Formatting instructions for NIPS 2018

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

2 Preliminaries

2.1 Markov Decision Processes

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 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 that action a is taken in state s, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 $\pi : \mathcal{S} \to \mathcal{A}$ 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 $\mathcal{R}(s_0, a_0)$, 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: $\max_\pi J(\pi) \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid \mathcal{M}, \pi]$. To this end, we define the optimal value function $Q^*(s, a)$ as the expected return obtained by taking action a in state s and following an optimal policy thereafter. Then, an optimal policy π^* is a policy that is greedy with respect to the optimal value function, i.e., $\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$ for all states s. It can be shown (e.g., [1]) that Q^* is the unique fixed-point of the optimal Bellman operator T 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 adopt the term Q-function to denote any plausible value function, i.e., any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ uniformly bounded by $\frac{R_{max}}{1-\gamma}$.

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

cite Maillard

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

2.2 Variational Inference

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

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})p(\boldsymbol{w})}{p(D)} = \frac{p(D|\boldsymbol{w})p(\boldsymbol{w})}{\int_{\boldsymbol{w}} p(D|\boldsymbol{w})p(\boldsymbol{w})}, \tag{1}$$

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 $p(\boldsymbol{w}|D)$ with a simpler distribution $q_{\boldsymbol{\xi}}(\boldsymbol{w})$. The latter is chosen in a parametric family, with variational parameters $\boldsymbol{\xi}$, as the minimizer of the Kullback-Leibler (KL) divergence w.r.t. p:

CITE

$$\min_{\boldsymbol{\xi}} KL\left(q_{\boldsymbol{\xi}}(\boldsymbol{w}) \mid\mid p(\boldsymbol{w} \mid D)\right) \tag{2}$$

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

$$\text{ELBO}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{w} \sim q_{\boldsymbol{\xi}}}\left[\log p(D|\boldsymbol{w})\right] - KL\left(q_{\boldsymbol{\xi}}(\boldsymbol{w}) \mid\mid p(\boldsymbol{w})\right) \tag{3}$$

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

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 τ to solve. Then, our main intuition is that it is possible to design an algorithm that efficiently explores τ so as to quickly adapt the prior distribution in a Bayesian fashion to put all probability mass over the optimal Q-function of τ.

We consider a parametric family of Q-functions $\mathcal{Q} = \left\{Q_{\boldsymbol{w}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \mid \boldsymbol{w} \in \mathbb{R}^K\right\}$. For simplicity, we assume each function in $\mathcal{Q}$ to be uniformly bounded by $\frac{R_{max}}{1-\gamma}$[1]. Then, we can reduce our prior distribution over Q-functions to a prior distribution over weights $p(\boldsymbol{w})$. Assume that we are given a dataset $D = \left\{(s_i, a_i, s_i', r_i) \mid i = 1, 2, \dots N\right\}$ of samples from some task τ 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 $p(D|\boldsymbol{w})$. 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 $\|B(\boldsymbol{w})\|_D^2$. Then, given a prior $p(\boldsymbol{w})$ we know from PAC-Bayesian theory that the optimal Gibbs posterior takes the form []:

Cite some-body

Cite Catoni 2007

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

for some parameter $\Lambda > 0$. Since Λ 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 $e^{-\Lambda\|B(\boldsymbol{w})\|_D^2}$ can

[1]In practice, this is easily achieved by truncation.

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

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

where the last objective is the well-knwon *evidence lower bound* (ELBO) []. Intuitively, the approximate posterior trades-off between placing probability mass over those weights $\boldsymbol{w}$ that have low TD error (first term), and staying close to the prior distribution (second term). Assuming that we are able to compute the gradients of (5) w.r.t. the variational parameters, our objective can be easily optimized with any stochastic optimization algorithm. Notice, however, that differentiating w.r.t. $\boldsymbol{\xi}$ typically requires differentiating $\|B(\boldsymbol{w})\|_D^2$ w.r.t. $\boldsymbol{w}$ (e.g., when using the reparameterization trick []). 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 algorithm are actually semi-gradient methods, i.e., they do not differentiate the targets (see, e.g., Chapter 11 of []). However, in our case ...

Cite

Cite Deepmind

Cite Sutton

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

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

Cite MM

$$\underset{a}{\mathrm{mm}}\, Q_{\boldsymbol{w}}(s,a) = \frac{1}{\kappa} \log \frac{1}{|\mathcal{A}|} \sum_{a} e^{\kappa Q_{\boldsymbol{w}}(s,a)} \tag{6}$$

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

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

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

Cite somebody

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

[2]If the prior and approximate posterior were in the same family of distributions we could simply set $\boldsymbol{\xi}$ to the prior parameters, however this does not always hold in practice.

Algorithm 1 Variational Transfer

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

Estimate prior $p(\boldsymbol{w})$ from $\mathcal{W}_s$
Initialize variational parameters: $\boldsymbol{\xi} \leftarrow \operatorname{argmin}_{\boldsymbol{\xi}} KL(q_{\boldsymbol{\xi}}||p)$
Initialize replay buffer: $D = \emptyset$
repeat
  Sample initial state: $s_0 \sim p_0^{(\tau)}$
  while s_h is not terminal **do**
    Sample weights: $\boldsymbol{w} \sim q_{\boldsymbol{\xi}}(\boldsymbol{w})$
    Take action $a_h = \operatorname{argmax}_a Q_{\boldsymbol{w}}(s_h, a)$
    Observe transition $s_{h+1} \sim \mathcal{P}^{(\tau)}(\cdot|s_h, a_h)$
    Collect reward $r_h = \mathcal{R}^{(\tau)}(s_h, a_h)$
    Add sample to the replay buffer: $D \leftarrow D \cup \langle s_h, a_h, r_h, s_{h+1}\rangle$
    Sample mini-batch $D' = \langle s_i, a_i, r_i, s'_i\rangle_{i=1}^{M_D}$ from D and $\mathcal{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_{M_{\mathcal{W}}}\}$ from $q_{\boldsymbol{\xi}}$
    Approximate ELBO: $\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{M_{\mathcal{W}}} \sum_{\boldsymbol{w} \in \mathcal{W}} \|B(\boldsymbol{w})\|_{D'}^2 - \frac{\lambda}{N} KL\left(q_{\boldsymbol{\xi}} \,||\, p\right)$
    Compute the gradient $\nabla_{\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{\xi})$
    Update $\boldsymbol{\xi}$ in the direction of $\nabla_{\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{\xi})$ using any stochastic optimizer (e.g., ADAM)
  end while
until forever

3.2 Gaussian Variational Transfer

3.3 Mixture of Gaussian Variational Transfer

4 Theoretical Analysis

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

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 [] that the mellow Bellman operator is a contraction 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 $\kappa \to \infty$, 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.

Cite MM

Theorem 1. *Let V be the fixed-point of the optimal Bellman operator T, and Q the corresponding action-value function. 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 $\widetilde{V}$ be the fixed-point of the mellow Bellman operator $\widetilde{T}$ with parameter $\kappa > 0$ and denote by $\beta > 0$ the inverse temperature of the induced Boltzmann distribution (as in []). Let ν be a probability measure over the state-space. Then, for any $p \geq 1$:*

Cite MM

$$\left\| V - \widetilde{V} \right\|_{\nu,p}^p \leq \frac{2R_{max}}{(1-\gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}|\, e^{-\beta g}} \right\|_{\nu,p}^p \tag{7}$$

5 Related Works

6 Experiments

6.1 Gridworld

6.2 Classic Control

6.3 Maze Navigation

7 Conclusion

References

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

A Proofs

Theorem 1. *Let V be the fixed-point of the optimal Bellman operator T, and Q the corresponding action-value function. 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 $\widetilde{V}$ be the fixed-point of the mellow Bellman operator $\widetilde{T}$ with parameter $\kappa > 0$ and denote by $\beta > 0$ the inverse temperature of the induced Boltzmann distribution (as in []). Let ν be a probability measure over the state-space. Then, for any $p \geq 1$:* Cite MM

$$\left\| V - \widetilde{V} \right\|_{\nu,p}^{p} \leq \frac{2R_{max}}{(1-\gamma)^2} \left\| 1 - \frac{1}{1 + |\mathcal{A}|\, e^{-\beta g}} \right\|_{\nu,p}^{p} \tag{7}$$

Proof. We begin by noticing that:

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

where the first inequality follows from Minkowsky's inequality and the second one from the contraction property of the mellow Bellman operator. This implies that:

$$\left\| V - \widetilde{V} \right\|_{\nu,p}^{p} \leq \frac{1}{1-\gamma} \left\| TV - \widetilde{T}V \right\|_{\nu,p}^{p} \tag{8}$$

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

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

Recall that applying the mellow-max is equivalent to computing an expectation under a Boltzmann distribution with inverse temperature β induced by κ []. Thus, we can write: Cite MM

$$\begin{aligned}
\left| \max_a Q(s,a) - \operatorname*{mm}_a Q(s,a) \right| &= \left| \sum_a \pi^*(a|s) Q(s,a) - \sum_a \pi_\beta(a|s) Q(s,a) \right| \\
&= \left| \sum_a Q(s,a) \left(\pi^*(a|s) - \pi_\beta(a|s) \right) \right| \\
&\leq \sum_a |Q(s,a)| \left| \pi^*(a|s) - \pi_\beta(a|s) \right| \\
&\leq \frac{R_{max}}{1-\gamma} \sum_a \left| \pi^*(a|s) - \pi_\beta(a|s) \right| \qquad (9)
\end{aligned}$$

where π^* is the optimal (deterministic) policy w.r.t. Q and π_β is the Boltzmann distribution induced by Q with inverse temperature β:

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

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

$$\begin{aligned}\sum_a |\pi^*(a|s) - \pi_\beta(a|s)| &= |\pi^*(a_1(s)|s) - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi^*(a|s) - \pi_\beta(a|s)| \\ &= |1 - \pi_\beta(a_1(s)|s)| + \sum_{a \neq a_1(s)} |\pi_\beta(a|s)| \\ &= 2\,|1 - \pi_\beta(a_1(s)|s)| \end{aligned} \tag{10}$$

Finally, let us bound this last term:

$$\begin{aligned}|1 - \pi_\beta(a_1(s)|s)| &= \left|1 - \frac{e^{\beta Q(s,a_1(s))}}{\sum_{a'} e^{\beta Q(s,a')}}\right| \\ &= \left|1 - \frac{e^{\beta(Q(s,a_1(s)) - Q(s,a_2(s)))}}{\sum_{a'} e^{\beta(Q(s,a') - Q(s,a_2(s)))}}\right| \\ &= \left|1 - \frac{e^{\beta g(s)}}{\sum_{a'} e^{\beta(Q(s,a') - Q(s,a_2(s)))}}\right| \\ &= \left|1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + \sum_{a' \neq a_1(s)} e^{\beta(Q(s,a') - Q(s,a_2(s)))}}\right| \\ &\leq \left|1 - \frac{e^{\beta g(s)}}{e^{\beta g(s)} + |\mathcal{A}|}\right| \\ &= \left|1 - \frac{1}{1 + |\mathcal{A}|\, e^{-\beta g(s)}}\right| \end{aligned} \tag{11}$$

Combining Eq. (9), (10), and (11), we obtain:

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

Taking the norm and plugging this into Eq. (8) concludes the proof. □

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

$$\|Q - Q^*\|_\nu^2 \leq \frac{\mathbb{E}_p\left[\|B(Q)\|_\nu^2\right]}{(1-\gamma)\delta} \tag{12}$$

Proof. First notice that:

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

which implies that:

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

Then we can write:

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

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

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

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

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

$$\mathbb{E}_q \left[\|B(\boldsymbol{w})\|_\nu^2 \right] \leq \mathbb{E}_q \left[\|B(\boldsymbol{w})\|_D^2 \right] - \mathbb{E}_q \left[v(\boldsymbol{w}) \right] + \frac{\lambda}{N} KL(q||p) + 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \tag{14}$$

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

Proof. From Hoeffding's inequality we have:

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

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

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

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

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

where $v(\boldsymbol{w}) \triangleq \mathbb{E}_\nu \left[Var_{\mathcal{P}} \left[b(\boldsymbol{w}) \right] \right]$. Thus:

$$\left| \|B(\boldsymbol{w})\|_\nu^2 + v(\boldsymbol{w}) - \|B(\boldsymbol{w})\|_D^2 \right| \leq 4 \frac{R_{max}^2}{(1-\gamma)^2} \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \tag{16}$$

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

Find a reference for this

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

Thus, multiplying both sides of (16) by $\lambda^{-1} N$ and applying the change of measure inequality with $f(\boldsymbol{w}) = \lambda^{-1} N \left| \|B(\boldsymbol{w})\|_\nu^2 + v(\boldsymbol{w}) - \|B(\boldsymbol{w})\|_D^2 \right|$, we obtain:

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

where the second inequality holds since the right-hand side of (16) does not depend on $\boldsymbol{w}$. Finally, we can explicitly write:

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

from which the lemma follows straightforwardly. □

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 $\widehat{\xi}$ is the minimizer of $ELBO(\xi) = \mathbb{E}_{q_\xi}\left[\|B(\boldsymbol{w})\|_D^2\right] + \frac{\lambda}{N}KL(q_\xi||p)$ for a dataset D of N samples. Define $v(\boldsymbol{w}) \triangleq \mathbb{E}_\nu\left[Var_{\mathcal{P}}\left[b(\boldsymbol{w})\right]\right]$. Then, for any $\delta > 0$, with probability at least $1-\delta$:*

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

Proof. Let us use Lemma 2 for the specific choice $q = q_{\widehat{\xi}}$. From Eq. (14), we have:

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

where the second inequality holds since $v(\boldsymbol{w}) > 0$, while the equality holds from the definition of $\widehat{\xi}$. We can now use Eq. (15) to bound $\mathbb{E}_{q_\xi}\left[\|B(\boldsymbol{w})\|_D^2\right]$, thus obtaining:

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

This concludes the proof. □