Homework 3: Review on MDP, RL, and Imitation Learning

CS 4789/5789: Introduction to Reinforcement Learning

(Due May 11 11:59 ET)

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 Review on Markov Decision Processes

Consider two infinite horizon MDPs as follows $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, r, P, \mu\}$ and $\widehat{\mathcal{M}} = \left\{\mathcal{S}, \mathcal{A}, \gamma, r, \widehat{P}, \mu\right\}$, where μ is the initial state distribution. Both MDPs share the same information except the transition, and let us assume that:

$$\max_{s,a} \|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq \epsilon.$$

Given a policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$, let us denote $\mathbb{P}_h^\pi(s;\mu)$ as the probability of π hitting s at time step h starting with μ under $\mathcal{M}$, and $\widehat{\mathbb{P}}_h^\pi(s;\mu)$ as the probability of π hitting s at time step h starting with μ under $\widehat{\mathcal{M}}$. Note that both have the same initial state distribution μ.

In this question, we will quantify the difference between $\widehat{\mathbb{P}}_h^\pi$ and $\mathbb{P}_h^\pi$. This question is extremely important as it translates the error in models to the differences between the resulting state distributions.

Recall that for any $s \in \mathcal{S}$, we define:

$$\mathbb{P}_h^\pi(s;\mu)$$
$$= \sum_{s_0 \in \mathcal{S}, a_0 \in \mathcal{A}, \ldots, s_{h-1} \in \mathcal{S}, a_{h-1} \in \mathcal{A}} \mu(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\ldots P(s_{h-1}|s_{h-2},a_{h-2})\pi(a_{h-1}|s_{h-1})P(s|s_{h-1},a_{h-1}),$$

i.e., the probability of π hitting s at time step h by following π starting from μ.

Q1: Use the Markov property, prove the following equality:

$$\forall h \geq 1, s' \in \mathcal{S} : \sum_{s \in \mathcal{S}} \mathbb{P}_{h-1}^\pi(s;\mu) \sum_{a \in \mathcal{A}} \pi(a|s)P(s'|s,a) = \mathbb{P}_h^\pi(s';\mu).$$

Q2: Let us simply focus on $h = 1$. Note that at $h = 0$, both MDPs have the same μ. Prove the following:

$$\left\|\widehat{\mathbb{P}}_1^\pi(\cdot;\mu) - \mathbb{P}_1^\pi(\cdot;\mu)\right\|_1 \leq \epsilon$$

where ℓ_1 norm here means that $\left\|\widehat{\mathbb{P}}_1^\pi(\cdot;\mu) - \mathbb{P}_1^\pi(\cdot;\mu)\right\|_1 := \sum_{s\in\mathcal{S}} \left|\widehat{\mathbb{P}}_1^\pi(s;\mu) - \mathbb{P}_1^\pi(s;\mu)\right|_1$. Here you may want to use the following inequality that we used quite a few times during the semester:

$$|\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim P} g(x)| \leq \mathbb{E}_{x\sim P} |f(x) - g(x)|.$$

Q3: Now we move on to any $h \geq 1$. Prove the following by using induction:

$$\left\|\widehat{\mathbb{P}}_h^\pi(\cdot;\mu) - \mathbb{P}_h^\pi(\cdot;\mu)\right\|_1 \leq h\epsilon.$$

Remark As we see now, there is an error amplification, i.e., while the two models disagree with each other by ϵ per time step, the error will accumulate and result a h amplification at time step h when we executing π under the two models for h many steps.

Q4: Finally, we are ready to bound $\|d_\mu^\pi - \widehat{d}_\mu^\pi\|_1$ where $\widehat{d}_\mu^\pi$ is the average state distribution defined with respect to model $\widehat{\mathcal{M}}$. Prove the following:

$$\left\|d_\mu^\pi - \widehat{d}_\mu^\pi\right\|_1 \leq \frac{\gamma}{1-\gamma}\epsilon.$$

Again we are experiencing an $1/(1-\gamma)$ error amplification.

Hint: how to calculate $\sum_{h=1}^\infty \alpha^h h$ where $\alpha \in (0,1)$? (Checkout Arithmetico–geometric sequences)

2 Global Optimality in Policy Optimization

In this section, let us study the global optimality of policy optimization. We will do that via the Conservative Policy Iteration algorithm. Again define the infinite horizon MDP as follows $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, r, P, \mu\}$, where μ is the initial state distribution. Given a policy class Π, denote the global optimal policy $\pi^\star$ from the policy class Π as follows:

$$\pi^\star = \arg\max_{\pi\in\Pi} \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h r(s_h, a_h) | a_h \sim \pi(\cdot|s_h), s_0 \sim \mu\right],$$

i.e., $\pi^\star$ maximizes the expected total reward with μ with μ as the initial state distribution. For notation simplicity, we will denote V_μ^π as the expected total reward of policy π starting from the initial distribution μ, i.e., $V_\mu^\pi = \mathbb{E}_{s_0\sim\mu} V^\pi(s_0)$. Also recall the average state distribution:

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

where $\mathbb{P}_h^\pi(s;\mu)$ is the probability of hitting state s at time step h when following π and starting from a state distribution μ as the initial state distribution.

Recall the conservative policy iteration algorithm. At iteration t, given π^t, CPI first finds the greedy policy $\pi' = \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s,a)$, and then updates policy as $\pi^{t+1}(\cdot|s) = (1-\alpha)\pi^t(\cdot|s) + \alpha\pi'(\cdot|s)$ (recall that π' is deterministic here, but deterministic policy is just a special stochastic policy); CPI terminates if $\mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s,\pi'(s)) \leq \epsilon \in \mathbb{R}^+$, where ϵ is the pre-defined error threshold. Just likely gradient descent algorithm, CPI in general only guarantees a local optimal solution, and cannot guarantee global optimality.

Thus, to ensure CPI finds a globally optimal solution, we need some additional information and assumption. Specifically, we will assume that we have access to an exploratory distribution $\nu \in \Delta(\mathcal{S})$, and we can reset based on ν instead of μ. Let us run CPI using ν as the new initial distribution (forget about μ for now):

1. $\pi' = \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\nu^{\pi^t}}[A^{\pi^t}(s,\pi'(s))]$,
2. Terminate if $\mathbb{E}_{s\sim d_\nu^{\pi^t}}[A^{\pi^t}(s,\pi'(s))] \leq \epsilon$,
3. Otherwise $\pi^{t+1}(\cdot|s) = (1-\alpha)\pi^t(\cdot|s) + \alpha\pi'(\cdot|s)$.

Note that the state distribution now is defined as d_ν^π rather than d_μ^π. I.e., our algorithm now leverages the given additional distribution ν. The question we want to ask is that under what condition of ν, the above modified CPI algorithm can find a globally optimal policy for the original MDP?

Assume the following conditions holds:

$$\max_{s,a} \frac{d_\mu^{\pi^\star}(s)}{\nu(s)} \leq C < \infty, \qquad \forall t : \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\nu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] = \mathbb{E}_{s\sim d_\nu^{\pi^t}}[\max_{a\in\mathcal{A}} A^{\pi^t}(s,a)]$$

The first condition says that ν happens to cover $d_\mu^{\pi^\star}$, and the second condition is the condition on Π and says that for all t, we happen to have $\arg\max_a A^{\pi^t}(s,a) \in \Pi$.

Before we dive into the questions, keep in mind that μ is the original given initial state distribution of the MDP, and we want to optimize $\max_{\pi\in\Pi} V_\mu^\pi$ with respect to the given initial state distribution μ. However, we are given with some additional distribution ν from where we can sample state and reset to that state. We modify CPI to use ν to help us find the global optimal solution $\arg\max_{\pi\in\Pi} V_\mu^\pi$.

Q1: Let us prove the following: if the above modified CPI algorithm terminates at iteration t, we must have that:

$$\epsilon \geq \left(\min_{s\in\mathcal{S}} \frac{d_\nu^{\pi^t}(s)}{d_\mu^{\pi^\star}(s)}\right) \mathbb{E}_{s\sim d_\mu^{\pi^\star}} \max_{a\in\mathcal{A}} A^{\pi^t}(s,a).$$

Q2: show the following: for any π^t and any $s \in \mathcal{S}$, we must have:

$$\frac{d_\nu^{\pi^t}(s)}{d_\mu^{\pi^\star}(s)} \geq (1-\gamma)\frac{\nu(s)}{d_\mu^{\pi^\star}(s)}$$

Q3: Finally, conclude by proving that under the original initial distribution μ, we have:

$$V_\mu^{\pi^\star} - V_\mu^{\pi^t} \leq \frac{C\epsilon}{(1-\gamma)^2}.$$

Remark At this stage, we see that as long as the additional distribution ν covers the optimal policy, i.e., C is finite, then CPI with the help from ν indeed finds a policy that is nearly global-optimal *under the original objective function defined with respect to the given initial distribution* μ. While finite C sounds like a bit strong assumption, there is not that much one could do in general unfortunately.

3 Compatible Function Approximation in Actor-Critic Framework

Recall that during the policy gradient lecture, we briefly touched the Actor-critic framework where we fit a function approximator to approximate the Q values. Let's formalize the idea here.

Imagine that we have $\pi_\theta : \mathcal{S} \mapsto \Delta(\mathcal{A})$ at hand, and we want to compute the policy gradient with respect to θ. Instead of using roll-out to get unbiased estimate of $Q^{\pi_\theta}(s,a)$, we will try to learn $Q^{\pi_\theta}(s,a)$. Specifically, we will use another parameterized function $Q_w(s,a)$ where $w \in \mathbb{R}^d$ (this Q_w is called critic, and π_θ is called actor. Here we'll assume that $Q_w(s,a)$ is differentiable with respect to w.) to approximate Q^{π_θ}, and we do that by minimizing the following least square objective:

$$\widehat{w} := \arg\min_w \mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$$

We then use $Q_{\widehat{w}}$ as the approximator of Q^{π_θ} to form the policy gradient:

$$\widehat{\nabla_\theta J_\theta} := \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a|s) Q_{\widehat{w}}(s,a).$$

In general, $\widehat{\nabla_\theta J_\theta} \neq \nabla_\theta J(\theta)$ as $Q_{\widehat{w}}$ is just a learned approximator (and it is learned under $d_\mu^{\pi_\theta}$ as the training distribution).

Q1: Recall $\widehat{w}$ is the minimizer of the least square loss function. Prove that for $\widehat{w}$, we must have:

$$\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left(Q_{\widehat{w}}(s,a) - Q^{\pi_\theta}(s,a)\right) \nabla_w Q_{\widehat{w}}(s,a) = 0,$$

where again $\nabla_w Q_{\widehat{w}}(s,a)$ is in short of $\nabla_w Q_w(s,a)|_{w=\widehat{w}}$.

Q2: Q_w and π_θ are called compatible if:

$$\forall s,a, \nabla_w Q_{\widehat{w}}(s,a) = \nabla_\theta \ln \pi_\theta(a|s).$$

Prove that under the compatible assumption, we have:

$$\widehat{\nabla_\theta J(\theta)} = \nabla_\theta J(\theta).$$

Q3: Consider a specific parameteriation: $Q_w(s,a) = w^\top \nabla_\theta \ln \pi_\theta(a|s)$. Clearly, this is a compatible setting, i.e., $\nabla Q_w(s,a) = \nabla_\theta \ln \pi_\theta(a|s)$ always. (a) Write out the closed-form of $\widehat{w}$:

$$\widehat{w} = \arg\min_w \mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left(w^\top \nabla_\theta \ln \pi_\theta(a|s) - Q^{\pi_\theta}(s,a)\right)^2,$$

where we assume that $\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a|s)(\nabla_\theta \ln \pi_\theta(a|s))^\top\right]$ is a full rank matrix. (b) Show that $\widehat{\nabla_\theta J_\theta}$ is indeed equal to the true policy gradient by substituting the $\widehat{w}$ in the formulation $\widehat{\nabla_\theta J_\theta} = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \nabla\theta \ln \pi_\theta(a|s) Q_{\widehat{w}}(s,a)$ with what you got from (a).

4 The Min-Max Imitation Learning Framework

Let us study a min-max optimization framework for Imitation Learning. Again define the infinite horizon MDP as follows $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, r, P, \mu\}$, where μ is the initial state distribution, and $r : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$.

We will consider two function approximator classes: policy class Π, and discriminator class $\mathcal{F}$ where $\pi \in \Pi$ is a policy, and $f \in \mathcal{F}$ is a discriminator such that $f : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$.

Denote $\pi^e : \mathcal{S} \mapsto \Delta(\mathcal{A})$ as our expert policy.

The min-max formulation of IL concerns the following optimization procedure:

$$\min_{\pi\in\Pi} \left[\max_{f\in\mathcal{F}} \left[\mathbb{E}_{s,a\sim d_\mu^\pi} f(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^e}} f(s,a)\right]\right].$$

Intuitively, this means that we want to find a policy π, such that the all discriminators $f \in \mathcal{F}$ cannot distinguish d_μ^π and $d_\mu^{\pi^e}$. Note that here $d_\mu^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ is the average state-action distribution induced by π.

Q1: First consider a fixed $\pi \in \Pi$. Let us first get a sense of how $\max_{f\in\mathcal{F}} \left[\mathbb{E}_{s,a\sim d_\mu^\pi} f(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^e}} f(s,a)\right]$ can be understood as a distribution divergence. To do that, let us assume that $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \mapsto [-1,1]\}$, i.e., $\mathcal{F}$ contains *all possible functions* that maps from (s,a) to a scalar between $[-1,1]$. Prove the following:

$$\left\|d_\mu^\pi(\cdot,\cdot) - d_\mu^{\pi^e}(\cdot,\cdot)\right\|_1 = \max_{f\in\mathcal{F}} \left[\mathbb{E}_{s,a\sim d_\mu^\pi} f(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^e}} f(s,a)\right]$$

Remark: so here we see that if our discriminator class $\mathcal{F}$ is rich enough to contain all possible functions that map from s,a to scalar in $[-1,1]$, $\max_{f\in\mathcal{F}} \left[\mathbb{E}_{s,a\sim d_\mu^\pi} f(s,a) - \mathbb{E}_{s,a\sim d_\mu^{\pi^e}} f(s,a)\right]$ is just the usual ℓ_1 of the difference between two distributions (which is equal to the Total variation distance up to a constant 2).

Q2: Assume that we approximately solve the min-max formulation, and we get a policy $\widehat{\pi}$ such that:

$$\max_{f \in \mathcal{F}} \left[\mathbb{E}_{s,a \sim d_\mu^{\widehat{\pi}}} f(s,a) - \mathbb{E}_{s,a \sim d_\mu^{\pi^e}} f(s,a) \right] \leq \delta \in \mathbb{R}^+,$$

where δ may be some small number due to possible statistical and optimization error. Let us prove the following: assume that $r \in \mathcal{F}$, and $\mathcal{F}$ is symmetric, i.e., if $f \in \mathcal{F}$, then $-f \in \mathcal{F}$ as well, we must have:

$$\left| V_\mu^{\widehat{\pi}} - V_\mu^{\pi^e} \right| \leq \frac{\delta}{1-\gamma}.$$

Remark: this shows that as long as the discriminator class $\mathcal{F}$ is rich enough to contain r, and being symmetric, the min-max formulation can learn a policy that performs almost as good as the expert.