Homework III

Deadline: 2023-1-7

1. (10 pts) Prove the convergence of Q-Learning (i.e., Theorem 4 of lecture5.pdf).

2. (10pts) Reproduce the figure in pg. 22 of lecture5.pdf.

3. (5 pts) Given a policy π, let v_π be the state value that corresponds to π and $\mathcal{T}_\pi$ be the corresponding Bellman operator. Recall the definition of Bellman error (under infinity norm) for a vector v:

$$\mathrm{BE}(v) = \|v - \mathcal{T}_\pi v\|_\infty.$$

Show that

$$\|v - v_\pi\|_\infty \leq \frac{\mathrm{BE}(v)}{1-\gamma}.$$

4. (5 pts) Consider the following softmax parameterization

$$\pi_\theta(a|s) = \frac{\exp(\theta_a)}{\sum_{a' \in \mathcal{A}} \exp(\theta_{a'})},$$

where $\theta = (\theta_a)_{a \in \mathcal{A}}$. Calculate $\nabla_\theta \log \pi_\theta(a|s)$.

5. (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.

6. (5 pts) Show the following expression for the Fisher information matrix in NPG for policy optimization (see pg. 8 of lecture8.pdf):

$$F(\theta) = \mathbb{E}_{\tau \sim \mathbb{P}_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)(\nabla_\theta \log \pi_\theta(a_t|s_t))^T\right].$$

7. (5 pts) Show the statement "$J(\theta)$ and $L_{\theta_{\mathrm{old}}}(\theta)$ matches at θ_{old} up to first derivative" in pg. 11 of lecture8.pdf.