Achieving optimization invariance w.r.t. monotonous transformations of the objective function and orthogonal transformations of the representation

Ilya Loshchilov
TAO, INRIA Saclay
LRI, Univ. Paris-Sud, Orsay, France
ilya.loshchilov@gmail.com

Marc Schoenauer
TAO, INRIA Saclay
LRI, Univ. Paris-Sud, Orsay, France
marc.schoenauer@inria.fr

Michèle Sebag
CNRS, LRI UMR 8623
LRI, Univ. Paris-Sud, Orsay, France
michele.sebag@inria.fr

1 Introduction

Estimation of distribution- and population-based optimization algorithms [1, 2] have been thoroughly investigated for their ability to handle ill-posed optimization problems. Maintaining a distribution on the search space or a population of candidate solutions indeed enforces the search robustness with respect to the (moderate) noise and multi-modality of the objective function, in contrast to classical optimization methods such as quasi-Newton methods.

The main limitation of the above stochastic optimization methods is due to their sample complexity, i.e. the large number of objective evaluations they require; this limitation hinders their application to computationally expensive problems such as optimal design in numerical engineering, where one evaluation might require solving structural problems using Finite Element Methods. For this reason, stochastic optimization has often-wise been coupled with learning surrogate models, that is, local approximations of the objective function, which enable to replace a significant fraction of the true objective evaluations, with the inexpensive evaluation of the surrogate function. Among the key issues of surrogate-based stochastic optimization are the exploration vs exploitation trade-off, namely the control of the surrogate learning module (update schedule, learning hyper-parameters).

The present paper focuses on coupling self-adaptive surrogate learning with the covariance matrix adaptation evolution strategy algorithm (CMA-ES) [2], a state-of-the art algorithm for continuous black-box optimization. The contribution is twofold. Firstly, we show that a tight coupling of CMA-ES and Ranking SVM [3], referred to as ACM-$*$ achieves optimization invariance with respect to both orthogonal transformations of the search space, and monotonous transformation of the objective function, while adaptively adjusting the update schedule and the learning hyper-parameters.

Secondly, ACM-$*$ is assessed comparatively to the quasi-Newton BFGS method [4], most specifically, its high precision arithmetic implementation. It is shown that ACM-$*$ matches the high-precision BFGS performances when considering an appropriately scaled (quadratic) objective function, and outperforms the high-precision BFGS by a multiplicative factor ranging in 2..3 otherwise, on the black-box optimization benchmarking (BBOB) noiseless test-bed [5].

The paper briefly describes CMA-ES for the sake of self-containedness before giving an overview of ACM-$*$ and discussing its invariance properties. The following sections are devoted to the experimental validation of the approach comparatively to high-precision BFGS and the paper concludes with some perspectives for further research.

2 Covariance matrix adaptation evolution strategy (CMA-ES)

Let f denote the objective function to be minimized, defined on $\mathbb{R}^n$

$$f : \mathbb{R}^n \mapsto \mathbb{R}$$

The so-called (μ_w, λ)-CMA-ES [2] maintains a Gaussian distribution on $\mathbb{R}^n$, which is iteratively used to generate λ candidate solutions, and updated based on the best (in the sense of f) μ solutions among the λ ones. Formally, every candidate solution $\boldsymbol{x}_{t+1}$ at time $t+1$ is drawn from the current Gaussian distribution:

$$\boldsymbol{x}_{t+1} \sim \mathcal{N}\left(\boldsymbol{m}_t, {\sigma_t}^2 \boldsymbol{C}_t\right) = \boldsymbol{m}_t + \sigma_t \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_t), \tag{1}$$

where $\boldsymbol{C}_t \in R^{n \times n}$ is a covariance matrix, σ_t is a perturbation step-size, and the center $\boldsymbol{m}_t$ of the distribution is the current best estimate of the optimum. The new distribution center $\boldsymbol{m}_{t+1}$ is computed as the weighted sum of the best (w.r.t. f) μ solutions as follows[1]:

$$\boldsymbol{m}_{t+1} = \sum_{i=1}^{\mu} w_i \boldsymbol{x}_{t+1}^{(i:\lambda)}, \tag{2}$$

where $\boldsymbol{x}_{t+1}^{(i:\lambda)}$ denotes the i-th best solution out of the λ ones generated at time step $t+1$, and $\sum_{i=1}^{\mu} w_i = 1$. The covariance matrix $\boldsymbol{C}_{t+1}$ is adjusted using both the local information about the search direction, given by $\frac{1}{\sigma_t}(\boldsymbol{x}_{t+1}^{(i:\lambda)} - \boldsymbol{m}_t)$, and the global information stored in the so-called evolution path $\boldsymbol{p}_{t+1}$ of the distribution center $\boldsymbol{m}$. For positive learning rates c_1 and c_μ ($c_1 + c_\mu \leq 1$) the update of the covariance matrix reads as follows:

$$\boldsymbol{C}_{t+1} = (1 - c_1 - c_\mu)\boldsymbol{C}_t + c_1 \underbrace{\boldsymbol{p}_{t+1} \cdot \boldsymbol{p}_{t+1}{}^{tr}}_{\text{rank-one update}} + c_\mu \underbrace{\sum_{i=1}^{\mu} \frac{w_i}{{\sigma_t}^2} (\boldsymbol{x}_{t+1}^{(i:\lambda)} - \boldsymbol{m}_t) \cdot (\boldsymbol{x}_{t+1}^{(i:\lambda)} - \boldsymbol{m}_t)^{tr}}_{\text{rank-}\mu\text{ update}} \tag{3}$$

The step-size σ_{t+1} is likewise updated to best align the distribution of the actual evolution path of σ, and an evolution path under random selection.

Interestingly, the recently proposed Information-Geometric Optimization (IGO [6]) framework shows that the pure rank-μ update of CMA-ES is a special case of IGO when considering the family of all Gaussian distributions P_θ in R^n and performing a natural gradient ascent of θ.

CMA-ES performances and robustness are explained from its invariances:
$*$ with respect to monotonous transformations of the objective function f. By construction, it does not make any difference whether one considers the minimization of f, or that of $g \circ f$, for any strictly increasing function $g : R \mapsto R$. In particular, CMA-ES yields the same performances regarding the optimization of f and f^3 (contrarily to BGFS, see below).
$*$ with respect to angle preserving transformations of the search space (rotation, reflection, translation)[2], due to the adaptation of the covariance matrix C.

3 ACM-$*$

The proposed ACM-$*$ algorithm is meant to extend CMA-ES to expensive optimization through surrogate model learning. The focus is on preserving CMA-ES invariance properties, and enforcing the self-adaptation of the surrogate learning module.

Surrogate model learning

The CMA-ES invariance w.r.t. monotonous transformations of the objective function is preserved in ACM-$*$ through a rank-based surrogate learning approach. Specifically, a ranking support vector machine (Ranking SVM [3]) is used to learn the surrogate model $\hat{f}$ from the available evidence $\mathcal{E} = \{(x_i, f(x_i)), i = 1, \ldots \ell\}$. Let us assume with no loss of generality that the training points in $\mathcal{E}$ are ordered by increasing value of $f(x_i)$. For the sake of computational efficiency, a linear number of ranking constraints ($x_i \prec x_{i+1},\ i = 1 \ldots \ell - 1$) is used. By construction, the surrogate model thus learned is invariant under the composition of f with any strictly increasing function g.

The invariance of the surrogate model w.r.t. orthogonal transformations of the search space is enforced by using a Radial Basis Function (RBF) kernel, which precisely involves the inverse of the covariance matrix adapted by CMA-ES. Formally, the rank-based surrogate model is learned using the kernel K_C defined as: $K_C(\boldsymbol{x}_i,\boldsymbol{x}_j) = \mathrm{e}^{-\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{C}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)}{2\sigma^2}}$, which corresponds to rescaling the training (and test, see below) data using the transformation $\boldsymbol{x} \rightarrow \boldsymbol{C}^{-1/2}(\boldsymbol{x}-\boldsymbol{m})$. Through using this kernel, the surrogate learning module benefits from the CMA-ES efforts in identifying the local curvature of the optimization landscape.

Surrogate model exploitation

A key issue in surrogate-based optimization concerns the updating schedule of the surrogate model, specifically how many optimization iterations (a.k.a surrogate life-length $\hat{n}$) should a surrogate model be used before being rebuilt. The proposed approach first optimizes f by CMA-ES for one iteration and uses the candidate solutions to build $\hat{f}$. It thereafter proceeds by iterating the following steps: i) optimize $\hat{f}$ by CMA-ES for $\hat{n}$ iterations; ii) optimize f by CMA-ES for one iteration and use the candidate solutions to assess the accuracy of $\hat{f}$; iii) depending on the fraction of incorrectly ordered pairs, adjust $\hat{n}$; iv) use the candidate solutions for relearning $\hat{f}$. In this way, $\hat{n}$ is set to 0 when $\hat{f}$ gives random prediction, thus falling back to the original CMA-ES. This procedure also allows ACM-$*$ to be efficiently parallelized on λ CPU that is often not the case for surrogate-assisted algorithms which evaluate only one point per iteration (e.g., some trust-region based algorithms).

Surrogate model hyper-parameters

The quality of the surrogate model $\hat{f}$ is sensitive to the learning hyper-parameters, e.g., the number and distribution of training points controls whether the model is global or local. The selection of the hyper-parameters is handled through launching an internal CMA-ES for one iteration, aimed at minimizing the surrogate model error and searching the hyper-parameter space. Along this line, ACM-$*$ achieves the lifelong learning/optimization of i) the learning hyper-parameters; ii) the surrogate model and its life-length. The user is only required to provide the range of variation of the learning hyper-parameters.

4 Goal of experiments: comparative assessment with high-precision BFGS

The experimental validation of the proposed scheme aims at assessing ACM-$*$ comparatively to the well-known quasi-Newton BFGS algorithm [4] on ill-conditioned optimization problems. The literature indeed shows that BFGS suffers from numerical problems in such cases (see [7] for an extensive discussion). Specifically, on the 24 noiseless benchmark problems of the BBOB suite, BFGS is shown to reach a first performance level (10^{-2}) very quickly comparatively to other algorithms, and then plateaus [8]. Our conjecture is that this shortcoming is due to an insufficient precision in estimating the gradient of f (and thus the Hessian of f). This conjecture is tested by considering a high-precision arithmetic version of BFGS, referred to as pBFGS (implemented in C++, integrated in the high-precision arithmetic package ARPREC [9])[3].

5 Experiments

The self-adaptive surrogate learning ACM-$*$ is used on the top of the bi-population with restart mode of CMA-ES, using its active update variant [10], considering the (non-surrogate-based) CMA-ES, the baseline implementation of BFGS, and the 32-decimal digit precision pBFGS as baselines[4], both with restart mode.

Fig. 1-Left reports the median results (out of 15 runs) on the 20-dimensional Rosenbrock function $f_{Rosenbrock}(x) = \sum_{i=1}^{n-1}\left(100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2\right)$ (uniform initialization in $[-5, 5]^{20}$ for all algorithms). ACM-$*$ outperforms the CMA-ES mode by a factor of about 2.5-3.0. As the Rosenbrock function is not ill-conditioned, pBFGS performs similarly to BFGS (not shown here) and ACM-$*$. Rescaling the objective function (considering f^2 or f^4) however adversely affects pBFGS, slowing down the convergence by a factor of about 3 for f^4. In the meanwhile, ACM-

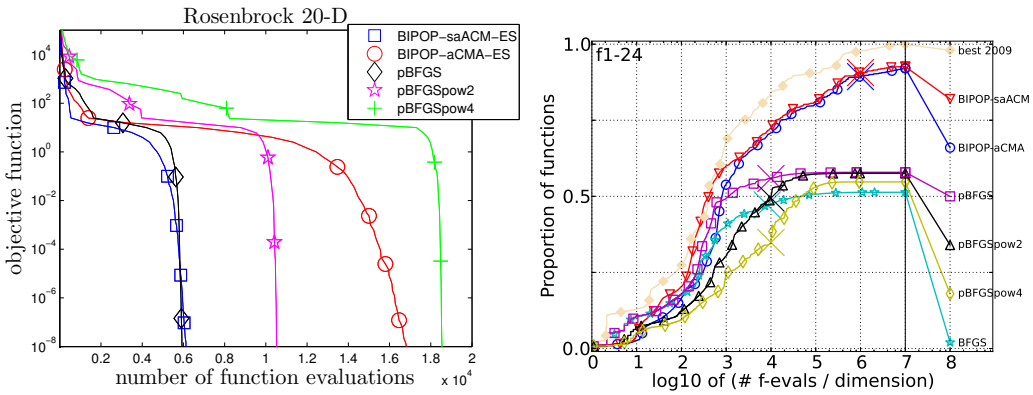

Figure 1: **Left**: Median objective value (out of 15 independent runs) reached for a number of objective evaluations for ACM-$*$, BFGS and pBFGS on 20-dimensional Rosenbrock function $f_{Rosenbrock}$ and its different scaled variants: $f = f^2_{Rosenbrock}$ (legend pBFGSpow2) and $f = f^4_{Rosenbrock}$ (legend pBFGSpow4) .
Right: Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension for 50 targets in $10^{[-8..2]}$ for all 24 noiseless functions in 20-D. The "best 2009" line indicates the BBOB 2009 "portfolio oracle", the aggregation of the best algorithm result for each function.

$*$ and CMA-ES-based approaches show their invariance w.r.t. monotonous transformations of the objective function.

BBOB framework provides 24 noiseless [5] and 30 noisy [11] benchmark problems with different properties: separable, non-separable, unimodal, multi-modal, ill-conditioned, deceptive, functions with and without weak global structure. Figure 1-Right shows the benchmarking results on 24 noiseless BBOB problems with 15 instances/runs for each problem and 50 uniformly generated target f values per problem (24$\times$50=1200 target function values for y-axis). The results show that ACM-$*$ outperforms its CMA-ES baseline and all other benchmarked algorithms in most cases. The high-precision version of BFGS, pBFGS, significantly improves on BFGS thanks to its robust performance on ill-conditioned problems. Monotonous transformations of the objective function (f^2 (pBFGSpow2) and f^4 (pBFGSpow4)) however significantly affect the pBFGS results.

6 Discussion and future work

This paper emphasizes the merits of invariance w.r.t. monotonous transformation of the objective function, regarding surrogate-assisted optimization. A second result concerns the advantages of high-precision BFGS for solving ill-conditioned optimization problems. High-precision computations however require the source code to be rewritten not only on the part of the optimization algorithm, but often also on the part of the objective function (as was the case in our study). The latter is impossible in standard black-box scenarios; often-wise it is intractable even when the objective source code is available. A third result is that high-precision arithmetic does not prevent BFGS results from degrading when the scaling of the objective function differs from the "desirable" quadratic BFGS scaling.

Further work will investigate a tighter coupling of CMA-ES and Ranking SVM (typically, relating the surrogate model life-length and the perturbation step size). Alternative comparison-based surrogate models will also be considered, such as Gaussian Processes for ordinal regression [12]. Finally, as shown by [13], quasi-Newton methods can be interpreted as approximations of Bayesian linear regression under varying prior assumptions; a prospective research direction is to replace the linear regression by ordinal regression-based Ranking SVM or Gaussian Processes in order to derive a version of BFGS invariant w.r.t. monotonous transformations of the objective function f.

References

[1] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals OR*, 134(1):19–67, 2005.

[2] N. Hansen, S.D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.

[3] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.

[4] D. F. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[5] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions. Research Report RR-6829, INRIA, 2009.

[6] L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *ArXiv e-prints*, June 2011.

[7] M.J.D. Powell. Updating conjugate directions by the BFGS formula. *Mathematical Programming*, 38(1):29–46, 1987.

[8] R. Ros. Benchmarking the BFGS algorithm on the BBOB-2009 function testbed. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, pages 2409–2414. ACM, 2009.

[9] D.H. Bailey, H. Yozo, X.S. Li, and B. Thompson. ARPREC: An arbitrary precision computation package. Technical report, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2002.

[10] N. Hansen and R. Ros. Benchmarking a weighted negative covariance matrix update on the BBOB-2010 noiseless testbed. In *GECCO '10: Proceedings of the 12th annual conference comp on Genetic and evolutionary computation*, pages 1673–1680, New York, NY, USA, 2010. ACM.

[11] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-Parameter Black-Box Optimization Benchmarking 2009: Noisy Functions Definitions. Research Report RR-6869, INRIA, 2009.

[12] W. Chu and Z. Ghahramani. Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144. ACM, 2005.

[13] P. Hennig and M. Kiefel. Quasi-newton methods: A new direction. *arXiv preprint arXiv:1206.4602*, 2012.

Notes

[1] An active CMA-ES variant also uses the worst candidate solutions to update $\boldsymbol{m}_{t+1}$.

[2] If the initial candidate solutions are transformed accordingly.

[3] The source code is available at `https://sites.google.com/site/highprecisionbfgs/`.

[4] For gradient approximations by finite differences $\epsilon = 10^{-20}$ is used in pBFGS instead of $\epsilon = 10^{-8}$ in BFGS.