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 Background

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

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

3.1 Algorithm

3.2 Gaussian Variational Transfer

3.3 Mixture of Gaussian Variational Transfer

4 Theoretical Analysis

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 of Theorems