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

Search as an optimization problem

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

$$g\left(\epsilon, R^{(t)}\right) = \begin{cases} (1 + \epsilon) R^{(t)} & R^{(t)} \geq 0 \\ (1 - \epsilon) R^{(t)} & R^{(t)} < 0 \end{cases}$$

PPO-Clip as a Regularizer

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