Homework 1: Dynamic Programming & Sample Complexity

CSE 599: Reinforcement Learning and Bandits

Instructions

Do any two of the five problems.

1 The Discounted State Distribution

1. Show that:
$$(I - \gamma P^\pi)^{-1}\mathbb{1} = (1-\gamma)^{-1}\mathbb{1}$$
where $\mathbb{1}$ is the vector of all ones.
2. Write an expression for $\Pr(s_t = s', a_t = a' | s_0 = s, a_0 = a)$ in terms of the transition model P. You should write this as a matrix of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$, where the $(s,a),(s',a')$ entry is $\Pr(s_t = s', a_t = a' | s_0 = s, a_0 = a)$.
3. Show that:
$$[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s', a_t = a' | s_0 = s, a_0 = a)$$
This is often referred to as the *discounted state visitation distribution*.

2 Bellman Consistency of the Variance

For any policy π in an MDP M, show that:
$$\Sigma^\pi = \gamma^2 \mathrm{Var}_P(V^\pi) + \gamma^2 P^\pi \Sigma^\pi\,,$$
where P is the transition model in the MDP M (and we have dropped the M subscripts).

Variance and the Doob martingale: If you are familiar with martingales, you may find it natural to think about the concepts above in terms of the Doob martingale based on the random variable $Z = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)$. If you are not familiar with martingales, then not to worry as the above will give you insights into this concept.

3 A Crude Value Bound

Let us now prove a crude bound on the optimal action value function (the proof of this case is not covered in the notes).

Let $\delta > 0$. Show that with probability greater than $1 - \delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}\,.$$

4 Component-wise Bounds

Provide a proof of the one of cases we needed in order to prove our sample complexity result. Show that:

$$Q^\star - \widehat{Q}^\star \geq \gamma(I - \gamma\widehat{P}^{\widehat{\pi}^\star})^{-1}(P - \widehat{P})V^\star$$

5 A worst-case example

Provide an example that shows the worst case bound from Lecture 1, on the suboptimality of the greedy policy itself, is (nearly) tight. In particular, specify an MDP M (the transition model P and the reward function r), such that for every γ and ε, you show there is vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $\|Q - Q^\star\|_\infty = \varepsilon$ and such that:

$$V^{\pi_Q} \leq V^\star - \frac{\varepsilon}{1-\gamma}\mathbb{1}.$$

where $\mathbb{1}$ denotes the vector of all ones. In other words, you should be specifying your Q as a function of Q^*, ε and γ. (Note that Q^* will be a function of γ).

(*Hint:* It is possible to do this with just two states and two actions, so that $Q \in \mathbb{R}^4$. The idea of this simple "worst-case" MDP is that it should give you insight into how errors accumulate. It might help to think of a two state MDP where one (suboptimal) action is absorbing at one of the two states.)