Proximal Policy Optimization (PPO)

Konpat Preechakul
Chulalongkorn University
November 2019

Policy gradient fails

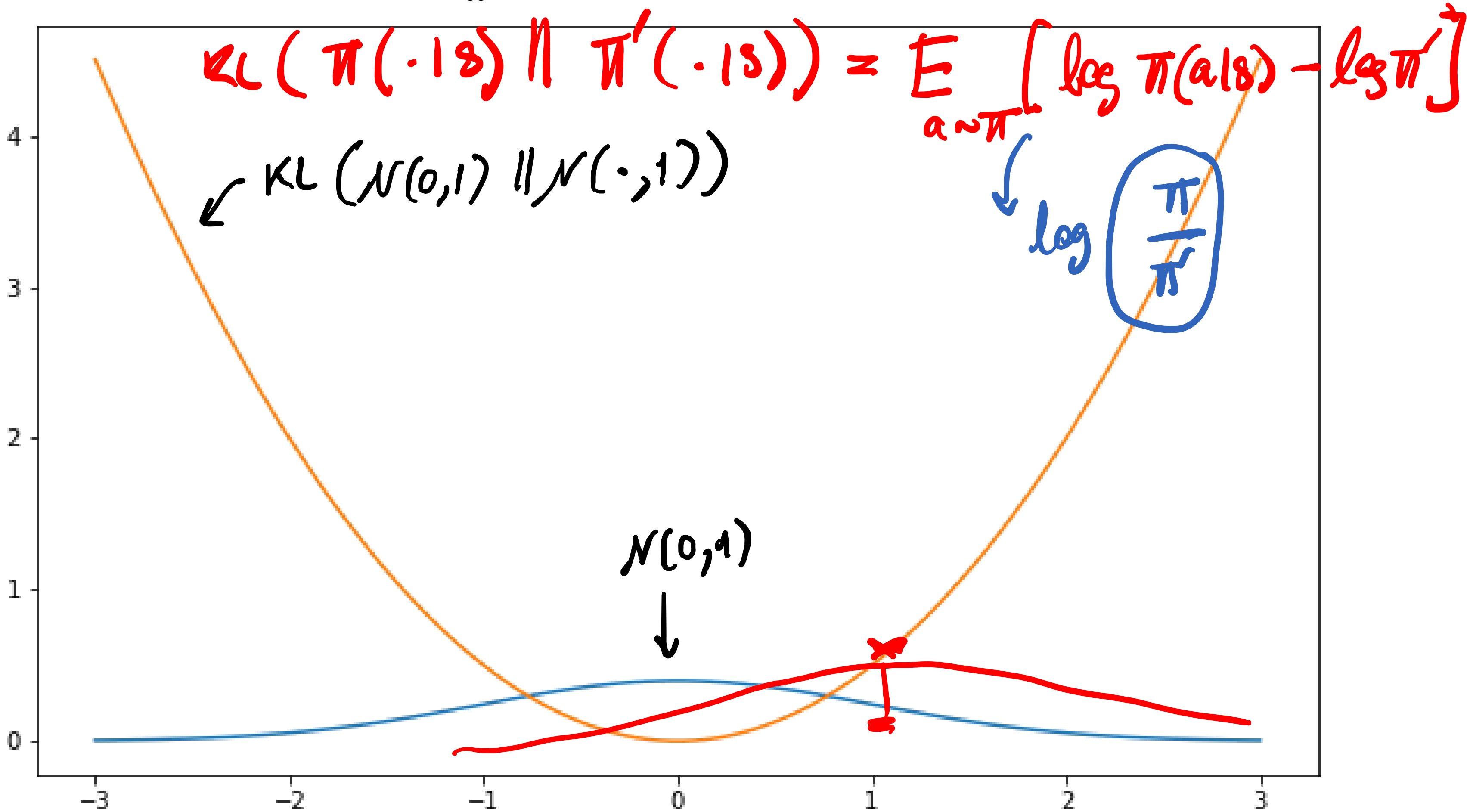
A, Q

- Critic is not oracle, it has its flaws
- Critic is prone to forgetting
- How to know when to trust? confidence
- **How to limit the trust?**
- Not updating too much …
- KL divergence:

$$\mathrm{KL}(P\|Q) = \sum_x P(x)[\log P(x) - \log Q(x)]$$

KL Divergence in picture

$$\mathrm{KL}(P\|Q) = \sum_x P(x)[\log P(x) - \log Q(x)]$$



Trust region policy optimization

$$d^* = \arg\max_d J(\theta + d) \quad \text{s.t. } \mathrm{KL}(\theta \| \theta + d) = c$$

- A constrained optimization

- We relax it using **Lagrangian:**

$$\mathcal{L}(d, \lambda) = J(\theta + d) + \lambda(\mathrm{KL}(\theta \| \theta + d) - c)$$

- Optimal d is at the critical point

$$\nabla_{d,\lambda} \mathcal{L} = 0$$

Policy improvement guarantee

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

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P'(s_t)}\left[\mathcal{A}_\pi(\pi')\right]$$

Lower bound:

$$J(\theta') - J(\theta) = \sum_t \mathbb{E}_{s_t \sim P(s_t)}\left[\mathcal{A}_\pi(\pi')\right] - \sum_t \epsilon t \mathcal{O}\left(\frac{r_{\max}}{1-\gamma}\right)$$

$\approx KL(\pi' \| \pi)$

$$E_{s,a \sim \pi}\left[\boxed{\frac{\pi'(a|s)}{\pi(a|s)}} A^\pi(s,a)\right]$$

$$\text{s.t.} \quad \sqrt{\frac{1}{2}\mathrm{KL}^{\max}(\pi \| \pi')} = \epsilon$$

Constrained optimization is hard

- Using a fixed constant is not likely to work:

$$\overline{J}(\theta + d) = J(\theta + d) - \beta \mathrm{KL}(\theta \| \theta + d)$$

- Lagrangian involves exotic terms like "inverse"

$$\alpha \underbrace{\left(\nabla_\theta^2 \mathrm{KL}\right)^{-1}}_{\text{exotic}} \nabla_\theta J(\theta)$$

Is there an easy way to constrain KL?

Goal

$KL(\theta \| \theta')$

$\underbrace{\theta \quad \theta'}_{\varepsilon} \rightarrow$

$\nabla J = 0$

- Design objective function J
- That has "zero" gradient
- When the constraint is breached

$\underbrace{\theta \quad \theta}_{> \varepsilon}$

∇J $\quad \boxed{\nabla KL}$ at $\theta = 0$

if $KL > \varepsilon \rightarrow \nabla J = 0$

else ∇J

$\nabla J > \varepsilon \rightarrow \theta \; \varepsilon$

- Optimize normally …

θ

$\theta' \rightarrow \theta$

$\xrightarrow[\varepsilon]{} \xrightarrow[\varepsilon]{} \xrightarrow[\varepsilon]{} \xrightarrow[\varepsilon]{} \xrightarrow[\varepsilon]{}$

First attempt

min, max(_ , _)

$\nabla J \leftarrow$

∇J_{KL}

if $\boxed{KL(\theta \| \theta') > \varepsilon :}$

if $\nabla J : \theta \rightarrow \leftarrow \theta' : \nabla J$

else : $\nabla J = 0$

else: ∇J

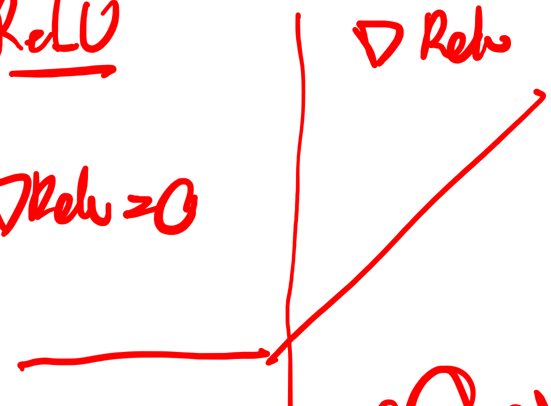
KL

$e^{-k} \nabla J$

ReLU

$\nabla ReLU = 0$

$\nabla ReLU$ 1



max(x, 0)

Possible pseudocode

$\theta' \leftarrow \theta$

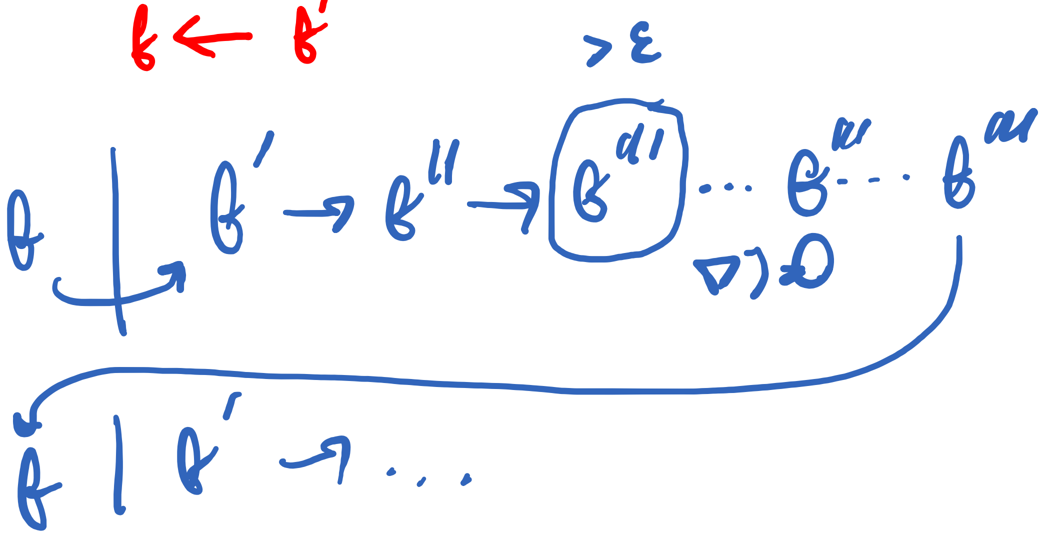
for until satisfied:

  collect data using π for t steps $\overbrace{(s_0, a_0, r_1 \ldots\ldots)}^{t}$

  for i to K:

    $\theta' \leftarrow \theta' + \alpha \nabla J_{KL}(\theta')$

  $\theta \leftarrow \theta'$



Is there a one-liner?

$$J(\theta) = E_{s,a\sim\pi}\left[\frac{\pi'(a|s)}{\pi(a|s)} \cdot A^{\pi}(s,a)\right]$$

$\pi' = \pi$: $\frac{\pi'}{\pi} = 1$

$\pi' \neq \pi$: $\frac{\pi'}{\pi} \neq 1 \rightsquigarrow \boxed{1-\varepsilon < \frac{\pi'}{\pi} < 1+\varepsilon}$

0.2

one liner clip, clamp

$\rightarrow \text{clip}\left(\frac{\pi'}{\pi}, 1-\varepsilon, 1+\varepsilon\right) \cdot A^{\pi}(s,a)$

$$J_{PPO}(\theta) = E_{s,a\sim\pi}\left[\min\left(\frac{\pi'}{\pi} \cdot A^{\pi}, \text{clip}(\ldots) \cdot A^{\pi}\right)\right]$$ PPO

Pseudocode

Algorithm 1 PPO, Actor-Critic Style

for iteration$=1, 2, \ldots$ **do**
  for actor$=1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

explore

$\uparrow \nabla J_{PPO}$

$\underbrace{\theta\ ,}_{\leq \ldots}\ \theta_{old}$

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1707.06347.

Results

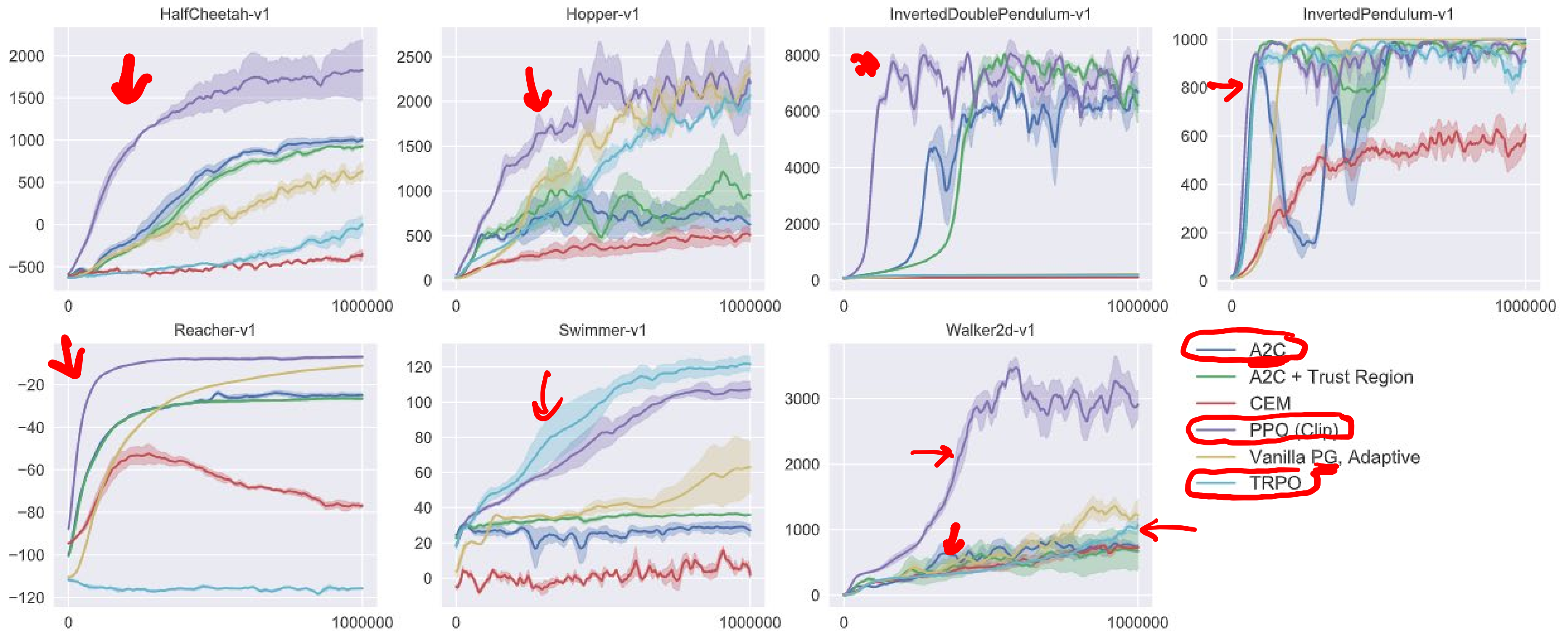


Figure 3: Comparison of several algorithms on several MuJoCo environments, training for one million timesteps.