





- **Reward is negative:** The objective reduces to

$$\max \left( \frac{\pi_{\theta}(a^{(t)} \mid s^{(t)})}{\pi_{\theta_{\text{old}}}(a^{(t)} \mid s^{(t)})}, (1 - \epsilon) \right) R^{(t)}$$

Then, the objective decreases with  $\pi_{\theta}(a^{(t)} \mid s^{(t)})$ .

Once  $\pi_{\theta}(a^{(t)} \mid s^{(t)}) < (1 - \epsilon)\pi_{\theta_{\text{old}}}(a^{(t)} \mid s^{(t)})$ ,  
the max kicks in, with a ceiling of  $(1 - \epsilon)R^{(t)}$ .

- **Reward is positive:** The objective reduces to

$$\min \left( \frac{\pi_{\theta}(a^{(t)} \mid s^{(t)})}{\pi_{\theta_{\text{old}}}(a^{(t)} \mid s^{(t)})}, (1 + \epsilon) \right) R^{(t)}$$

Then, the objective increases with  $\pi_{\theta}(a^{(t)} \mid s^{(t)})$ .

Once  $\pi_{\theta}(a^{(t)} \mid s^{(t)}) > (1 + \epsilon)\pi_{\theta_{\text{old}}}(a^{(t)} \mid s^{(t)})$ ,

the min kicks in, with a ceiling of  $(1 + \epsilon)R^{(t)}$ .

*PPO-Clip as a Regularizer*

**Key insight:** the new policy does **not** benefit  
by going far away from the old policy.







[OpenAI SpinningUp, Proximal Policy Optimization]

- Regularizer; similar to batch training jittering.
- $\epsilon$  is a trainable hyperparameter.

# PPO-Clip as a Regularizer

- **Reward is positive:** The objective reduces to

$$\min \left( \frac{\pi_{\theta}(a^{(t)} | s^{(t)})}{\pi_{\theta_{\text{old}}}(a^{(t)} | s^{(t)})}, (1 + \epsilon) \right) R^{(t)}$$

Then, the objective increases with  $\pi_{\theta}(a^{(t)} | s^{(t)})$ .  
Once  $\pi_{\theta}(a^{(t)} | s^{(t)}) > (1 + \epsilon)\pi_{\theta_{\text{old}}}(a^{(t)} | s^{(t)})$ ,  
the min kicks in, with a ceiling of  $(1 + \epsilon)R^{(t)}$ .

- **Reward is negative:** The objective reduces to


$$\max \left( \frac{\pi_{\theta}(a^{(t)} | s^{(t)})}{\pi_{\theta_{\text{old}}}(a^{(t)} | s^{(t)})}, (1 - \epsilon) \right) R^{(t)}$$

Then, the objective decreases with  $\pi_{\theta}(a^{(t)} | s^{(t)})$ .  
Once  $\pi_{\theta}(a^{(t)} | s^{(t)}) < (1 - \epsilon)\pi_{\theta_{\text{old}}}(a^{(t)} | s^{(t)})$ ,  
the max kicks in, with a ceiling of  $(1 - \epsilon)R^{(t)}$ .

**Key insight:** the new policy does not benefit  
by going far away from the old policy.

- Regularizer; similar to batch training jittering.
- $\epsilon$  is a trainable hyperparameter.



A misty, layered mountain landscape with a winding road in the foreground. The scene is hazy and atmospheric, with multiple ridges of mountains visible in the distance, each progressively lighter and more faded than the one in front. In the lower foreground, a dark, winding road curves through a valley. The overall color palette is muted, consisting of soft greys, greens, and whites, creating a sense of depth and tranquility.

# RESULTS