COMS6998-11: Homework 2

Akshay Krishnamurthy
akshay@cs.umass.edu
Due: Thursday, 4/21

Instructions: Turn in your homework to me by email by Thursday 4/21.

1. **Data re-use in NPG.** In this problem we will study one approach for reusing data in natural policy gradient. To do this we will consider an importance sampling estimator that we will define below. We work in the discounted tabular MDP setting $M = (\mathcal{S}, \mathcal{A}, P, R, \mu, \gamma)$ and consider the tabular softmax parametrization for natural policy gradient. Assume that $R : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$.

   To start off, let us consider two softmax policies π_1 and π_2 related by

$$\pi_2(a \mid s) = \pi_1(a \mid s) \cdot \frac{\exp(c(s,a))}{\sum_{a'} \pi_1(a \mid s) \exp(c(s,a))} \tag{1}$$

   where $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is some fixed function with $\sup_{s,a} |c(s,a)| \leq c_\star$. We estimate $Q^{\pi_2}(s,a)$ via geometric resampling and taking actions according to π_1, that is: we start with $s_0 = s, a_0 = a$, take actions via π_1 and stop at each time t with probability $1 - \gamma$. This means we stop immediately with probability $1 - \gamma$, and use the immediate reward $r_0 \sim R(s_0, a_0)$ below. Letting $t_\star$ denote the stopping time, we form the estimate

$$\widehat{Q}^{\pi_2}(s,a) = \left(\prod_{t=1}^{t_\star} \frac{\pi_2(a_\tau \mid s_\tau)}{\pi_1(a_\tau \mid s_\tau)} \right) \frac{r_{t_\star}}{1-\gamma}$$

   (Note that compared to the estimator we used in lecture, we have made two changes. First we are only using the last reward $r_{t_\star}$ rather than the sum of the rewards, although we are scaling it up by $1/(1-\gamma)$. Second, we are importance weighting by the density ratio of the two policies along the way. Note that we are not importance weighting with the 0^{th} term, since we are trying to estimate $Q(s_0, a_0)$, both policies will agree on this.)

   (a) First show that $\widehat{Q}^{\pi_2}(s,a)$ is an unbiased estimator for $Q^{\pi_2}(s,a)$, that is $\mathbb{E}[\widehat{Q}^{\pi_2}(s,a)] = Q^{\pi_2}(s,a)$.

   (b) Next, show that for all s, a

$$\exp(-2c_\star) \leq \frac{\pi_2(a \mid s)}{\pi_1(a \mid s)} \leq \exp(2c_\star)$$

   (c) Finally, provide high probability upper bounds on both $t_\star$ and the range of the estimator. That is find values $t_{\max}$ and $Q_{\max}$ (in terms of $c_\star$ and γ) such that

$$\Pr\left(t_\star \geq t_{\max} \text{ or } \widehat{Q}^{\pi_2}(s,a) \geq Q_{\max} \right) \leq \delta$$

   While we will not work through the details, this importance weighting estimator can be used in combination with NPG, since the NPG update takes a form similar to (1). Thus we can use the data collected from past iterates, or alternatively, we can just collect data once every few iterates. However, there is a competing force that prevents us from taking this to the extreme. If we collect data very infrequently, the iterates will be quite far apart, so $c_\star$ and hence $Q_{\max}$ will be large and we will have worse estimation error, since that degrades with the range of the random variable (e.g., from Hoeffding's inequality).

2. **Tabular RL with a generative model.** While we studied some harder settings for learning in MDPs in lecture, let us focus on a simpler setting here. This is called the "generative model" setting. There is a discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \mu, \gamma)$, where P, R, μ are all unknown. For learning, we have access to a *generative model*, Samp, that can be queried with a pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and, when queried, it returns a sample (r, s') where $r \sim R(s, a)$ and $s' \sim P(s, a)$. In this problem, we will use this generative model to learn a near-optimal policy for M.

   (a) First prove the simulation lemma for discounted MDPs. Suppose we have two MDPs $M_1 = (\mathcal{S}, \mathcal{A}, P_1, R_1, \mu, \gamma)$ and $M_2 = (\mathcal{S}, \mathcal{A}, P_2, R_2, \mu, \gamma)$ such that

$$\forall (s, a): \quad \|P_1(s, a) - P_2(s, a)\|_{\mathrm{TV}} \le \varepsilon, \qquad |\mathbb{E}_{r \sim R_1(s,a)}[r] - \mathbb{E}_{r \sim R_2(s,a)}[r]| \le \varepsilon.$$

   Then for any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ we have

$$\sup_{s,a} |Q^{\pi}_{M_1}(s, a) - Q^{\pi}_{M_2}(s, a)| \le \frac{2\varepsilon}{(1-\gamma)^2}$$

   Here recall that TV denotes the total variation distance, or $1/2$ of the ℓ_1 norm. That is $\|P_1(s, a) - P_2(s, a)\|_{\mathrm{TV}} = \frac{1}{2} \sum_{s'} |P_1(s'|s, a) - P_2(s'|s, a)|$. Also we use $Q^{\pi}_{M_1}$ to denote the action-value function of policy π in M_1.

   (b) Equipped with the simulation lemma, show that if we simply query Samp polynomially many times from each (s, a), we can estimate an MDP approximation $\widehat{M}$ that satisfies the preconditions of the simulation lemma with respect to the true MDP M.

   (c) Explain then how to find a policy $\hat{\pi}$ that is ϵ sub-optimal (with probability at least $1 - \delta$). What is the total sample complexity of this approach?

   Optional challenge question: can you get a better sample complexity?

3. **Generative models and Linear MDPs.** Now let us extend the above analysis to the Linear MDP setting. We have a MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and a feature map $\phi : \mathcal{S}, \mathcal{A} \to \mathbb{R}^d$ satisfying the Linear MDP property:

$$P(s' \mid s, a) = \langle \phi(s, a), \mu(s') \rangle, \qquad \mathbb{E}_{r \sim R(s,a)}[r] = \langle \phi(s, a), \theta^{\star} \rangle. \tag{2}$$

   (You may assume the standard normalizations for linear MDPs, that is $\|\phi(s, a)\|_2 \le 1, \|\theta^{\star}\|_2 \le 1$ and $\sup_g \|\int \mu(s) g(s) ds\|_2 \le \sqrt{d}$.) As above, we have access to a generative model Samp that takes as input s, a and produces a sample $s' \sim P(s, a)$ and $r \sim R(s, a)$. Using this generative model, we would like to compute a near-optimal policy. The twist is that we do not want the sample complexity scale with the number of states or actions.

   To do this, we will define a distribution $D \in \Delta(\mathcal{S} \times \mathcal{A})$ and we collect a dataset of n tuples the form (s, a, r, s') where $(s, a) \sim D$ and $(r, s') = \mathsf{Samp}(s, a)$. Let $\Sigma = \mathbb{E}_D[\phi(s, a)\phi(s, a)^{\top}]$ denote the feature covariance and assume $\sup_{s,a} \phi(s, a)^{\top} \Sigma^{-1} \phi(s, a) \le d$ (We will not prove this, but such a distribution always exists when the features form a compact set).

   In this problem, we define the Bellman backup operator $\mathcal{T}$ to map V-value functions to Q-value functions. That is $\mathcal{T} : (\mathcal{S} \to [0, \frac{1}{1-\gamma}]) \to (\mathcal{S} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}])$ and it is defined by $\mathcal{T}V : (s, a) \mapsto \mathbb{E}[r + \gamma V(s') \mid s, a]$. Under our setup, for any *fixed* function $V : \mathcal{S} \to [0, \frac{1}{1-\gamma}]$, we can estimate the bellman backup $\mathcal{T}V$ via linear regression:

$$\hat{w} \leftarrow \operatorname*{argmin}_{w} \frac{1}{n} \sum_{i=1}^{n} \left(\langle w, \phi(s_i, a_i) \rangle - r_i - \gamma V(s'_i) \right)^2$$
$$(\widehat{\mathcal{T}V})(s, a) = \langle \phi(s, a), \hat{w} \rangle$$

   Here $\widehat{\mathcal{T}V}$ is a function with the same type as a Q function, that is $\widehat{\mathcal{T}V} : \mathcal{S} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}]$. Owing to the linear MDP property, a concentration argument can be used to show that $\widehat{\mathcal{T}V}$ and $\mathcal{T}V$ are close. For this problem,

you can assume that, simultaneously for all functions of the form $V(s) = \max_a \langle \phi(s,a), \theta \rangle$ with $\theta \in \mathbb{R}^d$, with probability $1-\delta$ we have

$$\mathbb{E}_D \left[\left(\widehat{\mathcal{T}V}(s,a) - \mathcal{T}V(s,a) \right)^2 \right] \leq \frac{\Delta \log(1/\delta)}{n}$$

where Δ depends polynomially on d and $1/(1-\gamma)$.

(a) Suppose we perform the iteration: (1) $V^{(t)} : s \mapsto \max_a Q^{(t)}(s,a)$, (2) $Q^{(t+1)} \leftarrow \widehat{\mathcal{T}V^{(t)}}$, starting from $Q^{(1)} = 0$, for T iterations. Show that

$$\forall (s,a) : \left| Q^{(T)}(s,a) - Q^\star(s,a) \right| \leq \frac{1}{1-\gamma} \sqrt{\frac{d\Delta \log(1/\delta)}{n}} + \frac{\gamma^T}{1-\gamma}$$

(b) Use the above inequality to derive a sub-optimality bound for the policy $\hat{\pi} : s \mapsto \operatorname{argmax}_a Q^{(T)}(s,a)$. That is, obtain an upper bound on $J(\pi^\star) - J(\hat{\pi})$.

4. **Bellman rank.** Show that the feature learning linear MDP setting has V-type Bellman rank d. In this setting, we have a class of feature maps $\Phi : (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}^d$ and we assume that there exists a map $\phi^\star \in \Phi$ such that both rewards and transitions are linear in $\phi^\star$ (that is, they satisfy (2)). From here we define the class of Q-functions as $\{(s,a) \mapsto \langle \theta, \phi(s,a) \rangle : \theta \in \mathbb{R}^d, \phi \in \Phi\}$, which is the set of all linear functions on top of all features.