Note on Simulation Lemma

Wen Sun[1]

[1]Department of Computer Science, Cornell University

March 11, 2021

1 Simulation Lemma

Consider two infinite horizon MDP $\widehat{\mathcal{M}} = \{\mathcal{S}, \mathcal{A}, r, \widehat{P}, \gamma\}$ and $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, P, \gamma\}$. Consider any policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ (note that here we consider stochastic policy, i.e., a policy that maps from a state to a distribution over $\mathcal{A}$).

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., $\mathbb{P}^\pi_h(s; s_0) = \sum_a \mathbb{P}^\pi_h(s, a; s_0)$.

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

Lemma 1.

$$\begin{aligned} \left|V^\pi(s_0) - \widehat{V}^\pi(s_0)\right| &\leq \frac{\gamma}{1-\gamma} \mathbb{E}_{s,a\sim d^\pi_{s_0}} \left| \mathbb{E}_{s'\sim P(s,a)} \widehat{V}^\pi(s') - \mathbb{E}_{s'\sim \widehat{P}(s,a)} \widehat{V}^\pi(s') \right| \\ &\leq \frac{\gamma}{(1-\gamma)^2} \mathbb{E}_{s,a\sim d^\pi_{s_0}} \left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1 \end{aligned}$$

Proof. Using Bellman equation for $\widehat{V}^\pi$ and V^π, we have:

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

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), we have:

$$\begin{aligned} &\text{term a} = \gamma \mathbb{E}_{s_1\sim\mathbb{P}^\pi_1(\cdot;s_0)} \left[\widehat{V}^\pi(s_1) - V^\pi(s_1) \right] \\ &= \gamma \mathbb{E}_{s_1\sim\mathbb{P}^\pi_1(\cdot;s_0)} \left[\gamma \mathbb{E}_{a_1\sim\pi(\cdot|s_1)} [\mathbb{E}_{s'\sim\widehat{P}(s_1,a_1)} \widehat{V}^\pi(s') - \mathbb{E}_{s'\sim P(s_1,a_1)} \widehat{V}^\pi(s')] + \gamma \mathbb{E}_{a_1\sim\pi(\cdot|s_1)} [\mathbb{E}_{s'\sim P(s_1,a_1)} \widehat{V}^\pi(s') - \mathbb{E}_{s'\sim P(s_1,a_1)} V^\pi(s')] \right] \\ &= \gamma^2 \mathbb{E}_{s_1,a_1\sim\mathbb{P}^\pi_1(\cdot,\cdot;s_0)} \left[\mathbb{E}_{s'\sim\widehat{P}(s_1,a_1)} \widehat{V}^\pi(s') - \mathbb{E}_{s'\sim P(s_1,a_1)} \widehat{V}^\pi(s') \right] \\ &\qquad + \gamma^2 \mathbb{E}_{s_2\sim\mathbb{P}^\pi_2(\cdot;s_0)} \left[\widehat{V}^\pi(s_2) - V^\pi(s_2) \right] \end{aligned}$$

where last step we use the Markovian property again, i.e., $\mathbb{P}^\pi_2(s; s_0) = \sum_{s_1,a_1} \mathbb{P}^\pi_1(s_1, a_1; s_0) P(s|s_1, a_1)$.

Now combine the above derivations, we get:

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

Lastly, let us add absolute value on both sides, we get:

$$
\begin{aligned}
\left| \widehat{V}^\pi(s_0) - V^\pi(s_0) \right| &= \frac{\gamma}{1-\gamma} \left| \mathbb{E}_{s,a \sim d_{s_0}^\pi} \left[\mathbb{E}_{s' \sim \widehat{P}(s,a)} \widehat{V}^\pi(s') - \mathbb{E}_{s' \sim P(s,a)} \widehat{V}^\pi(s') \right] \right| \\
&\leq \frac{\gamma}{1-\gamma} \mathbb{E}_{s,a \sim d_{s_0}^\pi} \left| \mathbb{E}_{s' \sim \widehat{P}(s,a)} \widehat{V}^\pi(s') - \mathbb{E}_{s' \sim P(s,a)} \widehat{V}^\pi(s') \right|
\end{aligned}
$$

The second part of the lemma uses the fact that $\widehat{V}^\pi(s) \in [0, 1/(1-\gamma)]$ for all s (since our reward $r \in [0,1]$), and the inequality that:

$$
\left| \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim Q} f(x) \right| \leq \sup_x |f(x)| \, \|P - Q\|_1
$$

for any distributions P and Q and any function $f(x)$.

□

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:

$$
\left| \widehat{V}_0^\pi(s_0) - V_0^\pi(s_0) \right| \leq \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h \sim \mathbb{P}_h^\pi(\cdot,\cdot;s_0)} \left| \mathbb{E}_{s' \sim P(s_h,a_h)} \widehat{V}_{h+1}^\pi(s') - \mathbb{E}_{s' \sim \widehat{P}(s_h,a_h)} \widehat{V}_{h+1}^\pi(s') \right|
$$

where we should understand $\widehat{V}_H^\pi(s) = 0$ (as the episode ends at $H-1$), i.e., we can abstract this additional step H where we do not have any reward.