Answers to Assignment Putterman 6 for AML fall 2023

Gersi Doko

October 25, 2023

1 Problem 6.3 From Putterman 2005

$$|r(s,a)| \leq M \in \mathcal{R} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \implies ||v^*||_\infty \leq \frac{M}{1-\gamma}$$

Proof. Begin by observing that for any occupancy frequency $u \in \{u \in \mathcal{R}_+^{SA} | \sum_{a \in \mathcal{A}} (I - \gamma P_a^T) u(\cdot, a) = p_0\}$ $\sum_{(s,a)} u(s,a) = \frac{1}{1-\gamma}$. By solving the dual LP formulation found in Putterman 2005, we obtain...

$$u^* \in \arg\max_{u \in \mathcal{U}} r^T u$$

As noted in the book, $v^* = v_{\pi^*}$ for an optimal policy π^*, and $v_{\pi^*} = r^T u^*$. Therefore, since $|r(s,a)| \leq M$, we have...

$$||v^*||_\infty = ||v_{\pi^*}||_\infty = ||r^T u^*||_\infty \leq ||r||_\infty ||u^*||_1 \leq M ||u^*||_1 = M \sum_{(s,a)} u^*(s,a) = \frac{M}{1-\gamma}$$

□

2 Problem 6.11 From Sutton and Barto 2018

Problem 6.11 asks why Q-learning is considered off-policy, when something like SARSA is considered on-policy. Or at least this is how I choose to interpret the question. The answer is due to the inner maximization over actions that occurs in Q-learning. This maximization is used to collect the maximum value of the next state-action pair, and is not necessarily the action that an agent would take. In SARSA, I can use an epsilon greedy policy (or any other) to collect my online SARSA samples, however in Q-learning, I must use a greedy policy with respect to the learned Q function. This being said SARSA can also be adapted to the off policy setting, by using a behavior policy to collect samples, and then using importance sampling to correct for the difference in the behavior policy and the target policy.