Empirical Evaluation of Contextual Policy Search with a Comparison-based Surrogate Model and Active Covariance Matrix Adaptation

Alexander Fabisch
Robotics Innovation Center, DFKI GmbH
Alexander.Fabisch@dfki.de

ABSTRACT

Contextual policy search (CPS) is a class of multi-task reinforcement learning algorithms that is particularly useful for robotic applications. A recent state-of-the-art method is Contextual Covariance Matrix Adaptation Evolution Strategies (C-CMA-ES). It is based on the standard black-box optimization algorithm CMA-ES. There are two useful extensions of CMA-ES that we will transfer to C-CMA-ES and evaluate empirically: ACM-ES, which uses a comparison-based surrogate model, and aCMA-ES, which uses an active update of the covariance matrix. We will show that improvements with these methods can be impressive in terms of sample-efficiency, although this is not relevant any more for the robotic domain.

CCS CONCEPTS

• **Computing methodologies** → *Sequential decision making*;

KEYWORDS

multi-task learning, policy search, black-box optimization

ACM Reference Format:
Alexander Fabisch. 2019. Empirical Evaluation of Contextual Policy Search with a Comparison-based Surrogate Model and Active Covariance Matrix Adaptation. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3321935

1 INTRODUCTION AND STATE OF THE ART

Behaviors can be generated with reinforcement learning in robotics [3, 6]. A standard approach is policy search with movement primitives. Many episodic policy search algorithms are similar to black-box optimization. We are interested in contextual policy search: $\arg\max_{\boldsymbol{\omega}} \int_{\boldsymbol{s}} p(\boldsymbol{s}) \int_{\boldsymbol{\theta}} \pi_{\boldsymbol{\omega}}(\boldsymbol{\theta}|\boldsymbol{s}) \mathbb{E}\left[R(\boldsymbol{\theta}, \boldsymbol{s})\right] d\boldsymbol{\theta} d\boldsymbol{s}$, where $\boldsymbol{s} \in S$ is a context, $\pi_{\boldsymbol{\omega}}$ is a stochastic upper-level policy parameterized by $\boldsymbol{\omega}$ that defines a distribution of policy parameters for a given context [3]. The return R is extended to take into account the context, i.e., the context modifies the objective. During the learning process, we optimize $\boldsymbol{\omega}$, observe the current context $\boldsymbol{s}$, and select $\boldsymbol{\theta}_i \sim \pi_{\boldsymbol{\omega}}(\boldsymbol{\theta}|\boldsymbol{s})$. The deterministic problem formulation is: $\arg\min_{\boldsymbol{\omega}} \int_{\boldsymbol{s}} f_{\boldsymbol{s}}(g_{\boldsymbol{\omega}}(\boldsymbol{s}))d\boldsymbol{s}$, where $f_{\boldsymbol{s}}$ is a parameterized objective and we want to find an optimal function $g_{\boldsymbol{\omega}}$ from a parameterized class of functions. We call this contextual black-box optimization. Ideas from black-box optimization and policy search have been transerred to contextual problems. Relative entropy policy search [REPS, 13] was extended to C-REPS [10] although C-REPS is usually not robust against selection of its hyperparameters [5] and suffers from premature convergence [2]. Bayesian optimization has been extended to BO-CPS [12] and CMA-ES [8] to C-CMA-ES [1]. There are also extensions that tackle problems that are not relevant for standard black-box optimization, e.g., Contextual REPS has been extended to support active context selection [4]. In this work, we will build on one of the most promising algorithm: C-CMA-ES. It is more computationally efficient than BO-CPS and has only a few hyperparameters with good default values. Figure 1 illustrates how C-CMA-ES compares to C-REPS in a contextual optimization.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3321935

2 AC-ACM-ES

C-CMA-ES [1] is based on CMA-ES [8]. We transfer two extensions of CMA-ES to C-CMA-ES: active CMA-ES [9] and ACM-ES [11], which uses a surrogate model. We empirically tested hyperparameters of C-ACM-ES. We have two configurations: standard and aggressive exploitation of the surrogate model. We set the number of samples the surrogate after the model is accurate enough to be used to $n_{start} = 3000$ or $n_{start} = 100$. The number of samples tested with the surrogate model is set to $\lambda' = 3\lambda$ and $\lambda' = 10\lambda$ respectively. The polulation size is $\lambda = 50$. Larger values for n_{iter}, the number of iterations to train the surrogate model, improve the result. As a compromise between computational overhead and sample-efficiency, we select $n_{iter} = 1000$. c_{pow} is a parameter of the ranking SVM objective. Although in the original ACM-ES [11] the default value is 2, $c_{pow} = 1$ works better for C-ACM-ES.

3 EVALUATION

3.1 Contextual Black-box Optimization

This analysis is similar to the one of Abdolmaleki et al. [1] with additional objective functions. We will maximize $-f_{\boldsymbol{s}}$. We make standard benchmark functions contextual by defining $f_{\boldsymbol{s}}(\boldsymbol{\theta}) = f(\boldsymbol{\theta} + \boldsymbol{G}\boldsymbol{s})$, where components of the matrix $\boldsymbol{G}$ are sampled iid from $\mathcal{N}(0, 1)$. In our case $\boldsymbol{\theta} \in \mathbb{R}^{20}$ and $\boldsymbol{s} \in \mathbb{R}^{n_s}$ with $n_s = 1$ if not stated otherwise. Components of $\boldsymbol{s}$ are sampled from $[1, 2)$. To make results comparable to the one of Abdolmaleki et al. [1], we use the same sphere and Rosenbrock functions. In addition, we use the Ackley function and ellipsoidal, discus, and different powers from the COCO platform [7]. We compared several extensions of C-CMA-ES (see Table 1).

Table 1: Comparison of CPS methods, average objective of 20 runs. Best results are underlined.

Objective	Sphere	Rosenbrock	Ackley	Ellipsoidal	Diff. Powers	Discus
n_s	2	1	1	1	1	1
Test after generation	200	850	1100	800	600	850
Method	Average objective function value over contexts: $\frac{1}{\lvert S\rvert}\sum_{s\in S} f_s(x)$					
C-REPS	$-4.509 \cdot 10^{+01}$	$-1.255 \cdot 10^{+04}$	$-1.947 \cdot 10^{+01}$	$-2.944 \cdot 10^{+05}$	$-9.088 \cdot 10^{+02}$	$-1.288 \cdot 10^{+02}$
C-CMA-ES	$-1.815 \cdot 10^{-05}$	$-2.328 \cdot 10^{-03}$	$-8.762 \cdot 10^{-07}$	$-2.337 \cdot 10^{+02}$	$-1.562 \cdot 10^{-07}$	$-2.995 \cdot 10^{-10}$
aC-CMA-ES	$-1.348 \cdot 10^{-05}$	$-9.736 \cdot 10^{-01}$	$-8.773 \cdot 10^{-07}$	$-1.524 \cdot 10^{+02}$	$-3.038 \cdot 10^{-07}$	$-3.838 \cdot 10^{-10}$
C-ACM-ES+	$\underline{-1.294 \cdot 10^{-08}}$	$-1.445 \cdot 10^{+15}$	NaN	$-1.300 \cdot 10^{+16}$	$-7.111 \cdot 10^{+74}$	$-8.297 \cdot 10^{+27}$
aC-ACM-ES+	$-1.506 \cdot 10^{-01}$	$-3.227 \cdot 10^{+19}$	NaN	$-2.407 \cdot 10^{+18}$	$-8.717 \cdot 10^{+82}$	$-1.250 \cdot 10^{+24}$
C-ACM-ES	$-6.257 \cdot 10^{-04}$	$-3.656 \cdot 10^{-09}$	$\underline{-3.995 \cdot 10^{-09}}$	$-1.039 \cdot 10^{-10}$	$-2.464 \cdot 10^{-14}$	$\underline{-8.877 \cdot 10^{-12}}$
aC-ACM-ES	$-2.309 \cdot 10^{-04}$	$\underline{-3.899 \cdot 10^{-11}}$	$-1.813 \cdot 10^{-08}$	$\underline{-2.388 \cdot 10^{-11}}$	$\underline{-1.284 \cdot 10^{-14}}$	$-1.684 \cdot 10^{-11}$

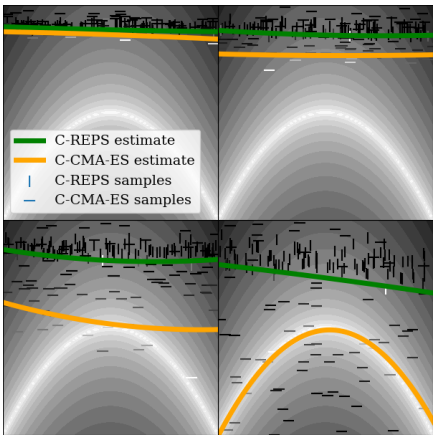

Figure 1: C-REPS vs. C-CMA-ES in a simple contextual function optimization. Values of the contextual objective are shown by background color. The optimum is a quadratic function. The x-axis represents context s and the y-axis the parameter x. The search distribution (mean indicated by sold line) is updated with 100 samples from the objective.

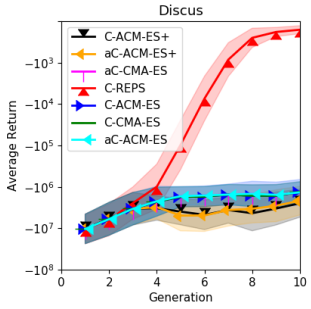

Figure 2: Learning curves for Discus function. Mean and standard deviation of 20 experiments are displayed for the first few generations.

NaN indicates divergence. We use C-REPS with the hyperparameter $\epsilon = 1$ and C-CMA-ES as baselines. The term aC-CMA-ES refers to active C-CMA-ES, C-ACM-ES uses the surrogate model, and aC-ACM-ES combines both. "+" indicates aggressive exploitation of the surrogate model. Variants of C-ACM-ES outperform vanilla C-CMA-ES. Although the surrogate model focuses on ordering the samples with the highest rank more correctly and aC-CMA-ES is often not better than C-CMA-ES, aC-ACM-ES performs best in most cases. Otherwise C-ACM-ES is better. On the sphere function, however, it is important to exploit the surrogate model as aggressively as possible to be better than C-CMA-ES. An interesting result is that C-REPS is often much faster in the early phase (see Figure 2). In the first 10 generations (500 evaluations) C-REPS outperforms all algorithms by orders of magnitude. This phase is interesting for learning in the real world. Although we see that C-REPS converges too early and variants of C-CMA-ES will continue making progress.

3.2 Conclusion

We demonstrated that the extensions active C-CMA-ES and C-ACM-ES can be combined and yield impressive results on contextual function optimization problems in comparison to C-CMA-ES. We have shown, however, that these results are actually not directly transferable to the domain of robotics, where we would like to learn successful upper-level policies in 100–1000 episodes at maximum.

ACKNOWLEDGMENTS

This work received funding from the EU's H2020 research and innovation program under grant agreement H2020-FOF 2016 723853.

REFERENCES

[1] A. Abdolmaleki, B. Price, N. Lau, L.P. Reis, and G. Neumann. 2017. Contextual Covariance Matrix Adaptation Evolutionary Strategies. In *IJCAI*. 1378–1385.
[2] A. Abdolmaleki, D. Simões, N. Lau, L.P. Reis, and G. Neumann. 2017. Learning a Humanoid Kick with Controlled Distance. In *RoboCup*. Springer, 45–57.
[3] M.P. Deisenroth, G. Neumann, and J. Peters. 2013. A Survey on Policy Search for Robotics. *Foundations and Trends in Robotics* 2, 1–2 (2013), 1–142.
[4] A. Fabisch and J.H. Metzen. 2014. Active Contextual Policy Search. *Journal of Machine Learning Research* 15 (2014), 3371–3399.
[5] A. Fabisch, J.H. Metzen, M.M. Krell, and F. Kirchner. 2015. Accounting for Task-Difficulty in Active Multi-Task Robot Control Learning. *KI - Künstliche Intelligenz* 29, 4 (2015), 369–377.
[6] L. Gutzeit, A. Fabisch, M. Otto, J.H. Metzen, J. Hansen, F. Kirchner, and E.A. Kirchner. 2018. The BesMan Learning Platform for Automated Robot Skill Learning. *Frontiers in Robotics and AI* 5 (2018), 43. https://doi.org/10.3389/frobt.2018.00043
[7] N. Hansen, A. Auger, O. Mersmann, T. Tusar, and D. Brockhoff. 2016. COCO: A Platform for Comparing Continuous Optimizers in a Black-Box Setting. *CoRR* abs/1603.08785 (2016).
[8] N. Hansen and A. Ostermeier. 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation* 9, 2 (2001), 159–195.
[9] G.A. Jastrebski and D.V. Arnold. 2006. Improving Evolution Strategies through Active Covariance Matrix Adaptation. In *CEC*. 2814–2821.
[10] A.G. Kupcsik, M.P. Deisenroth, J. Peters, and G. Neumann. 2013. Data-Efficient Generalization of Robot Skills with Contextual Policy Search. In *AAAI*.
[11] I. Loshchilov, M. Schoenauer, and M. Sebag. 2010. Comparison-Based Optimizers Need Comparison-Based Surrogates. In *PPSN*. Springer, 364–373.
[12] J.H. Metzen, A. Fabisch, and J. Hansen. 2015. Bayesian Optimization for Contextual Policy Search. In *MLPC-2015*. IROS, Hamburg.
[13] J. Peters, K. Mülling, and Y. Altun. 2010. Relative Entropy Policy Search. In *AAAI*.