Bayesian Optimization for Contextual Policy Search*

Jan Hendrik Metzen[1,2], Alexander Fabisch[1], and Jonas Hansen[1]

Abstract— **Contextual policy search allows adapting robotic movement primitives to different situations. For instance, a locomotion primitive might be adapted to different terrain inclinations or desired walking speeds. Such an adaptation is often achievable by modifying a relatively small number of hyperparameters; however, learning when performed on an actual robotic system is typically restricted to a relatively small number of trials. In black-box optimization, Bayesian optimization is a popular global search approach for addressing such problems with low-dimensional search space but expensive cost function. We present an extension of Bayesian optimization to contextual policy search. Preliminary results suggest that Bayesian optimization outperforms local search approaches on low-dimensional contextual policy search problems.**

I. INTRODUCTION

Contextual policy search (CPS) is a popular means for multi-task reinforcement learning in robotic control [1]. CPS learns a hierarchical policy, in which the lower-level policy is often a domain-specific behavior representation such as dynamical movement primitives (DMPs) [2]. Learning takes place on the upper-level policy, which is typically a conditional probability density $\pi(\theta|s)$ that defines a distribution over the parameter vectors θ of the lower-level policy for a given context s. This context vector s encodes properties of the environment or the task such as a desired walking speed for a locomotion behavior or a desired target position for a ball-throw behavior. The objective of CPS is to learn an upper-level policy which maximizes the expected return of the lower-level policy for a given context distribution.

CPS is typically based on local search based approaches such as cost-regularized kernel regression [3] and contextual relative entropy search (C-REPS) [4], [5]. From the field of black-box optimization, it is well-known that local search-based approaches are well suited for problems with a moderate dimensionality and no gradient-information. However, for the special case of relatively low-dimensional search spaces combined with an expensive cost function, which limits the number of evaluations of the cost functions, global search approaches like Bayesian optimization [6] are often superior, for instance for selecting hyperparameters [7]. Combining contextual policy search with pre-trained movement primitives[1] can also fall into this category as evaluating the cost function requires an execution of the behavior on the robot while only a small set of hyperparameters might have to be adapted. Bayesian optimization has been used for non-contextual policy search on locomotion tasks [8], [9] and robot grasping [10] before; we demonstrate an extension to contextual policy search.

II. METHOD

Bayesian optimization for contextual policy search (BO-CPS) learns internally a model of the expected return $\mathbb{E}\{R\}$ of a parameter vector θ in a context s. This model is learned by means of Gaussian process (GP) regression [11] from sample returns R_i obtained in rollouts at query points consisting of a context s_i determined by the environment and a parameter vector θ_i selected by BO-CPS. By learning a joint GP model over the context-parameter space, experience collected in one context is naturally generalized to similar contexts. Moreover, GPs allow estimating homo- and heteroscedastic noise models, which is important in robotic control tasks where sample returns R_i are typically noisy and the noise variance might differ for different behaviors.

The GP model provides both an estimate of the expected return $\mu_{GP}[R(s,\theta)]$ and the uncertainty $\sigma_{GP}[R(s,\theta)]$ of this estimate. Based on this information, the parameter vector for the given context is selected by maximizing a so-called acquisition function. These acquisition functions allow controlling the trade-off between exploitation (selecting parameters with maximal estimated return) and exploration (selecting parameters with high uncertainty). Common acquisition functions in Bayesian optimization are the probability of improvement (PI) and the expected improvement (EI) [6]. However, PI and EI both require the definition of an incumbent, which denotes the parameter vector with the maximal observed return. This notion of an incumbent is not easily extended to contextual policy search, where the incumbent is different for every context, returns are noisy, and each context will typically only be sampled once because of the continuous context space. Thus, we use the acquisition function GP-UCB instead, which does not require the definition of the incumbent. GP-UCB defines the acquisability of a parameter vector in a context as $\text{GP-UCB}(s,\theta) = \mu_{GP}[R(s,\theta)] + \kappa\sigma_{GP}[R(s,\theta)]$, where κ controls the exploration-exploitation trade-off.

BO-CPS selects parameters for a given context s_i by choosing $\theta_i = \arg\max_\theta \text{GP-UCB}(s_i,\theta)$. This maximization is performed by using the global maximizer DIRECT [12] to find the approximate global maximum, followed by L-BFGS [13] to refine it. Note that this maximization is performed only over the parameter space for fixed context

*This work was supported through two grants of the German Federal Ministry of Economics and Technology (BMWi, FKZ 50 RA 1216 and FKZ 50 RA 1217).

[1] JHM, AF, and JH are with the Robotics Research Group, Faculty of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany {jhm,afabisch}@informatik.uni-bremen.de

[2] JHM is with the Robotics Innovation Center, German Research Center for Artificial Intelligence (DFKI), 28359 Bremen, Germany

[1] A DMP might have been pre-trained, e.g., in simulation or via some kind of imitation learning for a certain fixed context.

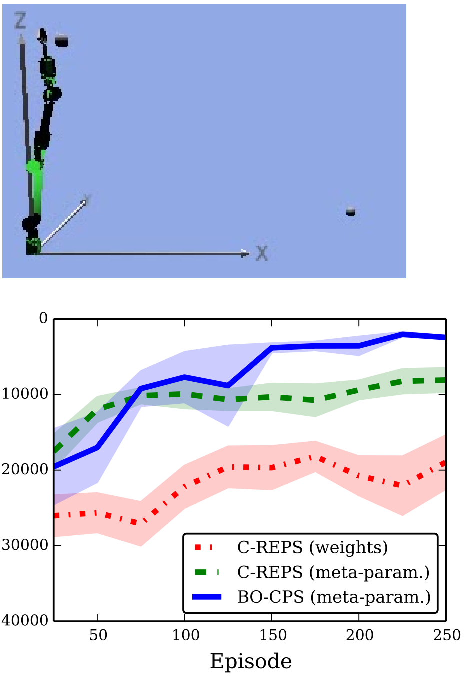

Fig. 1. Learning curves on the simulated robot arm COMPI: the offline performance is evaluated each 25 episodes on 16 test contexts distributed equally in the target area. Shown are mean and its standard error.

s_i. In principle, it would be desirable to perform the maximization over the context space as well, yielding an active learning approach. However, this would require addressing the incommensurability of returns between different context [14], which means that the achievable level of performance in different contexts can vary considerably. One approach for this might be active context selection based on multi-armed bandit learning [15] or entropy search [16]; however, we leave this to future work.

III. Results

We present preliminary results in a simulated robotic control task, in which the robot arm COMPI [17] (see Figure 1) is used to throw a ball at a target on the ground encoded in a context vector. The target area is $[1, 2.5]m \times [-1, 1]m$. We compare three learning approaches, which all start from a manually initialized DMP where the start and goal position are defined in joint space and the weights are set to zero: (1) C-REPS learning all weights of the DMP, resulting in a high dimensional search space, (2) C-REPS learning only two meta-parameters, and (3) BO-CPS learning the same two meta-parameters. These two meta-parameters are the execution time τ of the DMP, which determines how far the ball is thrown, and the final angle g_0 of the first joint, which determines the rotation of the arm around the z-axis.

For (1), we use a Gaussian upper-level policy for mapping context onto DMP weights, with the mean being a quadratic function of the context and initial variance 100. For (2), we also used a Gaussian upper-level policy for mapping context onto meta-parameters, with the mean being a quadratic function of context, initial mean $(g_0, \tau)^T = (0, 1)^T$, initial variance $\left((0.1\pi)^2, 1^2\right)$, and parameter space $g_0 \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and $\tau \in [0.2, 5]$. In (1) and (2), the upper level policy is updated after 25 episodes based on the last 100 samples. For (3), an anisotropic RBF kernel (a length scale for each dimension) is used for the GP and GP-UCB's exploration parameter is set to $\kappa = 1.0$. The reward $r = -||\boldsymbol{s} - \boldsymbol{b}_s||^2 - 0.01\sum_t \boldsymbol{v}_t^2$ includes the squared distance between the goal $\boldsymbol{s}$ and the position $\boldsymbol{b}_s$ hit by the ball as well as the squared sum of joint velocities during DMP execution.

Figure 1 shows the corresponding learning curves: although C-REPS performs initially better than BO-CPS, BO-CPS obtains a considerably higher reward than C-REPS after 150 episodes. Moreover, learning only meta-parameters is considerably faster than learning all weights. Future work is to evaluate the approach on real robots and investigate scalability to more than two context and parameter dimensions. Moreover, a combination with hierarchical approaches and intrinsic motivation [18] would be interesting.

References

[1] M. P. Deisenroth, G. Neumann, and J. Peters, "A Survey on Policy Search for Robotics," *Foundations and Trends in Robotics*, vol. 2, no. 1-2, pp. 1–142, 2013.

[2] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors," *Neural Computation*, vol. 25, pp. 1–46, 2013.

[3] J. Kober, A. Wilhelm, E. Oztop, and J. Peters, "Reinforcement learning to adjust parametrized motor primitives to new situations," *Autonomous Robots*, vol. 33, no. 4, pp. 361–379, 2012.

[4] J. Peters, K. Mülling, and Y. Altun, "Relative Entropy Policy Search," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta, Georgia, USA: AAAI Press, 2010.

[5] A. G. Kupcsik, M. P. Deisenroth, J. Peters, and G. Neumann, "Data-Efficient Generalization of Robot Skills with Contextual Policy Search," in *27th AAAI Conference on Artificial Intelligence*, June 2013.

[6] E. Brochu, V. M. Cora, and N. De Freitas, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," Tech. Rep.

[7] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2951–2959.

[8] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans, "Automatic gait optimization with gaussian process regression," 2007, pp. 944–949.

[9] R. Calandra, N. Gopalan, A. Seyfarth, J. Peters, and M. P. Deisenroth, "Bayesian gait optimization for bipedal locomotion," in *Proceedings of Learning and Intelligent OptimizatioN Conference (LION8)*, 2014.

[10] O. B. Kroemer, R. Detry, J. Piater, and J. Peters, "Combining active learning and reactive control for robot grasping," *Robot. Auton. Syst.*, vol. 58, no. 9, pp. 1105–1116, Sept. 2010.

[11] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[12] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, "Lipschitzian optimization without the Lipschitz constant," *Journal of Optimization Theory and Applications*, vol. 79, no. 1, pp. 157–181, Oct. 1993.

[13] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited-Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, vol. 16, pp. 1190–1208, 1995.

[14] A. Fabisch, J. H. Metzen, M. M. Krell, and F. Kirchner, "Accounting for Task-Difficulty in Active Multi-Task Robot Control Learning," *KI - Künstliche Intelligenz*, pp. 1–9, May 2015.

[15] A. Fabisch and J. H. Metzen, "Active contextual policy search," *Journal of Machine Learning Research*, vol. 15, pp. 3371–3399, 2014.

[16] P. Hennig and C. J. Schuler, "Entropy Search for Information-Efficient Global Optimization," *JMLR*, vol. 13, pp. 1809–1837, 2012.

[17] COMPI - compliant robot arm. [Online]. Available: http://robotik.dfki-bremen.de/en/research/robot-systems/compi.html

[18] J. H. Metzen and F. Kirchner, "Incremental Learning of Skill Collections based on Intrinsic Motivation," *Frontiers in Neurorobotics*, vol. 7, no. 11, pp. 1–12, 2013.