Search as an optimization problem

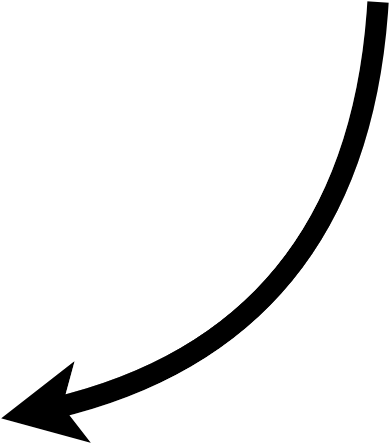
sampling distribution over the previous time *t*

clipping function

likelihood of the new policy choosing a new action vs. the old policy

$$\text{maximize}_\theta \, \mathbb{E}_{\left(a^{(t)},\, s^{(t)}\right) \sim \pi_{\theta_{\text{old}}}} \left[\min \left(\frac{\pi_\theta \left(a^{(t)} \mid s^{(t)}\right)}{\pi_{\theta_{\text{old}}} \left(a^{(t)} \mid s^{(t)}\right)} R^{(t)},\ g\left(\epsilon, R^{(t)}\right) \right) \right]$$

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



[OpenAI SpinningUp, *Proximal Policy Optimization*]

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

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