

---

# Factored Contextual Policy Search with Bayesian Optimization

---

Peter Karkus<sup>1,2</sup>

Andras Kupcsik<sup>2</sup>

David Hsu<sup>1,2</sup>

Wee Sun Lee<sup>2</sup>

<sup>1</sup>NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore

<sup>2</sup>School of Computing, National University of Singapore

karkus@comp.nus.edu.sg

## Abstract

Scarce data is a major challenge to scaling robot learning to truly complex tasks, as we need to generalize locally learned policies over different "contexts". Bayesian optimization approaches to contextual policy search (CPS) offer data-efficient policy learning that generalize over a context space. We propose to improve data-efficiency by factoring typically considered contexts into two components: target-type contexts that correspond to a desired outcome of the learned behavior, e.g. target position for throwing a ball; and environment type contexts that correspond to some state of the environment, e.g. initial ball position or wind speed. Our key observation is that experience can be directly generalized over target-type contexts. Based on that we introduce Factored Contextual Policy Search with Bayesian Optimization for both passive and active learning settings. Preliminary results show faster policy generalization on a simulated toy problem.

## 1 Introduction

Robots and other artificial agents act in complex environments where reinforcement learning (RL) is commonly used to learn desired behaviors by interacting with the environment. RL, however, is limited by two main factors: learned policies are difficult to generalize over different problem settings; and they require large amount of data that is often expensive to obtain, due to various costs of interaction with the environment. In order to perform truly complex tasks we need learning algorithms that generalize well over a context space with little data.

Factorization is generally used in computer science to decompose problem structures into components that are easier to address. We introduce a factorization of contexts that reduces data requirements for a class of learning tasks, where the context space exhibits specific properties.

We focus on the contextual policy search (CPS) problem, a formulation of RL, where the concept of *context* was introduced to generalize over task settings. More specifically, we use a Bayesian optimization approach to CPS based on previous work of Metzen et al. [10, 8]. A key assumption of CPS approaches is that the optimal policy parameters are similar for similar contexts. We observe however, that it is possible to directly generalize experience with some policy parameters over specific context types, which ultimately reduces data requirements.

In the CPS problem the agent is learning an upper level policy  $\pi(\theta|s)$  that defines a distribution over the parameter vectors  $\theta$  of a lower-level policy for context  $s$ . Executing the lower level policy with parameters  $\theta$  in context  $s$  results in an observed trajectory  $\tau(s, \theta)$  and reward  $R(s, \tau(s, \theta))$ .<sup>1</sup>

The context  $s = \{s^e, s^t\}$  may encode two types of task parameters,

---

<sup>1</sup>We assume that the reward function is fully known or that it can be evaluated for arbitrary  $s^t$  targets.

- desired target of the behavior,  $s^t$ , that only affects how we evaluate a trajectory,  $R(s^e, s^t, \tau(s^e, \theta))$ , e.g. for a ball throwing task the target coordinates of the ball
- properties of the environment  $s^e$  that influence the trajectory,  $\tau(s^e, \theta)$ , e.g. initial position of the ball, initial state of the robot arm or wind speed.

The key insight of this work is that a single trial provides useful information for *all* target-type contexts.

Consider a robot learning to throw balls at different targets. First it is asked to aim at target  $s_1^t$ . It chooses  $\theta_1$  according to the upper level policy, execute the throw and observes that the ball hits target  $s_2^t \neq s_1^t$ . Even if the trial is poor and results in a low reward the agent gains valuable information: in the future when it is asked to aim at  $s_2^t$  it knows that  $\theta_1$  parameters will result in hitting  $s_2^t$  and yield a high reward.<sup>2</sup> Previous work on CPS relies on reward values and an assumed correlation to generalize over contexts. The proposed factorization enables each trial to be evaluated for all target-type contexts leading to faster generalization.

We introduce Factored Contextual Policy Search with Bayesian Optimization (FCPS-BO), a variant of BO-CPS [10] with factored contexts. We also extend FCPS-BO to the active learning setting based on ACES [8]. In active learning the agent may choose the context in each episode during training, unlike the standard (passive) formulation of CPS, where the context is given.

## 2 Related work

There are a number of CPS approaches that generalize over a discrete or continuous context space. One group of work interpolates policy parameters over contexts given a number of local policies [1, 9]. This approach is limited to problems where local policies are available or easy to learn but inefficient if learning local policies are challenging themselves.

The second group of work jointly learns local policies and generalizes over the context space. Reward-Weighted Regression (RWR) [11] - and its kernelized variant Cost-Regularized Regression (CrKR) [5] - use a linear function to represent the upper-level policy and perform updates by weighted linear regression. Contextual Relative Entropy Policy Search (C-REPS) [6] represents the upper-level policy as a probability distribution over the lower-level policy parameters conditioned on the context, while updates are performed using information theoretic insights. CrKR and C-REPS were applied to a variety of real-world RL tasks: playing table tennis [5], throwing darts [5] and playing hockey [6]. Although these tasks all involve target-type contexts they are not leveraged to evaluate previous experience in new query contexts.

Recently Metzen et al. introduced BO-CPS [10] and ACES [8] that use a Bayesian optimization approach for CPS in passive and active learning settings, respectively. They directly learn a Gaussian Process (GP) model of the expected reward given a parameter vector  $\theta$  and context  $s$ . These methods offer high data-efficiency and handle noisy observations but are limited to low dimensional parameter spaces and are yet to be evaluated on real world learning tasks. Our method extends these approaches with the concept of factored contexts.

To the best of our knowledge there is no work that explicitly factors the context space. Similar ideas are implicitly used by Kober et al. who apply CrKR to learn a contextual policy for discrete targets while performing a higher-level task [4]. They map experience gained in one context to another; however, they do so for the purpose of estimating discrete outcome probabilities and not for improving the policy.

Kupcsik et al. introduced GP-REPS [7], a model-based policy search approach that iteratively learns a transition model of the system using a GP. The GP prediction is then used to generate trajectories offline for updating the policy. The authors discuss the option of generating additional samples for artificial contexts - which is a similar idea to our mapping method, -but they do not define an explicit factorization and generate multiple samples for each observed trajectory rather than explicitly mapping to the query context.

---

<sup>2</sup>Note that the argument above does not hold for environment-type contexts, such as wind speed: an imperfect throw at wind speed  $w_1$  does not tell anything about the optimal parameters for any other wind speed  $w_2$  apart from the assumed correlation.

### 3 Factored Contextual Policy Search with Bayesian Optimization

The Bayesian optimization approach to CPS trains a GP to directly map from policy parameters and contexts to the expected reward,  $E[R(s, \theta)]$ , given all previous experience,  $D = \{s_i, \theta_i, R_i\}$ . The GP posterior is an estimate of the reward  $\mu[R^*(s, \theta)]$  with uncertainty  $\sigma[R^*(s, \theta)]$ .

In each episode the agent is given a query context  $s_q$  and needs to choose an appropriate parameter vector  $\theta^*$ . We use the acquisition function,

$$\text{GP-UCB}(s_q, \theta) = \mu[R^*(s_q, \theta)] + \kappa\sigma[R^*(s_q, \theta)], \quad (1)$$

where  $\kappa$  controls the exploration-exploitation trade-off. The policy parameters are selected by simply optimizing the acquisition function given the query context,

$$\theta^* = \arg \max_{\theta} \text{GP-UCB}(s_q, \theta). \quad (2)$$

In order to leverage factored contexts we maintain a dataset  $D = \{s_i^e, \theta_i, \tau_i\}$  with trajectories  $\tau_i$  observed when executing the policy with  $\theta_i$  parameters in a context  $s_i^e$ . Note that we store full trajectories instead of rewards.<sup>3</sup> At the query stage, given  $s_q = \{s_q^t, s_q^e\}$ , we evaluate all previous experience in light of the current target-type query context  $s_q^t$ . We compute the reward for all previous trajectories,

$$R_i^q = R^t(s_q^t, s_i^r, \tau_i), \quad (3)$$

where  $s_i^r$  and  $\tau_i$  are recorded data points and  $s_q^t$  is the target context for the query point. We form the dataset  $D^q = \{s_i^e, \theta_i, R_i^q\}$  that is specific to the query context. The optimal  $\theta^*$  parameters for context  $s_q$  are then found by optimizing the GP prediction given data  $D^q$ . By evaluating all previous experience for the current query, we directly generalize over target-type contexts instead of solely relying on assumed correlation as is the case for environment-type contexts.

### 4 Active learning setting

In the active learning setting the agent chooses both parameters  $\theta$  and context  $s$  for the next trial. Optimizing GP-UCB for  $\theta$  and  $s$  would not lead to a good result as the achievable rewards for different contexts may be different. Instead we follow ACES [8] and use an Entropy Search based acquisition function<sup>4</sup> which aims to choose the most informative query points for global optimization [3].

In addition to a GP mapping from contexts and parameters to expected rewards, ACES maintains an explicit probability distribution  $p_{opt}(\theta|s)$ , the probability of  $\theta$  being the global optimal in context  $s$ . The value of  $p_{opt}$  is approximated on a set of heuristically chosen representer points using Monte Carlo integration. ACES predicts  $p_{opt}[s_q, \theta_q](\theta|s)$ , the distribution after a hypothetical query at  $\{s_q, \theta_q\}$ . To do that we need to predict the GP posterior after a query. The new posterior is approximated by first sampling from the current GP posterior at the query point and then adding the sample to the GP training data.

The most informative query point is chosen by maximizing the change in relative entropy of  $p_{opt}$  integrated over the context space. More specifically we define a loss function for a query point at context  $s$ ,

$$L^s(s_q, \theta_q) = L(p_{opt}[s_q, \theta_q](\theta|s)) - L(p_{opt}(\theta|s)), \quad (4)$$

where  $L(p_{opt})$  is the negative relative entropy between  $p_{opt}$  and a uniform measure. The acquisition function is then the sum of loss functions over different contexts,

$$\text{ACES}(s_q, \theta_q) = \sum_{i=1}^N L^{s_i}(s_q, \theta_q), \quad (5)$$

where  $\{s_i\}_{i=1}^N$  is a set of randomly chosen representer points. The next query point  $\{s_q, \theta_q\}$  is selected by minimizing  $\text{ACES}(s_q, \theta_q)$ . For further details we refer to [8] and [3].

<sup>3</sup>In a practical implementation, if  $\tau$  is too large or not observed, we may only store a mapping of the trajectory  $f(\tau)$  sufficient for computing the reward for arbitrary target-type contexts. E.g. for a ball throwing task we may only need to store the coordinates where the ball hit rather than the full trajectory.

<sup>4</sup>Entropy search could be used in the passive setting as well. We plan to do such experiments in the future.

We integrate factored contexts as follows. In each iteration we map previous experience to all  $N$  representer points in the context space,  $s_i$ , resulting in a set of GP models,  $GP_i, i = 1..N$ . When the loss function  $L^s(s_q, \theta_q)$  is evaluated for  $s = s_i$  we predict the change in  $p_{opt}$  using the corresponding  $GP_i$ . This way we directly generalize over the target-type context space, similarly to the passive learning case. Note that the choice on the target-type query context  $s_q^t$  is indifferent if we ignore rewards during training, and thus we only need to select  $\{s_q^e, \theta_q\}$  by minimizing  $ACES(s_q^e, \theta_q)$ .

## 5 Results

We run FCPS-BO on a simulated toy cannon problem (Figure 1a), where evaluating previous trials for all targets is trivially beneficial. The task is inspired by [2] and is similar in nature to the ball throwing task of [10, 8]. A cannon is placed in the center of a 3D coordinate system and has to shoot at targets on the ground in the range of  $[-11, 11] \times [-11, 11]$ m. The contextual policy maps from 2D targets to 3D launch parameters: horizontal orientation  $[0, 6.28]$ , vertical angle  $[0.01, 1.37]$  and velocity  $[0.1, 5] \frac{m}{s}$ . The reward is a negative quadratic function of the distance between the target and where the projectile hits, plus two terms penalizing larger launch speeds and vertical angles.

To increase the difficulty of the problem we randomly place hills in the environment. The learning agent is unaware of the hills and the target coordinates carry no information on the target elevation. Furthermore, we add Gaussian noise ( $\sigma_n = 1^\circ$ ) to the desired launch angle during training.

Note that evaluating previous trials for the given context is trivially beneficial in our problem setting. Shooting in a direction opposite to one target results in extremely low reward, but the same launch parameters may achieve high reward when another target is in the shooting direction.

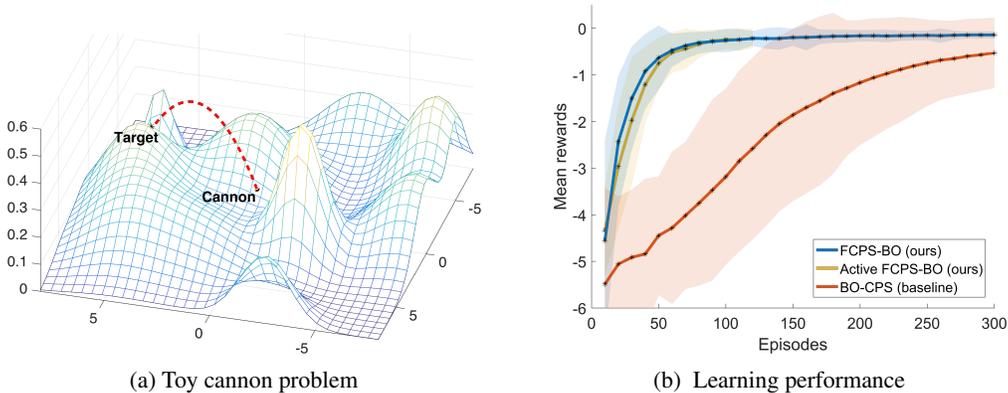


Figure 1: a) Example for an environment with a target and the optimal trajectory. b) Mean rewards using FCPS-BO (blue), FCPS-BO in active setting (yellow), and the baseline BO-CPS (orange). The shaded area corresponds to two standard deviations.

We run trials in 50 randomly generated environments and evaluate policies offline, with test targets placed on an  $8 \times 8$  grid (Figure 1b). Our factored approach, FCPS-BO, achieves near optimal performance after less than 100 episodes, while the baseline BO-CPS requires over 300 episodes. We also evaluate FCPS-BO in active learning setting; however, we observe no benefit over the passive setting. This may be due to the aggressive approximations used in the active approach; or due to our problem setting, where it is not more difficult to learn a good policy for one target than for another. Further investigation is to be conducted in the future.

## 6 Conclusion

We introduced FCPS-BO, a Bayesian optimization approach to CPS that factors the context space to generalize policies with little amount of data. We present an extension of FCPS-BO to the active learning setting, and show the benefit of factorization on a toy problem. Addressing scalability to higher dimensions and evaluation on more realistic problems remain future work. We also plan to apply factorization to standard policy search methods, such as C-REPS and GP-REPS.

## References

- [1] B. C. da Silva, G. Konidaris, A. G. Barto, and B. Castro. Learning Parameterized Skills. *ICML*, pages 1679–1686, 2012.
- [2] B. C. da Silva, G. Konidaris, and A. Barto. Active Learning of Parameterized Skills. *ICML*, 32:1737–1745, 2014.
- [3] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Machine Learning Research*, 13:1809–1837, 2012.
- [4] J. Kober and J. Peters. Learning elementary movements jointly with a higher level task. *IEEE International Conference on Intelligent Robots and Systems*, pages 338–343, 2011.
- [5] J. Kober, A. Wilhelm, E. Oztop, and J. Peters. Reinforcement learning to adjust parametrized motor primitives to new situations. *Autonomous Robots*, 33(4):361–379, 2012.
- [6] A. Kupcsik, M. P. Deisenroth, J. Peters, A. P. Loh, P. Vadakkepat, and G. Neumann. Data-Efficient Generalization of Robot Skills with Contextual Policy Search. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1401–1407, 2013.
- [7] A. Kupcsik, M. P. Deisenroth, J. Peters, A. P. Loh, P. Vadakkepat, and G. Neumann. Model-based contextual policy search for data-efficient generalization of robot skills. *Artificial Intelligence*, 2014.
- [8] J. H. Metzen. Active Contextual Entropy Search. *arXiv preprint arXiv:1511.04211*, 2015.
- [9] J. H. Metzen, A. Fabisch, L. Senger, J. de Gea Fernandez, and E. A. Kirchner. Towards Learning of Generic Skills for Robotic Manipulation. *KI-Künstliche Intelligenz*, 28:15–20, 2014.
- [10] J. H. Metzen, A. Fabisch, and J. Hansen. Bayesian Optimization for Contextual Policy Search. *Proceedings of the Second Machine Learning in Planning and Control of Robot Motion Workshop, Hamburg*, 2015.
- [11] J. Peters and S. Schaal. Applying the episodic natural actor-critic architecture to motor primitive learning. *ESANN*, pages 295–300, 2007.