

---

# Transfer of Value Functions via Variational Methods

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We consider the problem of transferring value functions in reinforcement learning.  
2 We propose an approach that uses the given source tasks to learn a prior distribution  
3 over optimal value functions and provide an efficient variational approximation  
4 of the corresponding posterior in a new target task. We show our approach to be  
5 general, in the sense that it can be combined with complex parametric function  
6 approximators and distribution models, while providing two practical algorithms  
7 based on Gaussians and Gaussian mixtures. We theoretically analyze them by  
8 deriving a finite-sample analysis and provide a comprehensive empirical evaluation  
9 in four different domains.

## 10 1 Introduction

11 Recent advancements have allowed reinforcement learning (RL) [35] to achieve impressive results in  
12 a wide variety of complex tasks, ranging from Atari [26] through the game of Go [34] to the control  
13 of sophisticated robotics systems [16, 23, 22]. The main limitation is that these RL algorithms still  
14 require an enormous amount of experience samples before successfully learning such complicated  
15 tasks. One of the most promising solutions to alleviate this problem is transfer learning, which  
16 focuses on reusing past knowledge available to the agent in order to reduce the sample-complexity  
17 for learning new tasks. In the typical settings of transfer in RL [37], the agent is assumed to have  
18 already solved a set of *source tasks* generated from some unknown distribution. Then, given a *target*  
19 *task* (which is drawn from the same distribution, or a slightly different one), the agent can rely  
20 on knowledge from the source tasks to speed up the learning process. This reuse of knowledge  
21 constitutes a significant advantage over plain RL, where the agent learns each new task from scratch  
22 independently of any previous learning experience. Several algorithms have been proposed in the  
23 literature to transfer different elements involved in the learning process: experience samples [21, 36],  
24 policies/options [11, 18], rewards [17], features [5], parameters [10, 15], and so on. We refer the  
25 reader to [37, 19] for a thorough survey on transfer in RL.

26 Under the assumption that tasks follow a specific distribution, an intuitive choice for designing a  
27 transfer algorithm is to attempt at characterizing the uncertainty over the target task. Then, an ideal  
28 algorithm would leverage prior knowledge from the source tasks to interact with the target task  
29 to reduce the uncertainty as quickly as possible. This simple intuition makes Bayesian methods  
30 appealing approaches for transfer in RL, and many previous works have been proposed in this  
31 direction. In [39], the authors assume tasks share similarities in their dynamics and rewards and  
32 propose a hierarchical Bayesian model for the distribution of these two elements. Similarly, in [20],  
33 the authors assume that tasks are similar in their value functions and design a different hierarchical  
34 Bayesian model for transferring such information. More recently, [10], and its extension [15], consider  
35 tasks whose dynamics are governed by some hidden parameters, and propose efficient Bayesian  
36 models for quickly learning such parameters in new tasks. However, most of these algorithms require  
37 specific, and sometimes restrictive, assumptions (e.g., on the distributions involved or the function  
38 approximators adopted), which might limit their practical applicability. The importance of having

transfer algorithms that alleviate the need for strong assumptions and that easily adapt to different contexts motivates us to take a more general approach.

Similarly to [20], we assume tasks to share similarities in their value functions and use the given source tasks to learn a distribution over such functions. Then, we use this distribution as a prior for learning the target task and we propose a variational approximation of the corresponding posterior that is computationally efficient. Leveraging on recent ideas from randomized value functions [27, 3], we design a Thompson Sampling-based algorithm which efficiently explores the target task by repeatedly sampling from the posterior and acting greedily w.r.t. (with respect to) the sampled value function. We show that our approach is very general, in the sense that it can work with any parametric function approximator and with any prior/posterior distribution models (in this paper we focus on the Gaussian and Gaussian mixture models). In addition to the algorithmic contribution, we also give a theoretical contribution by providing a finite-sample analysis of our approach and an experimental contribution showing its empirical performance on four domains with increasing level of difficulty.

## 2 Preliminaries

We consider a distribution  $\mathcal{D}$  over tasks, where each task  $\mathcal{M}_\tau$  is modeled as a discounted Markov Decision Process (MDP). We define an MDP as a tuple  $\mathcal{M}_\tau = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_\tau, \mathcal{R}_\tau, p_0, \gamma \rangle$ , where  $\mathcal{S}$  is the state-space,  $\mathcal{A}$  is a finite set of actions,  $\mathcal{P}_\tau(\cdot|s, a)$  is the distribution of the next state  $s'$  given that action  $a$  is taken in state  $s$ ,  $\mathcal{R}_\tau : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function,  $p_0$  is the initial-state distribution, and  $\gamma \in [0, 1)$  is the discount factor. We assume the reward function to be uniformly bounded by a constant  $R_{max} > 0$ . A deterministic policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  is a mapping from states to actions. At the beginning of each episode of interaction, the initial state  $s_0$  is drawn from  $p_0$ . Then, the agent takes the action  $a_0 = \pi(s_0)$ , receives a reward  $\mathcal{R}_\tau(s_0, a_0)$ , transitions to the next state  $s_1 \sim \mathcal{P}_\tau(\cdot|s_0, a_0)$ , and the process is repeated. The goal is to find the policy maximizing the long-term return over a possibly infinite horizon:  $\max_\pi J(\pi) \triangleq \mathbb{E}_{\mathcal{M}_\tau, \pi} [\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_\tau(s_t, a_t)]$ . To this end, we define the optimal value function of task  $\mathcal{M}_\tau$ ,  $Q_\tau^*(s, a)$ , as the expected return obtained by taking action  $a$  in state  $s$  and following an optimal policy thereafter. Then, an optimal policy  $\pi_\tau^*$  is a policy that is greedy with respect to the optimal value function, i.e.,  $\pi_\tau^*(s) = \operatorname{argmax}_a Q_\tau^*(s, a)$  for all states  $s$ . It can be shown (e.g., [28]) that  $Q_\tau^*$  is the unique fixed-point of the optimal Bellman operator  $T_\tau$  defined by  $T_\tau Q(s, a) = \mathcal{R}_\tau(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}_\tau} [\max_{a'} Q(s', a')]$  for any value function  $Q$ . From now on, we adopt the term  $Q$ -function to denote any plausible value function, i.e., any function  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  uniformly bounded by  $\frac{R_{max}}{1-\gamma}$ . In the following, to avoid cluttering the notation, we will drop the subscript  $\tau$  whenever there is no ambiguity.

We consider a parametric family of  $Q$ -functions,  $\mathcal{Q} = \{Q_w : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \mid w \in \mathbb{R}^d\}$ , and we assume each function in  $\mathcal{Q}$  to be uniformly bounded by  $\frac{R_{max}}{1-\gamma}$ . When learning the optimal value function, a quantity of interest is how close a given function  $Q_w$  is to the fixed-point of the Bellman operator. A possible measure is its Bellman error (or Bellman residual), defined by  $B_w \triangleq TQ_w - Q_w$ . Notice that  $Q_w$  is optimal if and only if  $B_w(s, a) = 0$  for all  $s, a$ . If we assume the existence of a distribution  $\nu$  over  $\mathcal{S} \times \mathcal{A}$ , a sound objective is to directly minimize the squared Bellman error of  $Q_w$  under  $\nu$ , denoted by  $\|B_w\|_\nu^2$ . Unfortunately, it is well-known that an unbiased estimator of this quantity requires two independent samples of the next state  $s'$  for each  $s, a$  (e.g., [25]). In practice, the Bellman error is typically replaced by the TD error  $b(w)$ , which approximates the former using a single transition sample  $\langle s, a, s', r \rangle$ ,  $b(w) = r + \gamma \max_{a'} Q_w(s', a') - Q_w(s, a)$ . Finally, given a dataset  $D = \langle s_i, a_i, r_i, s'_i \rangle_{i=1}^N$  of  $N$  samples, the squared TD error is computed as  $\|B_w\|_D^2 = \frac{1}{N} \sum_{i=1}^N (r_i + \gamma \max_{a'} Q_w(s'_i, a') - Q_w(s_i, a_i))^2 = \frac{1}{N} \sum_{i=1}^N b_i(w)^2$ . Whenever the distinction is clear from the context, with a slight abuse of terminology, we refer to the squared Bellman error and squared TD error as Bellman error and TD error, respectively.

## 3 Variational Transfer Learning

In this section, we describe our variational approach to transfer in RL. In Section 3.1, we start by introducing our algorithm from a high-level perspective, in such a way that any choice of prior and posterior distributions is possible. Then, in Sections 3.2 and 3.3, we propose practical implementations based on Gaussians and mixtures of Gaussians, respectively. We conclude with some considerations on how to optimize the proposed objective in Section 3.4.

---

**Algorithm 1** Variational Transfer

---

**Require:** Target task  $\mathcal{M}_\tau$ , source  $Q$ -function weights  $\mathcal{W}_s$ , batch size  $M$

```
1: Estimate prior  $p(\mathbf{w})$  from  $\mathcal{W}_s$ 
2: Initialize variational parameters:  $\xi \leftarrow \operatorname{argmin}_\xi KL(q_\xi||p)$ 
3: Initialize replay buffer:  $D = \emptyset$ 
4: repeat
5:   Sample initial state:  $s_0 \sim p_0$ 
6:   while  $s_h$  is not terminal do
7:     Sample weights:  $\mathbf{w} \sim q_\xi(\mathbf{w})$ 
8:     Take action  $a_h = \operatorname{argmax}_a Q_{\mathbf{w}}(s_h, a)$ 
9:     Observe transition  $s_{h+1} \sim \mathcal{P}_\tau(\cdot|s_h, a_h)$  and collect reward  $r_{h+1} = \mathcal{R}_\tau(s_h, a_h)$ 
10:    Add sample to the replay buffer:  $D \leftarrow D \cup \langle s_h, a_h, r_{h+1}, s_{h+1} \rangle$ 
11:    Sample mini-batch  $D' = \langle s_i, a_i, r_i, s'_i \rangle_{i=1}^M$  from  $D$ 
12:    Estimate the gradient  $\nabla_\xi \mathcal{L}(\xi)$  using  $D'$ 
13:    Update  $\xi$  in the direction of  $-\nabla_\xi \mathcal{L}(\xi)$  using any stochastic optimizer (e.g., ADAM)
14:  end while
15: until forever
```

---

### 91 3.1 Algorithm

92 Let us observe that the distribution  $\mathcal{D}$  over tasks induces a distribution over optimal  $Q$ -functions.  
93 Furthermore, for any MDP, learning its optimal  $Q$ -function is sufficient for solving the problem.  
94 Thus, we can safely replace the distribution over tasks with the distribution over their optimal value  
95 functions. In our parametric settings, we reduce the latter to a distribution  $p(\mathbf{w})$  over weights.  
96 Assume, for the moment, that we know the distribution  $p(\mathbf{w})$  and consider a dataset  $D =$   
97  $\langle s_i, a_i, r_i, s'_i \rangle_{i=1}^N$  of samples from some task  $\mathcal{M}_\tau \sim \mathcal{D}$  that we want to solve. Then, we can  
98 compute the posterior distribution over weights given such dataset by applying Bayes theorem as  
99  $p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$ . Unfortunately, this cannot be directly used in practice since we do not have  
100 a model of the likelihood  $p(D|\mathbf{w})$ . In such case, it is very common to make strong assumptions on the  
101 MDPs or the  $Q$ -functions to get tractable posteriors. However, in our transfer settings, all distributions  
102 involved depend on the family of tasks under consideration and making such assumptions is likely  
103 to limit the applicability to specific problems. Thus, we take a different approach to derive a more  
104 general, but still well-grounded, solution. Notice that our final goal is to move the total probability  
105 mass over the weights minimizing some empirical loss measure, which in our case is the TD error  
106  $\|B_{\mathbf{w}}\|_D^2$ . Then, given a prior  $p(\mathbf{w})$ , we know from PAC-Bayesian theory that the optimal Gibbs  
107 posterior  $q$  minimizing an oracle upper bound on the expected loss takes the form (e.g., [8]):

$$q(\mathbf{w}) = \frac{e^{-\Lambda \|B_{\mathbf{w}}\|_D^2} p(\mathbf{w})}{\int e^{-\Lambda \|B_{\mathbf{w}'}\|_D^2} p(d\mathbf{w}')}, \quad (1)$$

108 for some parameter  $\Lambda > 0$ . Since  $\Lambda$  is typically chosen to increase with the number of samples  
109  $N$ , in the remaining, we set it to  $\lambda^{-1}N$ , for some constant  $\lambda > 0$ . Notice that, whenever the term  
110  $e^{-\Lambda \|B_{\mathbf{w}}\|_D^2}$  can be interpreted as the actual likelihood of  $D$ ,  $q$  becomes a classic Bayesian posterior.  
111 Although we now have an appealing distribution, the integral at the denominator of (1) is intractable  
112 to compute even for simple  $Q$ -function models. Thus, we propose a variational approximation  $q_\xi$  by  
113 considering a simpler family of distributions parameterized by  $\xi \in \Xi$ . Then, our problem reduces to  
114 finding the variational parameters  $\xi$  such that  $q_\xi$  minimizes the Kullback-Leibler (KL) divergence  
115 w.r.t. the Gibbs posterior  $q$ . From the theory of variational inference (e.g., [6]), this can be shown to  
116 be equivalent to minimizing the well-known (negative) *evidence lower bound* (ELBO):

$$\min_{\xi \in \Xi} \mathcal{L}(\xi) = \mathbb{E}_{\mathbf{w} \sim q_\xi} \left[ \|B_{\mathbf{w}}\|_D^2 \right] + \frac{\lambda}{N} KL(q_\xi(\mathbf{w}) || p(\mathbf{w})). \quad (2)$$

117 The approximate posterior balances between placing probability mass over those weights  $\mathbf{w}$  that  
118 have low expected TD error (first term), and staying close to the prior distribution (second term).  
119 Assuming that we can compute the gradients of (2) w.r.t. the variational parameters  $\xi$ , our objective  
120 can be optimized using any stochastic optimization algorithm, as shown in the next subsections.

121 We now highlight our general transfer procedure in Algorithm 1, while deferring a description of  
122 specific choices for the involved distributions to the next two subsections. Given a set of weights  $\mathcal{W}_s$   
123 from the source tasks' optimal  $Q$ -functions, we start by estimating the prior distribution (line 1), and

we initialize the variational parameters by minimizing the KL divergence w.r.t. such distribution (line 2).<sup>1</sup> Then, at each time step of interaction, we re-sample the weights from the current approximate posterior and act greedily w.r.t. the corresponding  $Q$ -function (lines 7,8). After collecting the new experience (lines 9-10), we draw a mini-batch of samples from the replay buffer (line 11), use this to estimate the objective function gradient (line 12), and update the variational parameters (line 13).

The key property of our approach is the weight resampling at line 7, which resembles the well-known Thompson sampling approach adopted in multi-armed bandits [7] and closely relates to the recent value function randomization [27, 3]. At each time we guess what is the task we are trying to solve based on our current belief and we act as if such guess were true. This mechanism allows an efficient adaptive exploration of the target task. Intuitively, during the first steps of interaction, the agent is very uncertain about the current task, and such uncertainty induces stochasticity in the chosen actions, allowing a rather informed exploration to take place. Consider, for instance, that actions that are bad on average for all tasks are improbable to be sampled, while this cannot happen in uninformed exploration strategies, like  $\epsilon$ -greedy, before learning takes place. As the learning process goes on, the algorithm will quickly figure out which task is solving, thus moving all the probability mass over the weights minimizing the TD error. From that point, sampling from the posterior is approximately equivalent to deterministically taking such weights, and no more exploration will be performed.

Finally, notice the generality of the proposed approach: as far as the objective  $\mathcal{L}$  is differentiable in the variational parameters  $\xi$ , and its gradients can be efficiently computed, any approximator for the  $Q$ -function and any prior/posterior distributions can be adopted. For the latter, we describe two practical choices in the next two sections.

### 3.2 Gaussian Variational Transfer

We now restrict to a specific choice of the prior and posterior families that makes our algorithm very efficient and easy to implement. We assume that optimal  $Q$ -functions (or better, their weights) follow a multivariate Gaussian distribution. That is, we model the prior as  $p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$  and we learn its parameters from the set of source weights using maximum likelihood estimation (with small regularization to make sure the covariance is positive definite). Then, our variational family is the set of all well-defined Gaussian distributions, i.e., the variational parameters are  $\Xi = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}, \boldsymbol{\Sigma} \succ 0\}$ . To prevent the covariance from becoming not positive definite, we consider its Cholesky decomposition  $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$  and learn the lower-triangular Cholesky factor  $\mathbf{L}$  instead. In this case, deriving the gradient of the objective is very simple. Both the KL between two multivariate Gaussians and its gradients have a simple closed-form expression. The expected log-likelihood, on the other hand, can be easily differentiated by adopting the reparameterization trick (e.g., [13, 29]). We report these results in Appendix B.2.

### 3.3 Mixture of Gaussian Variational Transfer

Although the Gaussian assumption of the previous section is very appealing as it allows for a simple and efficient way of computing the variational objective and its gradients, in practice it rarely allows us to describe the prior distribution accurately. In fact, even for families of tasks in which the reward and transition models are Gaussian, the  $Q$ -values might be far from being normally distributed. Depending on the family of tasks under consideration and, since we are learning a distribution over weights, on the chosen function approximator, the prior might have arbitrarily complex shapes. When the information loss due to the Gaussian approximation becomes too severe, the algorithm is likely to fail at capturing any similarities between the tasks. We now propose a variant to successfully solve this problem, while keeping the algorithm efficient and simple enough to be applied in practice.

Given the source tasks' weights  $\mathcal{W}_s$ , we model our estimated prior as a mixture with equally weighted isotropic Gaussians centered at each weight:  $p(\mathbf{w}) = \frac{1}{|\mathcal{W}_s|} \sum_{\mathbf{w}_s \in \mathcal{W}_s} \mathcal{N}(\mathbf{w} | \mathbf{w}_s, \sigma_p^2 \mathbf{I})$ . This model resembles a kernel density estimator [31] with bandwidth  $\sigma_p^2$  and, due to its nonparametric nature, it allows capturing arbitrarily complex distributions. Consistently with the prior, we model our approximate posterior as a mixture of Gaussians. Using  $C$  components, our posterior is  $q_{\xi}(\mathbf{w}) = \frac{1}{C} \sum_{i=1}^C \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ , with variational parameters  $\xi = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_C)$ . Once again,

<sup>1</sup>If the prior and approximate posterior were in the same family of distributions we could simply set  $\xi$  to the prior parameters. However, we are not making this assumption at this point.

we learn Cholesky factors instead of full covariances. Finally, since the KL divergence between two mixtures of Gaussians has no closed-form expression, we rely on an upper bound to such quantity, so that the negative ELBO still upper bounds the KL between the approximate and the exact posterior. Among the many upper bounds available, we adopt the one proposed in [12] (see Appendix B.3).

### 3.4 Minimizing the TD Error

From Sections 3.2 and 3.3, we know that differentiating the negative ELBO  $\mathcal{L}$  w.r.t.  $\xi$  requires differentiating  $\|B_w\|_D^2$  w.r.t.  $w$ . Unfortunately, the TD error is well-known to be non-differentiable due to the presence of the max operator. This issue is rarely a problem since typical value-based algorithms are semi-gradient methods, i.e., they do not differentiate the targets (see, e.g., Chapter 11 of [35]). However, our transfer settings are quite different from common RL. In fact, our algorithm is likely to start from  $Q$ -functions that are very close to an optimum and aims only to adapt the weights in some direction of lower error so as to quickly converge to the solution of the target task. Unfortunately, this property does not hold for most semi-gradient algorithms. Even worse, many online RL algorithms combined with complex function approximators (e.g., DQNs) are well-known to be unstable, especially when approaching an optimum, and require many tricks and tuning to work well [30, 38]. This property is clearly undesirable in our case, as we only aim at adapting already good solutions. Thus, we consider using a residual gradient algorithm [4]. To differentiate the targets, we replace the optimal Bellman operator with the mellow Bellman operator introduced in [2], which adopts a softened version of max called *mellowmax*:

$$\text{mm}_a Q_w(s, a) = \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_a e^{\kappa Q_w(s, a)} \quad (3)$$

where  $\kappa$  is a hyperparameter and  $|\mathcal{A}|$  is the number of actions. The mellow Bellman operator, which we denote as  $\tilde{T}$ , has several appealing properties: (i) it converges to the maximum as  $\kappa \rightarrow \infty$ , (ii) it has a unique fixed-point, and (iii) it is *differentiable*. Denoting by  $\tilde{B}_w = \tilde{T}Q_w - Q_w$  the Bellman residual w.r.t. the mellow Bellman operator  $\tilde{T}$ , we have that the corresponding TD error,  $\|\tilde{B}_w\|_D^2$ , is now differentiable w.r.t.  $w$ . Further considerations on how to better optimize it are given in Appendix B.1.

## 4 Theoretical Analysis

A first important question that we need to answer is whether replacing max with mellow-max in the Bellman operator constitutes a strong approximation or not. It has been proved [2] that the mellow Bellman operator is a non-expansion under the  $L_\infty$ -norm and, thus, has a unique fixed-point. However, how such fixed-point differs from the one of the optimal Bellman operator remains an open question. Since mellow-max monotonically converges to max as  $\kappa \rightarrow \infty$ , it would be desirable if the fixed point of the corresponding operator also monotonically converged to the fixed point of the optimal one. We confirm that this property actually holds in the following theorem.

**Theorem 1.** *Let  $Q^*$  be the fixed-point of the optimal Bellman operator  $T$ . Define the action-gap function  $g(s)$  as the difference between the value of the best action and the second best action at each state  $s$ . Let  $\tilde{Q}$  be the fixed-point of the mellow Bellman operator  $\tilde{T}$  with parameter  $\kappa > 0$  and denote by  $\beta_\kappa > 0$  the inverse temperature of the induced Boltzmann distribution (as in [2]). Then:*

$$\|Q^* - \tilde{Q}\|_\infty \leq \frac{2\gamma R_{max}}{(1-\gamma)^2} \left\| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g}} \right\|_\infty. \quad (4)$$

The proof is provided in Appendix A.1. Notice that  $\tilde{Q}$  converges to  $Q^*$  exponentially fast as  $\kappa$  (equivalently,  $\beta_\kappa$ ) increases and the action gaps are all larger than zero. Notice that this result is of interest even outside our specific settings.

The second question that we need to answer is whether we can provide any guarantee on our algorithm’s performance when given limited data. To address this point, we consider the two variants of Algorithm 1 from Section 3.2 and 3.3 with linear approximators. We assume only a finite dataset is available and provide a finite-sample analysis bounding the expected (mellow) Bellman error under the variational distribution minimizing the objective (2). Due to space constraints, we provide only

the result for mixtures of Gaussians, while referring the reader to Appendix A.2 for the Gaussian case.

**Theorem 2.** Fix a target task  $\mathcal{M}_\tau$ . Assume linearly parameterized value functions  $Q_{\mathbf{w}}(s, a) = \mathbf{w}^T \phi(s, a)$  with bounded weights  $\|\mathbf{w}\|_2 \leq w_{\max}$  and uniformly bounded features  $\|\phi(s, a)\|_2 \leq \phi_{\max}$ . Consider the mixture version of Algorithm 1 using  $C$  components, source task weights  $\mathcal{W}_s$ , and bandwidth  $\sigma_p^2$  for the prior. Denote by  $\hat{\xi} = (\hat{\mu}_1, \dots, \hat{\mu}_C, \hat{\Sigma}_1, \dots, \hat{\Sigma}_C)$  the variational parameters minimizing the objective of Eq. (2) on a dataset  $D$  of  $N$  i.i.d. samples distributed according to  $\tau$  and  $\nu$ . Let  $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \|\tilde{B}_{\mathbf{w}}\|_\nu^2$  and define  $v(\mathbf{w}^*) \triangleq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \frac{1}{N}\mathbf{I})} [v(\mathbf{w})]$ , with  $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [\text{Var}_{\mathcal{P}_\tau} [\tilde{b}(\mathbf{w})]]$ . Then, there exist constants  $c_1, c_2, c_3$  such that, with probability at least  $1 - \delta$  over the choice of the dataset  $D$ :

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_\nu^2 \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_\nu^2 + v(\mathbf{w}^*) + c_1 \sqrt{\frac{\log \frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + 2\lambda \varphi(\Delta)}{N} + \frac{c_3}{N^2}, \quad (5)$$

where,  $\Delta$  is a vector whose  $j$ -th component is  $\Delta_j = \frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\|$  and, for a vector  $\mathbf{x} = (x_1, \dots, x_d)$ ,  $\varphi(\mathbf{x}) \triangleq \sum_i \frac{e^{-x_i}}{\sum_j e^{-x_j}} x_i$  is the softmax function.

We refer the reader to Appendix A.2 for the proof and a specific definition of the constants. Four main terms constitute our bound: the approximation error due to the limited hypothesis space (first term), the variance (second and third terms), the distance to the prior (third term), and a constant term decaying as  $\mathcal{O}(N^2)$ . Our main result shows a remarkable property of using Alg. 1 with mixtures of Gaussians: in order to tighten the bound, it is enough to have at least one source task that is close to the optimal solution of the target task. In such case, the dominating error is due to the variance of the estimates, and, thus, the algorithm is expected to achieve good performance rather quickly, as new data is collected. Furthermore, as  $N \rightarrow \infty$  the only error terms remaining are the irreducible approximation error due to the limited functional space and the variance term  $v(\mathbf{w}^*)$ . The latter is due to the fact that we minimize a biased estimator of the Bellman error and can be removed in cases where double sampling of the next state is possible (e.g., in simulation). Finally, we want to point out that the only difference between the bound of Theorem 2 and the one for the Gaussian version of Alg. 1 (Theorem 3 in Appendix A.2) is, as expected, in the term bounding the distance to the prior. While in Theorem 2 we have the (smoothened) minimum distance to the source tasks' weights, in the Gaussian case we have the distance to the mean of such weights,  $\|\mathbf{w}^* - \mu_p\|_{\Sigma_p^{-1}}$ . This proves a remarkable advantage of using mixtures. In fact, the Gaussian version requires the source tasks to be, on average, similar to the target task in order to perform well, while the mixture version only requires this property for one of them. We verify this consideration from an empirical perspective in the next section.

## 5 Experiments

In this section, we provide an experimental evaluation of our approach in four different domains with increasing level of difficulty. In all experiments, we compare our Gaussian variational transfer algorithm (GVT) and the version using a  $c$ -component mixture of Gaussians ( $c$ -MGVT) to plain no-transfer RL (NT) with  $\epsilon$ -greedy exploration. To the best of our knowledge, no existing transfer algorithm is directly comparable to our approach from an experimental perspective. A comparative discussion of related works motivating this statement is provided in the next section.

**The Rooms Problem** We consider an agent navigating in the environment depicted in Figure 1. The agent starts in the bottom-left corner and must move from one room to another to reach the goal position in the top-right corner. The rooms are connected by small doors whose locations are unknown to the agent. The state-space is modeled as a  $10 \times 10$  continuous grid, while the action-space is the set of 4 movement directions (up, right, down, left). After each action, the agent moves by 1 in the chosen direction and the final position is corrupted by Gaussian noise  $\mathcal{N}(0, 0.2)$ . In case the agent hits a wall, its position remains unchanged. The reward is 1 when reaching the goal (after which the process terminates) and 0 otherwise, while the discount factor is  $\gamma = 0.99$ . In this experiment, we consider linearly parameterized  $Q$ -functions with 121 equally-spaced radial basis features.

We generate a set of 50 source tasks for the three-room environment of Figure 1 by sampling both door locations uniformly in the allowed space, and solve all of them by directly minimizing the

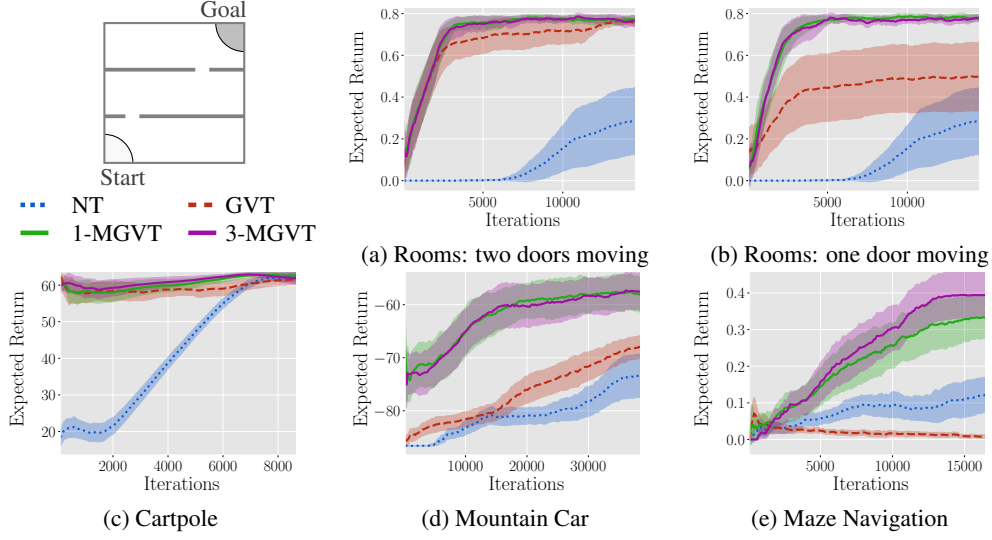


Figure 1: Expected return as a function of the number of iterations averaged over 20 independent runs. 95% confidence intervals are shown.

268 TD error as presented in Section 3.4. Then, we use our algorithms to transfer from 10 source tasks  
 269 sampled from the previously generated set. The average return over the last 50 learning episodes as a  
 270 function of the number of iterations is shown in Figure 1a. Each curve is the result of 20 independent  
 271 runs, each one resampling the target and source tasks, with 95% confidence intervals. Further details  
 272 on the parameters adopted in this experiment are given in Appendix C.1. As expected, the no-transfer  
 273 (NT) algorithm fails at learning the task in so few iterations due to the limited exploration provided by  
 274 an  $\epsilon$ -greedy policy. On the other hand, all our algorithms achieve a significant speed-up and converge  
 275 to the optimal performance in few iterations, with GVT being slightly slower. Interestingly, we notice  
 276 that there is no advantage in adopting more than 1 component for the posterior in MGVT. This result  
 277 is intuitive since, as soon as the algorithm figures out which is the target task, all the components  
 278 move towards the same region.

279 To better understand the differences between GVT and MGVT, we now consider transferring from a  
 280 slightly different distribution than the one from which target tasks are drawn. We generate 50 source  
 281 tasks again but this time with the bottom door fixed at the center and the other one moving. Then, we  
 282 repeat the previous experiment, allowing both doors to move when sampling target tasks. The results  
 283 are shown in Figure 1b. Interestingly, MGVT seems almost unaffected by this change, proving that it  
 284 has sufficient representation power to generalize to slightly different task distributions. The same  
 285 does not hold for GVT, which now is not able to solve many of the sampled target tasks, as can be  
 286 noticed from the higher variance. This result proves again that assuming Gaussian distributions can  
 287 pose severe limitations in our transfer settings.

288 **Classic Control** We now consider two well-known classic control environments: Cartpole and  
 289 Mountain Car [35]. For both, we generate 20 source tasks by uniformly sampling their physical  
 290 parameters (cart mass, pole mass, pole length for Cartpole and car speed for Mountain Car) and  
 291 solve them by directly minimizing the TD error as in the previous experiment. We parameterize  
 292  $Q$ -functions using neural networks with one layer of 32 hidden units for Cartpole and 64 for Mountain  
 293 Car. A better description of these two environments and their parameters is given in Appendix C.2.  
 294 In this experiment, we use a Double Deep Q-Network (DDQN) [38] to provide a stronger no-transfer  
 295 baseline for comparison. The results (same settings of Section 5) are shown in Figures 1c and 1d.  
 296 For Cartpole (Figure 1c), all transfer algorithms are almost zero-shot. This result is expected since,  
 297 although we vary the system parameters in a wide range, the optimal  $Q$ -values of states near the  
 298 balanced position are similar for all tasks. On the contrary, in Mountain Car (Figure 1d) the optimal  
 299  $Q$ -functions become very different when changing the car speed. This phenomenon hinders the  
 300 learning of GVT in the target task, while MGVT achieves a good jump-start and converges in fewer  
 301 iterations.

**Maze Navigation** In our last experiment, we consider a robotic agent navigating mazes. At the beginning of each episode, the agent is dropped to a random position in a  $10m^2$  maze and must reach a goal area in the smallest time possible. The robot is equipped with sensors detecting its absolute position, its orientation, the distance to any obstacle within  $2m$  in 9 equally-spaced directions, and whether the goal is present in the same range. The only actions available are *move forward* with speed  $0.5m/s$  or *rotate* (in either direction) with speed of  $\pi/8 rad/s$ . Each time step corresponds to  $1s$  of simulation. The reward is 1 for reaching the goal and 0 otherwise, while the discount factor is  $\gamma = 0.99$ . For this experiment, we design a set of 20 different mazes and solve them using a DDQN with two layers of 32 neurons and ReLU activations. Then, we fix a target maze and transfer from 5 source mazes uniformly sampled from such set (excluding the chosen target). To further assess the robustness of our method, we now consider transferring from the  $Q$ -functions learned by DDQNs instead of those obtained by minimizing the TD error as in the previous domains. From our considerations of Sections 3.4 and 4, the fixed-points of the two algorithms are different, which creates a further challenge for our method. We show the results for a fixed target maze in Figure 1e, while referring the reader to Appendix C.3 for the illustration of our mazes and additional results. Once again, MGVT achieves a remarkable speed-up over (no-transfer) DDQN. This time, using 3 components achieves slightly better performance than using only 1, which is likely due to the fact that the task distribution is much more complicated than in the previous domains. For the same reason, GVT shows negative transfer and performs even worse than DDQN.

## 6 Related Works

Our approach is mostly related to [20]. Although we both assume the tasks to share similarities in their value functions, [20] consider only linear approximators and adopt a hierarchical Bayesian model of the corresponding weights' distribution, which is assumed Gaussian. On the other hand, our variational approximation allows for more general distribution families and can be combined with non-linear approximators. Furthermore, [20] propose a Dirichlet process model for the case where weights cluster into different classes, which relates to our mixture formulation and proves the importance of capturing more complicated task distributions again. Finally, [20] considers the problem of jointly learning all given tasks, while we focus on transferring information from a set of source tasks to the target task. In [39], the authors propose a hierarchical Bayesian model for the distribution over MDPs. Unlike our approach and [20], they consider a distribution over transition probabilities and rewards, rather than value functions. In the same spirit of our method, they consider a Thompson sampling-based procedure which, at each iteration, samples a new task from the posterior and solves it. However, [39] consider only finite MDPs, which poses a severe limitation on the algorithm's applicability. On the contrary, our approach can handle high-dimensional tasks. In [10], the authors consider a family of tasks whose dynamics are governed by some hidden parameters and use Gaussian processes (GPs) to model such dynamics across tasks. Recently, [15] extended this approach by replacing GPs with Bayesian neural networks to obtain a more scalable approach. Both approaches result in a model-based algorithm that quickly adapts to new tasks by estimating their hidden parameters, while we propose a model-free method which does not require such assumptions.

## 7 Conclusion

We presented a variational method for transferring value functions in RL. We showed our approach to be general, in the sense that it can be combined with several distributions and function approximators, while providing two practical algorithms based on Gaussians and mixtures of Gaussians, respectively. We analyzed both from a theoretical and empirical perspectives, proving that the Gaussian version has severe limitations, while the mixture one is much better for our transfer settings. We evaluated the proposed algorithms in different domains, showing that both achieve excellent performance in simple tasks, while only the mixture version is able to handle complex environments.

Since our algorithm effectively models the uncertainty over tasks, a relevant future work is to design an algorithm that explicitly explores the target task to reduce such uncertainty (e.g., [14]). Furthermore, our variational approach could be extended to model a distribution over optimal policies instead of value functions (e.g., [32, 24]), which might allow better transferred behavior.



## References

- [1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.
- [2] Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pages 243–252, 2017.
- [3] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. *arXiv preprint arXiv:1802.04412*, 2018.
- [4] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.
- [5] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pages 4055–4065, 2017.
- [6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [7] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- [8] Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.
- [9] Vincent Cottet and Pierre Alquier. 1-bit matrix completion: Pac-bayesian analysis of a variational approximation. *Machine Learning*, 107(3):579–603, 2018.
- [10] Finale Doshi-Velez and George Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, volume 2016, page 1432. NIH Public Access, 2016.
- [11] Fernando Fernández and Manuela Veloso. Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 720–727. ACM, 2006.
- [12] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–317. IEEE, 2007.
- [13] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- [14] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.
- [15] Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in Neural Information Processing Systems*, pages 6250–6261, 2017.
- [16] Jens Kober and Jan R Peters. Policy search for motor primitives in robotics. In *Advances in neural information processing systems*, pages 849–856, 2009.
- [17] George Konidaris and Andrew Barto. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 489–496. ACM, 2006.
- [18] George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning.
- [19] Alessandro Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.
- [20] Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *ICML-27th International Conference on Machine Learning*, pages 599–606. Omnipress, 2010.
- [21] Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 544–551. ACM, 2008.

- 402 [22] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor  
403 policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.
- 404 [23] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David  
405 Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint*  
406 *arXiv:1509.02971*, 2015.
- 407 [24] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint*  
408 *arXiv:1704.02399*, 2017.
- 409 [25] Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-  
410 sample analysis of bellman residual minimization. In *Proceedings of 2nd Asian Conference on Machine*  
411 *Learning*, pages 299–314, 2010.
- 412 [26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare,  
413 Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through  
414 deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- 415 [27] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value  
416 functions. *arXiv preprint arXiv:1402.0635*, 2014.
- 417 [28] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley  
418 & Sons, Inc., New York, NY, USA, 1994.
- 419 [29] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approxi-  
420 mate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- 421 [30] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv*  
422 *preprint arXiv:1511.05952*, 2015.
- 423 [31] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons,  
424 2015.
- 425 [32] Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber.  
426 Policy gradients with parameter-based exploration for control. In *International Conference on Artificial*  
427 *Neural Networks*, pages 387–396. Springer, 2008.
- 428 [33] Yevgeny Seldin, François Laviolette, Nicolo Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. Pac-  
429 bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093,  
430 2012.
- 431 [34] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian  
432 Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go  
433 with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- 434 [35] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press  
435 Cambridge, 1998.
- 436 [36] Matthew E Taylor, Nicholas K Jong, and Peter Stone. Transferring instances for model-based reinforcement  
437 learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*,  
438 pages 488–505. Springer, 2008.
- 439 [37] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey.  
440 *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.
- 441 [38] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning.  
442 2016.
- 443 [39] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a  
444 hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*,  
445 pages 1015–1022. ACM, 2007.

## 446 A Proofs

### 447 A.1 Proof of Theorem 1

448 **Theorem 1.** Let  $Q^*$  be the fixed-point of the optimal Bellman operator  $T$ . Define the action-gap  
 449 function  $g(s)$  as the difference between the value of the best action and the second best action at each  
 450 state  $s$ . Let  $\tilde{Q}$  be the fixed-point of the mellow Bellman operator  $\tilde{T}$  with parameter  $\kappa > 0$  and denote  
 451 by  $\beta_\kappa > 0$  the inverse temperature of the induced Boltzmann distribution (as in [2]). Then:

$$\|Q^* - \tilde{Q}\|_\infty \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2} \left\| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g}} \right\|_\infty. \quad (4)$$

452 *Proof.* We begin by noticing that:

$$\begin{aligned} \|Q^* - \tilde{Q}\|_\infty &= \|TQ^* - \tilde{T}\tilde{Q}\|_\infty \\ &= \|TQ^* - \tilde{T}Q^* + \tilde{T}Q^* - \tilde{T}\tilde{Q}\|_\infty \\ &\leq \|TQ^* - \tilde{T}Q^*\|_\infty + \|\tilde{T}Q^* - \tilde{T}\tilde{Q}\|_\infty \\ &\leq \|TQ^* - \tilde{T}Q^*\|_\infty + \gamma \|Q^* - \tilde{Q}\|_\infty, \end{aligned}$$

453 where the first inequality follows from Minkowsky's inequality and the second one from the contrac-  
 454 tion property of the mellow Bellman operator. This implies that:

$$\|Q^* - \tilde{Q}\|_\infty \leq \frac{1}{1-\gamma} \|TQ^* - \tilde{T}Q^*\|_\infty. \quad (6)$$

455 Let us bound the norm on the right-hand side separately. In order to do that, we will bound the  
 456 function  $|TQ^*(s, a) - \tilde{T}Q^*(s, a)|$  point-wisely for any pair  $\langle s, a \rangle$ . By applying the definition of the  
 457 optimal and mellow Bellman operators, we obtain:

$$\begin{aligned} |TQ^*(s, a) - \tilde{T}Q^*(s, a)| &= |R(s, a) + \gamma \mathbb{E} \left[ \max_{a'} Q^*(s', a') \right] - R(s, a) - \gamma \mathbb{E} \left[ \text{mm}_{a'} Q^*(s', a') \right]| \\ &= \gamma \left| \mathbb{E} \left[ \max_{a'} Q^*(s', a') \right] - \mathbb{E} \left[ \text{mm}_{a'} Q^*(s', a') \right] \right| \\ &\leq \gamma \mathbb{E} \left[ \left| \max_{a'} Q^*(s', a') - \text{mm}_{a'} Q^*(s', a') \right| \right]. \end{aligned} \quad (7)$$

458 Thus, bounding this quantity reduces to bounding  $|\max_a Q^*(s, a) - \text{mm}_a Q^*(s, a)|$  point-wisely for  
 459 any  $s$ . Recall that applying the mellow Bellman operator is equivalent to computing an expectation  
 460 under a Boltzmann distribution with inverse temperature  $\beta_\kappa$  induced by  $\kappa$  [2]. Thus, we can write:

$$\begin{aligned} \left| \max_a Q^*(s, a) - \text{mm}_a Q^*(s, a) \right| &= \left| \sum_a \pi^*(a|s) Q^*(s, a) - \sum_a \pi_{\beta_\kappa}(a|s) Q^*(s, a) \right| \\ &= \left| \sum_a Q^*(s, a) (\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)) \right| \\ &\leq \sum_a |Q^*(s, a)| |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| \\ &\leq \frac{R_{\max}}{1-\gamma} \sum_a |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)|, \end{aligned} \quad (8)$$

461 where  $\pi^*$  is the optimal (deterministic) policy w.r.t.  $Q^*$  and  $\pi_{\beta_\kappa}$  is the Boltzmann distribution induced  
 462 by  $Q^*$  with inverse temperature  $\beta_\kappa$ :

$$\pi_{\beta_\kappa}(a|s) = \frac{e^{\beta_\kappa Q^*(s, a)}}{\sum_{a'} e^{\beta_\kappa Q^*(s, a')}}.$$

463 Denote by  $a_1(s)$  the optimal action for state  $s$  under  $Q^*$ . We can then write:

$$\begin{aligned}
\sum_a |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| &= |\pi^*(a_1(s)|s) - \pi_{\beta_\kappa}(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| \\
&= |1 - \pi_{\beta_\kappa}(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi_{\beta_\kappa}(a|s)| \\
&= 2|1 - \pi_{\beta_\kappa}(a_1(s)|s)|.
\end{aligned} \tag{9}$$

464 Finally, denoting with  $a_2(s)$  the second-best action in state  $s$ , let us bound this last term:

$$\begin{aligned}
|1 - \pi_{\beta_\kappa}(a_1(s)|s)| &= \left| 1 - \frac{e^{\beta_\kappa Q^*(s, a_1(s))}}{\sum_{a'} e^{\beta_\kappa Q^*(s, a')}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa(Q^*(s, a_1(s)) - Q^*(s, a_2(s)))}}{\sum_{a'} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa g(s)}}{\sum_{a'} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa g(s)}}{e^{\beta_\kappa g(s)} + \sum_{a' \neq a_1(s)} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&\leq \left| 1 - \frac{e^{\beta_\kappa g(s)}}{e^{\beta_\kappa g(s)} + |\mathcal{A}|} \right| \\
&= \left| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g(s)}} \right|.
\end{aligned} \tag{10}$$

465 Combining Eq. (8), (9), and (10), we obtain:

$$\left| \max_a Q(s, a) - \min_a Q(s, a) \right| \leq \frac{2R_{max}}{1 - \gamma} \left| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g(s)}} \right|.$$

466 Finally, using Eq. (7) we get:

$$\left| TQ^*(s, a) - \tilde{T}Q^*(s, a) \right| \leq \frac{2\gamma R_{max}}{1 - \gamma} \left| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g(s)}} \right|.$$

467 Taking the norm and plugging this into Eq. (6) concludes the proof.  $\square$

## 468 A.2 Proof of Theorem 2

469 We begin by proving some important lemmas. Then, we use them to derive a finite-sample analysis  
470 of Algorithm 1 with linearly parameterized value functions for both Gaussian distributions (Theorem  
471 3) and Gaussian mixture models (Theorem 2). We start by proving some important properties of the  
472 variational approximation introduced in Section 3.1. Our results generalize those of existing works  
473 that consider variational approximations of intractable Gibbs posteriors [1, 9]. From now on, we  
474 consider only  $Q$ -functions parameterized by weights  $\mathbf{w}$  and assume them to be uniformly bounded  
475 by  $\frac{R_{max}}{1 - \gamma}$ .

476 **Lemma 1.** *Let  $p$  and  $q$  be arbitrary distributions over weights  $\mathbf{w}$ , and  $\nu$  be a probability measure*  
477 *over  $\mathcal{S} \times \mathcal{A}$ . Consider a dataset  $D$  of  $N$  i.i.d. samples where state-action couples are distributed*  
478 *according to  $\nu$  and define  $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [\text{Var}_\mathcal{P} [\tilde{b}(\mathbf{w})]]$ . Then, for any  $\lambda > 0$  and  $\delta > 0$ , with*  
479 *probability at least  $1 - \delta$ , the following two inequalities hold simultaneously:*

$$\mathbb{E}_q \left[ \left\| \tilde{B}_\mathbf{w} \right\|_\nu^2 \right] \leq \mathbb{E}_q \left[ \left\| \tilde{B}_\mathbf{w} \right\|_D^2 \right] - \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1 - \gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \tag{11}$$

480

$$\mathbb{E}_q \left[ \left\| \tilde{B}_\mathbf{w} \right\|_D^2 \right] \leq \mathbb{E}_q \left[ \left\| \tilde{B}_\mathbf{w} \right\|_\nu^2 \right] + \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1 - \gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \tag{12}$$

481 *Proof.* From Hoeffding's inequality we have:

$$P \left( \left| \mathbb{E}_{\nu, \mathcal{P}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] - \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right| > \epsilon \right) \leq 2 \exp \left( - \frac{2N\epsilon^2}{\left( 2 \frac{R_{max}}{1-\gamma} \right)^4} \right)$$

482 which implies that, for any  $\delta > 0$ , with probability at least  $1 - \delta$ :

$$\left| \mathbb{E}_{\nu, \mathcal{P}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] - \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right| \leq 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}.$$

483 Under independence assumptions, the expected TD error can be re-written as:

$$\begin{aligned} \mathbb{E}_{\nu, \mathcal{P}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] &= \mathbb{E}_{\nu, \mathcal{P}} \left[ \frac{1}{N} \sum_{i=1}^N (r_i + \gamma \text{mm}_{a'} Q_{\mathbf{w}}(s'_i, a') - Q_{\mathbf{w}}(s_i, a_i))^2 \right] \\ &= \mathbb{E}_{\nu, \mathcal{P}} \left[ (R(s, a) + \gamma \text{mm}_{a'} Q_{\mathbf{w}}(s', a') - Q_{\mathbf{w}}(s, a))^2 \right] \\ &= \mathbb{E}_{\nu} \left[ \mathbb{E}_{\mathcal{P}} \left[ \tilde{b}(\mathbf{w})^2 \right] \right] \\ &= \mathbb{E}_{\nu} \left[ \text{Var}_{\mathcal{P}} \left[ \tilde{b}(\mathbf{w}) \right] + \mathbb{E}_{\mathcal{P}} \left[ \tilde{b}(\mathbf{w}) \right]^2 \right] \\ &= v(\mathbf{w}) + \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2, \end{aligned}$$

484 where  $v(\mathbf{w}) \triangleq \mathbb{E}_{\nu} \left[ \text{Var}_{\mathcal{P}} \left[ \tilde{b}(\mathbf{w}) \right] \right]$ . Thus:

$$\left| \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 + v(\mathbf{w}) - \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right| \leq 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \quad (13)$$

485 From the change of measure inequality [33], we have that, for any measurable function  $f(\mathbf{w})$  and  
486 any two probability measures  $p$  and  $q$ :

$$\log \mathbb{E}_p \left[ e^{f(\mathbf{w})} \right] \geq \mathbb{E}_q [f(\mathbf{w})] - KL(q||p).$$

487 Thus, multiplying both sides of (13) by  $\lambda^{-1}N$  and applying the change of measure inequality with

488  $f(\mathbf{w}) = \lambda^{-1}N \left| \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 + v(\mathbf{w}) - \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right|$ , we obtain:

$$\mathbb{E}_q [f(\mathbf{w})] - KL(q||p) \leq \log \mathbb{E}_p \left[ e^{f(\mathbf{w})} \right] \leq 4 \frac{R_{max}^2 \lambda^{-1}N}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}},$$

489 where the second inequality holds since the right-hand side of (13) does not depend on  $\mathbf{w}$ . Finally,  
490 we can explicitly write:

$$\mathbb{E}_q \left[ \left| \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 + v(\mathbf{w}) - \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right| \right] \leq \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}$$

491 from which the lemma follows straightforwardly.  $\square$

492 From Lemma 1 we can straightforwardly prove the following result which will be of fundamental  
493 importance in the remaining.

494 **Lemma 2.** Fix a task  $\mathcal{M}_{\tau}$ . Let  $p$  be a prior distribution over weights  $\mathbf{w}$ , and  $\nu$  be a probability  
495 measure over  $\mathcal{S} \times \mathcal{A}$ . Assume  $\hat{\xi}$  is the minimizer of (2) for a dataset  $D$  of  $N$  i.i.d. samples where  
496 state-action couples are distributed according to  $\nu$ . Define  $v(\mathbf{w}) \triangleq \mathbb{E}_{\nu} \left[ \text{Var}_{\mathcal{P}_{\tau}} \left[ \tilde{b}(\mathbf{w}) \right] \right]$ . Then, for  
497 any  $\delta > 0$ , with probability at least  $1 - \delta$ :

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] \leq \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_{\xi}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] + \mathbb{E}_{q_{\xi}} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(q_{\xi}||p) \right\} + 8 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}.$$

498 *Proof.* Let us use Lemma 1 for the specific choice  $q = q_{\hat{\xi}}$ . From Eq. (11), we have:

$$\begin{aligned}\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] &\leq \mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] - \mathbb{E}_{q_{\hat{\xi}}} [v(\mathbf{w})] + \frac{\lambda}{N} KL(q_{\hat{\xi}} \| p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] + \frac{\lambda}{N} KL(q_{\hat{\xi}} \| p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &= \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_{\xi}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right] + \frac{\lambda}{N} KL(q_{\xi} \| p) \right\} + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}},\end{aligned}$$

499 where the second inequality holds since  $v(\mathbf{w}) > 0$ , while the equality holds from the definition of  $\hat{\xi}$ .

500 We can now use Eq. (12) to bound  $\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 \right]$ , thus obtaining:

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] \leq \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_{\xi}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] + \mathbb{E}_{q_{\xi}} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(q_{\xi} \| p) \right\} + 8 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}.$$

501 This concludes the proof.  $\square$

502 It is worth noting the generality of Lemma 2: in bounding the expected Bellman error we do not need  
503 to assume any particular distribution, nor we have to assume any particular function approximator.

504 We are now ready to state our main result. We start from the Gaussian case and then straightforwardly  
505 extend the proof to the mixture one.

506 **Theorem 3.** Fix a target task  $\mathcal{M}_{\tau}$ . Assume linearly parameterized value functions  $Q_{\mathbf{w}}(s, a) =$   
507  $\mathbf{w}^T \phi(s, a)$  with bounded weights  $\|\mathbf{w}\|_2 \leq w_{max}$  and uniformly bounded features  $\|\phi(s, a)\|_2 \leq$   
508  $\phi_{max}$ . Consider the Gaussian version of Algorithm 1 with prior  $p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$  and denote by  
509  $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$  the variational parameter minimizing the objective of Eq. (2) on a dataset  $D$  of  $N$  i.i.d. samples  
510 distributed according to  $\tau$  and  $\nu$ . Let  $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2$  and define  $v(\mathbf{w}^*) \triangleq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \frac{1}{N} \mathbf{I})} [v(\mathbf{w})]$ ,  
511 with  $v(\mathbf{w}) \triangleq \mathbb{E}_{\nu} [Var_{\mathcal{P}} [\tilde{b}(\mathbf{w})]]$ . Then, there exist constants  $c_1, c_2, c_3$  such that, with probability at  
512 least  $1 - \delta$  over the choice of the dataset  $D$ :

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + v(\mathbf{w}^*) + c_1 \sqrt{\frac{\log \frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + \lambda \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p}^{-1}}{N} + \frac{c_3}{N^2}. \quad (14)$$

513 *Proof.* Using Lemma 2 with variational parameters  $\hat{\xi} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ , we have:

$$\begin{aligned}\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] &\leq \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_{\xi}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] + \mathbb{E}_{q_{\xi}} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(q_{\xi} \| p) \right\} + 8 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] + \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| p) \\ &\quad + 8 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}},\end{aligned} \quad (15)$$

514 where the second inequality is due to the fact that, since Lemma 2 contains an infimum over the  
515 variational parameters, we can upper bound its right-hand side by choosing any specific  $\xi$  from  $\Xi$ .  
516 Here, we choose  $\boldsymbol{\mu} = \mathbf{w}^*$  and  $\boldsymbol{\Sigma} = c\mathbf{I}$ , for some positive constant  $c > 0$ . Let us now bound these  
517 terms separately.

518 **Bounding the expected TD error** We have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \mathbb{E}_{\nu} \left[ (\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}})^2 \right] \right] \\
&= \mathbb{E}_{\nu} \left[ \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ (\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}})^2 \right] \right] \\
&= \mathbb{E}_{\nu} \left[ \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})}^2 \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] \right] + \mathbb{E}_{\nu} \left[ \text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] \right].
\end{aligned} \tag{16}$$

519 Let us bound these two terms point-wisely for each pair  $\langle s, a \rangle$ . For the first expectation, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ R(s, a) + \gamma \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] - \mathbf{w}^T \phi(s, a) \right] \\
&= R(s, a) + \gamma \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] \right] - \mathbf{w}^{*T} \phi(s, a).
\end{aligned} \tag{17}$$

520 To bound the second term, we adopt Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} e^{\kappa \mathbf{w}^T \phi(s', a')} \right] \right] \\
&\leq \mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ e^{\kappa \mathbf{w}^T \phi(s', a')} \right] \right].
\end{aligned} \tag{18}$$

521 Now, since we know that  $\mathbf{w}^T \phi(s', a') \sim \mathcal{N}(\mathbf{w}^{*T} \phi(s', a'), c \phi(s', a')^T \phi(s', a'))$ ,  $e^{\kappa \mathbf{w}^T \phi(s', a')}$  fol-  
522 lows a log-normal distribution with mean  $e^{\kappa \mathbf{w}^{*T} \phi(s', a') + \frac{1}{2} \kappa^2 c \phi(s', a')^T \phi(s', a')}$ . Thus:

$$\begin{aligned}
\mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ e^{\kappa \mathbf{w}^T \phi(s', a')} \right] \right] &= \mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} e^{\kappa \mathbf{w}^{*T} \phi(s', a') + \frac{1}{2} \kappa^2 c \phi(s', a')^T \phi(s', a')} \right] \\
&\leq \mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} e^{\kappa \mathbf{w}^{*T} \phi(s', a')} e^{\frac{1}{2} \kappa^2 c \phi_{max}^2} \right] \\
&= \mathbb{E}_{s'} \left[ \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} e^{\kappa \mathbf{w}^{*T} \phi(s', a')} \right] + \frac{1}{2} \kappa c \phi_{max}^2 \\
&= \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^{*T} \phi(s', a') \right] + \frac{1}{2} \kappa c \phi_{max}^2.
\end{aligned}$$

523 Plugging this into (18) and then into (17), we obtain:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] &\leq R(s, a) + \gamma \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^{*T} \phi(s', a') \right] + \frac{1}{2} \gamma \kappa c \phi_{max}^2 - \mathbf{w}^{*T} \phi(s, a) \\
&= \tilde{B}_{\mathbf{w}^*} + \frac{1}{2} \gamma \kappa c \phi_{max}^2.
\end{aligned}$$

524 This implies:

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})}^2 \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] &\leq \left( \tilde{B}_{\mathbf{w}^*} + \frac{1}{2} \gamma \kappa c \phi_{max}^2 \right)^2 \\
&\leq 2 \tilde{B}_{\mathbf{w}^*}^2 + \frac{1}{2} \gamma^2 \kappa^2 c^2 \phi_{max}^4,
\end{aligned}$$

525 where the second inequality follows from Cauchy-Schwarz inequality. Going back to (16), the first  
526 term can now be upper bounded by:

$$\mathbb{E}_{\nu} \left[ \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})}^2 \left[ \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + \frac{1}{2} \gamma^2 \kappa^2 c^2 \phi_{max}^4.$$

Let us now consider the variance term of (16) and derive a bound that holds point-wisely for any  $s, a$ .  
We have:

$$\begin{aligned}
\text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [\tilde{T}Q\mathbf{w} - Q\mathbf{w}] &= \text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ R(s, a) + \gamma \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] - \mathbf{w}^T \phi(s, a) \right] \\
&= \text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \gamma \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \phi(s', a') - \frac{1}{\gamma} \mathbf{w}^T \phi(s, a) \right] \right] \\
&= \text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \gamma \mathbb{E}_{s'} \left[ \text{mm}_{a'} \mathbf{w}^T \left( \phi(s', a') - \frac{1}{\gamma} \phi(s, a) \right) \right] \right] \\
&= \gamma^2 \text{Var}_{\mathcal{N}(\mathbf{w}^*, \mathbf{I})} \left[ \mathbb{E}_{s'} \left[ \text{mm}_{a'} \sqrt{c} \mathbf{w}^T \left( \phi(s', a') - \frac{1}{\gamma} \phi(s, a) \right) \right] \right].
\end{aligned}$$

From Cauchy-Schwarz inequality:

$$\begin{aligned}
\left| \sqrt{c} \mathbf{w}^T \left( \phi(s', a') - \frac{1}{\gamma} \phi(s, a) \right) \right| &\leq \sqrt{c} \|\mathbf{w}\| \left\| \phi(s', a') - \frac{1}{\gamma} \phi(s, a) \right\| \\
&\leq \sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma}.
\end{aligned}$$

Then, the random variable over which the variance is computed is limited in  $[-\sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma}, \sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma}]$  and the variance can be straightforwardly bounded using Popoviciu's inequality:

$$\text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [\tilde{T}Q\mathbf{w} - Q\mathbf{w}] \leq \gamma^2 \frac{1}{4} \left( 2\sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma} \right)^2 = c (\mathbf{w}_{\max} \phi_{\max} (1+\gamma))^2.$$

We can finally plug everything into (16), thus obtaining:

$$\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + \frac{1}{2} \gamma^2 \kappa^2 c^2 \phi_{\max}^4 + c (\mathbf{w}_{\max} \phi_{\max} (1+\gamma))^2.$$

**Bounding the KL divergence** We have:

$$\begin{aligned}
KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \parallel p) &= KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \parallel \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)) \\
&= \frac{1}{2} \left( \log \frac{|\boldsymbol{\Sigma}_p|}{c^d} + c \text{Tr}(\boldsymbol{\Sigma}_p^{-1}) + \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}^2 - d \right) \\
&\leq \frac{1}{2} d \log \frac{\sigma_{\max}}{c} + \frac{1}{2} d \frac{c}{\sigma_{\min}} + \frac{1}{2} \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}^2.
\end{aligned}$$

Now, putting all together into (15):

$$\begin{aligned}
\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] &\leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + \frac{1}{2} \gamma^2 \kappa^2 c^2 \phi_{\max}^4 + c (\mathbf{w}_{\max} \phi_{\max} (1+\gamma))^2 + \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [v(\mathbf{w})] \\
&\quad + \frac{\lambda}{N} d \log \frac{\sigma_{\max}}{c} + \frac{\lambda}{N} d \frac{c}{\sigma_{\min}} + \frac{\lambda}{N} \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}^2 + 8 \frac{R_{\max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}.
\end{aligned}$$

Since the bound holds for any  $c > 0$ , we can set it to  $1/N$ , thus obtaining:

$$\begin{aligned}
\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] &\leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + v(\mathbf{w}^*) + \frac{1}{N^2} \left( \frac{1}{2} \gamma^2 \kappa^2 \phi_{\max}^4 + \frac{\lambda d}{\sigma_{\min}} \right) \\
&\quad + \frac{1}{N} \left( \mathbf{w}_{\max}^2 \phi_{\max}^2 (1+\gamma)^2 + \lambda d (\log \sigma_{\max} + \log N) + \lambda \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}^2 \right) \\
&\quad + 8 \frac{R_{\max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}
\end{aligned}$$

Finally, defining the constants  $c_1 = \frac{8R_{\max}^2}{\sqrt{2}(1-\gamma)^2}$ ,  $c_2 = \mathbf{w}_{\max}^2 \phi_{\max}^2 (1+\gamma)^2 + \lambda d \log \sigma_{\max}$ , and

$c_3 = \frac{1}{2} \gamma^2 \kappa^2 \phi_{\max}^4 + \frac{\lambda d}{\sigma_{\min}}$ , we obtain:

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + v(\mathbf{w}^*) + c_1 \sqrt{\frac{\log \frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + \lambda \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}^2}{N} + \frac{c_3}{N^2}.$$

539

□



**Theorem 2.** Fix a target task  $\mathcal{M}_\tau$ . Assume linearly parameterized value functions  $Q_{\mathbf{w}}(s, a) = \mathbf{w}^T \phi(s, a)$  with bounded weights  $\|\mathbf{w}\|_2 \leq w_{\max}$  and uniformly bounded features  $\|\phi(s, a)\|_2 \leq \phi_{\max}$ . Consider the mixture version of Algorithm 1 using  $C$  components, source task weights  $\mathcal{W}_s$ , and bandwidth  $\sigma_p^2$  for the prior. Denote by  $\hat{\xi} = (\hat{\mu}_1, \dots, \hat{\mu}_C, \hat{\Sigma}_1, \dots, \hat{\Sigma}_C)$  the variational parameters minimizing the objective of Eq. (2) on a dataset  $D$  of  $N$  i.i.d. samples distributed according to  $\tau$  and  $\nu$ . Let  $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \|\tilde{B}_{\mathbf{w}}\|_\nu^2$  and define  $v(\mathbf{w}^*) \triangleq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \frac{1}{N}\mathbf{I})} [v(\mathbf{w})]$ , with  $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [\text{Var}_{\mathcal{P}_\tau} [\tilde{b}(\mathbf{w})]]$ . Then, there exist constants  $c_1, c_2, c_3$  such that, with probability at least  $1 - \delta$  over the choice of the dataset  $D$ :

$$\mathbb{E}_{q_{\hat{\xi}}} \left[ \|\tilde{B}_{\mathbf{w}}\|_\nu^2 \right] \leq 2 \|\tilde{B}_{\mathbf{w}^*}\|_\nu^2 + v(\mathbf{w}^*) + c_1 \sqrt{\frac{\log \frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + 2\lambda \varphi(\Delta)}{N} + \frac{c_3}{N^2}, \quad (5)$$

where,  $\Delta$  is a vector whose  $j$ -th component is  $\Delta_j = \frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\|$  and, for a vector  $\mathbf{x} = (x_1, \dots, x_d)$ ,  $\varphi(\mathbf{x}) \triangleq \sum_i \frac{e^{-x_i}}{\sum_j e^{-x_j}} x_i$  is the softmax function.

*Proof.* Similarly to the previous proof, we can apply Lemma 2 with variational parameters  $\hat{\xi} = (\hat{\mu}_1, \dots, \hat{\mu}_C, \hat{\Sigma}_1, \dots, \hat{\Sigma}_C)$ , while choosing the same specific parameters for the right-hand side:  $\mu_i = \mathbf{w}^*$  and  $\Sigma_i = c\mathbf{I}$  for all  $i = 1, \dots, C$ . Then, we obtain:

$$\begin{aligned} \mathbb{E}_{q_{\hat{\xi}}} \left[ \|\tilde{B}_{\mathbf{w}}\|_\nu^2 \right] &\leq \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_\xi} \left[ \|\tilde{B}_{\mathbf{w}}\|_\nu^2 \right] + \mathbb{E}_{q_\xi} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(q_\xi \| p) \right\} + 8 \frac{R_{\max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[ \|\tilde{B}_{\mathbf{w}}\|_\nu^2 \right] + \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| p) \\ &\quad + 8 \frac{R_{\max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \end{aligned} \quad (19)$$

The only difference w.r.t. Eq. (15) of Theorem 3 is the KL divergence term, which now contains a mixture distribution. From Theorem 4 we have:

$$KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| p) \leq KL(\chi^{(2)} \| \chi^{(1)}) + \sum_j \chi_j^{(2)} KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| \mathcal{N}(\mathbf{w}_j, \sigma_p^2 \mathbf{I})), \quad (20)$$

where the vectors  $\chi^{(1)}$  and  $\chi^{(2)}$  are the ones defined in Theorem 4. Notice that, since we reduced the posterior to one component, we can get rid of the index  $i$ . Using the definitions of these two vectors from Section 8 of [12], we have:

$$\begin{aligned} \chi_j^{(1)} &= \frac{1}{|\mathcal{W}_s|} \quad \forall j = 1, \dots, |\mathcal{W}_s| \\ \chi_j^{(2)} &= \frac{e^{-KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| \mathcal{N}(\mathbf{w}_j, \sigma_p^2 \mathbf{I}))}}{\sum_{j'} e^{-KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| \mathcal{N}(\mathbf{w}_{j'}, \sigma_p^2 \mathbf{I}))}} \quad \forall j = 1, \dots, |\mathcal{W}_s|. \end{aligned} \quad (21)$$

Since the KL divergence is:

$$KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \| \mathcal{N}(\mathbf{w}_j, \sigma_p^2 \mathbf{I})) = \frac{1}{2} \left( d \log \frac{\sigma_p^2}{c} + d \frac{c}{\sigma_p^2} + \frac{1}{\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\|^2 - d \right),$$

Eq. (21) can be rewritten as:

$$\chi_j^{(2)} = \frac{e^{-\frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\|^2}}{\sum_{j'} e^{-\frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_{j'}\|^2}} \quad \forall j = 1, \dots, |\mathcal{W}_s|.$$

Let us bound the two terms of (20) separately. For the first one, we have:

$$\begin{aligned} KL(\chi^{(2)} \| \chi^{(1)}) &= \sum_j \chi_j^{(2)} \log \frac{\chi_j^{(2)}}{\chi_j^{(1)}} \\ &= \sum_j \chi_j^{(2)} \log \chi_j^{(2)} - \sum_j \chi_j^{(2)} \log \frac{1}{|\mathcal{W}_s|} \\ &\leq \log |\mathcal{W}_s|, \end{aligned}$$

where the inequality holds since the first term is negative. For the second term of (20):

$$\begin{aligned} \sum_j \chi_j^{(2)} KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \parallel \mathcal{N}(\mathbf{w}_j, \sigma_p^2 \mathbf{I})) &= \frac{1}{2} \sum_j \chi_j^{(2)} \left( d \log \frac{\sigma_p^2}{c} + d \frac{c}{\sigma_p^2} + \frac{1}{\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\| - d \right) \\ &\leq \frac{1}{2} d \log \frac{\sigma_p^2}{c} + \frac{1}{2} d \frac{c}{\sigma_p^2} + \sum_j \chi_j^{(2)} \frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\| \\ &= \frac{1}{2} d \log \frac{\sigma_p^2}{c} + \frac{1}{2} d \frac{c}{\sigma_p^2} + \varphi(\Delta). \end{aligned}$$

where we defined the vector  $\Delta$  whose components are  $\Delta_j = \frac{1}{2\sigma_p^2} \|\mathbf{w}^* - \mathbf{w}_j\|$ . Putting the two terms together:

$$KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) \parallel p) \leq \log |\mathcal{W}_s| + \frac{1}{2} d \log \frac{\sigma_p^2}{c} + \frac{1}{2} d \frac{c}{\sigma_p^2} + \varphi(\Delta).$$

Notice that, from now on, one can simply apply the proof of Theorem 3 with  $\sigma_{max} = \sigma_{min} = \sigma_p^2$  and  $\frac{1}{2} \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\Sigma_p^{-1}}$  replaced by  $\varphi(\Delta)$ . Thus, by redefining the three constants to  $c_1 = \frac{8R_{max}^2}{\sqrt{2}(1-\gamma)^2}$ ,  $c_2 = \mathbf{w}_{max}^2 \phi_{max}^2 (1+\gamma)^2 + \lambda d \log \sigma_p^2 + 2\lambda \log |\mathcal{W}_s|$ , and  $c_3 = \frac{1}{2} \gamma^2 \kappa^2 \phi_{max}^4 + \frac{\lambda d}{\sigma_p^2}$ , we can write that, with probability at least  $1 - \delta$ :

$$\mathbb{E}_{q_{\xi}} \left[ \left\| \tilde{B}_{\mathbf{w}} \right\|_{\nu}^2 \right] \leq 2 \left\| \tilde{B}_{\mathbf{w}^*} \right\|_{\nu}^2 + v(\mathbf{w}^*) + c_1 \sqrt{\frac{\log \frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + 2\lambda \varphi(\Delta)}{N} + \frac{c_3}{N^2}.$$

569

□

## 570 B Additional Details on the Algorithms

### 571 B.1 Minimizing the TD Error

Although residual algorithms have guaranteed convergence, they are typically much slower than their semi-gradient counterpart. [4] proposed to project the gradient in a direction that achieves higher learning speed, while preserving convergence. This projection is obtained by including a parameter  $\psi \in [0, 1]$  in the TD error gradient:

$$\nabla_{\mathbf{w}} \left\| \tilde{B}_{\mathbf{w}} \right\|_D^2 = \frac{2}{N} \sum_{i=1}^N b_i(\mathbf{w}) \left( \gamma \psi \nabla_{\mathbf{w}} \text{mm}_{a'} Q_{\mathbf{w}}(s'_i, a') - \nabla_{\mathbf{w}} Q_{\mathbf{w}}(s_i, a_i) \right), \quad (22)$$

where  $b_i(\mathbf{w}) = r_i + \gamma \text{mm}_{a'} Q_{\mathbf{w}}(s'_i, a') - Q_{\mathbf{w}}(s_i, a_i)$ . Notice that  $\psi$  trades-off between the semi-gradient ( $\psi = 0$ ) and the full residual gradient ( $\psi = 1$ ). A good criterion for choosing such parameter is to start with values close to zero (to have faster learning) and move to higher values when approaching the optimum (to guarantee convergence).

### 580 B.2 Gaussian Variational Transfer

Under Gaussian distributions, all quantities of interest for using Algorithm 1 can be computed very easily. The KL divergence between the prior and approximate posterior can be computed in closed-form as:

$$KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) = \frac{1}{2} \left( \log \frac{|\Sigma_p|}{|\Sigma|} + \text{Tr}(\Sigma_p^{-1} \Sigma) + (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \Sigma_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) - d \right) \quad (23)$$

for  $\xi = (\boldsymbol{\mu}, \mathbf{L})$  and  $\Sigma = \mathbf{L}\mathbf{L}^T$ . Its gradients with respect to the variational parameters are:

$$\nabla_{\boldsymbol{\mu}} KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) = \Sigma_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) \quad (24)$$

585

$$\nabla_{\mathbf{L}} KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) = \Sigma_p^{-1} \mathbf{L} - (\mathbf{L}^{-1})^T \quad (25)$$

Finally, the gradients w.r.t. the expected likelihood term of the variational objective (2) can be computed using the reparameterization trick (e.g., [13, 29]):

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)} [\|B_{\mathbf{w}}\|_D^2] = \mathbb{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\nabla_{\mathbf{w}} \|B_{\mathbf{w}}\|_D^2] \quad \text{for } \mathbf{w} = \mathbf{L}\mathbf{v} + \boldsymbol{\mu} \quad (26)$$

588

$$\nabla_{\mathbf{L}} \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)} [\|B_{\mathbf{w}}\|_D^2] = \mathbb{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\nabla_{\mathbf{w}} \|B_{\mathbf{w}}\|_D^2 \cdot \mathbf{v}^T] \quad \text{for } \mathbf{w} = \mathbf{L}\mathbf{v} + \boldsymbol{\mu} \quad (27)$$

### 589 B.3 Mixture of Gaussian Variational Transfer

590 As mentioned in the main paper, for the mixture version of Alg. 1 we rely on the upper bound on the  
 591 KL divergence between two mixture of Gaussians presented in [12]. We report it here for the sake of  
 592 completeness.

593 **Theorem 4** ([12]). *Let  $p = \sum_i c_i^{(p)} f_i^{(p)}$  and  $q = \sum_j c_j^{(q)} f_j^{(q)}$  be two mixture of Gaussian distri-*  
 594 *butions, where  $f_i^{(p)} = \mathcal{N}(\boldsymbol{\mu}_i^{(p)}, \boldsymbol{\Sigma}_i^{(p)})$  denotes the  $i$ -th component of  $p$ ,  $c_i^{(p)}$  denotes its weight, and*  
 595 *similarly for  $q$ . Introduce two vectors  $\chi^{(1)}$  and  $\chi^{(2)}$  such that  $c_i^{(p)} = \sum_j \chi_{j,i}^{(2)}$  and  $c_j^{(q)} = \sum_i \chi_{i,j}^{(1)}$ .*  
 596 *Then:*

$$KL(p||q) \leq KL(\chi^{(2)}||\chi^{(1)}) + \sum_{i,j} \chi_{j,i}^{(2)} KL(f_i^{(p)}||f_j^{(q)}) \quad (28)$$

597 Our new algorithm replaces the KL with the above-mentioned upper bound. Each time we require its  
 598 value, we have to recompute the parameters  $\chi^{(1)}$  and  $\chi^{(2)}$  that tighten the bound. As shown in [12],  
 599 we can use a simple fixed-point procedure for this purpose. Finally, both terms in the objective are  
 600 now linear combinations of functions of the variational parameters of different components, and their  
 601 gradients easily derive from the ones of the Gaussian case.

602 Consider we have  $C$  components for the posterior family  $q_{\boldsymbol{\xi}}(\mathbf{w}) = \frac{1}{C} \sum_{i=1}^C \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  and a  
 603 prior distribution, constructed from the set of weights  $\mathcal{W}_s = \{\mathbf{w}_1, \dots, \mathbf{w}_{|\mathcal{W}_s|}\}$  of the sources' optimal  
 604  $Q$ -functions,  $p(\mathbf{w}) = \frac{1}{|\mathcal{W}_s|} \sum_{j=1}^{|\mathcal{W}_s|} \mathcal{N}(\mathbf{w}|\mathbf{w}_j, \sigma_p^2 \mathbf{I})$ . Then:

$$KL(q_{\boldsymbol{\xi}}(\mathbf{w}) || p(\mathbf{w})) \leq KL(\chi^{(2)}||\chi^{(1)}) + \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) || \mathcal{N}(\mathbf{w}|\mathbf{w}_j, \sigma_p^2 \mathbf{I})) \quad (29)$$

605 And substituting (29) in the negative ELBO in (2) we get the following upper bound.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\xi}) \leq \tilde{\mathcal{L}}(\boldsymbol{\xi}) &= \mathbb{E}_{\mathbf{w} \sim q_{\boldsymbol{\xi}}} [\|B_{\mathbf{w}}\|_D^2] \\ &+ \frac{\lambda}{N} KL(\chi^{(2)}||\chi^{(1)}) + \frac{\lambda}{N} \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) || \mathcal{N}(\mathbf{w}|\mathbf{w}_j, \sigma_p^2 \mathbf{I})) \end{aligned} \quad (30)$$

606 Finally, using this upper bound as objective of our optimization problem, we can then exploit the  
 607 linearity of the expectation operator to obtain

$$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{\xi}) &= \frac{1}{C} \sum_{i=1}^C \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} [\|B_{\mathbf{w}}\|_D^2] \\ &+ \frac{\lambda}{N} KL(\chi^{(2)}||\chi^{(1)}) + \frac{\lambda}{N} \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) || \mathcal{N}(\mathbf{w}|\mathbf{w}_j, \sigma_p^2 \mathbf{I})) \end{aligned} \quad (31)$$

608 that is easily differentiable with respect to  $\boldsymbol{\xi} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_C)$  using the Equations (24),  
 609 (25), (26), (27) derived for the Gaussian case.

## 610 C Additional Details on the Experiments

611 In the present section we provide details on the parameters adopted in all experiments and discuss  
 612 further results.

### 613 C.1 The Rooms Problem

614 In order to train the source tasks, we directly minimize the TD error based on the *mellow* Bellman  
 615 operator by stochastic gradient descent. We use a *batch size* of 50, a *buffer size* of 50000,  $\psi = 0.5$

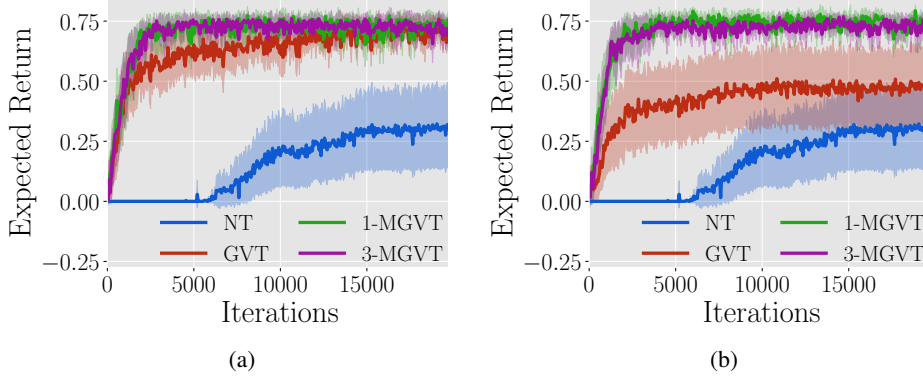


Figure 2: (a) Rooms Problem: Expected Return w.r.t. Greedy Policy, (b) Rooms Problem: Expected Return w.r.t. Greedy Policy in the generalization experiment

and a learning rate  $\alpha = 0.001$ . Additionally, we use an  $\epsilon$ -greedy policy for exploration, with  $\epsilon$  linearly decaying from 1 to 0.02 in a fraction of 0.7 the maximum number of iterations.

For the transfer algorithm GVT, we set a *batch size* of 50 and a *buffer size* of 10000. We use  $\psi = 0.5$ ,  $\lambda = 10^{-4}$  and 10 *weights* to estimate the expected TD error. For the learning rates,  $\alpha_\mu = 0.001$  for the mean of the posterior and  $\alpha_L = 0.1$  to learn its Cholesky factor  $L$ . Furthermore, we restrict the minimum value reachable by the eigenvalues of these factors to be  $\sigma_{min}^2 = 0.0001$ . In the case of MGVT we use, instead,  $\lambda = 10^{-6}$ ,  $\alpha_\mu = 0.001$  and  $\alpha_L = 0.1$ . Finally, we use a bandwidth  $\sigma_p^2 = 10^{-5}$  for the prior.

Besides the results that we show in Section 5, we present in this section further empirical evaluation.

Firstly, we show the results of the evaluation of the greedy performance. We compute this as the expected return obtained when the agent acts using the greedy policy w.r.t. to the estimated  $Q$ -function. In the case of GVT, we take the mean of the posterior as the estimated  $Q$ -function, while for MGVT, we compute the greedy expected return under all components' means and average them.

In Figure 2a we show the results when transferring from the prior with both doors moving. It is easily noticeable that both GVT and MGVT perform much better in comparison with the no transfer performance and that the mean behavior of our posterior distribution, indeed, converges to the actual optimal solution.

In Figure 2b we show the results when transferring from the prior with only one door moving. From this we can clearly appreciate that MGVT is able to quickly converge to the optimal solution in this more complicated setting, whereas GVT fails to adapt as consistently, thus the higher variance and distance to the optimal value. In this scenario, using a Gaussian to model the prior over-constrains the algorithm to stay close to part of the function space that cannot solve optimally the target tasks sampled from the modified distribution.

Furthermore, we investigate the exploratory behavior induced by our transfer algorithms and how they differ between each other and in comparison with  $\epsilon$ -greedy exploration. In Fig. 3, we show the results of running the no transfer (NT) algorithm, GVT and 1-MGVT for 2000 iterations and we represent as a scatter plot the positions visited by the agent.

Observing Fig. 3a, it is possible to understand the difference between the  $\epsilon$ -greedy exploration and the resulting behavior from GVT. It is noticeable that NT is not capable to lead the agent to the goal within the given iterations as most of the states visited are sparse within the first room, whereas GVT is able to concentrate more of its effort in looking for the door around the middle of the wall. After finding it, within the second room, the positions concentrate in the path leading to the goal given that the need for the exploration is less. This is not surprising as the value function should be equal for all tasks after crossing the door.

In the other case, we have Fig. 3b that shows a similar situation to that of GVT, but it is quite interesting to notice how sparser the exploration of 1-MGVT is with respect to GVT. Indeed, 1-MGVT is able to actually explore the right part of the first room within these iterations, which might

653 be seen as the result of the prior model being able to capture more information than the Gaussian;  
 654 hence, the higher speed-up in convergence and robustness to changes in the distribution from which  
 655 target tasks are drawn. Indeed, as 1-MGVT is able to allow for more flexible exploration, it is capable  
 656 to discover how to best solve the task much faster than GVT.

657 In Figure 4 we present the expected return as a function of the number of source tasks used for  
 658 GVT and MGVT. In particular, we show the resulting curves after 1000 iterations in Figure 4a and  
 659 after 1950 iterations in Figure 4b. It is interesting to notice the difference on performance between  
 660 MGVT and GVT whenever there is a small number of source tasks. MGVT clearly provides faster  
 661 adaptation in the presence of low prior knowledge as it can be discern from the gap seen in nearly  
 662 a 1000 iterations between the plots. It is as expected because approximating the prior Gaussian  
 663 distribution using maximum likelihood estimations with a low number of samples does not provide  
 664 enough precision. As the number of source tasks increases, as seen more clearly from Figure 4b, the  
 665 performances become closer for this environment.

666 Finally, we analyze the transfer performance as a function of how likely the target task is according  
 667 to the prior. We consider a two-room version of the environment of Figure 1. Unlike before, we  
 668 generate tasks by sampling the door position from a Gaussian with mean 5 and standard deviation  
 669 1.8, so that tasks with the door near the sides are very unlikely. Figure 5 shows the performance  
 670 reached by GVT and 1-MGVT at fixed iterations as a function of how likely the target task is  
 671 according to such distribution. As expected, GVT achieves poor performance on very unlikely  
 672 tasks, even after many iterations. In fact, estimating a single Gaussian distribution definitely entails  
 673 some information loss, especially about the unlikely tasks. On the other hand, MGVT keeps such  
 674 information and, consequently, performs much better. Perhaps not surprisingly, MGVT reaches the  
 675 optimal performance in  $4k$  iterations no matter what task is being solved.

## 676 C.2 Classic Control

### 677 C.2.1 Cartpole

678 For this environment we generate tasks by uniformly sampling the cart mass in the range  $[0.5, 1.5]$ ,  
 679 the pole mass in  $[0.1, 0.3]$  and the pole length in  $[0.2, 1.0]$ .

680 During the training of the source tasks, we use a *batch size* of 150 and a *buffer size* of 50000.  
 681 Specifically, for DDQN we use a *target update frequency* of 500, *exploration fraction* of 0.35 and a  
 682 learning rate  $\alpha = 0.001$ . We use a Multilayer Perceptron (MLP) with ReLU as activation function  
 683 and a single hidden layer of 32 neurons.

684 For the transfer experiments, we set the *batch size* to 500, the number of *weights* sampled to  
 685 approximate the expected TD error to 5,  $\lambda = 0.001$  and  $\psi = 0.5$ . We use  $\alpha_\mu = 0.001$  as the learning  
 686 rate for the mean of the Gaussian posterior. For its the Cholesky factor L we use  $\alpha_L = 0.0001$  and  
 687 set the limit that the minimum eigenvalue may reach to  $\sigma_{min}^2 = 0.0001$ . Additionally, for MGVT  
 688 we set the variance of the prior components  $\sigma_p^2 = 10^{-5}$  and leave the learning rates of the posterior  
 689 components' means and Cholesky factor the same as GVT.

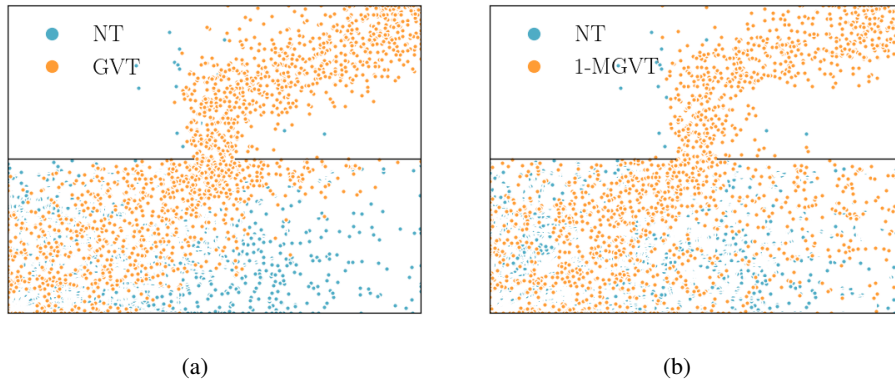


Figure 3: Two-Rooms Problem: (a)  $\epsilon$ -greedy vs. GVT, and (b)  $\epsilon$ -greedy vs. 1-MGVT

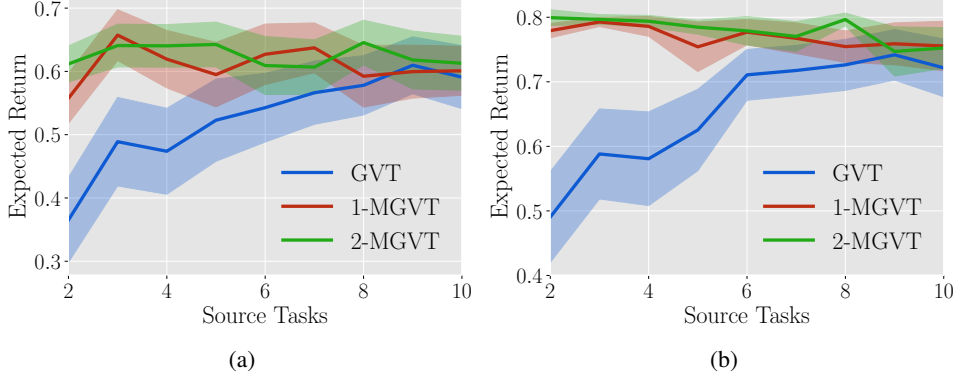


Figure 4: Expected return w.r.t. to the number of source tasks (a) 1000 iterations, (b) 1950 iterations

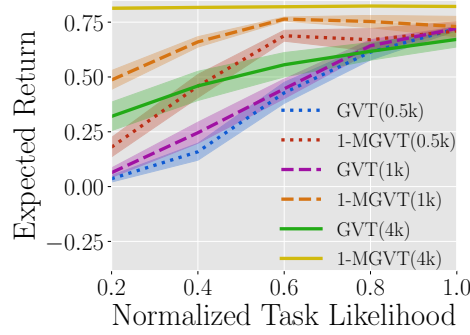


Figure 5: Expected return as a function of the task likelihood.

In Figure 6a, we show the greedy performance of DDQN, GVT, 1-MGVT and 3-MGVT, as in the previous section. Once again, we can see how the transfer methods provide a significant jump-start w.r.t. the DDQN evaluation.

### C.2.2 Mountain Car

We generate tasks sampling uniformly the base speed of the actions in the range  $[0.001, 0.0015]$ .

For the sources, we train the tasks using DDQN with a *target update frequency* of 500, a *batch size* of 32, a *buffer size* of 50000 and learning rate  $\alpha = 0.001$ . Moreover, we set the *exploration fraction* to 0.15. We use an MLP with single hidden layer of 64 neurons with ReLU activation function.

For the transfer experiments, we set the *batch size* to 500, and use 10 *weights* to approximate the expected TD error,  $\lambda = 10^{-5}$  and  $\psi = 0.5$ . For the learning rates, we use  $\alpha_\mu = 0.001$  for the means of the Gaussians. In the case of the Cholesky factors  $L$ , we use  $\alpha_L = 0.0001$  and allow the eigenvalues to reach a minimum value of  $\sigma_{min}^2 = 0.0001$ . In the case of MGVT, additionally, we set the prior covariance to be  $\sigma_p^2 = 10^{-5}$ .

In Figure 6b, we show the greedy performance obtained during the executions, as in the previous sections. In this plot, we can observe directly how fast is MGVT able to converge its mean to the actual optimal performance and that even if GVT struggles to learn in comparison with MGVT, still provides a clear advantage w.r.t. DDQN. Also, it is worth noting that the variances seen are result that different speeds of the car result in different time to reach the goal and, thus, different optimal performances.

Clearly, this variation in the values among tasks allows MGVT to excel w.r.t. to the other methods. The richer prior of MGVT allows to exploit efficiently the previous knowledge on the possible values attainable from the tasks and, therefore, quickly resolve how to best optimize for the target. This fact is a clear contrast with GVT as the search is more likely constrained to the mean values and moving away from that is, in fact, slower.

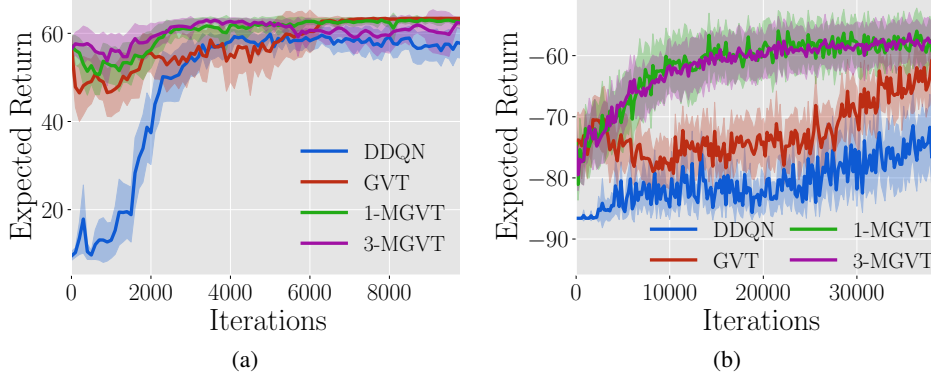


Figure 6: (a) Cartpole: Expected Return w.r.t. Greedy Policy, (b) Mountain Car: Expected Return w.r.t. Greedy Policy

### 714 C.3 Maze Navigation

715 The mazes adopted in the experiments of Section 5 are shown in Fig. 7. Our 20 mazes have varying  
 716 degree of difficulty and are designed to hold few similarities that would be useful for transferring.  
 717 Moreover, we ensure 4 groups of mazes that are characterized by their goal position.

718 For the experiments we use as an approximator an MLP with two hidden layers of 32 neurons with  
 719 ReLU activation functions. For training the sources we use a DDQN with a *batch size* of 70, a  
 720 *buffer size* of 10000 and a *target update frequency* of 100, setting the *exploration fraction* to 0.1 and  
 721 learning rate to  $\alpha = 0.001$ .

722 In the transfer experiments we use  $\psi = 0.5$ , a *batch size* of 50, a *buffer size* of 50000 and use 10  
 723 sampled *weights* from the posterior to approximate the TD error. Moreover, we use  $\lambda = 10^{-6}$ . For  
 724 GVT, in particular, we use  $\alpha_\mu = 0.001$ ,  $\alpha_L = 10^{-7}$ , and set the minimum value reachable by its  
 725 eigenvalues to be  $\sigma_{min} = 0.0001$ . In the case of MGVT, we set  $\alpha_\mu = 0.001$  and  $\alpha_L = 10^{-6}$ . Finally,  
 726 we use  $\sigma_p^2 = 10^{-5}$  as the prior bandwidth.

727 Hereafter, we present additional results of transferring from 5 source tasks to the mazes shown in  
 728 Figure 7a, Figure 7g, Figure 7n, and Figure 7i using both GVT and MGVT. We show both the greedy  
 729 and online performances. All the curves result from averaging 20 independent runs and randomly  
 730 sampled sources.

731 In Figure 8, 9, 10, and 11 we can appreciate that in this more complex transfer setting, MGVT is able  
 732 to provide significant speed-up in a consistent manner in this subset of mazes. It is also noticeable the  
 733 bad performance obtained with GVT in all cases. The Gaussian prior model clearly fails to capture  
 734 enough information to transfer in this setting that result in a negative transfer effect.

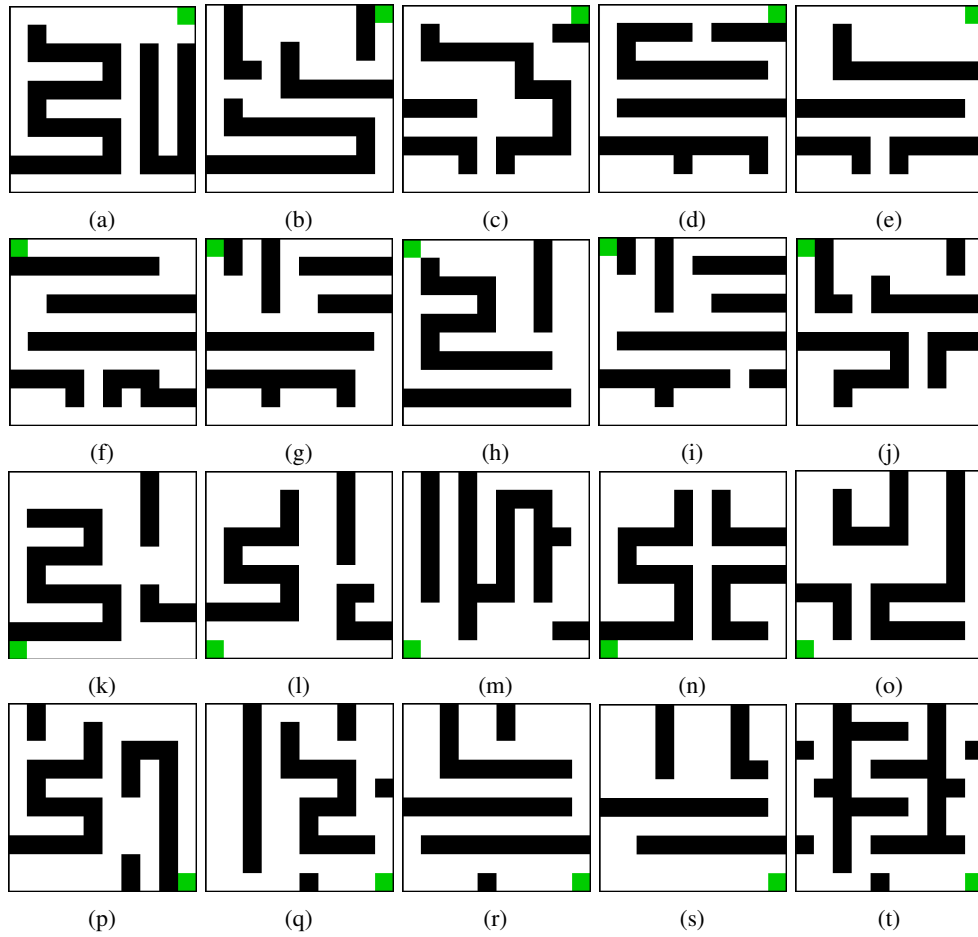
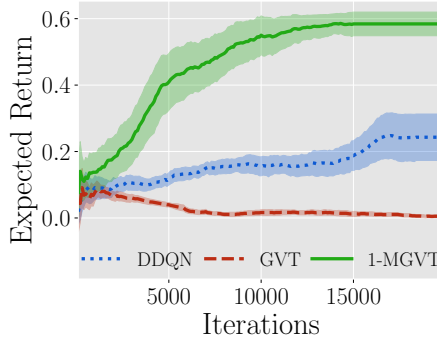
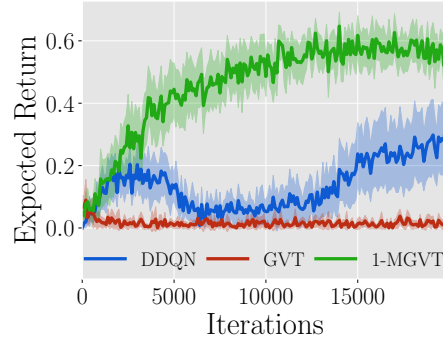


Figure 7: Set of mazes for the Maze Navigation task



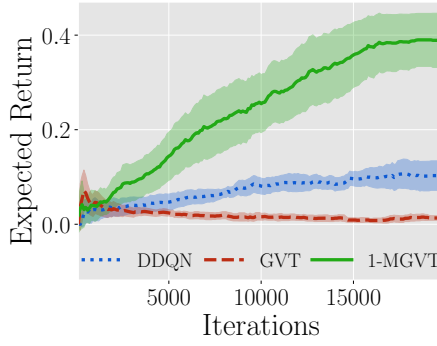


(a) Expected Return during learning

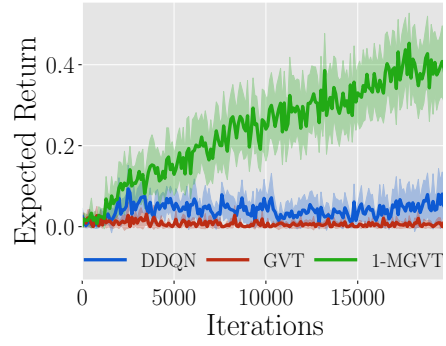


(b) Expected Return w.r.t. greedy policy

Figure 8: Performance in Maze 7a

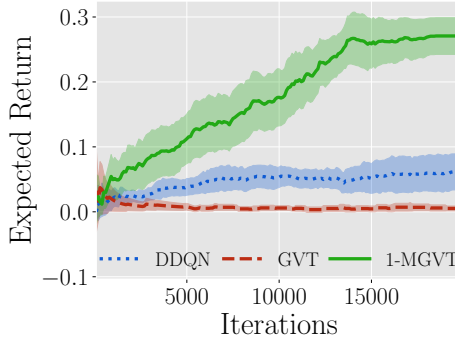


(a) Expected Return during learning

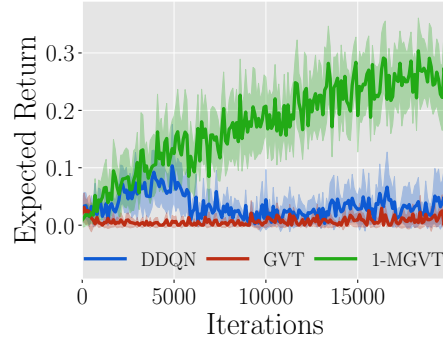


(b) Expected Return w.r.t. greedy policy

Figure 9: Performance in Maze 7g

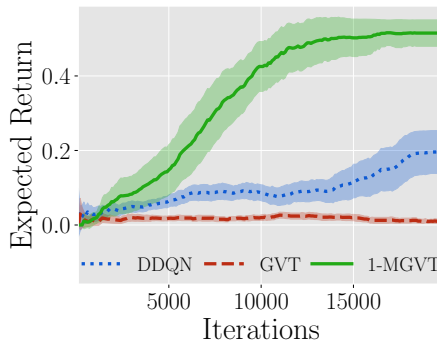


(a) Expected Return during learning

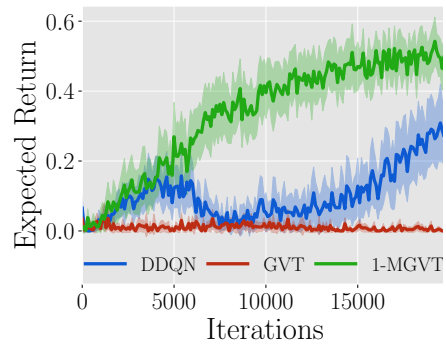


(b) Expected Return w.r.t. greedy policy

Figure 10: Performance in Maze 7n



(a) Expected Return during learning



(b) Expected Return w.r.t. greedy policy

Figure 11: Performance in Maze 7i