# 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

$$J^{\mathrm{PPO}}(\pi) = \mathbb{E}_{s\sim d_0}\left[\mathbb{E}_{a\sim\pi^{\mathrm{ref}}(\cdot|s)}\Big[R^{\mathrm{clip}}(w(s,a), A(s,a))\Big] - \beta\,\mathrm{KL}\big(\pi(\cdot\mid s)\,\|\,\pi^{\mathrm{ref}}(\cdot\mid s)\big)\right],$$
$$\text{with}\qquad w(s,a) := \frac{\pi(a\mid s)}{\pi^{\mathrm{ref}}(a\mid s)},$$

where $\pi^{\mathrm{ref}}$ 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

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

where $\mathrm{Clip}(w, [1-\epsilon, 1+\epsilon]) = \min(\max(w, 1-\epsilon), 1+\epsilon)$ 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 $J^{\mathrm{PPO}}(\pi)$ is attained by a clipped exponential tilting of $\pi^{\mathrm{ref}}$:

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

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

## 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 $R^{\mathrm{clip}}(w, A)$ in (1) is equivalent to*

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

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 intution is as follows.

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

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

   This increases $\pi(a\mid s)$ 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 $w(s,a) \to \infty$ when $A(s,a) > 0$, and $w(s,a) \to 0$ 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.

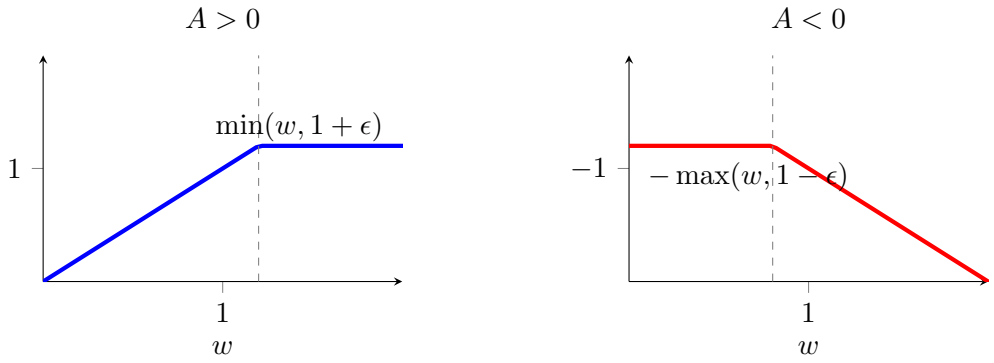



Figure 1: Clipped objective as a function of $w$ for positive and negative advantages.

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:*

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

*Then we have the identity:*

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

*where* $A^+ = \mathbb{E}[\max(\mathbf{A}, 0)]$ *and* $A^- = \mathbb{E}[\min(\mathbf{A}, 0)]$*, 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* $A^+ > 0$ *and* $A^- < 0$*, 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, maximizing the expected clip function yields the extreme solution $w^* = 1 + \text{sign}(\mathbb{E}[\mathbf{A}])\epsilon$.

**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 $\mathbb{E}[\mathbf{A}] = 0$.

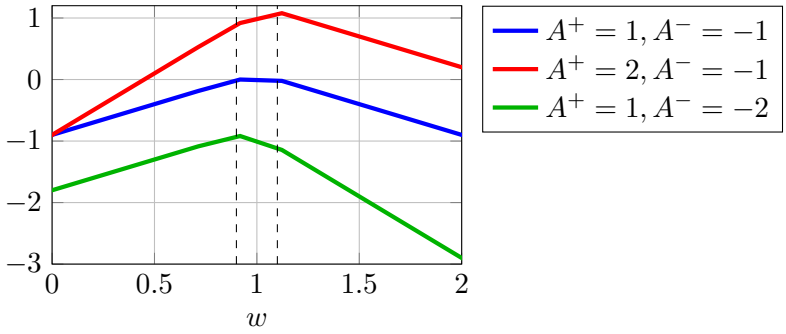



Figure 2: Plot of $\min(w, 1+\epsilon)A^+ + \max(w, 1-\epsilon)A^-$. Code here.

**Proposition 2.2.** *Assume $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differentiable strictly convex function, whose minimum is attained inside $[1-\epsilon, 1+\epsilon]$. Consider*

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

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

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

*and the optimum is attained by*

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

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

See the proof in Appendix.

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

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

where we assume a stochastic advantage $\mathbf{A}(s,a) = \mathbf{A}(s,a,\xi)$, with $\xi$ 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

$$\mathrm{KL}(\pi(\cdot \mid s)||\pi^{\mathrm{ref}}(\cdot \mid s)) = \sum_a \pi^{\mathrm{ref}}(a \mid s)\phi_{\mathrm{KL}}(w(a,s)),$$
$$\text{with} \qquad \phi_{\mathrm{KL}}(w) = w\log w - w + 1,$$

where $\phi_{\mathrm{KL}}$ 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 of $J^{\mathrm{PPO}}(\pi)$ is attained by a clipped exponentially tilted distribution.

**Theorem 2.3.** *Consider*

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

*where $\Delta$ is the set of policy distributions. Then optimal solution $\pi^*$ is obtained by*

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

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

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

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 $b = A + \epsilon\,|A|$:

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

□

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

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

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 regularized objective function is

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

Because $A^+ \ge 0$ and $A^- \le 0$, 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. To establish this, note that since $\phi$ attains its minimum at a point inside $[1-\epsilon, 1+\epsilon]$, we have $\nabla\phi(w) \ge 0$ for $w > 1+\epsilon$ and $\nabla\phi(w) \le 0$ for $w < 1-\epsilon$. Hence, we have:

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

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

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

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

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

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

$$\max_{w\in\mathbb{R}} w\mathbb{E}[\mathbf{A}] - \phi(w) \qquad s.t. \quad w \in [1-\epsilon, 1+\epsilon]. \tag{2}$$

This is a one dimensional interval constrained concave optimization. It is easy to show that its optimal solution is $w^* = \text{Clip}(\nabla\phi^{-1}(\mathbb{E}[\mathbf{A}]), \quad [1-\epsilon, 1+\epsilon])$. See Lemma 3.1 below. □

**Lemma 3.1.** *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a strictly convex function with unconstrained minimizer $x^0$. Then the solution to*

$$\min_x f(x) \qquad s.t. \quad x \in [a, b],$$

*where $a < b$, is $x^* = \text{Clip}(x^0, [a, b])$.*

*Proof.* Since $f$ is strictly convex, it has a unique unconstrained minimizer $x^0$. If $x^0 \in [a, b]$, then $x^* = x^0$ is optimal. If $x^0 < a$, then $f'(x) > 0$ for all $x \in [a, b]$ (by strict convexity), so $x^* = a$ minimizes $f$ on $[a, b]$. Similarly, if $x^0 > b$, then $x^* = b$. Hence $x^* = \text{Clip}(x^0, [a, b])$. □

*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

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

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

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

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 $s$. □

**Theorem 3.2.** *Consider the unregularized PPO optimization problem ($\beta = 0$):*

$$\min_{\pi \in \Delta} \mathbb{E}_{s\sim d_0}\left[\mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}\left[\mathcal{R}^{\text{clip}}(w(s,a), \mathbf{A}(s,a))\right]\right].$$

*For a given state $s$, let $\lambda(s)$ be the Lagrange multiplier associated with the normalization constraint $\sum_a \pi(a|s) = 1$. The optimal density ratio $w^*(s,a)$ is discrete and is given by:*

$$w^*(s,a) = \begin{cases} 0, & \text{if } \lambda(s) > \mu^+(s,a), \\ 1-\epsilon, & \text{if } A(s,a) \le \lambda(s) < \mu^+(s,a), \\ 1+\epsilon, & \text{if } \mu^-(s,a) \le \lambda(s) < A(s,a), \\ +\infty, & \text{if } \lambda(s) < \mu^-(s,a), \end{cases}$$

*where $A(s,a) = \mathbb{E}[\mathbf{A}(s,a)]$ is the expected advantage, and $\mu^+(s,a) = \mathbb{E}[\max(\mathbf{A}(s,a), 0)]$ and $\mu^-(s,a) = \mathbb{E}[\min(\mathbf{A}(s,a), 0)]$ are the expected positive and negative parts of the advantage.*

See the proof in Appendix.

*Proof of Theorem 3.2.* With $\beta = 0$, the Lagrangian for a fixed state-action pair $(s,a)$ is

$$L(w) = \mathcal{R}^{\text{clip}}(w, \mathbf{A}(s,a)) - \lambda(s)w,$$

subject to the non-negativity constraint $w \ge 0$. The function $\mathcal{R}^{\text{clip}}(w, \mathbf{A})$ is a concave, piecewise linear function. We maximize $L(w)$ by comparing the marginal gain (the subgradient of $\mathcal{R}^{\text{clip}}$) with the constant marginal cost $\lambda(s)$.

Recall the subgradients of the gain function $\mathcal{R}^{\text{clip}}(w, \mathbf{A})$ in its three linear segments:

1. For $w \in [0, 1-\epsilon)$, the slope is $\mu^+ = \mathbb{E}[\max(\mathbf{A}, 0)]$.
2. For $w \in (1-\epsilon, 1+\epsilon)$, the slope is $A = \mathbb{E}[\mathbf{A}]$.
3. For $w \in (1+\epsilon, \infty)$, the slope is $\mu^- = \mathbb{E}[\min(\mathbf{A}, 0)]$.

Note that by concavity, $\mu^+ \ge A \ge \mu^-$. We analyze the four regimes for $\lambda(s)$:

**Case 1: High Cost ($\lambda(s) > \mu^+$).** Here, the marginal cost $\lambda(s)$ strictly exceeds the marginal gain everywhere, including the steepest initial segment.

$$\partial_w L(w) \le \mu^+ - \lambda(s) < 0 \quad \forall w \ge 0.$$

The objective is strictly decreasing; hence the optimum is at the lower boundary $w^* = 0$.

**Case 2: Moderate Cost ($A \le \lambda(s) < \mu^+$).** Here, the gain exceeds the cost in the first segment ($\mu^+ > \lambda$), pushing $w$ up. However, in the second segment, the cost exceeds the gain ($\lambda \ge A$). The function increases up to $1-\epsilon$ and decreases thereafter.

$$w^* = 1 - \epsilon.$$

**Case 3: Low Cost ($\mu^- \le \lambda(s) < A$).** Here, the gain exceeds the cost in both the first and second segments ($\mu^+ > \lambda$ and $A > \lambda$), pushing $w$ past $1-\epsilon$. However, in the third segment (the tail), the cost exceeds the gain ($\lambda \ge \mu^-$). The function increases up to $1+\epsilon$ and decreases thereafter.

$$w^* = 1 + \epsilon.$$

**Case 4: Negative/Tiny Cost ($\lambda(s) < \mu^-$).** Here, the marginal gain exceeds the cost everywhere, even in the rightmost tail.

$$\partial_w L(w) \ge \mu^- - \lambda(s) > 0 \quad \forall w \ge 0.$$

The objective is unbounded and strictly increasing, implying $w^* \to +\infty$. (Note: In practice, $\lambda(s)$ is determined by normalization; if this case holds for all actions, the problem is ill-posed). □

**Proposition 3.3.** *Assume $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differentiable strictly convex function minimized at $w = 1$. Consider the non-negative optimization problem:*

$$\max_{w \ge 0} \mathbb{E}[\min(w\mathbf{A}, \quad \mathbf{A} + \epsilon\,|\mathbf{A}|)] - \phi(w) - \lambda w.$$

*Let $\rho = (\phi')^{-1}$ denote the inverse of the derivative of $\phi$. The optimal solution $w^*$ is given by:*

$$w^* = \begin{cases} \max\big(0,\ \rho(\mu^+ - \lambda)\big), & \text{if } \lambda > \mu^+ - \phi'(1-\epsilon) \quad \textit{(Left Tail)}, \\ \rho(\mu^- - \lambda), & \text{if } \lambda < \mu^- - \phi'(1+\epsilon) \quad \textit{(Right Tail)}, \\ \mathrm{Clip}\big(\rho(\mathbb{E}[\mathbf{A}] - \lambda),\ [1-\epsilon, 1+\epsilon]\big), & \textit{otherwise} \quad \textit{(Stable Regime)}. \end{cases} \tag{3}$$

*where $\mu^+ = \mathbb{E}[\max(\mathbf{A}, 0)]$ and $\mu^- = \mathbb{E}[\min(\mathbf{A}, 0)]$. (Note: We assume standard PPO setting where $1-\epsilon > 0$).*

See the proof in Appendix.

*Proof of Proposition 3.3.* The objective function is $L(w) = \mathcal{R}^{\text{clip}}(w, \mathbf{A}) - \phi(w) - \lambda w$. We maximize $L(w)$ subject to $w \ge 0$. Since $L$ is strictly concave, the solution is unique. We analyze the unconstrained gradient conditions and apply the domain constraints.

**Case 1: Left Tail Regime ($w < 1-\epsilon$).** This regime is active when the marginal cost at the boundary $1-\epsilon$ is too high for the solution to enter the interval:

$$\phi'(1-\epsilon) + \lambda > \mu^+ \iff \lambda > \mu^+ - \phi'(1-\epsilon).$$

In this region, the marginal gain is constant at $\mu^+$. The optimization becomes:

$$\max_{w \in [0, 1-\epsilon)} w\mu^+ - \phi(w) - \lambda w.$$

The unconstrained maximum of this sub-problem satisfies $\phi'(w) = \mu^+ - \lambda$, yielding candidate $w_{left} = \rho(\mu^+ - \lambda)$. Due to the constraint $w \ge 0$, the optimal solution is:

$$w^* = \max(0, w_{left}) = \max(0, \rho(\mu^+ - \lambda)).$$

(Note: $w^*$ will be 0 if $\lambda \geq \mu^+ - \phi'(0)$).

**Case 2: Right Tail Regime ($w > 1+\epsilon$).** This regime is active when the marginal cost at $1+\epsilon$ is too low:

$$\phi'(1+\epsilon) + \lambda < \mu^- \iff \lambda < \mu^- - \phi'(1+\epsilon).$$

Here the marginal gain is $\mu^-$. The candidate is $w_{right} = \rho(\mu^- - \lambda)$. Since we assume $1+\epsilon > 0$, and this regime implies $w_{right} > 1+\epsilon$, the constraint $w \geq 0$ is naturally satisfied.

$$w^* = \rho(\mu^- - \lambda).$$

**Case 3: Stable Regime ($w \in [1-\epsilon, 1+\epsilon]$).** This regime holds when $\lambda$ is between the tail thresholds. The unconstrained target is $w_{mid} = \rho(\mathbb{E}[\mathbf{A}] - \lambda)$. Since the optimal solution lies within $[1-\epsilon, 1+\epsilon]$ and we assume $1-\epsilon > 0$, the non-negativity constraint is satisfied automatically. The solution is simply clipped to the interval bounds:

$$w^* = \text{Clip}(w_{mid}, [1-\epsilon, 1+\epsilon]).$$

□

**Theorem 3.4** (Global Optimum of PPO Objective)**.** *Consider the optimization problem*

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

*where $\Delta$ is the set of valid policy distributions. Let $\mu^+(s,a) = \mathbb{E}[\max(\mathbf{A}(s,a), 0)]$ and $\mu^-(s,a) = \mathbb{E}[\min(\mathbf{A}(s,a), 0)]$. The optimal policy is given by $\pi^*(a \mid s) = \pi^{\text{ref}}(a \mid s) w^*(s,a)$, where the density ratio $w^*(s,a)$ is determined by the normalization constant $\lambda(s)$ according to three regimes:*

$$w^*(s,a) = \begin{cases} \exp\left(\frac{\mu^+(s,a) - \lambda(s)}{\beta}\right), & \text{if } \lambda(s) > \mu^+(s,a) - \beta \log(1-\epsilon), \\ \exp\left(\frac{\mu^-(s,a) - \lambda(s)}{\beta}\right), & \text{if } \lambda(s) < \mu^-(s,a) - \beta \log(1+\epsilon), \\ \text{Clip}\left(\exp\left(\frac{A(s,a) - \lambda(s)}{\beta}\right), [1-\epsilon, 1+\epsilon]\right), & \text{otherwise}, \end{cases}$$

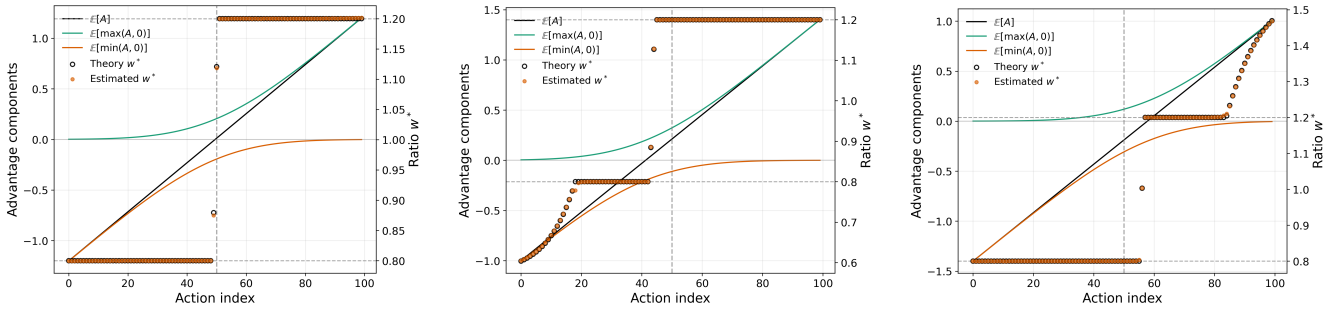
*where $A(s,a) = \mathbb{E}[\mathbf{A}(s,a)]$ is the expected advantage, and $\lambda(s)$ is the unique scalar satisfying $\sum_a \pi^*(a \mid s) = 1$.*

Figure 3 provides a toy numerical illustration of the theorem, plotting both the Gaussian-smoothed advantage components and the corresponding optimal density ratio $w^*$ against the *action index* under three mean-advantage scenarios.

See the proof in Appendix.

*Proof of Theorem 3.4.* The PPO objective with the KL penalty can be decomposed state-wise. For a fixed state $s$, we maximize the Lagrangian with respect to the density ratio $w(s,a) \overset{\text{def}}{=} \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)}$. The Lagrangian for state $s$ (neglecting the constant expectation over $d_0$) is:

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




**(a)** $A(s,a) = a$ **(b)** $A(s,a) = a + 0.2$ **(c)** $A(s,a) = a - 0.2$

Figure 3: **Toy illustration of Theorem 3.4 with a unified action-index axis.** We use a uniform reference policy $\pi^{\text{ref}}$ and PPO parameters $\beta = 0.1$, $\epsilon = 0.2$ (with Gaussian noise of standard deviation $\sigma = 0.5$ in the computation of $\mu^{\pm}$). In each panel, the left axis plots $\mathbb{E}[A]$, $\mu^+ = \mathbb{E}[\max(A,0)]$, and $\mu^- = \mathbb{E}[\min(A,0)]$, while the right axis overlays the optimal density ratio $w^*$: theory (open circles) and numerical PPO optimum (filled dots). The dashed vertical line marks the midpoint action index ($x = 50$). The three cases correspond to $A(s,a) = a$, $A(s,a) = a + 0.2$, and $A(s,a) = a - 0.2$, where $a$ is a linear grid over actions in $[-1.2, 1.2]$.

where $\phi_{\text{KL}}(w) = w \log w - w + 1$. Since the optimization is pointwise over actions, for each $(s,a)$ we solve:

$$\max_{w \geq 0} \ \mathbb{E}_\xi \left[\min(w\mathbf{A}, \mathbf{A} + \epsilon|\mathbf{A}|)\right] - \beta\phi_{\text{KL}}(w) - \lambda(s)w.$$

This matches the form of **Proposition 3.3**, with the specific penalty function $\phi(w) = \beta\phi_{\text{KL}}(w)$. We compute the necessary derivatives and inverses:

- The derivative of the penalty is $\phi'(w) = \beta \frac{d}{dw}(w \log w - w) = \beta \log w$.
- The inverse derivative function is $\rho(y) = (\phi')^{-1}(y) = \exp(y/\beta)$.

Substituting these into the three-case solution formula from Proposition 3.3:

1. **Left Tail Regime:** Occurs when $\lambda(s) > \mu^+(s,a) - \beta \log(1-\epsilon)$. The solution is $w^* = \rho(\mu^+ - \lambda) = \exp\left(\frac{\mu^+(s,a) - \lambda(s)}{\beta}\right)$. (Note: Since $\exp(\cdot) > 0$, the non-negativity constraint is satisfied).

2. **Right Tail Regime:** Occurs when $\lambda(s) < \mu^-(s,a) - \beta \log(1+\epsilon)$. The solution is $w^* = \rho(\mu^- - \lambda) = \exp\left(\frac{\mu^-(s,a) - \lambda(s)}{\beta}\right)$.

3. **Stable Regime:** Occurs when $\lambda(s)$ lies between the bounds. The solution is the clipped version of $\rho(\mathbb{E}[\mathbf{A}] - \lambda)$:

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

Thus, the optimal policy is a mixture of clipped and unclipped exponential tiltings, governed by the magnitude of the normalization constant $\lambda(s)$ relative to the tail slopes $\mu^+$ and $\mu^-$. □

**Theorem 3.5.** *Consider the unregularized PPO optimization problem ($\beta = 0$). For a given state $s$, let $\lambda(s)$ be the Lagrange multiplier for the normalization constraint. The optimal density ratio $w^*(s,a)$ is given by:*

$$w^*(s,a) \in \begin{cases} \{0\}, & \text{if } \lambda(s) > \mu^+(s,a), \\ \{1-\epsilon\}, & \text{if } A(s,a) < \lambda(s) \le \mu^+(s,a), \\ [1-\epsilon, 1+\epsilon], & \text{if } \lambda(s) = A(s,a), \\ \{1+\epsilon\}, & \text{if } \mu^-(s,a) \le \lambda(s) < A(s,a), \\ \{+\infty\}, & \text{if } \lambda(s) < \mu^-(s,a). \end{cases}$$

See the proof in Appendix.

*Proof of Theorem 3.5.* We maximize the Lagrangian $L(w) = \mathcal{R}^{\text{clip}}(w, \mathbf{A}) - \lambda w$. The optimality condition requires that the slope of the cost, $\lambda$, belongs to the subgradient set of the gain function $\mathcal{R}^{\text{clip}}$.

Recall the subgradients of $\mathcal{R}^{\text{clip}}$:

- $w \in [0, 1-\epsilon)$: slope is $\mu^+$.
- $w \in (1-\epsilon, 1+\epsilon)$: slope is $A$.
- $w \in (1+\epsilon, \infty)$: slope is $\mu^-$.

We analyze the case where the cost exactly balances the expected advantage:

**Case: Balanced Cost ($\lambda(s) = A(s,a)$).** In the middle interval $w \in (1-\epsilon, 1+\epsilon)$, the marginal gain is $A$ and the marginal cost is $\lambda = A$. The net derivative is $L'(w) = A - \lambda = 0$. Since the derivative is zero throughout this entire open interval, the objective function is constant (flat) in this region. Checking the boundaries:

- To the left ($w < 1-\epsilon$), the slope is $\mu^+ - A \ge 0$ (objective rises or stays flat up to $1-\epsilon$).
- To the right ($w > 1+\epsilon$), the slope is $\mu^- - A \le 0$ (objective falls or stays flat after $1+\epsilon$).

Thus, the maximum is attained at every point in the closed interval $[1-\epsilon, 1+\epsilon]$.

The other cases follow the strict inequality logic derived previously (e.g., if $\lambda > A$, the cost exceeds the gain in the middle segment, pushing the optimum to the left boundary $1-\epsilon$). □

## References

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