

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

may vary from algorithm to algorithm: In Coordinate Descent, the basis is fixed to the canonical basis of $\mathbb{R}^n$ during the process. In ACiD, the basis is updated by Adaptive Encoding (Hansen, 2008), which is the generalization of the covariance matrix adaptation in CMA-ES. We deliberately term this sampling method as *deterministic* orthogonal sampling due to the fact that the update of the orthonormal basis is completely deterministic and it is easier to distinguish this sampling method from the *random* orthogonal sampling proposed here.

3 Mirrored Orthogonal Sampling

In this section, we elaborate the mirrored orthogonal sampling technique.

3.1 The Proposed Method

This new method is motivated by the following observation: In mirrored sampling, half of the mutation vectors (the mirrored ones) completely depend on the other half (the original ones). Mirrored sampling ensures a significant difference between these two halves of mutations. In addition, the mirrored mutation is anti-parallel to the original one and thus a mirrored pair would never miss one half of the search space, no matter how the search space is partitioned[1]. However, within each half of mutations, the diversity is still not regulated such that many mutations might be "squeezed" in a narrow direction. Thus, the mirrored sampling technique can still generate unrepresentative samples as described in Sec. 2.2.

In order to alleviate this issue, we consider the deterministic orthogonal sampling method (Sec. 2.5), where the mutations are selected from a precomputed orthogonal basis and thus the minimal distance between samples is enlarged. The disadvantage is that deterministic search directions are used and only one of the orthogonal vectors can be used in one evolution cycle, which limits it usability for the general (μ, λ)-ES. Instead of just picking vectors in an orthogonal basis, it is proposed here to create uniformly random orthogonal vectors, in the sense that each vector is stochastic (instead of being deterministic) and uniformly random (meaning that each search direction is sampled with the same probability). The definition of such samples is given as:

Definition 1. The *uniform random orthogonal vectors* are defined as a set of random vectors $\{\boldsymbol{\mathcal{O}}_1, \boldsymbol{\mathcal{O}}_2, \ldots, \boldsymbol{\mathcal{O}}_k\} \subset \mathbb{R}^n$ ($k \leq n$), satisfying the following three properties:

1. Orthogonality: $\forall i \neq j \in \{1, 2, \ldots, k\}, \ \langle \boldsymbol{\mathcal{O}}_i, \boldsymbol{\mathcal{O}}_j \rangle = 0$.
2. χ-distributed norm: $\forall i \in \{1, 2, \ldots, k\}, \ \|\boldsymbol{\mathcal{O}}_i\| = \sqrt{\langle \boldsymbol{\mathcal{O}}_i, \boldsymbol{\mathcal{O}}_i \rangle} \sim \chi(n)$.
3. Uniformity: for each vector $\boldsymbol{\mathcal{O}}_i$, its normalization $\boldsymbol{\mathcal{O}}_i / \|\boldsymbol{\mathcal{O}}_i\|$ distributes *uniformly* on the unit sphere.

Remarks: 1) The norm of the sample vector is restricted to χ distribution for mimicking the behavior of the standard Gaussian vector. 2) The uniform distribution on the unit sphere is equivalent to the *rotation-invariant* property with respect to an arbitrary rotation matrix[2] $\mathbf{R}$: the random vector $\mathbf{x}$ and the rotated one $\mathbf{x}' = \mathbf{R}\mathbf{x}$ are identically distributed. 3) Throughout this paper, the dot product is taken for the inner product, namely $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$.

[1] Note that the mirrored pair can stay on the partition boundary. However, this situation has only zero measure in $\mathbb{R}^n$.

[2] A n dimensional rotation matrix $\mathbf{R}$ satisfies conditions $\mathbf{R}^{-1} = \mathbf{R}^\top$ and $\det \mathbf{R} = 1$. All such matrices form a so-called special orthogonal group $\mathrm{SO}(n)$.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

The new mutation method is named *random orthogonal sampling*. For clarity, we shall refer the default mutation operator (Eq. (1)) in CMA-ES as *simple random sampling*. In addition, the random orthogonal samples are rescaled and rotated according to the covariance matrix $\mathbf{C}$ before they are added to the parental point $\mathbf{m}$, as with the default mutation operator:

$$\mathbf{x}_{2i-1} \leftarrow \mathbf{m} + \sigma \mathbf{C}^{\frac{1}{2}} \boldsymbol{\mathcal{O}}_i, \quad 1 \leq i \leq \lambda/2. \tag{2}$$

The $\mathbf{x}$'s are the new search points and σ denotes the step size. The implementation of the random orthogonal sampling algorithm and the validity of the implementation are discussed in the following section. Consider two i.i.d. random vectors $\mathbf{x}$ and $\mathbf{y}$ drawn from a standard normal distribution. The expected value of the inner product of these two vectors is given as:

$$\mathbb{E}\left\{\langle \mathbf{x}, \mathbf{y} \rangle\right\} = \sum_{i=1}^{n} \mathbb{E}\left\{x_i y_i\right\} = 0.$$

This indicates two independent standard normal vectors are orthogonal to each other in expectation. Intuitively, by generating random orthogonal samples, the mutations are derandomized such that the variance of the angle formed by a pair of mutations vanishes. Therefore, the search directions are guaranteed to be uncorrelated so that mutations are spread over the search space evenly. In the next step, we combine the mirroring technique with random orthogonal sampling to generate the remaining half of the mutations:

$$\mathbf{x}_{2i} \leftarrow \mathbf{m} - \sigma \mathbf{C}^{\frac{1}{2}} \boldsymbol{\mathcal{O}}_i, \quad 1 \leq i \leq \lambda/2. \tag{3}$$

Note that only using random orthogonal sampling is not sufficient for exploration due

Algorithm 2 Mirrored orthogonal sampling

procedure MIRRORED-ORTHOGONAL-SAMPLING($\mathbf{m}, \sigma, \mathbf{C}, \lambda$)
  $\mathbf{B}, \mathbf{D} \leftarrow$ EIGEN-DECOMPOSITION($\mathbf{C}$)
  if $\lambda \mod 2 \neq 0$ **and** $\mathbf{z}_{\text{last}}$ is not set **then**
    $\mathbf{x}_\lambda \leftarrow \mathbf{m} - \sigma \mathbf{B}\mathbf{D}\mathbf{z}_{\text{last}}$ ▷ $\mathbf{z}_{\text{last}}$: the unused mutation from the last iteration
    $\lambda' \leftarrow \lambda - 1$ ▷ One offspring is already created
    Unset the static variable $\mathbf{z}_{\text{last}}$. ▷ Unset $\mathbf{z}_{\text{last}}$ once it is used
  else
    $\lambda' \leftarrow \lambda$
  end if
  $p \leftarrow \lceil \lambda'/2 \rceil$
  $\{\mathbf{z}_i\}_{i=1}^{p} \leftarrow$ ORTHOGONAL-SAMPLING(p) ▷ sub-procedure, see Alg. 4
  for $i = 1 \rightarrow p$ **do**
    $\mathbf{x}_{2i-1} \leftarrow \mathbf{m} + \sigma \mathbf{B}\mathbf{D}\mathbf{z}_i$
    $\mathbf{x}_{2i} \leftarrow \mathbf{m} - \sigma \mathbf{B}\mathbf{D}\mathbf{z}_i$ ▷ Mirroring
  end for
  if $\lambda' \mod 2 \neq 0$ **then** ▷ Save the unused mutation to the next iteration
    Set the static variable $\mathbf{z}_{\text{last}} \leftarrow \mathbf{z}_p$
  end if
  return $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\lambda\}$
end procedure

to the fact that random orthogonal vectors are only capable of spanning one orthant

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

of the space, no matter how they are realized (just consider the canonical basis in 3-D). Combining Eq. (2) and (3), the new sampling approach is completed and is called *mirrored orthogonal sampling*. In addition, any ES algorithm equipped with it is denoted as $(\mu \stackrel{+}{,} \lambda_{\mathrm{m}}^{\mathrm{o}})$-ES here. The detailed algorithm of the mirrored orthogonal sampling method is given in Alg. 2. Note that an algorithm for generating random orthogonal Gaussian vectors (which is explained in the following) is invoked in line 10 and replaces the direct sampling of the Gaussian distribution. The remainder of this algorithm is basically the same as mirrored sampling (Alg. 1).

Compared to mirrored sampling, which ensures the difference within any mirrored pair, the orthogonalization method is exploited to guarantee the significant differences among mutations. Therefore, it is straightforward to compare the performance of mirrored orthogonal sampling to that of mirrored sampling / simple random sampling. Such a comparison is presented in the experimental results (Sec. 5).

3.2 Implementation of Random Orthogonal Sampling

In order to implement the random orthogonal sampling technique, the well-known Gram-Schmidt process (Björck, 1994) is exploited to generate the orthogonal sample points. The Gram-Schmidt process is a method for orthonormalizing a set of vectors in an inner product space, most commonly the Euclidean space $\mathbb{R}^n$. It takes a finite, linearly independent set $\mathcal{S} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ ($k \leq n$) and generates an orthogonal set $\mathcal{S}' = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ that spans k-dimensional subspace of $\mathbb{R}^n$. The pseudo code of the Gram-Schmidt process is listed in Alg. 3.

Algorithm 3 Gram-Schmidt Orthonormalization

```
1:  procedure GRAM-SCHMIDT({v_1, v_2, ..., v_k})
2:      for i = 2 → k do
3:          for j = 1 → i − 1 do
4:              v_i ← v_i − (v_i^⊤ v_j / ||v_j||^2) v_j        ▷ Orthogonalizing v_i to v_j
5:          end for
6:      end for
7:      for i = 1 → k do
8:          v_i ← v_i / ||v_i||                               ▷ Normalization
9:      end for
10:     return {v_1, v_2, ..., v_k}
11: end procedure
```

Let p equal $\lambda/2$ again. In the first step, we sample p i.i.d. vectors from the standard normal distribution and record their norms (lengths), i.e.,

$$\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_p\}, \quad \mathbf{s}_i \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I}), \quad L_i = \|\mathbf{s}_i\|, \quad i = 1, \ldots, p. \tag{4}$$

Note that the Gram-Schmidt process is an orthonormalization method, which produces orthogonal unit vectors. Therefore, the original mutation lengths have to be manually recorded before applying the Gram-Schmidt process, such that the mutation length can be restored later. Then, processing $\mathcal{S}$ by the Gram-Schmidt process would give us a collection $\mathcal{S}'$ of random orthonormal vectors,

$$\mathcal{S}' = \{\mathbf{s}'_1, \ldots, \mathbf{s}'_p\} = \text{GRAM-SCHMIDT}(\mathcal{S}). \tag{5}$$

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

Note that each vector in $\mathbf{s}'_1, \ldots, \mathbf{s}'_p$ has unit length and they are orthogonal to each other. It is not very hard to see from Alg. 3 that among all the resulting vectors, the direction of $\mathbf{s}'_1$ remains unchanged and the direction of $\mathbf{s}'_i$ depends on the set $\{\mathbf{s}_k\}_{k=1}^{i-1}$. Therefore, intuitively, the output vectors of the Gram-Schmidt process, $\{\mathbf{s}'_i\}_{i=1}^{p}$ are uniformly distributed on the unit sphere because the input vectors $\{\mathbf{s}_k\}_{k=1}^{p}$ are independent and identically distributed. Finally, we rescale all the $\mathbf{s}'_i$ by their corresponding original length:

$$\mathbf{z}_i = L_i \mathbf{s}'_i, \quad i = 1, \ldots, p. \tag{6}$$

A special situation takes place if p is greater than the dimensionality n: it is simple not possible to generate more than n distinct orthogonal vectors in $\mathbb{R}^n$. In this case, only n mutation samples are created using Equations 4, 5 and 6, and the remaining $p - n$ samples are created using simple random sampling. The detailed procedure of orthogonal sampling is described in Alg. 4. Lines $3 - 6$ correspond to Eq. (4). Through lines $7 - 17$, the Gram-Schmidt process is invoked and the number of samples p is handled properly. The advantage of this implementation is that there is no additional parameter to be considered. As for the time complexity, the extra cost are spent in calling the Gram-Schmidt process, which is $O(nk^2), k = \min\{p, n\}$.

Algorithm 4 Orthogonal sampling

```
1:  procedure ORTHOGONAL-SAMPLING(p)
2:      for i = 1 → p do
3:          s_i ← N(0, I)                          ▷ generate standard normal vectors
4:          L_i ← ||s_i||                          ▷ store the length
5:      end for
6:      k ← min{p, n}                              ▷ number of inputs for Gram-Schmidt
7:      {s'_1, ..., s'_k} ← GRAM-SCHMIDT({s_1, ..., s_k})   ▷ sub-procedure, see Alg. 3
8:      for i = 1 → k do
9:          z_i ← L_i s'_i                         ▷ rescale the length
10:     end for
11:     if k < p then                              ▷ more than n samples are needed
12:         for i = 1 → p − k do
13:             z_{k+i} ← s_{k+i}                  ▷ copy the standard normal vectors
14:         end for
15:     end if
16:     return {z_1, z_2, ..., z_p}
17: end procedure
```

To justify this implementation, it is possible to check the generated samples according to Def. 1. the orthogonality and restriction on the vectors length are immediately satisfied. The rotation-invariance of the vectors can be shown as follows. Firstly, the standard normal vectors are rotation-invariant, meaning that for every $\mathbf{s}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it has the same distribution as $\mathbf{R}\mathbf{s}_i$, where $\mathbf{R}$ is rotation matrix taken from $\mathrm{SO}(n)$. Second, the orthogonalization formula of the Gram-Schmidt process, which is encoded in Alg. 3, reads as follows:

$$\mathbf{s}'_i = \mathbf{s}_i - \sum_{j=1}^{i-1} \frac{\langle \mathbf{s}_i, \mathbf{s}_j \rangle}{\|\mathbf{s}_j\|^2} \mathbf{s}_j, \quad i = 1, \ldots, p,$$

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

Now if an arbitrary rotation operator $\mathbf{R} \in \mathrm{SO}(n)$ is applied on $\mathbf{s}'_i$, the resulting vector is,

$$\mathbf{s}''_i = \mathbf{R}\mathbf{s}'_i = \mathbf{R}\mathbf{s}_i - \sum_{j=1}^{i-1} \frac{\langle \mathbf{R}\mathbf{s}_i, \mathbf{R}\mathbf{s}_j \rangle}{\|\mathbf{R}\mathbf{s}_j\|^2} \mathbf{R}\mathbf{s}_j, \quad i = 1, \ldots, p. \tag{7}$$

Note that it is valid to put $\mathbf{R}$ in the norm and the inner product (e.g., $\|\mathbf{R}\mathbf{s}_j\|$) because such matrices preserve the inner product. Finally, $\mathbf{R}\mathbf{s}_i$ is identically distributed as $\mathbf{s}_i$ and it also holds for the rest terms in the right-hand-side of Eq. (7). Therefore, $\mathbf{s}''_i$ is identically distributed as $\mathbf{s}'_i$ and therefore it is rotation-invariant. A more rigorous proof can be found in (Eaton, 1983, Proposition 7.2).

3.3 Recombination and Pairwise Selection

When weighted recombination and cumulative step size adaptation are also used in an evolution strategy algorithm (e.g., CMA-ES), mirrored sampling causes an undesired bias of the step size σ, where the step size is reduced in the much faster rate compared to the original CMA-ES (Auger et al., 2011b; Brockhoff et al., 2010). This bias could result in the premature convergence of optimization (Brockhoff et al., 2010). The bias occurs if any mirrored pair of mutations $\mathbf{m} + \sigma\mathbf{z}_i, \mathbf{m} - \sigma\mathbf{z}_i$ is selected together and then the two mutations cancel each other during the recombination, which is called pairwise cancellation here. To see its side effect, please consider a population $\{\mathbf{z}_i\}_{1 \le i \le \lambda}$ where $\mathbf{z}_i \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})$ and equal weights are used in recombination. Under the random selection, the distribution of selected vectors is still normal so that the recombination $\langle \mathbf{z} \rangle$ is still normally distributed as follows:

$$\langle \mathbf{z} \rangle = \frac{1}{\mu} \sum_{i=1}^{\mu} \mathbf{z}_{i:\lambda} \sim \frac{1}{\mu} \boldsymbol{\mathcal{N}}(\mathbf{0}, \mu^2 \mathbf{I}) \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})$$

In the mirrored case, if one pair of mirrored mutations is selected, then such a pair would disappear in the summation above. The recombined mutation of mirrored sampling is then distributed as follows:

$$\langle \mathbf{z}_m \rangle = \frac{1}{\mu} \sum_{i=1}^{\mu-2} \mathbf{z}_{i:\lambda} \sim \boldsymbol{\mathcal{N}} \left(\mathbf{0}, \left(1 - \frac{2}{\mu}\right)^2 \mathbf{I} \right)$$

It is now obvious to see that the variance of recombined mutation is reduced under the random selection. The more mirrored pairs are selected, the more undesirable bias will be generated. In addition, the cumulative step size adaptation mechanism (CSA) updates the step size according to the exponential change of the length of the accumulated vector $\langle \mathbf{z} \rangle$, namely the accumulation of realized steps (Hansen and Ostermeier, 2001). This is the reason why the step size is quickly reduced in CMA-ES when mirrored sampling is naively plugged in.

To fix this undesirable effect, the *pairwise selection* heuristic introduced in Auger et al. (2011b) is adopted here. Pairwise selection prevents the pairwise cancellation by allowing only the better mutation among the mirrored pair to contribute to the weighted recombination. The effect of combining pairwise selection and mirroring is presented by solid curves in Fig. 2a, in which no bias in step size adaptation can be observed. In the following sections, pairwise selection is used in the ES whenever mirrored sampling or mirrored orthogonal sampling is used.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

3.4 Application to the CMA-ES Algorithm

We apply the *mirrored orthogonal sampling* technique to the CMA-ES (Hansen and Ostermeier, 2001). In addition to the recombination problem discussed in the last section, some tuning is required to find default settings of control parameters for the new sampling technique. The step size control mechanism, *Cumulative Step-size Adaptation* (CSA), is exploited in CMA-ES. In the CSA technique, the damping factor d_σ controls the adaptation speed of the step size σ and is originally developed for i.i.d. Gaussian mutations. However, the mutations generated by mirrored orthogonal sampling are no longer independently distributed. Therefore, the damping factor d_σ needs to be optimized for the newly proposed technique. The default setting of the damping factor in (Hansen, 2006) is $d_\sigma = 1 + 2\max\{0, \sqrt{(\mu_{\text{eff}} - 1)/(n+1)} - 1\} + c_\sigma$. Note that μ_{eff} is defined as the variance effective selection mass (Hansen and Ostermeier, 2001) of the recombination weights $\{w_i\}_{i=1}^{\mu}$ and computed according to $\mu_{\text{eff}} = \left(\sum_{i=1}^{\mu} w_i^2\right)^{-1}$. c_σ is the cumulation constant used for the evolution path and usually $c_\sigma \ll 1$. For other default parameters in CMA-ES and their explanation, please refer to (Hansen, 2006).

We tune the damping factor under the default λ setting, which is the rounded logarithm of dimensionality . The tuning approach follows the approach proposed in (Brockhoff et al., 2010) to choose the new d_σ setting. First, every strategy parameter except d_σ is initialized by its default value. Second, multiple d_σ values are evaluated according to an experiment performed on the sphere function $f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2$, where the performance can be assumed to be a unimodal function of d_σ, such that a unique optimum value for d_σ can be determined. An example of this second step for $(\mu/\mu_{\text{w}}, \lambda^{\text{o}}_{\text{m}})$-CMA-ES is shown in Fig. 2a. Finally, all the tuning curves from step 2 are collected and the feasible ranges of d_σ are chosen according to three criteria (Brockhoff et al., 2010):

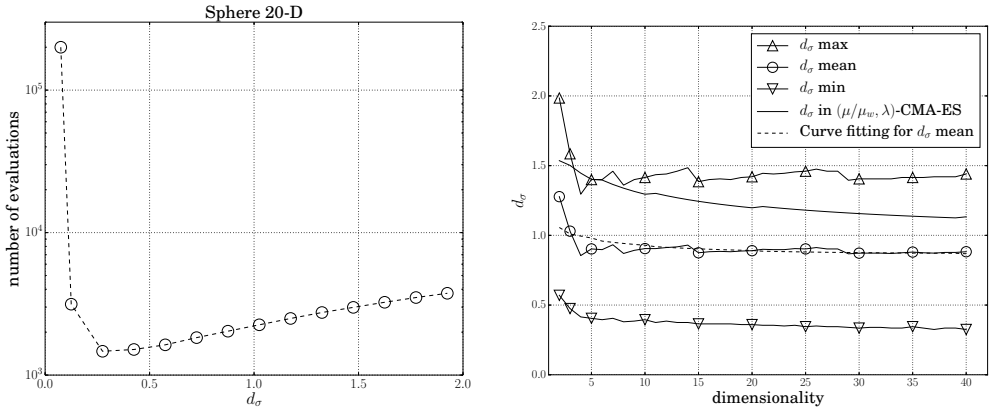


Figure 2: (a): Plot of the number of function evaluations needed to reach the termination criterion of function value 10^{-10} on the 20-D Sphere function, against the candidate d_σ value for $(\mu/\mu_{\text{w}}, \lambda^{\text{o}}_{\text{m}})$-CMA-ES. The values shown are averaged over 64 runs. (b): The feasible range of d_σ (shown by maximum and minimum values), the mean of the feasible range (d_σ mean) and the curve fitting for the mean values over the dimensionality. The default setting for d_σ in $(\mu/\mu_{\text{w}}, \lambda)$-CMA-ES is also illustrated by the solid curve.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

1. Decreasing the selected d_σ from the feasible value by a factor of two leads to a better performance than increasing it by a factor of two.
2. Decreasing the selected d_σ by a factor of three never leads to an observed failure.
3. The selected d_σ should never lead to a performance that is two times slower than the optimal performance in the tuning graph.

The resulting feasible d_σ ranges are labeled as "d_σ max" and "d_σ min" in Fig. 2b. The mean value of the feasible range for each dimension is then selected as the new d_σ setting for the mirrored orthogonal sampling. The result is shown as "d_σ mean". The modified damping factor d_σ is found by fitting the functional form $a + b(\sqrt{(\mu_{\text{eff}} + c)/(n + d)} + e) + c_\sigma$ to the mean values, which results in:

$$d_\sigma = 1.5 - 0.63 \left(\sqrt{\frac{\mu_{\text{eff}} + 0.157}{n + 1.65}} + 0.87 \right) + c_\sigma. \tag{8}$$

Using the same method, we also modify the damping factor for the mirrored sampling technique. The new damping value reads:

$$d_\sigma = 1 - 0.78 \frac{\mu_{\text{eff}}}{\lambda} + c_\sigma. \tag{9}$$

4 Performance analysis

In this section, we analyze the possible performance improvement introduced by mirrored orthogonal sampling. We first give the theoretical analysis for the single-parent evolution strategy and then investigate the multi-parent strategies empirically on the sphere function.

4.1 Theoretical Aspects

The theoretical analysis is twofold. First, the progress rate analysis for $(1, \lambda)$-ES, introduced in (Beyer, 1993), is applied to analyze mirrored sampling. In addition, such analysis gives a straightforward explanation why mirrored orthogonal sampling improves performance. There are no analytical results for mirrored orthogonal sampling yet while its empirical results are compared to random and mirrored sampling. Second, the progress rate analysis is applied again to provide an analytical results about the worst case performance of mirrored orthogonal sampling. This will (partially) explain the advantages of the new sampling method. For the analysis in the following, we will only consider the $(1, \lambda)$-ES with isotropic mutations on the sphere function, which is defined as:

$$f(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*), \quad \mathbf{x} \in \mathbb{R}^n,$$

which has the global minimum $\mathbf{x}^*$. In addition, for the simplicity of our deviation, it is also assumed that the population size λ is **even** in the following analysis. In practice, when λ is odd, the corresponding progress rate can be bounded from below by using $\lambda - 1$ in the analysis and also be bounded from above by using $\lambda + 1$.

Note that although some results (e.g., Fig. 4b) can be equivalently obtained, using the theoretical framework of *convergence rate analysis* (Brockhoff et al., 2010), we did not adopt such an analysis approach because the progress rate analysis gives more insight into why the proposed sampling method outperforms its counterparts. The link between progress rate and convergence rate is elaborated in (Auger and Hansen, 2011). For the convergence rate analysis on the mirrored sampling, please see (Auger et al., 2011a,b).

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

4.1.1 Mirrored Sampling

We will begin with the analysis of the $(1, \lambda_{\mathrm{m}})$-ES in order to show the reason why it outperforms random sampling and this analysis serves as a baseline for the comparison to mirrored orthogonal sampling, which is investigated here by the Monte Carlo simulation. The basics of the analysis are shown in Fig. 3a, following the same treatment as in Bäck (1995). Let $\mathbf{P}$ be the current parent which is at a distance R from the optimum $\mathbf{O}$, namely $\|\mathbf{PO}\| = R$. The mutation distribution is depicted as the hypersphere centered at $\mathbf{P}$ (of radius $\sigma\sqrt{n}$), which represents the mean length of isotropic Gaussian vectors: $\mathbf{z} = \boldsymbol{\mathcal{N}}(\mathbf{0}, \sigma^2\mathbf{I})$. The mirrored mutation is then indicated as $-\mathbf{z}$. The progress made by a mutation $\mathbf{z}$ vector is $R - r$, where r is the new distance to the optimum $\mathbf{O}$ after mutation. Furthermore, in the $(1, \lambda)$-ES, only the best progress among a population of mutations $\{\mathbf{z}_i\}_{i=1}^{\lambda}$ is selected and thus the overall progress of the $(1, \lambda)$-ES is $R - r_{1:\lambda}$ ($r_{1:\lambda}$ is the smallest order statistic among $\{r_i\}_{i=1}^{\lambda}$). The progress rate is defined as the expected progress (Beyer, 2001):

$$\varphi_{1,\lambda} = \mathbb{E}\left\{R - r_{1:\lambda}\right\}.$$

This expectation can be calculated using the following observations: The progress of each mutation can be measured by the projection of $\mathbf{z}$ onto line $\mathbf{PO}$, which is denoted as z in Fig. 3a. Then the smallest order statistic $r_{1:\lambda}$ must be associated with the largest order statistic of the projections: $z_{\lambda:\lambda}$. Note that, for each mutation $\mathbf{z}$, the following relation is established between R, r and z:

$$\|\mathbf{z}\|^2 - z^2 = r^2 - (R - z)^2.$$

Using this relation, the progress rate can be expressed as:

$$\begin{aligned}
\varphi_{1,\lambda} &= \mathbb{E}\left\{R - \sqrt{(R - z_{\lambda:\lambda})^2 + \|\mathbf{z}_{\lambda:\lambda}\|}\right\} \\
&\simeq \mathbb{E}\left\{R - \sqrt{(R - z_{\lambda:\lambda})^2 + \sigma^2 n}\right\} \qquad (10a) \\
&= R\mathbb{E}\left\{1 - \sqrt{1 + \left(\frac{\sigma^2 n}{R^2} - 2\frac{z_{\lambda:\lambda}}{R}\right)}\right\} \\
&\simeq R\mathbb{E}\left\{1 - \left(1 + \frac{\sigma^2 n}{2R^2} - \frac{z_{\lambda:\lambda}}{R}\right)\right\}, \qquad (10b)
\end{aligned}$$

where $\mathbf{z}_{\lambda:\lambda}$ represents the mutation vector that has the largest projection onto direction $\mathbf{PO}$. Note that in the first approximation (Eq. (10a)), $\|\mathbf{z}_{\lambda:\lambda}\|$ is replaced by the mean value of $\|\mathbf{z}\|$, namely $\sigma\sqrt{n}$ and in the second approximation (Eq. (10b)), the first-order Taylor approximation is taken for the square root. In addition, the distribution of $z_{\lambda:\lambda}$ can be easily obtained due to the invariance properties of isotropic Gaussian vectors: z is normally distributed as $\mathcal{N}(0, \sigma^2)$, regardless of the actual direction of $\mathbf{PO}$.

For the mirrored sampling, if z_i is the projection of mutation $\mathbf{z}_i$ onto $\mathbf{PO}$, then the projection of its mirrored mutation $-\mathbf{z}_i$ is $-z_i$ by symmetry. Thus, the set of the projections of all the mutations of mirrored sampling can be written as $\{z_i, -z_i\}_{1 \le i \le \lambda/2}$. Let $P^{\mathrm{m}}_{\lambda:\lambda}(Z \le z)$ denote the cumulative probability distribution (CDF) of the largest order statistic among $\{z_i, -z_i\}_{1 \le i \le \lambda/2}$. Suppose for every $z \ge 0$, in order to facilitate the condition in $P^{\mathrm{m}}_{\lambda:\lambda}(Z \le z)$, namely the largest order statistic is less than or equals to z, we must have $z_i \le z, -z_i \le z$ for all the z_i, which indicates $-z \le z_i \le z$ for all the z_i. The intuition is that all random mutation points are required to be sampled less than

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

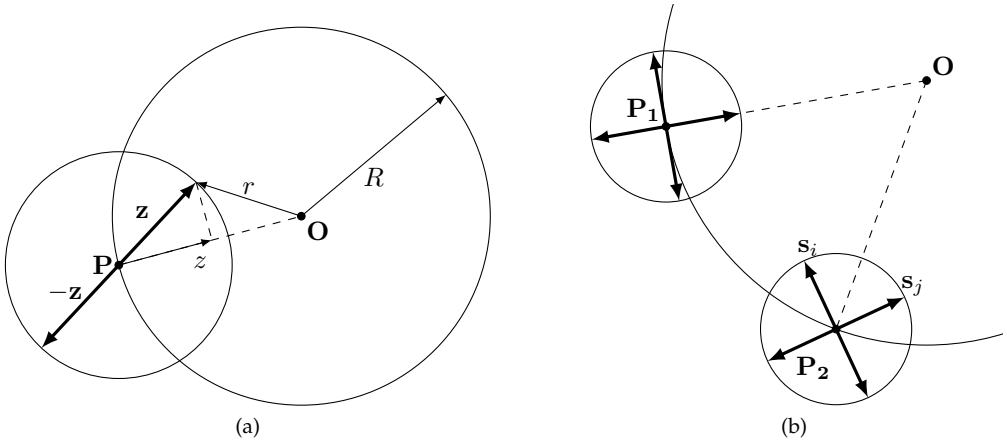


Figure 3: (a). Schematic diagram for the progress rate analysis on the sphere function. The mutations are centered at $\mathbf{P}$, which is at distance R from the optimum $\mathbf{O}$. (b) In $2D$, the diagram shows the best case ($\mathbf{P}_1$) of progress and the worst case ($\mathbf{P_2}$) for mirrored orthogonal sampling on the sphere function.

or equal to z. In addition, because mirrored mutations are generated by reversing the signs of random mutations, every random mutation also needs to be bigger than $-z$, otherwise the mirrored counterpart of an outlier would be larger than z and fails the condition. The argument reads,

$$\begin{aligned} P^{\mathrm{m}}_{\lambda:\lambda}(Z \leq z) &= \left[\Pr\left(-z < Z \leq z\right)\right]^{\lambda/2} \\ &= \left[\Phi\left(\frac{z}{\sigma}\right) - \Phi\left(-\frac{z}{\sigma}\right)\right]^{\lambda/2} \\ &= \left[2\Phi\left(\frac{z}{\sigma}\right) - 1\right]^{\lambda/2}, \quad \forall z \geq 0. \end{aligned}$$

Note that $\Phi(\cdot)$ stands for the CDF of a standard normal random variable. Then, in case of $z < 0$, the cumulative probability should be always 0. The reason is that if an original mutation is negative, then its mirrored counterpart would be positive. Therefore the largest order statistics could not be negative ever. In total, the CDF of the largest order statistic is summarized as:

$$P^{\mathrm{m}}_{\lambda:\lambda}(Z \leq z) = \begin{cases} \left[2\Phi\left(\frac{z}{\sigma}\right) - 1\right]^{\lambda/2} & \forall z \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

And its probability density function is:

$$p^{\mathrm{m}}_{\lambda:\lambda}(z) = \begin{cases} \lambda p(\frac{z}{\sigma}) \left[2\Phi\left(\frac{z}{\sigma}\right) - 1\right]^{\lambda/2-1} & \forall z \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

where $p(\cdot)$ denotes the probability density function (PDF) of a standard normal distribution. This density can be compared to the largest order statistic among the same projections of random samples (Beyer, 1993):

$$p_{\lambda:\lambda}(z) = \lambda p\left(\frac{z}{\sigma}\right) \Phi\left(\frac{z}{\sigma}\right)^{\lambda-1}.$$

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

In 5-D with $\lambda = 10$, we plot the CDF and density function of mirrored sampling and random sampling in Fig. 4a. It is clear from the figure that the the distribution of the largest projection for mirrored sampling is shifted to the right, compared to that for the Gaussian sampling and therefore the corresponding distributions of projections is shifted towards larger values. This advantage would affect the progress rate (as shown in the following) and is the main reason why the mirrored sampling technique has a better performance than the simple random sampling. Substituting density function $p^{\mathrm{m}}_{\lambda:\lambda}$ into Eq. (10b) and using the normalized quantities as in Beyer (1993),

$$\varphi^* = \varphi\frac{n}{R}, \quad \sigma^* = \sigma\frac{n}{R},$$

the normalized progress rate of $(1, \lambda_{\mathrm{m}})$-ES can be obtained:

$$\begin{aligned}
\varphi^*_{1,\lambda_{\mathrm{m}}} &= \frac{n}{R}\left(R\int_0^\infty \left(\frac{z}{R} - \frac{\sigma^2 n}{2R^2}\right) p^{\mathrm{m}}_{\lambda:\lambda}(z)\,\mathrm{d}z\right) \\
&= \frac{n}{R}\int_0^\infty z p^{\mathrm{m}}_{\lambda:\lambda}(z)\,\mathrm{d}z - \frac{(\sigma^*)^2}{2} \\
&= \lambda\frac{n}{R}\int_0^\infty z p\left(\frac{z}{\sigma}\right)\left[2\Phi\left(\frac{z}{\sigma}\right) - 1\right]^{\lambda/2-1}\,\mathrm{d}z - \frac{(\sigma^*)^2}{2} \\
&= \sigma\frac{n}{R}\left(\lambda\int_0^\infty z' p(z')\left[2\Phi\left(z'\right) - 1\right]^{\lambda/2-1}\,\mathrm{d}z'\right) - \frac{(\sigma^*)^2}{2} \\
&= c_{1,\lambda_{\mathrm{m}}}\sigma^* - \frac{(\sigma^*)^2}{2}. \qquad (12)
\end{aligned}$$

In the equation above, the integral about the normalized largest projection $z' = z/\sigma$ computes its expectation and it is known as the *progress coefficient* from (Beyer, 1993). We denote it by $c_{1,\lambda_{\mathrm{m}}}$ here. It can be compared to the progress coefficient of random sampling, which reads:

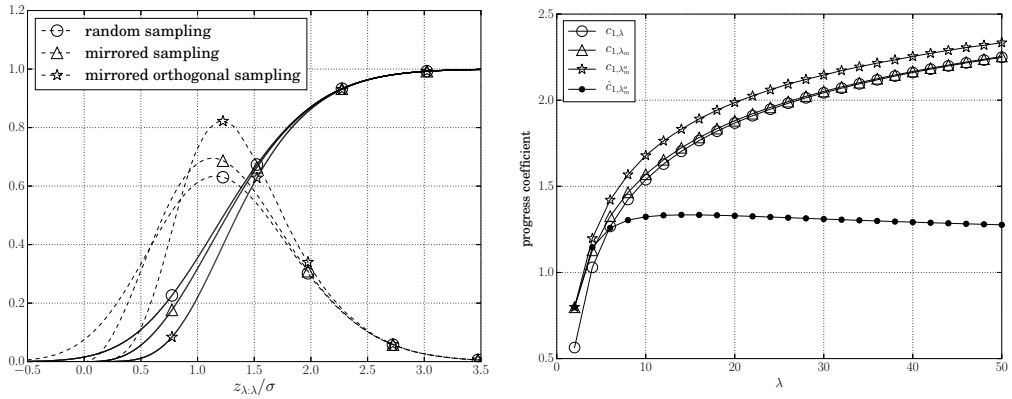
$$c_{1,\lambda} = \lambda\int_{-\infty}^{\infty} z p(z)\Phi(z)^{\lambda-1}\,\mathrm{d}z.$$

Note that the progress rate of random sampling can be easily obtained by replacing $c_{1,\lambda_{\mathrm{m}}}$ in Eq. (12) with $c_{1,\lambda}$.

Numerically, we plot the progress coefficients of random sampling and mirrored sampling against population size in Fig. 4b. The mirrored sampling (the curve marked by triangles) shows a small yet obvious advantage compared to the random sampling for small population sizes. In larger populations, these two converging curves imply that mirrored sampling provides no speed-up to the ES algorithm. Thus, the application of mirrored sampling should be limited to the small population setting.

For mirrored orthogonal sampling, we would like to use the same approach as for the mirrored sampling analysis above. However, it is hard to analytically obtain the CDF and the density function of the largest projection onto $\mathbf{PO}$ of the mirrored orthogonal sampling. Therefore, we compute its CDF and density function empirically by Monte-Carlo simulation. For the simulation, the population size λ is set to $2N$. The mirrored orthogonal samples are projected onto $\mathbf{PO}$ and the largest projections are stored, from which the CDF is estimated. The results are also summarized in Fig. 4. In Fig. 4a, the CDF of mirrored orthogonal sampling (the solid curve marked by stars) is more likely to distribute samples towards bigger values compared to the CDF of mirrored sampling. As a consequence, in Fig. 4b, the progress coefficients of mirrored

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251



(a) (b)

Figure 4: (a): The CDFs (solid) and PDFs (dashed) of the largest projection (normalized) onto $\mathbf{PO}$ for random, mirrored and mirrored orthogonal sampling. The dimension n is set to 5 and $\lambda = 10$ for all curves. 10^6 trials are used in the estimation for mirrored orthogonal sampling. For the rest sampling method, the curves plot the corresponding analytical results. (b): Progress coefficients against population size λ for random sampling, mirrored sampling and mirrored orthogonal sampling. The dimensionality n is set to $\lambda/2$ for all curves. The curve marked by black dots is the lower bound of the progress coefficients for mirrored orthogonal sampling.

orthogonal sampling are significantly bigger than those of mirrored sampling, even in a large population.

4.1.2 Mirrored orthogonal sampling: the worst case analysis

The worst case analysis of mirrored orthogonal sampling is conducted when the population is set to $2n$. We will call such population setting as "full mutations". Under this condition, the progress rate is maximized (as will be explained later) and it is possible to provide analytical results. The progress under the condition $\lambda < 2n$ will be also discussed later.

In $2D$ with $\lambda = 4$, the worst case (together with best case) of progress for $(1, \lambda^{\text{o}}_{\text{m}})$ is shown in Fig. 3b. Suppose there is no step size σ (σ=1) involved here for simplification. In the mutations centered at $\mathbf{P_1}$, there is one mutation pointing to the optimum $\mathbf{O}$ and therefore this mutation performs optimally. We call this mutation scenario the best case of progress. The progress coefficient in this case is the expectation of the standard norm mutation length. It serves as the upper bound of the progress coefficient and is the same for random, mirrored and mirrored orthogonal sampling.

The worst case of progress is indicated by the mutations centered at $\mathbf{P_2}$ in which the angle formed by the line segment $\mathbf{P_2O}$ and mutation $\mathbf{s}_i$ is the same as the one ($\pi/4$ as shown in the figure) formed by $\mathbf{P_2O}$ and $\mathbf{s}_j$. In this scenario, the expected projections of $\mathbf{s}_i$ and $\mathbf{s}_j$ are the same. It is not possible to make the expected projection of one mutation smaller without rendering the expected projection of the other one larger. For example, if we rotate $\mathbf{s}_j$ a little bit clockwise, then its projection becomes smaller. However, in the meanwhile $\mathbf{s}_i$ is also rotated and its projection gets larger. Consequently, the largest projection of all the mutations becomes larger. Therefore,

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

among all the possible mutation scenarios, $\mathbf{P}_2$ gives the lower bound of the largest projection of mutations onto $\mathbf{P_2O}$. Recall from Eq. (**??**) that the progress made by $(1, \lambda)$-ES is determined by the largest projection. Thus, the scenario $\mathbf{P}_2$ is the worst case of progress.

Under the "full" mutation condition, we generalize the worst case for arbitrary dimensions. Let the mirrored orthogonal samples be denoted as $\{\boldsymbol{\mathcal{O}}_i, -\boldsymbol{\mathcal{O}}_i\}_{1\leq i\leq \lambda/2}$. The unit vectors along the orthogonal mutations are defined as:

$$\mathbf{u}_i = \frac{\boldsymbol{\mathcal{O}}_\mathbf{i}}{\|\boldsymbol{\mathcal{O}}_i\|}. \tag{13}$$

Combining the unit vectors for mirrored mutations, all the unit vectors are $\{\mathbf{u}_i, -\mathbf{u}_i\}_{1\leq i\leq \lambda/2}$. The worst case of progress is defined by the following conditions: for all the unit vectors, the linear combination with equal weights (denoted as $\mathbf{d}$ in the following) of $\lambda/2 = n$ unit vectors points to the optimum $\mathbf{O}$ and also to the reverse direction of the gradient of the sphere function, which reads:

$$\mathbf{d} = \sum_{k=1}^{\lambda/2} a_k \mathbf{u}_k = -\alpha \nabla f(\mathbf{x}), \quad \alpha > 0, a_k = \pm 1,$$

where a_k is a sign operator to select among $\mathbf{u}_k, -\mathbf{u}_k$. Then the scalar projection of mutation $\boldsymbol{\mathcal{O}}_i$ onto $\mathbf{d}$ is expressed as:

$$\mathrm{proj}_\mathbf{d}(\boldsymbol{\mathcal{O}}_i) = \frac{\langle \boldsymbol{\mathcal{O}}_i, \mathbf{d}\rangle}{\|\mathbf{d}\|} = \frac{\sum_{k=1}^{\lambda/2} a_k \langle \boldsymbol{\mathcal{O}}_i, \mathbf{u}_k\rangle}{\left\|\sum_{k=1}^{\lambda/2} a_k \mathbf{u}_k\right\|} = \frac{\sum_{k=1}^{\lambda/2} a_k \langle \boldsymbol{\mathcal{O}}_i, \boldsymbol{\mathcal{O}}_k\rangle / \|\boldsymbol{\mathcal{O}}_k\|}{\left\|\sum_{k=1}^{\lambda/2} a_k \mathbf{u}_k\right\|} = \frac{a_k \|\boldsymbol{\mathcal{O}}_i\|}{\sqrt{n}}.$$

Note that we substitute the expression of $\mathbf{u}_i$ (Eq. (13)) in the derivation above. The projections of all the mutations onto $\mathbf{d}$ can be summarized as:

$$\mathrm{proj}_\mathbf{d} = \left\{ \frac{\|\boldsymbol{\mathcal{O}}_i\|}{\sqrt{n}}, -\frac{\|\boldsymbol{\mathcal{O}}_i\|}{\sqrt{n}} \right\}_{i=1}^{\lambda/2}.$$

The largest order statistic of all the projections is the maximum of $\mathrm{proj}_\mathbf{d}$:

$$\begin{aligned}
\max\{\mathrm{proj}_\mathbf{d}\} &= \max_{1\leq i\leq \lambda/2} \left\{ \frac{\|\boldsymbol{\mathcal{O}}_i\|}{\sqrt{n}}, -\frac{\|\boldsymbol{\mathcal{O}}_i\|}{\sqrt{n}} \right\} \\
&= \frac{1}{\sqrt{n}} \max_{1\leq i\leq \lambda/2} \{\|\boldsymbol{\mathcal{O}}_i\|\} \\
&= \frac{z}{\sqrt{n}}.
\end{aligned}$$

We denote the maximal mutation length as z above. Note that the $\|\boldsymbol{\mathcal{O}}_i\|$ are independently distributed according to $\chi(n)$ (see Alg. 4). Therefore, the density function of the maximal mutation length among $\lambda/2$ mutations reads:

$$p_{\frac{\lambda}{2}:\frac{\lambda}{2}}(z) = \frac{\lambda}{2} p_\chi(z) \left(F_\chi(z)\right)^{\frac{\lambda}{2}-1},$$

where $p_\chi(\cdot), F_\chi(\cdot)$ denote the density and CDF of the $\chi(n)$ distribution, respectively. The worst case progress coefficient of mirrored orthogonal sampling, which is the ex-

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

pectation of $z/\sqrt{n}$, is denoted as $\hat{c}_{1,\lambda_{\mathrm{m}}^{\mathrm{o}}}$ and derived as follows:

$$\begin{aligned}\hat{c}_{1,\lambda_{\mathrm{m}}^{\mathrm{o}}} &= \int_0^\infty \frac{z}{\sqrt{n}} p_{\frac{\lambda}{2}:\frac{\lambda}{2}}(z)\,\mathrm{d}z \\ &= \frac{\lambda}{2\sqrt{n}} \int_0^\infty z p_\chi(z)\left(F_\chi(z)\right)^{\frac{\lambda}{2}-1}\,\mathrm{d}z \\ &= \sqrt{n}\int_0^\infty z p_\chi(z)\left(F_\chi(z)\right)^{n-1}\,\mathrm{d}z. \end{aligned} \tag{14}$$

The last equation results from the fact that we picked the special population size $\lambda = 2n$ from the previous analysis setting. Eq. (14) is numerically evaluated and plotted in Fig. 4b. The curve for the worst case is above 1 and roughly stays constant when λ increases. It provides a non-zero lower bound of the progress coefficient of mirrored orthogonal sampling with "full mutations", which indicates no matter in what scenario, the mirrored orthogonal sampling with "full mutations" is going to guarantee positive progress on the sphere function. To compare, for random sampling, the lower bound of the progress coefficient is zero because it is possible to have all the mutations generated as in Fig. 1, where no mutation makes progress. For mirrored sampling, the lower bound of the progress coefficient is also zero because it is possible that all the mutations are generated in a tangent space of the local gradient, in which all the vectors are orthogonal to the gradient. Thus, the non-zero lower bound of mirrored orthogonal sampling with "full mutations" is its main advantage over the random and mirrored sampling.

In the case that mirrored orthogonal sampling does not use "full mutations", namely $\lambda < 2n$, the progress rate would be reduced in contrast to the "full mutations" case. This is because it can now happen that some subspace could not be covered when $\lambda < 2n$. Therefore, it is possible that the subspace in which the progress can be made is simply unexplored.

4.2 Empirical Aspects

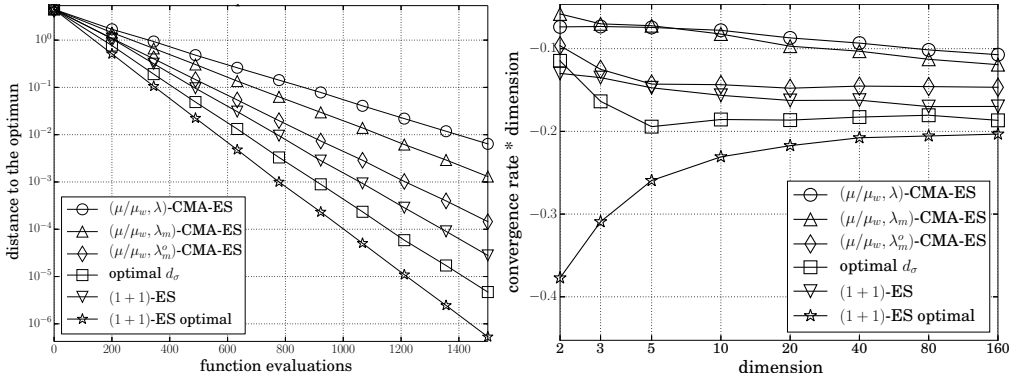
For the multi-parental variants of ES, we only consider their empirical convergence rates here. Similar to the convergence rate estimation in (Loshchilov et al., 2011), the effect of the mirrored orthogonal sampling technique on the sphere function is investigated empirically by incorporating it into the well-known CMA-ES algorithm.

On the 20-D sphere function, the convergence rates of the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}}^{\mathrm{o}})$-CMA-ES and other comparable ES variants are illustrated in Fig. 5a. The empirical convergence rate is estimated as the average slope of convergence curve over 200 runs. For all the CMA-ES variants tested here, the default settings of population size are applied (Hansen, 2006): $\lambda = 4 + \lfloor 3\ln n \rfloor, \mu = \lfloor \lambda/2 \rfloor$. The legend "$(1+1)$-ES" represents the $(1+1)$-ES with $1/5$ success rule step size control while the "$(1+1)$-ES optimal" is for the $(1+1)$-ES with scale-invariant step size setting $\sigma = \frac{1.2}{n}\|\mathbf{x}^{(k)}\|$, which proves to be the optimal step size setting on the sphere function (Loshchilov et al., 2011). The pairwise selection is always used if the mirroring operation is present in the sampling procedure. The mirrored sampling CMA-ES with modified d_σ (Eq. (9)) is denoted as "$(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES ". The curve labeled by "$(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}}^{\mathrm{o}})$-CMA-ES " stands for the mirrored orthogonal CMA-ES with modified d_σ (Eq. (8)). In addition, "optimal d_σ" represents the mirrored orthogonal CMA-ES using the optimal d_σ tuning on the sphere function, corresponding to the minimal value of the tuning curve in Fig. 2a. Due to the empirical results, the convergence of $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}}^{\mathrm{o}})$-CMA-ES (marked by diamond) is slower but close to that of

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

the $(1+1)$-ES (marked by upside-down triangle) while the $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES using the optimal parameter settings gradually catches the convergence rates of the optimal $(1+1)$-ES in high dimensions.

The relation between the empirical convergence rate and the dimensionality is shown in Fig. 5b. The algorithms tested here is the same as Fig. 5a. It is obvious that there is a leap of convergence rates between the CMA-ES and its mirrored orthogonal competitor. The advantages of the mirrored orthogonal CMA-ES over the mirrored CMA-ES are significant and preserved even for large dimensions. The upper limit of the $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES on the sphere function is shown by the convergence rates achieved under the optimal d_σ tuning, which is even better than $(1+1)$-ES for almost all the dimensions. However, the optimal d_σ setting on the sphere function turned out to be not robust when considering other fitness functions and therefore is not used.



(a) Sphere 20-D (b)

Figure 5: The comparison of empirical convergence rates on the sphere function. All the results are estimated over 200 runs. The suggested λ setting $4 + \lfloor 3 \ln N \rfloor$ (Hansen, 2006) is used for all the CMA-ES variants (a): Plot of the average distance (over 200 runs) to the global optimum against the number of function evaluations for four ES algorithms: $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES with tuned d_σ and optimal d_σ, $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES, standard $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES and $(1+1)$-ES in dimension 20. (b): Plot of convergence rate $\times$ dimensionality against the dimensionality for different algorithms on the sphere function, using 1500 function evaluation.

5 Experimental Validation

The mirrored orthogonal version of CMA-ES with pairwise selection has been tested on the noiseless Black-Box Optimization Benchmark (BBOB) (Hansen et al., 2010). By using the automatic comparison procedures provided in this benchmark, the BBOB results of $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES are compared to those of $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES and $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES.

5.1 Experimental Settings

The three algorithms, $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES, $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES and $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES are benchmarked on BBOB-2012[3] and their results are compared and processed by

[3]The exact version is v11.06.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

the post-processing procedure of BBOB.

The BBOB parameter settings of the experiment are the same for all the tested ES variants. The initial global step size σ is set to 1. The maximum number of function evaluations is set to $10^4 \times n$. The initial solution (initial parent) is a uniformly sampled in the hyper-box $[-4, 4]^n$. The dimensions tested in the experiment are $n \in \{2, 3, 5, 10, 20, 40\}$.

In addition, two independent but similar experiments are conducted. In the first experiment, the default population size setting, rounded logarithm of dimensionality, is used to configure all three algorithms. The result of this experiment is denoted as **small population** in the following. In this experiment, the strategy parameters are the same except for the d_σ setting. The default setting $d_\sigma = 1 + 2\max\{0, \sqrt{(\mu_w - 1)/(n + 1)}\} + c_\sigma$ is used in the standard CMA-ES while the modified d_σ, as stated in Eq. (8) and (9), are used for mirrored and mirrored orthogonal sampling, respectively. Another experiment exploits a relatively large population size, namely $2N$, the result of which is denoted as **large population**. In this experiment, the strategy parameters used are exactly the same for the three ES variants. The modified d_σ is not used because it is tuned under the default population setting instead of the large population setting.

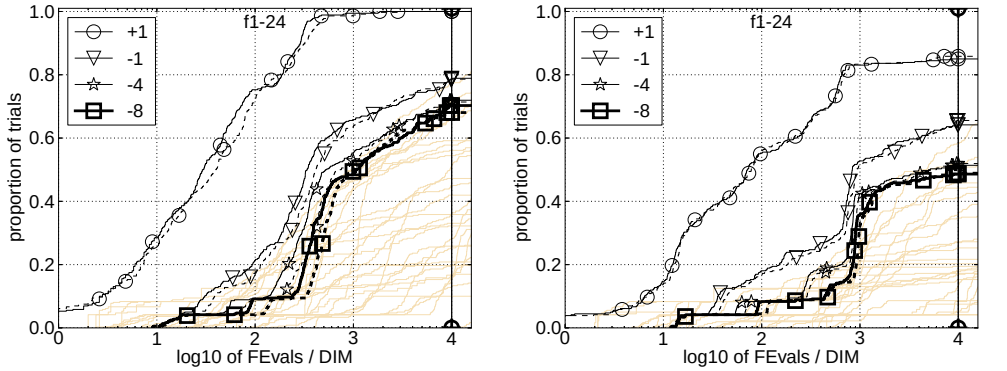
5.2 Results and Discussion

The BBOB noiseless testbed (Hansen et al., 2009) contains 24 test functions which are classified into several groups as separable, ill-conditioned or multi-modal functions. Due to space limitations, only the comparisons of the aggregated empirical cumulative distributions (ECDFs) of run length over all the test functions are presented here. The ECDFs of run length estimates the cumulative distribution of the function evaluations consumed in ESs, with respect to a given precision target.

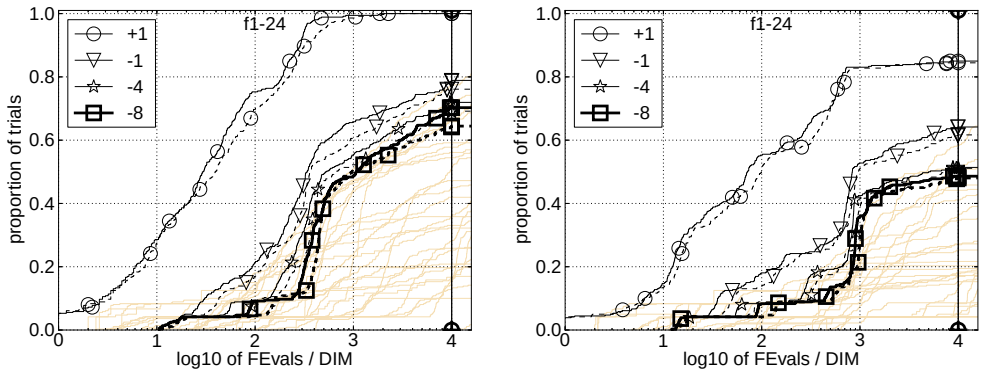
Small population. The results under the default small population setting are shown in Fig. 6. The comparison between the mirrored orthogonal sampling and the mirrored sampling is shown in Fig. 6a, 6b. Four different target precision values (10^k with $k \in \{1, -1, -4, -8\}$) are presented. On the left side, the comparisons under 5-D indicate a big performance improvement by the mirrored orthogonal sampling, which holds for all the target precisions. On the right side, the situation in 20-D still shows small advantages of the mirrored orthogonal sampling technique over the other algorithms. As for comparison between the mirrored orthogonal sampling and the standard CMA-ES, Fig. 6c, 6d gives the results. The comparison here shows approximately the same results as in Fig. 6a, 6b. The improvement introduced by mirrored orthogonal sampling is decreasing when the dimensionality increases.

Large population. For the cases where the population size is linearly related to the dimensionality, we are mainly interested in validating the theoretical performance advantage of the mirrored orthogonal sampling (Sec. 4.1.1 and 4.1.2). Thus, the results of the original $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES is not shown here. The results are illustrated in Fig. 7. From the comparisons between the ECDFs of 5-D (left half) to that of 20-D (right half), it is obvious that the amount of the improvement is still significant when the dimensionality goes large. The more detailed results in 5-D, which are shown in Fig. 8, indicate that the mirrored orthogonal sampling technique outperforms its mirrored counterpart on almost all the test functions: highly-conditioned functions $f10$-$f14$, multi-modal functions with adequate global structure $f15$-$f19$, separable functions $f1$-$f5$ and multi-modal functions with weak global structure $f20$-$f24$. The detailed results in 10-D, as summarized in Fig. 9, shows roughly the same comparisons as that in 5-D expect that it is hard to judge which algorithm is better from the ECDFs of the multi-modal functions

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251



(a) 5-D: $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES ,$(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES (b) 20-D: $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES , $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES



(c) 5-D: $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES , $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES (d) 20-D: $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES , $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES

Figure 6: Left column: $n = 5$. Right column: $n = 20$. For the small population, sub-figures (a), (b) show the ECDFs of run lengths averaged over all the test functions ($f1$-24) for $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES (solid lines) and $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES (dashed lines) while sub-figures (c), (d) compare that for $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES (solid lines) and $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES (dashed lines). ECDF of run lengths (the number of function evaluations divided by dimension) for each algorithm needed to reach a target value are counted for four target precisions: $f_{\mathrm{opt}} + \Delta f$ with $\Delta f = 10^k$, where $k \in \{1, -1, -4, -8\}$ is given by the value in the legend. The vertical black line indicates the maximum normalized run length. Light beige lines show the ECDF of run lengths for target value $\Delta f = 10^{-8}$ of all algorithms benchmarked during BBOB-2009.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

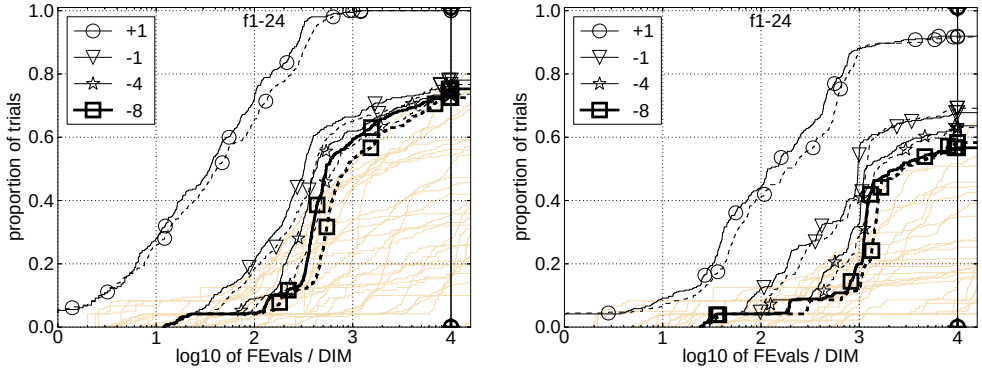


Figure 7: Left: $n = 5$. Right: $n = 20$. For the large population, the empirical cumulative distributions (ECDF) of run lengths (the number of function evaluations divided by dimension) for $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES (solid lines) and $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES (dashed lines) needed to reach a target value. All the figure settings are the same as in Fig. 6.

with adequate global structure $f15$-$f19$ (Fig. 9c).

The better experimental results for a large population suggest that the newly proposed mirrored orthogonal sampling technique would be most suitable in the case where the population size is about two times the dimensionality.

6 Discussion and Conclusion

In this paper, we propose a new mutation operator, the mirrored orthogonal sampling to generated evenly distributed samples for evolution strategies. Several approaches, including the mirrored sampling, to achieve derandomized sampling are briefly introduced. By the theoretical analysis, we have shown that the performance improvement given by the mirrored sampling vanishes in a large population setting by theoretical analysis. As a remedy to this limitation, random orthogonal samples are introduced as a possible improvement of mirrored sampling. Pairwise selection is also used to avoid the undesired bias caused by the mirroring operation. The resulting algorithm, called the mirrored orthogonal sampling, is applied to the CMA-ES after some parameter tuning. The performance of random, mirrored and mirrored orthogonal sampling are compared both analytically and empirically. On the sphere function, the $(1, \lambda)$-ES with the mirrored orthogonal sampling is just a little bit slower than the $(1+1)$-ES with $1/5$ rule, as shown in the empirical analysis. Finally, we tested the $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES on the BBOB benchmark regarding its performance for small population size and for large population size. The results reveal the advantages of the mirrored orthogonal sampling over mirrored sampling and over the standard $(\mu/\mu_{\mathrm{w}}, \lambda)$-CMA-ES. In particular on highly conditioned and multi-modal functions the competitiveness of the new mirrored orthogonal sampling becomes significant. As discussed in the theoretical analysis (Sec. 4.1), the proposed method is well-suited for the problem where the dimensionality is larger than or similar to half of the population size. However, in very high dimensions, the advantage of the new method gradually diminishes.

Some interesting future directions can be identified, based on the suggested new method of generating mutations. First, the pairwise selection method is chosen here for avoiding the undesired bias. A more advanced idea introduced in Auger et al. (2011b), *selective mirroring*, is also suitable option for being using in mirrored orthogonal sam-

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

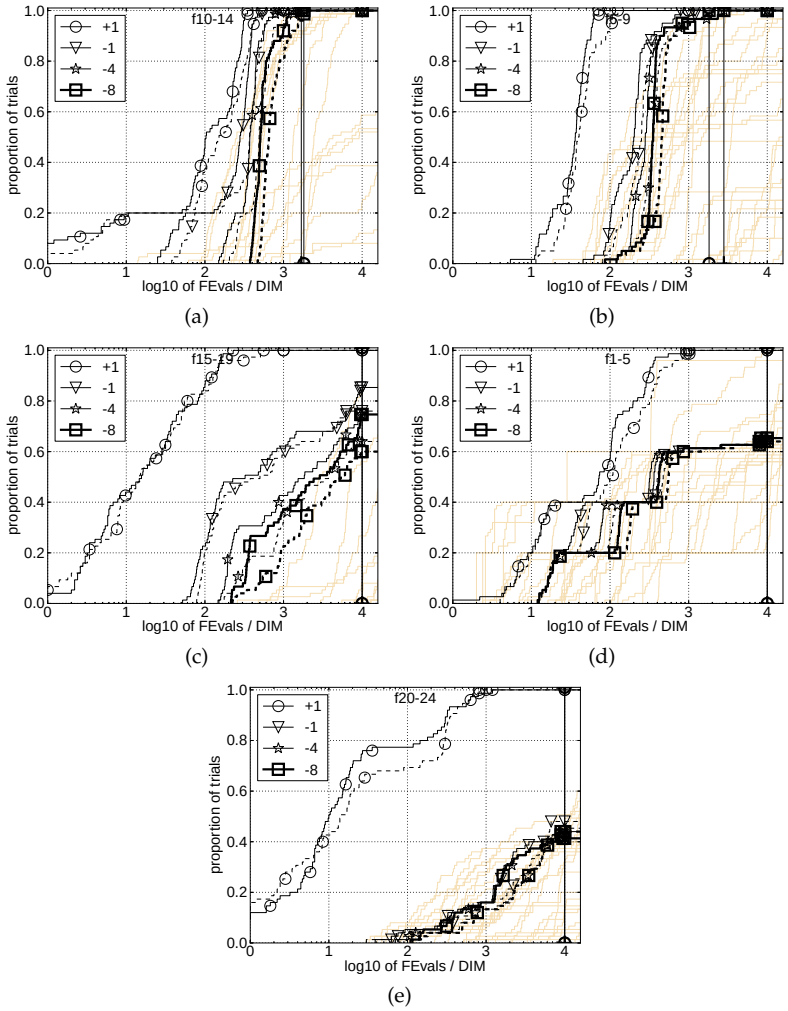


Figure 8: The figures show the details of Fig. 7, left: in 5-D, the ECDFs of run lengths for $(\mu/\mu_{\mathrm{w}}, \lambda^{\mathrm{o}}_{\mathrm{m}})$-CMA-ES (solid lines) and $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES (dashed lines) are shown for each function class: (a) functions with high conditioning, (b) functions with low or moderate conditioning, (c) multi-modal functions with adequate global structure, (d) separable functions and (e) multi-modal functions with weak global structure. The figure settings are the same as Fig. 6.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

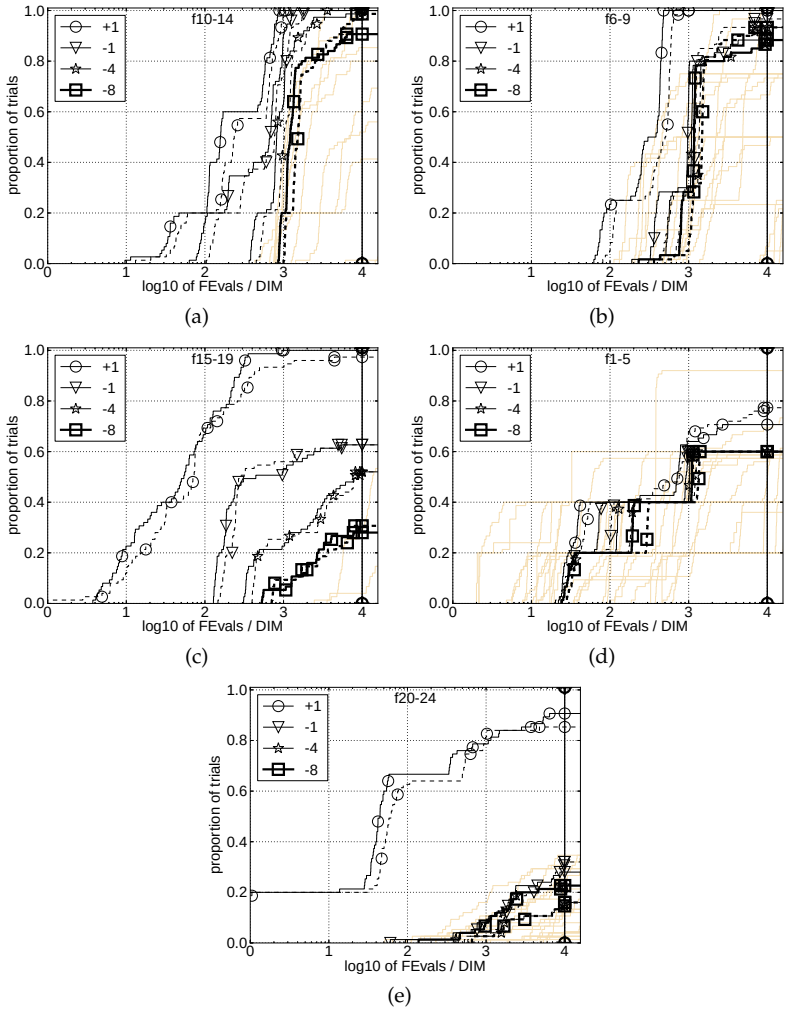


Figure 9: The figures show the details of Fig. 7, right: in 20-D, the ECDFs of run lengths for $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}}^{\mathrm{o}})$-CMA-ES (solid lines) and $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{m}})$-CMA-ES (dashed lines) are shown for each function class: (a) functions with high conditioning, (b) functions with low or moderate conditioning, (c) multi-modal functions with adequate global structure, (d) separable functions asnd (e) multi-modal functions with weak global structure. The figure settings are the same as Fig. 6.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

pling. More work is needed to identify the best possible selection method for mirrored orthogonal sampling.

Second, some more parameter tuning should be done. The learning rates c_1, c_μ for rank-one and rank-μ update of the covariance matrix remain unchanged from their suggested settings. It is important to adapt those parameters to the new sampling technique to obtain the best possible speed-up of the algorithm.

Third, concerning the progress rate analysis (Sec. 4.1), deriving the distribution function of the uniform random orthogonal vectors still remains an open problem. The exact progress rate formula for mirrored orthogonal sampling is unknown. This is planned as another part of the further work. Finally, it would be interesting to apply the mirrored orthogonal sampling to more recent CMA-ES variants such as the active CMA-ES.

Acknowledgment

The authors would like to gratefully thank for the financial support by the Dutch Research Project (NWO) PROMIMOOC (project number 650.002.001).

References

Auger, A., Brockhoff, D., and Hansen, N. (2010). Mirrored Variants of the (1,2)-CMA-ES Compared on the Noiseless BBOB-2010 Testbed. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, GECCO '10, pages 1551–1558, New York, NY, USA. ACM.

Auger, A., Brockhoff, D., and Hansen, N. (2011a). Analyzing the Impact of Mirrored Sampling and Sequential Selection in Elitist Evolution Strategies. In Beyer, H.-G. and Langdon, W. B., editors, *Proceedings of the 11th workshop proceedings on Foundations of genetic algorithms*, FOGA '11, pages 127–138, ACM, New York.

Auger, A., Brockhoff, D., and Hansen, N. (2011b). Mirrored Sampling in Evolution Strategies with Weighted Recombination. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11, pages 861–868, New York, NY, USA. ACM.

Auger, A. and Hansen, N. (2011). Theory of Evolution Strategies: A New Perspective. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, 1:289–325.

Bäck, T. (1995). Order Statistics for Convergence Velocity Analysis of Simplified Evolutionary Algorithms. volume 3 of *Foundations of Genetic Algorithms*, pages 91 – 102. Elsevier.

Bäck, T., Foussette, C., and Krause, P. (2013). *Contemporary Evolution Strategies*. Springer Publishing Company, Incorporated.

Beyer, H.-G. (1993). Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1,^+\lambda)$-Theory. *Evolution Computation*, 1(2):165–188.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies*. Springer-Verlag Berlin Heidelberg, 1 edition.

Björck, A. (1994). Numerics of Gram-Schmidt orthogonalization. *Linear Algebra and its Applications*, 197-198:297–316.

Brockhoff, D., Auger, A., Hansen, N., Arnold, D. V., and Hohm, T. (2010). Mirrored Sampling and Sequential Selection for Evolution Strategies. In Schaefer, R., Cotta, C., Kolodziej, J., and Rudolph, G., editors, *Proceedings of the 11th international conference on Parallel problem solving from nature: Part I*, PPSN'10, pages 11–21, Springer-Verlag, Berlin.

Dick, J. and Pillichshammer, F. (2010). *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, New York, NY, USA.

Evolutionary Computation Just Accepted MS.
https://doi.org/10.1162/evco_a_00251

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley, New York.

Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In Lozano, J. A., Larrañaga, P., Inza, I., and Bengoetxea, E., editors, *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, pages 75–102. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hansen, N. (2008). Adaptive Encoding: How to Render Search Coordinate System Invariant. In *Parallel Problem Solving from Nature - PPSN X*, pages 205–214, Dortmund, Germany.

Hansen, N., Auger, A., Finck, S., and Ros, R. (2010). Real-Parameter Black-Box Optimization Benchmarking 2010: Experimental Setup. Technical Report RR-7215, INRIA.

Hansen, N., Finck, S., Ros, R., and Auger, A. (2009). Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions. Technical Report RR-6829, INRIA.

Hansen, N., Müller, S. D., and Koumoutsakos, P. (2003). Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evol. Comput.*, 11(1):1–18.

Hansen, N. and Ostermeier, A. (2001). Completely Derandomized Self-Adaptation in Evolution Strategies. *Evol. Comput.*, 9(2):159–195.

Kimura, S. and Matsumura, K. (2005). Genetic Algorithms Using Low-discrepancy Sequences. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, GECCO '05, pages 1341–1346, New York, NY, USA. ACM.

Loshchilov, I., Schoenauer, M., and Sebag, M. (2011). Adaptive Coordinate Descent. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 885–892. ACM.

Niederreiter, H. (1992). *Random Number Generation and quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Rahnamayan, S., Tizhoosh, H. R., and Salama, M. M. (2006). Opposition-Based Differential Evolution Algorithms. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pages 2010–2017. IEEE.

Rosenbrock, H. H. (1960). An Automatic Method for Finding the Greatest or Least Value of a Function. *The Computer Journal*, 3(3):175–184.

Schwefel, H.-P. (1993). *Evolution and Optimum Seeking: The Sixth Generation*. John Wiley & Sons, Inc., New York, NY, USA.

Teytaud, O. and Gelly, S. (2007). DCMA: yet another derandomization in covariance-matrix-adaptation. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 955–963. ACM.

Wang, H., Emmerich, M., and Bäck, T. (2014). Mirrored Orthogonal Sampling with Pairwise Selection in Evolution Strategies. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pages 154–156, New York, NY, USA. ACM.