Lecture 7: Advanced PG

So far, we saw:

- Policy gradient methods
  (policy, maybe an on-policy V(s))
  on-policy
- Q-learning methods
  (no policy, just a value fn)
  off-policy

Can we build a hybrid of these methods?

Should be possible!

Why?

How can we get started in a principled manner??

Performance Difference Lemma

For any two policies $\pi(a|s)$ & $\mu(a|s)$, the following holds:

$$J(\pi) - J(\mu) = \mathbb{E}_{\tau \sim P_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^\mu(s_t, a_t) \right]$$

$J(\pi)$: return of policy π

Proof:

$$J(\pi) = \mathbb{E}_{\tau \sim P_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

$$= \mathbb{E}_{\tau \sim P_\pi(\tau)} \left[V^\pi(s_0) \right] = \mathbb{E}_{\tau \sim P_\mu(\tau)} \left[V^\pi(s_0) \right]$$

$$= \mathbb{E}_{\tau \sim P_\mu(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t V^\pi(s_t) - \sum_{t=1}^{\infty} \gamma^t V^\pi(s_t) \right]$$

$$= \mathbb{E}_{\tau \sim P_\mu(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t \left(V^\pi(s_t) - \gamma V^\pi(s_{t+1}) \right) \right]$$

What does the performance difference lemma give us?

$$A^{\mu} \iff \pi$$

Policy iteration

$$\max_{\pi} J(\pi) = \max_{\pi} \left(J(\pi) - J(\mu) \right)$$

$$= \max_{\pi} \mathbb{E}_{s \sim d^{\pi}} \left[A^{\mu}(s, a) \right]$$

Why is this useful?

A practical off-policy policy gradient algorithm

$$J(\pi) - J(\mu) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^\mu(s_t, a_t) \right]$$

$J(\mu)$: data collection policy

$A^\mu(s_t, a_t)$: estimated via a critic

Option 1: use states from μ, A^π instead

Option 2: use states from π, A^μ instead

Which one should we use?

An approximation

$$\max_{\pi} \; J(\pi) - J(\mu)$$

$$\mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^\mu(s_t, a_t) \right]$$

$$= \sum_{t=0}^{\infty} \mathbb{E}_{s_t \sim p_\pi(s_t)} \left[\gamma^t A^\mu(s_t, a_t) \right]$$

$a_t \sim \pi(\cdot|s_t)$

ignore

utilize

$$\mathbb{E}_{s_t \sim p_\mu(s_t)} \left[\mathbb{E}_{a_t \sim \pi(\cdot|s_t)} \left[\gamma^t A^\mu(s_t, a_t) \right] \right]$$

only term involved in maximization

Different ways to solve this maximization when π & μ are close!!

Generic optimization problem:

$$\max_{\pi} \quad \mathbb{E}_{s \sim d^{\mu}(s)} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\mu}(s,a) \right] \right] - \alpha D(\pi, \mu)$$

what is this, precisely?

We want $P_{\pi}(s_t)$ & $P_{\mu}(s_t)$ to be close!

$P_{\pi}(s_t) - P_{\mu}(s_t) = ??$

say π & μ are deterministic & agree with probability $1-\varepsilon$, then:

$$P_{\pi}(s_t) = (1-\varepsilon)^t \, P_{\mu}(s_t) + \left(1 - (1-\varepsilon)^t\right) P_{diff}(s_t)$$

$$\Rightarrow \left| P_{\pi}(s_t) - P_{\mu}(s_t) \right| = \left(1 - (1-\varepsilon)^t\right) \left| P_{\mu}(s_t) - P_{diff}(s_t) \right|$$

$$\leq 2 \cdot \varepsilon \cdot t$$

This means that:

$$D(\pi, \mu) = D_{TV}(\pi(\cdot|s), \mu(\cdot|s))$$

total variations distance.

$$D_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}$$

&

$$D_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(Q \| P)}$$

} Pinsker's inequality

Each of these choices gives a different algorithm:

① AWR → $D_{KL}(\pi \| \mu)$

② PPO / TRPO → $D_{KL}(\mu \| \pi)$

Advantage-weighted regression (AWR)

[Peng, Kumar, Li, Levine, 2019]

$$\max_{\pi} \; \mathbb{E}_{s \sim d^{\mu}(s)} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\mu}(s,a) \right] - \alpha D_{KL}(\pi \| \mu) \right]$$

Easier problem:

$$\max_{p(x)} \; \mathbb{E}_{x \sim p(x)} \left[f(x) \right] - \alpha D_{KL}(p \| q)$$

$$\Rightarrow \quad p^*(x) \propto q(x) \exp\left(\frac{f(x)}{\alpha} \right)$$

So, $\pi^*(a|s) \propto \mu(a|s) \exp\left[\frac{A^{\mu}(s,a)}{\alpha} \right]$

How do I use this oracle solution??