CS 542 Stats RL
Homework 4

November 3, 2025

1. (4 pts) Let $\mathcal{X}$ be a finite and discrete space, and $p, q \in \Delta(\mathcal{X})$ are two distributions over $\mathcal{X}$. $f : \mathcal{X} \to [0, 1]$ is a function. Let $X_1, \ldots, X_n$ be sampled i.i.d. from q. Recall that the importance sampling estimator for $\mathbb{E}_p[f]$ is

$$v = \frac{1}{n} \sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} f(X_i).$$

Assume that $\|p/q\|_\infty := \max_x p(x)/q(x) \leq C < \infty$. This means $\frac{p(X_i)}{q(X_i)} f(X_i)$ are i.i.d. random variables with range $[0, C]$. If we use Hoeffding's inequality, we'd conclude that to guarantee $|v - \mathbb{E}_p[f]| \leq \epsilon$ with high probability (i.e., w.p. $\geq 1 - \delta$), we will need $n = O(C^2 \ln(1/\delta)/\epsilon^2)$ samples.

Prove an improved result that we should only need $n = O(C \ln(1/\delta)/\epsilon^2)$. Hint: check out Bernstein's inequality given by Lemma 7.37 of `https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf`.[1] Show that $\text{Var}[\frac{p(X_i)}{q(X_i)} f(X_i)] = O(C)$. [2]

[1] In Eq.(7.38), σ^2 is the variance of X_i; this is stated in Lemma 7.26.

[2] In general, for a random variable with bounded range $[0, C]$, the worst-case variance is $O(C^2)$.

2. Low-rank/linear MDPs (6 pts)

Low-rank/linear MDPs have been a popular setting in recent theoretical RL works. In this problem you will be asked to establish some essential properties of linear MDPs. First, a low-rank MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ is one such that for any (s, a, s'), we have $P(s'|s, a) = \phi(s, a)^\top \psi(s')$, where ϕ and ψ are two maps from (s, a) and s' respectively to d-dimensional real vectors. In other words, the transition matrix P has low rank and can be factorized into the product of two matrices, $\Phi \times \Psi$, where Φ has $\phi(s, a)^\top$ as its rows and Ψ has $\psi(s')$ as its columns.

Two further common assumptions for this model:

- $R(s, a) = \phi(s, a)^\top \theta_R, \ \forall (s, a)$, that is, reward is linear in ϕ.
- $d_0(s) = \psi(s)^\top \eta_0, \ \forall s$, that is, the initial distribution is linear in ψ.

The above model is known as a low-rank MDP. A linear MDP refers to the situation where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is known to the learner (ψ is unknown).

A useful special case of the model is where $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times d}$ and $\Psi \in \mathbb{R}^{d\times|\mathcal{S}|}$ are both row-stochastic, i.e., each row of Φ represents a distribution over d discrete possible outcomes, and each row of Ψ (denoted as ψ_i) represents a distribution over $\mathcal{S}$. [3] We also assume that d_0 is a probability mixture of $\{\psi_i\}$, i.e., $\eta_0 \in \Delta([d])$. This model is sometimes known as low-rank/linear MDPs with *simplex features*.

Let $\mathcal{F} := \{(s, a) \mapsto \phi(s, a)^\top \theta : \theta \in \mathbb{R}^d)\}$, i.e., the linear function space w.r.t. feature map ϕ. Prove the following:

1. For any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and any π, $\mathcal{T}^\pi f, \mathcal{T} f \in \mathcal{F}$. (Remark: this directly implies closure of $\mathcal{F}$ under $\mathcal{T}$ and $\mathcal{T}^\pi$, a.k.a. completeness, and is quite a bit stronger.)
2. For any policy π, let d_t^π be the t-th step state distribution induced by π from d_0. Show that d_t^π is linear in ψ, i.e., there exists $\eta \in \mathbb{R}^d$, such that $d_t^\pi(s) = \psi(s)^\top \eta, \forall s$.

   In the simplex feature setting, show further that $\eta \in \Delta([d])$, i.e., d_t^π is a probability mixture of $\{\psi_i\}_{i=1}^d$.
3. Recall that the concentrability condition states that a data distribution μ (often used in offline learning) satisfies

$$\forall s, a, \pi, t, \frac{d_t^\pi(s, a)}{\mu(s, a)} \leq C. \tag{1}$$

   Now consider a low-rank MDP M with simplex features. Construct a distribution μ, such that concentrability is satisfied with $C = d \times |\mathcal{A}|$. (Hint: the only property of d_t^π that matters is what you proved in the previous problem, i.e., it is a probability mixture of $\{\psi_i\}_{i=1}^d$.)

[3] Under such an assumption, the transition dynamics can be interpreted as the following: $s' \sim P(\cdot|s, a) \ \Leftrightarrow \ z \sim \phi(s, a), s' \sim \psi_z(\cdot)$, i.e., a latent variable $z \in [d]$ ($[d]$ is a shorthand for $\{1, 2, \ldots, d\}$) is sampled from $\phi(s, a) \in \Delta([d])$, and then the next state s' is drawn from the "emission distribution" ψ_z, which is the z-th row of Ψ.

(Optional; 3 pts) Prove Q2(3) without the simplex feature assumption (i.e., general low-rank MDPs). For simplicity you can assume that for Eq.(1) the $\forall \pi, t$ only considers policies from a finite class Π and all $t \leq T_0$ for some finite T_0. Hint: look up barycentric spanner.

3. Pessimism in face of uncertainty (5 pts)

Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ be the true MDP, and we want to compute a good policy. As usual we do not have direct access to M, and are instead given the following items:

- An approximate model $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma, d_0)$, which is also a valid MDP. Rewards are bounded in $[0, R_{\max}]$ in both M and $\widehat{M}$.
- A set $K \subseteq \mathcal{S} \times \mathcal{A}$ and two numbers ϵ_R, ϵ_P. It is guaranteed that $\forall (s,a) \in K$,

$$|R(s,a) - \widehat{R}(s,a)| \leq \epsilon_R, \quad \|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq \epsilon_P.$$

  However, there is no guarantee on the accuracy of $\widehat{R}$ and $\widehat{P}$ on $(s,a) \notin K$.

Design an algorithm that computes a good policy $\hat{\pi}$ and provide the following kind of guarantee about $J_M(\hat{\pi}) := \mathbb{E}_{s \sim d_0}[V_M^{\hat{\pi}}(s)]$: Show that for any policy π such that $d^\pi(s,a) = 0 \; \forall (s,a) \notin K$, $J_M(\pi) - J_M(\hat{\pi})$ can be upper-bounded by some function of ϵ_R and ϵ_P, which goes to 0 when $\epsilon_R = \epsilon_P = 0$.

Make your guarantee more general by providing an upper bound on $J_M(\pi) - J_M(\hat{\pi})$ for an *arbitrary* policy π, where the upper bound can depend on ϵ_R, ϵ_P, and a term that measures the violation of the aforementioned condition that $d^\pi(s,a) = 0 \; \forall (s,a) \notin K$.

Hints:

1. The situation could arise when the approximate model is estimated from *incomplete* data, where you only have enough samples for $(s,a) \in K$ but not elsewhere, and you are essentially asked to make the best effort with this incomplete dataset,[4] i.e., you are asked to *exploit* existing information.
2. The idea is to use pessimism, which we briefly talked about at the end of the FQI section. Here you are asked to perform a similar analysis for the tabular case yourself.

[4] To make your life easier, the accuracy of $\widehat{R}$ and $\widehat{P}$ on K is given directly, and you do not need to turn sample size into these accuracy parameters. The example of incomplete data, therefore, is only to provide you with some intuitions.