

The Clipping Function in PPO

Qiang Liu

Proximal Policy Optimization (PPO) [1] famously uses a clipped surrogate objective to mitigate the high variance issue associated with the vanilla policy gradient. The full PPO objective is

$$J^{\text{PPO}}(\pi) = \mathbb{E}_{s \sim d_0} \left[\mathbb{E}_{a \sim \pi^{\text{ref}}(\cdot | s)} \left[R^{\text{clip}}(w(s, a), A(s, a)) \right] - \beta \text{KL}(\pi(\cdot | s) \parallel \pi^{\text{ref}}(\cdot | s)) \right],$$

with $w(s, a) := \frac{\pi(a | s)}{\pi^{\text{ref}}(a | s)},$

where π^{ref} is the previous policy, w is the density ratio, d_0 is the state distribution, and $A(s, a)$ is the advantage. The PPO clipping function is defined as

$$R^{\text{clip}}(w, A) = \min(wA, \text{Clip}(w, [1 - \epsilon, 1 + \epsilon])A), \quad (1)$$

where $\text{Clip}(w, [1 - \epsilon, 1 + \epsilon]) = \min(\max(w, 1 - \epsilon), 1 + \epsilon)$ clips w to the interval $[1 - \epsilon, 1 + \epsilon]$.

It is intuitively known that the clipped objective induces a clipped density ratio at the optimal solution. We analyze this phenomenon. In particular, we show that the maximum of $J^{\text{PPO}}(\pi)$ is attained by a clipped exponential tilting of π^{ref} :

$$\pi^*(a | s) = \pi^{\text{ref}}(a | s) \text{Clip} \left(\exp \left(\frac{A(s, a) - \lambda(s)}{\beta} \right), [1 - \epsilon, 1 + \epsilon] \right),$$

where $\lambda(s)$ is chosen for each s to ensure that $\sum_a \pi^*(a | s) = 1$.

1 Understanding PPO Clipping

The form in (1) is not the simplest. The simpler expression below can help shed intuition more easily.

Proposition 1.1. *The $R^{\text{clip}}(w, A)$ in (1) is equivalent to*

$$R^{\text{clip}}(w, A) = \min(wA, A + \epsilon |A|).$$

See the proof in Appendix.

Hence, it simply caps the value of wA at a relative upper bound $A^{\epsilon+} = A + \epsilon |A|$, which is triggered when $w > 1 + \epsilon$ for $A > 0$, or when $w < 1 - \epsilon$ for $A < 0$. The intuition is:

Thus, wA is capped by the relative upper bound

$$A^{\epsilon+} := A + \epsilon |A|.$$

This cap is active when $w > 1 + \epsilon$ for $A > 0$, or when $w < 1 - \epsilon$ for $A < 0$. The intuition is as follows.

1. Policy optimization can be viewed as maximizing $w(s, a)A(s, a)$ on each data point, where

$$w(s, a) := \frac{\pi(a | s)}{\pi^{\text{ref}}(a | s)}.$$

This increases $\pi(a | s)$ for samples with positive advantage $A(s, a) > 0$, and decreases it when $A(s, a) < 0$.

2. Without any constraint or regularization, the optimal behavior would push $w(s, a) \rightarrow \infty$ when $A(s, a) > 0$, and $w(s, a) \rightarrow 0$ when $A(s, a) < 0$.
3. PPO clipping *gently* encourages w to stay within the range $[1 - \epsilon, 1 + \epsilon]$ by removing the incentive to further increase wA once it exceeds the cap $A^{\epsilon+}$. For each data point, maximizing wA is only beneficial up to $A + \epsilon|A|$. Beyond this point, changes in w no longer improve the objective, which discourages excessively large or small density ratios without explicitly constraining them.

See also the OpenAI Spinning Up PPO documentation for an intuitive discussion of the same form.

2 Maximizing the PPO-Clip Objective

Flattening the loss outside the interval $[1 - \epsilon, 1 + \epsilon]$ only removes the incentive to further increase or decrease w ; it does not impose a hard constraint on the density ratio.

To see this explicitly, consider maximizing $R^{\text{clip}}(w, A)$ for a fixed advantage A . In this case, the clipping function reduces to

$$R^{\text{clip}}(w, A) = \begin{cases} \min(w, 1 + \epsilon)A, & \text{if } A \geq 0, \\ \max(w, 1 - \epsilon)A, & \text{if } A \leq 0. \end{cases}$$

Hence, when $A \geq 0$, any $w^* \in [1 + \epsilon, \infty)$ maximizes $R^{\text{clip}}(w, A)$, while when $A \leq 0$, any $w^* \in (-\infty, 1 - \epsilon]$ is optimal. The clipping operation alone therefore does not prevent w from drifting arbitrarily far outside the clipping interval.

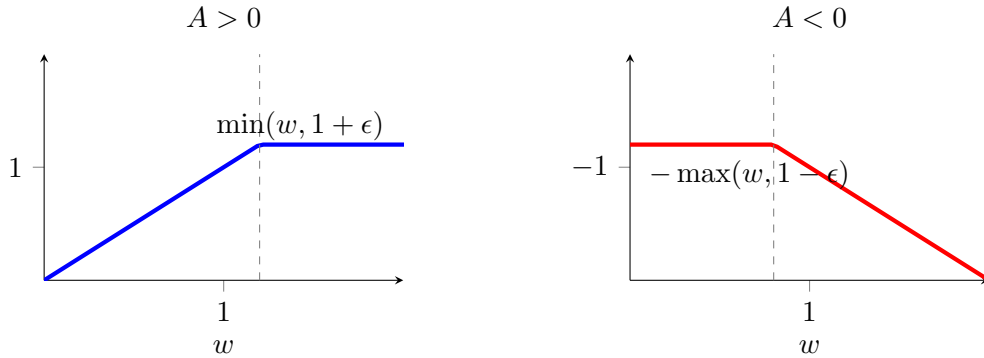


Figure 1: Clipped objective as a function of w for positive and negative advantages.

In practice, however, several additional mechanisms bias the solution toward smaller or more uniform density ratios:

- **Optimization bias.** Optimization typically initializes at $w = 1$, corresponding to the reference policy. Gradient-based updates from this point tend to remain in a moderate regime.
- **Coupling across (s, a) .** The ratios $w(s, a)$ are produced by a shared neural network and are therefore correlated. Updates induced by positive- and negative-advantage samples interact, which implicitly favors smaller deviations.
- **Stochastic advantages.** For a fixed (s, a) , the advantage $A(s, a)$ is noisy and may take either sign across samples. This variability penalizes large values of w that would otherwise exploit a fixed-sign advantage.
- **Explicit penalty terms.** Regularizers such as KL penalties directly discourage deviation from the reference policy.

Below, we analyze the last two effects in detail and show that, under mild conditions, the optimum of the PPO clipping objective necessarily lies in $[1 - \epsilon, 1 + \epsilon]$.

PPO-Clip on Stochastic Advantages If A is a random variable that takes both positive and negative values with nonzero probability, then the expected objective combines the positive and negative clipping terms. This coupling makes the objective strictly concave in the tails and ensures that the optimal solution lies in $[1 - \epsilon, 1 + \epsilon]$.

Proposition 2.1. *Let \mathbf{A} be a real-valued random variable. Consider the problem of maximizing the expected clipped objective:*

$$\max_{w \in \mathbb{R}} \left\{ \mathcal{R}^{\text{clip}}(w, \mathbf{A}) \stackrel{\text{def}}{=} \mathbb{E} [\min(w\mathbf{A}, \mathbf{A} + \epsilon|\mathbf{A}|)] \right\}.$$

Then we have the identity:

$$\mathcal{R}^{\text{clip}}(w, \mathbf{A}) = \min(w, 1 + \epsilon) \cdot \mu^+ + \max(w, 1 - \epsilon) \cdot \mu^-,$$

where $\mu^+ = \mathbb{E}[\max(\mathbf{A}, 0)]$ and $\mu^- = \mathbb{E}[\min(\mathbf{A}, 0)]$, and the optimum is attained if

$$w^* = \begin{cases} 1 + \epsilon, & \text{if } \mathbb{E}[\mathbf{A}] > 0, \\ 1 - \epsilon, & \text{if } \mathbb{E}[\mathbf{A}] < 0, \\ \text{any } w \in [1 - \epsilon, 1 + \epsilon], & \text{if } \mathbb{E}[\mathbf{A}] = 0. \end{cases}$$

In addition, if $\mu^+ > 0$ and $\mu^- < 0$, then the optimum must satisfy the condition above.

See the proof in Appendix.

Figure 2 shows the plot of $\mathcal{R}^{\text{clip}}(w, \mathbf{A})$, which combines the positive and negative parts.

Formally, we may write $w^* \in 1 + \text{sign}(\mathbb{E}[\mathbf{A}])\epsilon$, where $\text{sign}(\cdot)$ is interpreted as the subdifferential of the absolute function $|\cdot|$.

Since $\mathbb{E}[\mathbf{A}] = 0$ is rare to happen exactly, minimizing the expected clip function yields the extrem solution $w^* = 1 + \text{sign}(\mathbb{E}[\mathbf{A}])\epsilon$.

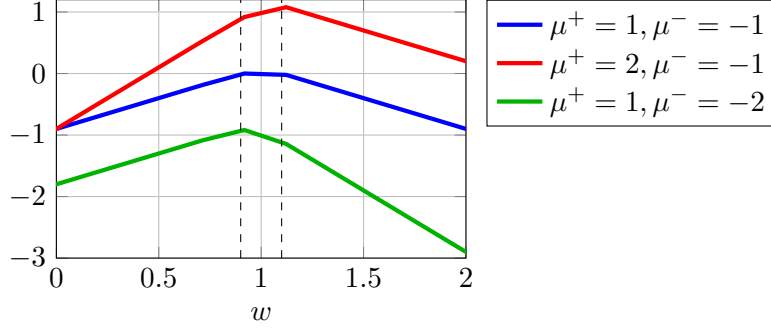


Figure 2: Plot of $\min(w, 1 + \epsilon)\mu^+ + \max(w, 1 - \epsilon)\mu^-$. Code here.

Maximum of Regularized PPO Clipping Further, introducing a strictly convex regularization term $\phi(w)$ biases the solution toward $w = 1$. The resulting optimizer is a clipped, monotone transform of $\mathbb{E}[\mathbf{A}]$, and the regularizer resolves the degeneracy when $\mathbb{E}[\mathbf{A}] = 0$.

Proposition 2.2. Assume $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable strictly convex function, whose minimum is attained at 1 (i.e., $\nabla\phi(1) = 0$). Consider

$$\max_w \mathbb{E}[\min(w\mathbf{A}, \mathbf{A} + \epsilon|\mathbf{A}|)] - \phi(w).$$

Then this problem is equivalent to the interval constrained optimization problem:

$$\max_{w \in \mathbb{R}} w\mathbb{E}[\mathbf{A}] - \phi(w) \quad \text{s.t.} \quad w \in [1 - \epsilon, 1 + \epsilon],$$

and the optimum is attained by

$$w^* = \text{Clip}(\nabla\phi^{-1}(\mathbb{E}[\mathbf{A}]), [1 - \epsilon, 1 + \epsilon]),$$

where $\nabla\phi^{-1}$ is the inverse function of $\nabla\phi$, which is also the derivative of the convex conjugate of ϕ .

See the proof in Appendix.

Maximum of Full PPO Objective Now consider the full PPO objective with KL divergence regularization:

$$J^{\text{PPO}}(\pi) = \mathbb{E}_{s \sim d_0} \left[\mathbb{E}_{a \sim \pi^{\text{ref}}(\cdot | s)} \left[\mathcal{R}^{\text{clip}}(w(s, a), \mathbf{A}(s, a)) \right] - \beta \text{KL}(\pi(\cdot | s) || \pi^{\text{ref}}(\cdot | s)) \right].$$

where we assume a stochastic advantage $\mathbf{A}(s, a) = \mathbf{A}(s, a, \xi)$, with ξ denoting an additional random source. That is, conditional on (s, a) , the advantage $\mathbf{A}(s, a)$ is a random variable.

Note that we can write the KL divergence into

$$\text{KL}(\pi(\cdot | s) || \pi^{\text{ref}}(\cdot | s)) = \sum_a \pi^{\text{ref}}(a | s) \phi_{\text{KL}}(w(a, s)),$$

$$\text{with} \quad \phi_{\text{KL}}(w) = w \log w - w + 1,$$

where ϕ_{KL} is strictly convex and is minimized at $w = 1$. Because $\nabla\phi(w) = \log w$ and $\nabla\phi^{-1}(w) = \exp(w)$, we can show that the optimum is $J^{\text{PPO}}(\pi)$ is attained by a clipped exponentially tilted distribution.

Theorem 2.3. *Consider*

$$\min_{\pi} J^{\text{PPO}}(\pi) \quad \text{s.t.} \quad \pi \in \Delta,$$

where Δ is the set of policy distributions. Then optimal solution π^* is obtained by

$$\pi^*(a | s) = \pi^{\text{ref}}(a | s) \text{Clip} \left(\exp \left(\frac{A(s, a) - \lambda(s)}{\beta} \right), [1 - \epsilon, 1 + \epsilon] \right),$$

where we write $A(s, a) = \mathbb{E}[\mathbf{A}(s, a) | s, a]$, and $\lambda(s)$ is a chosen such that $\sum_a \pi^*(a | s) = 1$ for each s .

See the proof in Appendix.

3 Proofs

Proof of Proposition 1.1. Note that

$$\text{Clip}(w, [1 - \epsilon, 1 + \epsilon])A = \text{Clip}(wA, [A - \epsilon|A|, A + \epsilon|A|]),$$

where we push A into the clip function.

Further, note that for $a \leq b$,

$$\min(x, \text{Clip}(x, [a, b])) = \min(x, b),$$

which shows that taking the minimum of x and its clipping to $[a, b]$ removes the lower bound a .

Hence, we obtain the simplified form by taking $x = wA$, $a = A - \epsilon|A|$ and $b = A + \epsilon|A|$:

$$\begin{aligned} R^{\text{clip}}(w, A) &= \min(wA, \text{Clip}(wA, [A - \epsilon|A|, A + \epsilon|A|])) \\ &= \min(wA, A + \epsilon|A|) \end{aligned}$$

□

Proof of Proposition 2.1. We have $\mathbf{A} = \max(\mathbf{A}, 0) + \min(\mathbf{A}, 0)$, and hence

$$\begin{aligned} \mathcal{R}^{\text{clip}}(w, \mathbf{A}) &= \mathbb{E} [\min(w, 1 + \epsilon) \max(\mathbf{A}, 0) + \max(w, 1 - \epsilon) \min(\mathbf{A}, 0)] \\ &= \min(w, 1 + \epsilon) \mathbb{E} [\max(\mathbf{A}, 0)] + \max(w, 1 - \epsilon) \mathbb{E} [\min(\mathbf{A}, 0)] \\ &= \min(w, 1 + \epsilon) \cdot \mu^+ + \max(w, 1 - \epsilon) \cdot \mu^-. \end{aligned}$$

We now get a simple concave three-piecewise linear function, and a simple case by case analysis gives the result. See Figure 1 for visualization. □

Proof of Proposition 2.2. The regularization objective function is

$$L(w) = \min(w, 1 + \epsilon)\mu^+ + \max(w, 1 - \epsilon)\mu^- - \phi(w).$$

Because $\mu^+ \geq 0$ and $\mu^- \leq 0$, this is a strictly concave function.

A key observation is that L cannot attain its optimum outside the interval $[1 - \epsilon, 1 + \epsilon]$. Indeed, we have $\nabla L(w) < 0$ for $w > 1 + \epsilon$ and $\nabla L(w) > 0$ for $w < 1 - \epsilon$, which forces the minimizer to lie within this interval:

1. When $w > 1 + \epsilon$, we have

$$\nabla L(w) = \mu^- - \nabla \phi(w) < \mu^- - \nabla \phi(1 + \epsilon) \leq -\nabla \phi(1 + \epsilon) < -\nabla \phi(1) = 0,$$

where we use that $\nabla \phi$ is strictly increasing and $\nabla \phi(1) = 0$ because 1 is the minimum of ϕ .

2. Similarly, when $w < 1 - \epsilon$, we have

$$\nabla L(w) = \mu^+ - \nabla \phi(w) > \mu^+ - \nabla \phi(1 - \epsilon) > \nabla \phi(1) = 0.$$

Therefore, the optimization is equivalent to the interval constrained optimization:

$$\max_{w \in \mathbb{R}} w \mathbb{E}[\mathbf{A}] - \phi(w) \quad s.t. \quad w \in [1 - \epsilon, 1 + \epsilon],$$

and it is standard to show that its optimal solution is $w^* = \text{Clip}(\mathbb{E}[\mathbf{A}], [1 - \epsilon, 1 + \epsilon])$.

□

Proof of Theorem 2.3. We will apply Proposition 2.2, with the complication that we need to handle the probability constraint. If we consider the constraint on w , it imposes

$$w(a, s) \geq 0, \quad \sum_a \pi^{\text{ref}}(a|s) w(a, s) = 1.$$

Because KL divergence involves the log function, which creates a barrier at $w = 0$ (that is, $\phi_{\text{KL}}(0) = +\infty$), it is not going to achieve $w = 0$ at optimum and we can safely drop it in the analysis. We only need to consider the normalization constraint. The Lagrangian is

$$L(\pi, \lambda) = \mathbb{E}_{s \sim d_0} \left[\mathbb{E}_{a \sim \pi^{\text{ref}}(\cdot|s)} \left[\mathcal{R}^{\text{clip}}(w(s, a), \mathbf{A}(s, a)) - \beta \phi_{\text{KL}}(w(s, a)) - \lambda(s)(w(s, a) - 1) \right] \right],$$

where $\lambda(s)$ is the Lagrangian multiplier for $\sum_a \pi(a|s) = 1$ for each s .

Using Proposition 2.2, for each (s, a) , the optimal value of $w(s, a)$ of the Lagrangian is obtained by

$$w^*(s, a) = \text{Clip} \left(\exp \left(\frac{A(s, a) - \lambda(s)}{\beta} \right), [1 - \epsilon, 1 + \epsilon] \right).$$

Plugging $w^*(s, a) = \frac{\pi^*(a|s)}{\pi^{\text{ref}}(a|s)}$ yields the result, and note that $\lambda(s)$ is chosen to ensure the normalization condition for each s . □

References

- [1] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.