TRPO/PPO

Policy gradients with value functions as baselines.

$$J(\theta) = \mathop{E}_{\tau \sim p_\theta(\tau)} \left[\sum_t \gamma^t r(s_t, a_t) \right]$$

$$\nabla_\theta J(\theta) \propto E_{s \sim p_\theta, a \sim \pi_\theta} \left[A^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \right]$$

Policy Iteration Perspective

$$J(\theta') - J(\theta)$$

$$= J(\theta') - E_{s_0 \sim p(s_0)} \left[V^{\pi_\theta}(s_0) \right]$$

$$= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} \left[V^{\pi_\theta}(s_0) \right]$$

$$= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}(s_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}(s_t) \right]$$

$$= J(\theta') + \mathop{E}_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t \left(\gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t) \right) \right]$$

$$= \mathop{E}_{\tau \sim p_{\theta'}(\tau)} \left\{ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right\}$$

$$= \mathop{E}_{\tau \sim p_{\theta'}(\tau)} \left\{ \sum_{t=0}^{\infty} \gamma^t \left[r(s_t, a_t) + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t) \right] \right\}$$

$$= \mathop{E}_{\tau \sim p_{\theta'}(\tau)} \left[\gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

$$J(\theta') - J(\theta) = \mathop{E}_{\tau \sim p_{\theta'}(\tau)} \left[\gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

Think about

$$\max_{\theta'} J(\theta') - J(\theta)$$

Key Idea in PPO / TRPO:

Use data from $p_\theta(\tau)$ to approximate $J(\theta') - J(\theta)$

Importance Sampling

$$E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

$$= \sum_t E_{s_t \sim p_{\theta'}(s_t)} \left[E_{a_t \sim \pi_{\theta'}(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

$$= \sum_t E_{s_t \sim p_{\theta'}(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta} \right]$$

$$\approx \sum_t E_{s_t \sim p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} A^{\pi_\theta} \right]$$

as long as $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \leq$

$$= \mathop{E}_{s_t \sim p_\theta(s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

TRPO Objective

such that

$$D_{KL}(\pi_{\theta'}(a_t|s_t), \pi_\theta(a_t|s_t)) \leq \epsilon.$$

PPO

$$r_t(\theta') = \frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)}$$

$$L_t^{CLIP}(\theta') = \min\left(r_t(\theta') \gamma^t A^{\pi_\theta}(s_t, a_t), \right.$$

$$\left. \text{clip}\left(r_t(\theta'), 1-\epsilon, 1+\epsilon\right) \gamma^t A^{\pi_\theta}_{(s_t, a_t)} \right)$$