CS234 Problem Session

Week 6: Feb 17

1) [CA Session] Conservative Policy Iteration

Let us consider an MDP with a fixed start state s_0.

Let us consider the conservative policy update rule:

$$\pi_{new}(s,a) = (1-\alpha)\pi(s,a) + \alpha\pi'(s,a)$$

for some $\alpha \in [0,1]$.

(a) What is $\pi_{new}(s,a)$ when $\alpha = 1$?

Recall that $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$.

Let $P(s_t;\pi)$ be the distribution over states at time t while following π from the start state s_0. Recall that the discounted stationary state distribution of a policy π is $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s;\pi)$. We now define the policy advantage of some policy π' with respect to a policy π as $\mathbb{A}^\pi(\pi') = \mathbb{E}_{s\sim d^\pi}[\mathbb{E}_{a\sim\pi'(s)}[A^\pi(s,a)]]$. Recall Lemma 1 from assignment 2.

Lemma 1: For all policies π', π, we have that $V^{\pi'}(s_0) - V^\pi(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}[\mathbb{E}_{a\sim\pi'(s)}[A^\pi(s,a)]]$

(b) How does $V^{\pi'}(s_0) - V^\pi(s_0)$ differ from the policy advantage $\mathbb{A}^\pi(\pi')$? A high-level description in words will suffice.

(c) Compute a simplified expression for $\mathbb{A}^{\pi}(\pi_{new})$ in terms of the policy advantage of π'.

With π_{new}, at any given timestep, the probability that we select an action according to π' is α. Let us define the random variable c_t as the number of actions chosen from π' before time t.

(d) Let us denote $\rho_t = Pr(c_t \geq 1)$. Compute an expression for ρ_t in terms of α and t.

Now let $\epsilon = max_s|\mathbb{E}_{a\sim\pi'(s)}[A^{\pi}(s,a)]|$.

(e) Prove that $\mathbb{E}_{s\sim P(s_t;\pi_{new})}[\sum_a \pi_{new}(s,a)A^{\pi}(s,a)] \geq \alpha\mathbb{E}_{s\sim P(s_t;\pi)}[\sum_a \pi'(s,a)A^{\pi}(s,a)] - 2\alpha\rho_t\epsilon$.

(f) Now let us lower bound the improvement of our policy. Please prove that the following equation holds:

$$V^{\pi_{new}}(s_0) - V^{\pi}(s_0) \geq \frac{\alpha}{1-\gamma}(\mathbb{A}^{\pi}(\pi') - \frac{2\alpha\gamma\epsilon}{1-\gamma(1-\alpha)})$$

2) [Breakout Rooms] Trajectory Likelihoods

Suppose π_1 and π_2 are two different stochastic policies. We now observe a trajectory $H = (S_0, A_0, R_0, S_1, ...S_{T-1}, A_{T-1}, R_{T-1})$. Assume the rewards are finite and denote $R(s, a, s', r) = Pr(R_t = r | S_t = s, A_t = a, S_{t+1} = s')$.

(a) Simplify $\frac{Pr(H|\pi_1)}{Pr(H|\pi_2)}$ using terms from the MDP definition. Your final answer should be able to be computed without needing to know the transition function, the reward function, or the reward distribution.

3) [Breakout Rooms] Off Policy Actor Critic Policy Gradients

We will derive an expression for the policy gradient for a new objective function, J'. This new objective is similarto one used in off-policy actor-critics. Assume there is a fixed policy π_b. Let

$$d'(s) = \sum_{t=0}^{L-1} Pr(S_t = s|\pi_b)$$

The objective function J' is defined as

$$J'(\theta) = \sum_{s \in S} d'(s) E[R_t|S_t = s, \theta]$$

Derive an expression for the policy gradient for this objective. The terms in your answer should only be terms used in defining an MDP (including the reward function defined as R(s, a)). Note that θ are not the parameters of π_b, but the parameters of another policy π.

.