Homework 2: Performance Difference Lemma, Simulation Lemma, and Absorbing MDPs

CS 6789: Foundations of Reinforcement Learning

Due Oct 30 6pm

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 Performance Difference Lemma (20 points)

Consider finite horizon MDPs $\{\mathcal{S}, \mathcal{A}, H, \{r_h\}_h, \{P_h\}_h, s_0\}$. Denote $\mathbb{P}_h^\pi(s,a)$ as the state-action distribution induced by $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ at time step h, i.e., the probability of π visiting (s,a) at time step h starting from s_0 at $h = 0$.

Let us denote V^π as the expected total reward of policy π starting at s_0, i.e.,

$$V^\pi = \mathbb{E}\left[\sum_{h=0}^{H-1} r(s_h, a_h) | a_h \sim \pi(s_h)\right].$$

1.1 Classic PDL proof (10 points)

Prove the following equality for two arbitrary stochastic stationary policies π and $\widetilde{\pi}$:

$$V^\pi - V^{\widetilde{\pi}} = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\mathbb{P}_h^\pi}\left[A_h^{\widetilde{\pi}}(s,a)\right],$$

where A^π is the advantage function of a policy π, i.e., $A_h^\pi(s,a) = Q_h^\pi(s,a) - V_h^\pi(s)$

1.2 Extension of PDL (10 points)

Now let us consider a function $f : \mathcal{S} \times \mathcal{A} \to [0, H]$. Note that f here is not necessarily corresponding to any value Q function Q^π (think about f as some function approximation of some $Q^\star$ for instance). Denote π_f as the stationary policy $\pi_f(s) = \text{argmax}_a f(s,a)$ for all s (you can think about π_f as a stochastic policy which puts all probability mass at argmax_a).

Recall that we consider finite horizon H. Let's assume that at $h = H-1$, for any s, a, the transition $P_{H-1}(\cdot|s,a)$ always transits to a special state $\hat{s}$ (i.e., $P_{H-1}(\hat{s}|s,a) = 1$) at time step H, and $f(\hat{s}, a') = 0$ for all $a' \in \mathcal{A}$.

Prove the following equality as well:

$$V^{\pi_f} - \max_a f(s_0, a) = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi_f}} \left[r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} \max_{a'} f(s', a') - f(s,a) \right].$$

2 Simulation Lemma and Model-based Learning (20 points)

Consider two finite horizon MDPs $\mathcal{M} := \{\mathcal{S}, \mathcal{A}, H, \{r_h\}_h, \{P_h\}_h, s_0\}$ and $\widetilde{\mathcal{M}} := \{\mathcal{S}, \mathcal{A}, H, r_h, \{\widetilde{P}_h\}_h, s_0\}$. Consider an arbitrary stochastic stationary policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$. Similarly to the above problem, we define V^π as the expected total reward of π under $\mathcal{M}$, and $\widetilde{V}^\pi$ as the expected total reward of π under $\widetilde{\mathcal{M}}$. We denote $\mathbb{P}_h^\pi$ as the state-action distribution of π under $\mathcal{M}$ at step h.

2.1 Proving Simulation Lemma (10 points)

Prove the following equality

$$V^\pi - \widetilde{V}^\pi = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\mathbb{P}_h^\pi} \left[\mathbb{E}_{s'\sim P_h(\cdot|s,a)} \widetilde{V}_{h+1}^\pi(s') - \mathbb{E}_{s'\sim \widetilde{P}_h(\cdot|s,a)} \widetilde{V}_{h+1}^\pi(s') \right]$$

Note that $\widetilde{V}_{h+1}^\pi(s')$ is the expected total reward of π under $\widetilde{\mathcal{M}}$ from time $h+1$.

2.2 Model-based Learning (10 points)

Imagine the following situation: we have $\mathcal{M}$ as the true real MDP, and we have $\widetilde{\mathcal{M}}$ as some learned model that supposes to approximate the real MDP $\mathcal{M}$. Given $\widetilde{\mathcal{M}}$, the natural thing to do is to compute the optimal policy under $\widetilde{\mathcal{M}}$, i.e.,

$$\widetilde{\pi}^\star = \text{argmax}_{\pi\in\Pi} \widetilde{V}^\pi,$$

where $\Pi \subset \{\pi : \mathcal{S} \to \Delta(\mathcal{A}\}$ is a pre-defined policy class. Let us also denote the true optimal policy $\pi^\star$ under the real model as:

$$\pi^\star = \text{argmax}_{\pi\in\Pi} V^\pi.$$

A natural question is that what is the performance of $\widetilde{\pi}^\star$ under the real model $\mathcal{M}$, compared to $\pi^\star$ under $\mathcal{M}$?

Let's prove the following inequality:

$$V^{\pi^\star} - V^{\widetilde{\pi}^\star} \leq H \sum_{h=0}^{H-1} \left[\mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi^\star}} \|\widetilde{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 + \mathbb{E}_{s,a\sim\mathbb{P}_h^{\widetilde{\pi}^\star}} \|\widetilde{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \right]$$

The above inequality shows the challenging of model-based approach. As we can see, in order for $\widetilde{\pi}^\star$ to be near optimal, we need to make sure $\widetilde{P}$ is accurate under both $\mathbb{P}^{\pi^\star}$ and $\mathbb{P}^{\widetilde{\pi}^\star}$, and unfortunately we have no way to know what $\mathbb{P}^{\pi^\star}$ looks like.

3 Absorbing MDPs (60 points)

Consider an infinite horizon discounted MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, r, P, s_0\}$. Let us assume $\mathcal{S}$ and $\mathcal{A}$ are discrete here, and $r(s,a) \in [0,1]$. We will consider stationary policy π here, and recall $d^\pi(s,a)$ is the state-action visitation of π under transition P, i.e.,

$$d^\pi(s,a) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s,a),$$

where $\mathbb{P}_h^\pi(s,a)$ is the probability of π hitting (s,a) at time step h under model P starting at s_0 at $h=0$.

Let us also assume that we are given a dataset $\{s_i, a_i, s_i'\}_{i=1}^N$ that consists of N state-action-next state triples, where $s_i' \sim P(\cdot|s_i,a_i)$. Following what we did in the tabular UCBVI lecture, we will define two statistics: $N(s,a) = \sum_{i=1}^N \mathbf{1}\{s_i, a_i = s, a\}$ and $N(s,a,s') = \sum_{i=1}^N \mathbf{1}\{s_i, a_i, s_i' = s, a, s'\}$, and we define $\widehat{P}$ as:

$$\widehat{P}(s'|s,a) = \frac{N(s,a,s')}{N(s,a)}, \forall s, a, s'.$$

For simplicity, let us assume $N(s,a) \geq 1$ for all s, a. We define a *known* set $\mathcal{K} \subset \mathcal{S} \times \mathcal{A}$ as follows:

$$\mathcal{K} = \{s, a \in \mathcal{S} \times \mathcal{A} : N(s,a) \geq k\},$$

where $k \in \mathbb{N}^+$ is some pre-defined constant integer and let us think about k as some reasonably big number. So from the lecture, we know that for any $s, a \in \mathcal{K}$, we have a reasonably good estimate of $\widehat{P}(\cdot|s,a)$, i.e., we can easy formulate the following high-probability event: *with probability at least* $1-\delta$*, for all* $s, a \in \mathcal{K}$*, we have that:*

$$\left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1 \leq \sqrt{\frac{S\ln(SA/\delta)}{k}}, \forall s, a \tag{1}$$

Again let's think about k as a big number here, so that $\widehat{P}(\cdot|s,a)$ is pair-wise accurate for any $s, a \in \mathcal{K}$ (refer to Ch. 2 of the monograph for a detailed treatment). For now, let us just condition on the above inequality being true.

Of course, outside $\mathcal{K}$, we cannot really guarantee a good accurate model. The question we want to study here is that *given the current data, how can we encourage the agent to visit unknown state-action pairs*?

Let us build the following absorbing MDP structure. Let us define an absorbing state $s^\dagger$ (note that $s^\dagger$ is some additional state we add to $\mathcal{S}$). Let us define the following two absorbing MDPs.

$$P^\dagger(\cdot|s,a) = \begin{cases} P(\cdot|s,a) & s,a \in \mathcal{K}, \\ \delta_{s^\dagger} & s = s^\dagger \text{ or } s,a \notin \mathcal{K}, \end{cases}, \quad \widehat{P}^\dagger(\cdot|s,a) = \begin{cases} \widehat{P}(\cdot|s,a) & s,a \in \mathcal{K}, \\ \delta_{s^\dagger} & s = s^\dagger \text{ or } s,a \notin \mathcal{K}, \end{cases}$$

where $\delta_{s^\dagger}$ is a delta distribution that has probability 1 at $s^\dagger$. Namely, in both models, whenever we hit a state-action pair not from $\mathcal{K}$, we transit to $s^\dagger$ immediately, and then we will just self-loop at $s^\dagger$ forever.

Let us define a reward function for both MDPs,

$$r^\dagger(s,a) = \begin{cases} r(s,a) & s,a \in \mathcal{K}, \\ 1 & s = s^\dagger \text{ or } s,a \notin \mathcal{K}. \end{cases}$$

At this stage, we have setup all things we want. Let's see how we can use the absorbing structures we constructed above.

3.1 Probabilities and State action Visitations (30 points)

Given a policy π, let us denote $d^\pi_{P^\dagger}(s,a)$ as the state-action visitation under $P^\dagger$ and $d^\pi(s,a)$ as the state-action visitation under the true model P. We denote $d^\pi_{P^\dagger}(s) = \sum_{a\in\mathcal{A}} d^\pi_{P^\dagger}(s,a)$.

1. (10 points) Prove that:

$$d^\pi(s,a) \geq d^\pi_{P^\dagger}(s,a), \forall s,a \in \mathcal{K}.$$

2. (10 points) Recall that for $P^\dagger$, we have an extra absorbing state $s^\dagger$. Prove that:

$$d^\pi_{P^\dagger}(s^\dagger) = \frac{\gamma}{1-\gamma} \sum_{s,a\in(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}} d^\pi_{P^\dagger}(s,a),$$

   where $d^\pi_{P^\dagger}(s^\dagger) = \sum_a d^\pi_{P^\dagger}(s^\dagger,a)$
3. (10 points) Let us consider $s,a \in (\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}$. Prove that:

$$d^\pi_{P^\dagger}(s,a) \leq d^\pi(s,a).$$

Hint: Remember that $\mathcal{M}$ does not contain the state $s^\dagger$. That is $\forall s,a \in \mathcal{S}\times\mathcal{A}$, $P(s^\dagger|s,a) = 0$. It may be helpful to think about what kind of (s,a) an agent needs to be in for $P^\dagger(s^\dagger|s,a) \neq 0$.

3.2 Optimism from the Absorbing MDPs (15 points)

Let us consider an arbitary stationary stochastic policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$. Denote $V^{\dagger,\pi}$ as the policy π's expected discounted total reward under $P^\dagger$ and $r^\dagger$. Prove the following inequality: for any π.

$$V^{\dagger,\pi} \geq V^\pi.$$

Namely we have found an upper bound of V^π.

3.3 Not Too Much Optimism (15 points)

Being optimism itself is not enough for efficient learning, afterall, $+\infty$ is always a valid upper bound of V^π. We want the gap between $V^{\dagger,\pi}$ and V^π not too big. Prove the following:

$$V^{\dagger,\pi} \leq V^\pi + \frac{1}{(1-\gamma)^2} \sum_{s,a\in(\mathcal{S}\times\mathcal{A})\setminus\mathcal{K}} d^\pi(s,a),$$

i.e., the gap is upper bounded by the probability of π escaping the known set $\mathcal{K}$.

Hint: first write $V^{\dagger,\pi}$ and V^{π} using the state-action distributions. Second, group state-action pairs into three groups: the known, the unknown, and the additional $s^{\dagger}$. Reason each term, and use the results you proved in Section 3.1.

3.4 Summary

(No question in this section) So in summary, we have seen another approach of providing optimism: we construct absorbing MDP where for any unknown state-action pair, we simply let it be absorbed into $s^{\dagger}$ and give the maximum possible reward one for the unknown state-action pair and $s^{\dagger}$. We also show that the optimism gap is related to the probability of escaping the known set $\mathcal{K}$. We will continue in Homework 3 to see how this construction gives an a polynomial sample complexity exploration algorithm in tabular MDPs.