

---

# Formatting instructions for NIPS 2018

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1       The abstract paragraph should be indented 1/2 inch (3 picas) on both the left- and  
2       right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points.  
3       The word **Abstract** must be centered, bold, and in point size 12. Two line spaces  
4       precede the abstract. The abstract must be limited to one paragraph.

## 5   1   Introduction

## 6   2   Preliminaries

### 7   2.1   Markov Decision Processes

8       We define a Markov decision process (MDP) as a tuple  $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, p_0, \gamma \rangle$ , where  $\mathcal{S}$  is  
9       the state-space,  $\mathcal{A}$  is a finite set of actions,  $\mathcal{P}(\cdot|s, a)$  is the distribution of the next state  $s'$  given  
10       that action  $a$  is taken in state  $s$ ,  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function,  $p_0$  is the initial-state  
11       distribution, and  $\gamma \in [0, 1)$  is the discount factor. We assume the reward function to be uniformly  
12       bounded by a constant  $R_{max} > 0$ . A deterministic policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  is a mapping from states  
13       to actions. At the beginning of each episode of interaction, the initial state  $s_0$  is drawn from  $p_0$ .  
14       Then, the agent takes the action  $a_0 = \pi(s_0)$ , receives a reward  $\mathcal{R}(s_0, a_0)$ , transitions to the next  
15       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  
16       long-term return over a possibly infinite horizon:  $\max_{\pi} J(\pi) \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | \mathcal{M}, \pi]$ . To this end,  
17       we define the optimal value function  $Q^*(s, a)$  as the expected return obtained by taking action  $a$   
18       in state  $s$  and following an optimal policy thereafter. Then, an optimal policy  $\pi^*$  is a policy that  
19       is greedy with respect to the optimal value function, i.e.,  $\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$  for all states  
20        $s$ . It can be shown (e.g., [1]) that  $Q^*$  is the unique fixed-point of the optimal Bellman operator  $T$   
21       defined by  $TQ(s, a) = \mathcal{R}(s, a) + \gamma \mathbb{E}_{\mathcal{P}}[\max_{a'} Q(s', a')]$  for any value function  $Q$ . From now on, we  
22       adopt the term  $Q$ -function to denote any plausible value function, i.e., any function  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$   
23       uniformly bounded by  $\frac{R_{max}}{1-\gamma}$ .

24       When learning the optimal value function, a quantity of interest is how close a given  $Q$ -function  
25       is to the fixed-point of the Bellman operator. This is given by its Bellman residual, defined by  
26        $B(Q) \triangleq TQ - Q$ . Notice that  $Q$  is optimal if, and only if,  $B(Q)(s, a) = 0$  for all  $s, a$ . Furthermore,  
27       if we assume the existence of a distribution  $\nu$  over  $\mathcal{S} \times \mathcal{A}$ , the squared Bellman error of  $Q$  is  
28       defined as the expected squared Bellman residual of  $Q$  under  $\nu$ ,  $\|B(Q)\|_{\nu}^2 = \mathbb{E}_{\nu}[B^2(Q)]$ . Although  
29       minimizing the empirical Bellman error is an appealing objective, it is well-known that an unbiased  
30       estimator requires two independent samples of the next state  $s'$  of each  $s, a$  (e.g., []). In practice,  
31       the empirical Bellman error is typically replaced by the TD error, which approximates the former  
32       using a single transition sample. Given a dataset of  $N$  samples, the TD error is computed as  
33        $\|B(Q)\|_D^2 = \frac{1}{N} \sum_{i=1}^N (r_i + \gamma \max_{a'} Q(s'_i, a') - Q(s_i, a_i))^2$ .

cite Maillard

## 2.2 Variational Inference

When working with Bayesian approaches, the posterior distribution of hidden variables  $\mathbf{w} \in \mathbb{R}^K$  given data  $D$ ,

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)} = \frac{p(D|\mathbf{w})p(\mathbf{w})}{\int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})}, \quad (1)$$

is typically intractable for many models of interest (e.g., when working with deep neural networks) due to difficulties in computing the integral of Eq. (1). The main intuition behind variational inference [] is to approximate the intractable posterior  $p(\mathbf{w}|D)$  with a simpler distribution  $q_{\xi}(\mathbf{w})$ . The latter is chosen in a parametric family, with variational parameters  $\xi$ , as the minimizer of the Kullback-Leibler (KL) divergence w.r.t.  $p$ :

$$\min_{\xi} KL(q_{\xi}(\mathbf{w}) || p(\mathbf{w} | D)) \quad (2)$$

It is well-known that minimizing the KL divergence is equivalent to maximizing the so-called *evidence lower bound* (ELBO), which is defined as:

$$\text{ELBO}(\xi) = \mathbb{E}_{\mathbf{w} \sim q_{\xi}} [\log p(D|\mathbf{w})] - KL(q_{\xi}(\mathbf{w}) || p(\mathbf{w})) \quad (3)$$

Intuitively, the best approximation is the one that maximizes the expected log-likelihood of the data, while minimizing the KL divergence w.r.t. the prior  $p(\mathbf{w})$ .

## 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 Gaussian prior/posterior and mixture of Gaussian prior/posterior, respectively.

### 3.1 Algorithm

We begin with a simple consideration: the distribution  $\mathcal{D}$  over tasks clearly induces a distribution over optimal  $Q$ -functions. Since, for any MDP, learning its optimal  $Q$ -function is sufficient for solving the problem, one can safely replace the distribution over tasks with the distribution over their optimal value functions. Furthermore, assume we know such distribution and we are given a new task  $\tau$  to solve. Then, our main intuition is that it is possible to design an algorithm that efficiently explores  $\tau$  so as to quickly adapt the prior distribution in a Bayesian fashion to put all probability mass over the optimal  $Q$ -function of  $\tau$ .

We consider a parametric family of  $Q$ -functions  $\mathcal{Q} = \{Q_{\mathbf{w}} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \mid \mathbf{w} \in \mathbb{R}^K\}$ . For simplicity, we assume each function in  $\mathcal{Q}$  to be uniformly bounded by  $\frac{R_{max}}{1-\gamma}$ <sup>1</sup>. Then, we can reduce our prior distribution over  $Q$ -functions to a prior distribution over weights  $p(\mathbf{w})$ . Assume that we are given a dataset  $D = \{(s_i, a_i, s'_i, r_i) \mid i = 1, 2, \dots, N\}$  of samples from some task  $\tau$  we want to solve. Then, the posterior distribution over weights given such dataset can be computed by applying Bayes theorem as in Eq. 1. Unfortunately, this cannot be directly used in practice since we do not have a model of the likelihood  $p(D|\mathbf{w})$ . In such case, it is very common to make strong assumptions on the MDPs or the  $Q$ -functions so as to get a tractable posterior []. On the other hand, we take a PAC-Bayesian approach to derive a more general and meaningful posterior form. Recall that our final goal is move all probability mass over the weights minimizing some empirical loss measure, which in our case is the TD error  $\|B(\mathbf{w})\|_D^2$ . Then, given a prior  $p(\mathbf{w})$  we know from PAC-Bayesian theory that the optimal Gibbs posterior takes the form []:

$$q(\mathbf{w}) = \frac{e^{-\Lambda \|B(\mathbf{w})\|_D^2} p(\mathbf{w})}{\int e^{-\Lambda \|B(\mathbf{w}')\|_D^2} p(d\mathbf{w}')} \quad (4)$$

for some parameter  $\Lambda > 0$ . Since  $\Lambda$  is typically chosen to increase with the number of samples  $N$ , we set it to  $\lambda^{-1}N$ , for some constant  $\lambda > 0$ . Notice that, whenever the term  $e^{-\Lambda \|B(\mathbf{w})\|_D^2}$  can

<sup>1</sup>In practice, this is easily achieved by truncation.

CITE

Cite some-body

Cite Catoni 2007

74 be interpreted as the actual likelihood,  $q$  becomes a classic Bayesian posterior. Unfortunately, the  
 75 integral at the denominator of  $q$  is still intractable to compute even for simple  $Q$ -function models.  
 76 Thus, we propose a variational approximation  $q_\xi$  in a simpler family of distributions parameterized  
 77 by  $\xi \in \Xi$ . Then, our problem reduces to finding the variational parameters  $\xi$  such that  $q_\xi$  minimizes  
 78 the KL divergence w.r.t.  $q$ :

$$\min_{\xi \in \Xi} KL(q_\xi(\mathbf{w}) \parallel q(\mathbf{w})) = \min_{\xi \in \Xi} \mathbb{E}_{\mathbf{w} \sim q_\xi} [\|B(\mathbf{w})\|_D^2] - \frac{\lambda}{N} KL(q_\xi(\mathbf{w}) \parallel p(\mathbf{w})) \quad (5)$$

79 where the last objective is the well-known *evidence lower bound* (ELBO) []. Intuitively, the ap- Cite  
 80 proximate posterior trades-off between placing probability mass over those weights  $\mathbf{w}$  that have  
 81 low TD error (first term), and staying close to the prior distribution (second term). Assuming that  
 82 we are able to compute the gradients of (5) w.r.t. the variational parameters, our objective can be  
 83 easily optimized with any stochastic optimization algorithm. Notice, however, that differentiating  
 84 w.r.t.  $\xi$  typically requires differentiating  $\|B(\mathbf{w})\|_D^2$  w.r.t.  $\mathbf{w}$  (e.g., when using the reparameterization  
 85 trick []). Unfortunately, the TD error is well-known to be non-differentiable due to the presence of Cite Deep-  
 86 the max operator. This rarely represents a problem since typical value-based algorithm are actually mind  
 87 semi-gradient methods, i.e., they do not differentiate the targets (see, e.g., Chapter 11 of []). However,  
 88 in our case ... Cite Sutton

89 What is a good motivation for the fact that we need a residual algorithm?

90 To solve this issue, we replace the optimal Bellman operator with the mellow Bellman operator  
 91 introduced in [], which adopts a softened version of max called *mellowmax*: Cite MM

$$\text{mm}_a Q_w(s, a) = \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_a e^{\kappa Q_w(s, a)} \quad (6)$$

92 where  $\kappa$  is a hyperparameter and  $|\mathcal{A}|$  is the number of actions. The mellow Bellman operator, which  
 93 we denote as  $\tilde{T}$ , has several appealing properties that make it suitable for our settings: (i) it converges  
 94 to the maximum as  $\kappa \rightarrow \infty$ , (ii) it has a unique fixed point, and (iii) it is *differentiable*. Denoting by  
 95  $\tilde{B}(\mathbf{w}) = \tilde{T}Q_w - Q_w$  the Bellman residual w.r.t. the mellow Bellman operator  $\tilde{T}$ , we have that the  
 96 corresponding TD error,  $\|\tilde{B}(\mathbf{w})\|_D^2$ , is now differentiable with respect to  $\mathbf{w}$ .

97 Here we need to talk about residual algorithms and their improved gradient

98 Our main algorithm is shown in Alg. 1. We start by estimating a prior distribution from the given  
 99 set of source  $Q$ -functions (line 1) and we initialize the variational parameters by minimizing the KL  
 100 divergence w.r.t. such distribution<sup>2</sup> (line 2). Then, at each time step of interaction, we re-sample the  
 101 weights from the current approximate posterior and act greedily w.r.t. the corresponding  $Q$ -function  
 102 (lines 7,8). This resembles the well-known Thompson sampling adopted in multi-armed bandits [and Cite  
 103 allows our algorithm to efficiently explore the target task. In some sense, at each time we guess what  
 104 is the task we are trying to solve based on our current belief and we act as if such guess were actually  
 105 true. After collecting and storing the new experience (lines 9-11), we draw a batch of samples from  
 106 the replay buffer and a batch of weights from the posterior (line 12). We use these to approximate the  
 107 ELBO, compute its gradient, and finally update the variational parameters (lines 13-15).

108 The main advantage of our approach is that it exploits knowledge from the source tasks to perform an  
 109 efficient adaptive exploration. Intuitively, during the first steps of interaction, our algorithm has no  
 110 idea about what is the current task. However, it can rely on the learned prior to take early informed  
 111 decisions. As the learning process goes on, it will quickly figure out which task is being solved, thus  
 112 moving all probability mass over the weights minimizing the TD error. From that point, sampling  
 113 from the posterior is approximately equivalent to deterministically taking the best weights, and no  
 114 more exploration will be performed.

### 115 3.2 Gaussian Variational Transfer

116 We now restrict ourselves to a specific choice of the prior and posterior families that makes our  
 117 algorithm very efficient and easy to implement. We assume that optimal  $Q$ -functions according to our

<sup>2</sup>If the prior and approximate posterior were in the same family of distributions we could simply set  $\xi$  to the prior parameters, however this does not always hold in practice.

---

**Algorithm 1** Variational Transfer

---

**Require:** Target task  $\tau$ , source  $Q$ -function weights  $\mathcal{W}_s$ , batch sizes  $M_D$  and  $M_{\mathcal{W}}$ , prior weight  $\lambda$

---

```

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  $s_{h+1} \sim \mathcal{P}^{(\tau)}(\cdot | s_h, a_h)$ 
10:    Collect reward  $r_h = \mathcal{R}^{(\tau)}(s_h, a_h)$ 
11:    Add sample to the replay buffer:  $D \leftarrow D \cup \langle s_h, a_h, r_h, s_{h+1} \rangle$ 
12:    Sample batch  $D' = \langle s_i, a_i, r_i, s'_i \rangle_{i=1}^{M_D}$  from  $D$  and  $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{M_{\mathcal{W}}}\}$  from  $q_{\xi}$ 
13:    Approximate ELBO:  $\mathcal{L}(\xi) = \frac{1}{M_{\mathcal{W}}} \sum_{\mathbf{w} \in \mathcal{W}} \|B(\mathbf{w})\|_{D'}^2 - \frac{\lambda}{N} KL(q_{\xi} || p)$ 
14:    Compute the gradient  $\nabla_{\xi} \mathcal{L}(\xi)$ 
15:    Update  $\xi$  in the direction of  $\nabla_{\xi} \mathcal{L}(\xi)$  using any stochastic optimizer (e.g., ADAM)
16:   end while
17: until forever

```

---

task distribution (or better, their weights) follow a multivariate Gaussian law. 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, e.g., 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 going 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. Under Gaussian distributions, all quantity of interest for using Alg. 1 can be computed very easily. The KL divergence between the prior and approximate posterior can be computed in closed-form as:

$$KL(q_{\xi}(\mathbf{w}) || p(\mathbf{w})) = \frac{1}{2} \left( \log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}|} + \operatorname{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) - K \right) \quad (7)$$

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

$$\nabla_{\boldsymbol{\mu}} KL(q_{\xi}(\mathbf{w}) || p(\mathbf{w})) = \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) \quad (8)$$

Cite matrix  
cookbook

$$\nabla_{\mathbf{L}} KL(q_{\xi}(\mathbf{w}) || p(\mathbf{w})) = \boldsymbol{\Sigma}_p^{-1} \mathbf{L} - (\mathbf{L}^{-1})^T \quad (9)$$

Finally, the gradients w.r.t. the expected likelihood term of the variational objective (5) can be computed using the reparameterization trick (e.g., []):

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)} [\|B(\mathbf{w})\|_D^2] = \mathbb{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\nabla_{\mathbf{w}} \|B(\mathbf{w})\|_D^2] \quad \text{for } \mathbf{w} = \mathbf{L}\mathbf{v} + \boldsymbol{\mu} \quad (10)$$

Cite deep-  
mind and  
another

$$\nabla_{\mathbf{L}} \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)} [\|B(\mathbf{w})\|_D^2] = \mathbb{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\nabla_{\mathbf{w}} \|B(\mathbf{w})\|_D^2 \cdot \mathbf{v}^T] \quad \text{for } \mathbf{w} = \mathbf{L}\mathbf{v} + \boldsymbol{\mu} \quad (11)$$

### 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 follow a Gaussian law, the  $Q$ -values might be far from it.

## 4 Theoretical Analysis

In this section, we theoretically analyze our variational transfer algorithm...

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

Cite MM

146 **Theorem 1.** *Let  $V$  be the fixed-point of the optimal Bellman operator  $T$ , and  $Q$  the corresponding*  
 147 *action-value function. Define the action-gap function  $g(s)$  as the difference between the value of*  
 148 *the best action and the second best action at each state  $s$ . Let  $\tilde{V}$  be the fixed-point of the mellow*  
 149 *Bellman operator  $\tilde{T}$  with parameter  $\kappa > 0$  and denote by  $\beta > 0$  the inverse temperature of the*  
 150 *induced Boltzmann distribution (as in []). Let  $\nu$  be a probability measure over the state-space. Then,*  
 151 *for any  $p \geq 1$ :*

Cite MM

$$\|V - \tilde{V}\|_{\nu,p}^p \leq \frac{2R_{max}}{(1-\gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g}} \right\|_{\nu,p}^p \quad (12)$$

## 152 5 Related Works

## 153 6 Experiments

### 154 6.1 Gridworld

### 155 6.2 Classic Control

### 156 6.3 Maze Navigation

## 157 7 Conclusion

## 158 References

- 159 [1] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley  
 160 & Sons, Inc., New York, NY, USA, 1994.

## 161 A Proofs

162 **Theorem 1.** *Let  $V$  be the fixed-point of the optimal Bellman operator  $T$ , and  $Q$  the corresponding*  
 163 *action-value function. Define the action-gap function  $g(s)$  as the difference between the value of*  
 164 *the best action and the second best action at each state  $s$ . Let  $\tilde{V}$  be the fixed-point of the mellow*  
 165 *Bellman operator  $\tilde{T}$  with parameter  $\kappa > 0$  and denote by  $\beta > 0$  the inverse temperature of the*  
 166 *induced Boltzmann distribution (as in []). Let  $\nu$  be a probability measure over the state-space. Then,*  
 167 *for any  $p \geq 1$ :*

Cite MM

$$\|V - \tilde{V}\|_{\nu,p}^p \leq \frac{2R_{max}}{(1-\gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g}} \right\|_{\nu,p}^p \quad (12)$$

168 *Proof.* We begin by noticing that:

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

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

$$\|V - \tilde{V}\|_{\nu,p}^p \leq \frac{1}{1-\gamma} \|TV - \tilde{T}V\|_{\nu,p}^p \quad (13)$$

171 Let us bound the norm on the right-hand side separately. In order to do that, we will bound the  
 172 function  $|TV(s) - \tilde{T}V(s)|$  point-wisely for any state  $s$ . By applying the definition of the optimal  
 173 and mellow Bellman operators, we obtain:

$$\begin{aligned} |TV(s) - \tilde{T}V(s)| &= \left| \max_a \{R(s, a) + \gamma \mathbb{E}[V(s')]\} - \min_a \{R(s, a) + \gamma \mathbb{E}[V(s')]\} \right| \\ &= \left| \max_a Q(s, a) - \min_a Q(s, a) \right| \end{aligned}$$

174 Recall that applying the mellow-max is equivalent to computing an expectation under a Boltzmann  
 175 distribution with inverse temperature  $\beta$  induced by  $Q$  []. Thus, we can write:

Cite MM

$$\begin{aligned} \left| \max_a Q(s, a) - \min_a Q(s, a) \right| &= \left| \sum_a \pi^*(a|s) Q(s, a) - \sum_a \pi_\beta(a|s) Q(s, a) \right| \\ &= \left| \sum_a Q(s, a) (\pi^*(a|s) - \pi_\beta(a|s)) \right| \\ &\leq \sum_a |Q(s, a)| |\pi^*(a|s) - \pi_\beta(a|s)| \\ &\leq \frac{R_{max}}{1-\gamma} \sum_a |\pi^*(a|s) - \pi_\beta(a|s)| \end{aligned} \quad (14)$$

176 where  $\pi^*$  is the optimal (deterministic) policy w.r.t.  $Q$  and  $\pi_\beta$  is the Boltzmann distribution induced  
 177 by  $Q$  with inverse temperature  $\beta$ :

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

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

$$\begin{aligned}
\sum_a |\pi^*(a|s) - \pi_\beta(a|s)| &= |\pi^*(a_1(s)|s) - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi^*(a|s) - \pi_\beta(a|s)| \\
&= |1 - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi_\beta(a|s)| \\
&= 2 |1 - \pi_\beta(a_1(s)|s)|
\end{aligned} \tag{15}$$

179 Finally, let us bound this last term:

$$\begin{aligned}
|1 - \pi_\beta(a_1(s)|s)| &= \left| 1 - \frac{e^{\beta Q(s, a_1(s))}}{\sum_{a'} e^{\beta Q(s, a')}} \right| \\
&= \left| 1 - \frac{e^{\beta(Q(s, a_1(s)) - Q(s, a_2(s)))}}{\sum_{a'} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta g(s)}}{\sum_{a'} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + \sum_{a' \neq a_1(s)} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&\leq \left| 1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + |\mathcal{A}|} \right| \\
&= \left| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g(s)}} \right|
\end{aligned} \tag{16}$$

180 Combining Eq. (14), (15), and (16), we obtain:

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

181 Taking the norm and plugging this into Eq. (13) concludes the proof.  $\square$

182 **Lemma 1.** Let  $p$  and  $\nu$  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  $\delta > 0$ ,  
183 with probability at least  $1 - \delta$  over the choice of a  $Q$ -function  $Q$ , the following holds:  
184

$$\|Q - Q^*\|_\nu^2 \leq \frac{\mathbb{E}_p \left[ \|B(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \tag{17}$$

185 *Proof.* First notice that:

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

186 which implies that:

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

187 Then we can write:

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

188 Settings the right-hand side equal to  $\delta$  and solving for  $\epsilon$  concludes the proof.  $\square$

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

$$\|Q - \tilde{Q}\|_\nu^2 \leq \frac{\mathbb{E}_p \left[ \|\tilde{B}(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \quad (18)$$

192 **Lemma 2.** Assume  $Q$ -functions belong to a parametric space of functions bounded by  $\frac{R_{max}}{1-\gamma}$ . Let  $p$   
 193 and  $q$  be arbitrary distributions over the parameter space  $\mathcal{W}$ , and  $\nu$  be a probability measure over  
 194  $\mathcal{S} \times \mathcal{A}$ . 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  
 195  $\delta > 0$ , with probability at least  $1 - \delta$ , the following two inequalities hold simultaneously:

$$\mathbb{E}_q \left[ \|B(\mathbf{w})\|_\nu^2 \right] \leq \mathbb{E}_q \left[ \|B(\mathbf{w})\|_D^2 \right] - \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1 - \gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (19)$$

196

$$\mathbb{E}_q \left[ \|B(\mathbf{w})\|_D^2 \right] \leq \mathbb{E}_q \left[ \|B(\mathbf{w})\|_\nu^2 \right] + \mathbb{E}_q [v(\mathbf{w})] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1 - \gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (20)$$

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

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

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

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

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

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

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

$$\left| \|B(\mathbf{w})\|_\nu^2 + v(\mathbf{w}) - \|B(\mathbf{w})\|_D^2 \right| \leq 4 \frac{R_{max}^2}{(1 - \gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (21)$$

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

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

203 Thus, multiplying both sides of (21) by  $\lambda^{-1}N$  and applying the change of measure inequality with  
 204  $f(\mathbf{w}) = \lambda^{-1}N \left| \|B(\mathbf{w})\|_\nu^2 + v(\mathbf{w}) - \|B(\mathbf{w})\|_D^2 \right|$ , 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 ref-  
erence for  
this



205 where the second inequality holds since the right-hand side of (21) does not depend on  $\mathbf{w}$ . Finally,  
 206 we can explicitly write:

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

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

208 **Lemma 3.** *Let  $p$  be a prior distribution over the parameter space  $\mathcal{W}$ , and  $\nu$  be a probability measure*  
 209 *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*  
 210 *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*  
 211 *least  $1 - \delta$ :*

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

212 *Proof.* Let us use Lemma 2 for the specific choice  $q = q_{\hat{\xi}}$ . From Eq. (19), we have:

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

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

214 We can now use Eq. (20) to bound  $\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| B(\mathbf{w}) \right\|_D^2 \right]$ , thus obtaining:

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

215 This concludes the proof.  $\square$