

---

# 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  $\mu$  over  $\mathcal{S} \times \mathcal{A}$ , the expected squared Bellman error  
28       of  $Q$  is defined as the expected squared Bellman residual of  $Q$  under  $\mu$ ,  $\mathbb{E}_{\mu}[B^2(Q)]$ . 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        $\frac{1}{N} \sum_{i=1}^N (r_i + \gamma \max_{a'} Q(s'_i, a') - Q(s_i, a_i))^2$ .

cite Maillard

## 34 2.2 Variational Inference

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

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

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

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

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

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

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

## 46 3 Variational Transfer Learning

### 47 3.1 Algorithm

### 48 3.2 Gaussian Variational Transfer

### 49 3.3 Mixture of Gaussian Variational Transfer

## 50 4 Theoretical Analysis

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

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

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

$$\|V - \tilde{V}\|_{\nu, p}^p \leq \frac{2R_{max}}{(1 - \gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g}} \right\|_{\nu, p}^p \quad (4)$$

65 **5 Related Works**

66 **6 Experiments**

67 **6.1 Gridworld**

68 **6.2 Classic Control**

69 **6.3 Maze Navigation**

70 **7 Conclusion**

71 **References**

- 72 [1] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley  
73 & Sons, Inc., New York, NY, USA, 1994.

## 74 A Proofs

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

Cite MM

$$\|V - \tilde{V}\|_{\nu,p}^p \leq \frac{2R_{max}}{(1-\gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g}} \right\|_{\nu,p}^p \quad (4)$$

81 *Proof.* We begin by noticing that:

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

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

$$\|V - \tilde{V}\|_{\nu,p}^p \leq \frac{1}{1-\gamma} \|TV - \tilde{T}V\|_{\nu,p}^p \quad (5)$$

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

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

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

Cite MM

$$\begin{aligned} \left| \max_a Q(s, a) - \min_a Q(s, a) \right| &= \left| \sum_a \pi^*(a|s) Q(s, a) - \sum_a \pi_\beta(a|s) Q(s, a) \right| \\ &= \left| \sum_a Q(s, a) (\pi^*(a|s) - \pi_\beta(a|s)) \right| \\ &\leq \sum_a |Q(s, a)| |\pi^*(a|s) - \pi_\beta(a|s)| \\ &\leq \frac{R_{max}}{1-\gamma} \sum_a |\pi^*(a|s) - \pi_\beta(a|s)| \end{aligned} \quad (6)$$

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

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

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

$$\begin{aligned}
\sum_a |\pi^*(a|s) - \pi_\beta(a|s)| &= |\pi^*(a_1(s)|s) - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi^*(a|s) - \pi_\beta(a|s)| \\
&= |1 - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi_\beta(a|s)| \\
&= 2 |1 - \pi_\beta(a_1(s)|s)|
\end{aligned} \tag{7}$$

92 Finally, let us bound this last term:

$$\begin{aligned}
|1 - \pi_\beta(a_1(s)|s)| &= \left| 1 - \frac{e^{\beta Q(s, a_1(s))}}{\sum_{a'} e^{\beta Q(s, a')}} \right| \\
&= \left| 1 - \frac{e^{\beta(Q(s, a_1(s)) - Q(s, a_2(s)))}}{\sum_{a'} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta g(s)}}{\sum_{a'} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&= \left| 1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + \sum_{a' \neq a_1(s)} e^{\beta(Q(s, a') - Q(s, a_2(s)))}} \right| \\
&\leq \left| 1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + |\mathcal{A}|} \right| \\
&= \left| 1 - \frac{1}{1 + |\mathcal{A}| e^{-\beta g(s)}} \right|
\end{aligned} \tag{8}$$

93 Combining Eq. (6), (7), and (8), we obtain:

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

94 Taking the norm and plugging this into Eq. (5) concludes the proof.  $\square$

95 **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  $\delta > 0$ ,  
96 with probability at least  $1 - \delta$  over the choice of a  $Q$ -function  $Q$ , the following holds:  
97

$$\|Q - Q^*\|_\nu^2 \leq \frac{\mathbb{E}_p \left[ \|B(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \tag{9}$$

98 *Proof.* First notice that:

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

99 which implies that:

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

100 Then we can write:

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

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

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

$$\|Q - \tilde{Q}\|_\nu^2 \leq \frac{\mathbb{E}_p \left[ \|\tilde{B}(Q)\|_\nu^2 \right]}{(1 - \gamma)\delta} \quad (10)$$

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

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

109

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

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

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

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

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

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

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

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

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

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

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

116 Thus, multiplying both sides of (13) by  $\lambda^{-1}N$  and applying the change of measure inequality with  
 117  $f(\mathbf{w}) = \lambda^{-1}N \left| \|B(\mathbf{w})\|_\nu^2 + v(\mathbf{w}) - \|B(\mathbf{w})\|_D^2 \right|$ , we obtain:

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

Find a ref-  
erence for  
this

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

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

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

121 **Lemma 3.** Let  $p$  be a prior distribution over the parameter space  $\mathcal{W}$ , and  $\nu$  be a probability measure  
 122 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  
 123 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  
 124 least  $1 - \delta$ :

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

125 *Proof.* Let us use Lemma 2 for the specific choice  $q = q_{\hat{\xi}}$ . From Eq. (11), we have:

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

126 where the second inequality holds since  $v(\mathbf{w}) > 0$ , while the equality holds from the definition of  $\hat{\xi}$ .  
 127 We can now use Eq. (12) to bound  $\mathbb{E}_{q_{\hat{\xi}}} \left[ \left\| B(\mathbf{w}) \right\|_D^2 \right]$ , thus obtaining:

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

128 This concludes the proof.  $\square$