Active Transfer Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 Introduction

2 Preliminaries

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: $\max_\pi J(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid \mathcal{M}, \pi]$. 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., [4]) 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 $\frac{R_{max}}{1-\gamma}$.

We define the Bellman error (or Bellman residual) of a Q-function Q as $B(Q) \triangleq TQ - Q$. Notice that a Q-function Q is optimal if, and only if, $||B(Q)||_\infty = 0$.

Multitask Settings We represent tasks τ as MDPs with shared state and action spaces, but with potentially different values for all other parameters. We assume the existence of a distribution $\mathcal{D}$ over tasks, i.e., $\tau \sim \mathcal{D}$, and we suppose that we are able to sample from such distribution.

3 Approach

We start by noticing that the task distribution $\mathcal{D}$ clearly induces a distribution over optimal Q-functions. Then, our goal is to estimate such distribution from the set of source tasks and use it as a prior for speeding up the learning process in the target task.

We assume states and actions to be drawn from a fixed joint distribution μ. We define the set of Q-functions of our interest as the set $\mathcal{Q}^\epsilon$ of all Q functions whose Bellman error $B(Q)$ is in some ϵ-ball defined by the l_p-norm $|| \cdot ||_{p,\mu}$:

$$\mathcal{Q}^\epsilon = \left\{ Q \in \mathcal{Q} \mid ||B(Q)||_{p,\mu}^p \leq \epsilon \right\} \tag{1}$$

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

Given a dataset of N samples $D = \langle s_i, a_i, r_i, s'_i \rangle_{i=1}^N$, we can approximate the l_p-norm of the Bellman error of a Q-function Q as:

$$||B(Q)||_{p,D}^p = \frac{1}{N} \sum_{i=1}^{N} \left| r_i + \gamma \max_{a'} Q(s'_i, a') - Q(s_i, a_i) \right|^p \tag{2}$$

Theorem 1. *Let Q be a Q-function with empirical Bellman error, computed on a dataset D of N i.i.d. samples, given by $||B(Q)||_{p,D}^p = \hat{q}$. Then, for any $\epsilon >= 0$:*

$$P\left(Q \in \mathcal{Q}^\epsilon \mid ||B(Q)||_{p,D}^p = \hat{q}\right) \leq exp\left(-\frac{2N \max\{\epsilon - \hat{q}, 0\}^2}{\left(\frac{2R_{max}}{1-\gamma}\right)^{2p}}\right) \tag{3}$$

Proof. Assume $\hat{q} > \epsilon$. Then:

$$\begin{aligned} P\left(Q \in \mathcal{Q}^\epsilon \mid ||B(Q)||_{p,D}^p = \hat{q}\right) &= P\left(||B(Q)||_{p,\mu}^p \leq \epsilon \mid ||B(Q)||_{p,D}^p = \hat{q}\right) \\ &= P\left(||B(Q)||_{p,\mu}^p - \hat{q} \leq \epsilon - \hat{q} \mid ||B(Q)||_{p,D}^p = \hat{q}\right) \end{aligned}$$

Notice that $\mathbb{E}[\hat{q}] = ||B(Q)||_{p,\mu}^p$ and that $\epsilon - \hat{q} < 0$ by assumption. Then, we can apply Hoeffding's inequality to write:

$$P\left(Q \in \mathcal{Q}^\epsilon \mid ||B(Q)||_{p,D}^p = \hat{q}\right) \leq exp\left(-\frac{2N(\epsilon - \hat{q})^2}{\left(\frac{2R_{max}}{1-\gamma}\right)^{2p}}\right)$$

Finally, when $\hat{q} \leq \epsilon$, the probability can be straightforwardly upper-bounded by 1. Combining the two results concludes the proof. □

Theorem 2. *Let Q be a Q-function with empirical Bellman error, computed on a dataset D of N i.i.d. samples, given by $||B(Q)||_{p,D}^p = \hat{q}$. Then, for any $\epsilon >= 0$:*

$$P\left(||B(Q)||_{p,D}^p = \hat{q} \mid Q \in \mathcal{Q}^\epsilon\right) \leq \frac{\epsilon}{\hat{q}}$$

Proof.

$$\begin{aligned} P\left(||B(Q)||_{p,D}^p = \hat{q} \mid Q \in \mathcal{Q}^\epsilon\right) &= P\left(||B(Q)||_{p,D}^p = \hat{q} \mid ||B(Q)||_{p,\mu}^p \leq \epsilon\right) \\ &\leq P\left(||B(Q)||_{p,D}^p \geq \hat{q} \mid ||B(Q)||_{p,\mu}^p \leq \epsilon\right) \\ &\leq \frac{E[||B(Q)||_{p,D}^p]}{\hat{q}} \\ &\leq \frac{\epsilon}{\hat{q}} \end{aligned}$$

The first inequality is straightforward, the second one is from Markov's inequality, while the third one is due to the fact that $Q \in \mathcal{Q}^\epsilon$. □

3.1 Regularized Bellman Residual Minimization with Gaussian Priors

Consider solving the target task given a dataset D of N samples. The posterior distribution over optimal Q-functions is:

$$P(Q^* \mid D) \propto P(D \mid Q^*) P(Q^*) \tag{4}$$

Bounding the likelihood $P(D \mid Q^*)$ (FIND A BETTER DERIVATION) and taking the maximum of the log-posterior, we obtain the following optimization problem:

$$\min_{Q \in \mathcal{Q}} ||B(Q)||_{p,D}^p - \log P(Q) \tag{5}$$

Then, let us specify a particular hypothesis space $\mathcal{Q}$. We consider Q-functions $Q_{\boldsymbol{w}}$ parameterized by the vector $\boldsymbol{w}$. Our optimization problem becomes:

$$\min_{\boldsymbol{w}} ||B(\boldsymbol{w})||_{p,D}^{p} - \log P(\boldsymbol{w}) \tag{6}$$

We model the prior distribution over the optimal parameters $\boldsymbol{w}$ as a Gaussian. That is, we assume:

$$P(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7}$$

Then, our optimization, adopting the l_2-norm, becomes:

$$\min_{\boldsymbol{w}} ||B(\boldsymbol{w})||_{D}^{2} + ||\boldsymbol{w} - \boldsymbol{\mu}||_{\boldsymbol{\Sigma}} = \min_{\boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{N} |y_i - Q_{\boldsymbol{w}}(s_i, a_i)|^2 + ||\boldsymbol{w} - \boldsymbol{\mu}||_{\boldsymbol{\Sigma}} \tag{8}$$

where $y_i = r_i + \gamma \max_{a'} Q_{\boldsymbol{w}}(s_i', a')$ and $||\boldsymbol{x}||_{\boldsymbol{A}} \triangleq \boldsymbol{x}^T \boldsymbol{A}^{-1} \boldsymbol{x}$ for $\boldsymbol{A}$ positive definite matrix.

Linear model We assume a linear model for the Q-functions: $Q_{\boldsymbol{w}}(s, a) = \boldsymbol{w}^T \boldsymbol{\phi}(s, a)$. Here $\boldsymbol{\phi}$ is a K-dimensional feature vector. Then, the solution to the optimization problem of Eq. (8) can be computed in closed form as follows:

$$\boldsymbol{w}^* = \left(\boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\boldsymbol{A}^T \boldsymbol{b} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right) \tag{9}$$

where $\boldsymbol{A}$ is an $N \times K$ matrix containing the feature vectors at each data point (s_i, a_i) and $\boldsymbol{b}$ is an N-dimensional vector containing their targets y_i.

4 Related Works

List of any paper that relates to our approach together with a brief description:

- [3]: the authors propose a method for efficient exploration via randomized value functions. The optimal Q-function is computed by bayesian LSVI and, after each update, parameters are sampled from the posterior. Then, the agent follows a greedy policy with respect to the sampled Q-function. Regret bounds are provided.
- [2]: the authors extend the idea of randomized value functions to drive exploration in deep RL. A posterior distribution over Q-functions is approximated via bootstrapping. In each episode, the agent acts greedily with respect to a Q-function sampled from the approximated posterior.
- [1]: the authors build on top of bootstrapped DQNs to provide UCB-like exploration bonuses. In a previous (?) version of the paper, exploration bonuses based on information gain are also proposed.

5 Experiments

6 Conclusion

References

[1] Richard Y Chen, John Schulman, Pieter Abbeel, and Szymon Sidor. Ucb and infogain exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

[2] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016.

[3] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

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

A Proofs