6.7950 Fall 2022: - Recitation 8 Handout

1 Performance difference lemma

1.1 Context

In lecture, we discussed a bound for the difference in value when following two different policies π and $\tilde{\pi}$. Before we derive this bounds, we first used the performance difference lemma, that was stated as follows

$$V(\tilde{\pi}) - V(\pi) = \sum_{s,a} d^{\tilde{\pi}}(s,a) A^{\pi}(s,a) = \sum_{s} d^{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(s,a) A^{\pi}(s,a)$$

For this recitation, we will discuss some nuances about this result in order to make it better defined in a wider range of situations. It should also serve as practice for manipulating these types of expressions as well as motivation for discussing how different authors define certain terms. Finally, this useful result is going to be proven.

1.2 Improving and clarifying the problem statement

First, let's disambiguate $V(\pi)$ from V^{π}: the first one assumes some distribution μ for s_0, while the second one assume it's value is deterministic, so these terms can be related as

$$V(\pi) = V_{\mu}(\pi) = \mathbb{E}_{s_0 \sim \mu}[V^{\pi}(s_0)] = \mathbb{E}_{s_t, a_t}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

To make the notation expectation more clear, we can define a trajectory τ containing all states and action over time as in $\tau = (s_0, a_0, s_1, a_1, \ldots)$ with the associated probability dependent on the MDP, π and μ as

$$P_{\mu}^{\pi}(\tau) = P_{\mu}^{\pi}(s_0, a_0, s_1, a_1, \ldots) = \mu(s_o)\pi(a_0|s_0)P(s_1|s_0, a_0)\pi(a_1|s_1)$$

We can now more clearly define V (after leaving the dependency on μ implicit) as

$$V(\pi) = \mathbb{E}_{\tau|\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

Note that some author choose to represent $V(\pi)$ as $\eta(\pi)$ to make this distinction even more clear.

Still, the theorem was stated based on the stationary distribution d^{π} for a certain policy π. You may recall that a distribution of states over time under a Markov process may achieve a stationary distribution after many iterations, so for short episodic problems this choice may seem limiting, but it was done for teaching

purposes in order to simplify presentation. In reality, we can replace this notion of stationary distribution with an analogous one that works in a more general settings.

We can define an unnormalized discounted state visitation distribution $d^{\pi}_{s_0}$ after starting at s_0 and following policy π and the equivalent considering the initial state distribution μ as

$$d^{\pi}_{s_0}(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi)$$

$$d^{\pi}_{\mu}(s) = \mathbb{E}_{s_0 \sim \mu}[d^{\pi}_{s_0}(s)] = \sum_{t=0}^{\infty} \gamma^t \sum_{s_0} P(s_t = s | s_0, \pi) \mu(s_0)$$

where we don't use a different notation from the stationary distribution mentioned in lecture because the same results apply.

To solidify our understanding of this formulation, let's solve the problems below

1. Rewrite $V(\pi)$ using d^{π}_{μ}
2. Show that $d\pi_{\mu}$ is not a valid distribution and find the actual distribution $\overline{d}\pi_{\mu}$
3. Rewrite $V(\pi)$ as an expectation over the rewards r using $\overline{d}\pi_{\mu}$. What makes this form potentially appealing and how would you make it even more so?

Solution:

1. The result follows from applications of expectation and the definitions presented

$$\begin{aligned}
V(\pi) &= \mathbb{E}_{\tau|\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\
&= \mathbb{E}_{s_0 \sim \mu}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{s_t} \sum_{a_t} r(s_t, a_t) P(s_t | s_0, \pi) \pi(a_t | s_t)\right] \\
&= \mathbb{E}_{s_0 \sim \mu}\left[\sum_{s_t} \sum_{t=0}^{\infty} \gamma^t P(s_t | s_0, \pi) \sum_{a_t} \pi(a_t | s_t) r(s_t, a_t)\right] \\
&= \mathbb{E}_{s_0 \sim \mu}\left[\sum_{s_t} d^{\pi}_{s_0}(s_t) \sum_{a_t} \pi(a_t | s_t) r(s_t, a_t)\right] \\
&= \sum_{s_t} d^{\pi}_{\mu}(s_t) \sum_{a_t} \pi(a_t | s_t) r(s_t, a_t)
\end{aligned}$$

2. If we add all elements of d^{π}_{μ}, we have that

$$\sum_{s} d^{\pi}_{\mu}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_{s} P(s_t = s | s_0, \pi) \mu(s_0) = \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$$

So we can turn it into a valid probability distribution by multiplying it by $1-\gamma$ as in

$$\overline{d}_\mu^\pi(s) = (1-\gamma) d_\mu^\pi(s)$$

3. Using the results from the previous items

$$\begin{aligned} V(\pi) &= \sum_{s_t} d_\mu^\pi(s_t) \sum_{a_t} \pi(a_t|s_t) r(s_t, a_t) \\ &= \frac{1}{1-\gamma} \sum_{s_t} \overline{d}_\mu^\pi(s_t) \sum_{a_t} \pi(a_t|s_t) r(s_t, a_t) \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s\sim\overline{d}_\mu^\pi} \mathbb{E}_{a\sim\pi}[r(s,a)] \end{aligned}$$

This is an interesting expression because it's similar to the one that appears on undiscounted problems. If we decided to write a "normalized" value function $\overline{V}(\pi) = (1-\gamma)V(\pi)$ then both settings would be even more similar. This is actually a definition that some authors use, so it's important to pay attention to the definitions inside each context.

1.3 Proof

Let's rewrite the performance difference lemma for the policies π and $\tilde{\pi}$ as

$$V_\mu(\tilde{\pi}) - V_\mu(\pi) = \mathbb{E}_{\tau|\tilde{\pi},\mu}\left[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)\right] \tag{1}$$

Note that the dependency on a distribution μ is usually omitted

Task: Prove the performance difference lemma. *Hint:* It might be easier to start on the RHS of the equation

Solution: If we expand the expression for A^π, we will notice a telescopic summation where most terms cancel

$$
\begin{aligned}
\mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t,a_t)\right] &= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)+\gamma V^{\pi}(st+1)-V^{\pi}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)-V^{\pi}(s_0)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]+\mathbb{E}_{s_0\tilde{\mu}}\left[V^{\pi}(s_0)\right] \\
&= V(\tilde{\pi})-V(\pi)
\end{aligned}
$$