Note on Performance Difference Lemma

Wen Sun[1]

[1]Department of Computer Science, Cornell University

March 14, 2021

1 Performance Difference Lemma

Consider an infinite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, P, \gamma\}$. Consider any policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ and two policies, $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ and $\pi' : \mathcal{S} \mapsto \Delta(\mathcal{A})$.

Recall that $\mathbb{P}^\pi_h(s, a; s_0)$ is the probability of π reaching (s, a) at time step h starting from s_0. Denote $\mathbb{P}^\pi_h(s; s_0)$ as the probability of π reaching s at time step h from s_0, i.e., $\mathbb{P}^\pi_h(s; s_0) = \sum_a \mathbb{P}^\pi_h(s, a; s_0)$. Recall that $d^\pi_{s_0}(s, a) := (1-\gamma)\sum_{h=0}^\infty \gamma^h \mathbb{P}^\pi_h(s, a; s_0)$, and we denote $d^\pi_{s_0}(s) = \sum_a d^\pi_{s_0}(s, a)$.

Let us denote $V^\pi(s_0) = \mathbb{E}\left[\sum_{h=0}^\infty \gamma^h r(s_h, a_h)|\pi, P\right]$

Lemma 1 (PDL (Kakade and Langford, 2002))**.** *For any state $s_0 \in \mathcal{S}$, we have*

$$V^\pi(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_{s_0}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)} A^{\pi'}(s, a)\right],$$

where $A^{\pi'}(s, a) = Q^{\pi'}(s, a) - V^{\pi'}(s)$.

Proof.

$$\begin{aligned}
&V^\pi(s_0) - V^{\pi'}(s_0) \\
&= V^\pi(s_0) - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi'}(s')\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi}(s')\right] - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi'}(s')\right] \\
&\quad + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^\pi(s_1) - V^{\pi'}(s_1)\right] \\
&\quad + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma\mathbb{E}_{s'\sim P(s_0,a_0)} V^{\pi'}(s')\right] - V^{\pi'}(s_0) \\
&= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^\pi(s_1) - V^{\pi'}(s_1)\right] + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[Q^{\pi'}(s_0, a_0) - V^{\pi'}(s_0)\right] \\
&= \underbrace{\gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\mathbb{E}_{s_1\sim P(s_0,a_0)}\left[V^\pi(s_1) - V^{\pi'}(s_1)\right]}_{\text{term a}} + \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[A^{\pi'}(s_0, a_0)\right]
\end{aligned}$$

For term a, note that by Markovian property, $\mathbb{P}^{\pi}(s_1; s_0) = \sum_{a_0} \pi(a_0|s_0) P(s_1|s_0, a_0) = \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} P(s_1|s_0, a_0)$, we can apply the same operation (i.e., recursion) for the term $V^{\pi}(s_1) - V^{\pi'}(s_1)$, we have:

$$
\begin{aligned}
&\text{term a} = \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi}(\cdot; s_0)} \left[V^{\pi}(s_1) - V^{\pi'}(s_1) \right] \\
&= \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi}(\cdot; s_0)} \left[\gamma \mathbb{E}_{a_1 \sim \pi(\cdot|s_1)} \mathbb{E}_{s_2 \sim P(s_1, a_1)} \left[V^{\pi}(s_2) - V^{\pi'}(s_2) \right] + \mathbb{E}_{a_1 \sim \pi(\cdot|s_1)} A^{\pi'}(s_1, a_1) \right] \\
&= \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi}(\cdot; s_0)} \left[V^{\pi}(s_2) - V^{\pi'}(s_2) \right] + \gamma \mathbb{E}_{s_1, a_1 \sim \mathbb{P}_1^{\pi}(\cdot, \cdot; s_0)} A^{\pi'}(s_1, a_1)
\end{aligned}
$$

where last step we use the Markovian property again, i.e., $\mathbb{P}_2^{\pi}(s; s_0) = \sum_{s_1, a_1} \mathbb{P}_1^{\pi}(s_1; s_0) \pi(a_1|s_1) P(s|s_1, a_1)$. Note that at this stage, we can apply the same operation (i.e., recursion) to the term $V^{\pi}(s_2) - V^{\pi'}(s_2)$.

Now combine the above derivations, and repeat the recursion step for infinitely many times, we get:

$$
\begin{aligned}
&V^{\pi}(s_0) - V^{\pi'}(s_0) \\
&= \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} A^{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s_1, a_1 \sim \mathbb{P}_1^{\pi}(\cdot, \cdot; s_0)} A^{\pi'}(s_1, a_1) + \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi}(\cdot; s_0)} [V^{\pi}(s_2) - V^{\pi'}(s_2)] \\
&\ldots \\
&= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s, a \sim \mathbb{P}_h^{\pi}(\cdot, \cdot; s_0)} A^{\pi'}(s, a) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s, a \sim d_{s_0}^{\pi}} A^{\pi'}(s, a).
\end{aligned}
$$

□

Exercise: Derive a similar result for finite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, H, s_0\}$. We should have something like:

$$
V_0^{\pi}(s_0) - V_0^{\pi'}(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi}(\cdot, \cdot; s_0)} A_h^{\pi'}(s_h, a_h),
$$

where $A_h^{\pi}(s, a) = Q_h^{\pi}(s, a) - V_h^{\pi}(s)$.

References

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, 2002.