Homework 1: Linear Programming & Sample Complexity

CS 6789: Foundations of Reinforcement Learning

0 Instructions

For each question in this HW, please list all your collaborators and reference materials (beyond those specified on the website) that were used for this homework. Please add your remarks in a "Question 0".

1 The (Discounted) State-Action Visitation Measure (25 Points)

1. (10 Points) Show that:
$$(I - \gamma P^\pi)^{-1}\mathbb{1} = (1-\gamma)^{-1}\mathbb{1}$$
where $\mathbb{1}$ is the vector of all ones.
2. (5 Points) 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. (10 Points) Show that:
$$[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{h=0}^{\infty} \gamma^t \mathbb{P}_h^\pi(s_h = s', a_h = a' | s_0 = s)$$
This rows of this matrix are often referred to as *discounted state-action visitation measures* (or state-action visitation distributions); we can view the (s,a)-th row of this matrix as an induced distribution over states and actions when following π after starting with $s_0 = s$ and $a_0 = a$.

2 Linear Programming for MDPS (25 Points)

1. (10 Points) Consider the following linear programming that we covered in the lecture:
$$\min_{V \in \mathbb{R}^S} \sum_s \mu(s) V(s), \quad s.t., V(s) \geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s'), \forall a \in \mathcal{A}, s \in \mathcal{S}.$$
Here we assume $\mu(s) > 0$ for all s. Prove that $V^\star$ is the unique solution to the above LP.

2. (5 points) Let us now consider a modified definition of the average state-action visitation measure: $d^{\pi}_{s_0}(s,a) := (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}^{\pi}_h(s_h = s, a_h = a|s_0)$, with respect to for a fixed start state s_0 and a stationary policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$. [1] Prove that:

$$\sum_a d^{\pi}_{s_0}(s,a) = (1-\gamma)\delta(s_0) + \gamma \sum_{s',a'} d^{\pi}_{s_0}(s',a')P(s|s',a'), \forall s$$

   Here $\delta(s_0)$ is the delta distribution, i.e., $\delta(s_0) = 1$ and 0 for any other state.
3. (No answer needed) Observe that we can write $V^{\pi}(s_0) = \frac{1}{1-\gamma} d^{\pi}_{s_0} \cdot r$ where we can view $d^{\pi}_{s_0}$ and r as vectors of length $|\mathcal{S}| \cdot |\mathcal{A}|$, i.e. the value is a linear functions of the state-action measure.
4. (10 Points) Consider the following polytope:

$$\mathcal{K} = \{v \in \Delta(\mathcal{S} \times \mathcal{A}) : \sum_a v(s,a) = (1-\gamma)\delta(s_0) + \gamma \sum_{s',a'} v(s',a')P(s|s',a'), \forall s\}.$$

   Consider any $v \in \mathcal{K}$. Denote the stationary policy as $\pi_v(a|s) = \frac{v(s,a)}{\sum_{a' \in \mathcal{A}} v(s,a)}, \forall s, a$. Prove that we have $v(s,a) = d^{\pi}_{s_0}(s,a), \forall s, a$.
   (Hint: Directly work on $v(s,a) - d^{\pi}_{s_0}(s,a)$, and use recursion. The whole process consists of straight equalities.)
5. (No answer needed) Equipped with what you just showed, read (and feel free to interpret) the formulation of the dual LP in the book.

3 Bellman Consistency of the Variance (25 Points)

For any policy π in an MDP M, let us define

$$\Sigma^{\pi}(s,a) \triangleq \mathbb{E}\left[\left|\sum_{t \geq 0} \gamma^t r(s_t,a_t) - Q^{\pi}(s,a)\right|^2 \Big| s_0 = s, a_0 = a\right]$$

as the variance of the sum of discounted rewards for the sequence of state-action pairs, $\{(s_0,a_0),(s_1,a_1),\ldots\}$. Furthermore, let us define

$$\mathrm{Var}_{y\sim\rho}(f(y)) \triangleq \mathbb{E}_{y\sim\rho}\left[|f(y) - \mathbb{E}_{y\sim\rho}[f(y)]|^2\right]$$

as the variance of a real-valued function $f : \mathrm{Y} \to \mathbb{R}$ under the probability distribution ρ. Given these definitions, show that for any policy π, the variance Σ^{π} satisfies the following Bellman-like recursion.

$$\Sigma^{\pi} = \gamma^2 \mathrm{Var}_P(V^{\pi}) + \gamma^2 P^{\pi}\Sigma^{\pi}\,,$$

where P is the transition model in the MDP M (and we have dropped the M subscripts).

[1] Note that the modification from the definition in Problem 1 is that here we are starting at a fixed state s_0 and then follow π, while the latter starts with s_0, a_0 and then we follow π. We could denote the latter definition by $d^{\pi}_{s_0,a_0}(s,a)$.

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.

Minimax Optimal Sample Complexity: The Bellman consistency condition for the variance is a key lemma in obtaining the minimax optimal sample complexity. This lemma, along with the "Weighted Sum of Deviations" Lemma (see the book) provide much of the insights for how to achieve minimax optimal sample complexity. For a mastery of the material, please read Chapter 2 and the proof sketch in the slides.

4 A Worst-case Example of ℓ_∞ Error Amplification (25 Points)

Provide an example that shows the worst case bound from Lecture 2, 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 $\|Q - Q^\star\|_\infty = \epsilon$ 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.)