

Specializing Versatile Skill Libraries using Local Mixture of Experts

Onur Celik¹, Dongzhuoran Zhou¹, Ge Li¹, Philipp Becker¹, and Gerhard Neumann¹

¹Autonomous Learning Robots, KIT, Germany

{celik, ge.li, philipp.becker, gerhard.neumann}@kit.edu
dongzhuoran.zhou@outlook.com

Abstract: A long-cherished vision in robotics is to equip robots with skills that match the versatility and precision of humans. For example, when playing table tennis, a robot should be capable of returning the ball in various ways while precisely placing it at the desired location. A common approach to model such versatile behavior is to use a Mixture of Experts (MoE) model, where each expert is a contextual motion primitive. However, learning such MoEs is challenging as most objectives force the model to cover the entire context space, which prevents specialization of the primitives resulting in rather low-quality components. Starting from maximum entropy reinforcement learning (RL), we decompose the objective into optimizing an individual lower bound per mixture component. Further, we introduce a curriculum by allowing the components to focus on a local context region, enabling the model to learn highly accurate skill representations. To this end, we use local context distributions that are adapted jointly with the expert primitives. Our lower bound advocates an iterative addition of new components, where new components will concentrate on local context regions not covered by the current MoE. This local and incremental learning results in a modular MoE model of high accuracy and versatility, where both properties can be scaled by adding more components on the fly. We demonstrate this by an extensive ablation and on two challenging simulated robot skill learning tasks. We compare our achieved performance to LaDiPS and HiREPS, a known hierarchical policy search method for learning diverse skills¹.

Keywords: Curriculum, Versatile Skill Learning, Episodic Hierarchical RL

1 Introduction

Human motor skills are precise and versatile, which allows us to perform motor tasks in different ways while achieving a consistent movement quality. For example, when playing table tennis, we can hit the ball in various ways while still targeting a specific landing point of the ball on the opponent’s side of the table. Another example is grasping an object which lies behind an obstacle. We can grasp even small objects precisely with different grasp types while avoiding the obstacle. Such versatile skills are crucial if we want to employ robots in unstructured and dynamically changing environments. Such skills, often represented as movement primitives, were already successfully learned for challenging robot learning tasks, e.g., the ball-in-a-cup [1, 2] task, by a variety of policy search algorithms [3]. Yet, most of these algorithms cannot find multiple, versatile, and precise solutions to the multi-modal solution space, as they usually assume a Gaussian policy [4, 2, 3, 5].

In this paper, we model versatile behavior with contextual skill libraries of motion primitives [6, 7], formalized by Mixtures of Experts (MoEs). Here the context defines task properties, e.g., reaching different goal positions or different friction parameters of an object to manipulate [5]. Our goal is to learn versatile skills, i.e., different movement styles to solve a given context. Given a context,

¹Code available at https://github.com/Onur4229/SVSL_LMOE

the MoE first selects a component, i.e., a motion primitive, to execute. Subsequently, the component adjusts the primitive’s parameters and a controller executes the primitive. Such models are already commonly used [8, 9]. Yet, the quality and versatility of the learned skill libraries remain limited using existing algorithms. Most algorithms are based on expectation-maximization (EM) [10] techniques [8], and suffer from local optima and mode averaging [10] which prevents the single components from specializing in a local context region. Moreover, existing algorithms [8, 9] do not explicitly optimize the versatility of the library. Hence, they often yield degenerated libraries with only a single movement style. We propose a new objective for learning contextual, precise, and versatile MoE models based on a maximum entropy formulation. We also introduce a learnable context distribution, which provides a curriculum for each component of the MoE model. We use a variational lower bound [11] to decompose the objective into individual updates for the components and their related local context distributions, allowing the components to specialize in local regions of the context space and preventing mode averaging. Due to the curriculum, the MoE does not have to cover the whole context space during training, which prevents the averaging effects that negatively affect most other approaches. Yet, not covering the whole context space leads to poor performance for some contexts, which is also undesirable. Thus, we introduce a heuristic-free mechanism to add new components during training until the whole context space is covered. Hence our algorithm provides a modular approach that learns highly precise and versatile skills that cover the whole context space.

We evaluate our approach in a simulated beer pong and a table tennis environment. Both environments allow different motion styles to solve the tasks, which are discovered by our algorithm. Moreover, we present ablation studies showing the importance of the single elements and hyperparameters of our algorithm.

2 Related Work

Contextual Episodic Policy Search. Episodic policy search [3] aims at maximizing the expected return by optimizing the parameters θ of a controller, e.g., a motion primitive [12, 6]. Most approaches use a stochastic search distribution $\pi(\theta)$ over the parameter space and aim to optimize the expected return under this search distribution [3], i.e.,

$$\max_{\pi(\theta)} \mathbb{E}_{\pi(\theta)}[R(\theta)],$$

where $R(\theta) = \sum_t r_t$ is the summed reward over a whole episode obtained when executing controller parameter θ . As it is common in the literature, we will denote $\pi(\theta)$ as our policy even though it only indirectly chooses the control actions of the agent by selecting the controller parameters θ . Different optimization schemes such as policy gradients Sehnke et al. [13], natural gradients [14], stochastic search strategies [15, 16] or trust-region optimization techniques [8, 17] have been used. Researchers extended these approaches to the contextual setting [18, 4], where the search distribution $\pi(\theta|c)$ now depends on a context vector c which describes the task, e.g., a goal location to reach. The contextual objective is typically formalized as

$$\max_{\pi(\theta|c)} \mathbb{E}_{p(c)} [\mathbb{E}_{\pi(\theta|c)}[R(\theta, c)]],$$

where $p(c)$ is the given distribution over context vector and the rewards now also depends on the context. Klink et al. [2] introduce a curriculum into contextual policy search. By having an adaptable distribution over the contexts, they allow the agent to concentrate on easy-to-solve contexts first and then generalize to the whole context space.

Versatile Skill Learning. The Hierarchical Relative Entropy Policy Search (HiREPS) algorithm [8] extends the classical Relative Entropy Policy Search (REPS) approach [19] to MoEs, which allows learning versatile skills in a contextual episodic policy search setting. In a similar approach, Layered Direct Policy Search (LaDiPS) [9] also uses MoE policies, but builds on Model-Based Relative Entropy Policy Search (MORE) [17] instead of REPS. Both HiREPS and LaDiPS address the same problems as our approach, yet there are also considerable differences. First, HiREPS jointly optimizes the whole mixture model and introduces an additional constraint, which bounds the entropy loss of the responsibilities in each iteration. This constraint is crucial for obtaining versatile and well-performing solutions. LaDiPS uses separate updates for the different parts of the mixture but also relies on additional constraints, where the entropy of the gating is lower bounded with a

constant value. In contrast, for our approach, the objective and separate updates of the individual mixture parts follow naturally from the maximum entropy formulation. Second, neither HiREPS nor LaDiPS uses a curriculum for training. Thus, in both approaches, the MoE always has to cover the whole context space and, hence, the components cannot specialize.

Variational Inference. Our work is also related to several recent advances in variational inference [11, 20, 21]. It is well known that maximum entropy RL is equivalent to inference in an appropriate probabilistic model [22]. Similar to previous works [23, 24], we exploit this relation and draw inspiration from recent research into variational inference and density estimation for Gaussian mixture models and MoEs [11, 20, 21]. We reformulate the lower bound objectives introduced in those approaches for our maximum entropy RL setting and extend them with a curriculum for the mixture components.

Related Step-Based Approaches. In the step-based setting, the policy does not learn a function from contexts to parameters of an episodic controller but directly maps from system states to actions and the policy updates are performed with the information from each time-step. Practitioners often use deep neural networks to parameterize step-based policies, giving rise to the field of deep RL. In this context, versatile policy learning is also a very active research area [25, 26, 27, 28]. These approaches use a similar MoE model where the mixture component is only chosen at the beginning of an episode. Yet, the component is chosen randomly without conditioning on a context or state variable. Moreover, these approaches reformulate a mutual information based objective into a maximum entropy objective while we develop a more direct maximum entropy maximization.

Curriculum Learning. Researchers also worked on introducing curricula into deep RL. In a first approach Ghosh et al. [29] proposed partitioning the initial state distribution using clustering. They then learn individual policies for each partition while keeping the partitioning fixed. Strictly, this is not a curriculum as the partitioning is not adjusted, yet it still allows specialization of the individual policies in different regions of the state space. To automatically generate and adapt a curriculum for deep RL approaches, Klink et al. [30] extended their approach from the episodic setting [2] to the step-based setting. Yet, neither of these approaches addresses versatility. While we follow an episode-based approach, both methodologies have their benefits and limitations [3] which are, however, not the focus of this paper. We offer further discussion and quantitative comparison to a step-based approach for a common benchmark task in the appendix, therefore.

Options. A related hierarchical approach is the options framework [31, 32, 33]. The options framework extends the standard MDP to a semi-MDP to include a temporal abstraction of low-level control policies. Given a termination condition, the executed low-level policy can be terminated and another can be turned on. Our policy structure can be seen as a simplification of the options framework where the option is only selected at the beginning of each episode. Yet, the options framework does not explicitly address learning versatile skills.

3 Specializing Versatile Mixture of Expert Models

To allow versatile solutions, we employ a Mixture of Experts (MoE) model as policy representation which is given as $\pi(\theta|c) = \sum_o \pi(o|c)\pi(\theta|c, o)$, where $\pi(o|c)$ is the gating distribution, assigning a probability to component o given the context c and $\pi(\theta|c, o)$ is the component distribution for component o , which adapts the motion primitive's parameters θ to the given context c . In this section, we derive a lower bound to optimize each component and its corresponding context distribution independently. In order to implement a curriculum over the context c , we also introduce a learned context distribution $\pi(c) = \sum_o \pi(c|o)\pi(o)$, which is also a mixture model specified by the component-wise context distribution $\pi(c|o)$ and the component weights $\pi(o)$. By applying Bayes' Rule to replace the gating distribution $\pi(o|c)$, we can now rewrite the general mixture of experts model as

$$\pi(\theta|c) = \sum_o \frac{\pi(c|o)\pi(o)}{\pi(c)}\pi(\theta|c, o). \quad (1)$$

This policy definition allows each component o to adjust its curriculum by explicitly optimizing for $\pi(c|o)$ (Section 3.3) and thus concentrating on a local region in the context space. We model the components $\pi(\theta|c, o)$ as a linear conditional Gaussian distribution and the component-wise state distribution $\pi(c|o)$ as a Gaussian. The prior weights $\pi(o)$ define a categorical distribution. Throughout

the next sections, we denote every probability distribution which is adaptable through the optimization process with π as it is part of our policy $\pi(\theta|c)$. Furthermore we show the full derivations for the next sections in the Appendix A.

3.1 Maximum Entropy Skill Learning with Curriculum

We consider a maximum entropy objective [22] for episodic policy search, i.e.,

$$\max_{\pi(\theta|c)} \mathbb{E}_{p(c)} [\mathbb{E}_{\pi(\theta|c)} [R(c, \theta)] + \alpha H[\pi(\theta|c)]] , \quad (2)$$

where $p(c)$ is the task specific context distribution, $R(c, \theta)$ is the reward function, $H(\pi(\theta|c)) = - \int_{\theta} \pi(\theta|c) \log \pi(\theta|c) d\theta$ the entropy and $\pi(\theta|c)$ is our MoE model. The reward maximization enforces preciseness while the entropy bonus enforces versatility. However, the standard maximum entropy objective does not allow each mixture component to create its own curriculum, since an optimization over the component-wise context distributions $\pi(c|o)$ is not given. Inspired by the work from Klink et al. [2], we extend and modify the objective to

$$\max_{\pi(\theta|c), \pi(c)} \mathbb{E}_{\pi(c)} [\mathbb{E}_{\pi(\theta|c)} [R(c, \theta)] + \alpha H[\pi(\theta|c)]] - \beta \text{KL}(\pi(c) \| p(c)) , \quad (3)$$

where α and β are scaling parameters, $H(\pi(\theta|c)) = - \int_{\theta} \pi(\theta|c) \log \pi(\theta|c) d\theta$ is the entropy and $\text{KL}(\pi(c) \| p(c)) = \int_c \pi(c) \log \frac{\pi(c)}{p(c)} dc$ denotes the KL-divergence. Note the difference in the optimization variables compared to Eq. (2). The Kullback-Leibler (KL) term ensures that the context distribution $\pi(c)$ is close to the task specific context distribution $p(c)$ while $\pi(c)$ can choose to have low probability in regions of the context space where the MoE model is performing poorly. Note that this objective is similar to the negative I-projection of the joint distribution $\pi(c, \theta)$ used in variational inference. Here, we also exploit the properties of the I-projection for learning the context distribution – the I-projection is mode seeking instead of mode-averaging and therefore allows for specialization on a local context area. Yet, the given objective is difficult to optimize for mixture models as the sum over the mixture components is appearing inside the log terms of the entropy and the KL. However, similar to [11], we can replace $\pi(\theta|c)$ in Obj. (3) with our mixture model and apply Bayes theorem to arrive at

$$\begin{aligned} & \max_{\pi(c, \theta)} \mathbb{E}_{\pi(o), \pi(c|o)} \left[\underbrace{\mathbb{E}_{\pi(\theta|c, o)} [R(c, \theta) + \alpha \log \pi(o|c, \theta)]}_{\text{augmented reward for component } o} + \underbrace{\beta \log p(c) + (\beta - \alpha) \log \pi(o|c)}_{\text{augmented reward for context distributions}} \right] \\ & + \alpha \mathbb{E}_{\pi(o), \pi(c|o)} [H[\pi(\theta|c, o)]] + \beta \mathbb{E}_{\pi(o)} [H[\pi(c|o)]] + \beta H[\pi(o)]. \end{aligned} \quad (4)$$

The exact derivations are given in the Appendix A. Note that Eq. (4) is equivalent to Eq. (3), yet, instead of the entropy for the whole mixture model, it now contains entropy terms for each hierarchy layer of the MoE model, i.e., $H[\pi(\theta|c, o)]$, $H[\pi(c|o)]$ and $H[\pi(o)]$, which are much simpler to compute. We also introduced log responsibilities $\pi(o|c, \theta) = \pi(\theta|c, o)\pi(o|c)/\pi(\theta|c)$ and $\pi(o|c) = \pi(c|o)\pi(o)/\pi(c)$, occurring in the augmented rewards. They return a high negative reward for component o , if the context-parameter pair (c, θ) or the context sample c is already covered by another component, pushing the component to uncovered regions of the parameter space or context space respectively. Yet, the regions for the components will still overlap due the entropy bonuses for $\pi(\theta|c, o)$ and $\pi(c|o)$. Without this reward augmentation, each component could be optimized completely independently in Eq. (4) using a max-entropy objective. However, in this case, all components would concentrate on learning the best solution irrespective of whether this solution has already been covered by another component. Yet, the log responsibilities still hinder us from optimizing each component $\pi(\theta|c, o)$ and its corresponding context distribution $\pi(c|o)$ independently, since the sum over o from the mixture models $\pi(c)$, $\pi(\theta|c)$ respectively appears in the log term. In the following sections we show, how we can overcome this limitation by introducing a lower bound inspired by variational inference [11]. As we consider each possible context as equally important, $p(c)$ is assumed uniformly distributed in a given interval in the following and thus, can be neglected in the objective.

3.2 Lower-Bound Decomposition for Component Distributions

In order to maximize the objective in Eq. (4) for each component $\pi(\theta|c, o)$ individually, we can first extract the terms which only depend on $\pi(\theta|c, o)$ for a specific o as

$$\max_{\pi(\theta|c, o)} \mathbb{E}_{\pi(c|o), \pi(\theta|c, o)} [R(c, \theta) + \alpha \log \pi(o|c, \theta)] + \alpha \mathbb{E}_{\pi(c|o)} [H[\pi(\theta|c, o)]] . \quad (5)$$

The responsibilities are still hindering us to optimize each component $\pi(\theta|\mathbf{c}, o)$ independently. However, similar to [11], we can obtain a tight lower-bound by introducing a variational distribution $\tilde{\pi}(o|\mathbf{c}, \theta)$ and replacing the responsibilities in Eq. (5) with $\tilde{\pi}(o|\mathbf{c}, \theta)$. This variational distribution is fixed during the optimization and can be computed according to the last policy model allowing us to update each component independently. It is easy to show that after the update of the variational distribution $\tilde{\pi}(o|\mathbf{c}, \theta) = \pi_{\text{old}}(o|\mathbf{c}, \theta)$, the introduced lower bound is tight. Please refer to the appendix. The resulting lower bound is a standard maximum entropy RL objective with an additional reward augmentation of $\alpha \log \tilde{\pi}(o|\mathbf{c}, \theta)$ and thus can be optimized with any suitable Policy Search algorithm. Here we use a maximum-entropy-adjusted version of contextual MORE [18].

3.3 Lower-Bound Decomposition for Context Distributions and Prior Weights

To update $\pi(\mathbf{c}|o)$ for a specific o , we extract the relevant terms from Obj. (4)

$$\max_{\pi(\mathbf{c}|o)} \mathbb{E}_{\pi(\mathbf{c}|o)} [L_c(o, \mathbf{c}) + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] + \beta H(\pi(\mathbf{c}|o)), \quad (6)$$

where $L_c(o, \mathbf{c}) = \mathbb{E}_{\pi(\theta|\mathbf{c}, o)} [\mathcal{R}(\mathbf{c}, \theta) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \theta)] + \alpha H[\pi(\theta|\mathbf{c}, o)]$ corresponds to the expected augmented maximum entropy objective of component $\pi(\theta|\mathbf{c}, o)$ in context \mathbf{c} and $\tilde{\pi}(o|\mathbf{c}) = \pi_{\text{old}}(o|\mathbf{c})$ is a second variational distribution, which we introduced to be able to optimize for each $\pi(\mathbf{c}|o)$ individually. Similarly to the previous section, the objective given in Eq. (6) is a tight lower bound to the original objective where the responsibilities $\pi(o|\mathbf{c})$ are used instead of $\tilde{\pi}(o|\mathbf{c})$. We approximate the integral over θ with a single sample, as we typically only have a single parameter sample per context available. Yet, as we still have the outer expectation $\mathbb{E}_{\pi(\mathbf{c}|o)}$ in Eq. (6) which we approximate by multiple context samples, the whole Monte-Carlo estimation of the expectations is still unbiased and with low variance. After the optimization step (Eq. 6), we obtain the optimal solution $\pi^*(\mathbf{c}|o)$ and tighten the bound by setting $\tilde{\pi}(o|\mathbf{c}) = \pi^*(o|\mathbf{c})$ and $\tilde{\pi}(o|\mathbf{c}, \theta) = \pi^*(o|\mathbf{c}, \theta)$. Again, this lower bound corresponds to an augmented maximum entropy RL objective and thus we can be updated it with any suitable policy search method. Like [11], we use an adjusted version of MORE [17].

Finally, we can formulate the objective for updating the component weights $\pi(o)$, which resembles a lower bound of the original Objective (4) and corresponds to the highest hierarchy in our update scheme. The objective is given as

$$\max_{\pi(o)} \sum_o \pi(o) [\mathbb{E}_{\pi(\mathbf{c}|o)} [L_c(o, \mathbf{c}) + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] + \beta H(\pi(\mathbf{c}|o))] + \beta H(\pi(o)), \quad (7)$$

which is a maximum entropy RL objective for categorical distributions. Here, we use REPS [19].

To summarize, we split the initial objective in Eq. (4) into different hierarchies, allowing us to optimize the different terms in our mixture model individually. Starting by first updating the components $\pi(\theta|\mathbf{c}, o)$ using the maximization problem in Eq. (5), we can optimize for the component-wise context distributions $\pi(\mathbf{c}|o)$ with Obj. (6) after tightening the bound. The components $\pi(\theta|\mathbf{c}, o)$ adjust the movement primitive parameters θ given a context \mathbf{c} , while the component-wise context distributions $\pi(\mathbf{c}|o)$ ensure that the components only see context samples from a local context region. Finally, we update the weight distribution $\pi(o)$ using Eq. (7).

Algorithmic Details and Addition of Components. We initialize our algorithm with only one component and incrementally add components, and their corresponding context distributions, randomly. We fix all components except for the newly added one and optimize it for K iterations to let it discover new solutions in yet undiscovered context regions. For updating $\pi(\theta|\mathbf{c}, o)$ (Obj. (5)) we sample from the local context distribution $\pi(\mathbf{c}|o)$. By also updating $\pi(\mathbf{c}|o)$ according to Obj. (6), the components can adjust their curriculum and search for their favored context regions. After K iterations we add a new component and repeat the procedure. Due to the augmented rewards, the new component will focus on undiscovered solutions and the local context distribution will cover uncovered areas of the context space. Note that such a simple adding procedure is only possible due to the mode-seeking properties of the I-projection, as the components do not need to average over multiple modes but can specialize on a local context region. We also fix the weights $\pi(o)$ to be uniformly distributed among all components during learning, since otherwise components which are already fully trained might dominate the optimization. By allowing to fine-tune all components every H iterations, the previously added components can adjust to the newly added ones. After finishing adding components, we update the weights $\pi(o)$ at the end of our optimization procedure.

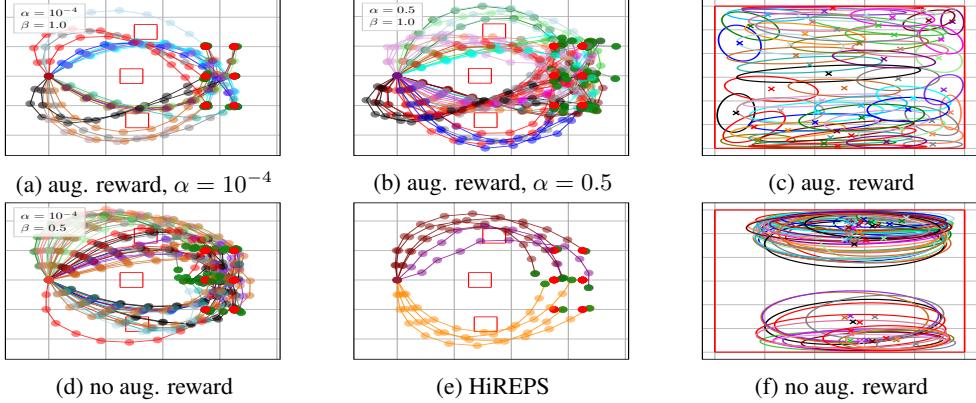


Figure 1: **Importance of the responsibilities in the augmented rewards and comparison to HiREPS.** A snapshot of the learned policies of the planar reaching task for a 2-dim context space (illustrated in (c) and (f)) by considering $\log \tilde{\pi}(o|\mathbf{c}, \theta)$, $\log \tilde{\pi}(o|\mathbf{c})$ ((a)+(b)+(c)) and neglecting the auxiliary distributions ((d)+(f)). In ((a)+(b)+(c)) we can learn more diverse solutions (a) + (b) and cover the whole 2d-context space (c) with different entropy bonuses, whereas the solutions in ((d)+(f)) show nearly the same, partially invalid solutions by going through the red rectangles (obstacles) (d) and are not able to cover the whole context space (f) leading to poor generalization performance. The solutions by HiREPS (e) indicate less versatility compared to the solutions (a) + (b). Note that each color indicates a different component and the red dots indicate the 6 chosen context vectors used for sampling.

The variables K and the number of components added in total are task-dependent. We describe the algorithm in more detail in the Appendix B.

4 Experiments

We start by investigating the importance of the different terms of the objectives derived in Section 3 and subsequently evaluate the versatility and precision of the learned skills on simulated robot beer pong and table tennis experiments. In Appendix C we report all hyperparameters and provide a comparison and analysis to a step-based policy search strategy (PPO [34]) in different scenarios.

4.1 Ablation Studies

We investigate the importance of the augmented rewards on a 10-link planar reaching task with two-dimensional context space. For this purpose, we update the component distributions $\pi(\theta|\mathbf{c}, o)$ and the corresponding context distributions $\pi(\mathbf{c}|o)$ by optimizing the objectives in (5, 6) with i) considering the responsibilities (as given by our algorithm) and ii) by setting $\log \tilde{\pi}(o|\mathbf{c}, \theta)$ and $\log \tilde{\pi}(o|\mathbf{c})$ to zero. Furthermore, we compare the solutions found by our method to the solutions found by the SOTA method HiREPS.

Versatility is Induced by the Augmented Rewards. In the 10-link planar reaching task (see Figure 1a) the robot has to reach the red dots with its end-effector in a 2-dim context-space, while avoiding the rectangle-shaped obstacles. By adding 60 components, we have trained both versions (i and ii) over 15 seeds/trials and have chosen the best performing parameter constellations (Fig. 1a) for i), Fig. 1d) for ii)). We then picked the first model and sampled for each of the shown six contexts vectors (red dots) 100 samples and plotted the corresponding mean of each sampled component. The plots for i) (Fig. 1a) + Fig. 1b)) show versatile – several modes to the same context – and precise – small distance of end-effector (green dot) to goal (red dot), while avoiding obstacles– solutions while fully covering the context space (Fig. 1c). A trend of covering more modes with higher α can also be seen in (Fig. 1b)). For the case without the responsibilities ii) (Fig. 1d)), the solutions are not precise and invalid –reaching through the obstacles–. As Fig. 1f)) shows, the context distributions focus on easy-to-solve context regions (top and bottom part) and do not cover the full context space, leading to extrapolated solutions from components that are not trained for these contexts (red dots).

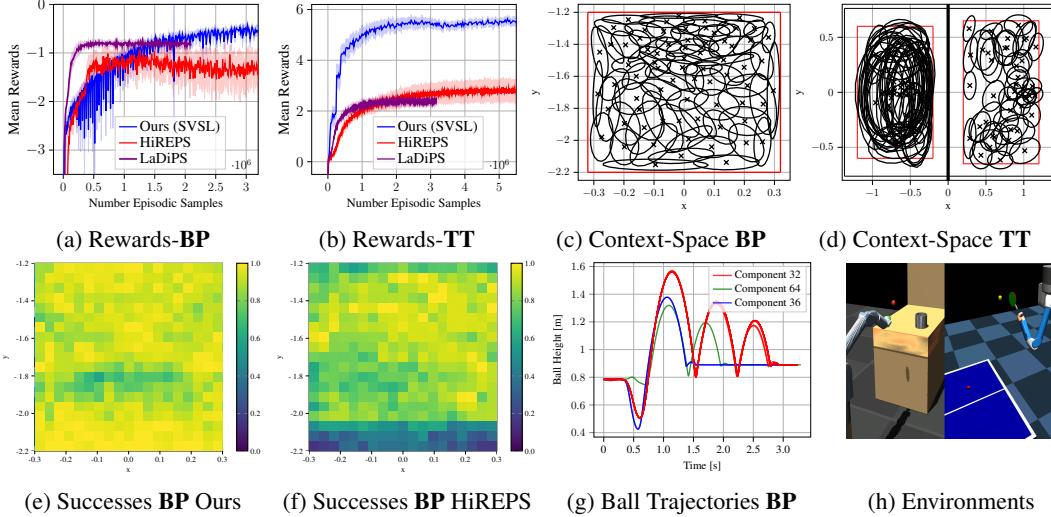


Figure 2: **Beer Pong (BP) and Table Tennis (TT) Experiments.** In the **BP** experiment –left of (h)– the robot has to throw the red ball into a cup on the table. In the **TT** experiment –right of (h)– the robot has to return the ball to the opponent’s table side. The learned mixture of expert (MoE) models show high quality of the skills, reflected in the high reward values (a + b) and dense coverage of the context spaces, i.e., the 2-dim. position of the cup (c) for **BP** and the 4-dim desired outgoing landing and incoming landing ball positions (d) for **TT**. For **BP** we also compare the success rates (92% in average) of our approach (e) and HiREPS (75% in average) (f) throughout the context space, where we consider a trial successful if the ball goes into the cup. For **BP** versatile skills induce versatile ball trajectories for a given context (g). Versatile strikes for the **TT** experiment are shown in Fig. 3)

Comparison to HiREPS. We pick the best performing model among 15 seeds/trials after hyperparameter optimization. In Fig. 1e) the mean solutions of the sampled components –sampled in the same way as before – can be seen. HiREPS shows less diverse and qualitative solutions, where only 3 of the 60 components were chosen in total.

4.2 Simulated Robotic Experiments

We test and compare our algorithm on simulated Beer Pong and Table Tennis environments in MuJoCo [35], where we have chosen HiREPS [8] and LaDiPS [9] as baselines. Both methods are suitable baselines since they are contextual policy search algorithms and consider optimizing a mixture model. Throughout our experiments we choose probabilistic movement primitives (ProMP) [6] as policy representation, where depending on the weights θ desired trajectories are generated and subsequently tracked with a PD-controller. In our experiments, given a context c , the components $\pi(\theta|c, o)$ adjust the weight vector and the length of the trajectory, which are summarized in the vector θ . We consider non-markovian rewards, in which the reward depends on the history of state and actions. This type of reward function is not applicable to common step-based RL methods which build on markovian properties. As for analyzing the results of the experiments, we focus on the questions i) how does our algorithm perform compared to SOTA baselines, ii) are we able to cover the whole context space, and iii) are we able to learn versatile skills?

Beer Pong. The goal of the Barrett WAM robot is to throw the red ball into the cup on the table. The 2-dim contexts resemble the position of the cup on the table. We incrementally add 70 components in our method, while HiREPS and LaDiPS directly start with 70 components. We have run all algorithms over 20 seeds/trials and report the performance in Fig. 2a)², where we plot the mean reward with two times the standard error. While HiREPS already converges after around half a million samples, our approach can quickly outperform it. Since LaDiPS uses intra-option learning, it outperforms our method in terms of sample efficiency. However, with the increasing number of components, we achieve a higher end-reward indicating that we can learn more qualitative so-

²To reflect the model’s true performance, the lastly added component was excluded from testing, since it is not fully trained yet and would not be chosen by the model if $\pi(o)$ would not be a uniform-distribution.

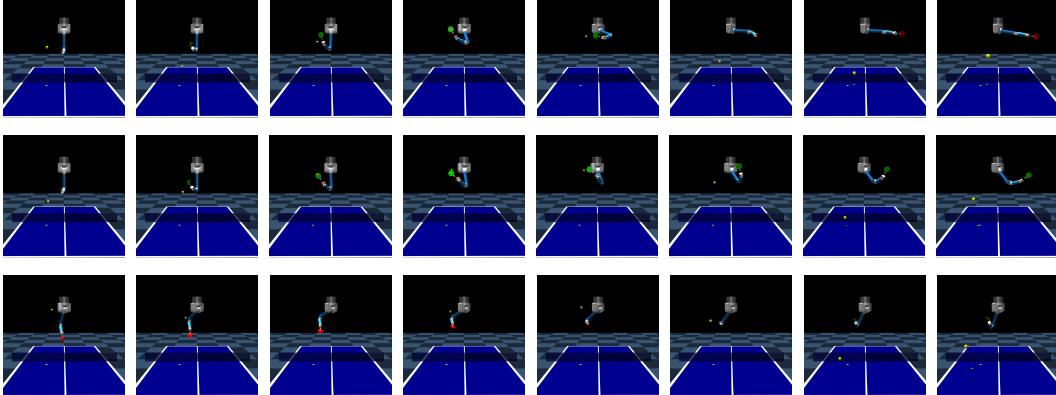


Figure 3: **Versatile Strikes for the Table Tennis (TT) Experiment** illustrated for a fixed context. The robot can hit the (yellow) ball in various ways, also indicated through the different colors of the racket sides. Note that the red and yellow dots on the table are markers for the serving and desired landing position respectively.

lutions. Our algorithm allows to cover the whole context space with the learned component-wise distributions (see Fig. 2c)) resulting in a high end-reward. This performance is reflected in Fig. 2e), where we have divided the context space into a fine grid and sampled 100 times for each context a component from which we have executed the mean. We repeated that for all 20 different models and plot the mean success rate of throwing the ball into the cup. We did the same procedure for HiREPS in Fig. 2f). We can observe that HiREPS has much darker rectangles, showing that the success rate in these context regions is low. Although versatility is encouraged in the joint space of the robot, different joint trajectories often yield different ball trajectories. Given one context, in Fig. 2g) we show the z-coordinates of the ball trajectories over time resulting from sampling 20 times from the MoE model. Each component leads to a different number of "ball-jumps". In Appendix C.2.1 we qualitatively show that we can learn more versatile solutions than LaDiPS and report that we can achieve a much higher expected mixture entropy (Fig. 4), which indicates that our solutions are versatile.

Table Tennis. The task of the robot is to return different incoming balls to desired targets on the opponent’s table side in different ways. We consider a four dimensional context space, including ball’s initial serve position and ball’s target landing position (right and left half of table Fig. 2d)). For both parts of the context, we fix the z position and vary the x and y coordinates. We incrementally add 50 components in our method, while HiREPS and LaDiPS directly start with 50 components. We have run all algorithms over 20 seeds/trials and report the performance in Fig. 2b)², in which the mean reward with two times the standard error is plotted. The component-wise context distributions are spread among the context space (Fig. 2d)) and allow each component to locally specialize on a context region. This high coverage of the context space allows for a high reward achievement, outperforming HiREPS and LaDiPS. In Fig. 3), we show three different striking skills sampled from our trained MoE model, for a fixed context, i.e., fixed serving and desired ball position. The first two skills use the green side of the racket to hit the ball (forehand), while the third skill uses the red side of the racket (backhand) to hit the ball. In contrast to the second strike, the first one performs a smash-like strike and ends it with the red side of the racket pointing to the camera.

5 Conclusion

We proposed a new objective for learning contextual and versatile Mixtures of Experts (MoE) models. We based our objective on a maximum entropy formulation to increase the versatility of the solutions and introduced a curriculum to allow the components to specialize. Our formulation also allows for easy online adaptation of the model complexity during training. We conducted an ablation to show the importance of the individual parts of our objective. Further, we showed that our method learns precise and versatile solutions and outperforms the baseline on sophisticated simulated robotic tasks. This work aims at presenting a mathematically well-founded method and demonstrates its general feasibility on various challenging tasks. Currently, the major drawback of our approach is sample efficiency, as we do not share experience between the components. We

intend to address this issue in future work, e.g., by intra-option learning. Another direction for future research is extending the approach to more complex models, such as non-linear mixtures of experts. We expect to need fewer components to cover the whole context space with more complex component model representations.

Acknowledgments

Calculations for this research were conducted on the high performance computer of the state of Baden-Württemberg.

References

- [1] J. Kober and J. Peters. Policy search for motor primitives in robotics. In *Learning Motor Skills*, pages 83–117. Springer, 2014.
- [2] P. Klink, H. Abdulsamad, B. Belousov, and J. Peters. Self-paced contextual reinforcement learning. In *Conference on Robot Learning*, pages 513–529. PMLR, 2020.
- [3] M. P. Deisenroth, G. Neumann, J. Peters, et al. A survey on policy search for robotics. *Foundations and trends in Robotics*, 2(1-2):388–403, 2013.
- [4] A. Abdolmaleki, D. Simões, N. Lau, L. P. Reis, and G. Neumann. Contextual direct policy search. *Journal of Intelligent & Robotic Systems*, 96(2):141–157, 2019.
- [5] A. Kupcsik, M. Deisenroth, J. Peters, and G. Neumann. Data-efficient generalization of robot skills with contextual policy search. In *Proceedings of the AAAI conference on artificial intelligence*, volume 27, 2013.
- [6] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann. Probabilistic movement primitives. *Advances in neural information processing systems*, 26:2616–2624, 2013.
- [7] A. J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. Technical report, 2002.
- [8] C. Daniel, G. Neumann, and J. Peters. Hierarchical relative entropy policy search. In *Artificial Intelligence and Statistics*, pages 273–281. PMLR, 2012.
- [9] F. End, R. Akour, J. Peters, and G. Neumann. Layered direct policy search for learning hierarchical skills. pages 6442–6448, 05 2017. doi:[10.1109/ICRA.2017.7989761](https://doi.org/10.1109/ICRA.2017.7989761).
- [10] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [11] O. Arenz, G. Neumann, and M. Zhong. Efficient gradient-free variational inference using policy search. In *International conference on machine learning*, pages 234–243. PMLR, 2018.
- [12] S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert. Learning movement primitives. In *Robotics research. the eleventh international symposium*, pages 561–572. Springer, 2005.
- [13] F. Sehnke, C. Osendorfer, T. Rückstiess, A. Graves, J. Peters, and J. Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 21(4):551–559, May 2010. doi:[10.1016/j.neunet.2009.12.004](https://doi.org/10.1016/j.neunet.2009.12.004).
- [14] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.
- [15] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.
- [16] S. Mannor, R. Y. Rubinstein, and Y. Gat. The cross entropy method for fast policy search. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 512–519, 2003.
- [17] A. Abdolmaleki, R. Lioutikov, N. Lua, L. Paulo Reis, J. Peters, and G. Neumann. Model-based relative entropy stochastic search. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 153–154, 2015.
- [18] V. Tangkaratt, H. van Hoof, S. Parisi, G. Neumann, J. Peters, and M. Sugiyama. Policy search with high-dimensional context variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

- [19] J. Peters, K. Mulling, and Y. Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.
- [20] O. Arenz, M. Zhong, and G. Neumann. Trust-region variational inference with gaussian mixture models. *Journal of Machine Learning Research*, 21(163):1–60, 2020.
- [21] P. Becker, O. Arenz, and G. Neumann. Expected information maximization: Using the i-projection for mixture density estimation. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=ByglL1HFDS>.
- [22] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.
- [23] G. Neumann et al. Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 817–824, 2011.
- [24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. URL <http://proceedings.mlr.press/v80/haarnoja18b.html>.
- [25] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *CoRR*, abs/1802.06070, 2018. URL <http://arxiv.org/abs/1802.06070>.
- [26] S. Kumar, A. Kumar, S. Levine, and C. Finn. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *arXiv preprint arXiv:2010.14484*, 2020.
- [27] T. Osa, V. Tangkaratt, and M. Sugiyama. Discovering diverse solutions in deep reinforcement learning. *arXiv preprint arXiv:2103.07084*, 2021.
- [28] V. Campos, A. Trott, C. Xiong, R. Socher, X. Giró-i Nieto, and J. Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, pages 1317–1327. PMLR, 2020.
- [29] D. Ghosh, A. Singh, A. Rajeswaran, V. Kumar, and S. Levine. Divide-and-conquer reinforcement learning. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=rJwelMbR->.
- [30] P. Klink, C. D’Eramo, J. R. Peters, and J. Pajarinen. Self-paced deep reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9216–9227. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf>.
- [31] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- [32] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [33] M. Riemer, M. Liu, and G. Tesauro. Learning abstract options. *arXiv preprint arXiv:1810.11583*, 2018.
- [34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.
- [35] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi:[10.1109/IROS.2012.6386109](https://doi.org/10.1109/IROS.2012.6386109).
- [36] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016.

- [37] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann. Stable baselines3.
<https://github.com/DLR-RM/stable-baselines3>, 2019.

A Derivations

In the following we derive our objectives from Section 3.

A.1 Maximum Entropy Skill Learning with Curriculum

We start our derivation from the KL-regularized maximum entropy objective

$$\max_{\pi(\mathbf{c}, \boldsymbol{\theta})} \mathbb{E}_{\pi(\mathbf{c})} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c})} [R(\mathbf{c}, \boldsymbol{\theta})] + \alpha H[\pi(\boldsymbol{\theta}|\mathbf{c})]] - \beta KL(\pi(\mathbf{c}) \parallel p(\mathbf{c})) \quad (8)$$

We use following identities

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{c}) &= \sum_o \frac{\pi(\mathbf{c}|o)\pi(o)}{\pi(\mathbf{c})}\pi(\boldsymbol{\theta}|\mathbf{c}, o) \\ \log \pi(\boldsymbol{\theta}|\mathbf{c}) &= \log \frac{\pi(\boldsymbol{\theta}|\mathbf{c}, o)\pi(o|\mathbf{c})}{\pi(o|\mathbf{c}, \boldsymbol{\theta})} \\ \pi(\mathbf{c}) &= \sum_o \pi(\mathbf{c}|o)\pi(o) \\ \log \pi(\mathbf{c}) &= \log \frac{\pi(\mathbf{c}|o)\pi(o)}{\pi(o|\mathbf{c})} \end{aligned}$$

and insert them into our objective in Equation (8), which leads to

$$\begin{aligned} L &= \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [R(\mathbf{c}, \boldsymbol{\theta}) - \alpha \log \pi(\boldsymbol{\theta}|\mathbf{c}, o) - \alpha \log \pi(o|\mathbf{c}) + \alpha \log \pi(o|\mathbf{c}, \boldsymbol{\theta})]] \\ &\quad + \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [-\beta \log \pi(\mathbf{c}|o) + \beta \log \pi(o|\mathbf{c}) - \beta \log \pi(o)]. \end{aligned} \quad (9)$$

Note that we have dropped the $\log p(\mathbf{c})$ term since we assume $p(\mathbf{c})$ to be uniformly distributed in a given interval.

By further rearranging the terms we can reformulate this objective as

$$\begin{aligned} L &= \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [R(\mathbf{c}, \boldsymbol{\theta}) - \alpha \log \pi(\boldsymbol{\theta}|\mathbf{c}, o) + \alpha \log \pi(o|\mathbf{c}, \boldsymbol{\theta})]] \\ &\quad + \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [-\beta \log \pi(\mathbf{c}|o) + (\beta - \alpha) \log \pi(o|\mathbf{c}) - \beta \log \pi(o)]. \end{aligned} \quad (10)$$

As stated in the paper, we can not optimize this objective. Therefore we introduce the auxiliary distributions $\tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})$ and $\tilde{\pi}(o|\mathbf{c})$ as

$$\begin{aligned} L &= \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [R(\mathbf{c}, \boldsymbol{\theta}) - \alpha \log \pi(\boldsymbol{\theta}|\mathbf{c}, o) + \alpha \log \pi(o|\mathbf{c}, \boldsymbol{\theta}) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})] \\ &\quad - \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})] + \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [-\beta \log \pi(\mathbf{c}|o) - \beta \log \pi(o) + (\beta - \alpha) \log \pi(o|\mathbf{c}) \\ &\quad + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c}) - (\beta - \alpha) \tilde{\pi}(o|\mathbf{c})]. \end{aligned}$$

We can rearrange the terms further to

$$\begin{aligned} L &= \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [R(\mathbf{c}, \boldsymbol{\theta}) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})] + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] \\ &\quad + \alpha \mathbb{E}_{\pi(o), \pi(\boldsymbol{\theta}|\mathbf{c}, o)} [\mathcal{H}(\pi(\boldsymbol{\theta}|\mathbf{c}, o))] + \beta \mathbb{E}_{\pi(o)} [\mathcal{H}(\pi(\mathbf{c}|o))] + \beta \mathcal{H}(\pi(o)) \\ &\quad + \alpha \mathbb{E}_{\pi(\mathbf{c}), \pi(\boldsymbol{\theta}|\mathbf{c})} [KL(\pi(o|\mathbf{c}, \boldsymbol{\theta}) || \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta}))] + (\beta - \alpha) \mathbb{E}_{\pi(\mathbf{c})} [KL(\pi(o|\mathbf{c}) || \tilde{\pi}(o|\mathbf{c}))]. \end{aligned} \quad (11)$$

A.2 Lower-Bound Decomposition for Component Distributions

By observing that not all terms depend on $\pi(\boldsymbol{\theta}|\mathbf{c}, o)$, we extract only the important ones from the objective in Equation (11) as

$$\begin{aligned} \tilde{L}_c &= \sum_o \pi(o)\mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [R(\mathbf{c}, \boldsymbol{\theta}) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})]] + \alpha \mathbb{E}_{\pi(o), \pi(\boldsymbol{\theta}|\mathbf{c}, o)} [\mathcal{H}(\pi(\boldsymbol{\theta}|\mathbf{c}, o))] \\ &\quad + \alpha \mathbb{E}_{\pi(\mathbf{c}), \pi(\boldsymbol{\theta}|\mathbf{c})} [KL(\pi(o|\mathbf{c}, \boldsymbol{\theta}) || \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta}))]. \end{aligned}$$

Since the KL is always positive, we can write the lower bound to this objective as

$$L_c = \sum_o \pi(o) \mathbb{E}_{\pi(\mathbf{c}|o)} [\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [\mathbf{R}(\mathbf{c}, \boldsymbol{\theta}) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})]] + \alpha \mathbb{E}_{\pi(o), \pi(\boldsymbol{\theta}|\mathbf{c}, o)} [\mathcal{H}(\pi(\boldsymbol{\theta}|\mathbf{c}, o))],$$

such that $\tilde{L}_c \geq L_c$ always holds. After maximizing L_c w.r.t. $\pi(\boldsymbol{\theta}|\mathbf{c}, o)$ we tighten the lower bound by updating the responsibilities as described in the paper.

A.3 Lower Bound Decomposition for Component-wise Context Distributions

By neglecting all terms which do not depend on $\pi(\mathbf{c}|o)$ in objective (11), we can write the objective for optimizing w.r.t. $\pi(\mathbf{c}|o)$ as

$$\begin{aligned} \tilde{L}_k &= \sum_o \pi(o) \mathbb{E}_{\pi(\mathbf{c}|o)} [L_c(o, \mathbf{c}) + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] + \beta \mathbb{E}_{\pi(o)} [\mathcal{H}(\pi(\mathbf{c}|o))] \\ &\quad + (\beta - \alpha) \mathbb{E}_{\pi(\mathbf{c})} [\text{KL}(\pi(o|\mathbf{c}) || \tilde{\pi}(o|\mathbf{c}))], \end{aligned}$$

where

$$L_c(o, \mathbf{c}) = \mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{c}, o)} [\mathbf{R}(\mathbf{c}, \boldsymbol{\theta}) + \alpha \log \tilde{\pi}(o|\mathbf{c}, \boldsymbol{\theta})] + \alpha \mathcal{H}(\pi(\boldsymbol{\theta}|\mathbf{c}, o)).$$

Again we can observe that the KL term is always positive and we can write the lower bound to \tilde{L}_k as

$$L_{c,k} = \sum_o \pi(o) \mathbb{E}_{\pi(\mathbf{c}|o)} [L_c(o, \mathbf{s}) + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] + \beta \mathbb{E}_{\pi(o)} [\mathcal{H}(\pi(\mathbf{c}|o))],$$

such that $\tilde{L}_k \geq L_{c,k}$ holds. By setting the auxillary distributions to the responsibilities of the updated model, we tighten the bound.

A.4 Lower Bound Decomposition for Prior Weights

Finally we can write down the objective for updating the prior weights $\pi(o)$ as

$$L_p = \sum_o L_{c,k}(o) + \beta \mathcal{H}(\pi(o)), \tag{12}$$

where

$$L_{c,k}(o) = \pi(o) \mathbb{E}_{\pi(\mathbf{c}|o)} [L_c(o, \mathbf{s}) + (\beta - \alpha) \log \tilde{\pi}(o|\mathbf{c})] + \beta \mathcal{H}(\pi(\mathbf{c}|o)).$$

Some context regions naturally lead to low reward due to for example high action regularizations. Still, these context regions should be discovered, even if the auxiliary rewards $\tilde{\pi}(o|\mathbf{c})$ are not sufficient. For this purpose, we calculate a value function, which reveals the mean reward in a context. If solely one component covers this context region, it will get a high value for updating its weight $\pi(o)$, although the task reward might be bad compared to other components in other regions. We use importance sampling, to calculate the Value function

$$V(\mathbf{c}) = \sum_o \frac{\pi(o|\mathbf{c})}{\hat{\pi}(o|\mathbf{c})} \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{c}, o) \mathbf{R}(\mathbf{c}, \boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{13}$$

where $\hat{\pi}(o|\mathbf{c})$ are the gating probabilities before starting to update the weights $\pi(o)$ at the end of our algorithm (see Algorithm 1). We completely resample the small replay buffer for all components before starting to update $\pi(o)$. We use the Nadaraya Watson predictor to get $V(\mathbf{c})$. Thus, the update rule for $\pi(o)$ is given by

$$L_p = \sum_o L_{c,k}(o) - V(\mathbf{c}) + \beta \mathcal{H}(\pi(o)). \tag{14}$$

Algorithm 1 Versatile Skill Learning

Input: $\alpha, \beta, \beta_w, N, K, H$
Output: $\pi(\theta|c)$

```
1: for  $k = 1$  to  $N$  do
2:    $\pi(\theta|c, o), \pi(c|o) \leftarrow$  randomly_add_component()
3:    $\pi(o) = 1/k, \forall o$ 
4:   for  $i = 1$  to  $K$  do
5:     if  $i \% H$  is 0 then
6:       update_all_components( $\alpha$ ) to obj. (5) each
7:       update_all_context_distributions( $\alpha, \beta$ ) to obj. (6) each
8:     else
9:        $\pi(\theta|c, o) \leftarrow$  update_component_k( $\alpha$ ) to obj. (5)
10:       $\pi(c|o) \leftarrow$  update_context_distribution_k( $\alpha, \beta$ ) to obj (6)
11:    end if
12:   end for
13:   if component_k in local optimum then
14:     delete_component_k()
15:   end if
16: end for
17:  $\pi(o) \leftarrow$  update_prior_weights( $\beta_w$ ) to obj. (14)
18: delete_redundant_components()
```

B Algorithmic Details

We describe algorithmic details in the algorithm box 1. We start with one component, which is randomly added and incrementally add components after training the lastly added component for K iterations. We update the component and context distributions individually. Note that there is no strict order on updating the components or context distributions first.

Every H iterations we fine-tune all components, if more than one component is available. This allows the already trained components to adjust for the newly added one. After finishing training the lastly added component k , we check if it converged to a local optimum. There are different ways on checking that. For example by comparing the entropy, or achieved task reward to already trained components. This step is not absolutely necessary since for the upcoming fine tune steps, the component might improve. For the Beer Pong and Table Tennis task, we have disabled the deletion step in line 13.

As last step we update the prior weights $\pi(o)$ and delete all components, which are below a threshold value (e.g. 10^{-5}) for $\pi(o)$.

Note that in some experiments, it might happen that highly versatile skills result in worse rewards, e.g. through action regularization, although the task is solved successfully. In order to avoid deleting these diverse skills from the skill library at the end of the optimization, a higher β value, β_w for the entropy of $\pi(o)$ in Equation 14 can be used. This higher β_w value will encourage to put weights on skills which are not achieving highest reward. Although those skills still will have lower weight compared to other components, they are kept in the skill set and might be useful. For example when the model should be adapted to an environmental change. We made use of a different β_w value only once in the Beer Pong task (see Experimental Details C).

Note that we have used a variant of MORE [17] for optimizing obj. 6, as described in [20] and a variant of Contextual MORE [18] for optimizing obj. 5 as described in [21].

Also note, that for updating each component we use a small replay buffer. The newly taken samples replace the oldest samples in the buffer. For the updates we simply use all samples in the buffer for each component.

C Experimental Details

We describe the different hyper parameters and environment specifications used in the experiments part here. Please note: We included a context punishment in all reward functions for all experiments for contexts which were sampled out of the context range defined by the environment. This encourages the context distributions $\pi(\mathbf{c}|o)$ to stay in the valid context region as defined from the environment. We report the reward functions and the environment specifications in the sub section for each experiment.

As for the hyper parameters for all experiments for our Algorithm, we provide a summary in Table 1. Following the notation from Algorithm 1, β_w is the entropy coefficient for $\pi(o)$ in Objective 14, N is the number of incrementally added components, K is the number of iterations one component is trained and H describes the number that every H . iteration all components are updated.

For the hyper parameters for all experiments for HiREPS and LaDiPS, we provide a summary in Table 2 and 3 respectively.

| Hyper Parameter (Ours) | α | β | β_w | N | K | H |
|---------------------------|--------------------------|---------|-----------|----|-----|----|
| Planar Reacher (Ablation) | varies (see section 4.1) | 1.0 | 1.0 | 60 | 350 | 50 |
| Beer Pong Task | 0.001 | 0.5 | 2.5 | 70 | 750 | 50 |
| Table Tennis Task | 0.00001 | 0.2 | 0.2 | 50 | 800 | 50 |

Table 1: Hyperparameters of our algorithm for all environments.

| Hyper Parameter (HiREPS) | ϵ | κ | N |
|---------------------------|------------|----------|----|
| Planar Reacher (Ablation) | 0.5 | 0.99 | 60 |
| Beer Pong Task | 0.3 | 0.9 | 70 |
| Table Tennis Task | 0.3 | 0.9 | 50 |

Table 2: Hyperparameters of HiREPS for all environments.

| Hyper Parameter (LaDiPS) | ϵ | ϵ_{gating} | α | N |
|--------------------------|------------|---------------------|-----------|----|
| Beer Pong Task | 0.1 | 0.05 | $\log 30$ | 70 |
| Table Tennis Task | 0.01 | 0.02 | $\log 30$ | 50 |

Table 3: Hyperparameters of LaDiPS.

C.1 Ablation Studies

For the Planar Reaching task we consider a two dimensional context space. Note that we have not used any Movement Primitive representation for the policy. Instead, the sampled θ from our search distributions directly represent the angles of each joint of the reacher task.

We consider a two dimensional context space x, y , where $x_{min} = 4.5$, $x_{max} = 7$ and $y_{min} = -6$, $y_{max} = 6$.

The reward function is given by:

$$R(\boldsymbol{\theta}, \mathbf{c}) = -\|\boldsymbol{\theta}\|_2^2 - 2 \cdot \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{c}\|_2^2 - l_c(\mathbf{c}) - l_o(\boldsymbol{\theta}),$$

where

$$l_c(\mathbf{c}) = \begin{cases} 10, & c \notin C \\ 0, & \text{else} \end{cases}, \quad l_o(\mathbf{c}) = \begin{cases} 3, & b_{1:10} \in O \\ 0, & \text{else} \end{cases},$$

where l_c are the costs if a context \mathbf{c} is not within the valid context range C and l_o are the costs, if one of the ten links b_i has a collision with the obstacles of the environment. $\mathbf{f}(\boldsymbol{\theta})$ are the forward kinematics of the planar reacher. Note that the angles $\boldsymbol{\theta}$ are always normalized into a range $[-\pi, \pi]$.

C.2 Beer Pong

For the Beer Pong task we consider a two dimensional context space x, y positions of the cup on the table, where $x_{min} = -0.32$, $x_{max} = 0.32$ and $y_{min} = -2.2$, $y_{max} = -1.2$. We use ProMPs as policy parameterizations. We also learn the length of the trajectory, which might lead to invalid, i.e. negative trajectory length. We use a punishment term, if the trajectories are out of a reasonable

range and do not execute this sample on the environment. We do the same for the context samples. Furthermore, we do not execute trajectories, which violate the joint constraints of the robot. Since these samples are not executed on the environment, they are not counted as "taken samples". For the case that those restrictions are not violated, the reward function is given as

$$R(\boldsymbol{\theta}, \mathbf{c}) = \begin{cases} -4 - \min(\|p_{c,top} - p_{b,1:T}\|_2^2) - 0.5\|p_{c,bottom} - p_{b,T}\|_2^2 - 10^{-4}\tau, & \text{if cond. 1} \\ -2 - \min(\|p_{c,top} - p_{b,1:T}\|_2^2) - 0.5\|p_{c,bottom} - p_{b,T}\|_2^2 - 10^{-4}\tau, & \text{if cond. 2} \\ \|p_{c,bottom} - p_{b,T}\|_2^2 + 1.5 \cdot 10^{-4}\tau, & \text{if cond. 3} \end{cases}$$

where $p_{c,top}$ is the position of the top edge of the cup, $p_{c,bottom}$ is the ground position of the cup, $p_{b,t}$ is the position of the ball at time point t and τ is the squared mean torque over all joints during one rollout. The different conditions are:

- cond. 1: The ball is not in the cup and had no table contact
- cond. 2: The ball is not in the cup and had table contact
- cond. 3: The ball is in the cup.

For the case that context samples were sampled out of the range we provide a high negative reward

$$R_c(\boldsymbol{\theta}, \mathbf{c}) = -30 - d_c^2,$$

where d_c^2 is the distance of the current context \mathbf{c} to the valid context region.

For the case that the duration of the trajectory was sampled out of the range we provide a negative reward as

$$R_\theta(\mathbf{c}, \boldsymbol{\theta}) = \begin{cases} -30 - 10 \cdot (l - 0.1)^2, & \text{if } l < 0.1 \\ -30 - 10 \cdot (l - 1.3)^2, & \text{if } l > 1.3 \end{cases},$$

where l is the duration of the Trajectory in seconds. If the joint constraint limits are violated by the ProMP's trajectory, we give a punishment of -30.

C.2.1 On Versatility of our Solutions on the Beer Pong Task

In this section we want to compare the learned solutions of LaDiPS [9] and our method. As LaDiPS uses experience-sharing between components, it outperforms our method in terms of sample efficiency, but does not find as qualitative solutions as our method, which is reflected in the end-rewards (see Fig. 2a), where we can clearly achieve a higher value. In addition to the more qualitative solutions, we are able to learn a mixture model with higher expected entropy, which can be seen in Fig. 4). Our method is able to achieve more qualitative skills with a much higher entropy compared to LaDiPS and HiREPS, indicating that the proposed solutions by our method are more versatile than the others. Please note that we have defined a fine grid of contexts and sampled for each context \mathbf{c} 1000 parameter vectors $\boldsymbol{\theta}$ from the mixture models to calculate the expected entropy. We have repeated this procedure for all 20 different trials/seeds and report the mean expected entropy together with two times the standard error. In addition to the expected mixture entropy we have conducted a qualitative comparison, where we have picked the first model for both, our method and LaDiPS and have let it run for twelve different contexts. For each context we have sampled 100 times from the mixture model and executed the mean of the sampled components on the environment. Note that we only have taken those ball trajectories into account that lead to a successful solution. For LaDiPS all observed solutions for all contexts showed only one mode, in which the ball bounced once on the table and then landed in the cup. For our method the same solution could be seen for six contexts. For three contexts, our method threw the ball directly into the cup, whereas for another three contexts three different ball trajectories could be observed: i) direct throw, ii) bouncing ones on the table ii) bouncing once on the table and once on the wall.

C.3 Table Tennis

For the Table Tennis task we consider a four dimensional context space $[\mathbf{b}^{des}, \mathbf{b}^{inc}] \in \mathbb{R}^4$. We have the desired landing position of outgoing ball $\mathbf{b}^{des} = [x^{des}, y^{des}] \in \mathbb{R}^2$ and the initial positions of the incoming ball $\mathbf{b}^{inc} = [x^{inc}, y^{inc}] \in \mathbb{R}^2$ where $x^{des} \in [-1.2, -0.2]$, $y^{des} \in$

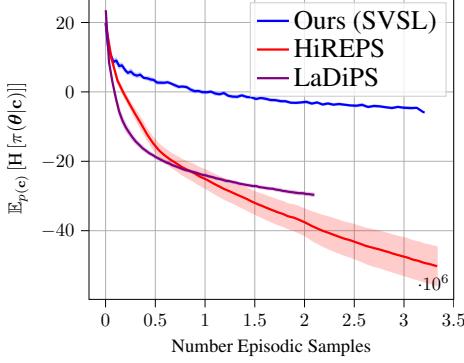


Figure 4: Expected Mixture Entropies.

$[-0.6, 0.6]$, $x^{inc} \in [-1.2, -0.2]$, $y^{inc} \in [-0.65, 0.65]$. We fix the initial ball velocity and z^{inc} position so we can transform the initial ball position to the incoming ball landing position on the table. We use ProMPs as policy parameterizations and learn the length of the trajectory, which might lead to invalid, i. e. negative trajectory length. We use a punishment term, if the length of trajectories are out of a reasonable range and do not execute this sample on the environment. We do the same for the context samples. Furthermore, we do not execute trajectories, which violate the joint constraints of the robot and return a punishment. Since these samples are not executed on the environment, they are not counted as "taken samples". For the case that those restrictions are not violated, the reward function is given as

$$R(\theta, c) = \begin{cases} 0.2 \cdot (1 - \tanh(\min_t (\|\mathbf{b}_t - \mathbf{r}_t\|_2^2))), & \text{if cond. 1} \\ 2 \cdot (1 - \tanh(\min_t (\|\mathbf{b}_t - \mathbf{r}_t\|_2^2))) \\ + (1 - \tanh(\min_t (\|\mathbf{b}^{des} - \mathbf{b}_{t,xy}\|_2^2))), & \text{if cond. 2} \\ 2 \cdot (1 - \tanh(\min_t (\|\mathbf{b}_t - \mathbf{r}_t\|_2^2))) \\ + 4 \cdot (1 - \tanh(\|\mathbf{b}^{des} - \mathbf{b}^{land}\|_2^2)) + 1_{\{b_x^{land} < 0\}}, & \text{if cond. 3} \end{cases}$$

where $\mathbf{b}_t \in \mathbb{R}^3$ is the position of the ball at time point t , $\mathbf{b}_{t,xy} \in \mathbb{R}^2$ is the x, y position of the ball at time point t , $\mathbf{r}_t \in \mathbb{R}^3$ is the position of the racket at time point t , $\mathbf{b}^{land} \in \mathbb{R}^2$ is the real x, y landing position point. The different conditions are:

- cond. 1: The ball had no racket contact.
- cond. 2: The ball had racket contact but then no table contact.
- cond. 3: The ball had racket contact and then table contact.

For the case that context samples were sampled out of the range we provide a high negative reward

$$R_c(\theta, c) = -20 \cdot d_c^2,$$

where d_c^2 is the distance of the current context c to the valid context region.

For the case that the duration of the trajectory was sampled out of the range we provide a negative reward as

$$R_\theta(c, \theta) = \begin{cases} -3\|l - 0.1\|, & \text{if } l < 0.1 \\ -3\|l - 5\|, & \text{if } l > 5 \end{cases},$$

where l is the duration of the Trajectory in seconds.

For the case that the trajectory was sampled out of the joint constraints we provide a negative reward as

$$R_c(\theta, c) = -1 \cdot \frac{1}{T} \sum_i^T d_{qpos,t},$$

where $d_{qpos,t}$ is the distance of the planned joint positions to the valid joint position region.

C.3.1 On Versatility of our Solutions on the Table Tennis Task

We compare the versatility of learned solutions of LaDiPS [9], HiREPS [8] and ours. We plot the expected mixture entropies of each algorithm in Fig. 5a. For each of the 1600 uniformly context samples we have sampled 1000 parameter vectors θ from the mixture model to calculate the mixture entropy. We repeat this for all models resulting from 20 seeds/trials and plot the average of the entropies in Fig. 5a. Please note that the fluctuations of the entropy curve in Fig. 5a of our algorithm are due to adding components during training.

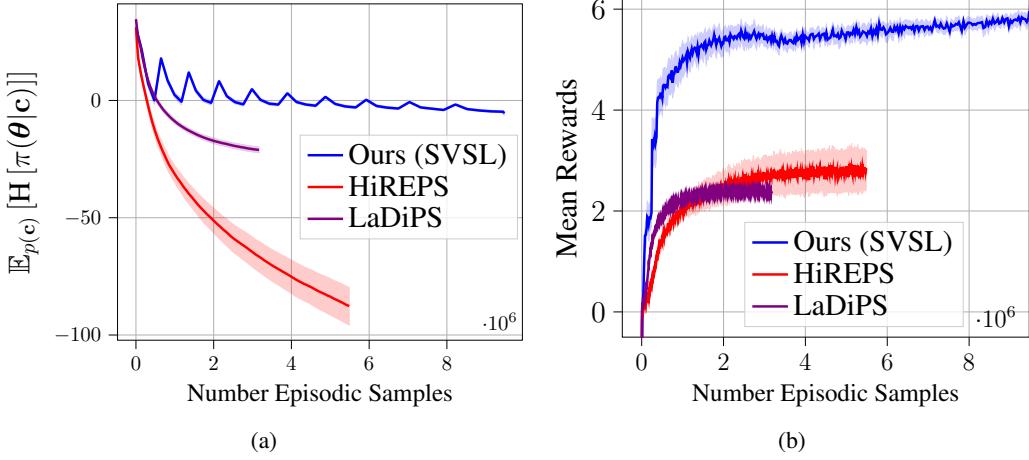


Figure 5: **Expected Mixture Entropies** (left) and **performance on the table tennis task** (right). Note that the performance graph is the same as in Fig. 2b, but with the full graph of our method to train 70 components. The graph in Fig. 2b was clipped in the x-axis to maintain overview.

We are able to achieve the highest entropy while clearly outperforming both methods in terms of the achieved task reward (see Fig. 5b), which indicates that we are able to learn more qualitative skills.

C.4 A Comparison Between Episodic- and Step-Based Policy Search

The step-based and the episodic policy search setting have their own benefits and limitations. For example, while in the step-based case we usually explore the action space using uncorrelated noise, the episodic setting uses correlated distributions directly in the parameter space of the controller [3]. Yet, step-based approaches can use the data more efficiently as they decompose the episodes into single time steps. Hence, it is interesting to compare these approaches.

We compare the performance of our algorithm against PPO, a well-known Deep Reinforcement Learning method [34]. We want to analyze when episodic exploration is beneficial over step-based exploration and vice-versa for different sparsity levels of the reward. For this purpose, we use the Reacher task from OpenAI gym [36]. In this task, the agent has to reach a goal position with its tip. In order to consider a more difficult task, we extend the original reacher set-up and show the considered environment in Figure 6). First, we increase the number of links from two to five and we fix the initial state position to be initialized on the horizontal to the left (see Fig. 6), instead of initializing on the horizontal to the right side as it is in the original version.

Furthermore, we fix the initial joint velocity to be zero. For the original 2-Link case, the goal reaching space is the circle around the agent's base, whereas in our case we consider the right half-circle around the agent's base, where the radius is the agent's total length. By initializing the agent on the

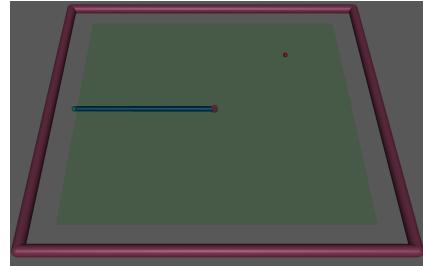


Figure 6: 5Link Reacher Set-up. The reacher's tip needs to reach the red goal position. Adapted from [36].

left side and only considering goal positions on the right side, we make the task harder. We compare our method to the popular PPO algorithm [34] and use the implementation from [37].

In order to examine the different policy search strategies we consider three different markovian reward functions in the same set-up. The first is given by

$$R_t(\mathbf{s}, \mathbf{a}) = -\|\mathbf{t}(\mathbf{s}) - \mathbf{g}\|_2 - \sum_i a_i^2, \quad (15)$$

where \mathbf{s} is the current state, \mathbf{a} is the current action, $\mathbf{t}(\mathbf{s})$ is the tip's current position and \mathbf{g} is the goal position. Except for \mathbf{g} , which is re-sampled at each episode, the values are given at the current time-step t . One episode lasts for $T = 200$ time-steps.

This reward function is also given by [36] and encodes task information, i.e. tip distance to the goal, in each time-step. In the following we only change the reward function.

The second reward function is given as

$$R_t(\mathbf{s}, \mathbf{a}) = \begin{cases} -\sum_i a_i^2, & \text{if } t < 190 \\ -\|\mathbf{t}(\mathbf{s}) - \mathbf{g}\|_2 - \sum_i a_i^2, & \text{if } t \geq 190 \end{cases}. \quad (16)$$

Characteristic for this reward function is that it will only return task information, i.e. tip-distance to goal, in the last 10 steps of the episode.

The third reward function is given as

$$R_t(\mathbf{s}, \mathbf{a}) = \begin{cases} -\sum_i a_i^2 - \sum_i v_i^2, & \text{if } t < 198 \\ -500 \cdot \|\mathbf{t}(\mathbf{s}) - \mathbf{g}\|_2 - \sum_i a_i^2 - 1000 \cdot \sum_i v_i^2, & \text{if } t \geq 198 \end{cases}, \quad (17)$$

where \mathbf{v} is the joint velocity at time-step t . This reward function is very similar to the reward function from Eq. (16). However, it only returns task information in the last two steps of the episode, which makes the task very difficult.

The reward functions in Equations (16, 17) can be motivated from the optimal control point of view. There, usually controllers with minimum energy consumption are sought giving rise to punishments to taken control actions in the first steps as it is done by the proposed reward functions in Equations (16, 17).

For PPO we use the same observation space as described in [36], but for the Objectives (16, 17) we augment this observation space with the current time-step t .

The context \mathbf{c} for our method is given through the two-dimensional goal position vector \mathbf{g} .

We report the average performance and two times the standard error over 20 seeds/trials for each experiment in Fig. 7). The performances for the reward function in Equation (15) are shown in Fig. 7a). Clearly, PPO outperforms our algorithm, indicating that task information in each time-step helps exploring in the raw action space to find solutions. The performances for the reward function in Equation (16) is shown in Fig. 7b). While PPO also performs well, we can observe that our algorithm can make use of the exploration in the parameter space and outperform PPO. This effect can be seen much more clearly for the reward function in Equation (17). Fig. 7c) shows that PPO has clear problems to solve this task. For the settings given by the reward functions in Equations (16, 17) exploration in the parameter space as done by episodic policy search methods lead to better performance, whereas exploration in the raw action space leads to converging to a local optimum.

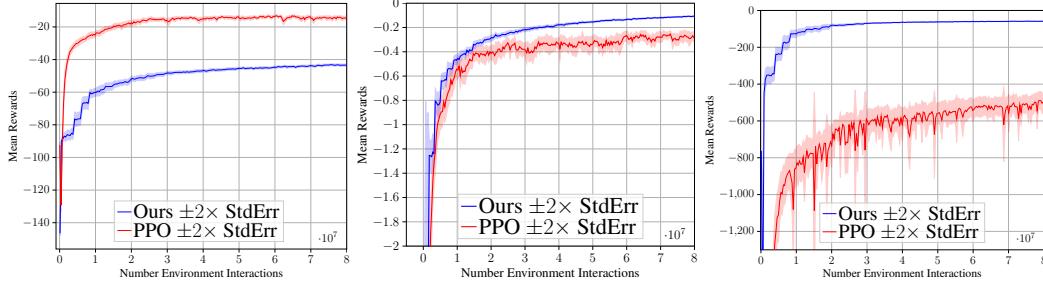
C.4.1 Hyperparameters

The hyperparameters for our algorithm are given in the following Table. Please note that we used punishments,

| Hyper Parameter (Ours) | α | β | β_w | N | K | H |
|------------------------|-----------|---------|-----------|----|-----|----|
| Reacher Reward (15) | $1e^{-6}$ | 15 | 15 | 40 | 250 | 50 |
| Reacher Reward (16) | 0.01 | 1 | 2.5 | 40 | 250 | 50 |
| Reacher Reward (17) | 1 | 25 | 25 | 40 | 800 | 50 |

Table 4: Hyperparameters of our algorithm for the reacher environment.

if contexts were sampled out of the context range (which is the right half circle around the base with radius 0.5), we provided a quadratic punishment to the distance of the sample to the valid context



(a) Performances for Reward (15) (b) Performance for Reward (16) (c) Performance for Reward (17)

Figure 7: A Comparison of Episodic Policy Search to Step-based Policy Search. We compare our episodic policy search method to the popular step-based approach PPO. We analyze the performance on three different reward types. For (a) we consider the reward in Equation (15). Here, task-information are returned in each time-step. For(b) we consider the reward in Equation (16). Here, task-information are returned in the last 10 steps of the episode. For (c) we consider the reward in Equation(17), where only at the last 2 time-steps reward information are returned.

bound. We multiplied this distance with a factor of 1000 for the case of reward function (15, 16) and with a factor of 50000 for the case of reward function (17).

We normalized the observations and rewards for PPO and used the same hyperparameters for all three experiments, which we summarized in the following table

| Hyp. Params (PPO) | α | N | GAE λ | γ | K | ϵ | network structure |
|---------------------|-----------|-------|---------------|----------|----|------------|-------------------|
| Reacher Reward (15) | $1e^{-4}$ | 16384 | 0.95 | 0.99 | 64 | 0.2 | [64, 64] |
| Reacher Reward (16) | $1e^{-4}$ | 16384 | 0.95 | 0.99 | 64 | 0.2 | [64, 64] |
| Reacher Reward (17) | $1e^{-4}$ | 16384 | 0.95 | 0.99 | 64 | 0.2 | [64, 64] |

Table 5: Hyperparameters of PPO for the reacher environment. Note that α is the learning rate, N is the number of rollouts, γ is the discount factor, K is the batch size, ϵ is the clip value for the importance ratio. The layer structure is for both: the policy network as well as the value function network.