CS234: Reinforcement Learning – Problem Session #3

Spring 2023-2024

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ where $\gamma \in [0, 1)$ and the state-action space is finite ($|\mathcal{S} \times \mathcal{A}| < \infty$). For any stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, recall that the discounted stationary-state distribution is defined such that, for any state $s \in \mathcal{S}$,

$$d^\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s),$$

where $\mathbb{P}^\pi(s_t = s)$ denotes the probability that the (random) state s_t encountered by policy π at timestep t is equal to s. Let $\beta \in \Delta(\mathcal{S})$ be an initial state distribution such that $\mathbb{P}^\pi(s_0 = s) = \beta(s)$ for all policies π and any state $s \in \mathcal{S}$.

1. Prove that for any state $s' \in \mathcal{S}$,

$$d^\pi(s') = (1 - \gamma)\beta(s') + \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a)\pi(a \mid s) d^\pi(s).$$

2. Show that for any two policies π, π', we have

$$||d^{\pi} - d^{\pi'}||_1 \leq \frac{2\gamma}{(1-\gamma)} \mathbb{E}_{s \sim d^{\pi}} \left[D_{\mathrm{TV}} \left(\pi(\cdot \mid s) \mid\mid \pi'(\cdot \mid s) \right) \right],$$

where $D_{\mathrm{TV}} \left(\pi(\cdot \mid s) \mid\mid \pi'(\cdot \mid s) \right) = \frac{1}{2} \sum_{a \in \mathcal{A}} |\pi(a \mid s) - \pi'(a \mid s)|$ is the total variation distance between policies π and π' at state s.

Hint: Use a "zero" term involving d^{π}.

3. Denote the stationary state-action visitation distribution $\chi^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ of a policy as $\chi^\pi(s,a) = d^\pi(s)\pi(a \mid s)$. Show that for any two policies π, π', we have

$$||\chi^\pi - \chi^{\pi'}||_1 \leq \frac{2}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\mathrm{TV}} \left(\pi(\cdot \mid s) \;||\; \pi'(\cdot \mid s) \right) \right].$$

4. Define $R_{\mathrm{MAX}} = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\mathcal{R}(s,a)|$ and show that

$$\mathbb{E}_{s_0 \sim \beta} \left[V^\pi(s_0) - V^{\pi'}(s_0) \right] \leq \frac{2R_{\mathrm{MAX}}}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\mathrm{TV}} \left(\pi(\cdot \mid s) \;||\; \pi'(\cdot \mid s) \right) \right].$$

Hint: Remember that $\mathbb{E}_{s_0 \sim \beta}[V^\pi(s_0)] = \mathcal{R}^\top \chi^\pi$, *where* $\mathcal{R} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ *is the vector of all MDP rewards, and recall Hölder's inequality.*