

# Importance Weighted Policy Learning and Adaptation

**Alexandre Galashov**  
agalashov@google.com

**Jakub Sygnowski**  
sygi@google.com

**Guillaume Desjardins**  
gdesjardins@google.com

**Jan Humplik**  
jhumplik@google.com

**Leonard Hasencler**  
leonardh@google.com

**Rae Jeong**  
raejeong@google.com

**Yee Whye Teh**  
ywteh@google.com

**Nicolas Heess**  
heess@google.com

**Abstract:** The ability to exploit prior experience to solve novel problems rapidly is a hallmark of biological learning systems and of great practical importance for artificial ones. In the meta reinforcement learning literature much recent work has focused on the problem of optimizing the learning process itself. In this paper we study a complementary approach which is conceptually simple, general, modular and built on top of recent improvements in off-policy learning. The framework is inspired by ideas from the probabilistic inference literature and combines robust off-policy learning with a behavior prior, or default behavior that constrains the space of solutions and serves as a bias for exploration; as well as a representation for the value function, both of which are easily learned from a number of training tasks in a multi-task scenario. Our approach achieves competitive adaptation performance on hold-out tasks compared to meta reinforcement learning baselines and can scale to complex sparse-reward scenarios.

**Keywords:** Meta-learning, Reinforcement Learning, Off-policy Reinforcement Learning

## 1 Introduction

Current reinforcement learning (RL) algorithms have achieved impressive results across a broad range of games and continuous control platforms. While effective, such algorithms all too often require millions of environment interactions to learn, requiring access to large compute as well as simulators or large amounts of demonstrations. This stands in stark contrast to the efficiency of biological learning systems [1], as well as the need for data-efficiency in real world systems, e.g. in robotics where environment interactions can be expensive and risky. In recent years, data efficient RL has thus become a key area of research and stands as one of the bottlenecks for RL to be applied in the real world [2]. Research in the area is multi-faceted and encompasses multiple overlapping directions. Recent developments in off-policy and model-based RL have dramatically improved stability and data-efficiency of RL algorithms which learn *tabula rasa* [e.g. 3, 4]. A rapidly growing body of literature, under broad headings such as *transfer learning*, *meta learning*, or *hierarchical RL*, aims to speed up learning by reusing knowledge acquired in previous instances of similar learning problems. *Transfer learning* typically follows a two step procedure: a system is first *pre-trained* on one or multiple training tasks, then a second step *adapts* the system on a downstream task. While transfer learning approaches allow significant flexibility in system design, the two-step process is often criticised for being sub-optimal. In contrast, *meta-learning* incorporates adaptation into the learning process itself. In gradient-based approaches, systems are explicitly trained such that they perform well on a downstream task after a few gradient descent steps [5]. Alternatively, in encoder-based approaches a mapping is learned from a data collected in a downstream task to a task representation [e.g 6, 7, 8, 9, 10, 11]. Because meta-learning approaches optimize the

adaptation process directly, they are expected to adapt faster to downstream tasks than transfer learning approaches. But performing this optimization can be algorithmically or computationally challenging, making it difficult to scale to complex and broader task distributions, especially since many approaches simultaneously solve not just the meta-learning but also a challenging multi-task learning problem.

Given the limitations of meta-learning, a number of recent works have raised the question whether transfer learning methods, potentially combined with data-efficient off-policy algorithms, are sufficient to achieve effective generalization as well as rapid adaptation to new tasks. For example, in the context of supervised meta learning, Raghu et al. [12] showed that learning good features and finetuning during adaptation led to results competitive with MAML. In reinforcement learning, Fakoor et al. [13] showed that direct application of TD3 [14] to maximize a multi-task objective along with a recurrent context and smart reuse of training data was sufficient to match performance of SOTA meta-learning methods on current benchmarks.

In this paper, we take a similar perspective and try to understand the extent to which fast adaptation can be achieved using a simple transfer framework, with the generality of gradient-based adaptation. Central to our approach is the behaviour prior recovered by multi-task KL-regularized objectives [15, 16]. We improve transfer performance by leveraging this prior in two important ways: first, as a regularizer which helps with exploration and restricts the space of solutions that need to be considered, and second as a proposal distribution for importance weighting, where the weights are learnt and given by the exponentiated Q-function. This avoids the need to learn an explicit parametric policy for the transfer task, instead the policy is obtained directly by tilting the prior with the learned, exponentiated action-value function. To further speed-up adaptation and avoid learning this Q-function de-novo, we make use of a particular parameterization of the action-value functions obtained during multi-task training: the Q-values are parameterized to be linear in some shared underlying feature space. Intuitively, this shared feature representation captures the commonalities in terms of both reward and transition dynamics. In practice, we found this value function representation together with the behaviour prior to generalize well to transfer tasks, drastically speeding-up the adaptation process. We show that across continuous control environments ranging from standard meta-RL benchmarks to more challenging environments with higher dimensional action spaces and sparse rewards, our method can match or outperform recent meta-learning approaches, echoing recent observations in [13].

Our paper is structured as follows. Section 2 provides the necessary background material and characterizes the multi-task reinforcement learning problem. Our method, based on importance weighting, is presented in Section 3 while Section 4 shows how our training algorithm can be adapted to improve transfer learning performance. Relevant work is discussed in Section 5 with experimental results presented in Section 6.

## 2 Background

We consider a multi-task reinforcement learning setup, where we denote a probability distribution over tasks as  $\mathcal{P}(\mathcal{T})$ . Each task  $\mathcal{T} \sim \mathcal{P}$  is a Markov Decision Process (MDP), i.e. a tuple  $\langle p_{\mathcal{T}}(s'|s, a), p_{\mathcal{T}}(s_0), r_{\mathcal{T}}(s, a), \mathcal{A}, \mathcal{S} \rangle$  described by (respectively) the transition probability, initial state distribution, reward function, action and state spaces, where  $\mathcal{A}$  and  $\mathcal{S}$  are identical across tasks. Furthermore, we assume that we are given finite i.i.d. samples of tasks split into training,  $\mathcal{T}_{train} = (\mathcal{T}_1, \dots, \mathcal{T}_n)$ , and test,  $\mathcal{T}_{test} = (\mathcal{T}_{n+1}, \dots, \mathcal{T}_{n+m})$  sets. For each task, denoted by  $i$ , we denote the task-specific policy as  $\pi_i$ , whereas  $\pi_0$  is a shared *behaviour prior* which regularizes the  $\pi_i$ 's. On top of that, we denote as  $p_i(s'|s, a), p_i(s_0), r_i(s, a)$ , the transition probability, initial state distribution and reward function for the task  $i$ .

The starting point in this paper is DISTRAL [15] which aims to optimize the following multi-task objective on the training set:

$$\mathcal{J}(\pi_0, \pi_1, \dots, \pi_n) = \sum_{i=1}^n \mathbb{E}_{\tau \sim \pi_i(\tau)} \left[ \sum_{t \geq 1} \gamma^t r_i(a_t, s_t) - \gamma^t \alpha \log \frac{\pi_i(a_t | s_t)}{\pi_0(a_t | s_t)} \right], \quad (1)$$

where  $\alpha$  is an inverse temperature parameter and  $\tau \sim \pi_i(\tau)$  denotes the sampling a trajectory from the task  $i$  using the policy  $\pi_i$ . The objective in (1) is optimized with respect to all  $\pi_i$  and

$\pi_0$  jointly. In particular, for each task  $i$  and for a fixed behaviour prior  $\pi_0$ , the optimization of the objective  $\mathcal{J}$  is equivalent to solving a regularized RL problem with augmented reward  $\tilde{r}_i(a_t, s_t; \pi_i) = r_i(a_t, s_t) - \alpha \log \frac{\pi_i(a_t | s_t)}{\pi_0(a_t | s_t)}$ . As for learning the behaviour prior  $\pi_0$ , optimizing (1) with respect to  $\pi_0$  amounts to minimizing the sum of KL divergences between the task-specific policies  $\pi_i$  and the prior:

$$\pi_0^*(a_t | s_t) = \arg \min_{\pi_0} \sum_i \text{KL}[\pi_i(a_t | s_t) || \pi_0(a_t | s_t)]. \quad (2)$$

The behaviour prior's role is to model behavior that is shared across the tasks. As shown in [16], a prior trained according to (1) with computational restrictions such as partial access to observations only (information asymmetry) can capture useful default behaviours (such as walking in some walking-related task). The prior regularizes the task-specific solutions and can transfer useful behavior between tasks, which can speed up learning.

Let  $\pi_i$  be the current policy for the task  $i$ . For a fixed behaviour prior  $\pi_0$ , we define the associated soft Q-function as

$$Q_i^{\pi_i}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim p_i(s' | s, a)} [\mathbb{E}_{a' \sim \pi_i(a' | s')} [Q_i^{\pi_i}(s', a')] - \alpha \text{KL}[\pi_i(\cdot | s') || \pi_0(\cdot | s')]] . \quad (3)$$

This function was considered in [17]. Note that if  $\pi_0$  is a uniform distribution, the definition in (3) is equivalent to the soft Q-function considered, for instance, in [4, 18]. Furthermore, the policy, which is a result of computing 1-step soft-greedy policy, defined as:

$$q(a | s) = \frac{\pi_0(a | s) \exp(Q_i^{\pi_i}(s, a) / \alpha)}{\int \pi_0(a | s) \exp(Q_i^{\pi_i}(s, a) / \alpha) da}, \quad (4)$$

will have higher soft Q-value on the task  $i$ , i.e.  $Q_i^q(s, a) \geq Q_i^{\pi_i}(s, a), \forall a, s$  (see [4]). Therefore, (4) gives us a principled way to perform policy improvement. A similar policy improvement step is used, for instance, in MPO [3] and Soft Actor Critic (SAC) [4]. In both cases, the authors optimize a parametric representation to fit the distribution in (4).

But instead of fitting a parametric policy, one can directly act according to the improved policy in (4). This can be potentially more efficient, since it avoids an additional step of learning policy with function approximation. However, sampling exactly from the distribution in (4) can only be done in a few special cases. Below, we propose a method which uses importance sampling to draw samples from a distribution, which approximates the distribution in (4).

### 3 Importance weighted policy learning

For each task  $i$  and for a fixed behaviour prior  $\pi_0$ , we consider the following. Firstly, we sample a set of actions from the behaviour prior:

$$\{a^{(k)}\}_{k=1}^K \sim_{iid} \pi_0(a | s) \quad (5)$$

We denote as  $\mathcal{A}_K = \{a^{(k)}\}_{k=1}^K$ , the set of sampled actions and as  $\Pi_K(s)$  the set of discrete action distributions defined on  $\mathcal{A}_K$  for a state  $s$ . For simplicity of notation, we will drop  $s$  from  $\Pi_K(s)$  and denote it as  $\Pi_K$ . We denote as  $Q_i$  the soft action value function for some policy  $\pi_i$  and reward function  $r_i$ . Then, we construct the following action distribution over  $\mathcal{A}_K$  for each state  $s$ :

$$\begin{aligned} \hat{q}_k &= \hat{q}(a = a^{(k)} | s) = \exp\left(\frac{Q_i(s, a^{(k)}) - Z(s)}{\alpha}\right) \text{ for } k = 1, \dots, K, \\ \hat{a} &\sim \hat{q}(a | s) = \text{Cat}(q_1, \dots, q_K) \end{aligned} \quad (6)$$

with a normalizing constant  $Z(s)$ :

$$Z(s) = \alpha \log \sum_{j=1}^K \exp\left(\frac{Q_i(s, a^{(j)})}{\alpha}\right)$$

Then, the resulting policy  $\hat{q} \in \Pi_K$  is a discrete approximation for the improved policy of the form  $q$  from (4). Note that the procedure 6 corresponds to a soft-max distribution over actions with respect to the exponent of the soft Q-function.

In the limit of  $K \rightarrow \infty$ , the procedure 5-6 is guaranteed to sample from the policy  $q$  from (4). The above sampling scheme gives rise to the *Importance Weighted Policy Learning* (IWPL) algorithm, which combines non-parametric policy evaluation and improvements steps, described below.

**Non-parametric policy evaluation** Let  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  be a function and  $\pi$  is a policy defined on  $\mathcal{A}$ . We define the soft Bellman backup operator:

$$\begin{aligned}\mathcal{T}^\pi Q(s_t, a_t) = \\ r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}} [\mathbb{E}_{a_{t+1} \sim \pi(\cdot | s_{t+1})} [Q(s_{t+1}, a_{t+1})] - \alpha K L[\pi(\cdot | s_{t+1}) || \pi_0(\cdot | s_{t+1})]]\end{aligned}$$

It is easy to see (as in [4]) that the Bellman iteration  $Q^{l+1} = \mathcal{T}^\pi Q^l, l \rightarrow \infty$  converges to the soft value function 3 for  $\pi$ . Then, for the policy  $q$  defined by eq.4 we consider an estimator for the Bellman operator induced by the importance weighting procedure 5-6 (with a new sampled set of actions  $\{a^{(k)}\}_{k=1}^K$ ):

$$\mathcal{T}_K^q Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}} \left[ \sum_{k=1}^K \hat{q}(a_k | s_{t+1}) \left( Q(s_{t+1}, a_k) - \alpha \log \frac{\hat{q}(a^{(k)} | s_{t+1})}{\pi_0(a^{(k)} | s_{t+1})} \right) \right]. \quad (7)$$

In the limit, this procedure would converge to the soft Q-function for  $q$ :  $Q^{l+1} = \mathcal{T}_K^q Q^l, l \rightarrow \infty, K \rightarrow \infty$ .

**Non-parametric policy improvement** Given the current proposal  $\pi_0$ , some old policy  $q^{old}$ , corresponding soft Q-function  $Q^{q^{old}}$ , we can obtain new policy  $q^{new}$  via (4). In this case, similar to [18] (Appendix B.2), we have:

$$Q^{q^{new}}(s, a) \geq Q^{q^{old}}(s, a), \forall s, a,$$

where  $Q^{q^{new}}$  is the soft Q-function corresponding to the  $q^{new}$ . To approximate the  $q^{new}$ , we resample new actions  $\{a^{(k)}\}_{k=1}^K$  via procedure 5 and apply procedure 6 to the  $Q^{q^{old}}$  and obtain the categorical distribution with following probabilities:

$$\hat{q}_k^{new} = \hat{q}_k^{new}(a = a^{(k)} | s) \sim \exp \left( \frac{Q^{q^{old}}(s; a^{(k)})}{\alpha} \right)$$

This describes a policy improvement procedure based on importance sampling.

**Behaviour prior (proposal) improvement** Given current policy  $q(a|s)$  of a form 4, corresponding approximation  $\hat{q}$  from (6), a new behaviour prior  $\hat{\pi}_0$  is obtained by maximizing the likelihood of obtaining samples from  $\hat{q}(a|s)$ :

$$\hat{\pi}_0(\cdot | s) = \arg \min_{\pi_0} \sum_{k=1}^K \hat{q}_k \log \pi_0(a_k | s)$$

**Temperature calibration** In the current formulation, IWPL requires us to choose the inverse temperature parameter in 1 and in 6. For varying reward scales, it could result in an unstable behaviour of the procedure 6. Some RL algorithms, such as REPS [19], MPO [3] therefore replace similar (soft) regularization terms with hard limits on KL or entropy. Here, we consider a hard-constraint version of objective (1):

$$\begin{aligned}\sum_i \mathbb{E}_{\tau \sim \pi_i(\tau)} \left[ \sum_{t \geq 1} \gamma^t r_i(a_t, s_t) \right] \\ \sum_i \mathbb{E}_{s \sim \pi_i(s)} \text{KL}[\pi_i(\cdot | s) || \pi_0(\cdot | s)] < \epsilon\end{aligned} \quad (8)$$

The parameter  $\epsilon$  defines the maximum average deviation of all the policies  $\pi_i$  from the behaviour prior  $\pi_0$ . Given  $\epsilon$ , we can adjust the inverse temperature  $\alpha$  to match this constraint. In many cases  $\epsilon$  is easier to choose than the inverse temperature  $\alpha$  since it does not, for instance, depend on the scale of the reward. The associated temperature parameter  $\alpha$  can be optimized by considering the Lagrangian for the objective 8, similar to REPS [19] and MPO [3].

---

**Algorithm 1** Distributed Importance Weighted Policy Learning (IWPL)

---

**Input:**

Behaviour prior  $\pi_0(a|s, \phi)$ , initial parameters  $\phi_0$   
 Q-function  $Q_{\theta_i}$ , initial parameters  $\theta_i^0$  for each task  $i$   
 Target networks with a separate set of parameters  $\theta', \phi'$   
 Target networks update period  $T$   
 Learning rates  $\beta_Q, \beta_{\pi_0}, \beta_\alpha$   
 Replay buffer  $\mathcal{B}$  containing data  $\mathcal{B}_i$  for each task  $i$   
 Training tasks indexes  $\mathcal{I} = \{1, \dots, n\}$   
 Define  $\theta = (\theta_1, \dots, \theta_n), \theta' = (\theta'_1, \dots, \theta'_n)$

**Steps:**
**Actor policy:**

**while** Not converged **do**

    Receive parameters from the learner

    Sample uniformly a training task  $i$  from  $\mathcal{I}$

    Sample full-episode trajectory  $\tau = (s_0, a_0, r_0, \dots, s_T, a_T, r_T) \sim \hat{q}_i(\tau)$ , using equations. (5,6)

$$\mathcal{B}_i = \mathcal{B}_i \cup \tau$$

**end while**

**Learner policy:**

**while** Learning **do**

    Sample uniformly (with replacement) a batch of tasks  $\mathcal{I}_b$  from  $\mathcal{I}$

**for** each task  $i$  from  $\mathcal{I}_b$  **do**

        Sample partial trajectory from replay buffer  $\mathcal{B}_i : \tau_{t:t+M} = (s_t, a_t, r_t, \dots, r_{t+M})$  for task  $i$

        Sample  $K$  actions  $(a_1^t, \dots, a_K^t)$  from  $\pi_0(a|s_t, \phi')$ , for each state  $s_t$

        Calculate the  $Q_{\theta'_i}(s_t, a_k^t)$ ,  $\forall t, k$

        Construct categorical distribution  $\hat{q}'_i$  as in (6) using  $Q_{\theta'}(s_t, a_k^t)$

        % Perform gradient update on the parameters

$$\theta_i \leftarrow \theta_i + \beta_Q \nabla_{\theta_i} \mathcal{J}_Q(\theta)$$

$$\phi \leftarrow \phi + \beta_{\pi_0} \nabla_\phi \mathcal{J}_{\pi_0}(\phi)$$

        Every  $T$  gradient steps, update target networks parameters  $\theta' \leftarrow \theta, \phi' \leftarrow \phi$ .

**end for**

**end while**

---

**Algorithm** The concrete algorithm is a combination of the steps above with parametric function approximation of the necessary quantities. We consider  $\pi_0(a|s, \phi)$  the approximation for the behaviour prior  $\pi_0$  and  $Q_{\theta_i(s,a)}$  an approximation for the soft value function for the task  $i$ . We denote as  $\phi'$  and as  $\theta'_i$  the other set of parameters which correspond to the target networks (see Mnih et al. [20]) - the networks which are kept fixed for some number of iterations. We denote as  $\hat{q}'_i$  the discrete policy coming from 6 associated with  $Q_{\theta'_i}(s, a)$  and  $\pi_0(a|s, \phi')$ . Then,  $Q_{\theta_i}(s, a)$  can be trained by minimizing the Bellman residual:

$$\mathcal{J}_Q(\theta) = \sum_i \mathbb{E}_{s,a \sim p_i(s,a)} \left[ \frac{1}{2} (Q_{\theta_i}(s, a) - \hat{Q}_i(s, a))^2 \right], \quad (9)$$

where  $\theta = (\theta_1, \dots, \theta_n)$  and:

$$\hat{Q}_i(s, a) = r_i(s, a) + \gamma \sum_{k=1}^K \hat{q}'_i(a_i^{(k)} | s_{t+1}) \left( Q_{\theta'_i}(s_{t+1}, a_i^{(k)}) - \alpha \log \frac{\hat{q}'_i(a_i^{(k)} | s_{t+1})}{\pi_0(a_i^{(k)} | s_{t+1}, \phi')} \right) \quad (10)$$

The behaviour prior  $\pi_0(a|s, \phi)$  is learned by minimizing:

$$\mathcal{J}_{\pi_0}(\phi) = - \sum_i \mathbb{E}_{s \sim p_i(s)} \left[ \sum_{k=1}^K \hat{q}'_i(a_k | s) \log \pi_0(a_k | s, \phi) \right] \quad (11)$$

The full algorithm is presented in Algorithm 1.

## 4 Importance weighted policy adaptation for transfer learning

Given pretrained action-value functions  $\{Q_i^*\}_{i=1}^n$  and a behaviour prior  $\pi_0^*$  from optimization of the objective 8 on the training set, we show how to leverage it to quickly solve tasks from the test set. We call this process adaptation. Below, we describe how adaptation is facilitate by two components of our method, behaviour and value transfer.

**Behaviour Transfer.** Given a pre-trained behaviour prior  $\pi_0^*$ , we can learn the solution to a new task by learning a new value function and sampling from the implicit policy defined by 6. This can be achieved by executing the procedure in Section 3 without the prior improvement step. Because the policy essentially is initialized from the behaviour prior, the latter constrains possible solutions and leads to sensible exploration. In order to obtain new optimal policy, we need to learn new optimal soft Q function, which can require considerable amount of samples when Q is naively parameterized by a neural network. Below, we propose a way to leverage the Q-functions learned for tasks in the training set to speed up transfer in terms of number of interactions with the environment.

**Value Transfer.** In order to acquire knowledge about the value function that can be leveraged for transfer we choose to represent the task specific value  $Q_i$  as a linear function of task-specific parameters  $w$  and shared features  $\psi$ :

$$Q_i(s, a; \Phi_i) = \psi(s, a; \theta)^T w_i, \quad (12)$$

where  $\psi_\theta : \mathbb{R}^S \times \mathbb{R}^A \rightarrow \mathbb{R}^d$  is a function mapping states and actions to a feature vector (with parameters  $\theta$  shared across tasks),  $w_i \in \mathbb{R}^d$  is a task-specific vector used to identify task-specific Q-values, and  $\Phi_i = \{\theta, w_i\}$ . During the adaptation phase, we initialize  $Q(s, a)$  as  $\psi(s, a; \theta^*)^\top \tilde{w}$ , with  $\tilde{w} \sim \mathcal{N}(0, I_d/d)$ , and adapt  $\tilde{w}$  using TD(0) learning. Furthermore, for some more challenging tasks, we replace (at training time) the task-specific vector  $w_i$  by a non-linear embedding of a structured goal descriptor  $g_i$  which is available during training but not during adaptation, i.e.  $Q_i(s, a, g_i; \Phi_i) = \psi(s, a; \theta)^\top f(g_i; \theta)$ , where  $f(g_i; \theta)$  is a learned embedding of goal  $g_i$  with parameters  $\theta$  shared across training tasks. At test time, we initialize the critic as before:  $\psi(s, a; \theta^*)^\top \tilde{w}$ . Since some RL problems can still be challenging multi-task learning problems, this "asymmetry" between learning and testing allows us to simplify the solution of the multi-task problem without affecting the applicability of the learned representation, in contrast to most of the meta-learning approaches which require that training and adaptation phase be matched. Then, our proposed method exploits both, behaviour prior and shared value features to derive an efficient off-policy transfer learning algorithm. Note that this approach does not require to have a finite or/and discrete set of tasks and could work also in the continuously parameterised task distributions, since we essentially allow the task-specific Q-function to depend on the task conditioning.

**Algorithm** Given the new task  $j$ , we will learn associated  $w$  to construct Q-function of the form 12. Let  $\pi_0(a|s, \phi)$  be a pretrained behaviour prior,  $\psi(s, a; \theta)$  be pretrained features for the Q-functions on the training set. We use similar notation as in Section 3, by denoting as  $w'$ , the target network parameters and as  $\hat{q}, \hat{q}'$  associated categorical distributions of form 6. Let  $Q_w(s, a; \theta)$  be the function approximator of the form 12 for the new task  $j$ . Then, the adaptation on the task  $j$  reduces to learning the Q-function by minimizing TD(0) Bellman residual:

$$\mathcal{J}(w; \theta) = \mathbb{E}_{s, a \sim p(s, a)} \left[ \frac{1}{2} (Q_w(s, a; \theta) - \hat{Q}_{w'}(s, a; \theta'))^2 \right], \quad (13)$$

where

$$\hat{Q}_{w'}(s, a; \theta') = r_j(s, a) + \gamma \sum_{k=1}^K \hat{q}'_j(a_j^{(k)} | s_{t+1}) \left( Q_{w'}(s_{t+1}, a_j^{(k)}; \theta') - \alpha \log \frac{\hat{q}'_j(a_j^{(k)} | s_{t+1})}{\pi_0(a_j^{(k)} | s_{t+1}, \phi)} \right). \quad (14)$$

Note that in addition to learning new  $w$ , it is also possible to finetune pre-trained features  $\psi(s, a; \theta)$ . It may be required if test tasks are too different from the training tasks. This scenario is discussed in *Generalization* part of Section 6. We call the resulted algorithm *Importance Weighted Policy Adaptation* (IWPA) which is described in Algorithm 2.

---

**Algorithm 2** Importance Weighted Policy Adaptation (IWPA)

---

**Input:**

Behaviour prior  $\pi_0(a|s; \phi)$  pre-trained on the training set.  
Shared features  $\psi(s, a; \theta)$  representing optimal training soft Q-functions 12

$\mathcal{I} = \{n + 1, \dots, n + m\}$  - indexes for the test set tasks.

$N$ : Number of adaptation episodes

$M$ : Number of gradient updates

Target networks parameters  $w', \theta'$

Target networks update period  $T$

$\beta_w, \beta_\theta$  - Learning rates

**for** Each test task  $j$  from  $\mathcal{I}$  **do**

    Initialize task specific critic parameters  $w \sim \mathcal{N}(0, I_d/d)$

    Define action-value function  $Q_w(s, a; \theta) = \psi(s, a; \theta)^T w$

    Denote as  $\hat{q}_w$  associated to  $\pi_0$  and  $Q_w$  categorical distribution of form 6

**for**  $n = 1 : N$  **do**

        Sample full-episode trajectory  $\tau = (s_0, a_0, r_0, \dots, s_T, a_T, r_T) \sim \hat{q}_w(\tau)$ , using eqs. (5,6)

**for**  $m = 1 : M$  **do**

            % Perform gradient update on the parameters for adaptation

$w \leftarrow w + \beta_w \nabla_w \mathcal{J}(w, \theta)$

            (Optionally) Finetune features,  $\theta \leftarrow \theta + \beta_\theta \nabla_\theta \mathcal{J}(w, \theta)$

            Every  $T$  gradient steps, update target networks parameters  $w' \leftarrow w, \theta' \leftarrow \theta$ .

**end for**

**end for**

**end for**

---

## 5 Related Work

The proposed algorithm has some similarities to recent off-policy RL methods. In both Maximum a Posteriori Policy Optimization (MPO) [3] and in Soft Actor Critic (SAC) [4], the authors propose to learn the parametric policy and fit it to the non-parametric improved policy as in eq. 4 (in MPO, the  $\pi_0$  is replaced by the parametric policy, whereas in SAC,  $\pi_0$  is replaced by the uniform distribution). Furthermore, as in our method, in SAC the authors use induced soft Q-function. The both methods collect the experience using the parametric policy. In contrast, in our method, we directly use the improved non-parametric policy to collect the experience as well as to construct the bootstrapped Q-function. Moreover, our method is explicitly build in the context of multi-task learning and makes use of behaviour prior with information asymmetry [16] which encourages structured exploration.

In recent work on Q-learning, there were many attempts to scale it up to high-dimensional and continuous action domains. In soft Q-learning [21], in the context of maximum entropy RL, the authors learn a parametric mapping from normally-distributed samples to ones drawn from a policy distribution, which converges to the optimal non-parametric policy induced by a soft Q function (in a similar way as in eq. 4 with a uniform  $\pi_0$ ). In Amortized Q-learning [22], the authors propose to learn a proposal distribution for actions and then select the one maximizing the Q-function. Unlike in our work, the authors do not regularize the induced non-parametric distribution to stay close to the proposal. Note that, in the limit of the temperature  $\tau \rightarrow 0$ , then our softmax operator over importance weights becomes a max, making our approach a strict generalization of AQL. Finally, Hunt et al. [23], propose to learn a proposal distribution which is good for transfer to a new task, in the context of successor features [24] while maximizing the entropy.

Transfer of knowledge from past tasks to future ones is a well-established problem in machine learning [25, 26] and has been addressed from several different angles. Meta learning approaches try to learn the adaptation mechanism by explicitly optimizing either for minimal regret during adaptation or for performance after adaptation. Gradient-based approaches, often derived from MAML, aim at learning initial network weights such that a few gradient steps from this initialization is sufficient to adapt to new tasks [5, 27, 28, 29]. Memory-based meta learning approaches model the adaptation procedure using recurrent networks [6, 7, 30, 11, 8]. One problem of meta learning approaches is the explicit optimization for adaptation on a new task, which may be computationally expensive. In addition, most of the meta-learning methods require the training and adaptation process to be

matched. It could restrict the class of problems which can be solved by this approach since some hard meta RL problems could also constitute hard multi-task problems. Our method allows to provide additional information at training time to facilitate this learning without affecting the adaptation phase.

Other transfer learning methods (ours included) do not explicitly optimize the algorithm for adaptation. A common approach is to use a neural network which shares some parameters across training tasks and fine-tunes the rest. Recent work [12] suggests that this yields performance comparable to the MAML-style training. Transfer learning with Successor Features [24] exploits a similar decomposition of the action-value function, but relies on Generalized Policy Improvement for efficient transfer, instead of our more general gradient-based adaptation. Another approach for reusing past experience is hierarchical RL which tries to compress the experience to a shared low-level controller or a set of options which are reused in later tasks [31, 32, 33, 34]. Finally, an approach we build upon is to distill past behavior into a prior policy [15, 16] from which we can bootstrap during adaptation. In Fakoor et al. [13], the authors propose a transfer learning approach based on fine-tuning a critic acquired via a multi-task objective. To speed-up adaptation, their method makes heavy use of off-policy data acquired during meta-training, and an adaptive trust region which regularizes the critic parameters based on task similarity.

## 6 Experiments

In this section, we empirically study the performance of our method in the following scenarios. Firstly, we assess how well the method performs in the multi-task scenario. Then, we demonstrate the methods ability to achieve competitive performance in adapting to hold-out tasks compared to meta reinforcement learning baselines on a few standard benchmarks. On top of that, we show that the method scales well to more challenging sparse reward scenarios and achieves superior adaptation performance on hold out tasks compared to considered baselines. Finally, we consider the case when the number of training tasks is very small. In this case the behaviour prior and value-function representation may overfit to the training tasks. We demonstrate that our method still generalizes to hold-out task when additional fine-tuning is allowed.

**Task setup.** We consider two standard meta reinforcement learning problems: 2D point mass navigation and half cheetah velocity task, described in Rakelly et al. [8]. In addition to these simple tasks, we design a set of sparse reward tasks, which are harder as control and exploration problems: *Go To Ring*: a quadruped body needs to navigate to a particular (unknown) position on a ring. *Move Box*: a sphere-like robot must move a box to a specific position. *Reach*: a simulated robotic arm is required to reach a particular (unknown) goal position. *GTT*: A humanoid body needs to navigate to a particular (unknown) position on a rectangle. For every task, we consider a set of training  $\mathcal{T}_1, \dots, \mathcal{T}_n$ , and held-out tasks  $\mathcal{T}_{n+1}, \dots, \mathcal{T}_{n+m}$ . For every task, the policy receives proprioceptive information, as well as the global position of the body and the unstructured task identifier (a number from 1 to  $n$ ). For the *Move Box* task, we provide additional global position of the target as task observation on training distribution to facilitate learning. We do not provide this information when working on test tasks. For more environment details, please refer to Appendix B.

**Multi-task training.** We first demonstrate our method ability to solve multi-task learning problems. As baseline, we consider SVG(0) [35], an actor-critic algorithm with additional Retrace off-policy correction [36] for learning the Q-function as described in [37]. We refer to this algorithm as *RS(0)*. We further consider a continuous-action version of DISTRAL [15] built on top of *RS(0)*, where we learn a behaviour prior alongside the policy and value function, similar to [16]. This prior exhibits information asymmetry of observations with respect to the policy and the value function (it receives less information) which makes it to learn useful default behaviour speeding up the learning. In Appendix B, we specify the information provided to the behaviour prior and the policy. Furthermore, we consider MPO [3] algorithm as well as its version with behaviour prior, which we call MPO + DISTRAL. The latter simply uses KL-regularization to the learned prior (alongside the policy learning) in the M-step as soft constraint as well as soft Q-function. In our method, IWPL, we also use the behaviour prior with information asymmetry between Q-function, which receives task-specific information.

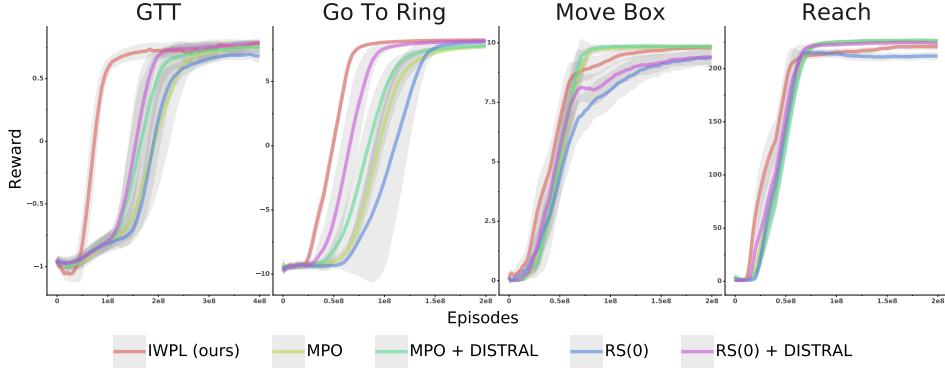


Figure 1: Multi-task training results.

For each of the models, we optimize hyperparameters and report the best found configuration with 3 random seeds. The experiments are run in a distributed setup with 64 actors that generate experience and a single learner somewhat similar to Espeholt et al. [38] using. We use a replay buffer of size  $10^6$  and control the number of times an individual experience tuple is considered by the learner. This ensures soft-synchronicity between actor and learner and ensures a fair comparison between models that differ with respect to the compute cost of inference and learning. For more details, please refer to the Appendix A.

The results are given on Figure 1. We can see that our method achieves competitive performance compared to the baselines. Note that it has larger gains in tasks where the control problem is harder. This effect of behaviour prior was observed in [16] and presumably is amplified for IWPL, where there is no intermediate parametric policy in the loop. It immediately samples the useful actions from the prior which is learned faster than the agent policy due to the restricted set of observations as discussed in [16]. Interestingly, we do not observe a difference between MPO and MPO+DISTRAL, presumably because the effect of the behaviour prior is reduced by the hard KL constraint to the previous policy.

**Adaptation performance.** Next, we investigate performance of our method in adapting to hold-out tasks. The main criteria is the data efficiency in terms of a number of episodes on a new task. As discussed in Section 4, we want to leverage the behaviour prior as well as learned shared representation for the action-value function. Therefore, we consider two variants of our method, IWPA described in Section 4. We refer to "Shared Q + IW" as the version which leverages both behaviour prior and action-value function, and "IW", which leverages only behaviour prior and learns action value function from scratch without making assumption 12. As natural baseline, we consider RS(0) + DISTRAL agent as in multi-task learning where for learning Q-function we use TD(0) as in IWPA. Starting from this, we call "Shared Q", the agent which leverages both behaviour prior and action-value function and "DISTRAL" which leverages only behaviour prior.

We pre-train "RS(0) + DISTRAL" agent with Q-function parameterisation 12 on the training set, choose best performing hyperparameter and freeze pretrained  $\pi_0$  and action-value features  $\psi$  for each task. Then we apply all four proposed adaptation methods to these behaviour prior and action-value features. The reason to use one algorithm for pretraining is to isolate the adaptation performance from the multi-task performance studied above. Empirically, we found that models trained based on IWPL lead to similar results, but we decided to report the results pretrained using "RS(0) + DISTRAL" because this agent was already considered in [16].

In addition, we consider two meta-reinforcement learning baselines: a re-implementation of RL2 [6], [7] as well as a re-implementation of PEARL [8]. For both implementations we build upon RS(0) as the base algorithm. In our implementation of PEARL (denoted as PEARL\*), we use simple LSTM to encode the context. As reported in Rakelly et al. [8], this variant is slower to learn but eventually achieves similar to PEARL performance. Despite this change, our results achieve comparable performance to those presented in Section 6.3 of [8]. On top of that, we also consider a baseline which learns to solve the test tasks "From Scratch" and corresponds to RS(0) algorithm without pre-training and behaviour prior. For more details, see Appendix A.

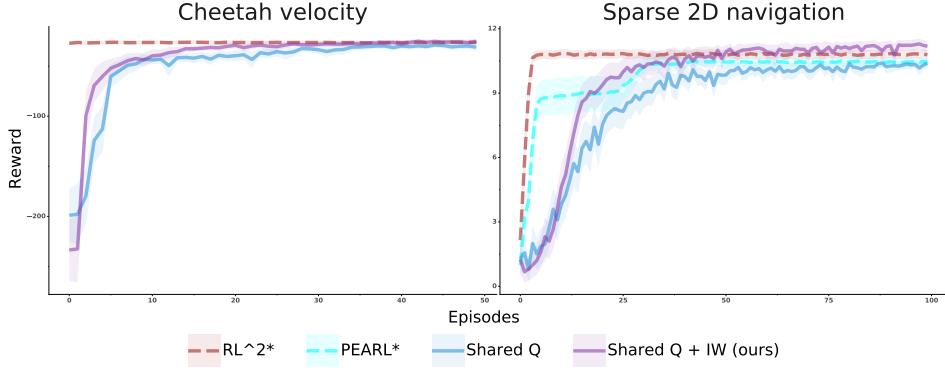


Figure 2: Adaptation performance on standard benchmarks after meta-training. Our method (not using meta-learning) achieves comparable results to other meta-learning baselines.

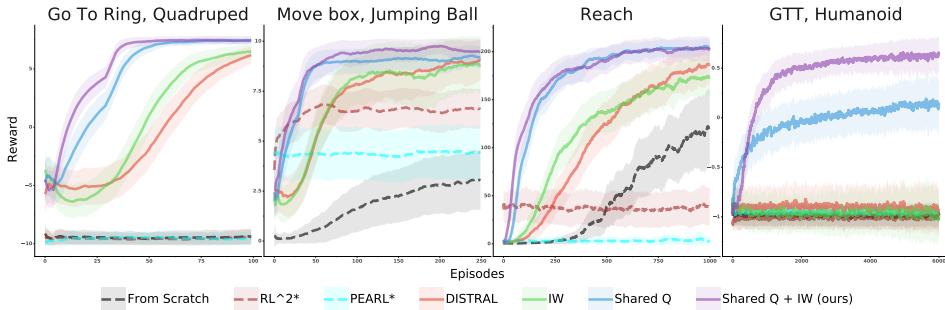


Figure 3: Adaptation performance of different methods on sparse reward tasks after meta-training.

We start by presenting test-time adaptation performance on two standard continuous control tasks used in [8]: *half-cheetah velocity* and *Sparse 2D navigation*. Note, that for *Sparse 2D navigation* task, PEARL receives dense reward during training whereas our agent is trained with sparse rewards. It additionally demonstrates that our method can be employed in more difficult scenarios. The results are presented in Figure 2. While RL2 and PEARL converge faster in absolute terms, IWPA remains competitive and converges quickly despite not optimizing the adaptation process directly.

Going further, we present the results on complex sparse reward tasks. Results on these tasks are depicted on Figure 3. Our proposed method achieves gains in adaptation time with respect to the baseline DISTRAL. Furthermore, we note that using shared features for the value function provides a significant gain. It is important to note that using shared features without the behaviour prior fails to learn fast, because the behaviour prior plays a crucial role in facilitating exploration (see Appendix D). On top of that, we observe that IWPA similarly to multi-task results section, provides bigger gains on harder to control problems, like GTT humanoid. Note that this is a very challenging task: humanoid needs to locate a target and only receives a reward when successful. Furthermore, the humanoid may fail at any moment and the episodes will terminate. It makes it extremely hard to learn without any prior knowledge. We note that both RL2 and PEARL failed to achieve optimal performance on these tasks. This could be for a variety of reasons, including the sparsity of the rewards and the complexity of learning a single policy that has to operate over long time horizons.

**Generalization** An efficient transfer learning method should be robust to low data regime. Here we show that in case, when a few of training tasks are available, the method is still be able to generalize if we allow for the additional finetuning of the shared features for the Q-function after 20 episodes of interaction on a new task. For each of the sparse reward tasks, we consider a version which has few training tasks. We trained IWPL on these and compare it to the IWPL trained in large tasks regime. The results are given in Figure 4. As we see, the method trained in a low tasks regime fails to generalize in most of the tasks, whereas the additional finetuning helps to recover the final performance and still be able to do it faster than learning from scratch.

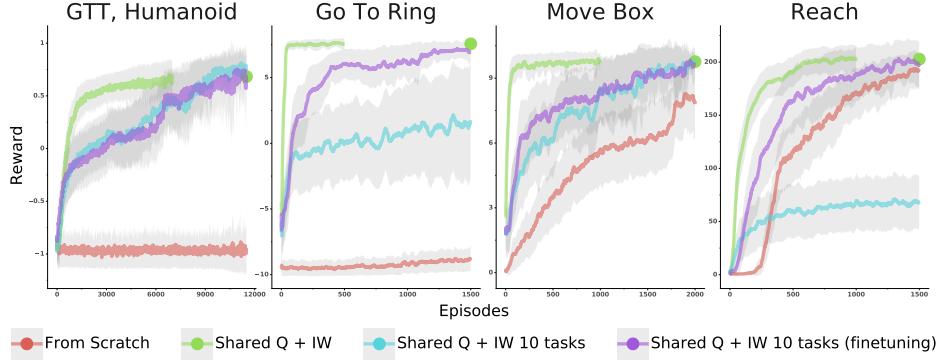


Figure 4: Generalization results. We report the performance of learning from scratch as well as Shared Q + IW architecture trained in high task regime. On top of that, we show the performance of the architectures trained in the low task regime with and without a finetuning of value function features. We denote by a point the final performance of the early-stopped Shared Q + IW experiment.

## 7 Discussion

We have presented a novel method for multi-task learning as well as for adaptation to new hold-out tasks which does not explicitly meta-learn the adaptation process and yet can match the adaptation speed of common meta-reinforcement learning algorithms. Instead of explicit meta-learning, we relied on feature reuse and bootstrapping from a behavioral prior. The behavior prior can be seen as an informed proposal for a task distribution that is then specialized to a particular task by a learned action-value function. This scheme can be easily integrated into different actor-critic algorithms for data efficient off-policy learning at training and test time. It further does not strictly require to execute test time adaptation as an inner loop during training thus adding extra flexibility.

## References

- [1] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.
- [2] G. Dulac-Arnold, D. Mankowitz, and T. Hester. Challenges of real-world reinforcement learning, 2019.
- [3] A. Abdolmaleki, J. T. Springenberg, J. Degrave, S. Bohez, Y. Tassa, D. Belov, N. Heess, and M. Riedmiller. Relative entropy regularized policy iteration. *arXiv preprint arXiv:1812.02256*, 2018.
- [4] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1861–1870, 2018.
- [5] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- [6] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Rl<sup>2</sup>: Fast reinforcement learning via slow reinforcement learning, 2016.
- [7] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.
- [8] K. Rakelly, A. Zhou, D. Quillen, C. Finn, and S. Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *arXiv preprint arXiv:1903.08254*, 2019.
- [9] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning, 2019.
- [10] P. A. Ortega, J. X. Wang, M. Rowland, T. Genewein, Z. Kurth-Nelson, R. Pascanu, N. Heess, J. Veness, A. Pritzel, P. Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.
- [11] J. Humplík, A. Galashov, L. Hasenclever, P. A. Ortega, Y. W. Teh, and N. Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.
- [12] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.
- [13] R. Fakoor, P. Chaudhari, S. Soatto, and A. J. Smola. Meta-q-learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJeD3CEFPH>.
- [14] S. Fujimoto, H. Van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.
- [15] Y. Teh, V. Bapst, W. M. Czarnecki, J. Quan, J. Kirkpatrick, R. Hadsell, N. Heess, and R. Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4496–4506, 2017.
- [16] A. Galashov, S. Jayakumar, L. Hasenclever, D. Tirumala, J. Schwarz, G. Desjardins, W. M. Czarnecki, Y. W. Teh, R. Pascanu, and N. Heess. Information asymmetry in KL-regularized RL. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=S11qMn05Ym>.
- [17] R. Fox, A. Pakman, and N. Tishby. Taming the noise in reinforcement learning via soft updates, 2015.

- [18] K. Hausman, J. T. Springenberg, Z. Wang, N. Heess, and M. Riedmiller. Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*, 2018.
- [19] J. Peters, K. Mülling, and Y. Altün. Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1607–1612. AAAI Press, 2010.
- [20] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning, 2013.
- [21] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies, 2017.
- [22] T. V. de Wiele, D. Warde-Farley, A. Mnih, and V. Mnih. Q-learning in enormous action spaces via amortized approximate maximization, 2020.
- [23] J. J. Hunt, A. Barreto, T. P. Lillicrap, and N. Heess. Composing entropic policies using divergence correction, 2018.
- [24] A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P. van Hasselt, and D. Silver. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pages 4055–4065, 2017.
- [25] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [26] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.
- [27] C. Finn, K. Xu, and S. Levine. Probabilistic model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9516–9527. Curran Associates, Inc., 2018. URL <http://papers.nips.cc/paper/8161-probabilistic-model-agnostic-meta-learning.pdf>.
- [28] A. Gupta, R. Mendonca, Y. Liu, P. Abbeel, and S. Levine. Meta-reinforcement learning of structured exploration strategies. *arXiv preprint arXiv:1802.07245*, 2018.
- [29] A. Nichol and J. Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2:2, 2018.
- [30] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.
- [31] E. Brunskill and L. Li. Pac-inspired option discovery in lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 316–324, 2014.
- [32] N. Heess, G. Wayne, Y. Tassa, T. Lillicrap, M. Riedmiller, and D. Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv:1610.05182*, 2016.
- [33] D. Tirumala, H. Noh, A. Galashov, L. Hasenclever, A. Ahuja, G. Wayne, R. Pascanu, Y. W. Teh, and N. Heess. Exploiting hierarchy for learning and transfer in kl-regularized rl. *arXiv preprint arXiv:1903.07438*, 2019.
- [34] M. Wulfmeier, A. Abdolmaleki, R. Hafner, J. T. Springenberg, M. Neunert, T. Hertweck, T. Lampe, N. Siegel, N. Heess, and M. Riedmiller. Regularized hierarchical policies for compositional transfer in robotics. *arXiv preprint arXiv:1906.11228*, 2019.
- [35] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa. Learning continuous control policies by stochastic value gradients. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2944–2952. Curran Associates, Inc., 2015.

- [36] R. Munos, T. Stepleton, A. Harutyunyan, and M. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems 29*, pages 1054–1062. 2016.
- [37] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. van de Wiele, V. Mnih, N. Heess, and J. T. Springenberg. Learning by playing solving sparse reward tasks from scratch. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4344–4353, 2018.
- [38] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. 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 1407–1416, Stockholm, Sweden, 10–15 Jul 2018. PMLR.

## A Experimental details

For all the models, we use similar architectures for all the components. Each agent has actor, critic and optionally behaviour prior networks. For all the methods, except for  $RL^2$  [7] and PEARL [8], actor, critic and behaviour prior networks are 2 dimensional multi-layer perceptron with ELU activation followed by one-dimensional linear layer. On top of that, for each of the networks, we use a layer normalizing inputs. For  $RL^2$  [7], the actor and critic networks are 2-dimensional multi layer perceptrons with ELU activations, followed by an LSTM with elu activations. In PEARL [8], actor and critic networks have similar structure as other methods and the encoder network is an LSTM followed by one-dimensional stochastic layer encoding Gaussian distribution. Actor and behaviour prior are represented by Gaussian distributions as well.

### A.1 Multi-task training experiment

We consider the following hyperparameter ranges:

- Learning rates:  $1e - 3, 1e - 4, 5e - 4, 5e - 5, 1e - 5$
- Initial inverse temperature  $\alpha$ :  $100, 10, 1, 1e - 1, 1e - 2, 1e - 3, 1e - 4$
- Epsilon  $\epsilon$ :  $100, 50, 30, 10, 5, 1, 0.1, 0.01, 0.001, 0.0001$
- KL-cost (inverse temperature) for DISTRAL baseline  $\alpha$ :  $1, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.0$

For the multi-task experiments, we found that the following values worked best for all the architectures:

- Learning rate:  $5e - 4$
- Epsilon  $\epsilon$ : 50

The best hyperparameters for RS(0) + DISTRAL for multi-task experiment:

- Go to Target, Humanoid:  $\alpha = 0.001$
- Go to Ring, Qudruped:  $\alpha = 0.001$
- Move box, Jumping Ball:  $\alpha = 0.001$
- Reach:  $\alpha = 0.1$

The best hyperparameters for IWPL for multi-task experiment:

- Go to Target, Humanoid:  $\epsilon = 100, \alpha = 0.1$
- Go to Ring, Qudruped:  $\epsilon = 30, \alpha = 0.01$
- Move box, Jumping Ball:  $\epsilon = 1, \alpha = 0.01$

- Reach:  $\epsilon = 50, \alpha = 0.0001$

To have a fair comparison, we optimize E-step epsilon as well as KL cost for MPO [3]. We consider the same ranges as above and the best hyperparameters are:

- Go to Target, Humanoid:  $\epsilon = 0.01, \alpha = 0.0001$
- Go to Ring, Qudruped:  $\epsilon = 0.1, \alpha = 0.001$
- Move box, Jumping Ball:  $\epsilon = 0.1, \alpha = 0.01$
- Reach:  $\epsilon = 0.001, \alpha = 0.0001$

For all the experiments, we use batch size of 512 and we split trajectories into chunks of size 10. For multi-task experiments, on Figure 1, we report 3 random seeds for each model with the best hyperparameters. Shading under the curves corresponds to 95% confidence interval within these evaluations. We split the data on the X-axis by chunks 200000 timesteps and the reward in these chunks is averaged. Then, we apply the rolling window smoothing with a window size of 200.

## A.2 Adaptation experiment

For the adaptation experiment, we train the Shared Q + DISTRAL architecture on each of the tasks. We found that the same combination of learning rate of  $5e - 4$  and of KL-cost of 0.01 worked the best, so we use the same values for pre-training for all the tasks. We run 3 random seeds of pre-training and take the best performing seed to use for adaptation, therefore producing behaviour prior  $\pi_0$  and shared features  $\psi$ . Then, for each task, we consider a small validation set consisting of 3 tasks which we use to choose the best adaptation hyperparameters. As for adaptation hyperparameter ranges, we consider only:

- Initial inverse temperature  $\alpha$ : 100, 10, 1,  $1e - 1$ ,  $1e - 2$ ,  $1e - 3$ ,  $1e - 4$
- KL-cost (inverse temperature) for DISTRAL baseline  $\alpha$ : 1, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.0

For all the adaptation experiments we use learning rate of  $5e - 4$  and epsilon of 30.

The best adaptation hyperparameters for IW and shared Q + IW:

- Sparse 2d navigation:  $\alpha = 1$ .
- Half-cheetah:  $\alpha = 0.01$
- Go to Target, Humanoid:  $\alpha = 1$ .
- Go to Ring, Qudruped:  $\alpha = 0.1$
- Move box, Jumping Ball:  $\alpha = 0.1$
- Reach:  $\alpha = 0.1$

The best adaptation hyperparameters for DISTRAL and DISTRAL + Shared Q:

- Sparse 2d navigation:  $\alpha = 0.1$
- Half-cheetah:  $\alpha = 0.1$
- Go to Target, Humanoid:  $\alpha = 0.1$
- Go to Ring, Qudruped:  $\alpha = 0.1$
- Move box, Jumping Ball:  $\alpha = 0.1$
- Reach:  $\alpha = 0.1$

As for baselines,  $RL^2$  [7] and PEARL [8], we use a learning rate of  $5e - 4$  and for PEARL we optimize a bottleneck cost from a range 10., 1., 0.1, 0.01, 0.001, 0.0001. We use bottleneck layer dimension of 5. The bottleneck costs per tasks are given here:

- Sparse 2d navigation: 0.001
- Half-cheetah: 0.001

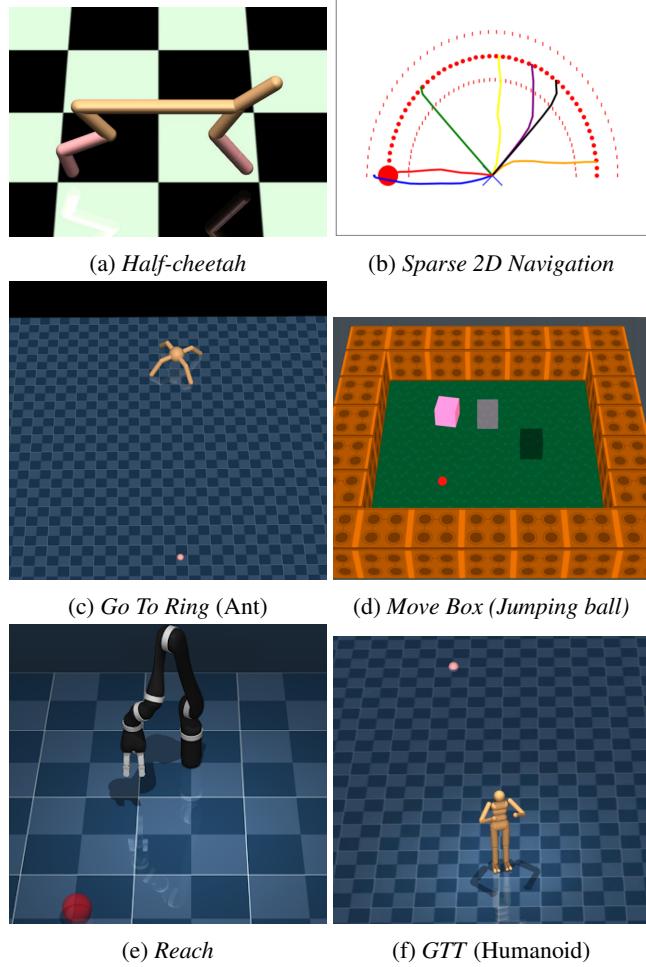


Figure 5: Tasks visualization.

- Go to Target, Humanoid: 0.1
  - Go to Ring, Qudruped: 0.0001
  - Move box, Jumping Ball: 1.
  - Reach: 0.01

**Adaptation protocol** We use a fixed protocol for adaptation on all the tasks for gradient-based methods. After each unroll of sub-trajectory of size 10, we apply 1 gradient update to the adapted parameters and after each episode we apply 50 gradient updates. The gradient updates performed by sampling trajectories from a local replay buffer with batch size of 128. Furthermore, for each task we act according to the behaviour prior (where appropriate) for a few exploration episodes.

- Sparse 2d navigation: 5 episodes.
  - Half-cheetah: 2 episodes.
  - Go to Target, Humanoid: 20 episodes.
  - Go to Ring, Qudruped: 5 episodes.
  - Move box, Jumping Ball: 5 episodes.
  - Reach: 5 episodes.

Curves from Figures 2 and 3 plot average episodic return during adaptation, averaged over 30 test tasks with 3 independent runs each (seeds). For each task and seed, we estimate average episodic

return by averaging over the last 3 episodes. Shading under the curves corresponds to 95% confidence interval within these evaluations. Results on *Sparse 2D navigation* shown in Figure 2 are smoothed using a rolling window of 5. No smoothing is applied for *Half-cheetah velocity*. For Figure 3 we use a rolling window of 30.

## B Environment Details

On *Go To Ring*, the agent receives a reward of 10 on achieving the target and is given an *immobility penalty* of -0.005 for each time step. The episode is terminated either by achieving a target or after 10 seconds (with 20 steps per second). The task distribution is defined by  $\alpha \in [0, 2\pi]$  and  $r \in [3, 7]$  which are sampled uniformly at each meta episode. At training time, we provide only task id as task-specific information. The walker is randomly spawn at each episode in the rectangle from  $[-8, 8]$ . The number of training tasks is 100, number of test tasks is 30. We provide proprioception, global position and orientation for both behaviour prior and the agent, whereas the task identifier is provided only to the agent at training time.

For *Reach*, we use a simulated Jaco robot which has to achieve a target specified in a cube with size of 0.4. Once the Jaco is within the radius of 0.05 of the target, it receives a reward of 1. The episode is terminated after 10 seconds (with 25 steps per second). At training time, we provide only task id as task-specific information. Number of training tasks is 300, number of test tasks is 30. We provide proprioception, global position and orientation for both behaviour prior and the agent, whereas the task identifier is provided only to the agent at training time.

For *Move Box*, the reward of 10 is only given once the box is on the target. The episode is terminated either after putting the box on a target or after 20 seconds (20 steps per second). The task distribution is defined by a tuple of box and target positions, which are kept fixed for the entire meta episode. These positions are sampled uniformly in the room of size 8x8 and on maximum relative distance of 2. At training time, we provide global target position as task information. Number of training tasks is 100, number of test tasks is 30. We provide proprioception, global position and orientation for both behaviour prior and the agent, whereas the global target position is provided only to the agent at training time.

For *GTT*, the agent receives the reward of 1.0 on achieving the target and is given an *immobility penalty* of -0.005 for each time step and a penalty of -1.0 if the agent (humanoid) touches the floor with the upper body or knees. The episode is terminated either by achieving a target or after 10 seconds (with 20 steps per second). The task distribution is defined by a target position sampled uniformly on the rectangle of size 8x8. At training time, we provide only task id as task-specific information. At training time, the walker position is randomly initialized in the room at each episode, whereas for the test time, the walker initial position is kept fixed for the entire meta-episode. Number of training tasks is 100, number of test tasks is 30. We provide proprioception, global position and orientation for both behaviour prior and the agent, whereas the task identifier is provided only to the agent at training time.

## C Additional Results

In Section 4 “Value Transfer”, we describe how IWPA can make use of privileged information during meta-training by mapping features  $\psi$  to task specific Q-values  $Q_i$ , via an inner product with task features  $f(g_i; w)$ . Figure 6 reports meta-training performance of “Shared Q” with either  $Q_i(s, a; \Phi_i) = \psi(s, a; \phi)^T w_i$  (referred as Task id) or  $Q_i(s, a, g_i; \Phi_i) = \psi(s, a; \phi)^T f(g_i; w)$  (referred as Task description), where  $g_i$  is a structured task descriptor. The latter yields a qualitative difference on Move Box, where this information represents a global position of a target location. This confirms that using rich privileged information during meta-training, is important to scale meta and transfer learning approaches to more challenging domains.

## D Ablations

The method IWPA described in Section 4 and in Algorithm 2 relies on both behaviour prior  $\pi_0$  and learnt Q-function features  $\psi$ . Furthermore, based on the transfer learning results presented in Figure 3,

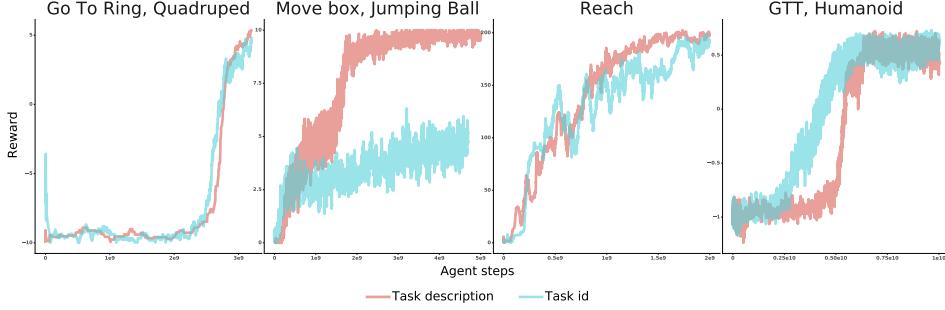


Figure 6: Meta-training performance of Shared Q method with types of task specification available at meta-training.

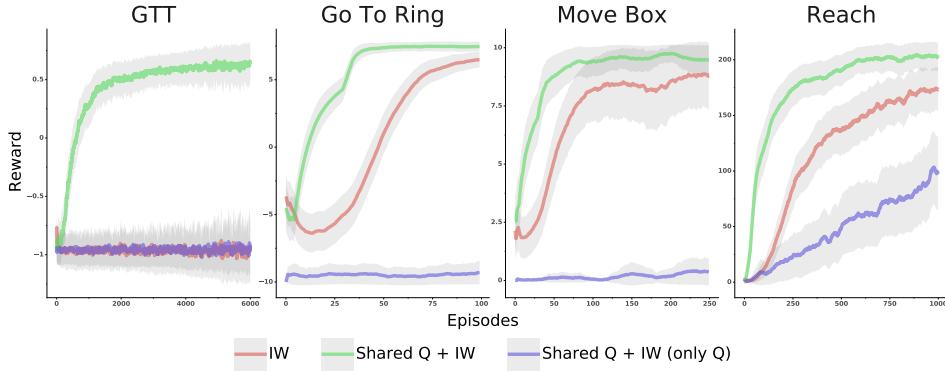


Figure 7: Ablation demonstrating "Shared Q + IW" architecture where both value function and the prior policy are reloaded. In "Shared Q + IW", only the Q-function is reloaded, and in "IW", only the behaviour prior is reloaded.

it may seem that state-action value function features are a crucial component for the transfer. In this section, we provide an ablation, where we show that without a behaviour prior, these features only do not transfer. Therefore, the combination of both, behaviour prior and value features is important. The results are given in Figure 7. As we can see, the architecture which uses both components, "Shared Q + IW" works very well, whereas the one which reloads only the value features fails to learn.