Homework III

Deadline: 2023-12-17

1. (10 pts) Recall the definition of state visitation measure
$$d_\mu^\pi(s) = \mathbb{E}_{s_0\sim\mu}\left[d_{s_0}^\pi(s)\right] = \mathbb{E}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left[s_t = s|s_0,\pi\right]\right],$$
where $(s_0, a_0, s_1, a_1, \cdots)$ is trajectory starting from initial distribution μ and then following policy π. Let T obey the geometric distribution, i.e., $\mathbb{P}\left[T = t\right] = \gamma^t(1-\gamma), \quad t = 0, 1\cdots$. Show that
$$\mathbb{P}\left[s_T = s\right] = d_\mu^\pi(s).$$
Then suggest a way to sample from d_μ^π.

2. (20 pts) Implement and test the Projected Policy Gradient method and the Softmax Policy Gradient method in Lecture 7 for the Gridworld problem in Homework I (Question 7, use $\gamma = 0.9$ and uniform distribution for μ). The action/advantage values and visitation measure in the policy gradient should be evaluated exactly based on the transition model. Display the convergence plots ($V^*(\mu) - V^k(\mu)$ *vs* # of iterations) of the two algorithms in a figure. Can you observe the finite iteration convergence of the Projected Policy Gradient method?

3. Consider the soft policy iteration algorithm in Lecture 8 (page 35).

   - (10 pts) Show the policy improvement property of the algorithm:
   $$V_\lambda^{\pi_{k+1}}(s) \geq V_\lambda^{\pi_k}(s), \quad \forall s.$$
   - (10 pts) Show the γ-rate convergence of the algorithm:
   $$\|V_\lambda^* - V_\lambda^{\pi_k}\|_\infty \leq \gamma^k\|V_\lambda^* - V_\lambda^{\pi_0}\|_\infty.$$

4. Consider the entropy regularized state value function $V_\lambda^{\pi_\theta}(\mu)$ (see Lecture 8 for the definition of the entropy regularized state value function) under a parameterized policy π_θ. Assume $\sum_{a\in\mathcal{A}}\pi_\theta(a|s) = 1$ for all θ.

   - (10 pts) Show the following policy gradient expression for $\nabla_\theta V_\lambda^{\pi_\theta}(\mu)$:
   $$\nabla_\theta V_\lambda^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_\lambda^{\pi_\theta}(s,a)\nabla_\theta\log\pi_\theta(a|s)\right],$$
   where $A_\lambda^{\pi_\theta}(s,a) = Q_\lambda^{\pi_\theta}(s,a) - \lambda\log\pi_\theta(a|s) - V_\lambda^{\pi_\theta}(s)$.

- (15 pts) Consider the natural policy gradient method with entropy regularization:

$$\theta^{+} = \theta + \eta \cdot F(\theta)^{\dagger} \nabla_{\theta} V_{\lambda}^{\pi_\theta}(\mu),$$

where

$$F(\theta) = \mathbb{E}_{s \sim d_{\mu}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \left(\nabla_\theta \log \pi_\theta(a|s) \right)^{T} \right].$$

Show that the update under the softmax parameterization in the policy space can be expressed as:

$$\pi_{\theta^{+}}(a|s) \propto [\pi_\theta(a|s)]^{\left(1 - \frac{\lambda \eta}{1-\gamma}\right)} \exp\left(\frac{\eta}{1-\gamma} Q_{\lambda}^{\pi_\theta}(s,a) \right).$$

For which value of η does this update reduce to the soft policy iteration?