Homework 1: Linear Programming & Sample Complexity

CS 6789: Foundations of Reinforcement Learning

Due Feb 21 11:59pm ET

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 (20 Points)

1. (5 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^π. 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:
$$[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{h=0}^{\infty} \gamma^t \mathbb{P}_h^\pi(s_h = s', a_h = a' | s_0 = s, a_0 = a)$$
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 (20 Points)

1. (5 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 $\pi(a|s) = \frac{v(s,a)}{\sum_{a' \in \mathcal{A}} v(s,a)}, \forall s, a$. 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 (20 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. For $V^{\pi} \in \mathbb{R}^{|S|}$, we define the vector $\mathrm{Var}_P(V^{\pi}) \in \mathbb{R}^{|S||A}$ as:

$$\mathrm{Var}_P(V^{\pi})(s,a) \triangleq \mathrm{Var}_{P(\cdot|s,a)}(V^{\pi}) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[(V^{\pi}(s') - \mathbb{E}_{s''\sim P(\cdot|s,a)} V^{\pi}(s''))^2]$$

Given these definitions, show that for any policy π, the variance Σ^{π} satisfies the following Bellman-like recursion.

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

[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 $d^{\pi}_{s_0,a_0}(s,a)$.

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.
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 (20 Points)

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 $\|Q - Q^\star\|_\infty = \epsilon$ 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.)

5 Learning Transition Models (20 Points)

In this problem, we are going to bound the model error, i.e., $\|\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a)\|_1$ for all s, a, where $\widehat{P}$ is the learned model and $P^\star$ is the ground truth model. Let us assume that we have a dataset $\mathcal{D} = \{s_i, a_i, s_i'\}_{i=1}^N$ which we use to learn $\widehat{P}$ as follows:

$$\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^N \mathbf{1}\{(s_i, a_i, s_i') = (s, a, s')\}}{N(s,a)}, \quad N(s,a) = \sum_{i=1}^N \mathbf{1}\{(s_i, a_i) = (s,a)\},$$

where for simplicity let us assume that $N(s,a) > 0, \forall s, a$. We will assume that given (s,a), the next state s' is sampled from $P^\star(\cdot|s,a)$ in an i.i.d fashion.

Q1 (5 points): Prove that for all s, a, $\|\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a)\|_1 = \max_{f:\mathcal{S}\mapsto[-1,1]}(\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a))^\top f$, where we treat $P(\cdot|s,a)$ as a vector in $\mathbb{R}^S$ and f as a vector in $\mathbb{R}^S$ as well.

Q2 (5 points): Now let us consider a fixed function $f \in \mathcal{S} \mapsto [-1, 1]$. Prove that with probability at least $1 - \delta$, for ALL s, a, we have:

$$\left|(\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a))^\top f\right| \leq 2\sqrt{\frac{\ln(SA/\delta)}{N(s,a)}}.$$

Here you will use Hoeffding's inequality and union bound.

Note that in Q2, the result only applies to the fixed function f we choose. Now we want to derive the model error that works for all $f \in \mathcal{S} \mapsto [-1, 1]$ so that using the result in Q1, we will get the model error under the ℓ_1 norm. From now on, since we only have finite number of states, let us consider f as a vector in $[-1, 1]^S$, i.e., a vector whose length is S and each element of the vector is $[-1, 1]$.

If the space $[-1, 1]^S$ were discrete, then we would have done by just applying a union bound over all possible vectors there. Unfortunately, $[-1, 1]^S$ is a continuous space, so we cannot do a union bound directly over that space. We will work out the standard solution to deal with this situation, which is the ϵ-net covering argument. Below is the standard ϵ-net definition.

Definition 1 (ϵ-net (under ℓ_2 norm))**.** Consider a compact set $\mathcal{R}$ (for simplicity, let us just assume $\mathcal{R} \subset \mathbb{R}^d$ for some d). An ϵ-net $\mathcal{N}_\epsilon \subset \mathcal{R}$ is a discrete set such that for any point $v \in \mathcal{R}$, there exists a point $v' \in \mathcal{N}_\epsilon$, such that $\|v - v'\|_2 \leq \epsilon$.

Basically, we can think about $\mathcal{N}_\epsilon$ as a type of discretization of the space $\mathcal{R}$ where the discretization resolution is roughly ϵ (i.e., any point in $\mathcal{R}$ is always covered by a point in $\mathcal{N}_\epsilon$ that is ϵ close in the ℓ_2 norm metric).

Now let us consider $\mathcal{R}$ to be a ball in $\mathbb{R}^d$ with radius equal to $R \in \mathbb{R}^+$, i.e., $\mathcal{R} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$. The following proposition shows that there always exists an ϵ-net whose size is at most $(1 + 2R/\epsilon)^d$.

Lemma 1. *For $\mathcal{R} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$, there exists an ϵ-net $\mathcal{N}_\epsilon$ such that $|\mathcal{N}_\epsilon| \leq (1 + 2R/\epsilon)^d$.*

We are not going to prove this lemma, but intuitively the exponent d is not hard to see: if we discrete each dimension, then roughly the number of points on the discrete grid is $\exp(d)$. In general computing the smallest ϵ-net is hard, but approximation algorithm exists. Note that in this problem, we only use an ϵ-net inside the analysis, i.e., we only need its existence, and there is no need to construct it explicitly.

The benefit of ϵ-net is that it is a finite set, which allows us to apply a union bound over all points inside it. We also know that any point in $\mathcal{R}$ is ϵ-close to a point in $\mathcal{N}_\epsilon$, so hopefully via tuning the magnitude of ϵ, we are able to build a uniform convergence bound for ALL points in $\mathcal{R}$. This covering argument is a very standard tool in learning theory to derive uniform convergence bounds over a continuous set.

Q3 (5 points): Denote $\mathcal{N}_\epsilon$ as the ϵ-net of $[-1,1]^S$. Show that with probability at least $1-\delta$, for ALL s, a and ALL $f \in [-1,1]^S$, we have:

$$\left|(\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a))^\top f\right| \leq 2\sqrt{\frac{S\ln(SA(1+2\sqrt{S}/\epsilon)/\delta)}{N(s,a)}} + 2\epsilon.$$

Q4 (5 points): Since $\epsilon > 0$ is a free parameter, now we can set ϵ to be a small number. Particularly, we can set $\epsilon = 1/(2N)$ where N is the total number of samples and we know that $N \geq N(s,a)$ (this means that $1/N \leq 1/N(s,a) \leq 1/\sqrt{N(s,a)}$. Now substitute $\epsilon = 1/(2N)$ into the above bound, prove the following: with probability at least $1-\delta$, for ALL s, a, we have:

$$\begin{aligned}
&\|\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a)\|_1 = \text{argmax}_{f\in[-1,1]^S}\left|(\widehat{P}(\cdot|s,a) - P^\star(\cdot|s,a))^\top f\right| \\
&\leq O\left(\sqrt{\frac{S\ln(SA(1+\sqrt{S}N)/\delta)}{N(s,a)}}\right) = \widetilde{O}\left(\sqrt{\frac{S}{N(s,a)}}\right),
\end{aligned}$$

where $O(\cdot)$ ignores absolute constants, and $\widetilde{Q}(\cdot)$ notation ignores log terms.