Robust Policy Gradient *

Tianbing Xu (Baidu Research), Qiang Liu (UT, Austin)

September 2018

1 Formulation of Robust Policy Gradient

We have two robots, Attacker and Defender. p_{π_θ} is the distribution of the trajectory under Defender's policy π_θ.

$$p_{\pi_\theta}(\tau) = p(s_0) \prod_t \pi_\theta(a_t|s_t) T(s_{t+1}|s_t, a_t) \tag{1}$$

where the trajectory $\tau = \{s_t, a_t, r_t, s_{t+1}\}_{t=0}^T$ and T is the transition model.

Given distribution p_{π_θ}, Attacker is to perturb the trajectory, end up with the trajectory distribution q, such that $\mathbb{KL}(q||p_{\pi_\theta}) \leq \epsilon$. The expected reward is $\mathbb{E}_{\tau\sim q}[R(\tau)]$. If we wan to be robust with p_{π_θ}, we can frame the problem into

$$\max_{\pi_\theta} \min_q \big\{\mathbb{E}_{\tau\sim q}[R(\tau)] + \frac{1}{\alpha}\mathbb{KL}(q \mid\mid p_{\pi_\theta})\big\}. \tag{2}$$

which is equivalent to

$$\min_{\pi_\theta} J_\alpha(\pi_\theta) = \log \mathbb{E}_{p_{\pi_\theta}}[\exp(-\alpha R(\tau))]. \tag{3}$$

Taking the gradient gives

$$\nabla_\theta J_\alpha(\pi_\theta) = \mathbb{E}_{q^*_{\pi_\theta}}\big[\sum_{t=1}^T \nabla_\theta \log \pi_\theta(a_t|s_t)\big]. \tag{4}$$

where $q^*_{\pi_\theta}(\tau) \propto \exp(-\alpha R(\tau)) p_{\pi_\theta}(\tau)$, which gives larger weights on these trajectories with lower rewards.

Given the old policy parameter in the last iteration t, this policy gradient could be approximated as,

$$\nabla_\theta J_\alpha(\pi_\theta) \approx \mathbb{E}_{q^*_{\pi_{\theta_t}}}\big[\sum_{t=1}^T \nabla_\theta \log \pi_\theta(a_t|s_t)\big].$$

*Working in Process

Or, equivalently,

$$\nabla_\theta J_\alpha(\pi_\theta) = \frac{\mathbb{E}_{p_{\pi_\theta}}[\nabla_\theta \log p_{\pi_\theta}(\tau)\exp(-\alpha R(\tau))]}{\mathbb{E}_{p_{\pi_\theta}}[\exp(-\alpha R(\tau))]} \tag{5}$$

Therefore, it is possible to approximate the gradient and have policy update

$$\theta_{t+1} \leftarrow \theta_t - \eta \frac{\sum_{\tau_i \sim p_{\pi_\theta}} \left[\exp(-\alpha R(\tau^i)) \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i | s_t^i)\right]}{\sum_{\tau_i \sim p_{\pi_\theta}} \exp(-\alpha R(\tau^i))}.$$

To stabilize the policy update, we use trust region to bound the step size similar to TRPO,

$$\mathbb{KL}(\pi_{\theta_t} || \pi_{\theta_{t+1}}) \leq \delta$$

Thus, the step size is bounded as,

$$\eta \leq \sqrt{\frac{2\delta}{(\nabla_\theta J_\alpha(\theta))^T H^{-1}(\theta_t)(\nabla_\theta J_\alpha(\theta))}}$$

where $H(\theta)$ is the Fisher information matrix.

Algorithm 1 Robust Policy Gradient

1: initialize policy parameter θ_0 randomly
2: **for** iteration $t = 0$ to T **do**
3: Generate trajectories from π_{θ_t}.
4: Calculate the policy gradient,

$$\nabla_\theta J_\alpha(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_t}}[w(\tau)\nabla_\theta \log p_{\pi_\theta}(\tau)]$$

where $w(\tau) = \exp(-\alpha R(\tau))$.
5: Update the policy parameter,

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_\theta J_\alpha(\pi_\theta)$$

6: Adaptive adjust the learning rate to stabilize the policy,

$$\eta \leftarrow \sqrt{\frac{2\delta}{(\nabla_\theta J_\alpha(\theta))^T H^{-1}(\theta_t)(\nabla_\theta J_\alpha(\theta))}}$$

7: **end for**

2 Connection to minimization of the Tail Probability

Let R_π be the random reward from trajectory τ generated by policy π with the interaction of MDP. Assuming we ignore the constant factor of the transition probability, it is a random variable whose distribution depends on policy π. Typical methods are interested in maximizing the expected reward $J_0(\pi) := \mathbb{E}_\pi[R(\tau)]$; this method, however, does not capture the uncertainty on rewards. In this work, we are interested in robust methods that minimize the tail probability $p(R_\pi \leq \epsilon)$, for some small number $\epsilon > 0$.

This is difficult to do directly, so we instead consider the inequality:

$$P(R_\pi \leq \epsilon) \leq \frac{\mathbb{E}_{\tau\sim\pi}[\exp(-\alpha R(\tau))]}{\exp(-\alpha\epsilon)}. \tag{6}$$

which holds for any α. The best α should be chosen to minimize the upper bound, that is,

$$\alpha^* = \arg\min_\alpha \big\{ \log \mathbb{E}_\tau[\exp(-\alpha R(\tau))] + \alpha\epsilon \big\}. \tag{7}$$

This means there is a one-to-one correspondence between ϵ and the optimal α^*.

$$\min_\pi \left\{ J_\alpha(\pi) := \frac{1}{\alpha} \log \mathbb{E}_{\tau\sim\pi}[\exp(-\alpha R(\tau))] \right\}. \tag{8}$$

which is similar to the robust policy gradient objective (3). Furthermore, it recovers special cases. If $\alpha \to 0$, it approaches to the typical average reward $J_0(\pi)$:

$$\lim_{\alpha\to 0^+} J_\alpha(\pi) = -\mathbb{E}_{\tau\sim\pi}[R(\tau)].$$

If $\alpha \to \infty$, it is like a **MiniMax** problem (which is not unfortunately not well defined):

$$\lim_{\alpha\to+\infty} J_\alpha(\pi) = -\min R_\pi.$$

In practice, we should α to be a sufficiently small number so that the exponential does not explode.

3 Derivations

3.1 InEquality

Given

$$f(x) = \exp(-\alpha x), \alpha > 0, f(x) > 0$$

, then

$$P(X \leq \epsilon) \leq \frac{\mathbb{E}[\exp(-\alpha X)]}{\exp(-\alpha\epsilon)} \tag{9}$$

That is,

$$\begin{aligned} P(X \leq \epsilon) = P(f(X) \geq f(\epsilon)) = \mathbb{E}[1(f(X) \geq f(\epsilon))] \\ \leq \mathbb{E}[\frac{f(X)}{f(\epsilon)}] = \frac{\mathbb{E}[f(X)]}{f(\epsilon)} \end{aligned}$$

3.2 Optimal policy distribution

$$q^*_{\pi_\theta}(\tau) \leftarrow argmin_q \left\{ \mathbb{E}_q[R(\tau)] + \frac{1}{\alpha}\mathbb{KL}(q||p_{\pi_\theta}) \right\} \tag{10}$$

Define

$$q^*_{\pi_\theta}(\tau) = \frac{1}{\exp(J_\alpha(\theta))} p_{\pi_\theta}(\tau) * \exp(-\alpha R(\tau)),$$

where $\exp(J_\alpha(\theta))$ serves as the normalization constant, and

$$\begin{aligned} J_\alpha(\theta) = \log \int p_{\pi_\theta}(\tau)\exp(-\alpha R(\tau))d\tau \\ = \log \mathbb{E}_{p_{\pi_\theta}}[\exp(-\alpha R(\tau))] \end{aligned}$$

$$\begin{aligned} \mathbb{E}_q[R(\tau)] + \frac{1}{\alpha}\mathbb{KL}(q||p_{\pi_\theta}) \\ = -\frac{1}{\alpha}\mathbb{E}_q[\log\exp(-\alpha R(\tau))] + \frac{1}{\alpha}\mathbb{E}_q[\log q - \log p_{\pi_\theta}(\tau)] \\ = \frac{1}{\alpha}\mathbb{E}_q[\log q - \log p_{\pi_\theta} - \log\exp(-\alpha R(\tau))] \\ = \frac{1}{\alpha}\mathbb{E}_q[\log q - \log q^*_{\pi_\theta} - J_\alpha(\pi)] \\ = \frac{1}{\alpha}\mathbb{KL}(q||q^*_{\pi_\theta}) - \frac{1}{\alpha}\mathbb{E}_q[J_\alpha(\pi_\theta)] \\ = \frac{1}{\alpha}\mathbb{KL}(q||q^*_{\pi_\theta}) - J_\alpha(\pi_\theta). \end{aligned}$$

3.3 Robust Policy Gradient

Objective function,

$$\min_\theta J_\alpha(\pi_\theta) = \log \mathbb{E}_{p_{\pi_\theta}}[\exp(-\alpha R(\tau))]$$

The policy gradient,

$$
\begin{aligned}
\nabla_\theta J_\alpha(\pi_\theta) = \nabla_\theta \log \int p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau)) d\tau \\
= \int \frac{\nabla_\theta p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau))}{\int p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau)) d\tau} \\
= \int \frac{p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau))}{\int p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau)) d\tau} \nabla_\theta \log p_{\pi_\theta}(\tau) \\
= \mathbb{E}_{q^*_{\pi_\theta}}[\log p_{\pi_\theta}(\tau)] = \mathbb{E}_{q^*_{\pi_\theta}}[\sum_t \log \pi_{\pi_\theta}(a_t|s_t)]
\end{aligned}
$$

Another equivalent formula for Policy Gradient Estimation,

$$
\nabla_\theta J_\alpha(\pi_\theta) = \frac{\mathbb{E}_{p_{\pi_\theta}}[\nabla_\theta \log p_{\pi_\theta}(\tau) \exp(-\alpha R(\tau))]}{\mathbb{E}_{p_{\pi_\theta}}[\exp(-\alpha R(\tau))]}
$$

References

[1] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, Shie Mannor Policy Gradient for Coherent Risk Measures *NIPS 2015*

[2] Yinlam Chow, Aviv Tamar, Shie Mannor, Marco Pavone Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach *NIPS 2015*

[3] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, Pieter Abbeel Trust Region Policy Optimization *ICML 2015*