Notes on TRPO

Yue Lin

December 31, 2024

1 Trust Region Policy Optimization

The objective is

$$J(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \cdot r\left(s_t, a_t\right) \right]. \tag{1}$$

The problem is $\max_\pi J(\pi)$. We want to develop an iterative method to solve this problem.

Lemma 1.1. *The performance difference between two policies is*

$$\begin{aligned} J\left(\pi'\right) &= J(\pi) + \mathbb{E}_{\pi'} \left[\sum_{t=0}^{\infty} \gamma^t \cdot A_\pi\left(s_t, a_t\right) \right] \\ &= J(\pi) + \sum_s d_{\pi'}(s) \sum_a \pi'(a \mid s) \cdot A_\pi(s, a), \end{aligned} \tag{2}$$

where $d_\pi(s)$ is the discounted state visitation frequencies and A_π is the advantage function under policy π.

Proof. Let $\operatorname{Pr}^\pi\left(\tau \mid s_0 = s\right)$ denote the probability of observing a trajectory τ when starting in state s and following the policy π. Using a telescoping argument, we have:

$$
\begin{aligned}
& V^{\pi}(s)-V^{\pi'}(s) \\
=& \mathbb{E}_{\tau \sim \operatorname{Pr}^{\pi}(\tau \mid s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right)\right]-V^{\pi'}(s) \\
=& \mathbb{E}_{\tau \sim \operatorname{Pr}^{\pi}(\tau \mid s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t\left(r\left(s_t, a_t\right)+V^{\pi'}\left(s_t\right)-V^{\pi'}\left(s_t\right)\right)\right]-V^{\pi'}(s) \\
\stackrel{(a)}{=}& \mathbb{E}_{\tau \sim \operatorname{Pr}^{\pi}(\tau \mid s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t\left(r\left(s_t, a_t\right)+\gamma V^{\pi'}\left(s_{t+1}\right)-V^{\pi'}\left(s_t\right)\right)\right] \\
\stackrel{(b)}{=}& \mathbb{E}_{\tau \sim \operatorname{Pr}^{\pi}(\tau \mid s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t\left(r\left(s_t, a_t\right)+\gamma \mathbb{E}\left[V^{\pi'}\left(s_{t+1}\right) \mid s_t, a_t\right]-V^{\pi'}\left(s_t\right)\right)\right] \\
\stackrel{(c)}{=}& \mathbb{E}_{\tau \sim \operatorname{Pr}^{\pi}(\tau \mid s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}\left(s_t, a_t\right)\right] \\
=& \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^{\pi}} \mathbb{E}_{a \sim \pi(\cdot \mid s')}\left[A^{\pi'}\left(s', a\right)\right],
\end{aligned}
$$

where (a) rearranges terms in the summation and cancels the $V^{\pi'}(s_0)$ term with the $-V^{\pi'}(s)$ outside the summation, and (b) uses the tower property of conditional expectations and the final equality follows from the definition of d_s^{π}.
$\square$

- If we can find a π' such that $\sum_a \pi'(a \mid s) \cdot A_{\pi}(s,a) \geq 0$ for all s, then update π by π' will make the objective larger or remain unchanged.
- The classic method, exact policy iteration, chooses $\pi'(s) = \arg\max_a A_{\pi}(s,a)$, so it can improve the policy in each iteration or at least not make it worse.
- But due to the unavoidable estimation error, the exact policy iteration may choose the suboptimal action so the true at some state s such that $\sum_a \pi'(a \mid s) \cdot A_{\pi}(s,a) < 0$.

Consider the following fuction:

$$
L_{\pi}(\pi') = J(\pi) + \sum_s d_{\pi}(s) \sum_a \pi'(a \mid s) \cdot A_{\pi}(s,a). \tag{3}
$$

Lemma 1.2. *$L_{\pi}(\pi')$ matches $J(\pi)$ to the first order:*

- $L_{\pi_0}(\pi_0) = J(\pi_0)$,
- $\nabla_{\theta} L_{\pi_{\theta_0}}(\pi_{\theta})|_{\theta=\theta_0} = \nabla_{\theta} J(\pi_{\theta})|_{\theta=\theta_0}$.

Theorem 1.3. *The following bound holds:*

$$J(\pi') \geq L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2, \tag{4}$$

where $\epsilon = \max_{s,a} |A_\pi(s,a)|$*,* $\alpha = D_{\mathrm{TV}}^{\max}(\pi, \pi')$ *and* $D_{\mathrm{TV}}(p\|q) = \frac{1}{2}\sum_i |p_i - q_i|$ *is the total variation divergence.*

Corollary 1.4. *From 1.3 we know that the following bound holds:*

$$J(\pi') \geq L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} \cdot D_{\mathrm{KL}}^{\max}(\pi, \pi'). \tag{5}$$

Assume that we have exact evaluation of A_π. Then we can have the following algorithm.

Algorithm 1 Policy iteration algorithm guaranteeing non-decreasing expected return η

1: Initialize π_0
2: **for** $i = 0, 1, 2, \ldots$ until convergence **do**
3: Compute all advantage values $A_{\pi_i}(s,a)$
4: Solve the constrained optimization problem:

$$\pi_{i+1} = \arg\max_\pi \left[L_{\pi_i}(\pi) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \cdot D_{\mathrm{KL}}^{\max}(\pi_i, \pi). \right]$$

5: **end for**

This is a minorization-maximization (MM) algorithm. The surrogate function is $M_i(\pi) = L_{\pi_i}(\pi) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \cdot D_{\mathrm{KL}}^{\max}(\pi_i, \pi)$ that minorizes J with equality at π_i.

Corollary 1.5. *Algorithm 1 generates a monotonically improving sequence of policies* $J(\pi_i) \leq J(\pi_j)$ *where* $i < j$*.*

Proof. We can see this by M_i:

- $\pi_{i+1} = \arg\max_\pi M_i(\pi_i)$, so $M_i(\pi_{i+1}) \geq M_i(\pi)$.
- $J(\pi_{i+1}) \geq M_i(\pi_{i+1})$ by (5), and $J(\pi_i) = M_i(\pi_i)$ because the divergence is 0.
- So $J(\pi_{i+1}) - J(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i) \geq 0$.

□

If π is parameterized by θ, then we know that by solving the following optimization problem, the objective J is guaranteed to be improved:

$$\max_\theta \left[L_{\pi_{\theta_0}}(\pi_\theta) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \cdot D_{\mathrm{KL}}^{\max}(\pi_{\theta_0}, \pi_\theta) \right] \tag{6}$$

In practice, if we use the coefficient $\frac{4\epsilon\gamma}{(1-\gamma)^2}$, the step sizes will be very small. (Why?) A robust approach to taking larger steps is to impose a **trust region constraint**, which limits the KL divergence between the new policy and the old policy:

$$\begin{aligned} \max_{\theta} \quad & L_{\pi_{\theta_0}}(\pi_\theta) \\ \text{s.t.} \quad & D_{\mathrm{KL}}^{\max}(\pi_{\theta_0}, \pi_\theta) \leq \delta. \end{aligned} \tag{7}$$

The constraints are too many, so we use a **heuristic approximation** which considers the average divergence:

$$\begin{aligned} \max_{\theta} \quad & L_{\pi_{\theta_0}}(\pi_\theta) = J(\pi_{\theta_0}) + \sum_s d_{\pi_{\theta_0}}(s) \sum_a \pi_\theta(a \mid s) \cdot A_{\pi_{\theta_0}}(s, a) \\ \text{s.t.} \quad & \mathbb{E}_{s \sim d_{\pi_{\theta_0}}} \left[D_{\mathrm{KL}}\left((\pi_{\theta_0}(\cdot \mid s) \| \pi_\theta(\cdot \mid s) \right) \right] \leq \delta. \end{aligned} \tag{8}$$

That is,

$$\begin{aligned} \max_{\theta} \quad & \sum_s d_{\pi_{\theta_0}}(s) \sum_a \pi_\theta(a \mid s) \cdot A_{\pi_{\theta_0}}(s, a) \\ \text{s.t.} \quad & \mathbb{E}_{s \sim d_{\pi_{\theta_0}}} \left[D_{\mathrm{KL}}\left((\pi_{\theta_0}(\cdot \mid s) \| \pi_\theta(\cdot \mid s) \right) \right] \leq \delta. \end{aligned} \tag{9}$$

Use importance sampling, then we have the trust region policy optimization problem:

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_{s \sim d_{\pi_{\theta_0}}, a \sim q} \left[\frac{\pi_\theta(a \mid s)}{q(a \mid s)} A_{\theta_0}(s, a) \right] \\ \text{s.t.} \quad & \mathbb{E}_{s \sim d_{\pi_{\theta_0}}} \left[D_{\mathrm{KL}}\left((\pi_{\theta_0}(\cdot \mid s) \| \pi_\theta(\cdot \mid s) \right) \right] \leq \delta. \end{aligned} \tag{10}$$