- **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 <u>far away</u> from the old policy.

[OpenAI SpinningUp, *Proximal Policy Optimization*]

- Regularizer; similar to batch training jittering.
- ϵ is a trainable hyperparameter.

PPO-Clip as a Regularizer

- **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 $(1+\epsilon)R^{(t)}$.

- **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 $(1-\epsilon)R^{(t)}$.

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

- Regularizer; similar to batch training jittering.
- ϵ is a trainable hyperparameter.

[OpenAI SpinningUp, *Proximal Policy Optimization*]

Results