Homework III

Deadline: 2025-01-05

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. (5 pts) Let V_τ^π and Q_τ^π be the value functions under the entropy regularization, and recall the definition of the Bellman optimality operator $\mathcal{T}_\tau$ in this case. Show that

$$\mathcal{T}_\tau V_\tau^\pi(s) - V_\tau^\pi(s) = \tau \mathrm{KL}(\pi(\cdot|s) \| \widehat{\pi}(\cdot|s)),$$

   where $\widehat{\pi}(\cdot|s) \propto \exp(Q_\tau^\pi(\cdot|s)/\tau)$.

4. 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:

$$\|V_\lambda^* - V_\lambda^{\pi_k}\|_\infty \leq \gamma^k \|V_\lambda^* - V_\lambda^{\pi_0}\|_\infty.$$

5. (20 pts) Reproduce the figure on page 26 of Lecture 9 for comparing different bandit algorithms.