

# Proximal Policy Optimization Algorithms

From AutoAugment

# Overview

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

# Pi - Actor

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

# Pi – N parallel Actors

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

# Loss (or what is called the loss)

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$

# Loss (or what is called the loss)

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \text{ so } r(\theta_{\text{old}}) = 1.$$

# Advantage

---

**Algorithm 1** PPO, Actor-Critic Style

---

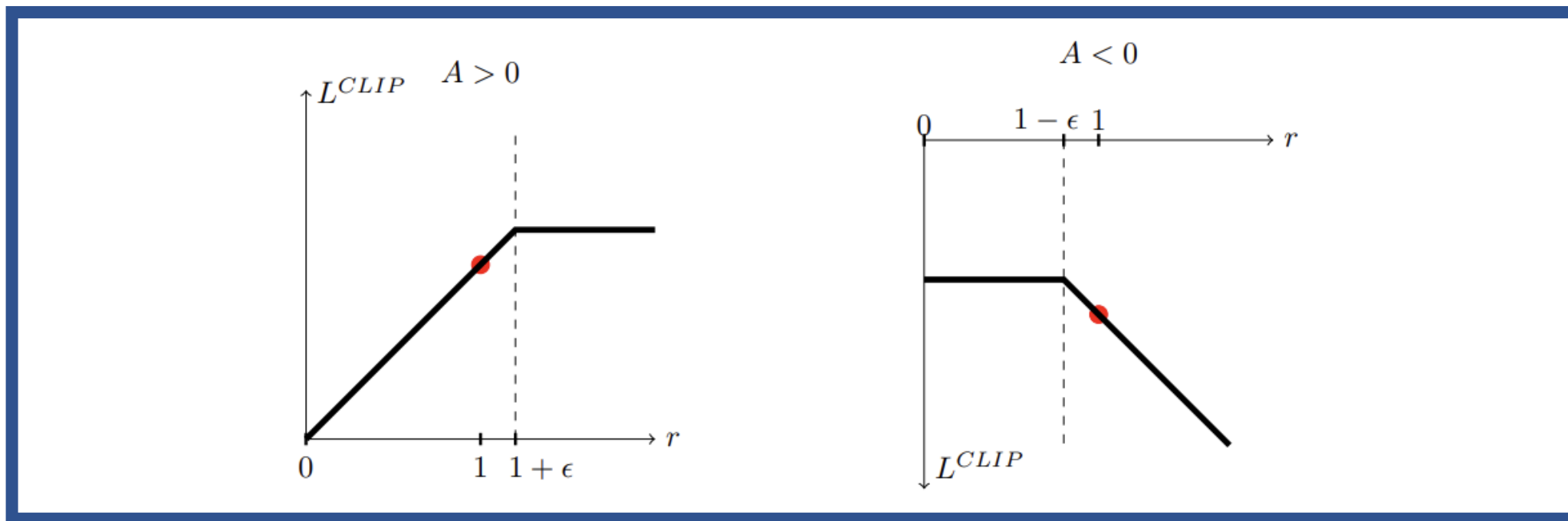
```
for iteration=1,2,... do
  for actor=1,2,...,N do
    Run policy  $\pi_{\theta_{\text{old}}}$  in environment for  $T$  timesteps
    Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$ 
  end for
  Optimize surrogate  $L$  wrt  $\theta$ , with  $K$  epochs and minibatch size  $M \leq NT$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

---

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \quad (11)$$

$$\text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (12)$$

# Loss (or what is called the loss)

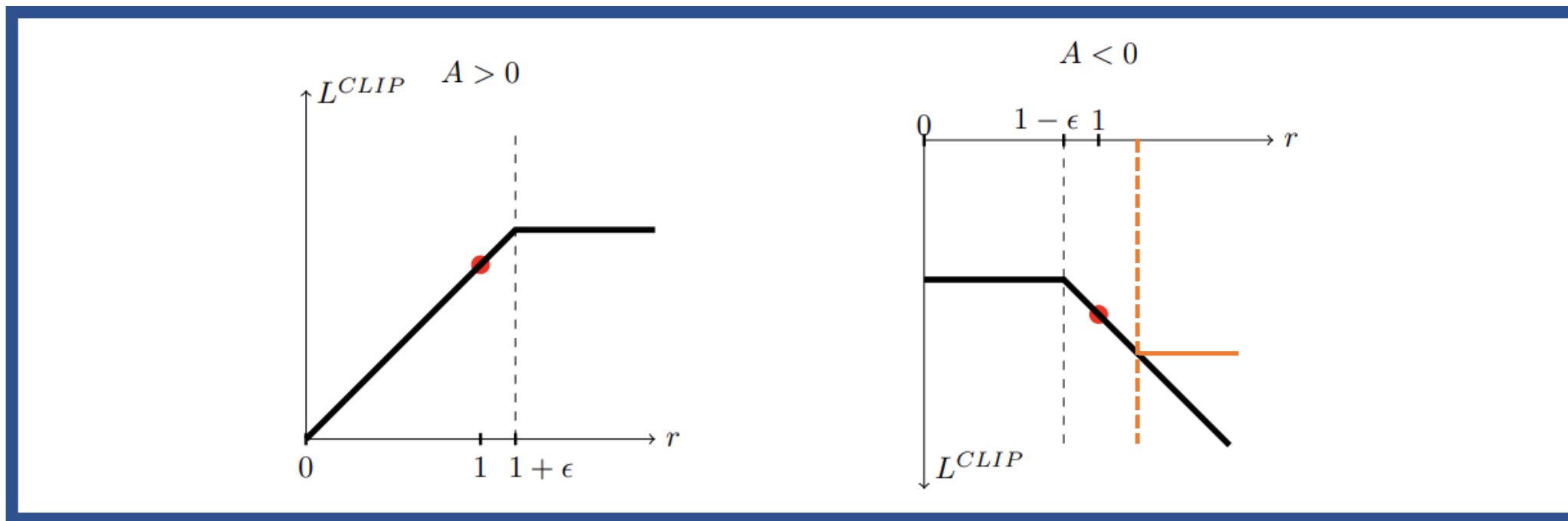


$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \text{ so } r(\theta_{\text{old}}) = 1.$$



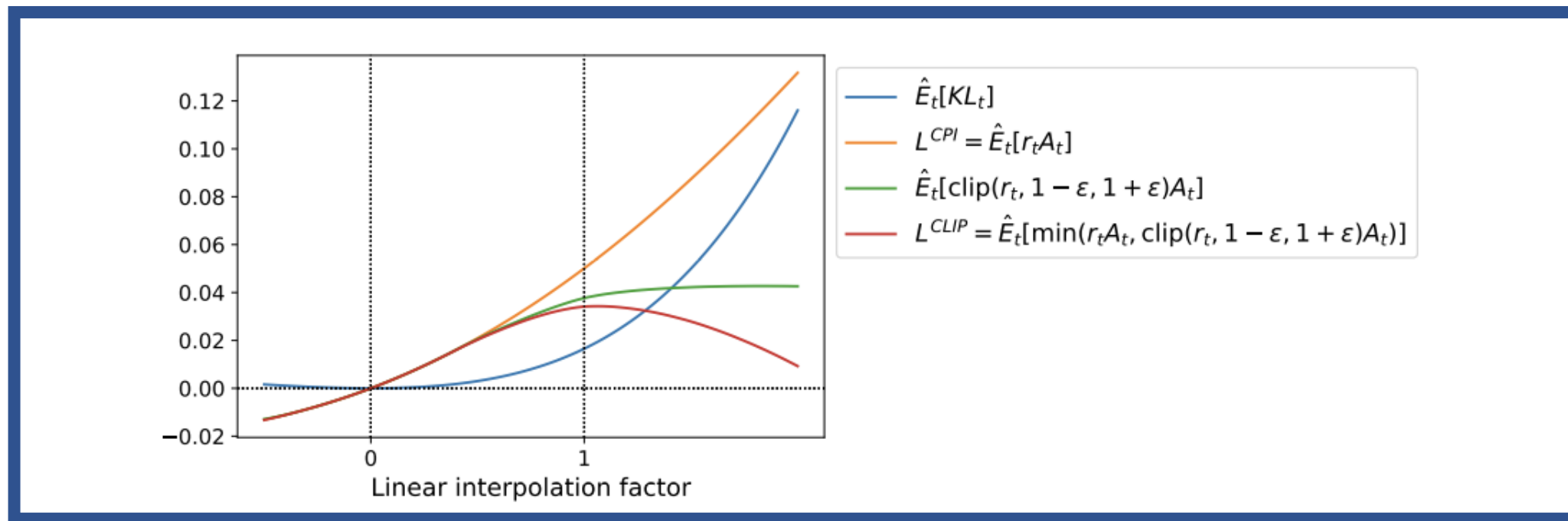
# Loss (or what is called the loss)



$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \text{ so } r(\theta_{\text{old}}) = 1.$$

# Loss (or what is called the loss)



$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \text{ so } r(\theta_{\text{old}}) = 1.$$

# Loss (or what is called the loss)

---

**Algorithm 1** PPO, Actor-Critic Style

---

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

---

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (7)$$


$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t [L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)], \quad (9)$$

# Loss (or what is called the loss)

---

**Algorithm 1** PPO, Actor-Critic Style

---



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

---


$\pi_{\text{old}} = \pi$

# Loss (or what is called the loss)

---

**Algorithm 1** PPO, Actor-Critic Style

---



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

---

Pi old = pi

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \text{ so } r(\theta_{\text{old}}) = 1.$$

?

# Test Results on Atari Domain 엥?

	A2C	ACER	PPO	Tie
(1) avg. episode reward over all of training	1	18	<b>30</b>	0
(2) avg. episode reward over last 100 episodes	1	<b>28</b>	19	1

