Homework 1: Statistical Foundations of Reinforcement Learning (Spring 2021)

University of California, Santa Barbara

Assigned on April 6, 2020 (Tuesday)

Due at 11:59 pm on Apr 26, 2020 (Monday)

Notes:

- Be sure to read "Policy on Academic Integrity" on the course syllabus.
- There are ***[100 points]*** in this homework, and a bonus ***[20 points]***.
- You need to submit your homework via Gradescope.
- Contract the instructor if you spot typos. Any updates or correction will be posted on the course Announcements page and piazza, so check there occasionally.

---

0. Acknowledgment *[0 points]*

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.

1. **List of Collaborators** List the names of all people you have collaborated with and for which question(s).
2. **List of Acknowledgements.** If you find an assignment's answer or use a another source for help, acknowledge for which question and provide an appropriate citation (there is no penalty, provided you include the acknowledgement). If not, then write "none".

***Attention:** Question 1-3 are taken from Wen Sun and Sham Kakade (Thanks both!). The indexing is slightly different from what we used in the lecture, specifically, we have $h = 0, 1, 2, ...$ in this homework. The initial states are S_0, A_0, rather than S_1, A_1 that we considered.

1 The (Discounted) State-Action Visitation Measure (20 Points)

(a) (5 Points) Show that:

$$(I - \gamma P^\pi)^{-1}\mathbf{1} = (1 - \gamma)^{-1}\mathbf{1}$$

where $\mathbf{1}$ is the vector of all ones.

(b) (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 $\Pr(S_t = s', A_t = a' | S_0 = s, A_0 = a)$.

(c) (10 Points) Show that:

$$[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}^\pi(S_h = s', A_h = a' | S_0 = s, A_0 = a)$$

(In the above $\mathbb{P}$ is the general operator for probability, the superscript π indicates that the distribution is induced by running policy π.)

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)

(a) (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.

(b) (10 points) (Scaled occupancy measure) Let us now consider a modified definition of the average state-action visitation measure: $d^\pi_{s_0}(s,a) := (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}^\pi_h(S_h = s, A_h = a | S_0 = s_0)$, 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.

(Check that after the rescaling, unlike the occupancy measure ν^π defined in the lectures, d^π is a valid probability distribution.)

Remark: 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.

(c) (10 points **Bonus**) Derive the dual LP from the primal LP by following the standard steps to derive a Lagrange dual. Write down the KKT conditions, and how you can convert the dual optimal solutions to the primal optimal solution.

(If you have not taken a course on convex optimization before, you may just skip this question.)

[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 $d^\pi_{s_0,a_0}(s,a)$. It is different from the "occupancy measure" from the lecture by a constant scaling factor.

(d) (10 Points **Bonus**) 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: You can directly work on $v(s,a) - d^{\pi}_{s_0}(s,a)$, and use recursion. The whole process consists of straight equalities. Note that the results apply to all policies, rather than only the optimal policy, that you have established via strong duality in Part (c).)

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. 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).

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 Bernstein's inequality for bounding the sample complexity of MDP with a Generative Model (20 Points)

A key step in the Attempt 1 (via Simulation Lemma and Uniform convergence) of Lecture 3 was to bound $\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1$. We applied the McDiarmid inequality. In this question we will explore the alternatives.

(a) (10 Points) Try obtaining a high probability bound of $\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1$ by applying the Bernstein inequality to bound $|P(s'|s,a) - \hat{P}(s'|s,a)|$ for every $s' \in \mathcal{S}$. For what ranges of target sub-optimality ϵ do you get the same sample complexity bound as in our McDiarmid inequality approach (in big O notation)?

(b) (10 Points) For Attempt 2, the key trick to save a factor of S was to construct the iid univariate random variables (see Slide 29). We applied Hoeffding's inequality there. Apply the Bernstein's inequality instead in this question to obtain an improved high probability bound of $\|Q^* - \hat{Q}^*\|_\infty$.

Then by the Q-error amplification lemma, work out the sample complexity bound.

Be sure to specify which random variable it is when you apply the Bernstein's inequality and what you are conditioning on.

(Hint: You could use the Bellman equation of the variance from Problem 3. If you are stuck, you may refer to Section 2.3 of the AJKS book, which should contain all details needed for you to complete this question.)

5 Simulation Lemma in more general settings (20 Points)

(a) (5 Points) In the lectures we considered the case when the reward function is known. I promised you that the unknown reward case does not change the sample complexity. We will work it out a simulation lemma in the case when the *expected* reward $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ needs to be estimated. Let $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{P}, \hat{r}, \gamma, \mu)$, and $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$.

Generalize the simulation lemma (Slide 16 in Lecture 3) to the case when $\hat{Q}^\pi$ is computed by $\hat{M}$.

(b) (5 points) Consider the generative model where we can sample S', R in $O(1)$ time with an arbitrary chosen s, a pair, and we collect N data points $S'_{i,s,a}, R_{i,s,a}$ for each s, a. Assume $0 \le R \le 1$, apply Hoeffding's inequality / union bound for estimating r from the observed data and derive the corresponding sample complexity bound for Attempt 1 (you do not have to reproduce the McDiarmid Inequality part of the argument, just say what difference it makes when r is replaced with $\hat{r}$ in $\hat{M}$ now.)

(c) (5 points) Sometimes it is convenient to make use of the more general MDP definition when $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, i.e., the reward depends on not just the state and action you take, but also the state you transition into. Does this affect the results you get in Part (a) and (b). And why?

(d) (5 points) Consider the finite-horizon episodic MDP setting with stationary transition dynamics / and reward design. Write down a simulation lemma for Q_h^π for all $h = 1, ..., H$. The approximate MDP $\hat{M}$ has $\hat{P}$, $\hat{r}$ and $\hat{\mu}$ being potentially different from that of M.

(Hint: Note that the Q functions are different at each time point h. You should go from $h = H$ and work backwards.)