Homework 2: Policy Gradient Methods & RMax

CSE 599: Reinforcement Learning and Bandits

Instructions

Do any two of the six problems.

1 Policy Gradient Theorem

1. For any function $f(s)$ (that is only a function of the states) show that:

$$\nabla V^{\pi_\theta} = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d^{\pi_\theta}}\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\Big(Q^{\pi_\theta}(s,a) - f(s)\Big)\nabla\log\pi_\theta(a|s)\right]$$

2. Use this to prove the third policy gradient expression in Theorem 4.3.
3. Bonus: Find the $f(\cdot)$ which leads to the minimum variance estimate of our gradient, assuming that our gradient estimate will be:

$$\widehat{\nabla V^{\pi_\theta}} = \frac{1}{1-\gamma}\Big(Q^{\pi_\theta}(s,a) - f(s)\Big)\nabla\log\pi_\theta(a|s)$$

   where $s \sim d^{\pi_\theta}$, $a \sim \pi_\theta(\cdot|s)$.

2 Policy Gradients for the Finite Horizon Case

1. Derive analogous expressions for the policy gradient theorem (in Theorem 4.3) for the (episodic) H stage, undiscounted finite horizon MDPs. In particular, derive two expressions analogous to the first two expressions in Theorem 4.3. An H stage finite horizon MDP is one which starts at some fixed state s_0, lasts for H steps, and the value is the undiscounted sum of the rewards in H steps.

3 Compatible Function Approximation and Estimation

Read the compatible function approximation section in the notes.

1. Provide a sample based estimation procedure for this approach. Use the same quantities $\widehat{Q}_t$ defined in the "Monte Carlo estimation and stochastic gradient descent" section of the notes, and explicitly setup the regression problem that you would solve.
2. Show that the samble based estimator you obtain will result in exactly the same estimator as that provided in the "Monte Carlo estimation and stochastic gradient descent" section. In particular, the estimate used by both procedures (on the same set of samples) will end up being identical to each other.

4 Effect of Distribution Mismatch in Approximate Policy Evaluation

Recall the notation $d_{\pi,\nu}$ where states are sampled according to d_π and actions according to $\nu(a \mid s)$. In the lecture, we showed the bound

$$\|f_k - Q^\pi\|_{d^{\pi,\pi}} \leq \gamma^{k/2}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \sum_{i=1}^{k} \gamma^{(k-i)/2}\|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\nu}}$$

1. Let $\rho(s,a) = \nu(s,a)/\pi(s,a)$. Show that the result above implies the bound:

   $$\|f_k - Q^\pi\|_{d^{\pi\nu}} \leq \max\{\epsilon, 1\},$$

   where

   $$\epsilon = \gamma^{k/2}\sqrt{\|\rho\|_\infty\|1/\rho\|_\infty}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \sum_{i=1}^{k} \gamma^{(k-i)/2}\sqrt{\|\rho\|_\infty\|1/\rho\|_\infty}\|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\nu}}$$

2. Is this bound improvable in general? Consider the case when ν is uniform and π is deterministic and show that the LHS can be as large as 1 for the iteration $f_k = \widehat{\mathcal{T}}^\pi_{\mathcal{F}} f_{k-1}$. Does the choice of $\mathcal{F}$ matter?

   Hint: What can you say about the behavior of f_k on states which are not visited by π, but are visited by ν? Does this matter?

5 Bandit algorithm variants

1. Consider the active arms elimination algorithm. At each round, it maintains a feasible action set A_t, and bounds $\mathrm{LCB}_t(a)$, $\mathrm{UCB}_t(a)$ on the expected reward for each $a \in A_t$. It acts by playing an action $a \in A_t$ uniformly at random.
   (a) Can you design an elimination strategy for removing an action from A_t, assuming you are given LCB_t and UCB_t? What are reasonable values to use for LCB_t and UCB_t based on our analysis of the UCB algorithm?
   (b) Derive a bound on the regret of this algorithm. How does it compare with UCB?

2. Consider a variant of the τ-GREEDY algorithm instead parametrized by ϵ. At round t, it has a history $(x_s, a_s, p_s, r_s)_{s=1}^{t-1}$. It computes the policy $\pi_t = \text{argmax}_{\pi \in \Pi} \sum_{s=1}^{t-1} \widehat{r}_s(\pi(x_s))$, where $\widehat{r}_s(a)$ is computed via IPS as in the lectures. The algorithm then samples an action according to $p_t(a) = (1-\epsilon)\pi_t(x_t) + \epsilon/K$, and adds the resulting sample along with the reward to its history. Modify the analysis of τ-GREEDY to obtain a regret bound for this scheme. Do you see any practical benefits of this alternative approach?

6 Off-policy evaluation of contextual bandit exploration algorithms

In the lecture, we saw how to evaluate a static policy, and by union bound, a large collection of policies using IPS and improvements. A natural question is whether these techniques can be also used to evaluate the regret we would incur when we run a contextual bandit exploration algorithm online, using previously collected data. For instance, we could use such a scheme to tune parameters such as τ (or ϵ above), policy class, etc. in our algorithms. Turns out this is not possible. To formalize this, an exploration algorithm $\mathcal{A}$ can be thought of as a mapping $\mathcal{A}(t, h_t, x_t)$ to a distribution over actions, where h_t is the history of interactions in the first $t-1$ rounds. Show that given a dataset $(x_t, a_t, r_t, p_t)_{t=1}^{T}$, and an arbitrary exploration algorithm $\mathcal{A}$, no estimator such $\hat{v}_T$ can be computed such that $\left|\hat{v}_T - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim D, a \sim \mathcal{A}(t, h_t, x_t)}[r(a) \mid x_t]\right| \leq 1/2$, with probability at least $1/3$, no matter how large T is.

Hint: Think of contextual bandit algorithms which branch along two very different behaviors based on some initial history.