Bayesian CMA-ES

A new approach

Benhamou Eric
LAMSADE
Paris, France
eric.benhamou@dauphine.eu

Saltiel David
ULCO-LISIC
Calais, France
david.saltiel@univ-littoral.fr

Verel Sebastien
ULCO-LISIC
Calais, France
verel@univ-littoral.fr

ABSTRACT

This paper introduces a novel theoretically sound approach for the celebrated CMA-ES algorithm. Assuming the parameters of the multi variate normal distribution for the minimum follow a conjugate prior distribution, we can derive the optimal update at each iteration step thanks to Bayesian statistics. Update formulae are very similar to the vanilla (μ/λ) CMA-ES. We also use variance contraction and dilatation to accommodate for local and global search. As a result, this Bayesian framework provides a justification for the update of the CMA-ES algorithm and gives a new version of CMA-ES assuming normal-Inverse Wishart prior that has similar convergence speed and efficiency as the vanilla (μ/λ) CMA-ES on test functions ranging from cone, Rastrigin to Schwefel 1 and 2 functions.

CCS CONCEPTS

• **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

CMA-ES, Bayesian, conjugate prior, normal-inverse-Wishart

ACM Reference Format:
Benhamou Eric, Saltiel David, and Verel Sebastien. 2020. Bayesian CMA-ES: A new approach. In *Proceedings of the Genetic and Evolutionary Computation Conference 2020 (GECCO '20).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The covariance matrix adaptation evolution strategy (CMA-ES) [5] and all its variants are arguably one of the most powerful real-valued derivative-free optimization algorithms, finding many applications in machine learning. It is a state-of-the-art optimizer for continuous black-box functions as shown by the various benchmarks of the COmparing Continuous Optimisers (http://coco.gforge.inria.fr) INRIA platform for ill-posed functions. It has led to a large number of papers and articles and we refer the interested reader to [1, 4–6] and the recent publications in GECCO 2019 [3, 7, 8].

Briefly, the $(\mu \;/\; \lambda)$ CMA-ES is an iterative black box optimization algorithm, that, at each of its iterations, samples λ candidate solutions according to a multivariate normal distribution, evaluates these solutions, retains μ candidates and adjusts the distribution parameter for the next iteration. Each iteration makes an initial guess or *prior* for the distribution parameters (mean and variance of the multi-variate normal), evaluates the fit function and updates parameters thanks to the revised guess or *posterior*. Recently, using Bayesian statistics, [2] showed that CMA-ES could be interpreted as an iterative prior-posterior update. The complexity of the approach has been to decipher the updates. First of all, in a regular Bayesian setting, all sample points should be taken. This is not the case in the (μ/λ) CMA-ES, as only μ out of λ generated paths are kept. Second and more importantly, the covariance matrix update does not read easily as a Bayesian prior posterior update as the update is done according to two paths: the isotropic and anisotropic evolution ones. The objective of this short work is to compare CMA ES and Bayesian CMA-ES. We use the normal inverse Wishart conjugate prior to reduce computation of the prior thanks to a mean field approach. We experiment that the Bayesian adapted CMA-ES performs well on convex and non convex functions and is comparable to the classical CMA-ES.

2 BAYESIAN UPDATE

A conjugate prior for the multi variate normal is normal-inverse-Wishart distribution, parametrized by $\mu_0, \kappa_0, \nu_0, \psi$ whose formula is $f(\boldsymbol{\mu}, \Sigma|\mu_0, \kappa_0, \nu_0, \psi) = \mathcal{N}\left(\boldsymbol{\mu}\middle|\mu_0, \frac{1}{\kappa_0}\Sigma\right)\mathcal{W}^{-1}(\Sigma|\psi, \nu_0)$. where $\mathcal{W}^{-1}$ is the inverse Wishart and $\mathcal{N}$ a multi variate normal. Because the update is the optimal one given the prior information, the Bayesian updates should work well in initial iterations. Numerical experiments confirm this intuition.

Conjugate prior means that the posterior distribution is also a normal-inverse-Wishart with updated parameters $\text{NIW}(\mu_0^\star, \kappa_0^\star, v_0^\star, \psi^\star)$ given by:

$$\begin{aligned}
\mu_0^\star &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0 + n}, \qquad \kappa_0^\star = \kappa_0 + n, \qquad v_0^\star = v_0 + n \\
\psi^\star &= \psi + \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^T + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{x} - \mu_0)(\overline{x} - \mu_0)^T
\end{aligned} \tag{1}$$

with $\overline{x}$ the sample mean. This is the key property to update mean and covariance in Bayesian CMA-ES. If we have a sampling density that is a d dimensional multivariate normal distribution $\sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ with unknown mean μ and covariance Σ distributed according to a Normal-Inverse-Wishart $(\mu, \Sigma) \sim \text{NIW}(\mu_0, \kappa_0, v_0, \psi)$ and if we observe $\mathcal{X} = (x_1, .., x_n)$ samples, then the posterior is a Normal-Inverse-Wishart $\text{NIW}(\mu_0^\star, \kappa_0^\star, v_0^\star, \psi^\star)$. To avoid simulating the prior and the posterior, we take the mean values of the normal Wishart distribution that is closed form (mean field approximation). We hence avoid simulation in simulations. Following [2], we use

GECCO '20, July 8–12, 2020, Cancun, Mexico
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

a contraction dilatation mechanism that is similar to exploration-exploitation trade-off in reinforcement learning to mimic the local and global search CMA-ES feature. It consists in artificially inflating and contracting the variance obtained by Bayesian update given by (1).

3 NUMERICAL RESULTS

We compare Bayesian CMA-ES (BCMA-ES) to standard CMA-ES on four 2-dimensional traditional functions: Sphere (also called the cone), Rastrigin, Schwefel 1 and 2. Because of the importance of the seed in these two algorithms, we take the average convergence and a confidence interval over 30 function evaluations. This is shown in the various figures 1, 2, 3, and 4. Standard CMA-ES used is the one provided in the open source python package pycma and displayed in *blue*, while BCMA-ES is in *red*. Full details of the implementation of the experiment is provided in supplementary materials with python source code.

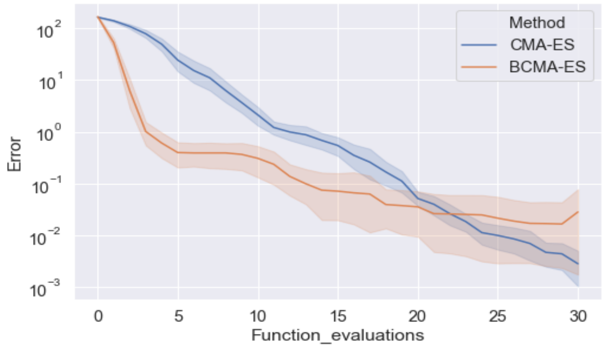

Figure 1: Convergence for the Sphere function

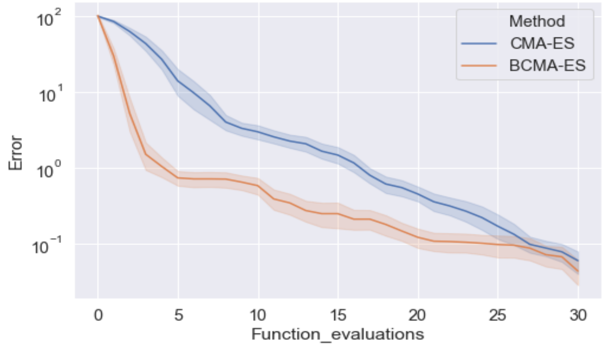

Figure 2: Convergence for the Schwefel 2 function

Overall, Bayesian CMA-ES achieves similar performance as standard CMA-ES on convex and non convex functions. BCMA-ES initial fast convergence is probably due to the optimal update. However, we observe that CMA-ES catches up rapidly and is able to continuously contract its variance which is not the case for BCMA-ES. Further research should be done to get more intuition on this interesting phenomenon and gives hints for additional improvement for BCMA-ES.

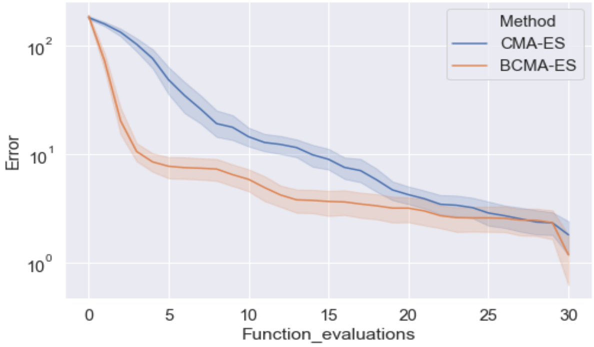

Figure 3: Convergence for the Rastrigin function

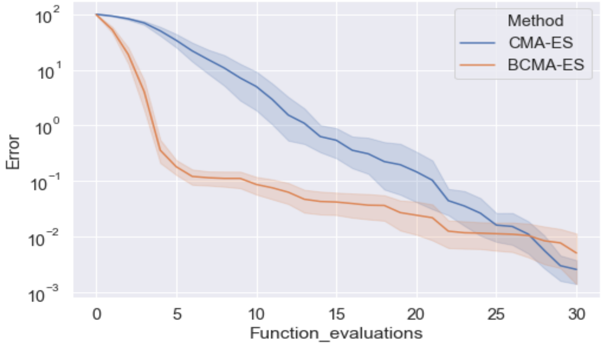

Figure 4: Convergence for the Schwefel 1 function

4 CONCLUSION

Here, we revisit CMA-ES and provide a Bayesian version of it. Taking a conjugate prior, we find the optimal update for the mean and variance. Numerical experiments show that this new version is competitive to standard CMA-ES on traditional functions such as Sphere, Schwefel 1, Rastrigin and Schwefel 2. The initial faster convergence is due to the Bayesian optimal posterior update. Further work should examine why CMA-ES continues increasing the rate of variance contraction while BCMA-ES does not achieve it.

REFERENCES

[1] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2016. CMA-ES and Advanced Adaptation Mechanisms. *GECCO, Denver* 2016 (2016), 533–562.

[2] Eric Benhamou, David Saltiel, Sébastien Vérel, and Fabien Teytaud. 2019. BCMA-ES: A Bayesian approach to CMA-ES. *arxiv* 1904.01401 (2019).

[3] Nikolaus Hansen. 2019. A Global Surrogate Assisted CMA-ES. In *GECCO 2019*. ACM, Prague, Czech Republic, 664–672.

[4] Nikolaus Hansen and Anne Auger. 2011. CMA-ES: evolution strategies and covariance matrix adaptation. *GECCO 2011* (2011), 991–1010.

[5] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation* 9, 2 (2001), 159–195.

[6] Christian Igel, Nikolaus Hansen, and Stefan Roth. 2007. Covariance Matrix Adaptation for Multi-objective Optimization. *Evol. Comput.* 15, 1 (March 2007), 1–28.

[7] Cheikh Touré, Nikolaus Hansen, Anne Auger, and Dimo Brockhoff. 2019. Uncrowded Hypervolume Improvement: COMO-CMA-ES and the Sofomore framework. In *GECCO 2019*. Prague, Czech Republic.

[8] Diederick Vermetten, Sander van Rijn, Thomas Back, and Carola Doerr. 2019. Online selection of CMA-ES variants. In *GECCO 2019*. Prague, 951–959.