Soft Learning

Peng Lingwei

September 23, 2019

Contents

1 Trust Region Policy Optimization

1.1 Basics

1. Initial state: $\rho_0 : S \to \mathbb{R}$;
2. Total reward: $\eta(\pi) = \mathbb{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$;
3. $Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1},a_{t+1},\ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$;
4. $V_\pi(s_t) = \mathbb{E}_{a_t,s_{t+1},\ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$;
5. $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$;
6. $\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right]$

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

7. $\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \cdots$;

$$\begin{aligned}
\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t,a_t)\right] &= \sum_{t=0}^{\infty} \sum_s P(s_t = s|\tilde{\pi}) \sum_a \tilde{\pi}(a|s) \gamma^t A_\pi(s,a) \\
&= \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\tilde{\pi}) \sum_a \tilde{\pi}(a|s) A_\pi(s,a) \\
&= \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)
\end{aligned}$$

8. Hard: $\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$ and
Easy: $L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$;

$$\nabla_\theta L_\pi(\pi_\theta) = \sum_s \rho_\pi(s) \sum_a A_\pi(s,a) \nabla_\theta \pi_\theta(a|s)$$

$$\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)|_{\theta=\theta_0} = \sum_s \rho_{\pi_{\theta_0}}(s) \sum_a A_{\pi_{\theta_0}}(s,a) \nabla_\theta \pi_\theta(a|s)|_{\theta=\theta_0}$$

Because policy gradient theorem:

$$\nabla_\theta \eta(\pi_\theta) = \sum_s \rho_{\pi_\theta}(s) \sum_a Q_{\pi_\theta}(s,a) \nabla_\theta \pi(a|s) = \sum_s \rho_{\pi_\theta}(s) \sum_a A_{\pi_\theta}(s,a) \nabla_\theta \pi(a|s)$$

Therefore $\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_0}$. We also have $L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta(\pi_{\theta_0})$.

9. Let $\bar{A}(s) = \mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\left[A_\pi(s,a)\right]$.
Bound $|\eta(\tilde{\pi}) - L_\pi(\tilde{\pi})| = \sum_{t=0}^{\infty} \gamma^t \left|\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\bar{A}(s_t)\right] - \mathbb{E}_{\tau\sim\pi}[\bar{A}(s_t)]\right|$.
Let n_t be the number of times that $a_i \neq \tilde{a}_i$ for $i < t$.

$$\begin{cases}
\mathbb{E}_{s_t\sim\tilde{\pi}}[\bar{A}(s_t)] = P(n_t = 0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t=0}[\bar{A}(s_t)] + P(n_t > 0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)] \\
\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)] = P(n_t = 0)\mathbb{E}_{s_t\sim\pi|n_t=0}[\bar{A}(s_t)] + P(n_t > 0)\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)] \\
\mathbb{E}_{s_t\sim\tilde{\pi}|n_t=0}[\bar{A}(s_t)] = \mathbb{E}_{s_t\sim\pi|n_t=0}[\bar{A}(s_t)]
\end{cases}$$

$$\begin{aligned}\left|\mathbb{E}_{s_t\sim\tilde{\pi}}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| =&\left|P(n_t>0)\{\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)]-\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]\}\right| \\ \leq& 2\left[1-(1-\alpha)^t\right]\max_s\left[\bar{A}(s)\right]\end{aligned}$$

$$\begin{aligned}\bar{A}(s) =& \mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\left[A_\pi(s,a)\right]=\mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})}\left[A_\pi(s,\tilde{a})-A_\pi(s,a)\right] \quad since \quad \mathbb{E}_{a\sim\pi}\left[A_\pi(s,a)\right]=0 \\ =& P(a\neq\tilde{a}|s)\mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})|a\neq\tilde{a}}\left[A_\pi(s,\tilde{a})-A_\pi(s,a)\right]\end{aligned}$$

Definition 1. *$(\pi,\tilde{\pi})$ is $\alpha-$coupled policy pair if $(a,\tilde{a})\sim(\pi,\tilde{\pi})(s)\Rightarrow P(a\neq\tilde{a}|s)\leq\alpha$.*

α-coupled policy $(\pi,\tilde{\pi})\Rightarrow\bar{A}(s)\leq 2\alpha\max_{s,a}|A_\pi(s,a)|\Rightarrow$

$$|\eta(\tilde{\pi})-L_\pi(\tilde{\pi})|\leq\sum_{t=0}^{\infty}\gamma^t\cdot 4\alpha\left[1-(1-\alpha)^t\right]\max_{s,a}|A_\pi(s,a)|=\frac{4\alpha^2\gamma}{(1-\gamma)(1-\gamma(1-\alpha))}\max_{s,a}|A_\pi(s,a)|$$

$$|\eta(\tilde{\pi})-L_\pi(\tilde{\pi})|\leq\frac{4\alpha^2\gamma}{(1-\gamma)^2}\max_{s,a}|A_\pi(s,a)|$$

10. $\pi'=\arg\max_{\pi'}L_{\pi_{old}}(\pi')$ and $\pi_{new}(a|s)=(1-\alpha)\pi_{old}(a|s)+\alpha\pi'(a|s)$, then π_{new} and π_{old} are α-coupled.

11. Define total variation divergence $D_{TV}(p\|q)=\frac{1}{2}\sum_i|p(i)-q(i)|$, If $D_{TV}(p\|q)\leq\alpha$, and $(i,j)\sim(p,q)$, then $P(i=j)\geq 1-\alpha$. (No proof.)

12. $D_{TV}^2(p\|q)\leq D_{KL}(p\|q)\leq\alpha$ (no proof). Define $D_{KL}^{\max}=\max_s D_{KL}(\pi(s)\|\tilde{\pi}(s))$.Let $\tilde{\pi}=\arg\max_{\tilde{\pi}}L_\pi(\tilde{\pi})-\frac{4\epsilon\gamma}{(1-\gamma)^2}D_{KL}^{\max}(\pi,\tilde{\pi})$, then

$$\eta(\tilde{\pi})\geq L_\pi(\tilde{\pi})-\frac{4\epsilon\gamma}{(1-\gamma)^2}D_{KL}^{\max}(\pi,\tilde{\pi})\geq L_\pi(\pi)=\eta(\pi)$$

13. $\max_\theta L_{\theta_{old}}(\theta)-CD_{KL}^{\max}(\theta_{old},\theta)$. If C is given by the theory, the step sizes would be very small. So, we use a const constraint instead

$$\max L_{\theta_{old}}(\theta),\quad s.t.\quad D_{KL}^{\max}(\theta_{old},\theta)\leq\delta$$

The problem is still hard to solve, so we go further. Define $\bar{D}_{KL}^{\rho}(\theta_1,\theta_2)=\mathbb{E}_{s\sim\rho}\left[D_{KL}(\pi_{\theta_1(s)}\|\pi_{\theta_2}(s))\right]$

$$\max L_{\theta_{old}}(\theta),\quad s.t.\quad \bar{D}_{KL}^{\rho}(\theta_{old},\theta)\leq\delta$$

14. Sample-Based estimation:

$$\max_\theta\mathbb{E}_{s\sim\rho_{old},a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)}Q_{\theta_{old}}(s,a)\right],\quad s.t.\quad \mathbb{E}_{s\sim\rho_{old}}\left[D_{KL}(\pi_{\theta_{old}}(s)\|\pi_\theta(s))\right]\leq\delta$$