CS 598 Stats RL
Homework 3

October 27, 2020

- **Submission deadline: Nov 11 (Wed) before class.**
- Submission website: compass2g.

1. Bisimulation and completeness
Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP and $\phi : \mathcal{S} \to \mathcal{S}_\phi$ be a state abstraction. Let $\mathcal{F}^\phi$ be the set of all possible functions over $\mathcal{S} \times \mathcal{A}$ with value range $[0, V_{\max}]$ ($V_{\max} = R_{\max}/(1-\gamma)$) that are piece-wise constant under ϕ. (That is, for any $f \in \mathcal{F}^\phi$, $\forall s^{(1)}, s^{(2)}$ such that $\phi(s^{(1)}) = \phi(s^{(2)})$, we always have $f(s^{(1)}, a) = f(s^{(2)}, a), \forall a \in \mathcal{A}$.

Prove that the following two conditions are equivalent:

1. ϕ is a bisimulation for M.
2. F^ϕ is closed under Bellman update, that is, $\forall f \in \mathcal{F}^\phi$, $\mathcal{T} f \in \mathcal{F}^\phi$, where $\mathcal{T}$ is the Bellman update operator of M.

Hint: (1) $\Rightarrow$ (2) is straightforward. For (2) $\Rightarrow$ (1), try to prove that $\neg$ (1) $\Rightarrow \neg$ (2). That is, if ϕ is not a bisimulation, you should be able to find $f \in \mathcal{F}^\phi$ such that $\mathcal{T} f \notin \mathcal{F}^\phi$.

2. Linear MDPs
Linear MDPs have been a popular setting in recent theoretical RL works. In this problem you will be asked to establish some essential properties of linear MDPs. First, a linear MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ is one such that for any (s, a, s'), we have $P(s'|s, a) = \phi(s, a)^\top \psi(s')$, where ϕ and ψ are two feature maps that map (s, a) and s' respectively to d-dimensional real vectors. (**Caution:** ϕ here has nothing to do with the ϕ in the first problem. The (re)use of the notation is merely a coincidence and follows the convention in the literature.) In other words, the transition matrix P has low rank and can be factorized into the product of two matrices, $\Phi \times \Psi$, where Φ has $\phi(s, a)^\top$ as its rows and Ψ has $\psi(s')$ as its columns. It is also typically assumed that

- $R(s, a) = \phi(s, a)^\top \theta_R$, that is, reward is also linear in ϕ.
- $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is known to the learner, but ψ is unknown. (When ϕ is also unknown, it is often called a low-rank MDP.)

Let $\mathcal{F} := \{(s, a) \mapsto \phi(s, a)^\top \theta : \theta \in \mathbb{R}^d)\}$, i.e., the linear function space w.r.t. feature map ϕ. Prove the following:

1. $\mathcal{F}$ is closed under $\mathcal{T}$ and $\mathcal{T}^\pi$ for any π.
2. For any policy π, let d_t^π be the t-th step state distribution induced by π from d_0. If $d_0(s') = \psi(s')^\top \theta_0$, i.e., d_0 is linear in ψ as feature, then d_t^π is also linear in ψ.

3. Error Propagation in Policy Evaluation

Suppose we have an algorithm that approximates Q^π for a given π by producing a sequence of functions $f_1, f_2, \ldots, f_k, \ldots$. Let ϵ be such that for any $t \geq 1$, $\|f_k - \mathcal{T}^\pi f_{k-1}\|_{2,\mu} \leq \epsilon$ for a fixed μ distribution. In other words, a small ϵ implies that the algorithm is very close to the procedure $f_k \leftarrow \mathcal{T}^\pi f_{k-1}$, i.e., the analogy of value iteration for solving Q^π.

Let d_t^π be the t-th step state-action distribution induced by π from initial state distribution d_0. Assume that there exists $C < \infty$, such that for any t, $\max_{s,a} d_t^\pi(s,a)/\mu(s,a) \leq C$. Prove a bound on $|\mathbb{E}_{s\sim d_0}[f_k(s,\pi)] - \mathbb{E}_{s\sim d_0}[Q^\pi(s,\pi)]|$ as a function of π.

Hint: $|\mathbb{E}_{s\sim d_0}[f_k(s,\pi)] - \mathbb{E}_{s\sim d_0}[Q^\pi(s,\pi)]| = |\mathbb{E}_{s\sim d_0}[f_k(s,\pi) - (\mathcal{T}^\pi f_{k-1})(s,\pi) + (\mathcal{T}^\pi f_{k-1})(s,\pi) - (\mathcal{T}^\pi Q^\pi)(s,\pi)]|$.

Remark: This problem here is similar to the FQI analysis we did in class, but simplified in the sense that it does not care about the details of the algorithm and merely concerns how the per-iteration error ϵ affects the final accuracy; for any concrete algorithm, if we can prove a bound on ϵ (using, say, sample size, among other factors), we can directly plug the bound into this analysis to obtain a final guarantee. Error propagation in policy evaluation is also simpler than that in policy optimization because we only need to care about the distribution induced by π, the policy of interest, whereas in FQI we need to consider the distributions induced by many different policies.

Optional: What if we do not assume that $\max_{s,a,t} d_t^\pi(s,a)/\mu(s,a) \leq C$, but only assume that $\max_{s,a} d^\pi(s,a)/\mu(s,a) \leq C'$? Hint: upper bound $\max_{s,a} d_t^\pi(s,a)/\mu(s,a)$ using an expression that depends on both C' and t.