Homework 3: Statistical Foundations of Reinforcement Learning (Spring 2021)

University of California, Santa Barbara

Assigned on May 17, 2021 (Monday)

Due at 11:59 pm on June 2, 2021 (Wednesday)

Notes:

- Be sure to read "Policy on Academic Integrity" on the course syllabus.
- There are ***[100 points]*** in this homework, and a bonus ***[30 points]***.
- You only have to do **either Q4 or Q5**. If you do both, then Q5 will be considered bonus 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".

***Acknowledgment:** Q1 and some parts of Q2 are taken from Wen Sun and Sham Kakade.

1 Exploration in Absorbing MDPs (30 points)

Let's first recall Problem 3 in HW2 where we constructed absorbing MDPs and have shown that the absorbing MDP can be used to provide optimism. We complete the rest exercise regarding absorbing MDPs in this section.

Specifically, we will see the Explore-or-Terminate phenomena, where the algorithm either has already found a near optimal policy—hence we terminate, or the algorithm can guarantee to explore new state-action pairs outside the current known set.

1.1 Exploration or Termination

Note that we know $\widehat{P}^\dagger$, as it uses our learned model $\widehat{P}$ and it has deterministic absorbing structure at unknown state-action pairs. So we can plan in this model using $r^\dagger$. Let's define $\widehat{\pi}^\star$ as the optimal policy of $\widehat{P}^\dagger$ and $r^\dagger$ (imagine we run PI under $\widehat{P}^\dagger$ and $r^\dagger$ until convergence).

We will consider two cases.

In case one, when we execute $\widehat{\pi}^\star$ under the real model P, $\widehat{\pi}^\star$ does not escape the known set $\mathcal{K}$ with big probability, i.e., $\sum_{s,a\in(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}} d^{\widehat{\pi}^\star}(s,a) \leq \epsilon$ for some small $\epsilon \in \mathbb{R}^+$.

Q (a) (15 points): Let us prove the following inequality:

$$V^{\pi^\star} \leq V^{\widehat{\pi}^\star} + \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\epsilon + \frac{1}{(1-\gamma)^2}\sqrt{\frac{S\ln(SA/\delta)}{k}}\right).$$

This says that if k is big enough, and the escape probability is small, then we find a near-optimal policy $\widehat{\pi}^\star$.

(**Hint:** here we need to consider the difference between $\widehat{V}^{\dagger,\pi}$ and $V^{\dagger,\pi}$ using simulation lemma)

Q (b) (15 points): On the other hand, we encounter the case where $\widehat{\pi}^\star$ escapes, i.e.,

$$\sum_{s,a\in(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}} d^{\widehat{\pi}^\star}(s,a) \geq \epsilon.$$

Let us assume that we have sampling oracle $\mathcal{O}(\pi)$ that samples an i.i.d state-action pair from d^π, i.e., $(s,a) \sim d^\pi$

Prove the following conclusion. *With probability at least $1-\delta$, after $M = \Omega\left(\frac{kSA}{\epsilon} + \frac{\ln(1/\delta)}{\epsilon^2}\right)$ many sampling oracle $\mathcal{O}(\widehat{\pi}^\star)$ calls, at least one state-action pair from the unknown set $(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}$ becomes known, i.e., it will be sampled more than k times.*

(**Hint**: First let's ask ourselves that after M many calls of $\mathcal{O}(\widehat{\pi}^\star)$, how many (s,a) samples out of M samples will fall into $(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}$? Now, if I have N many (s,a) samples falling into $(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}$, can I say anything useful here using the Pigeon hole principle here?)

Remark: To this end, we have seen that using $\widehat{P}^\dagger$ and $r^\dagger$, we plan a policy $\widehat{\pi}^\star$ that either is near optimal if it stays inside the known set with big probability, or we guarantee to explore, i.e., making at least one previous unknown state-action pair to be known, by executing $\widehat{\pi}^\star$ in the real MDP for polynomially many times. With the new known set, we can repeat the above whole process again which will either terminate with a near optimal policy or will guarantee to identify another known state-action pair. Note that the total number of state-action pairs that can be made to the known set is at most SA. Thus, we can expect that the algorithm will find a near optimal policy in polynomial sample complexity.

2 High probability bounds for Importance Sampling *[20 points]*

Suppose we want to evaluate the expected value of a function f under a distribution P, that is $\mu(f) := \mathbb{E}_{X\sim p} f(X)$. If we have access to samples from p, an estimate of this quantity can be

obtained using a sample average, that is, $\widehat{\mu}(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$. In many cases, samples from p are either not available, or can be expensive to obtain (e.g. running a particle accelerator for validating theoretical models in physics). Often we have access instead to samples from another distribution q which is easy to sample from (such as a from a high-fidelity simulator). An important technique to use the samples from q to evaluate $\mu(f)$ is called importance sampling, where we estimate:

$$\widehat{\mu}(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i). \tag{1}$$

We are assuming for now that p and q are the probability mass functions of a distribution, with X taking at most M distinct values. That is, we compute a similar sample average as when we have access to samples from p, but additionally reweight each sample by the ratio of probabilities under p and q. In this question, we will study some basic properties of this importance sampling technique.

(a) ***[5 points]*** Show that $\widehat{\mu}(f)$ is not an unbiased estimator of $\mu(f)$ in general with an example. Based on the example, suggest a regularity condition relating p and q under which the estimator is unbiased.

(Hint: do expectations always exists?)

(b) ***[5 points]*** What is the variance of $\widehat{\mu}(f)$? Which distribution q minimizes this variance among the ones that gives rise to unbiased estimators?

(Hint: $f(x_i)$ can be 0 too. You may need to use the method of Lagrange multiplier.)

(c) ***[5 points]*** Suppose that $f(X) \in [0, 1]$ for all x. Using Hoeffding's inequality, show that with probability at least $1 - \delta$

$$|\widehat{\mu}(f) - \mu(f)| \leq \max_x \frac{p(x)}{q(x)} \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$

(d) ***[5 points]*** Combine Part (b) and (c) to obtain an improved bound using the Bernstein inequality?

(e) (Bonus ***[5 points]***) The bounds above depend on $\max_x \frac{p(x)}{q(x)}$, which could be very large. Now consider two trivial regression estimators:$\hat{f}_1(x) \equiv 0$ or $\hat{f}_2(x) \equiv 0$. Use SWITCH by thresholding the importance weights at τ. Show that the expectation of the estimator is an upper / lower bounds of $\widehat{\mu}(f)$ respectively when we use $\hat{f}_1$ and $\hat{f}_2$. Construct a Hoeffding's style lower bound fo the lower bound estimate and upper bound of the upper bound estimate.

3 Importance Sampling and Off-Policy Policy Gradients *[20 points]*

Let's instantiate the general importance sampling above with RL-specific quantities and consider the OPE estimator under the contextual bandits model. Let the target policy be π_θ which is parameterized by $\theta \in \mathbb{R}^d$. We will consider two settings regarding the logging policy.

Setting 1 (Unknown logging policy) The logging policy μ is unknown, but we know it is a randomized policy and the offline dataset (X_i, A_i, R_i, μ_i) for $i = 1, ..., n$ where μ_i is the probability of taking the action A_i.

Setting 2 (known logging policy) We know θ_0 and that the logging policy is π_{θ_0}, i.e., $\mu_i := \pi_{\theta_0}(A_i|X_i)$. So we could just work with the dataset (X_i, A_i, R_i) for $i = 1, ..., n$.

The importance sampling estimator is the following

$$\hat{v}^{\pi_\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_\theta(A_i|X_i)}{\mu_i} R_i$$

Notice that $X_i \in \mathbb{R}^d$ are drawn iid from some unknown distribution, and $R_i = r(X_i, A_i)$ deterministically (treating random $R_i|X_i, A_i$ is straightforward, so let's simplify it a bit by considering deterministic reward). We assume that $0 \leq r(X_i, A_i) \leq 1$. Furthermore, assume π_θ is differentiable in θ.

(a) ***[5 points]*** Derive the gradient of the $\hat{v}^{\pi_\theta}$ with respect to θ. Does it make a difference whether we are in Setting 1 or Setting 2?

(b) ***[5 points]*** Review the lecture where we learned the Policy Gradient Theorem, which implies the following in the contextual bandits setting:

$$\nabla_\theta v^{\pi_\theta} = \mathbb{E}_{\pi_\theta}\left[\nabla \log(\pi_\theta) r(X, A)\right].$$

Derive an importance sampling estimator to the policy gradient theorem to construct an unbiased estimator to $\nabla_\theta v^\pi$ directly. What is the connection to Part (a)?

(c) ***[5 points]*** Consider the following "conservative" policy update algorithm where we would like to find a good policy without moving too far away from the current policy in terms of KL-divergence

$$\max_\theta \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_\theta(A_i|X_i)}{\pi_{\theta_0}(A_i|X_i)} R_i - \lambda \frac{1}{n}\sum_{i=1}^{n}\sum_{i} \mathrm{KL}(\pi_\theta(\cdot|X_i)\|\pi_{\theta_0}(\cdot|X_i)).$$

Look up the definition of KL-divergence, derive the gradient of this objective function under Setting 2. Can you compute this gradient in Setting 1 too?

Remark: The KL-divergence regularization is slightly different from the standard entropy regularization because the base-measure is not under π_θ. However, this regularization had been shown to be critical in obtaining efficient learning bounds for policy-gradient and natural policy gradients methods in recent reinforcement learning literature.

(d) ***[5 points]*** What if we did not even log the probabilities $\mu_1, ..., \mu_n$? This situation is quite typical in practice. Show that the following modified objective function is always a lower bound of the importance sampling objective $\hat{v}^{\pi_\theta}$

$$\frac{1}{n}\sum_{i=1}^{n} \left(1 + \log(\frac{\pi_\theta(A_i|X_i)}{\mu_i})\right) R_i$$

Clearly you still need μ_i to evaluate the above expression, but if all you want is to maximize this lower bound, do you still need the knowledge of the logging policy from either Setting 1 or Setting 2? (Hint: Use $\log(1 + x) \leq x$ for all $x \geq -1$.)

Remark: Maximizing the expression in Part (d) can be interpreted as a model-based approach that aims at maximizing a reward-weighted log-likelihood. For example, if R_i is drawn from a Bernoulli distribution with its parameter p_i jointly determined by X_i, A_i, then the log-likelihood $\log P_\theta(R_i|X_i, A_i) = R_i \log p_\theta(X_i, A_i) + (1 - R_i)\log(1 - p_\theta(X_i, A_i))$.

4 Uniform convergence in offline learning [30 Points + 20 Bonus Points]

Finally we will prove a uniform convergence bound for an infinite-sized policy class, so that we can say something about the statistical guarantees of offline policy learning.

Consider the same OPE estimator as in Q3

$$\hat{v}^{\pi_\theta} := \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_\theta(A_i|X_i)}{\mu_i}.$$

We make three assumptions:

- π_θ is L-Lipschitz in θ, i.e., $\left|\pi_\theta(a|s) - \pi_{\theta'}(a|s)\right| \leq L\|\theta - \theta'\|_2 \;\; \forall \theta, \theta' \in \Theta, \forall a \in \mathcal{A}, s \in \mathcal{S}$.
- The logging policy obeys that $\mu(a|s) \geq d_m$.
- The reward is deterministic and bounded $0 \leq R_i = r(X_i, A_i) \leq 1$.

(a) ***[10 points]*** Let the policy class be indexed by $\theta \in \Theta$ where $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq B\}$. Use the covering number of the Euclidean ball (Lemma 7.5 AJKS) to derive the the ε-covering number for $\Pi := \{\pi_\theta \mid \theta \in \Theta\}$, i.e., the smallest number of elements of $\tilde{\Pi} \subset \pi$ such that for any $\pi \in \Pi$, there exists $\tilde{\pi} \in \tilde{\Pi}$ obeying $|v^\pi - v^{\tilde{\pi}}| \leq \varepsilon$.

(b) ***[10 points]*** By combining your results in Q2(c) and a covering number argument, obtain a uniform convergence bound that holds with probability at least $1 - \delta$, i.e.,

$$\sup_{\theta \in \Theta} |\hat{v}^{\pi_\theta} - v^{\pi_\theta}| \leq \epsilon$$

with ϵ parameterized by n, d_m, d, L, B, δ.

(Hint: you could choose the ε in the covering number appropriately after obtaining the overall bound that depends on ε.)

(c) ***[10 points]*** Apply the above uniform convergence bound to upper bound the sample complexity of offline learning, i.e., how large a sample size n do we need such that we are guaranteed to find an ϵ-suboptimal policy with probability at least $1 - \delta$.

(d) ***[Bonus 10 points]***. Obtain a stronger uniform convergence bound via the same covering number argument, but now with Bernstein inequality (Your results in Q2(d)!). Show that with n samples,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{v}^{\pi_\theta}$$

obeys that

$$v^{\pi_{\theta^*}} - v^{\pi_{\hat{\theta}}} \leq \tilde{O}\left(\sqrt{\frac{\mathbb{E}_\mu\left(\frac{\pi_{\theta^*}(A_i|X_i)^2}{\mu(A_i|X_i)^2}\right)}{n}} + \sqrt{\frac{\mathbb{E}_\mu\left(\frac{\pi_{\hat{\theta}}(A_i|X_i)^2}{\mu(A_i|X_i)^2}\right)}{n}} + \frac{1}{d_m n}\right.$$

where $\tilde{O}$ hides logarithmic factors and constants (e.g., B, L, d).

You could think about θ^* as the parameter for the optimal policy within the class.

(e) ***[Bonus 10 points]***. Consider the alternative policy optimization rule either using "pessimism" update rule

$$\textbf{Optimism: } \overline{\theta} = \arg\max_{\theta\in\Theta} \hat{v}^{\pi_\theta} + b_\theta$$

or

$$\textbf{Pessimism: } \underline{\theta} = \arg\max_{\theta\in\Theta} \hat{v}^{\pi_\theta} + b_\theta$$

where b_θ is the uniform confidence bound specific to policy π_θ.

Derive the high-probability bounds of $v^{\pi_{\theta^*}} - v^{\pi_{\overline{\theta}}}$ and $v^{\pi_{\theta^*}} - v^{\pi_{\underline{\theta}}}$.

Inspect the results and discuss which rule is preferred under what circumstance?

Remark: We have simplified the discussion by considering only the contextual bandits case. We note that θ^* might not need to be the optimal policy. The optimal policy might not be "measurable" given the logging policy. Moreover, we could use SWITCH-like ideas to construct stronger upper / lower bounds in the finite sample setting.

5 Duality of FQI and TMIS via simulation lemma (30 points)

In this question, assume the reward function $r_t(s,a)$ is known for all $t = 1, ..., H$ and the initial state distribution d_1 is known. Consider the tabular setting and that all state and actions are visited for at least once.

The FQI-based OPE estimator is

$$\hat{v}^\pi_{FQI} = \sum_{i=1}^{n} \sum_{s,a} d_1(s)\pi_1(a|s)\hat{Q}^\pi_1(s,a)$$

where $\hat{Q}^\pi_t$ is recursively estimated via the FQI updates backwards from $t = H, H-1, ..., 1$.

$$\hat{Q}^\pi_t = \arg\min_{q\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \sum_{i=1}^{n} (r_t(S^{(i)}_t, A^{(i)}_t) + \sum_{a\in\mathcal{A}} \pi_{t+1}(a'|S^{(i)}_{t+1})\hat{Q}^\pi_{t+1}(S^{(i)}_{t+1}, a') - q(S^{(i)}_t, A^{(i)}_t))^2.$$

We define $\hat{Q}_{H+1} \equiv 0$ trivially.

The TMIS-estimator is

$$\hat{v}^\pi_{TMIS} = \sum_{t=1}^{H} \sum_{s,a} \hat{d}^\pi_t(s)\pi_t(a|s)r(s,a)$$

where $\hat{d}^\pi_t$ is recursively estimated from $t = 1, 2, 3, ..., H$ by

$$\hat{d}^\pi_{t+1} = \sum_{s,a} \hat{P}_t(\cdot|s,a)\pi_t(a|s)\hat{d}^\pi_t(s)$$

where $\hat{P}_t(\cdot|s,a)$ is the empirically estimated probabilities.

In this question, we will show that FQI and TMIS are equivalent.

(a) ***[5 points]*** Show that the update rules of FQI is equivalent to running Policy Evaluation on the empirically estimated MDP.

(Hint: We have done this in the class. It involves taking the gradient and set it to 0.)

(b) ***[5 points]*** Extend the simulation lemma that you derived from HW1 Q5(d) to the non-stationary episodic MDP case.

(c) ***[10 points]*** Expand $\hat{v}^{\pi}_{FQI} - v^{\pi}$ by substituting the simulation lemma into $\hat{Q}^{\pi}_1(s,a) - Q^{\pi}_1(s,a)$.

(d) ***[10 points]*** Apply the simulation lemma again, but this time swap the true MDP M and the estimated MDP $\hat{M}$. Construct $\hat{d}^{\pi}_t$ from the expression and then write down an expansion of $\hat{v}^{\pi}_{TMIS} - v^{\pi}$.