Policy-Based Reinforcement Learning

- Policy Gradient
- Actor-Critic (Discrete actions)
- A3C (Asynchronous Advantage Actor-Critic)
- TRPO & PPO
- DDPG (Deep Deterministic Policy Gradient)
  - TD3
  - SAC

I-Chen Wu

Trust Region Policy Optimization (TRPO)

- TRPO is a policy optimization algorithm
  - can replace gradient descent
- There are many gradient descent methods
  - Original gradient descent method
  - Natural gradient descent method
  - Stochastic gradient descent method
- TRPO is similar to natural gradient descent method
- TRPO can be combined with A2C, called ACKTR

TRPO

- Consider a Markov decision process (MDP), defined by the tuple

$$(S, A, P, r, \rho_0, \gamma)$$

  - S is a finite set of states, A is finite set of actions
  - $P: S \times A \times S \rightarrow \mathbb{R}$ is the transition probability distribution
  - r is reward function
  - $\rho_0: S \rightarrow \mathbb{R}$ is the distribution of initial state (implicitly, $s_0 \sim \rho_0$)
  - $\gamma \in (0, 1)$ is discounted factor
- Let π be a stochastic policy $\pi: S \times A \rightarrow [0, 1]$
- The return function of reinforcement learning is

$$\eta(\pi) \coloneqq E_{s_0 \sim \rho_0}[V_\pi(s_0)] = \mathbb{E}_{s_0, a_0, \ldots \sim \sim \rho_0, \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$$

TRPO

- Starting point:

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]$$

  - Proposed in 2002 by Kakade & Langford
  - Note: for simplicity, $\sim\rho_0$ is omitted later.
- This implies that we can derive “return of new policy” from “advantage of old policy”
  - Advantage $A_\pi(s_t,a_t) \coloneqq Q_\pi(s_t,a_t) - V_\pi(s_t)$

Appendix (Proof of the previous equation)

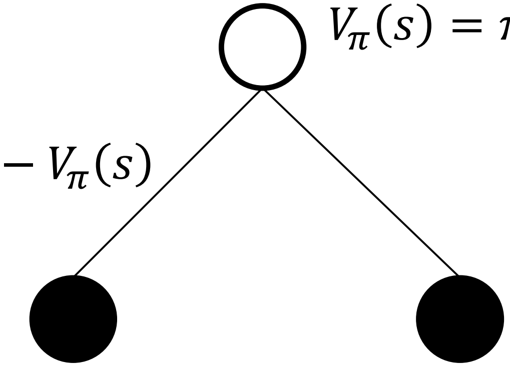
Since $A_\pi(s,a) = E_{s' \sim P(s'|s,a)}[r(s) + \gamma V_\pi(s') - V_\pi(s)]$,

we have

$$
\begin{aligned}
& E_{s_0, a_0, \ldots \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right] \\
&= E_{s_0, a_0, \ldots \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\big)\right] \\
&= E_{s_0, a_0, \ldots \sim \tilde{\pi}}\left[-V_\pi(s_0) + \sum_{t=0}^{\infty} \gamma^t r(s_t)\right] \\
&= -E_{s_0}[V_\pi(s_0)] + E_{s_0, a_0, \ldots \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] \\
&= -\eta(\pi) + \eta(\tilde{\pi}) \qquad \square
\end{aligned}
$$

TRPO

- Advantage $A_\pi(s_t, a_t) \coloneqq Q_\pi(s_t, a_t) - V_\pi(s_t)$
- Can evaluate the current action compared to average value



TRPO

- Expanding $\eta(\tilde{\pi})$, we get

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{s_0,a_0,\ldots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]$$

$$= \eta(\pi) + \sum_{t=0}^{\infty}\left(\sum_{s}\left(P(s_t = s|\tilde{\pi})\sum_{a}\tilde{\pi}(a|s)\gamma^t A_\pi(s,a)\right)\right)$$

$$= \eta(\pi) + \sum_{s}\left(\left(\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\tilde{\pi})\right)\left(\sum_{a}\tilde{\pi}(a|s)A_\pi(s,a)\right)\right)$$

Called density of s, denoted $\rho_{\tilde{\pi}}(s)$

$$= \eta(\pi) + \sum_{s}\left(\rho_{\tilde{\pi}}(s)\left(\sum_{a}\tilde{\pi}(a|s)A_\pi(s,a)\right)\right)$$

  - Convert the view from each time point t to each state s

TRPO

$$\rho_{\tilde{\pi}}(s) \coloneqq \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\tilde{\pi}) = \underbrace{P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \cdots}_{\text{unnormalized discounted visitation frequencies}}$$

- Denote the un-normalized discounted visitation frequencies by $\rho_{\tilde{\pi}}(s)$, then the return of $\tilde{\pi}$ become

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

- This implies
  - any policy update $\pi \to \tilde{\pi}$ that has a nonnegative expected advantage at every state s, is guaranteed to increase the policy performance η
  - *or*: If all $\sum_a \tilde{\pi}(a|s) A_\pi(s,a)$ are non-negative for the new policy $\tilde{\pi}$, the policy performance η must be improved.

TRPO

$$\rho_{\tilde{\pi}}(s) := \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\tilde{\pi}) = \underbrace{P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \cdots}_{\text{unnormalized discounted visitation frequencies}}$$

- Denote the un-normalized discounted visitation frequencies by $\rho_{\tilde{\pi}}(s)$, then the return of $\tilde{\pi}$ become

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a|s) A_{\pi}(s,a)$$

- This implies
  - any policy update $\pi \rightarrow \tilde{\pi}$ that has a nonnegative expected advantage at every state s, is guaranteed to increase the policy performance η
  - *or*: If all $\sum_a \tilde{\pi}(a|s) A_{\pi}(s,a)$ are non-negative for the new policy $\tilde{\pi}$, the policy performance η must be improved.

TRPO

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

- However, due to the complex dependency of $\rho_{\tilde{\pi}}(s)$ on $\tilde{\pi}$ makes above equation difficult to optimize directly
- Instead, introducing local approximation to η:

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

  - L ignores changes in state visitation density due to changes in the policy
- Key: try to maximize $L_\pi(\tilde{\pi})$ instead of $\eta(\tilde{\pi})$.
  - Question: why is it fine to replace $\rho_{\tilde{\pi}}(s)$ by $\rho_\pi(s)$?

TRPO

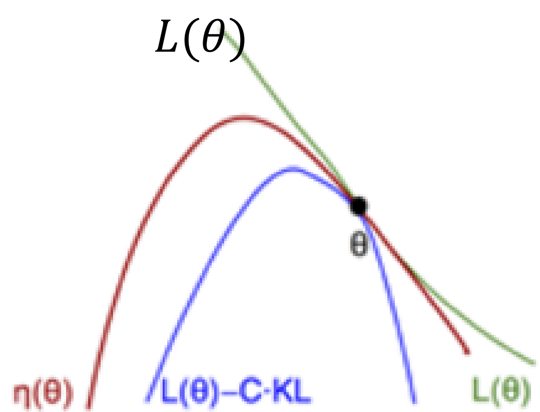
- If we have a policy π_θ, which is differentiable w.r.t. θ, then L_π matches η to first order. i.e., for any parameter θ_{old}

$$L_{\pi_{\theta_{old}}}(\pi_{\theta_{old}}) = \eta(\pi_{\theta_{old}}),$$

$$\nabla_\theta L_{\pi_{\theta_{old}}}(\pi_\theta)\Big|_{\theta=\theta_{old}} = \nabla_\theta \eta(\pi_\theta)\Big|_{\theta=\theta_{old}}$$

Proved in next page

- This implies that a step small enough that improves $L_{\pi_{old}}$ will also improve η.



- Sutton's proof by induction for

$$\frac{\partial \eta(\pi_\theta)}{\partial \theta} = \sum_s \rho^\pi(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} Q^\pi(s,a)$$

For the start-state formulation:

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &\stackrel{\text{def}}{=} \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) \qquad \forall s \in \mathcal{S} \\
&= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} Q^\pi(s,a) \right] \\
&= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} \left[\mathcal{R}_s^a + \sum_{s'} \gamma \mathcal{P}_{ss'}^a V^\pi(s') \right] \right] \\
&= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \sum_{s'} \gamma \mathcal{P}_{ss'}^a \frac{\partial}{\partial \theta} V^\pi(s') \right] \qquad (7) \\
&= \sum_x \sum_{k=0}^{\infty} \gamma^k Pr(s \to x, k, \pi) \sum_a \frac{\partial \pi(x,a)}{\partial \theta} Q^\pi(x,a),
\end{aligned}$$

- Sutton's proof by induction for

$$\frac{\partial \eta(\pi_\theta)}{\partial \theta} = \sum_s \rho^\pi(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} Q^\pi(s,a)$$

$$= \sum_s \rho^\pi(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} A^\pi(s,a)$$

(Why? $\sum_a \pi_\theta(a|s) V^\pi(s) = 1$)

$$= \frac{\partial L(\pi_\theta)}{\partial \theta}$$

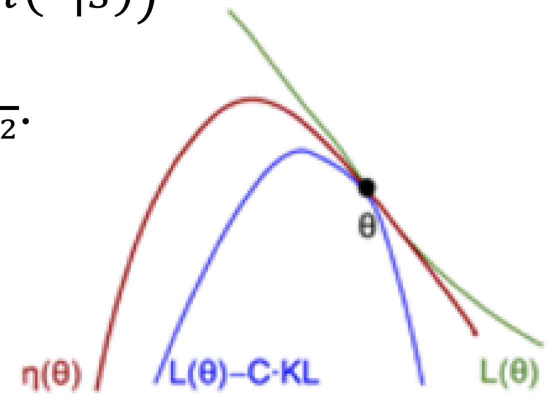
TRPO (next five pages can be skipped)

- The main result in this paper is the following theorem:
- Let $\alpha = D_{TV}^{max}(\pi, \tilde{\pi})$, then the following bound holds:

$$\eta(\tilde{\pi}) \geq L_{\pi_{old}}(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2$$

$$\text{where } \epsilon = \max_{s,a}|A_\pi(s,a)|$$

  - $D_{TV}(p,q) = \frac{1}{2}\sum_i |p_i - q_i|$ for discrete probability distribution p, q
  - $D_{TV}^{max}(\pi, \tilde{\pi}) = \max_s D_{TV}\big(\pi(\cdot|s) \parallel \tilde{\pi}(\cdot|s)\big)$
  - Note: we will use C to denote $\frac{4\epsilon\gamma}{(1-\gamma)^2}$.



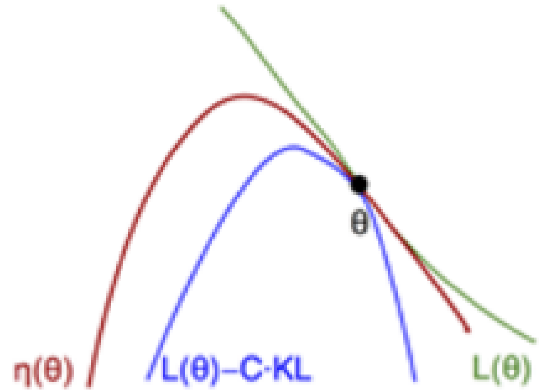
TRPO

- And $D_{TV}(p \parallel q)^2 \leq D_{KL}(p \parallel q)$.
- Let $D_{KL}^{max}(\pi, \tilde{\pi}) = \max_s D_{KL}\big(\pi(\cdot\,|s) \parallel \tilde{\pi}(\cdot\,|s)\big)$, then

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - C \cdot D_{KL}^{max}(\pi, \tilde{\pi})$$

$$\text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$$

  – When $\pi \to \tilde{\pi}$, $D_{KL}^{max}(\pi, \tilde{\pi}) \to 0$, so the lower bound is tight. How much we improve on $L_\pi(\tilde{\pi})$, how much the return $\eta(\tilde{\pi})$ also improve
  – When π is not close to $\tilde{\pi}$, the penalty is large since constant C is large, and the lower bound is meaningless.
- A kind of MM algorithm
  – Minorize-Maximization or
  – Majorize-Minimization



TRPO

$$\eta(\tilde{\pi}) \geq L_{\pi}(\tilde{\pi}) - C \cdot D_{KL}^{max}(\pi, \tilde{\pi})$$

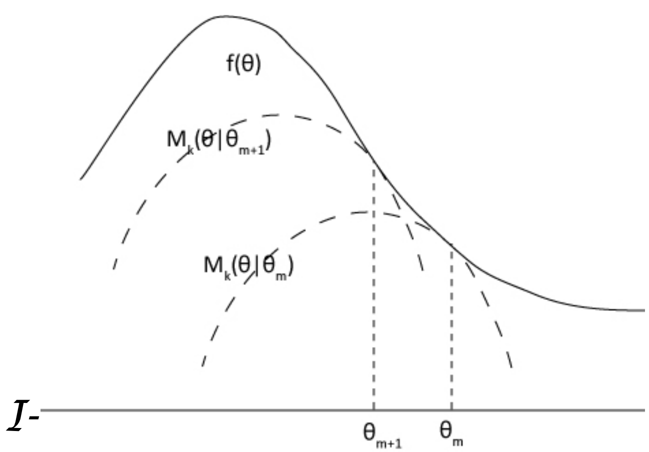
- We show that the improvement must be monotonically increasing (MM algorithm)
- Let $M_i(\pi) = L_{\pi_i}(\pi) - C \cdot D_{KL}^{max}(\pi_i, \pi)$:

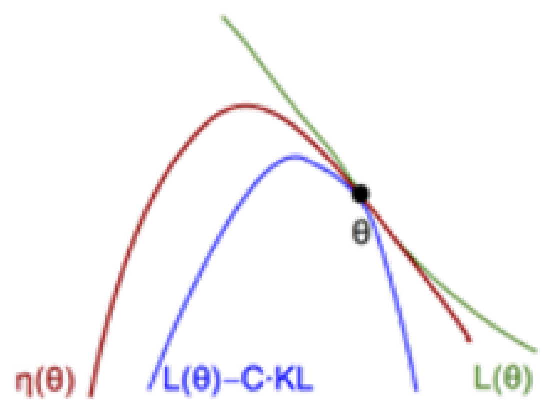
$$\eta(\pi) \geq M_i(\pi)$$
$$\eta(\pi_i) = M_i(\pi_i)$$
$$\eta(\pi) - \eta(\pi_i) \geq M_i(\pi) - M_i(\pi_i)$$

- Let $\pi_{i+1} = \text{argmax}_{\pi} M_i(\pi)$, then

$$\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i) \geq 0$$

and thus the return of next iteration is not worse than current one.





M =

TRPO

- Algorithm

Algorithm 1 Policy iteration algorithm guaranteeing non-decreasing expected return η

Initialize π_0.
for $i = 0, 1, 2, \ldots$ until convergence **do**
Compute all advantage values $A_{\pi_i}(s, a)$.
Solve the constrained optimization problem

$$\pi_{i+1} = \arg\max_{\pi} \left[L_{\pi_i}(\pi) - C D_{\mathrm{KL}}^{\max}(\pi_i, \pi) \right]$$

where $C = 4\epsilon\gamma/(1-\gamma)^2$

and $L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s, a)$

end for

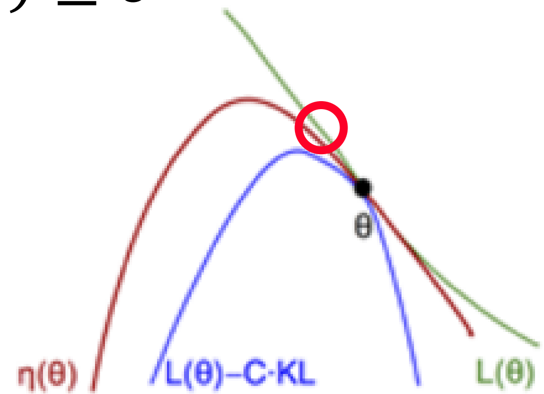
TRPO

- Problems:
  - In practice, the step size is very small
  - D_{KL}^{max} is hard to compute
  - How do we approximate the objective function and constraint?

TRPO

- In practice, if using the penalty coefficient C recommended by the theory above, the step size would be very small.
- One way to take larger steps in a robust way is to use a constraint on the KL divergence between the new policy and the old policy, i.e., a trust region constraint:

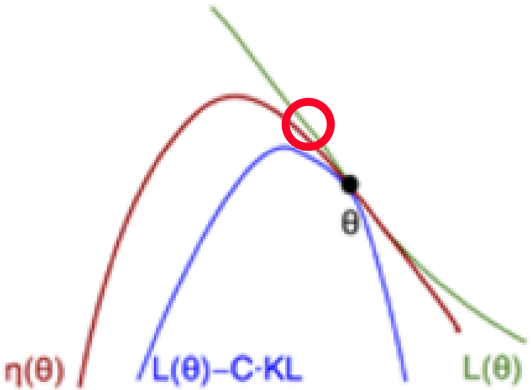
$$\max_{\theta} L_{\theta_{old}}(\theta)$$
$$\text{subject to } D_{KL}^{max}(\theta_{old}, \theta) \leq \delta$$



TRPO (can be skipped)

- Since the D_{KL}^{max} is hard to compute, we can use a heuristic approximation which considers the average KL divergence $\bar{D}_{KL}^{\rho}(\theta_{old}, \theta) \coloneqq \mathbb{E}_{s\sim\rho}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot\,|s) \parallel \pi_{\theta}(\cdot\,|s)\right)\right]$
- Thus, the problem becomes

$$\max_{\theta} L_{\theta_{old}}(\theta)$$
$$\text{subject to } \bar{D}_{KL}^{\rho}(\theta_{old}, \theta) \le \delta$$



TRPO

- Transform the problem: $\max_\theta L_{\theta_{old}}(\theta)$

$$\max_\theta \sum_s \rho_{\theta_{old}}(s) \sum_a \pi_\theta(a|s) A_{\theta_{old}}(s,a)$$

$$\text{subject to } \overline{D}_{KL}^{\rho}(\theta_{old},\theta) \leq \delta$$

1. Replace $\sum_s \rho_{\theta_{old}}(s)[\cdots]$ by expectation $\frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_{\theta_{old}}}[\cdots]$
2. Replace the sum over the actions by an importance sampling estimator. Using $\pi_{\theta_{old}}(a|s)$ to denote the sampling distribution, then the contribution of a single s_n to the loss function is:

$$\sum_a \pi_\theta(a|s_n) A_{\theta_{old}}(s_n,a) = \mathbb{E}_{a\sim\pi_{\theta_{old}}(a|s_n)}\left[\frac{\pi_\theta(a|s_n)}{\pi_{\theta_{old}}(a|s_n)} A_{\theta_{old}}(s_n,a)\right]$$

TRPO

- The problem at the beginning:

$$\max_\theta L(\pi_{\theta_{old}}) \;\; or$$

$$\max_\theta \sum_s \rho_{\theta_{old}}(s) \sum_a \pi_\theta(a|s) A_{\theta_{old}}(s,a)$$

$$\text{subject to } \bar{D}_{KL}^{\rho}(\theta_{old}, \theta) \le \delta$$

- And currently, we solve:

$$\max_\theta \mathbb{E}_{s\sim\rho_{\theta_{old}},\, a\sim\pi_{\theta_{old}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s,a)\right]$$

$$\text{subject to } \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot\,|s) \parallel \pi_\theta(\cdot\,|s)\right)\right] \le \delta$$

- In another form, maximize a surrogate objective:

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta\,old}(a_t|s_t)} \hat{A}_t\right]$$

  - CPI: conservative policy iteration
  - $\hat{A}_t$: can be any form of advantage, like GAE.

Proximal Policy Optimization (PPO)

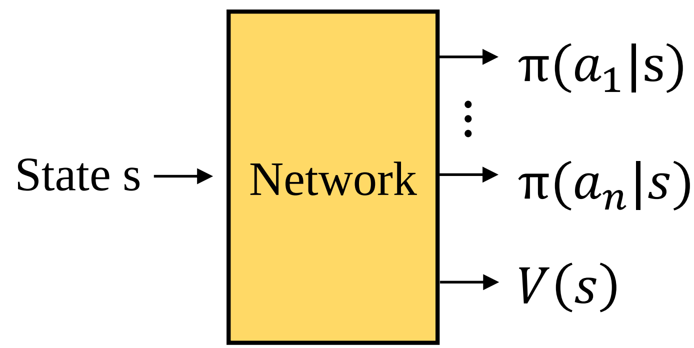
- Problems of TRPO:
  - still relatively complicated, and
  - not compatible with architectures that include noise (such as dropout) or parameter sharing
- Background:
  - In 2017, OpenAI release a new reinforcement learning algorithms, PPO.
  - PPO has some of the benefits of TRPO, but much simpler to implement, more general, and has better sample complexity.
  - attains the data efficiency and reliable performance of TRPO, while using only first-order optimization
- The experiments show that PPO outperforms other online policy gradient methods, like A2C or TRPO.
  - Although PPO is a little worse than ACER (Actor-Critic with Experience Replay), the implementation of PPO is much easier than ACER.

Generalized Advantage Estimation (GAE)

- Use the learned state-value function $V(s)$ to compute variance-reduced advantage-function estimators.
- PPO uses a truncated version of generalized advantage estimation

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$$

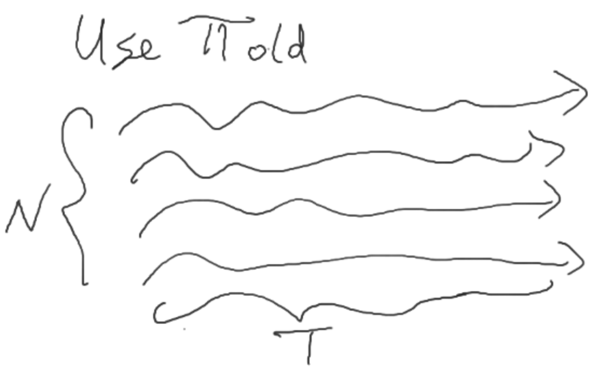
where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$



PPO Algorithm

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

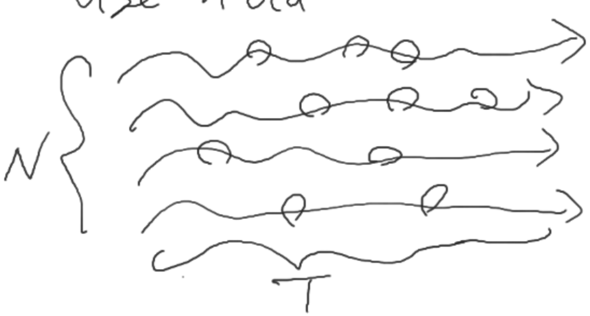


PPO Algorithm

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

Use π_{old}



Let $\pi_0 = \pi_{old}$

1. pick a batch with M
2. optimize θ.
   from $\pi_i \to \pi_{i+1}$
3. repeat 1.

$$\pi_0 \to \pi_1 \to \pi_2 \to \cdots \to \pi_K$$

Recall TRPO

- Recall: TRPO maximizes a surrogate objective: $\max_\theta L^{CPI}(\theta)$ (with small change on $\pi_\theta(a|s)$)

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right] = \hat{\mathbb{E}}_t\left[r_t(\theta)\hat{A}_t\right]$$

  - CPI: conservative policy iteration
  - $\hat{A}_t$: can be any form of advantage, like GAE.
- Let $r_t(\theta)$ denote the probability ratio (not reward)

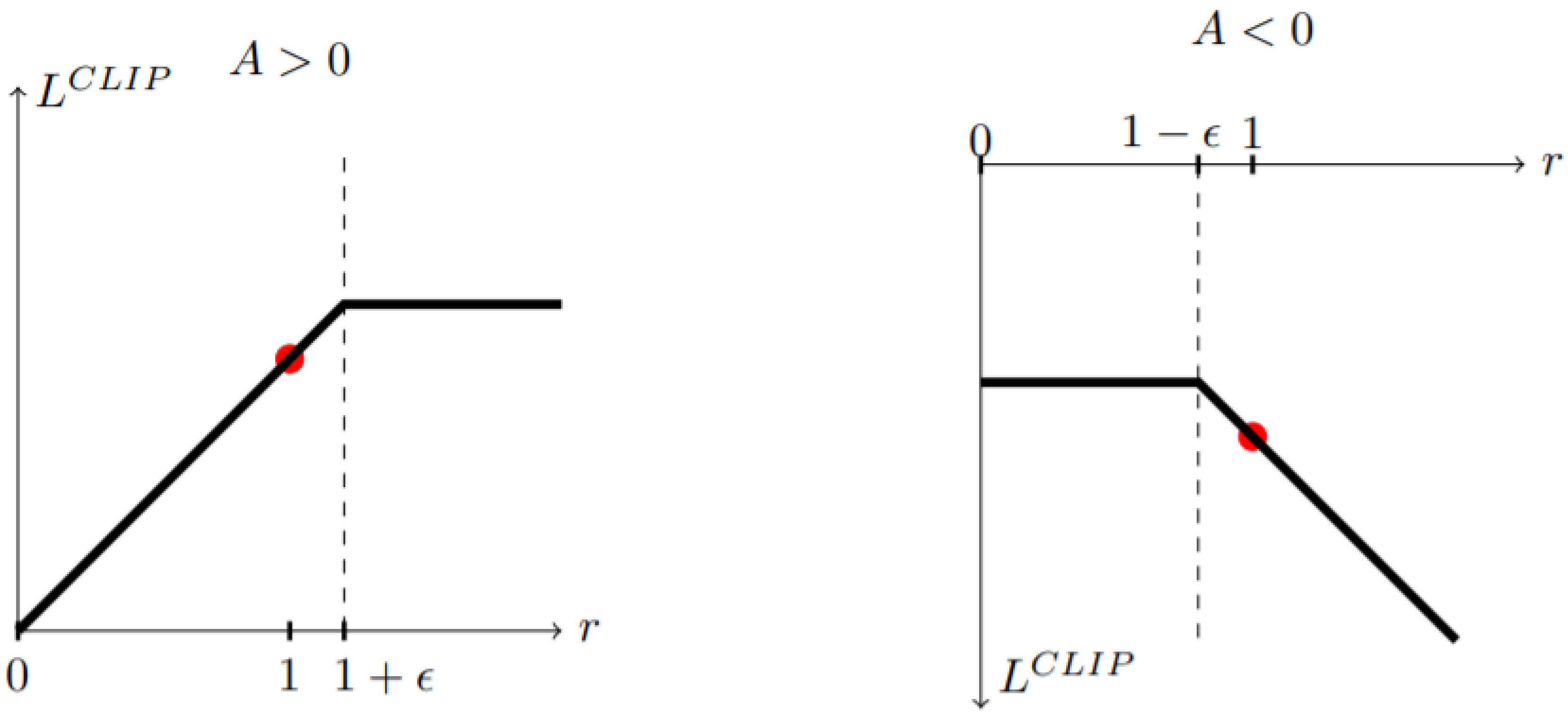
$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

  - $r(\theta_{old}) = 1$
  - Note: π_θ can be any of π_i in PPO

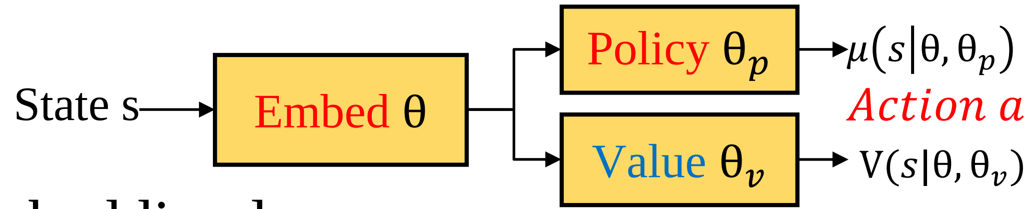
PPO Clip

- Without constraint, the step size of L^{CPI} would be large
- Hence, we consider modifying the objective, to penalize changes to the policy that move $r_t(\theta)$ away from 1
- The main objective proposed in PPO is:
  $L^{CLIP} = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$
  - ϵ is a hyper-parameter
  - First term implies that the min is L^{CPI}
  - Second term modifies the surrogate objective by clipping the probability ratio
  - The final objective is a lower bound on L^{CPI}

- If clipped, $\nabla_\theta L^{CLIP}$ becomes 0, and then drop the gradient



PPO



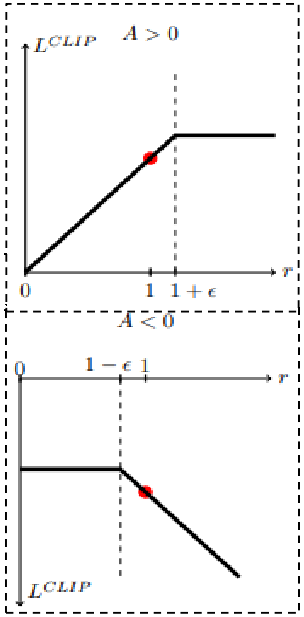
- Use one network with same embedding layer: policy and value
  - Value: estimates value of current state by TD-like learning
    - Value loss: $L_t^{VF}(\theta) = \left(V_\theta(s_t) - V_t^{target}\right)^2$
  - Policy: output probability of actions
    - Policy obj.: $L_t^{CLIP}(\theta) = \widehat{E}_t\left[\min\left(r_t(\theta)\widehat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_t\right)\right]$

      where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$,

      $\widehat{A}_t$ is generalized advantage estimation (GAE)

      $\widehat{A}_t = \sum_{n=0}^{\infty}(\gamma\lambda)^n\delta_{t+n}^V$,

      where $\delta_t^V = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$ [TD error]



  - Total objective (usually version): maximize

    $L_t^{CLIP+VF+S}(\theta) = \widehat{E}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]$

    - Augment with an entropy bonus (S) to ensure sufficient exploration

Experiments - PPO