A Few Notes on Trust Region Policy Optimization

David Meyer

dmm@{1-4-5.net,uoregon.edu,...}

Last update: April 26, 2018

1 Introduction

One of the goals of this note is to prove Theorem 1 of Schulman, J. et al., "Trust Region Policy Optimization [1]. This is the proof of the *Policy Improvement Bound* described in [1], The proof begins with a lemma from Kakade & Langford [2] that shows that the difference in policy performance $\eta(\tilde{\pi}) - \eta(\pi)$ can be decomposed as a sum of per-timestep *advantages*.

First, define $\eta(\pi)$ as follows. Let π denote a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, and let $\eta(\pi)$ be the expected discounted reward under π. Then define $\eta(\pi)$:

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, s_1, a_1, \cdots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right] \text{ where}$$
$$s_0 \sim \rho(s_0), a_t \sim \pi(a_t \mid s_t), s_{t+1} \sim P(s_{t+1} \mid s_t, a_t)$$

Lemma 1.1 *Given two policies π and $\tilde{\pi}$*

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t) \right]$$

where the expectation is taken over trajectories $\tau \coloneqq (s_0, a_0, s_1, a_1, \cdots)$ and the notation $\mathbb{E}_{\tau \sim \tilde{\pi}}\big[\cdots\big]$ means that actions are sampled from $\tilde{\pi}$ to generate τ.

Proof. First, recall that the advantage $A_\pi(s, a)$ of an action a in state s is defined as follows:

$$A_\pi(s, a) = \mathbb{E}_{s' \sim P(s' \mid s, a)} \big[r(s) + \gamma V_\pi(s') - V_\pi(s) \big]$$

From here we can just work out the result:

$$
\begin{aligned}
\mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] &= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\left[\gamma^t(r(s_t)+\gamma V_\pi(s_{t+1})-V_\pi(s_t))\right]\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\left[\gamma^t r(s_t)+\gamma^{t+1}V_\pi(s_{t+1})-\gamma^t V_\pi(s_t)\right]\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)+\sum_{t=0}^{\infty}\gamma^{t+1}V_\pi(s_{t+1})-\sum_{t=0}^{\infty}\gamma^t V_\pi(s_t)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)+\sum_{t=1}^{\infty}\gamma^{t}V_\pi(s_{t})-\sum_{t=0}^{\infty}\gamma^t V_\pi(s_t)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)+\sum_{t=1}^{\infty}\gamma^{t}V_\pi(s_{t})-\left(V_\pi(s_0)+\sum_{t=1}^{\infty}\gamma^t V_\pi(s_t)\right)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)+\left(\sum_{t=1}^{\infty}\gamma^{t}V_\pi(s_{t})-\sum_{t=1}^{\infty}\gamma^t V_\pi(s_t)\right)-V_\pi(s_0)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)-V_\pi(s_0)\right] \\
&= -\mathbb{E}_{s_0}\left[V_\pi(s_0)+\mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]\right] \\
&= -\eta(\pi)+\eta(\tilde{\pi}) && \text{\# Definition } \eta(\pi)
\end{aligned}
$$

1.1 A Bit of Intuition

Another way to see this result: we can see that $\gamma V_\pi(s_{t+1}) - V_\pi(s_t) = -V_\pi(s_0)$ by expanding the first few terms:

Table 1: Expansion of terms

t	$\gamma^t(r(s_t)+\gamma V_\pi(s_{t+1})-V_\pi(s_t))$
0	$r(s_0)+\color{red}{\gamma V_\pi(s_1)}-V_\pi(s_0)$
1	$\gamma r(s_1)+\color{red}{\gamma^2 V_\pi(s_2)}-\color{blue}{\gamma V_\pi(s_1)}$
2	$\gamma^2 r(s_2)+\color{red}{\gamma^3 V_\pi(s_3)}-\color{blue}{\gamma^2 V_\pi(s_2)}$
3	$\gamma^3 r(s_3)+\color{red}{\gamma^4 V_\pi(s_4)}-\color{blue}{\gamma^3 V_\pi(s_3)}$
$\vdots$	$\vdots$
∞	$\sum_{t=0}^{\infty}\gamma^t r(s_t)-V_\pi(s_0)$

Notice that in Table 1 the red term at time t minus the blue term at time $t+1$ equals zero. For example, at times $t = 0$ and $t = 1$ we have $\gamma V_\pi(s_1) - \gamma V_\pi(s_1) = 0$. More generally, at time t we have the term $\gamma^t V_\pi(s_t)$ and at time $t+1$ we have the term $\gamma^t V_\pi(s_t)$ whose difference is again zero. So in the limit we are left with $\sum_{t=0}^{\infty} \gamma^t r(s_t) - V_\pi(s_0)$.

Armed with this result we can now define the expected advantage of $\tilde{\pi}$ over π at state s as $\bar{A}(s) = \mathbb{E}_{a \sim \tilde{\pi}(\cdot|s)}\Big[A_\pi(s,a)\Big]$. Now Lemma 1.1 can be written as follows:

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{A}_\pi(s_t, a_t)\right]$$

and that $L_\pi(\tilde{\pi})$ can be written as

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{A}_\pi(s_t, a_t)\right]$$

The difference between $\eta(\tilde{\pi})$ and $L_\pi(\tilde{\pi})$ is whether the states are sampled using π or $\tilde{\pi}$.

To bound the difference between $\eta(\tilde{\pi})$ and $L_\pi(\tilde{\pi})$ we need to bound the difference arising at each timestep. To do this, we first need to introduce a measure of how much π and $\tilde{\pi}$ agree. The approach taken in this paper is to *couple* the policies so that they define a joint distribution over pairs of actions.

References

[1] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," *CoRR*, vol. abs/1502.05477, 2015.

[2] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)* (C. Sammut and A. Hoffman, eds.), (San Francisco, CA, USA), pp. 267–274, Morgan Kauffman, 2002.