A Concise Proof of TRPO

Ziniu Li

18/12/2019

1 Notation

- $\{\mathcal{S}, \mathcal{A}, P(s'|s,a), \rho_0(s), \gamma\}$: state space, action space, system state transition matrix, initial state distribution, and discount factor.
- A_π: advantage function of policy π.
- $\rho_\pi(s)$: state occupancy of policy π. $\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$.
- $P_\pi(s'|s)$: state transition matrix based on π. $P_\pi(s'|s) = \sum_a \pi(a|s) P(s'|s,a)$.
- $\eta(\pi)$: the performance of π. $\eta(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{\rho_0,\pi}[\gamma^t r_t|\pi]$.

(Trust Region Policy Optimization[1]) Let $\epsilon = \max_{s,a} |A_\pi(s,a)|$, $\alpha = \max_s D_{TV}\big(\tilde{\pi}(\cdot|s), \pi(\cdot|s)\big)$.

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{1}$$

where,

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

2 Proof

$$\begin{aligned}
\eta_\pi(\tilde{\pi}) &= \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a) && \text{(Policy Difference Lemma)} \\
&= \eta(\pi) + \sum_s \big(\rho_\pi(s) + \rho_{\tilde{\pi}}(s) - \rho_\pi(s)\big) \sum_a \tilde{\pi}(a|s) A_\pi(s,a) && \text{(Permutation)} \\
&= L_\pi(\tilde{\pi}) + \sum_s (\rho_{\tilde{\pi}}(s) - \rho_\pi(s)) \sum_a \tilde{\pi}(a|s) A_\pi(s,a) && \text{(Definition of } L_\pi(\pi)\text{)} \\
&= L_\pi(\tilde{\pi}) + \sum_s (\rho_{\tilde{\pi}}(s) - \rho_\pi(s)) \sum_a (\tilde{\pi}(a|s) - \pi(a|s)) A_\pi(s,a) && (\sum_s \pi(a|s) A_\pi(s,a) = 0) \\
&\geq L_\pi(\tilde{\pi}) - \sum_s |\rho_{\tilde{\pi}}(s) - \rho_\pi(s)| \sum_a |\tilde{\pi}(a|s) - \pi(a|s)| \max_{s,a} |A_\pi(s,a)| && \text{(Triangle inequality)} \\
&= L_\pi(\tilde{\pi}) - 2\sum_s |\rho_{\tilde{\pi}}(s) - \rho_\pi(s)| D_{TV}(\tilde{\pi}, \pi)\epsilon && \text{(Definition of } D_{TV} \text{ and } \epsilon\text{)} \\
&\geq L_\pi(\tilde{\pi}) - 2\sum_s |\rho_{\tilde{\pi}}(s) - \rho_\pi(s)| \max_s D_{KL}(\tilde{\pi}, \pi)\epsilon && \text{(Maximum over state space)} \\
&= L_\pi(\tilde{\pi}) - 4\alpha\epsilon D_{TV}(\rho_{\tilde{\pi}}, \rho_\pi) && \text{(Definition of } D_{TV}\text{)}
\end{aligned}$$

In the following, we analyze the state-distribution discrepancy based on the Lemma 1 in the framework of value discrepancy analysis in [2].

Let $M_{\tilde{\pi}} = (\mathcal{I} - \gamma P_{\tilde{\pi}})^{-1}$ and $M_\pi = (\mathcal{I} - \gamma P_\pi)^{-1}$.

$$\begin{aligned} \rho_{\tilde{\pi}} - \rho_\pi &= [(\mathcal{I} - \gamma P_{\tilde{\pi}})^{-1} - (\mathcal{I} - \gamma P_\pi)^{-1}]\rho_0 \\ &= (M_{\tilde{\pi}} - M_\pi)\rho_0 \end{aligned} \tag{2}$$

Let's dive into the term $M_{\tilde{\pi}} - M_\pi$

$$\begin{aligned} M_{\tilde{\pi}} - M_\pi &= M_{\tilde{\pi}}(M_\pi^{-1} - M_{\tilde{\pi}}^{-1})M_\pi \\ &= \gamma M_{\tilde{\pi}}(P_{\tilde{\pi}} - P_\pi)M_\pi \end{aligned} \tag{3}$$

Thus, back to the Equation (2):

$$\rho_{\tilde{\pi}} - \rho_\pi = \gamma M_{\tilde{\pi}}(P_{\tilde{\pi}} - P_\pi)M_\pi\rho_0 \tag{4}$$

It is easy to show $M_{\tilde{\pi}}$ and M_π is bounded.

$$\begin{aligned} M_{\tilde{\pi}} &= ||\sum_{t=0}^{\infty} \gamma^t P_{\tilde{\pi}}^t|| \leq \sum_{t=0}^{\infty} \gamma^t ||P_{\tilde{\pi}}||_1^t \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma} \\ M_\pi &= ||\sum_{t=0}^{\infty} \gamma^t P_\pi^t|| \leq \sum_{t=0}^{\infty} \gamma^t ||P_\pi||_1^t \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma} \end{aligned} \tag{5}$$

Now, we can show that $||\rho_{\tilde{\pi}} - \rho_\pi||_1$ is bounded.

$$\begin{aligned} ||\rho_{\tilde{\pi}} - \rho_\pi||_1 &\leq \frac{\gamma}{(1-\gamma)^2} \sum_{s,s'} |P_{\tilde{\pi}}(s'|s) - P_\pi(s'|s)|\rho_0(s) \\ &\leq \frac{\gamma}{(1-\gamma)^2} \sum_{s,s'} \rho_0(s) \big| \sum_a P(s'|s,a)\big(\tilde{\pi}(a|s) - \pi(a|s)\big)\big| \\ &\leq \frac{\gamma}{(1-\gamma)^2} \sum_{s,s'} \rho_0(s) \sum_a P(s'|s,a)\big|\tilde{\pi}(a|s) - \pi(a|s)\big| \\ &\leq \frac{\gamma}{(1-\gamma)^2} \sum_{s} \rho_0(s) \sum_a \big|\tilde{\pi}(a|s) - \pi(a|s)\big| \\ &\leq \frac{2\gamma}{(1-\gamma)^2} \sum_{s} \rho_0(s) D_{TV}(\tilde{\pi}, \pi) \\ &\leq \frac{2\gamma}{(1-\gamma)^2}\alpha \end{aligned} \tag{6}$$

Back to our target,

$$\eta_\pi(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{7}$$

References

[1] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1889–1897, 2015.

[2] Tian Xu, Ziniu Li, and Yang Yu. On value discrepancy of imitation learning. *arXiv:1911.07027*, 2019.