Lecture 8: PPO, DDPG / TD3

So far:

1) Perf. difference lemma:

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

$$= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^\pi(s) \\ a \sim \pi(\cdot|s)}} \left[A^\mu(s,a) \right]$$

$\max_\pi \left(J(\pi) - J(\mu) \right)$

2) Advantage weighted regression (AWR):

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

$d^\mu(s)$

Approximate $d^\mu(s)$ by → this is ok if π & μ are close in action dist. ⇒

Use $D_{KL}(\pi(\cdot|s) \| \mu(\cdot|s))$ on states from μ

Gives us a nice closed form:

$$\pi^*(\cdot|s) \propto \mu(\cdot|s)\, e^{A(s,\cdot)/\alpha}$$

PPO (proximal policy optimization)

$d^{\pi} \rightarrow d^{\mu}$

$$\max_{\theta} \; \mathbb{E}_{s \sim d^{\mu}(s)} \Big[A^{\mu}(s,a) \Big] - \alpha D_{KL}\big(\mu(\cdot|s) \,\|\, \pi_{\theta}(\cdot|s)\big)$$

$a \sim \pi_{\theta}(\cdot|s)$ → $\mu(a|s)$

what if I want to learn on actions from μ too??

another way to compute this is via IS

$$\mathbb{E}_{x \sim p(x)}\big[f(x)\big] = \mathbb{E}_{x \sim q(x)}\left[\frac{p(x)}{q(x)} f(x)\right]$$

different dist.

$\mu(a|s)$

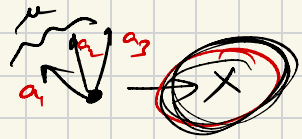
How can we do this for policy grad?

"Importance weighted PG"

$$\nabla_{\theta} \; \mathbb{E}_{\substack{s \sim d^{\mu}(s) \\ a \sim \mu(\cdot|s)}} \left[\frac{\pi_{\theta}(a|s)}{\mu(a|s)} A^{\mu}(s,a)\right] - \alpha D_{KL}\big(\mu(\cdot|s) \,\|\, \pi_{\theta}(\cdot|s)\big)$$

$$\mathbb{E}\left[\frac{d^{\pi}(s)}{d^{\mu}(s)} \cdot A^{\mu}(s,a)\right]$$

$\frac{\pi_\theta(a_1|s)}{\mu(a_1|s)} \cdot A^\mu + \frac{\pi_\theta(a_2|s)}{\mu(a_2|s)} \cdot A^\mu + \dots$



Do I need the KL term?

Answer: No!!

$$\max_\theta \left[\mathbb{E}_{\substack{s\sim d^\mu \\ a\sim\mu}} \left[\frac{\pi_\theta(a|s)}{\mu(a|s)} \cdot A^\mu \right] - \alpha D(\mu \| \pi_\theta) \right]$$

- Can I just throw out the KL?

Not really as the update will be too large!

$\Rightarrow$ π_θ is close to μ for all s, a $\Rightarrow$ $\frac{\pi_\theta(a|s)}{\mu(a|s)}$ is close to 1

But $D_{KL}(\mu \| \pi)$ effectively keeps π & μ close.

So, we can keep $\frac{\pi_\theta(a|s)}{\mu(a|s)}$ close to 1.

$\varepsilon = 0.1, 0.2$ (hyperparameter)

$$\nabla_\theta L^{CLIP}(\theta) = \mathbb{E}_{\substack{s\sim d^\mu(s) \\ a\sim\mu(\cdot|s)}} \left[\text{clip}\left(\frac{\pi_\theta(a|s)}{\mu(a|s)}, 1-\varepsilon, 1+\varepsilon \right) A^\mu \right]$$

$\frac{\pi_\theta(a|s)}{\mu(a|s)} > 1+\varepsilon$

$\mathbb{E}_{s,a\sim\mu} \left[\frac{\pi_\theta(a|s)}{\mu(a|s)} \cdot A^\mu(s,a) \right]$ $\qquad (1-\varepsilon)\cdot A^\mu$

$s, a \rightarrow \left| \frac{\pi_\theta(a|s)}{\mu(a|s)} < 1-\varepsilon \right. \Rightarrow \text{clip}(\dots) = 1-\varepsilon$

$r(\theta) = \pi_\theta(a|s) / \pi_{old}(a|s)$

Any problems with the clip objective?

$$\Rightarrow L^{CLIP}(\theta) = \hat{\mathbb{E}}\left[clip\left(r(\theta), 1-\varepsilon, 1+\varepsilon \right) \cdot A^{\pi_{old}} \right]$$

① Can make an update even when we get clipped on the lower side!

$$\nabla_\theta L^{new}(\theta) = \hat{\mathbb{E}}\left[\min\left(\underbrace{r(\theta) \cdot A^{\pi_{old}}}_{\text{unclipped}}, \underbrace{clip(r(\theta), 1-\varepsilon, 1+\varepsilon) \cdot A^{\pi_{old}}}_{\text{clipped}} \right) \right]$$

$(1-\varepsilon) \cdot A^{\pi_{old}}$

when $r(\theta) < 1-\varepsilon$ & $A^{\pi_{old}}(s,a) > 0$

action is low prob under π_θ — but the action is a good action

you see some signal.

but when $r(\theta) < 1-\varepsilon_{low}$ & $A^{\pi_{old}}(s,a) < 0$

you don't see signal.

instability

Is this good or bad?

✓

$\Sigma\uparrow$

Likewise if

highly likely under π but they are bad actions

$r(\theta) > 1+\varepsilon$ but $A^M(s,a) < 0$

you see some signal

"unless" terms

but when $r(\theta) > 1+\varepsilon$ but $A^M(s,a) > 0$

there is no signal.

↳ conservative choice

⇒ A^M can be erroneous

⇒ IS are high variance

not in the RL sense

This hurts "exploration"!!

sometimes useful to have ε_{high}, ε_{low}

"asymmetric clip"

$1-\varepsilon_{low}$

$1+\varepsilon_{high}$

used in DAPO (in LLMs)

$\varepsilon_{high} > \varepsilon_{low}$

$J(\pi) - J(\mu) = \mathbb{E}_{\pi}[A^{\mu}]$ $\quad \mathbb{E}_{\mu}[A^{\pi}]$

Option 2 for using the perf. difference lemma

use states from μ, & compute $A^{\pi}(s,a)$

$Q^{\pi}(s,a)$ — can do this as we can learn Q-functions

can be estimated using data from any policy

for any policy π,

$\forall s,a: \quad Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{\substack{s' \sim p(\cdot|s,a) \\ a' \sim \pi(\cdot|s')}}[Q^{\pi}(s',a')]$

"train a Q-function"

We know how to implement this via ideas like value-based methods.

$Q^{\pi}_{\phi}(s,a) \Rightarrow$

→ use target nets

→ soft/hard target updates

→ everything else remains the same

value-based RL

$\pi_{\theta}(\cdot|s)$

$\max_{\theta} \mathbb{E}_{\mu}[Q^{\pi}_{\phi}(s,a)]$

Loss for training the Q-functions:

$r(s,a) + \gamma \max_{a'} Q(s',a')$

$$\mathcal{L}(\phi) = \mathbb{E}_{s,a,r,s' \sim RBL}\left[\left(Q_\phi(s,a) - y(s,a)\right)^2\right]$$

replay buffer

$$y(s,a) = r(s,a) + \gamma \mathbb{E}_{s',a' \sim \pi}\left[Q^{tgt}_{\bar{\phi}}(s',a')\right]$$

How do we compute this?

$$\mathbb{E}_{\substack{s' \sim p(\cdot|s,a) \\ a' \sim \pi(\cdot|s')}}\left[Q_{\bar{\phi}}(s',a')\right] \approx$$

1) sample $a'_0 \sim \pi_\theta(\cdot|s')$

2) query the tgt net. with it $Q_{\bar{\phi}}(s',a')$

$s' \in RB$

$a' \sim \pi_\theta(\cdot|s')$

How do we do the policy update?

"π_θ"