Mirrored Orthogonal Sampling with Pairwise Selection in Evolution Strategies

Hao Wang
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
h.wang@liacs.leidenuniv.nl

Michael Emmerich
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
m.t.m.emmerich@liacs.leidenuniv.nl

Thomas Bäck
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
t.h.w.baek@liacs.leidenuniv.nl

ABSTRACT
In this paper, an improvement of the mirrored sampling method, called mirrored orthogonal sampling, is proposed. The convergence rates on the sphere function are estimated. It is also applied to Covariance Matrix Adaptation Evolution Strategy (CMA-ES). The resulting algorithm, termed (µ/µw, λo)-CMA-ES, is benchmarked on the Black-box optimization benchmark (BBOB). The newly proposed technique is found to outperform both the standard (µ/µw, λ)-CMA-ES and its mirrored variant.

Categories and Subject Descriptors
I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search—heuristic methods; G.1.6 [Numerical Analysis]: Optimization—global optimization, unconstrained optimization

General Terms
Algorithms

Keywords
Evolution strategies, derandomization, orthogonal sampling, mirrored sampling, empirical study

1. INTRODUCTION
The recently proposed mirrored sampling technique is a simple and effective derandomized sampling method to accelerate the convergence speed for evolution strategies and has been theoretically proven to accelerate the convergence [1, 5]. The purpose of this paper is to introduce an improvement of the mirrored sampling technique, named mirrored orthogonal sampling to generate more evenly distributed samples.

2. MIRRORED SAMPLING
The mirrored sampling technique is introduced by Auger et al. [5]. Instead of generating λ independent and identically distributed (i.i.d.) search points, only half of the i.i.d. samples are created, namely \( z_i \sim \mathcal{N}(0, \sigma^2 \mathbf{C}) \), \( 1 \leq i \leq \lambda/2 \), where \( \sigma \) is the global step-size and \( \mathbf{C} \) is the covariance matrix. Each mutation vector \( z_i \) is used to produce two offspring, the usual one \( x_i = m + z_i \), and the mirrored offspring \( x_i = m - z_i \). Those two offspring are symmetric or mirrored to the parental point \( m \). The mirrored sampling method is described in Algorithm 1. Following the notation in [5], we denote the ES algorithm using mirrored sampling as \((\mu + \lambda_m)\)-ES.

Algorithm 1 MIRRORED-SAMPLING(\(m, \sigma, \mathbf{C}, \lambda\))

1. \( B, D \leftarrow \text{EIGEN-DECOMPOSITION}(\mathbf{C}) \)
2. \( \text{for } i = 1 \text{ to } \lambda/2 \text{ do} \)
3. \( z_i \leftarrow \mathcal{N}(0, \mathbf{I}) \)
4. \( x_{2i-1} \leftarrow m + \sigma B z_i \)
5. \( x_{2i} \leftarrow m - \sigma B z_i \)
6. \( \text{end for} \)

3. PROPOSED ALGORITHM

3.1 Mirrored Orthogonal Sampling
It is proposed here to first generate half of the mutations as random orthogonal samples. Specifically, each pair of mutation vectors are ensured to be orthogonal to each other while their directions are still random. The simplest way to realize orthogonal sampling is to orthonormalize a collection of i.i.d. Gaussian vectors. The Gram-Schmidt process [4] is chosen as the orthonormalization procedure. After the orthonormalization, the lengths of the samples are restored to the values before the process. Note that the maximal number of orthogonal samples is just the dimensionality. If \( \lambda/2 > N \), the remaining \( \lambda/2 - N \) samples are created as i.i.d. samples. Then we create the remaining half of the mutations by mirroring. The resulting sampling algorithm is called mirrored orthogonal sampling and denoted as \((\mu + \lambda_m)\)-ES here. The detailed procedure is described in Algorithm 2.

3.2 Recombination and Pairwise Selection
When working with the weighted recombination, mirroring related methods cause an undesired reduction of the variance of recombined mutation, leading to an undesired
uniformly distributed in the hyperbox \([s_1, \ldots, s_N] \). The (1 + 1)-ES uses 1 variant is illustrated in Figure 1a and 1b. 1a shows the optimum \(X^\ast\) and the advantage even holds for large dimensions.

3.3 Application to the CMA-ES Algorithm

We apply the mirrored orthogonal sampling technique to the well-known CMA-ES [6] (Covariance Matrix Adaptation Evolution Strategy). The damping factor \(d_\sigma\) to control the updating speed of step-size should be optimized for the new sampling method because it is originally developed for i.i.d. Gaussian mutations. Thus, after the optimization on the sphere function (details are not shown here), we modified it to \(d_\sigma = 1.49 - 0.6314 \cdot (\sqrt{\mu_{\text{eff}}} + 0.1572)/(N + 1.647) + 0.869 + c_\sigma\). See [6] for the original setting of \(d_\sigma\).

3.4 Empirical Convergence Rates

Given the starting point \(X^{(0)}\), the distance to the global optimum \(X^{(k)}\) of generation \(k\) and the global optimum \(X^\ast\), the convergence rate of evolution strategies can be measured as [3],

\[
\frac{1}{T_k} \ln \frac{\|X^{(k)} - X^\ast\|}{\|X^{(0)} - X^\ast\|}
\]

where \(T_k\) is the total number of function evaluations performed until generation \(k\). On the sphere function, the convergence of \((\mu/\mu_w, \lambda_0)-CMA-ES\) and other comparable ES variants are illustrated in Figure 1a and 1b. 1a shows the details for 10-D. The \((1+1)\)-ES uses 1/5 success step-size control while the \((1+1)\)-ES optimal uses optimal \(\sigma\) setting. “\(d_\sigma = \text{optimal}\)” denotes \((\mu/\mu_w, \lambda_0)\)-CMA-ES using optimal \(d_\sigma\) tuning on the sphere function. The mirrored orthogonal CMA-ES is significantly better than the mirrored version and the advantage even holds for large dimensions.

4. EXPERIMENTAL SETTING

The standard experimental procedure of BBOB is adopted. \((\mu/\mu_w, \lambda_0)\)-CMA-ES, \((\mu/\mu_w, \lambda_\ast)\)-CMA-ES and \((\mu/\mu_w, \lambda)\)-CMA-ES are benchmarked on BBOB. The initial global step-size \(\sigma\) is set to 1. The maximum number of function evaluations is set to \(10^5 \times N\). The initial solution vector is a uniformly distributed in the hyperbox \([-4, 4]^N\). The dimensions tested in the experiment are \(N \in \{2, 3, 5, 10, 20, 40\}\).

5. RESULTS

Small population. The results under the default population setting (logarithm of the dimensionality) are shown in Figure 1c and 1d. \((\mu/\mu_w, \lambda_0)\)-CMA-ES is shown in dashed curves while \((\mu/\mu_w, \lambda_\ast)\)-CMA-ES is shown in solid curves. When \(N = 5\), the performance leap of mirrored orthogonal sampling is significant. When \(N = 20\), there is still a small advantage.

Large population. When the population size is linearly related to the dimensionality, the comparison between the mirrored orthogonal sampling and its mirrored counterpart is illustrated in Figure 1e and 1f. The improvement brought by mirrored orthogonal sampling becomes more significant than the small population case in 5-D and further holds for 20-D.

6. CONCLUSION

The mirrored orthogonal sampling improves the mirrored sampling both theoretically and experimentally. On the sphere function, mirrored orthogonal sampling is much faster than the mirroring and just a little bit slower than the \((1+1)\)-ES with 1/5 rule. As for the BBOB comparisons, the results also reveal its advantages over the mirrored counterpart and the standard \((\mu/\mu_w, \lambda)\)-CMA-ES, although such advantages seem to decrease in higher dimensions. The better BBOB results for the large population suggest that the mirrored orthogonal sampling may be more suitable for increasing the exploration effects of a large population.

7. REFERENCES


Figure 1: (a): The comparison of empirical convergence in 10-D, measured by the distance to the optimum. (b): The comparison of empirical convergence rates against the dimensionality. (c) and (d): Empirical cumulative distributions (ECDF) of run lengths (the number of function evaluations divided by dimension) for \((\mu/\mu_w, \lambda_o)\)-CMA-ES (solid lines) and \((\mu/\mu_w, \lambda_m)\)-CMA-ES (dashed lines) needed to reach a target value \(f_{opt} + \Delta f\) with \(\Delta f = 10^k\), where \(k \in \{1, -1, -4, -8\}\) is given by the first value in the legend. The default population size setting is used here. The vertical black line indicates the maximum number of 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. The comparisons in two dimensions, \(N = 5\) (c) and \(N = 20\) (d) are shown. The ECDFs are estimated using small population setting. (e) and (f): The same comparisons as (c), (d) except the large population setting is used. (e): \(N = 5\) and (f): \(N = 20\).