CSC 696H Homework 2

Chicheng Zhang

October 2021

- This homework is due on Nov 2 in class.
- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel unable to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- Feel free to use existing theorems from the course notes / the textbook.

Problem 1

Finish the proof of the last claim in the UCB-VI analysis; specifically, recall that its bonus function used at episode k for step h is defined as $b_h^k(s,a) := H\sqrt{\frac{l}{N_h^k(s,a)+1}}$, and $\xi_h^k(s,a) := \frac{2H^2Sl}{N_h^k(s,a)+1}$, for $l = 8\ln(HSAK)$. Using a similar reasoning as in the UCB analysis in multi-armed bandits, prove that:

1. $\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} b_h^k(s_h^k,a_h^k) = O\left(\sqrt{H^4SAKl}\right)$.
2. $\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \xi_h^k(s_h^k,a_h^k) = O\left(H^3S^2Al^2\right)$.

Problem 2

In class, we introduced the performance difference lemma:

Lemma 1 (Performance difference lemma)**.** *Given an undiscounted episodic MDP $M = (\mathbb{P}_h, r_h)_{h=0}^{H-1}$, and two deterministic history-independent policies $\pi = (\pi_h)_{h=0}^{H-1}$ and $\pi' = (\pi'_h)_{h=0}^{H-1}$, we have:*

$$\forall h, s : V_h^{\pi}(s) - V_h^{\pi'}(s) = \mathbb{E}\left[\sum_{i=h}^{H-1} -A_i^{\pi}(s_i, \pi'(s_i)) \mid M, \pi', s_h = s\right],$$

where $A_i^{\pi}(s,a) = Q_i^{\pi}(s,a) - V_i^{\pi}(s)$ is the advantage function of policy π at step i.

And we have given the following starting point of the proof: fix h, s; let $a := \pi_h(s)$, $a' := \pi'_h(s)$;

$$\begin{aligned} V_h^{\pi}(s) - V_h^{\pi'}(s) &= Q_h^{\pi}(s,a) - Q_h^{\pi'}(s,a') \\ &\overset{(*)}{=} \left(Q_h^{\pi}(s,a) - Q_h^{\pi}(s,a')\right) + \left(Q_h^{\pi}(s,a') - Q_h^{\pi'}(s,a')\right) \\ &= -A_h^{\pi}(s,\pi'_h(s)) + \left\langle \mathbb{P}_h(\cdot \mid s,a'), V_{h+1}^{\pi} - V_{h+1}^{\pi'} \right\rangle \\ &= \mathbb{E}\left[-A_h^{\pi}(s_h,\pi'_h(s_h)) + V_{h+1}^{\pi}(s_{h+1}) - V_{h+1}^{\pi'}(s_{h+1}) \mid s_h = s, M, \pi'\right]. \end{aligned}$$

1. Finish the proof of the performance difference lemma, using a similar reasoning as in the proof of the simulation lemma.
2. Given a history-independent policy $\pi = (\pi_h)_{h=0}^{H-1}$, and consider one of its greedy policies $\pi' = (\pi'_h)_{h=0}^{H-1}$, where for all h, s, we have $\pi'_h(s) \in \operatorname{argmax}_{a\in\mathcal{A}} Q_h^{\pi}(s,a)$. Is π' always better than π, in other words, is $V_h^{\pi'} \succeq V_h^{\pi}$ for all h? Why?
3. In line (*), we add and subtract $Q_h^{\pi}(s,a)$ and regroup the terms; another reasonable way to proceed is to add and subtract $Q_h^{\pi'}(s,a)$. If you do this instead, can you show a result similar to Lemma 1? Justify your answer.

Problem 3

Given a linear MDP $M = (r_h, \mathbb{P}_h)_{h=0}^{H-1}$ and its associated feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, such that there exist $(\alpha_h^\star)_{h=0}^{H-1} \subset \mathbb{R}^d$ and $(\mu_h^\star)_{h=0}^{H-1} \subset \mathbb{R}^{S\times d}$, and $\forall h$, $\forall s, a, s'$:

$$r_h(s,a) = \langle \alpha_h^\star, \phi(s,a)\rangle,$$

$$\mathbb{P}_h(s' \mid s,a) = \left\langle (\mu_h^\star)^{s'}, \phi(s,a)\right\rangle,$$

where $(\mu_h^\star)^{s'} \in \mathbb{R}^d$ denotes the s'-th row of $\mu_h^\star$ (recall the pictures in the class). Define the linear action value function class induced by feature map ϕ as $\mathcal{F} := \left\{f_\theta : \theta \in \mathbb{R}^d\right\}$, where $f_\theta(s,a) := \langle \theta, \phi(s,a)\rangle$. Verify that the linear Bellman completeness condition is satisfied: for any step h and $f \in \mathcal{F}$, we always have $\mathcal{T}_h^\star f \in \mathcal{F}$.

Problem 4

Recall that for real symmetric matrix $A \in \mathbb{R}^{d\times d}$, we use $A \succeq 0$ to denote that A is positive semidefinite, and use $A \succ 0$ to denote that A is positive definite.

In class, we have introduced the positive-semidefinite cone partial order: given two real symmetric matrices A and B in $\mathbb{R}^{d\times d}$, $A \succeq B$ iff $A - B \succeq 0$. Also, recall the Mahalanobis norm: given $v \in \mathbb{R}^d$ and $A \succeq 0$, denote by $\|v\|_A = \sqrt{v^\top A v}$.

Below, let A, B denote positive semidefinite matrices in $\mathbb{R}^{d\times d}$, and u, v denote vectors in $\mathbb{R}^d$. Prove the following basic linear algebra facts used in our analysis of RL with linear function approximation:

1. If $A \succeq B$, then $\|u\|_A \geq \|u\|_B$.
2. If $A \succ 0$, then $|\langle u, v\rangle| \leq \|u\|_A \cdot \|v\|_{A^{-1}}$.
3. If $\|u\|_2 \leq 1$, then $I \succeq uu^\top$.
4. If $A \succeq I$, then $I \succeq A^{-1}$. (Hint: you may want to consider A's eigendecomposition.)
5. If $A \succeq B \succ 0$, then $B^{-1} \succeq A^{-1}$. (Hint: first show $B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \succeq I$, and use the previous item.)

Problem 5

Given an undiscounted episodic MDP $M = (\mathbb{P}_h, r_h)_{h=0}^{H-1}$, suppose we are given a set of value functions and action value functions $(\hat{V}_h)_{h=0}^{H}$ and $(\hat{Q}_h)_{h=0}^{H}$ that satisfy a variant of *strong optimism* [SJ19]: $\hat{V}_H \equiv 0$, $\hat{Q}_h \equiv 0$, and

$$\forall h \in \{0, \dots, H-1\} : \begin{cases} \hat{V}_h(s) = \min\left(H, \max_{a \in \mathcal{A}} \hat{Q}_h(s,a)\right), & \forall s \\ \hat{Q}_h(s,a) \geq r_h(s,a) + \left\langle \mathbb{P}_h(\cdot \mid s,a), \hat{V}_{h+1} \right\rangle, & \forall s, a \end{cases}$$

Show that $(\hat{V}_h)_{h=0}^{H}$ and $(\hat{Q}_h)_{h=0}^{H}$ also satisfy optimism, i.e., for all $h = 0, 1, \dots, H$, $\hat{V}_h \succeq V_h^\star$ and $\hat{Q}_h \succeq Q_h^\star$.

Problem 6

Recall that the Kiefer-Wolfowitz experiment design theorem says: for any full-dimensional $\mathcal{X} \subset \mathbb{R}^d$ (i.e., $\mathrm{span}(\mathcal{X}) = \mathbb{R}^d$), there exists a distribution $\rho^\star$ over $\mathcal{X}$, whose induced "covariance matrix" $\Sigma_{\rho^\star} = \mathbb{E}_{x \sim \rho^\star}\left[xx^\top\right]$ satisfies that $\max_{x \in \mathcal{X}} \|x\|^2_{\Sigma_{\rho^\star}^{-1}} \leq d$; in other words, $\rho^\star$ ensures that all $x \in \mathcal{X}$ look relatively "typical" with respect to it.

1. Suppose $d = 1$ and $\mathcal{X} \subset \mathbb{R}^1$. How would you define $\rho^\star$?
2. Show that Kiefer-Wolfowitz Theorem is tight: for any $d \in \mathbb{N}_+$, find a full-dimensional set $\mathcal{X} \subset \mathbb{R}^d$, such that for any distribution ρ over $\mathcal{X}$, $\max_{x \in \mathcal{X}} \|x\|^2_{\Sigma_\rho^{-1}} = d$.

References

[SJ19] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32:1153–1162, 2019.