CMA-ES: A Function Value Free Second Order Optimization Method

Nikolaus Hansen

PGMO - COPI'14 manuscript No.
(will be inserted by the editor)

CMA-ES: A Function Value Free Second Order Optimization Method

Nikolaus Hansen

Received: date / Accepted: date

Abstract We give a bird's-eye view introduction to the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and emphasize relevant design aspects of the algorithm, namely its invariance properties. While CMA-ES is gradient and function value free, we show that using the gradient in CMA-ES is possible and can reduce the number of iterations on unimodal, smooth functions.

Keywords Optimization · Invariance · Evolution Strategies · CMA

Mathematics Subject Classification (2000) MSC 49M37 · MSC 65K05 · MSC 90C15

1 Introduction

We consider the problem to minimize an objective function

$$f : \mathbb{R}^n \to \mathbb{R} \ . \tag{1}$$

We do not assume to have any particular knowledge on the structure of f, thereby considering f as a black box. The *cost* of search is given to be the number of calls to the "black-box" f. In this context, we define a parameter vector $\boldsymbol{\theta}$ and a generic optimization procedure taking three steps

1. propose one or several new candidate solutions, depending on $\boldsymbol{\theta}$
2. evaluate the candidate solution(s) on f
3. update $\boldsymbol{\theta}$

This framework covers essentially any optimization procedure. We focus here on algorithms that sample from a probability distribution, and more specifically, sample in each iteration the same number of i.i.d. candidate solutions. In the continuous search space, a (multi-variate) normal distribution suggests itself as sample distribution, because it has maximum entropy (given that variances exist) and it is in a natural way detached from the given coordinate system. If the covariance matrix is a multiple of the identity, the distribution is isotropic.

Nikolaus Hansen
Inria
Research centre Saclay – Île-de-France
E-mail: nikolaus.hansen AT inria.fr

2 A Second Order Method

We consider a moderate search space dimensionality n, that is, $n \not\ll 10$ and $n \not\gg 100$. In this case, utilizing second-order information seems indispensable to achieve a competitive method. With a multivariate normal distribution, this can be naturally achieved by using a full covariance matrix to parametrize the distribution. We have $\boldsymbol{\theta}_t = (\mathbf{m}_t, \sigma_t, \mathbf{C}_t) \in \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^{n \times n}$, and at iteration t, for $i = 1, \ldots, \lambda$, new candidate solutions obey

$$\mathbf{x}_i = \mathcal{N}(\mathbf{m}_t, \sigma_t^2 \mathbf{C}_t) = \mathbf{m}_t + \sigma_t \, \mathcal{N}(\mathbf{0}, \mathbf{C}_t) \ , \tag{2}$$

where $\mathcal{N}(\mathbf{m}, \mathbf{C})$ denotes a normal distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$. Choosing $\mathbf{C}_t$ mimics in effect the linear coordinate system transformation $\mathbf{C}_t^{-1/2}$, because (2) is equivalent to

$$\mathbf{C}_t^{-1/2} \mathbf{x}_i = \mathbf{C}_t^{-1/2} \mathbf{m} + \sigma_t \, \mathcal{N}(\mathbf{0}, \mathbf{I}) \ . \tag{3}$$

On convex-quadratic functions, $\mathbf{C}$ resembles in the ideal case the inverse Hessian matrix.

3 Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

The CMA-ES method prescribes the updates of $\mathbf{m}_t, \sigma_t, \mathbf{C}_t$ and of some additional hidden variables (evolution paths). The updates of $\mathbf{m}$ and $\mathbf{C}$ can be (to a large extend) derived from information geometry, as they follow the natural gradient in $\boldsymbol{\theta}$-space [7]. The updates can be motivated equally well from the maximum likelihood principle [2]. In contrast, the overall step-size σ_t is updated with the aim to achieve $\mathbf{C}^{-1}$-conjugate (orthogonal) movements of the mean $\mathbf{m}$. The update procedure is detailed in [1].

The main **governing design principle** of the CMA-ES algorithm is invariance (see also the contribution of A. Auger in the present volume). Namely, we find for CMA-ES the same invariance properties as for the Nelder-Mead Simplex Downhill method [6]: invariance to affine transformations of the search space (including translations and rotations) and invariance to order-preserving transformations of the f-value. The former is a natural consequence of a "full" second order method. The latter sets the two methods apart from other derivative-free or gradient-based methods to be in essence *function value free*.

Invariances mean generalization of behavior on single functions to the entire set of functions belonging to an respective equivalence class [2], thereby making previous observations meaningful for prediction.

Most **internal parameters** of CMA-ES are not left to the users choice, because reasonable settings do not depend on the given problem f (in contrast to optimal settings). However, the user must choose a suitable representation (scaling or transformation of parameters used in f) and corresponding values for $\mathbf{m}_0$ and σ_0. The population size λ is an optional parameter to modify, as well as termination settings.

4 Current Developments

We describe two recent developments in the CMA-ES method. First, in implementing the idea of so-called active CMA [5], i.e. using also *negative* weights for the update of the covariance matrix, we choose slowly decreasing weights, proportional to

$$-\log k + \log \frac{\lambda + 1}{2} \tag{4}$$

for $k = \lceil (\lambda+1)/2 \rceil, \ldots, \lambda$. Positive weights are by default set with the same equation for $k = 1, \ldots, \lfloor \lambda/2 \rfloor$. For $\lambda = 15$, the effective parent number becomes 4.5 for positive and 5.9 for negative update. Further-

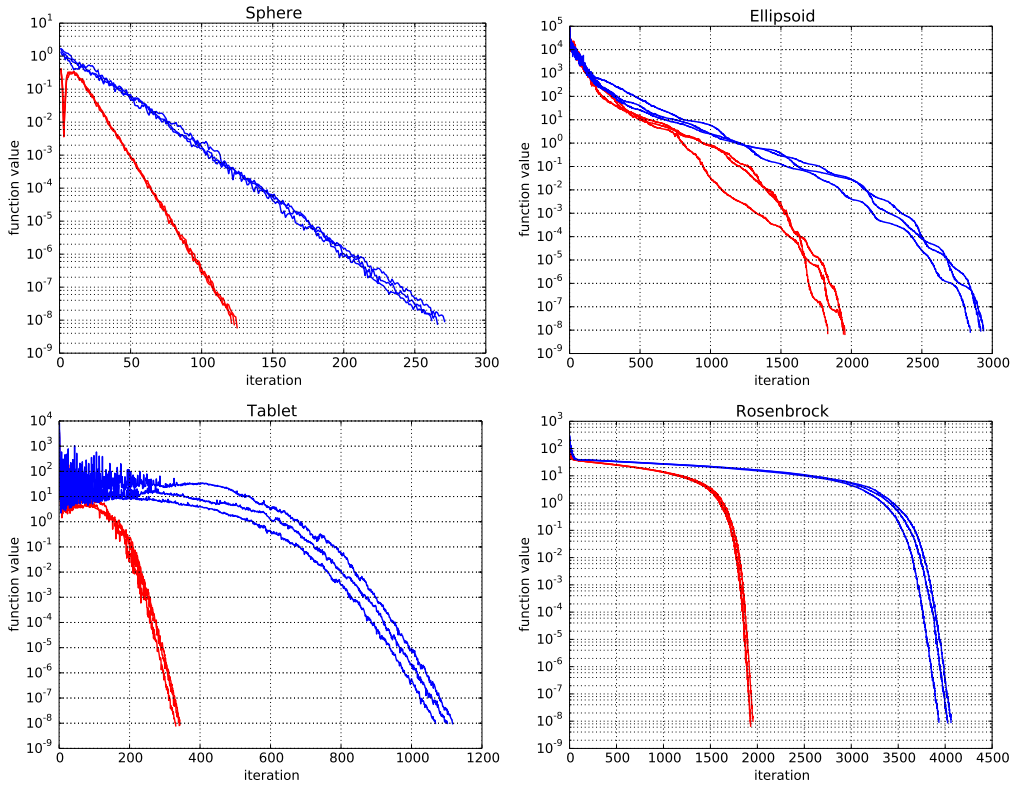

Fig. 1 Three runs in dimension $n = 40$ with default setting (blue) and with injected gradient in replacement of one sample (red). Initial values are $m_0 = -0.2$ in each component and $\sigma_0 = 0.1$, and $\lambda = 15$. The gradient reduces the number of iterations by a factor of between 1.5 and 3. The spike at the early stage on the Sphere function shows a possibly much faster convergence rate, then prevented by the slow decrement of the step-size.

more, the steps used with negative weights are normalized to constant length [4] and positive definiteness is guarantied by modulating the learning rate accordingly.

Second, we use the gradient of the function to generate one of the samples along $\mathbf{C}\,\nabla f(\mathbf{m}_t)$ [4]. Figure 1 shows results on four different functions [3]. Using the gradient reduces the number of iterations by a factor between 1.5 and 3. The effect of using the gradient is more pronounced with active CMA, and on the Tablet function, where active CMA is most relevant. Using the gradient also reduces the effect of λ on the number of iterations (not shown).

References

1. N. Hansen. The CMA evolution strategy: a comparing review. In J.A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.
2. N. Hansen and A. Auger. Principled design of continuous stochastic search: From theory to practice. In Yossi Borenstein and Alberto Moraglio, editors, *Theory and Principled Methods for the Design of Metaheuristics*, Natural Computing Series, pages 145–180. Springer Berlin Heidelberg, 2013.
3. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
4. Nikolaus Hansen. Injecting external solutions into CMA-ES. *ArXiv e-prints*, arXiv:1110.4181, 2011.
5. G. A. Jastrebski and D. V. Arnold. Improving evolution strategies through active covariance matrix adaptation. In *IEEE Congress on Evolutionary Computation (CEC 2006), proceedings*, pages 2814–2821. IEEE Press, 2006.
6. John Ashworth Nelder and R Mead. A simplex method for function minimization. *The Computer Journal*, pages 308–313, 1965.
7. Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles (2011v1; 2013v2). *ArXiv e-prints*, arXiv:1106.3708v2, 2013.