Homework III

Deadline: 2026-01-15

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., $\mathbb{P}[T=t] = \gamma^t(1-\gamma), \quad t = 0, 1\cdots$. Show that

$$\mathbb{P}[s_T = s] = d_\mu^\pi(s).$$

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 8, 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 28).

   - (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. (10 pts) NPG under the entropy regularization admits the following form (same as the one given in Lecture 8 expect the constant):

$$\pi^+(a|s) \propto \pi(a|s)\exp\left(\frac{\eta}{\eta\tau+1}A_\tau^\pi(s,a)\right).$$

Show that this is equivalent to

$$\pi^+(\cdot|s) = \arg\max_p \left\{\eta\left(\langle Q_\tau^\pi(s,\cdot), p\rangle + \tau H(p)\right) - \mathrm{KL}(p\|\pi(\cdot|s))\right\},$$

where $H(p) = -\sum_a p_a \log p_a$.

5. (20 pts) Reproduce the figure on page 25 of Lecture 9 for comparing different bandit algorithms.