Proximal Policy Optimization

$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$,
$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \dots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$,
$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$

Policy Evaluation

$V_{i+1}^{\pi_k}(s) = \sum_{s'} P(s'|s, \pi_k(s))[r + \gamma V_i^{\pi_k}(s')]$

Policy Improvement

$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s, a)[r + \gamma V^{\pi_k}(s')]$

n – step return:

$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n \hat{v}(s_{t+n})$

$\nabla J(\theta) = \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(a|s, \theta)}{\pi(a|s, \theta)} \right]$

$\nabla \mathrm{J}(\theta) \propto \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla \pi(a|s, \theta)$

$\nabla J(\theta) = \mathbb{E}_\pi [\nabla_\theta \log \pi_\theta (a_t | s_t) \hat{A}_t]$

Expected discounted reward for a stochastic policy π_θ:

$\eta(\theta) = \mathbb{E}_{s_0, a_0, \dots}[\sum_{t=0}^{\infty} \gamma^t r(s_t)]$

$a_t \sim \pi(a_t|s_t, \theta)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t) \; \forall \; t \geq 0$

State ‘s’ distribution: $\rho(s)$

$\eta(\beta) = \eta(\theta) + \mathbb{E}_{s_0, a_0, \dots \sim \pi_\beta}[\sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t)]$

$\eta(\beta) = \eta(\theta) + \sum_s \rho_\beta(s) \sum_a \pi_\beta(a|s) A_{\pi_\theta}(s, a)$

Complex dependency of $\rho_\beta(s)$on β is difficult to optimize

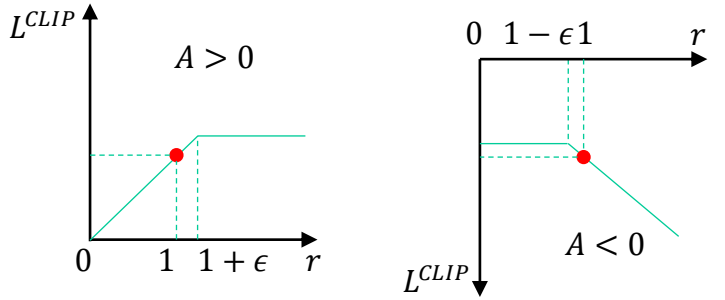
Local approximation to η:

$J_{\pi_\theta}(\beta) = \eta(\theta) + \sum_s \rho_{\pi_\theta}(s) \sum_a \pi_\beta(a|s) A_{\pi_\theta}(s, a)$

Doesn’t specify a step size

- Performance function (at $timestep = t$) optimization problem in policy gradient methods:
  - $\nabla J(\theta) = \mathbb{E}_\pi [\nabla_\theta \ln \pi_\theta (a_t | s_t) \hat{A}_t]$ ← Gradient equation
  - $J(\theta) = \mathbb{E}_\pi [\ln \pi_\theta (a_t | s_t) \hat{A}_t]$ ← Maximize this cost function
  - $\theta_{t+1} = \theta_t + \alpha \hat{A}_t \nabla \ln \pi(a_t|s_t, \theta_t) \Rightarrow \theta_{new} \gg \theta_{old}$
  - How to set step size for update?
- Conservative policy (Kakade & Langford, 2002):
  - $\pi_{new}(a|s) = (1 - \alpha)\pi_{old}(a|s) + \alpha \pi'(a|s)$
- Trust Region (at $timestep = t$) Methods:
  - Maximize, $J(\theta) = \mathbb{E}_\pi \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right]$
  - $s.t. \; \hat{\mathbb{E}} \left[KL[\pi_{\theta_{old}}, \pi_\theta] \right] \leq \delta$
- For more flexibility:
  - Maximize $\mathbb{E}_\pi \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t - \underbrace{C D_{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)]}_{\text{Penalty}} \right]$
  - Difficult to choose a single value of C for different problems. (Experimentally demonstrated)

- Clipped Surrogate Objective
  - $\mathbb{E}_\pi \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t, \mathrm{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right]$



- Adaptive KL Penalty Coefficient (Worse Performance Experimentally)
  - $\mathbb{E}_\pi \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t - C D_{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)] \right]$
  - $d = \hat{\mathbb{E}}_t \left[D_{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)] \right]$
  - If $d < \frac{d_t}{1.5}$, $C \leftarrow \frac{C}{2}$ elseif $d > 1.5 d_t, C \leftarrow 2C$

Algorithm 1: PPO, Actor – Critic Style

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

https://www.profrkumar.co.in/

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).