CS234: Reinforcement Learning – Problem Session #4

Winter 2022-2023

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, \gamma \rangle$ where $\gamma \in [0, 1)$, $\beta \in \Delta(\mathcal{S})$ is the initial state distribution, and the state-action space is finite ($|\mathcal{S} \times \mathcal{A}| < \infty$). For any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, recall that the discounted stationary state distribution is defined for any state $s' \in \mathcal{S}$ as

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

Solution: For those who showed up on Zoom, my typo was swapping the L_1-norm with the L_∞-norm. Apologies for the confusion.

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

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

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

Solution: This result is given as Lemma 3 of Achiam et al. [2017]. Applying the definitions for the visitation distributions of π and π', we have

$$\begin{aligned}
||d^\pi - d^{\pi'}||_1 &= \sum_{s' \in \mathcal{S}} |d^\pi(s') - d^{\pi'}(s')| \\
&= \sum_{s' \in \mathcal{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) - (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)| \\
&= \sum_{s' \in \mathcal{S}} \gamma | \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) \left(\pi(a \mid s) d^\pi(s) - \pi'(a \mid s) d^{\pi'}(s) \right) | \\
&= \sum_{s' \in \mathcal{S}} \gamma | \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) \left(\pi(a \mid s) d^\pi(s) - \pi'(a \mid s) d^\pi(s) + \pi'(a \mid s) d^\pi(s) - \pi'(a \mid s) d^{\pi'}(s) \right) | \\
&\leq \sum_{s' \in \mathcal{S}} \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) d^\pi(s) |\pi(a \mid s) - \pi'(a \mid s)| + \sum_{s' \in \mathcal{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) - d^{\pi'}(s)| \\
&= \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \underbrace{\sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s, a)}_{=1} d^\pi(s) |\pi(a \mid s) - \pi'(a \mid s)| + \gamma \sum_{s \in \mathcal{S}} |d^\pi(s) - d^{\pi'}(s)| \underbrace{\sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s, a)\pi'(a \mid s)}_{=1} \\
&= \gamma \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} |\pi(a \mid s) - \pi'(a \mid s)| + \gamma ||d^\pi - d^{\pi'}||_1 \\
&= \gamma \sum_{s \in \mathcal{S}} d^\pi(s) \cdot 2 \cdot \frac{1}{2} \sum_{a \in \mathcal{A}} |\pi(a \mid s) - \pi'(a \mid s)| + \gamma ||d^\pi - d^{\pi'}||_1 \\
&= 2\gamma \mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right] + \gamma ||d^\pi - d^{\pi'}||_1 \implies ||d^\pi - d^{\pi'}||_1 \leq \frac{2\gamma}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right].
\end{aligned}$$

2. 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

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

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

Solution: Applying the definition of the stationary state-action distribution, we have

$$\begin{aligned}
||\chi^\pi - \chi^{\pi'}||_1 &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\chi^\pi(s,a) - \chi^{\pi'}(s,a)| \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |d^\pi(s)\pi(a \mid s) - d^{\pi'}(s)\pi'(a \mid s)| \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |d^\pi(s)\pi(a \mid s) - d^\pi(s)\pi'(a \mid s) + d^\pi(s)\pi'(a \mid s) - d^{\pi'}(s)\pi'(a \mid s)| \\
&\leq \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} |\pi(a \mid s) - \pi'(a \mid s)| + \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi'(a \mid s) |d^\pi(s) - d^{\pi'}(s)| \\
&= \sum_{s \in \mathcal{S}} d^\pi(s) \cdot 2 \cdot \frac{1}{2} \sum_{a \in \mathcal{A}} |\pi(a \mid s) - \pi'(a \mid s)| + \sum_{s \in \mathcal{S}} |d^\pi(s) - d^{\pi'}(s)| \underbrace{\sum_{a \in \mathcal{A}} \pi'(a \mid s)}_{=1} \\
&= 2\mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right] + ||d^\pi - d^{\pi'}||_1 \\
&\leq 2\mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right] + \frac{2\gamma}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right] \\
&= \frac{2}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right].
\end{aligned}$$

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

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

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.

Solution: This result appears as a corollary of Lemma 2 in [Abel et al., 2019], where Pinsker's inequality is used to express the result in terms of the expected KL-divergence between the two policies instead of the total variation distance.
Leveraging the hint and the previous part, we see that

$$\begin{aligned}
\mathbb{E}_{s_0 \sim \beta} \left[V^\pi(s_0) - V^{\pi'}(s_0) \right] &= \mathcal{R}^\top \chi^\pi - \mathcal{R}^\top \chi^{\pi'} \\
&= \mathcal{R}^\top \left(\chi^\pi - \chi^{\pi'} \right) \\
&\leq |\mathcal{R}^\top \left(\chi^\pi - \chi^{\pi'} \right)| \\
&\leq \underbrace{||\mathcal{R}||_\infty}_{=R_{\text{MAX}}} ||\chi^\pi - \chi^{\pi'}||_1 \\
&\leq \frac{2R_{\text{MAX}}}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[D_{\text{TV}} \left(\pi(\cdot \mid s) \,||\, \pi'(\cdot \mid s) \right) \right].
\end{aligned}$$

References

David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L Littman, and Lawson LS Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3134–3142, 2019.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.