Simplified PPO-Clip Objective

Joshua Achiam

July 30, 2018

The objective for PPO-Clip is given by Schulman et al. as

$$L_{\theta_k}^{CLIP}(\theta) \doteq \underset{s,a\sim\theta_k}{\mathrm{E}} \left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\theta_k}(s,a), \;\; \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1-\epsilon, 1+\epsilon \right) A^{\theta_k}(s,a) \right) \right],$$

where θ_k are the parameters of the policy at iteration k and ϵ is a small hyperparameter.

Proposition 1. The PPO-Clip objective can be simplified to

$$L_{\theta_k}^{CLIP}(\theta) = \underset{s,a\sim\theta_k}{\mathrm{E}} \left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\theta_k}(s,a), \;\; g\left(\epsilon, A^{\theta_k}(s,a)\right) \right) \right],$$

where

$$g(\epsilon, A) = \begin{cases} (1+\epsilon)A & A \geq 0 \\ (1-\epsilon)A & \text{otherwise} \end{cases}$$

Proof. Suppose $\epsilon \in (0,1)$. Define

$$F(r, A, \epsilon) \doteq \min\left(rA, \;\; \text{clip}(r, 1-\epsilon, 1+\epsilon)A\right).$$

When $A \geq 0$,

$$
\begin{aligned}
F(r, A, \epsilon) &= \min\left(rA, \quad \textbf{clip}(r, 1-\epsilon, 1+\epsilon)A\right) \\
&= A \min\left(r, \quad \textbf{clip}(r, 1-\epsilon, 1+\epsilon)\right) \\
&= A \min\left(r, \quad \left\{ \begin{array}{ll} 1+\epsilon & r \geq 1+\epsilon \\ r & r \in (1-\epsilon, 1+\epsilon) \\ 1-\epsilon & r \leq 1-\epsilon \end{array} \right\}\right) \\
&= A \left\{ \begin{array}{ll} \min(r, 1+\epsilon) & r \geq 1+\epsilon \\ \min(r, r) & r \in (1-\epsilon, 1+\epsilon) \\ \min(r, 1-\epsilon) & r \leq 1-\epsilon \end{array} \right\} \\
&= A \left\{ \begin{array}{ll} 1+\epsilon & r \geq 1+\epsilon \\ r & r \in (1-\epsilon, 1+\epsilon) \\ r & r \leq 1-\epsilon \end{array} \right\} \\
&= A \min(r, (1+\epsilon)) \\
\therefore F(r, A, \epsilon) &= \min(rA, (1+\epsilon)A)
\end{aligned}
$$

When $A < 0$,

$$
\begin{aligned}
F(r, A, \epsilon) &= \min\left(rA, \quad \textbf{clip}(r, 1-\epsilon, 1+\epsilon)A\right) \\
&= A \max\left(r, \quad \textbf{clip}(r, 1-\epsilon, 1+\epsilon)\right) \\
&= A \max\left(r, \quad \left\{ \begin{array}{ll} 1+\epsilon & r \geq 1+\epsilon \\ r & r \in (1-\epsilon, 1+\epsilon) \\ 1-\epsilon & r \leq 1-\epsilon \end{array} \right\}\right) \\
&= A \left\{ \begin{array}{ll} \max(r, 1+\epsilon) & r \geq 1+\epsilon \\ \max(r, r) & r \in (1-\epsilon, 1+\epsilon) \\ \max(r, 1-\epsilon) & r \leq 1-\epsilon \end{array} \right\} \\
&= A \left\{ \begin{array}{ll} r & r \geq 1+\epsilon \\ r & r \in (1-\epsilon, 1+\epsilon) \\ 1-\epsilon & r \leq 1-\epsilon \end{array} \right\} \\
&= A \max(r, (1-\epsilon)) \\
\therefore F(r, A, \epsilon) &= \min(rA, (1-\epsilon)A)
\end{aligned}
$$

Summarizing, for all cases,

$$F(r, A, \epsilon) = \min(rA, g(\epsilon, A)),$$

where, as before, $g(\epsilon, A) = (1+\epsilon)A$ for $A \geq 0$, and $g(\epsilon, A) = (1-\epsilon)A$ for $A < 0$.

$\square$

The intuition we get from this: if a given state-action pair has negative advantage A, the optimization wants to make $\pi_\theta(a|s)$ smaller, but no additional benefit to the objective function is conferred by making $\pi_\theta(a|s)$ smaller than $(1-\epsilon)\pi_{\theta_k}(a|s)$. If a state-action pair has positive advantage A, the optimization wants to make $\pi_\theta(a|s)$ larger, but no additional benefit is gained by making $\pi_\theta(a|s)$ larger than $(1+\epsilon)\pi_{\theta_k}(a|s)$.