Paper Reading 2018-10-27

TRPO

PPO

公式预警

MDP: $\left\langle S, A, P, r, \gamma, \rho_0 \right\rangle$

S : is the finite set of states

A : is the finite set of actions

$P : S \times A \times S \to R$: is the transition probability distribution

$r : S \to R$: is the reward function

$\rho_0 : S \to R$: is the distribution of the initial state s_0

$$Q_\pi(s_t, a_t) = E_{s_{t+1}, a_{t+1}, \ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$$

$$V_\pi(s_t) = E_{a_t, s_{t+1}, a_{t+1}, \ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$$

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s_t)$$

$$\eta(\pi) = E_{s_0, a_0, \ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t)]$$

Trust Region Policy Optimization(TRPO)---Review: REINFORCE

$$\nabla\eta(\theta) = E_\pi[\gamma^t Q_\pi(S_t, A_t)\frac{\nabla\pi(A_t \mid S_t,\theta)}{\pi(A_t \mid S_t,\theta)}]$$

$$= E_\pi[\gamma^t G_t \frac{\nabla\pi(A_t \mid S_t,\theta)}{\pi(A_t \mid S_t,\theta)}]$$

$$\theta' = \theta + \alpha\gamma^t G_t \frac{\nabla\pi(A_t \mid S_t,\theta)}{\pi(A_t \mid S_t,\theta)}$$

$$\theta' = \theta + \alpha\gamma^t G_t \nabla\log\pi(A_t \mid S_t,\theta)$$

With Baseline:

$$\sum_a b(s)\nabla_\theta\pi(a \mid s,\theta) = 0$$

$$\nabla\eta(\theta) \propto \sum_s d_\pi(s)\sum_a (Q_\pi(s,a) - b(s))\nabla\pi(a \mid s,\theta)$$

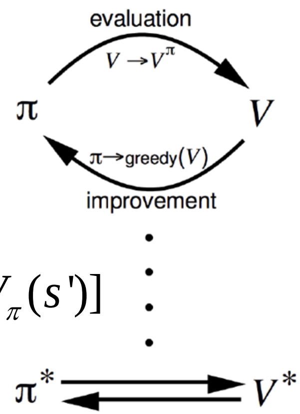
$$\theta' = \theta + \alpha\gamma^t (G_t - b(S_t))\nabla\log\pi(A_t \mid S_t,\theta)$$

$b(S_t) \approx V(S_t)$ can reduce the variance.

Trust Region Policy Optimization(TRPO)---Review: Dynamic Programing

- Policy Iteration:

$$V_\pi(s) \le Q_\pi(s,\pi'(s)) \le V_{\pi'}(s)$$



$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)[r+\gamma V_\pi(s')]$$

$$\pi'(s) \doteq \arg\max_a Q_\pi(s,a)$$

$$= \max_a \sum_{s',r} P(s',r|s,a)[r+\gamma V_\pi(s')]$$

- Value Iteration:

$$V_*(s) \doteq \max_\pi V_\pi(s) = \max_a \sum_{s',r} P(s',r|s,a)[r+\gamma V_*(s')]$$

$$\pi'(s) = \max_a \sum_{s',r} P(s',r|s,a)[r+\gamma V_*(s')]$$

$$\begin{aligned}
&V_\pi(s) \le Q_\pi(s,\pi'(s)) \\
&= E_{\pi'}[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s] \\
&\le E_{\pi'}[R_{t+1} + \gamma Q_\pi(S_{t+1},\pi'(s)) \mid S_t = s] \\
&= E_{\pi'}[R_{t+1} + \gamma E_{\pi'}[R_{t+2} + \gamma V_\pi(S_{t+2})] \mid S_t = s] \\
&= E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2})] \mid S_t = s] \\
&\ldots \\
&\le E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots] \mid S_t = s] \\
&= V_{\pi'}(s)
\end{aligned}$$

Trust Region Policy Optimization(TRPO)

Policy Improvement :

$$\eta(\pi) \geq \eta(\pi)$$

So how to find policy π ?

$$\eta(\pi) = \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} P(s_t = s \mid \pi) \sum_{a} \pi(a \mid s) \gamma^t A_\pi(s,a)$$

$$= \eta(\pi) + \sum_{s} \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi) \sum_{a} \pi(a \mid s) A_\pi(s,a)$$

$$= \eta(\pi) + \sum_{s} \rho_\pi(s) \sum_{a} \pi(a \mid s) A_\pi(s,a)$$

Policy Improvement :

$$\eta(\pi) \geq \eta(\pi)$$

Define:

$$L_\pi(\pi) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \pi(a \mid s) A_\pi(s,a)$$

We have:

$$L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta(\pi_{\theta_0})$$

$$\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)\big|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)\big|_{\theta=\theta_0}$$

A sufficiently small step that improves $L_{\pi_{\theta_0}}(\pi_{\theta_0})$ will also improve η

Trust Region Policy Optimization(TRPO)

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

Notice:

$$\pi' = \arg\max_{\pi'} L_{\pi_{old}}(\pi')$$
$$\pi_{new}(a|s)=(1-\alpha)\pi_{old}(a|s)+\alpha\pi'$$
$$\varepsilon = \max_s |E_{a\sim\pi'(a|s)}[A_{\pi}(s,a)]|$$

$$\eta(\pi_{new}) \ge L_{\pi_{old}}(\pi_{new})-\frac{2\varepsilon\gamma}{(1-\gamma)^2}\alpha^2$$

Then:

$$D_{TV}(p\|q)=\frac{1}{2}\sum_i |p_i-q_i|$$
$$D_{TV}^{\max}(\pi,\pi)=\max_s D_{TV}(\pi(\cdot|s)\|\pi(\cdot|s))$$
$$\alpha = D_{TV}^{\max}(\pi_{old},\pi_{new})$$
$$\varepsilon = \max_{s,a}|A_{\pi}(s,a)|$$

$$\eta(\pi_{new}) \ge L_{\pi_{old}}(\pi_{new})-\frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2$$

Trust Region Policy Optimization(TRPO)

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

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

Notice:

$$\pi' = \arg\max_{\pi'} L_{\pi_{old}}(\pi')$$

$$\pi_{new}(a|s) = (1-\alpha)\pi_{old}(a|s) + \alpha\pi'$$

$$\varepsilon = \max_s |E_{a \sim \pi'(a|s)}[A_\pi(s,a)]|$$

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{2\varepsilon\gamma}{(1-\gamma)^2}\alpha^2$$

Then:

$$D_{TV}(p \| q) = \frac{1}{2}\sum_i |p_i - q_i|$$

$$D_{TV}^{\max}(\pi, \pi) = \max_s D_{TV}(\pi(\cdot|s) \| \pi(\cdot|s))$$

$$\alpha = D_{TV}^{\max}(\pi_{old}, \pi_{new})$$

$$\varepsilon = \max_{s,a} |A_\pi(s,a)|$$

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2$$

Trust Region Policy Optimization(TRPO)

$$D_{TV}(p \| q)^2 \le D_{KL}(p \| q)$$

Define:

$$D_{KL}^{\max}(\pi, \pi) = \max_s D_{KL}(\pi(\cdot|s) \| \pi(\cdot|s))$$

$$\eta(\pi_{new}) \ge L_{\pi_{old}}(\pi_{new}) - \frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2 \qquad \eta(\pi_{new}) \ge L_{\pi_{old}}(\pi_{new}) - CD_{KL}^{\max}(\pi_{old}, \pi_{new})$$

Define:

$$M_i(\pi) = L_{\pi_i}(\pi) - CD_{KL}^{\max}(\pi_i, \pi)$$

We have:

$$\eta(\pi_{i+1}) \ge M_i(\pi_{i+1})$$

$$\eta(\pi_i) = M_i(\pi_i)$$

$$\eta(\pi_{i+1}) - \eta(\pi_i) \ge M_i(\pi_{i+1}) - M_i(\pi_i)$$

So when:

$$\pi_{i+1} = \arg\max_\pi M_i(\pi)$$

$$\eta(\pi_{i+1}) - \eta(\pi_i) \ge M_i(\pi_{i+1}) - M_i(\pi_i) \ge 0$$

Trust Region Policy Optimization(TRPO)

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

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

$$\pi_{i+1} = \arg\max_{\pi} \left[L_{\pi_i}(\pi) - C D_{\mathrm{KL}}^{\max}(\pi_i, \pi) \right]$$

where $C = 4\epsilon\gamma/(1-\gamma)^2$

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

end for

Trust Region Policy Optimization(TRPO)

$$\pi_{i+1} = \arg\max_\pi M_i(\pi)$$
$$= \arg\max_\pi [L_{\pi_i}(\pi) - CD_{KL}^{\max}(\pi_i, \pi)]$$
$$= \arg\max_\pi [\eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s,a) - CD_{KL}^{\max}(\pi_i, \pi)]$$
$$= \arg\max_\pi [\sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s,a) - CD_{KL}^{\max}(\pi_i, \pi)]$$

Trust Region Constraint:

$$\max imize_\pi \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s,a)$$

Subject to $\quad D_{KL}^{\max}(\pi_i, \pi) \le \delta$

Approximation:

$$\max imize_\theta \sum_s \rho_{\theta_{old}}(s) \sum_a \pi_\theta(a|s) A_{\theta_{old}}(s,a)$$

Subject to $\quad \overline{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \le \delta$

$$\overline{D}_{KL}^{\rho}(\theta_1, \theta_2) \triangleq E_{s \sim \rho}[D_{KL}(\pi_{\theta_1}(\cdot|s) \| \pi_{\theta_2}(\cdot|s))]$$

Trust Region Policy Optimization(TRPO)

$$\sum_a \pi_\theta(a|s)A_{\theta_{old}}(s,a) = E_{a \sim q}[\frac{\pi_\theta(a|s_n)}{q(a|s_n)} A_{\theta_{old}}(s_n,a)]$$

$$\max imize_\theta E_{s \sim \rho_{\theta_{old}}, a \sim q}[\frac{\pi_\theta(a|s_n)}{q(a|s_n)} Q_{\theta_{old}}(s_n,a)]$$

Subject to $\quad E_{s \sim \rho_{\theta_{old}}}[D_{KL}(\pi_{\theta_{old}}(\cdot|s) \| \pi_\theta(\cdot|s))] \le \delta$

Look Back: optimal step size, trust region

Proximal Policy Optimization Algorithms(PPO)

Policy Gradient Methods:

Gradient Estimator:

$$g = E_t[\nabla_\theta \log \pi_\theta(a_t | s_t) A_t]$$

Objective:

$$L^{PG} = E_t[\log \pi_\theta(a_t | s_t) A_t]$$

Trust Region Methods:

$$\max imize_\theta E_t[\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} A_t]$$

Subject to $\quad E_t[D_{KL}(\pi_{\theta_{old}}(\cdot | s) \| \pi_\theta(\cdot | s))] \leq \delta$

Objective:

$$\max imize_\theta E_t[\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} A_t] - \beta D_{KL}(\pi_{\theta_{old}}(\cdot | s) \| \pi_\theta(\cdot | s))]$$

Proximal Policy Optimization Algorithms(PPO)

Define: $$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$$

TRPO Objective: $$L^{CPI} = E_t[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} A_t] = E_t[r_t(\theta) A_t]$$

PPO Objective: $$L^{CLIP} = E_t[\min(r_t(\theta) A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon) A_t)]$$

谢谢！