
Variational Approximations for Transfer in Reinforcement Learning

Anonymous Author(s)

Affiliation

Address

email

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Recent advancements have allowed reinforcement learning (RL) [29] to achieve impressive results in a wide variety of complex tasks, ranging from Atari [22] through the game of Go [28] to the control of sophisticated robotics systems [14, 20, 19]. The main limitation is that these RL algorithms still require an enormous amount of experience samples before successfully learning such complicated tasks. One of the most promising solutions to reduce the need of samples is transfer learning, which focuses on reusing past knowledge available to the agent in order to reduce the sample-complexity for learning new tasks. In the typical settings of transfer in RL [31], the agent is assumed to have already solved a set of *source tasks* generated from some unknown distribution. Then, given a *target task* (which is drawn from the same distribution, or a slightly different one), the agent can rely on knowledge from the source tasks to speed up the learning process. This reuse of knowledge constitutes a significant advantage over plain RL, where the agent learns each new task from scratch independently of any previous learning experience. Several algorithms have been proposed in the literature to transfer different elements involved in the learning process: experience samples [18, 30], policies/options [10, 16], rewards [15], features [4], parameters [9, 13], and so on. We refer the reader to [31] for a thorough survey on transfer in RL.

Under the assumption that tasks follow a specific distribution, an intuitive choice for designing a transfer algorithm is to attempt at characterizing the uncertainty over the target task. Then, an ideal algorithm would leverage prior knowledge from the source tasks to interact with the target task to reduce the uncertainty as quickly as possible. This simple intuition makes Bayesian methods appealing approaches for transfer in RL, and many previous works have been proposed in this direction. In [33], the authors assume tasks share similarities in their dynamics and rewards and propose a hierarchical Bayesian model for the distribution of these two elements. Similarly, in [17], the authors assume that tasks are similar in their value functions and design a different hierarchical Bayesian model for transferring such information. More recently, [9], and its extension [13], consider tasks whose dynamics are governed by some hidden parameters, and propose efficient Bayesian models for quickly learning such parameters in new tasks. However, most of these algorithms require specific, and sometimes restrictive, assumptions (e.g., on the distributions involved or the function 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.

probabilistic?
distribu-
tional?

Similarly to [17], 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 [23], 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 give also 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}_τ 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 $s_1 \sim \mathcal{P}(\cdot|s_0, a_0)$, 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}_τ , $Q_\tau^*(s, a)$, as the expected return obtained by taking action a in state s and following an optimal policy thereafter. Then, an optimal policy π_τ^* 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., [24]) that Q_τ^* is the unique fixed-point of the optimal Bellman operator T_τ 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 τ when 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 ν over $\mathcal{S} \times \mathcal{A}$, a sound objective is to directly minimize the squared Bellman error of Q_w under ν , 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., [21]). 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 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 it can be used for any choice of prior and posterior distributions. 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 Sec. 3.4.

86 3.1 Algorithm

87 We begin with a simple consideration: the distribution \mathcal{D} over tasks clearly induces a distribution
 88 over optimal Q -functions. Moreover, for any MDP, learning its optimal Q -function is sufficient for
 89 solving the problem. Thus, one can safely replace the distribution over tasks with the distribution
 90 over their optimal value functions. In our parametric settings, we reduce the latter to a distribution
 91 $p(\mathbf{w})$ over weights.

Should we just make the assumption that Q -functions share knowledge in their weights?

92 Assume, for the moment, that we know the distribution $p(\mathbf{w})$ and consider a dataset $D =$
 93 $\{(s_i, a_i, s'_i, r_i) \mid i = 1, 2, \dots, N\}$ of samples from some task τ that we want to solve. Then, the
 94 posterior distribution over weights given such dataset can be computed by applying Bayes theorem
 95 as $p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$. Unfortunately, this cannot be directly used in practice since we do not
 96 have a model of the likelihood $p(D|\mathbf{w})$. In such case, it is very common to make strong assumptions
 97 on the MDPs or the Q -functions so as to get tractable posteriors. However, in our transfer settings
 98 all distributions involved depend on the family of tasks under consideration, and making such as-
 99 sumptions is likely to limit the applicability to specific problems. Thus, we take a different approach
 100 to derive a more general, but still well-grounded, solution. Recall that our final goal is move all
 101 probability mass over the weights minimizing some empirical loss measure, which in our case is the
 102 TD error $\|B(\mathbf{w})\|_D^2$. Then, given a prior $p(\mathbf{w})$, we know from PAC-Bayesian theory that the optimal
 103 Gibbs posterior q takes the form (e.g., [8]):

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

104 for some parameter $\Lambda > 0$. Since Λ is typically chosen to increase with the number of samples
 105 N , in the remaining we set it to $\lambda^{-1}N$, for some constant $\lambda > 0$. Notice that, whenever the term
 106 $e^{-\Lambda\|B(\mathbf{w})\|_D^2}$ can be interpreted as the actual likelihood of D , q becomes a classic Bayesian posterior.
 107 Although we now have an appealing distribution, the integral at the denominator of (1) is intractable
 108 to compute even for simple Q -function models. Thus, we propose a variational approximation q_ξ by
 109 considering a simpler family of distributions parameterized by $\xi \in \Xi$. Then, our problem reduces to
 110 finding the variational parameters ξ such that q_ξ minimizes the Kullback-Leibler (KL) divergence
 111 w.r.t. the Gibbs posterior q . From the theory of variational inference (e.g., [5]), this can be shown to
 112 be equivalent to minimizing the well-known (negative) *evidence lower bound* (ELBO):

Maybe say how this is obtained

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

113 Intuitively, the approximate posterior trades-off between placing probability mass over those weights
 114 \mathbf{w} that have low expected TD error (first term), and staying close to the prior distribution (second
 115 term). Assuming that we are able to compute the gradients of (2) w.r.t. the variational parameters ξ ,
 116 our objective can be easily optimized using any stochastic optimization algorithm.

117 We now highlight our general transfer procedure in Alg. 1, while deferring a description of specific
 118 choices for the distributions involved to the next two sections. Given a set of weights \mathcal{W}_s from
 119 the source tasks' optimal Q -functions, we start by estimating the prior distribution (line 1) and we
 120 initialize the variational parameters by minimizing the KL divergence w.r.t. such distribution¹ (line
 121 2). Then, at each time step of interaction, we re-sample the weights from the current approximate
 122 posterior and act greedily w.r.t. the corresponding Q -function (lines 7,8). After collecting and storing
 123 the new experience (lines 9-10), we draw a mini-batch of samples from the replay buffer (line 11), use
 124 this to estimate the objective function gradient (line 12), and, finally, update the variational parameters
 125 (line 13).

126 The key property of our approach is the weight resampling at line 7. This resembles the well-known
 127 Thompson sampling approach adopted in multi-armed bandits [7] and closely relates to the recent
 128 value function randomization [23]. In some sense, at each time we guess what is the task we are
 129 trying to solve based on our current belief and we act as if such guess were actually true. This allows
 130 an efficient adaptive exploration of the target task. Intuitively, during the first steps of interaction,
 131 the agent is very uncertain about the current task and such uncertainty induces stochasticity in the
 132 chosen actions, allowing rather informed exploration to take place. Consider, for instance, that actions

¹If the prior and approximate posterior were in the same family of distributions we could simply set ξ to the prior parameters. However, we are not making this assumption at this point.

Algorithm 1 Variational Transfer

Require: Target task τ , 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:  $s_0 \sim p_0^{(\tau)}$ 
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  $r_h = \mathcal{R}^{(\tau)}(s_h, a_h)$ 
10:    Add sample to the replay buffer:  $D \leftarrow D \cup \langle s_h, a_h, r_h, s_{h+1} \rangle$ 
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
```

that are bad on average for all tasks are very unlikely to be sampled, while this cannot happen in uninformed exploration strategies, like ϵ -greedy, before learning takes place. As the learning process goes on, the algorithm will quickly figure out which task is being solved, thus moving all 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 ξ , 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 $\Xi = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\mu} \in \mathbb{R}^K, \boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}, \boldsymbol{\Sigma} \succ 0\}$. 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., [12, 25]). Although these results are well-known in the literature, we report them in App. ?? to have a more self-contained description.

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, we believe that such assumption almost never holds in practice. In fact, even for families of tasks in which the reward and transition models are Gaussian, the Q -values might be far from it. 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 simple and efficient enough to be applied in practice.

Given the source tasks' weights \mathcal{W}_s , we model our estimated prior as a mixture of Gaussians with one equally weighted isotropic Gaussian 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 resembles a kernel density estimator [27] and, due to its nonparametric nature, it allows capturing arbitrarily complex distribution. Consistently with the prior, we model our approximate posterior as a mixture of Gaussians. However, we allow a different number of components (typically much less than the prior's) and we adopt full covariances instead of only diagonals, so that our posterior has the potential to match complex distributions with less components. 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, we learn Cholesky factors instead of full covariances.

Although this new model has the potential to capture much more complex distributions, it poses a major complication: the KL divergence between two mixture of Gaussians has no closed-form expression. To solve this issue, we rely on an upper bound to such quantity, so that negative ELBO still upper bounds the KL between the approximate and true posterior. Among the many upper bounds available in the literature, we adopt the one proposed in [11], which we report here for the sake of completeness. We refer the reader to the original paper for the proof.

Theorem 1 ([11]). *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 distributions, 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 similarly for q . Introduce two vectors $\chi^{(1)}$ and $\chi^{(2)}$ such that $c_i^{(p)} = \sum_j \chi_{j,i}^{(2)}$ and $c_j^{(q)} = \sum_i \chi_{i,j}^{(1)}$. Then:*

$$KL(p||q) \leq KL(\chi^{(2)}||\chi^{(1)}) + \sum_{i,j} \chi_{j,i}^{(2)} KL(f_i^{(p)}||f_j^{(q)}) \quad (3)$$

Our new algorithm replaces the KL with the above-mentioned upper bound. Each time we require its value, we have to recompute the parameters $\chi^{(1)}$ and $\chi^{(2)}$ that tighten the bound. As shown in [11], this can be achieved by a simple fixed-point procedure. Finally, both terms in the objective are now linear combinations of functions of the variational parameters of different components, and their gradients are easily derived from the ones of the Gaussian case. We report the derivation in App. ??.

3.4 Optimizing the TD error

From Sections 3.2 and 3.3, we know that differentiating the negative ELBO \mathcal{L} w.r.t. ξ requires differentiating $\|B(\mathbf{w})\|_D^2$ w.r.t. \mathbf{w} . Unfortunately, the TD error is well-known to be non-differentiable due to the presence of the max operator. This rarely represents a problem since typical value-based algorithms are actually semi-gradient methods, i.e., they do not differentiate the targets (see, e.g., Chapter 11 of [29]). However, our transfer settings are rather different than common RL. In fact, our algorithm is likely to start from Q -functions that are very close to an optimum, and the only thing that needs to be done is to adapt the weights in a direction of lower error (i.e., higher likelihood) so as to quickly converge to the solution of the task that is being solved. Unfortunately, this property cannot be guaranteed 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 [26, 32]. This is obviously an undesirable property in our case, as we only aim at adapting already good solutions. Thus, we consider using a residual gradient algorithm [3]. In order to differentiate the targets, we replace the optimal Bellman operator with the mellow Bellman operator introduced in [1], which adopts a softened version of max called *mellowmax*:

$$\text{mm}_a Q_{\mathbf{w}}(s, a) = \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_a e^{\kappa Q_{\mathbf{w}}(s, a)} \quad (4)$$

where κ 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 that make it suitable for our settings: (i) it converges to the maximum as $\kappa \rightarrow \infty$, (ii) it has a unique fixed point, and (iii) it is *differentiable*. Denoting by $\tilde{B}(\mathbf{w}) = \tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}}$ the Bellman residual w.r.t. the mellow Bellman operator \tilde{T} , we have that the corresponding TD error, $\|\tilde{B}(\mathbf{w})\|_D^2$, is now differentiable with respect to \mathbf{w} . Although residual

algorithms have guaranteed convergence, they are typically much slower than their semi-gradient counterpart. [3] proposed to project the gradient in a direction that achieves higher learning speed, while preserving convergence. This can be easily done by including a parameter $\psi \in [0, 1]$ in the TD error gradient such that:

$$\nabla_{\mathbf{w}} \left\| \tilde{B}(\mathbf{w}) \right\|_D^2 = \frac{2}{N} \sum_{i=1}^N b_i(\mathbf{w}) \left(\gamma \psi \nabla_{\mathbf{w}} \text{mm}_{a'} Q_{\mathbf{w}}(s'_i, a') - \nabla_{\mathbf{w}} Q_{\mathbf{w}}(s_i, a_i) \right) \quad (5)$$

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 ψ 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 converge).

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 [1] that the mellow Bellman operator is a non-expansion under the L_∞ -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 corresponding operator also monotonically converged to the optimal one. We confirm that this property actually holds in the following theorem.

Theorem 2. *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 [1]). Let ν be a probability measure over the state-action space. Then, for any $p \geq 1$:*

$$\left\| Q^* - \tilde{Q} \right\|_{\nu, p}^p \leq \frac{2\gamma R_{max}}{(1-\gamma)^2} \left\| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g}} \right\|_{\nu, p}^p \quad (6)$$

Correct measure on the rhs

The proof is provided in App. ?? . As expected, \tilde{Q} converges to Q^* exponentially fast when either κ (equivalently, β_κ) increases, or the action gaps are enlarged. 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 restrict to the case of linear approximators with Gaussian prior and posterior. We assume Alg. 1 is used with a finite dataset and provide a finite-sample analysis bounding the distance between the fixed-point of the mellow Bellman operator and Q -functions sampled from the variational distribution minimizing the objective (2). Our main results is given in the following theorem.

Theorem 3. *Fix a target task τ a let \tilde{Q} be the fixed-point of the corresponding mellow Bellman operator. Assume linearly parameterized value functions $Q_{\mathbf{w}}(s, a) = \mathbf{w}^T \phi(s, a)$ with bounded weights $\mathbf{w} \leq w_{max}$ and uniformly bounded features $\phi(s, a) \leq \phi_{max}$. Consider the Gaussian version of Alg. 1 with prior $p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and denote by $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ the variational parameter minimizing the objective of Eq. 2 on a dataset D of N samples. Let ν be a probability measure over $\mathcal{S} \times \mathcal{A}$ and $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \left\| \tilde{B}(\mathbf{w}) \right\|_\nu^2$. Define $v(\mathbf{w}^*) \triangleq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \mathbf{I})} [v(\mathbf{w})]$, with $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [\text{Var}_{\mathcal{P}} [b(\mathbf{w})]]$. Then, there exist constants c_1, c_2, c_3 such that, with probability at least $1 - 2\delta$ over the choice of weights $\mathbf{w} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and dataset D :*

$$\left\| Q_{\mathbf{w}} - \tilde{Q} \right\|_\nu^2 \leq \frac{1}{(1-\gamma)\delta} \left(2 \left\| \tilde{B}(\mathbf{w}^*) \right\|_\nu^2 + \frac{v(\mathbf{w}^*) + c_3 \sqrt{\log \frac{2}{\delta}}}{\sqrt{N}} + \frac{c_2 + \lambda \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}}{N} + \frac{c_1}{N^2} \right) \quad (7)$$

We refer the reader to App. ?? for the proof and a specific definition of the three constants. Four main terms constitute our bound: the approximation error due to the limited hypothesis space (first term), the variance (second term), the distance to the prior (third term), and a constant term decaying very

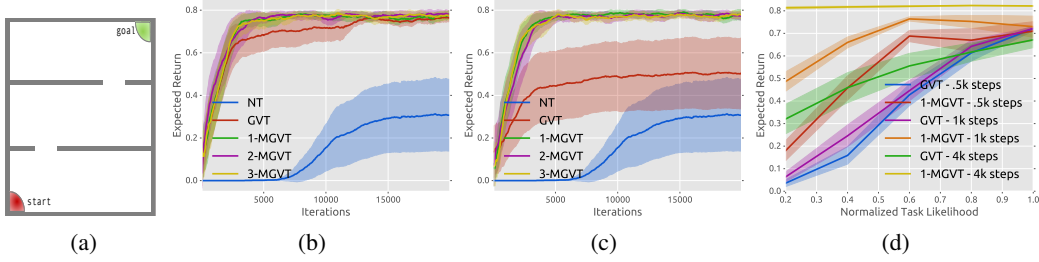


Figure 1: (a) the rooms environment, (b) transfer from 10 source tasks with both doors moving, (c) transfer from 10 source tasks with only one door moving, and (d) transfer performance as a function of how likely the target task is according to the prior.

fast (as $\mathcal{O}(\frac{1}{N^2})$). Intuitively, our bound demonstrates that, if one starts very close the prior, the only error incurred, besides the irreducible approximation error, is due to the variance of the estimates, and, thus, the algorithm is expected to achieve good performance rather quickly. Furthermore, as $N \rightarrow \infty$ the only error term remaining is the irreducible approximation error. Notice that the variance term $v(w^*)$ is due to the fact that we minimize a biased estimator of the Bellman error. As mentioned in the preliminaries, if double sampling of the next state is possible (e.g., in simulation), such term would disappear. Finally, we want to point out that the proof of this theorem relies on a very general result (Lemma 3 in App. ??) that does not require any specific choice of distribution of approximator. Leveraging on such results, it is straightforward to provide a finite-sample analysis of the mixture version of our algorithm. For the sake of completeness, we report the derivation in App. ??.

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 ϵ -greedy exploration. To the best of our knowledge, no existing transfer algorithm is directly comparable to our approach from an experimental perspective. Thus, we provide a discussion of related works in the next section.

Any better motivation?

5.1 The Rooms Problem

We consider an agent navigating in the rooms environment depicted in Fig. ?. 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 positions are unknown to the agent. The state-space is modeled as a 10×10 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 with 0.2 standard deviation. 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.

Since the agent does not know where the doors are located in advance and receives only very sparse feedback, it must efficiently explore the environment to figure out (i) their positions, and (ii) how to reach the goal. While this might be a complicated problem for plain RL, our transfer algorithm should be able to quickly figure out the door positions. In fact, notice that, although different, the optimal Q -functions for all tasks share some similarities. For example, once the agent has passed the last door before the goal, the Q -values are exactly the same in all tasks. This does not hold for positions nearby the start state. However, it is clear that there should be a preference over actions up and right, rather than down and left (which are worse in all tasks).

Good candidate to be removed

In order to prove that our guesses are correct, we generate a set of 50 source tasks for the three-room environment of Fig. ? by sampling both door positions uniformly in the allowed space, and solve all of them by directly minimizing the TD error as presented in Sec. 3.4. Then, we use our algorithms to transfer from 10 source tasks sampled from the previously generated set. The average return over

the last 50 learning episodes as a function of the number of iterations is shown in Fig. ?? . Each curve is the result of 20 independent runs, each resampling the target and source tasks, with 95% confidence intervals. Further details on the parameters adopted in this experiment are given in App. ?? . As expected, the no-transfer (NT) algorithm fails at learning the task in so few iterations due to the limited exploration provided by an ϵ -greedy policy. On the other hand, all our algorithms achieve a significant speed-up and are able to converge to the optimal performance in few iterations, with GVT being slightly slower. Interestingly, we notice that there is no advantage in adopting more than 1 component for the posterior in MGVT. This is intuitive since, when the algorithm quickly figures out which task is being solved, it will move all components in the same direction.

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

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

5.2 Classic Control

We now consider two well-known classic control environments: Cartpole and Mountain Car [29]. For both, we generate 20 source tasks by uniformly sampling their physical parameters (cart mass, pole mass, pole length for Cartpole and car speed for Mountain Car). We parameterize Q -functions using neural networks with 1-layer of 32 hidden units for Cartpole and 64 for Mountain Car. A better description of these two environments and their parameters is given in App. ?? . In this experiment, we use a Double Deep Q-Network [32] to provide a stronger no-transfer baseline for comparison. The results (same settings of Sec. 5.1) are shown in Fig. ?? and ?? . For Cartpole (Fig. ??), all transfer algorithms are almost zero-shot. This is expected since, although we vary the system parameters in a wide range, the optimal Q -values of states near the balanced position are almost the same for all tasks. On the contrary, in Mountain Car (Fig. ??) the optimal Q -functions become really different when changing the car speed. This causes GVT to struggle to solve the target task, while 1-MGVT achieves a good jump-start and convergences in less iterations.

5.3 Maze Navigation

For this setting we propose a maze environment, such as those in Fig. ?? , for a robotic agent to navigate. We created a set of 20 mazes—see App ??—of size 10×10 with continuous space within which we ensured four groups: each group with the goal position in one of the corners of the maze. In order to simulate a robotic agent, we set an action space that allows to control the linear and angular motion by a fixed distance of 0.5 (*move forward*) and a fixed angle of $\pm 22.5^\circ$ (*rotate counter- and clockwise*). Moreover, we enhanced the state space to include, in addition to the absolute position and orientation, a set of distance measurements to the nearest obstacles from the agent’s perspective with maximum range of 2 and a field of vision of 9 equally-spaced directions in $[-90, 90^\circ]$ relative to its orientation. In this way, we provide the agent with information that a robot would normally collect from LIDARs, sonars and other sensors alike. Finally, in the same field of perception, we provide a boolean vector indicating whether the goal is within the range of observation. Regarding rewards, the

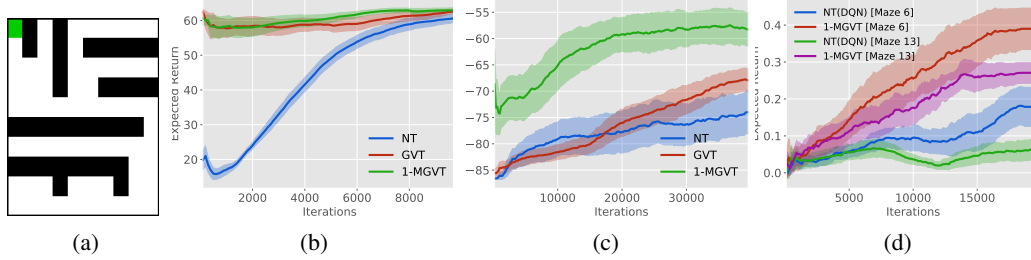


Figure 2: (a) Sample Maze (Maze 6), (b) transfer from 10 sources in Cartpole, (c) transfer from 10 source tasks in Mountain Car, and (d) transfer from 10 sources in Mazes 6 and 13—see App. ??.

agent gets 1 whenever it enters the goal area and 0 in any other situation and the MDP has a discount factor of $\gamma = 0.99$.

The main motivation for this set-up is that a robot is able to exploit locally-sensed information in addition to its estimation of the actual position—as opposed to the Rooms problem presented above—and it is such information that the robotic agent could learn to use and, also, exploit in different maze configurations leveraging its previous experience. Moreover, simple variations in the maze structure could, potentially, produce big differences in the value function with respect to the ones seen in previous tasks which makes the setting suitable to assess how robust our algorithm is to negative transfer.

The mazes were solved using Deep Q Networks (DQN) [17] with an MLP with hidden layers 32×32 and, in order to perform the knowledge transfer, the target maze would be excluded from the source tasks given to the algorithm. Finally, the 1-MGVT algorithm was used to solve the target tasks—Mazes 6 and 13. As before, 20 independent runs with randomized sources to transfer are averaged and shown in Fig. ??, in this case 5 source tasks were used for each trial. While these mazes pose a challenge for our algorithm, we can see that 1-MGVT is capable to still capture useful information from the sources and adapt to the target mazes which is demonstrated by the increased slope in the learning curve.

cite DQN?

actually, decide how to improve these results

6 Related Works

Our approach is mostly related to [17]. Although we both assume the tasks to share similarities in their value functions, [17] 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, [17] propose a Dirichlet process model for the case where weights cluster into different classes, which relates to our mixture formulation and proves again the importance of capturing more complicated task distributions. In [33], the authors propose a hierarchical Bayesian model for the distribution over tasks. Differently from our approach and [17], 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, [33] 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 [9], 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, [13] extended this approach by replacing GPs with Bayesian neural networks, so as to obtain a more scalable approach.

In the RL community, our approach is related to value function randomization[23], which extends the well-known LSTD [6] by adopting Bayesian linear regression to model the uncertainty over the predicted value function weights, and use that to perform a form of Thompson sampling. Such approach was recently extended by [2], where the approximator is replaced by a Bayesian neural network, leading to an algorithm capable of solving much more complicated problems. Both these algorithms rely on the Gaussian assumption and, since they work in plain RL settings, have no informative prior available. On the other hand, our variational approximation allows more complex

More comments about HiP-MDPs? Should we discuss more related works?

distributions (e.g., mixtures) to be adopted, while knowledge from the source tasks allows us to learn very informative priors.

7 Conclusion

References

- [1] Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pages 243–252, 2017.
- [2] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. *arXiv preprint arXiv:1802.04412*, 2018.
- [3] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.
- [4] 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.
- [5] 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.
- [6] Justin A Boyan. Least-squares temporal difference learning.
- [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] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] Jens Kober and Jan R Peters. Policy search for motor primitives in robotics. In *Advances in neural information processing systems*, pages 849–856, 2009.
- [15] 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.
- [16] George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning.
- [17] Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *ICML-27th International Conference on Machine Learning*, pages 599–606. Omnipress, 2010.
- [18] 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.
- [19] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

- 432 [20] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David
433 Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint*
434 *arXiv:1509.02971*, 2015.
- 435 [21] Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-
436 sample analysis of bellman residual minimization. In *Proceedings of 2nd Asian Conference on Machine*
437 *Learning*, pages 299–314, 2010.
- 438 [22] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare,
439 Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through
440 deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- 441 [23] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value
442 functions. *arXiv preprint arXiv:1402.0635*, 2014.
- 443 [24] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley
444 & Sons, Inc., New York, NY, USA, 1994.
- 445 [25] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approxi-
446 mate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- 447 [26] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv*
448 *preprint arXiv:1511.05952*, 2015.
- 449 [27] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons,
450 2015.
- 451 [28] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian
452 Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go
453 with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- 454 [29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press
455 Cambridge, 1998.
- 456 [30] Matthew E Taylor, Nicholas K Jong, and Peter Stone. Transferring instances for model-based reinforcement
457 learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*,
458 pages 488–505. Springer, 2008.
- 459 [31] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey.
460 *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.
- 461 [32] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning.
462 2016.
- 463 [33] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a
464 hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*,
465 pages 1015–1022. ACM, 2007.

466 A Proofs

467 **Theorem 2.** Let Q^* be the fixed-point of the optimal Bellman operator T . Define the action-gap
 468 function $g(s)$ as the difference between the value of the best action and the second best action at
 469 each state s . Let \tilde{Q} be the fixed-point of the mellow Bellman operator \tilde{T} with parameter $\kappa > 0$ and
 470 denote by $\beta_\kappa > 0$ the inverse temperature of the induced Boltzmann distribution (as in [1]). Let ν be
 471 a probability measure over the state-action space. Then, for any $p \geq 1$:

$$\|Q^* - \tilde{Q}\|_{\nu,p}^p \leq \frac{2\gamma R_{max}}{(1-\gamma)^2} \left\| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g}} \right\|_{\nu,p}^p \quad (6)$$

Correct mea-
sure on the
rhs

472 *Proof.* We begin by noticing that:

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

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

$$\|Q^* - \tilde{Q}\|_{\nu,p}^p \leq \frac{1}{1-\gamma} \|TQ^* - \tilde{T}Q^*\|_{\nu,p}^p \quad (8)$$

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

$$\begin{aligned} |TQ^*(s, a) - \tilde{T}Q^*(s, a)| &= |R(s, a) + \gamma \mathbb{E} [\max_{a'} Q^*(s', a')] - R(s, a) - \gamma \mathbb{E} [\text{mm}_{a'} Q^*(s', a')]| \\ &= \gamma |\mathbb{E} [\max_{a'} Q^*(s', a')] - \mathbb{E} [\text{mm}_{a'} Q^*(s', a')]| \\ &\leq \gamma \mathbb{E} [|\max_{a'} Q^*(s', a') - \text{mm}_{a'} Q^*(s', a')|] \end{aligned} \quad (9)$$

478 Thus, bounding this quantity reduces to bounding $|\max_a Q^*(s, a) - \text{mm}_a Q^*(s, a)|$ point-wisely
 479 for any s . Recall that applying the mellow-max is equivalent to computing an expectation under a
 480 Boltzmann distribution with inverse temperature β_κ induced by κ [1]. Thus, we can write:

Cite MM

$$\begin{aligned} |\max_a Q^*(s, a) - \text{mm}_a Q^*(s, a)| &= \left| \sum_a \pi^*(a|s) Q^*(s, a) - \sum_a \pi_{\beta_\kappa}(a|s) Q^*(s, a) \right| \\ &= \left| \sum_a Q^*(s, a) (\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)) \right| \\ &\leq \sum_a |Q^*(s, a)| |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| \\ &\leq \frac{R_{max}}{1-\gamma} \sum_a |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| \end{aligned} \quad (10)$$

481 where π^* is the optimal (deterministic) policy w.r.t. Q^* and π_{β_κ} is the Boltzmann distribution induced
 482 by Q^* with inverse temperature β_κ :

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

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

$$\begin{aligned}
\sum_a |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| &= |\pi^*(a_1(s)|s) - \pi_{\beta_\kappa}(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi^*(a|s) - \pi_{\beta_\kappa}(a|s)| \\
&= |1 - \pi_{\beta_\kappa}(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi_{\beta_\kappa}(a|s)| \\
&= 2|1 - \pi_{\beta_\kappa}(a_1(s)|s)|
\end{aligned} \tag{11}$$

484 Finally, let us bound this last term:

$$\begin{aligned}
|1 - \pi_{\beta_\kappa}(a_1(s)|s)| &= \left| 1 - \frac{e^{\beta_\kappa Q^*(s, a_1(s))}}{\sum_{a'} e^{\beta_\kappa Q^*(s, a')}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa(Q^*(s, a_1(s)) - Q^*(s, a_2(s)))}}{\sum_{a'} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa g(s)}}{\sum_{a'} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta_\kappa g(s)}}{e^{\beta_\kappa g(s)} + \sum_{a' \neq a_1(s)} e^{\beta_\kappa(Q^*(s, a') - Q^*(s, a_2(s)))}} \right| \\
&\leq \left| 1 - \frac{e^{\beta_\kappa g(s)}}{e^{\beta_\kappa g(s)} + |\mathcal{A}|} \right| \\
&= \left| \frac{1}{1 + \frac{1}{|\mathcal{A}|} e^{\beta_\kappa g(s)}} \right|
\end{aligned} \tag{12}$$

485 Combining Eq. (10), (11), and (12), we obtain:

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

486 Finally, using Eq. (9) we get:

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

487 Taking the norm and plugging this into Eq. (8) concludes the proof. \square

488 **Lemma 1.** Let p and ν denote probability measures over Q -functions and state-action pairs, respectively. Assume Q^* is the unique fixed-point of the optimal Bellman operator T . Then, for any $\delta > 0$,
489 with probability at least $1 - \delta$ over the choice of a Q -function Q , the following holds:
490

$$\|Q - Q^*\|_\nu^2 \leq \frac{\mathbb{E}_p \left[\|B(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \tag{13}$$

491 *Proof.* First notice that:

$$\begin{aligned}
\|Q - Q^*\| &= \|Q + TQ - TQ - TQ^*\| \\
&\leq \|Q - TQ\| + \|TQ - TQ^*\| \\
&\leq \|Q - TQ\| + \gamma \|Q - Q^*\| \\
&= \|B(Q)\| + \gamma \|Q - Q^*\|
\end{aligned}$$

492 which implies that:

$$\|Q - Q^*\| \leq \frac{1}{1 - \gamma} \|B(Q)\|$$

493 Then we can write:

$$P(\|Q - Q^*\| > \epsilon) \leq P(\|B(Q)\| > \epsilon(1 - \gamma)) \leq \frac{\mathbb{E}_p \left[\|B(Q)\|_\nu^2 \right]}{(1 - \gamma)\epsilon}$$

494 Settings the right-hand side equal to δ and solving for ϵ concludes the proof. \square

495 **Corollary 1.** Let p and ν denote probability measures over Q -functions and state-action pairs,
 496 respectively. Assume \tilde{Q} is the unique fixed-point of the mellow Bellman operator \tilde{T} . Then, for any
 497 $\delta > 0$, with probability at least $1 - \delta$ over the choice of a Q -function Q , the following holds:

$$\|Q - \tilde{Q}\|_\nu^2 \leq \frac{\mathbb{E}_p \left[\|\tilde{B}(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \quad (14)$$

498 **Lemma 2.** Assume Q -functions belong to a parametric space of functions bounded by $\frac{R_{max}}{1-\gamma}$. Let p
 499 and q be arbitrary distributions over the parameter space \mathcal{W} , and ν be a probability measure over
 500 $\mathcal{S} \times \mathcal{A}$. Consider a dataset D of N samples and define $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [\text{Var}_{\mathcal{P}} [b(\mathbf{w})]]$. Then, for any
 501 $\delta > 0$, with probability at least $1 - \delta$, the following two inequalities hold simultaneously:

$$\mathbb{E}_q \left[\|B(\mathbf{w})\|_\nu^2 \right] \leq \mathbb{E}_q \left[\|B(\mathbf{w})\|_D^2 \right] - \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (15)$$

502

$$\mathbb{E}_q \left[\|B(\mathbf{w})\|_D^2 \right] \leq \mathbb{E}_q \left[\|B(\mathbf{w})\|_\nu^2 \right] + \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (16)$$

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

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

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

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

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

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

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

$$\left| \|B(\mathbf{w})\|_\nu^2 + v(\mathbf{w}) - \|B(\mathbf{w})\|_D^2 \right| \leq 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (17)$$

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

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

509 Thus, multiplying both sides of (17) by $\lambda^{-1}N$ and applying the change of measure inequality with
 510 $f(\mathbf{w}) = \lambda^{-1}N \left| \|B(\mathbf{w})\|_\nu^2 + v(\mathbf{w}) - \|B(\mathbf{w})\|_D^2 \right|$, we obtain:

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

Find a reference for this

where the second inequality holds since the right-hand side of (17) does not depend on \mathbf{w} . Finally, we can explicitly write:

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

from which the lemma follows straightforwardly. \square

Lemma 3. Let p be a prior distribution over the parameter space \mathcal{W} , and ν be a probability measure over $\mathcal{S} \times \mathcal{A}$. Assume $\hat{\xi}$ is the minimizer of $ELBO(\xi) = \mathbb{E}_{q_\xi} \left[\left\| B(\mathbf{w}) \right\|_D^2 \right] + \frac{\lambda}{N} KL(q_\xi||p)$ for a dataset D of N samples. Define $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [Var_{\mathcal{P}} [b(\mathbf{w})]]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$:

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

Proof. Let us use Lemma 2 for the specific choice $q = q_{\hat{\xi}}$. From Eq. (15), we have:

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

where the second inequality holds since $v(\mathbf{w}) > 0$, while the equality holds from the definition of $\hat{\xi}$. We can now use Eq. (16) to bound $\mathbb{E}_{q_{\hat{\xi}}} \left[\left\| B(\mathbf{w}) \right\|_D^2 \right]$, thus obtaining:

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

This concludes the proof. \square

Theorem 3. Fix a target task τ and let \tilde{Q} be the fixed-point of the corresponding mellow Bellman operator. Assume linearly parameterized value functions $Q_{\mathbf{w}}(s, a) = \mathbf{w}^T \phi(s, a)$ with bounded weights $\mathbf{w} \leq w_{max}$ and uniformly bounded features $\phi(s, a) \leq \phi_{max}$. Consider the Gaussian version of Alg. 1 with prior $p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and denote by $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ the variational parameter minimizing the objective of Eq. 2 on a dataset D of N samples. Let ν be a probability measure over $\mathcal{S} \times \mathcal{A}$ and $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \left\| \tilde{B}(\mathbf{w}) \right\|_\nu^2$. Define $v(\mathbf{w}^*) \triangleq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \mathbf{I})} [v(\mathbf{w})]$, with $v(\mathbf{w}) \triangleq \mathbb{E}_\nu [Var_{\mathcal{P}} [b(\mathbf{w})]]$. Then, there exist constants c_1, c_2, c_3 such that, with probability at least $1 - 2\delta$ over the choice of weights $\mathbf{w} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and dataset D :

$$\left\| Q_{\mathbf{w}} - \tilde{Q} \right\|_\nu^2 \leq \frac{1}{(1-\gamma)\delta} \left(2 \left\| \tilde{B}(\mathbf{w}^*) \right\|_\nu^2 + \frac{v(\mathbf{w}^*) + c_3 \sqrt{\log \frac{2}{\delta}}}{\sqrt{N}} + \frac{c_2 + \lambda \left\| \mathbf{w}^* - \boldsymbol{\mu}_p \right\|_{\boldsymbol{\Sigma}_p^{-1}}}{N} + \frac{c_1}{N^2} \right) \quad (7)$$

Proof. Using Lemma 3 with variational parameters $\hat{\xi} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, we have:

$$\begin{aligned} \mathbb{E}_{q_{\hat{\xi}}} \left[\left\| B(\mathbf{w}) \right\|_\nu^2 \right] &\leq \inf_{\xi \in \Xi} \left\{ \mathbb{E}_{q_\xi} \left[\left\| B(\mathbf{w}) \right\|_\nu^2 \right] + \mathbb{E}_{q_\xi} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(q_\xi||p) \right\} + 2 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{N}} \\ &\leq \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\left\| B(\mathbf{w}) \right\|_\nu^2 \right] + \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} [v(\mathbf{w})] + 2 \frac{\lambda}{N} KL(\mathcal{N}(\mathbf{w}^*, c\mathbf{I}) || p) \\ &\quad + 2 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{N}} \end{aligned} \quad (18)$$

where the second inequality is due to the fact that, since Lemma 3 contains an infimum over the variational parameters, we can upper bound its right-hand side by choosing any specific ξ from Ξ . Here, we choose $\mu = \mathbf{w}^*$ and $\Sigma = c\mathbf{I}$, for some positive constant $c > 0$. Let us now bound these terms separately.

Bounding the expected TD error We have:

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\left\| \tilde{B}(\mathbf{w}^*) \right\|_{\nu}^2 \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\mathbb{E}_{\nu} \left[(\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}})^2 \right] \right] \\ &= \mathbb{E}_{\nu} \left[\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[(\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}})^2 \right] \right] \\ &= \mathbb{E}_{\nu} \left[\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})}^2 \left[\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] \right] + \mathbb{E}_{\nu} \left[\text{Var}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] \right]\end{aligned}\tag{19}$$

Let us bound these two terms point-wisely for each s, a . For the first expectation, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\tilde{T}Q_{\mathbf{w}} - Q_{\mathbf{w}} \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[R(s, a) + \gamma \mathbb{E}_{s'} \text{mm}_{a'} \mathbf{w}^T \phi(s', a') - \mathbf{w}^T \phi(s, a) \right] \\ &= R(s, a) + \gamma \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\mathbb{E}_{s'} \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] - \mathbf{w}^{*T} \phi(s, a)\end{aligned}\tag{20}$$

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

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\mathbb{E}_{s'} \text{mm}_{a'} \mathbf{w}^T \phi(s', a') \right] &= \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[\mathbb{E}_{s'} \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} e^{\kappa \mathbf{w}^T \phi(s', a')} \right] \\ &\leq \mathbb{E}_{s'} \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a'} \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, c\mathbf{I})} \left[e^{\kappa \mathbf{w}^T \phi(s', a')} \right]\end{aligned}\tag{21}$$

Now, since we know that $\mathbf{w}^T \phi(s', a') \sim \mathcal{N}(\mathbf{w}^{*T} \phi(s', a'), c \phi(s', a')^T \phi(s', a'))$, $e^{\kappa \mathbf{w}^T \phi(s', a')}$ follows 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:

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

Plugging this into 21 and then into 20, we obtain:

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

This implies:

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

where the second inequality follows from Cauchy-Schwarz inequality. Going back to 19, the first term can now be upper bounded by:

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

544 Let us now consider the variance term of 19 and derive a bound that holds point-wisely for any s, a .
 545 We have:

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

546 From Cauchy-Schwarz inequality:

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

547 Then, the random variable over which the variance is computed is limited in
 548 $[-\sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma}, \sqrt{c} \mathbf{w}_{\max} \phi_{\max} \frac{1+\gamma}{\gamma}]$ and the variance can be straightforwardly bounded using
 549 Popoviciu's inequality:

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

550 We can finally plug everything into 19, thus obtaining:

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

551 **Bounding the KL divergence** We have:

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

552 Now, putting all together into 18:

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

553 where $v(\mathbf{w}^*) = \mathbb{E}_{\mathcal{N}(\mathbf{w}^*, \mathbf{I})} [v(\mathbf{w})]$. Since the bound holds for any $c > 0$, we can set it to $1/N$, thus
 554 obtaining:

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

555 Finally, defining the constants $c_1 = \frac{1}{2} \gamma^2 \kappa^2 \phi_{\max}^2 + \frac{\lambda K}{\sigma_{\min}}$, $c_2 = \mathbf{w}_{\max}^2 \phi_{\max}^2 (1+\gamma)^2 + \lambda K \log(\sigma_{\max} +$
 556 $\log N)$, and $c_3 = 2 \frac{R_{\max}^2}{(1-\gamma)^2}$, we obtain:

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

c_2 actually contains N . Should we make this more explicit?

557 Let us now apply Corollary 1. We have that, with probability at least $1 - \delta$:

$$\|Q_{\mathbf{w}} - \tilde{Q}\|_{\nu}^2 \leq \frac{\mathbb{E}_{q_{\xi}} [\|B(\mathbf{w})\|_{\nu}^2]}{(1 - \gamma)\delta}$$

558 Thus, we probability at least $1 - 2\delta$:

$$\|Q_{\mathbf{w}} - \tilde{Q}\|_{\nu}^2 \leq \frac{1}{(1 - \gamma)\delta} \left(2\|\tilde{B}(\mathbf{w}^*)\|_{\nu}^2 + \frac{c_1}{N^2} + \frac{c_2 + \lambda \|\mathbf{w}^* - \boldsymbol{\mu}_p\|_{\Sigma_p^{-1}}}{N} + \frac{v(\mathbf{w}^*) + c_3 \sqrt{\log \frac{2}{\delta}}}{\sqrt{N}} \right)$$

559 \square

560 B Additional Details on the Algorithms

561 B.1 Gaussian Variational Transfer

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

$$KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) = \frac{1}{2} \left(\log \frac{|\Sigma_p|}{|\Sigma|} + \text{Tr}(\Sigma_p^{-1}\Sigma) + (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \Sigma_p^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_p) - K \right) \quad (22)$$

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

$$\nabla_{\boldsymbol{\mu}} KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) = \Sigma_p^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_p) \quad (23)$$

565

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

566 Finally, the gradients w.r.t. the expected likelihood term of the variational objective (2) can be
567 computed using the reparameterization trick (e.g., [12, 25]):

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)} [\|B(\mathbf{w})\|_D^2] = \mathbb{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\nabla_{\mathbf{w}} \|B(\mathbf{w})\|_D^2] \text{ for } \mathbf{w} = \mathbf{L}\mathbf{v} + \boldsymbol{\mu} \quad (25)$$

568

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

569 B.2 Mixture of Gaussian Variational Transfer

570 For the implementation of the Mixture of Gaussian Variational Transfer, we use the upper bound
571 on the KL divergence between two Mixtures of Gaussians, as in Theorem1, to obtain an upper
572 bound on the negative ELBO in Equation2. Consider we have C components for the posterior
573 family $q_{\xi}(\mathbf{w}) = \frac{1}{C} \sum_{i=1}^C \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_i, \Sigma_i)$ and a prior distribution, constructed from the set of weights
574 $\mathcal{W}_s = \{\mathbf{w}_1, \dots, \mathbf{w}_{|\mathcal{W}_s|}\}$ of the sources' optimal Q -functions, $p(\mathbf{w}) = \frac{1}{|\mathcal{W}_s|} \sum_{j=1}^{|\mathcal{W}_s|} \mathcal{N}(\mathbf{w} | \mathbf{w}_j, \sigma_p^2 \mathbf{I})$.

$$KL(q_{\xi}(\mathbf{w}) \parallel p(\mathbf{w})) \leq KL(\chi^{(2)} \parallel \chi^{(1)}) + \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_i, \Sigma_i) \parallel \mathcal{N}(\mathbf{w} | \mathbf{w}_j, \sigma_p^2 \mathbf{I})) \quad (27)$$

575 And substituting 27 in the negative ELBO in 2 we get the following upper bound.

$$\mathcal{L}(\xi) \leq \tilde{\mathcal{L}}(\xi) = \mathbb{E}_{\mathbf{w} \sim q_{\xi}} [\|B(\mathbf{w})\|_D^2] + KL(\chi^{(2)} \parallel \chi^{(1)}) + \frac{\lambda}{N} \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_i, \Sigma_i) \parallel \mathcal{N}(\mathbf{w} | \mathbf{w}_j, \sigma_p^2 \mathbf{I})) \quad (28)$$

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

$$\tilde{\mathcal{L}}(\boldsymbol{\xi}) = \frac{1}{C} \sum_{i=1}^C \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \left[\|B(\mathbf{w})\|_D^2 \right] + KL(\chi^{(2)} \|\chi^{(1)}\|) + \frac{\lambda}{N} \sum_{i=1}^C \sum_{j=1}^{|\mathcal{W}_s|} \chi_{j,i}^{(2)} KL(\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \parallel \mathcal{N}(\mathbf{w}|\mathbf{w}_j, \sigma_p^2 \mathbf{I})) \quad (29)$$

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 Eq. 23, 24, 25, 26 derived for the Gaussian case.

C Additional Details on the Experiments

In the present section we present the hyper-parameters and corresponding values that were used for the experiments presented in this paper.

For all the algorithms implemented for the empirical evaluation, we used a Replay Buffer (*Buffer_size*) with a fixed size and batches of a given size (*Batch_size*) sampled randomly from it to perform the gradient steps.

In the case of no-transfer-based algorithms we used a linearly scheduled ϵ -greedy exploration for which we could specify the starting and final values of ϵ , ϵ_s and ϵ_f respectively, and the *Exploration Fraction* of the maximum number of iterations for which the linear schedule would be used. As the learning procedure is done by gradient descent, we also have α as the learning rate. Additionally, in the case of the DDQNs we have the *Target Update Frequency* that specifies the number of iterations before setting the target neural network to the current parameters of the online network.

As presented in Sec. 3.4, for the minimization procedure of the TD error based on *mellow-max* we also have to set κ to control the approximation to the maximum operator and $\psi \in [0, 1]$ to control the proportion between the full residual gradient and the semi-gradient approach used.

For the transfer algorithms proposed here, namely, GVT and MGVT we have the following parameters. As presented in Sec. 3.1, we have $\lambda \geq 0$ to control the trade-off between the prior knowledge and the loss measure—i.e. the TD error—in the target task. Moreover, we allow to control the learning rates differently for the gradients related to the means and covariance matrices, the Cholesky factor \mathbf{L} to be precise: α_μ and α_L . In order to approximate the expectation of the TD error, we used a fixed number of *weights* sampled from the posterior distribution. Specifically, for MGVT it is necessary to specify the value of the variance σ_p^2 for the prior model and the number of components C for the posterior family to be used. Furthermore, as explained in Sec. 3.3, to perform the initialization of the posterior by minimizing the KL divergence we set the a maximum number of iterations (*UKL max iterations*) and the precision to reach during this optimization ε and its learning rate η . Finally, to avoid problems with the positive-definiteness of the covariances being learned, we specify a hyper-parameter σ_{min} as the minimum value the diagonal elements of the Cholesky factor can reach.

In the following sections we present the configurations used for the experiments presented in this paper. For all our experiments we set $\psi = 0.5$, *UKL_max_iterations* = 600, $\eta = 10^{-6}$, $\varepsilon = 0.001$, $\sigma_{min} = 0.0001$. Also, we used ADAM optimizer—with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ —for the gradients steps related to the means and simple SGD for gradients steps of the Cholesky factors \mathbf{L} in GVT and MGVT. For the non-transfer algorithms only ADAM was used with the same mentioned parameters.

C.1 The Rooms Problem

C.2 Classic Control

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

During the training of the source tasks, we used a *Batch_size* = 32, *Buffer_size* = 50000. For DDQN with a *Target_Update_Frequency* = 500 and *Exploration_fraction* = 0.35 with $\epsilon_s = 1$ and $\epsilon_f = 0.1$. The function approximator was a Multilayer Perceptron (MLP) with ReLU

620 as activation function and a hidden layer of 32 neurons. For the transfer experiments we set
621 $Batch_size = 500$ and $weights = 5$, $\lambda = 0.001$, $\alpha_\mu = 0.001$ and $\alpha_L = 0.0001$. In the case
622 of MGVT we set $\sigma_p^2 = 10^{-5}$.

623 In the case of Mountain Car, to generate tasks we sampled uniformly the base speed of the actions
624 in the range $[0.001, 0.0015]$. We trained DDQN with $Target_Update_Frequency = 500$ and
625 $Exploration_fraction = 0.15$ with $\epsilon_s = 1$ and $\epsilon_f = 0.1$ and used an MLP with single hidden
626 layer of 64 neurons with ReLU activation. For the transfer experiments we set the $Batch_size = 50$,
627 $weights = 10$, $\lambda = 10^{-5}$, $\alpha_\mu = 0.001$ and $\alpha_L = 0.0001$. For MGVT we set $\sigma_p^2 = 10^{-5}$.

628 C.3 Maze Navigation