Black-box Optimization Benchmarking of NIPOP-aCMA-ES and NBIPOP-aCMA-ES on the BBOB-2012 Noiseless Testbed

Ilya Loshchilov
TAO, INRIA Saclay
U. Paris Sud, F-91405 Orsay
Marc Schoenauer
TAO, INRIA Saclay
U. Paris Sud, F-91405 Orsay
Michèle Sebag
CNRS, LRI UMR 8623
U. Paris Sud, F-91405 Orsay
firstname.lastname@inria.fr

ABSTRACT

In this paper, we study the performance of NIPOP-aCMA-ES and NBIPOP-aCMA-ES, recently proposed alternative restart strategies for CMA-ES. Both algorithms were tested using restarts till a total number of function evaluations of $10^D$ was reached, where $D$ is the dimension of the function search space. We compared new strategies to CMA-ES with IPOP and BIPOP restart schemes, two algorithms with one of the best overall performance observed during the BBOB-2009 and BBOB-2010. We also present the first benchmarking of BIPOP-CMA-ES with the weighted active covariance matrix update (BIPOP-aCMA-ES).

The comparison shows that NIPOP-aCMA-ES usually outperforms IPOP-aCMA-ES and has similar performance with BIPOP-aCMA-ES, using only the regime of increasing the population size. The second strategy, NBIPOP-aCMA-ES, outperforms BIPOP-aCMA-ES in dimension 40 on weakly structured multi-modal functions thanks to the adaptive allocation of computation budgets between the regimes of restarts.

Categories and Subject Descriptors

G.1.6 [Numerical Analysis]: Optimization—global optimization, unconstrained optimization; F.2.1 [Analysis of Algorithms and Problem Complexity]: Numerical Algorithms and Problems

General Terms

Algorithms

Keywords

Benchmarking, black-box optimization, evolution strategy, CMA-ES, self-adaptation, restart strategies

1. INTRODUCTION

The CMA-ES algorithm is a stochastic optimizer, searching the continuous space $\mathbb{R}^D$ by sampling $\lambda$ candidate solutions from a multivariate normal distribution $[9, 8]$. It exploits the best $\mu$ solutions out of the $\lambda$ ones to adaptively estimate the local covariance matrix of the objective function, in order to increase the probability of successful samples in the next iteration. The information about the remaining (worst $\lambda - \mu$) solutions is used only implicitly during the selection process.

The CMA-ES has been extended to active $(\mu/\mu_1, \lambda)$-CMA-ES [11] and weighted active $(\mu/\mu_w, \lambda)$-CMA-ES (aCMA-ES [10]), where the information about worst $\lambda - \mu$ points can be also taken into account to reduce the variance of the mutation distribution in unpromising directions. However, aCMA-ES no longer guarantees the positive-definiteness of the covariance matrix, possibly resulting in algorithmic instability. The instability issues can however be numerically controlled during the search; as a matter of fact they are never observed on the BBOB benchmark suite.

Two versions of CMA-ES with restarts have been proposed to handle multi-modal functions: IPOP-CMA-ES [1] was ranked first on the continuous optimization benchmark at CEC 2005 [4, 3]; and BIPOP-CMA-ES [5] showed the best results together with IPOP-CMA-ES on the black-box optimization benchmark (BBOB) in 2009 and 2010.

The restart strategies of CMA-ES can be viewed as a noisy optimization problem of proper hyper-parameters of the CMA-ES in a 2D space (population size, initial step-size). In this paper we study the performance of two alternative restart strategies for CMA-ES, NIPOP-aCMA-ES and NBIPOP-aCMA-ES. The interested reader is referred to [12] for an in-depth presentation and discussion of these algorithms.

2. THE ALGORITHMS

2.1 The IPOP-aCMA-ES

A search for the global optima of multimodal function can be difficult if the number of local optima is high. For the specific case of the CMA-ES algorithm it has been observed that the probability and the overall number of function evaluations to reach the optima are very sensitive to the population size $[8]$. The default population size $\lambda_{\text{default}}$, tuned for uni-modal functions, is not always sufficiently large for multi-modal functions. This observation led to an idea to restart the CMA-ES, each time with larger population size [1] to perform a more global search. The restart $(\mu/\mu_w, \lambda)$-
CMA-ES with increasing population (IPOP-CMA-ES [1]) launches independent restarts and double the population size each time at least one of the stopping criterions is met. The IPOP-CMA-ES is an extension of the weighted active CMA-ES in IPOP restart scheme [10], which usually performs not worse than IPOP-CMA-ES on noiseless and noisy functions.

2.2 The BIPOP-aCMA-ES

In BIPOP-CMA-ES after the first single run with default population size, we restart the algorithm in one of two possible regimes and account the budget of function evaluations spent in the corresponding regime. Each time we restart the algorithm, we use the regime with smallest budget used so far.

Under the first regime we double the population size \( \lambda_{\text{large}} = 2^{\text{restart}} \lambda_{\text{default}} \) in each restart \( \text{restart} \) and use some fixed initial step-size \( \sigma_{\text{large}} = \sigma_{\text{default}}^0 \). This regime corresponds to the IPOP-CMA-ES.

Under the second regime we restart the CMA-ES with some small population size \( \lambda_{\text{small}} \) and step-size \( \sigma_{\text{small}}^0 \), where \( \lambda_{\text{small}} \) is set to

\[
\lambda_{\text{small}} = \left( \lambda_{\text{default}} \left( \frac{1}{2} \right)^{\lambda_{\text{default}}} \right)^{U(0,1)^2}.
\]

Here \( U(0,1) \) denote independently uniformly distributed numbers in \([0,1]\) and \( \lambda_{\text{small}} \in [\lambda_{\text{default}}, \lambda/2] \). The initial step-size is set to \( \sigma_{\text{small}}^0 = \sigma_{\text{default}}^0 \times 10^{-2U(0,1)} \).

In each restart, BIPOP-CMA-ES selects the restart regime with less function evaluations. Clearly, the second regime consumes less function evaluations than the doubling regime; it is therefore launched more often.

The BIPOP-CMA-ES, an extension of BIPOP-CMA-ES to the case of the weighted active covariance matrix update (weighted active (\(\mu/\mu_w, \lambda\))-CMA-ES in BIPOP restart scheme), will be for the first time benchmarked in this paper.

2.3 The NIPOP-aCMA-ES

In NIPOP-aCMA-ES in addition to increasing population size in each restart, we also decrease the initial step-size by some factor \( k_{\text{dec}} \). In this study we choose \( k_{\text{dec}} = 1.6 \) such that \( \sigma \) value after 9 restarts roughly corresponds to the minimum possible initial \( \sigma = 10^{-2}\sigma_{\text{default}} \) used for BIPOP-CMA-ES.

2.4 The NBIPOP-aCMA-ES

In NBIPOP-aCMA-ES as well as in BIPOP-aCMA-ES we have two restart regimes:
i). Double the population size and decrease the initial step-size by \( k_{\text{dec}} = 1.6 \) (NIPOP-aCMA-ES).
ii). Launch CMA-ES with default population size \( \lambda_{\text{default}} \) and \( \sigma^0 = \sigma_{\text{default}}^0 \times 10^{-2U(0,1)} \).

In contrast with BIPOP-CMA-ES, where both regimes have the same budget, the budget is adapted here according to the performance of the regime: the best solutions \( x_{A}^* \) and \( x_{B}^* \) found by regimes A and B are used as an estimator of their quality. We thus allocate \( k_{\text{budget}} = 2 \) times larger computation budget for regime A if it performs better than B (i.e., if \( x_{A}^* \) is better than \( x_{B}^* \)), and vice versa.

2.5 The Benchmarked Algorithms

For benchmarking we consider four CMA-ES algorithms in restart scenario: IPOP-aCMA-ES [10], BIPOP-aCMA-ES as an extension of BIPOP-CMA-ES [5], NIPOP-aCMA-ES and NBIPOP-aCMA-ES [12]. In \((\mu/\mu_w, \lambda)\)-CMA-ES part of these algorithms we use default parameters as given in [10] and [5].

The maximum budget of function evaluations is \(10^6D\) and the initial step-size \( \sigma_{\text{default}}^0 = 2.0 \).

3. RESULTS

Results from experiments according to [6] on the benchmark functions given in [2, 7] are presented in Figures 1, 2 and 3 and in Tables 1 and 2. The expected running time (ERT), used in the figures and table, depends on a given target function value, \( f_t = f_{\text{best}} + \Delta f \), and is computed over all relevant trials (on the first 15 instances) as the number of function evaluations executed during each trial while the best function value did not reach \( f_t \), summed over all trials and divided by the number of trials that actually reached \( f_t \) [6, 13]. Statistical significance is tested with the rank-sum test for a given target \( \Delta f_t \) (10\(^-8\)) as in Figure 1) using, for each trial, either the number of needed function evaluations to reach \( \Delta f_t \) (inverted and multiplied by \(-1\)), or, if the target was not reached, the best \( \Delta t \)-value achieved, measured only up to the smallest number of overall function evaluations for any unsuccessful trial under consideration.

All benchmarked here algorithms represent \((\mu/\mu_w, \lambda)\)-CMA-ES before the first restart occurs, therefore, the results are very similar for the uni-modal functions, where the optimum usually can be found without restarts. We show the results in 40-D instead of 20-D, because the difference between algorithms is more significant in higher dimensions.

NIPOP-aCMA-ES. On 6 out of 12 test functions (\(f_{15}, f_{16}, f_{17}, f_{18}, f_{23}, f_{24}\)) NIPOP-aCMA-ES obtains the best known results for BBOB-2009 and BBOB-2010 workshops. On \(f_{23}\) Katsuuras and \(f_{24}\) Lunacek bi-Rastrigin, NIPOP-aCMA-ES has a speedup of a factor from 2 to 3, as expected. It performs unexpectedly well on \(f_{16}\) Weierstrass functions, 7 times faster than IPOP-CMA-ES and almost 3 times faster than BIPOP-aCMA-ES. Overall, according to Fig. 3, NIPOP-aCMA-ES performs as well as BIPOP-aCMA-ES, while restricted to only one regime of increasing population size.

NBIPOP-aCMA-ES. Thanks to the first regime of increasing population size, NBIPOP-aCMA-ES inherits some results of NIPOP-aCMA-ES. However, on functions where the population size does not play any important role, it performs significantly better than BIPOP-CMA-ES. This is the case for \(f_{21}\) Gallagher 101 peaks and \(f_{22}\) Gallagher 21 peaks functions, where NBIPOP-aCMA-ES has a speedup of a factor of 6. It seems that the adaptive choice between two regimes works efficiently on all functions except on \(f_{16}\) Weierstrass, where NBIPOP-aCMA-ES incorrectly prefers small populations. This leads to a loss of a factor of 4 in comparison to NIPOP-aCMA-ES, while a factor of 1.5 is expected in the case of correct adaptation.

An interesting result is a comparatively good performance of NBIPOP-aCMA-ES on 5-dimensional \(f_5\) Skew Rastrigin Bueche multi-modal function, where NBIPOP-aCMA-ES is the only algorithm among 4 tested here, which is able to find the global optimum in 9 out of 15 runs. According to Fig. 3, NBIPOP-aCMA-ES performs better than BIPOP-aCMA-ES on weakly structured multi-modal functions, showing
overall best results for BBOB-2009 and BBOB-2010 workshops in dimensions 20 and 40.

4. CONCLUSION

In this paper, we have compared the recently proposed restart strategies for aCMA-ES, NIPOP-aCMA-ES and NBIPPOP-aCMA-ES with the IPOP-aCMA-ES and BIPOP-aCMA-ES. The main message of the paper is that the decreasing of initial step-size makes IPOP restart scenario more robust and sometimes even comparable to BIPOP scenario on noiseless functions. We also suppose that the adaptation of the computation budgets of different restart regimes is a promising idea for black-box optimization and should be further investigated.

5. ACKNOWLEDGMENTS

This work was partially funded by FUI of System@tic Paris-Region ICT cluster through contract DGT 117 407 Complex Systems Design Lab (CSDL).

6. REFERENCES

Figure 1: Expected running time (ERT in number of $f$-evaluations) divided by dimension for target function value $10^{-8}$ as $\log_{10}$ values versus dimension. Different symbols correspond to different algorithms given in the legend of $f_1$ and $f_{24}$. Light symbols give the maximum number of function evaluations from the longest trial divided by dimension. Horizontal lines give linear scaling, slanted dotted lines give quadratic scaling. Black stars indicate statistically better result compared to all other algorithms with $p < 0.01$ and Bonferroni correction number of dimensions (six). Legend: $\circ$: BIPOP-aCMA, $\triangledown$: IPOP-aCMA, $\star$: NBIPOP-aCMA, $\Box$: NIPOP-aCMA.
Figure 2: Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension (FEvals/D) for 50 targets in $10^{−8...2}$ for all functions and subgroups in 5-D. The “best 2009” line corresponds to the best ERT observed during BBOB 2009 for each single target.
Figure 3: Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension (FEvals/D) for 50 targets in $10^{1-8.2}$ for all functions and subgroups in 40-D. The “best 2009” line corresponds to the best ERT observed during BBOB 2009 for each single target.
Table 1: Expected running time (ERT in number of function evaluations) divided by the respective best ERT

<table>
<thead>
<tr>
<th></th>
<th>NBIPOP-</th>
<th>BIPOP-a</th>
<th>IPOP-aC</th>
</tr>
</thead>
<tbody>
<tr>
<td>#succ</td>
<td>5.7(3)</td>
<td>6.9(3)</td>
<td>6.9(3)</td>
</tr>
<tr>
<td>f11</td>
<td>2.1</td>
<td>5.6</td>
<td>3.6</td>
</tr>
<tr>
<td>f10</td>
<td>5.9(3)</td>
<td>6.1(3)</td>
<td>1.5</td>
</tr>
<tr>
<td>f9</td>
<td>7.3</td>
<td>5.9</td>
<td>4.2</td>
</tr>
<tr>
<td>f8</td>
<td>5.9</td>
<td>3.6</td>
<td>0.76</td>
</tr>
<tr>
<td>f7</td>
<td>5.6</td>
<td>1.5</td>
<td>0.53</td>
</tr>
<tr>
<td>f6</td>
<td>5.9</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>f5</td>
<td>5.9</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>f4</td>
<td>5.9</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>f3</td>
<td>5.9</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>f2</td>
<td>5.9</td>
<td>1.2</td>
<td>0.96</td>
</tr>
</tbody>
</table>

The expected running time is calculated using the formula $\Delta f_t = \frac{f_{opt}}{f_{opt}^{\text{best}}}$, where $f_{opt}$ is the optimal function value and $f_{opt}^{\text{best}}$ is the best function value found. The uncertainties are given in parentheses. The results are obtained using the epsilon-BBO algorithm with different population sizes and dimensions.
| Table 2: Expected running time (ERT in number of function evaluations) divided by the respective best ERT measured during BBOB-2009 (given in the respective first row) for different $\Delta_f$ values in dimension 40. The central 80% range divided by two is given in braces. The median number of conducted function evaluations is additionally given in italics, if $ERT = \infty$. #suc is the number of trials that reached the final target $f_{\text{opt}} + 10^{-8}$. Best results are printed in bold. |