The Clipping Function in PPO

Qiang Liu

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

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

where π^{ref} is the previous policy, w is the density ratio, d_0 is the state distribution, and $\tilde{A}(s,a)$ is a stochastic advantage signal, which may depend on additional randomness conditioned on (s,a). The PPO clipping function is defined as

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

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 often intuitively stated that the clipped surrogate objective induces a clipped density ratio at the optimal solution. This intuition is only partially correct: the optimal density ratio is indeed truncated, but in a subtle, state- and action-dependent manner.

Specifically, we give an explict formula of the global maximizer of $J^{\mathrm{PPO}}(\pi)$ as follows:

$$\pi^*(a\mid s) = \pi^{\mathrm{ref}}(a\mid s)\,w^*_\lambda(s,a),$$

where the optimal density ratio $w^*(s,a)$ takes the form of an *adaptive, leaky* clipping of an exponential tilt,

$$w^*_\lambda(s,a) = \mathrm{Clip}\Bigg(w_{\mathrm{mid}},\ \ \underbrace{\min\big(1-\epsilon,\,w^+_{\mathrm{tail}}\big)}_{\text{dynamic floor}},\ \ \underbrace{\max\big(1+\epsilon,\,w^-_{\mathrm{tail}}\big)}_{\text{dynamic ceiling}}\Bigg). \tag{2}$$

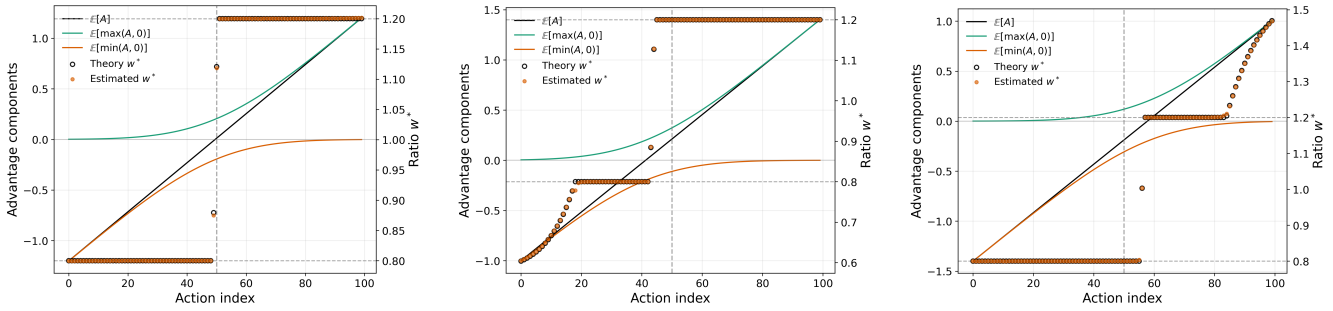
Here,

$$w_{\mathrm{mid}} = \exp\left(\frac{A(s,a)-\lambda(s)}{\beta}\right),\quad w^+_{\mathrm{tail}} = \exp\left(\frac{A^+(s,a)-\lambda(s)}{\beta}\right),\quad w^-_{\mathrm{tail}} = \exp\left(\frac{A^-(s,a)-\lambda(s)}{\beta}\right),$$

with

$$A(s,a) = \mathbb{E}[\mathbf{A}(s,a)\mid s,a],\qquad A^+(s,a) = \mathbb{E}[\max(\mathbf{A}(s,a),0)\mid s,a],\qquad A^-(s,a) = \mathbb{E}[\min(\mathbf{A}(s,a),0)\mid s,a].$$

The scalar $\lambda(s)$ is a normalization constant chosen so that $\sum_a \pi^*(a\mid s) = 1$ for each state s.



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

Figure 1: Illustration of (3) with a uniform reference π^{ref}. We model a noisy advantage signal by $\mathbf{A}(s,a) \sim \mathcal{N}(A(s,a), \sigma^2)$ with $\sigma = 0.2$, where $A(s,a)$ is linear in a, and set $\beta = 0.1$. Panel (a) shows a setting with no "leaky" clipping, while panels (b) and (c) demonstrate that a constant shift of the advantage can induce the leaky behavior.

Thus, $\pi^*(\cdot \mid s)$ is an exponentially tilted version of $\pi^{\text{ref}}(\cdot \mid s)$, with density ratio $w_{\text{mid}} = \exp\big((A(s,a) - \lambda(s))/\beta\big)$ that is *adaptively* truncated. Unlike the common intuition that PPO simply clips w_{mid} to $[1-\epsilon,\, 1+\epsilon]$, the effective clipping interval is

$$\Big[\min\big(1-\epsilon,\, w^+_{\text{tail}}\big),\ \max\big(1+\epsilon,\, w^-_{\text{tail}}\big)\Big],$$

whose (state-, action-dependent) floor and ceiling depend on the positive and negative truncated means $A^+(s,a)$ and $A^-(s,a)$ of the stochastic advantage $\mathbf{A}(s,a)$. In particular, when $w^+_{\text{tail}} \geq 1-\epsilon$ and $w^-_{\text{tail}} \leq 1+\epsilon$, this reduces to the standard clipping of w_{mid} to $[1-\epsilon,\, 1+\epsilon]$; otherwise the constraint becomes "leaky" and the allowable range expands accordingly. This more subtle behavior arises from the interaction between the clipping objective and the policy normalization constraint $\sum_a \pi(a \mid s) = 1$.

This structure implies a loss of shift invariance with respect to the advantage: replacing $\mathbf{A}(s,a)$ by $\mathbf{A}(s,a) + c$ for a constant c generally changes the optimal policy. This contrasts with classical policy gradient objectives, where constant shifts of the advantage serve as baselines and leave the policy update unchanged.

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*

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

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

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

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.

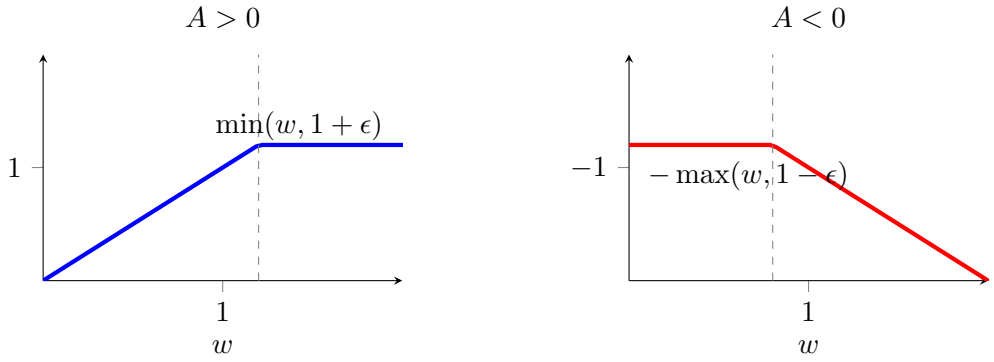


Figure 2: Clipped objective as a function of w for positive and negative advantages.

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

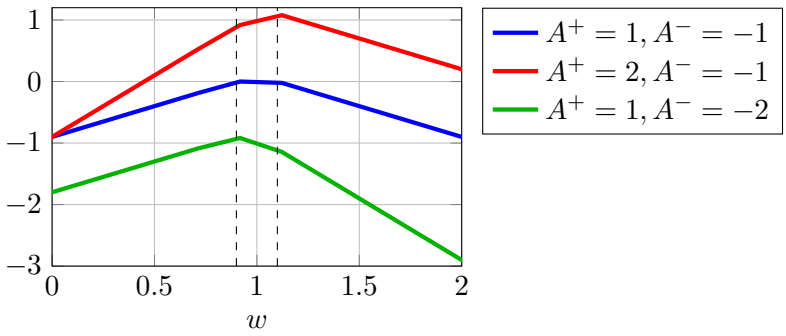


Figure 3: Plot of $\min(w, 1+\epsilon)A^+ + \max(w, 1-\epsilon)A^-$. Code here.

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 ϕ.

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 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

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

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

The key complication is that the policy probabilities must satisfy the normalization constraint $\sum_a \pi^*(a \mid s) = 1$ for each s. When this constraint cannot be satisfied while clipping to $[1-\epsilon, 1+\epsilon]$, the optimum instead exhibits the more subtle "leaky" clipping behavior in (3).

Theorem 2.3 (Global Optimum of PPO Objective)**.** *Consider the optimization problem*

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

where Δ is the set of valid policy distributions, and the KL penalty coefficient is positive ($\beta > 0$).

The optimal policy is given by

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

where the density ratio $w^*(s,a)$ is determined by a normalization constant $\lambda(s)$ according to

$$w^*(s,a) = \mathrm{Clip}\Bigg(w_{\mathrm{mid}}, \quad \underbrace{\min(1-\epsilon,\ w^+_{\mathrm{tail}})}_{\textit{Dynamic Floor}}, \quad \underbrace{\max(1+\epsilon,\ w^-_{\mathrm{tail}})}_{\textit{Dynamic Ceiling}}\Bigg), \tag{3}$$

where

$$w_{\mathrm{mid}} = \exp\left(\frac{A(s,a)-\lambda(s)}{\beta}\right), \quad w^+_{\mathrm{tail}} = \exp\left(\frac{A^+(s,a)-\lambda(s)}{\beta}\right), \quad w^-_{\mathrm{tail}} = \exp\left(\frac{A^-(s,a)-\lambda(s)}{\beta}\right),$$

and $A(s,a) = \mathbb{E}[\mathbf{A}(s,a) \mid s,a]$ and $A^+(s,a) = \mathbb{E}[\max(\mathbf{A}(s,a),0) \mid s,a]$ and $A^-(s,a) = \mathbb{E}[\min(\mathbf{A}(s,a),0) \mid s,a]$, and $\lambda(s)$ is a scalar chosen to satisfy $\sum_a \pi^*(a \mid s) = 1$ for each s.

See the proof in Appendix.

Remark A more explicit formula is

$$w^*(s,a) = \begin{cases} w^+_{\mathrm{tail}} & \text{if } w^+_{\mathrm{tail}} < 1-\epsilon \\ w^-_{\mathrm{tail}} & \text{if } w^-_{\mathrm{tail}} > 1+\epsilon \\ \mathrm{Clip}(w_{\mathrm{mid}}, \ \ [1-\epsilon, 1+\epsilon]) & \text{otherwise} \end{cases}$$

Remark For each state s, the constant $\lambda(s)$ is defined as a solution to

$$Z(\lambda(s)) = 1, \qquad Z(\lambda(s)) := \sum_a \pi^{\mathrm{ref}}(a \mid s)\, w^*_\lambda(s,a),$$

where

$$w^*_\lambda(s,a) = \begin{cases} \exp\left(\dfrac{A^+(s,a)-\lambda(s)}{\beta}\right), & \text{if } \lambda(s) > A^+(s,a) - \beta\log(1-\epsilon), \\ \exp\left(\dfrac{A^-(s,a)-\lambda(s)}{\beta}\right), & \text{if } \lambda(s) < A^-(s,a) - \beta\log(1+\epsilon), \\ \mathrm{Clip}\left(\exp\left(\dfrac{A(s,a)-\lambda(s)}{\beta}\right), [1-\epsilon,\ 1+\epsilon]\right), & \text{otherwise.} \end{cases} \tag{4}$$

Lemma 2.4. *For each fixed state s, the function $Z(\lambda(s))$ is continuous and monotonically non-increasing in $\lambda(s)$, with*

$$Z(+\infty) = 0, \qquad Z(-\infty) = +\infty.$$

Hence, there exists at least one solution to $Z(\lambda(s)) = 1$.

Moreover, although $\lambda(s)$ may not be unique, the corresponding weight $w^*_\lambda(s,a)$ is unique.

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 2 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 ϕ 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 ϕ.

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:

$$\max_{w\in\mathbb{R}} w\mathbb{E}[\mathbf{A}] - \phi(w) \qquad s.t. \quad w \in [1-\epsilon, 1+\epsilon]. \tag{5}$$

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 $x^* = \text{Clip}(x^0, [a, b])$. □

Lemma 3.2. *Assume $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differentiable strictly convex function minimized at $w = 1$. Consider the non-negative optimization problem:*

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

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

$$w^*(s,a) = \text{Clip}\left(w_{mid}, \ \underbrace{\max(0, \ \min(1-\epsilon, \ w^+_{\text{tail}}))}_{\textit{Dynamic Floor}}, \ \underbrace{\max(1+\epsilon, \ w^-_{\text{tail}})}_{\textit{Dynamic Ceiling}}\right),$$

where

$$w_{mid} = (\phi')^{-1}(A - \lambda), \qquad w^+_{\text{tail}} = (\phi')^{-1}(A^+ - \lambda), \qquad w^-_{\text{tail}} = (\phi')^{-1}(A^- - \lambda),$$

and $A^+ = \mathbb{E}[\max(\mathbf{A}, 0)]$ and $A^- = \mathbb{E}[\min(\mathbf{A}, 0)]$, and $A = \mathbb{E}[\mathbf{A}]$.

Explicitly, we have

$$w^* = \begin{cases} \max\big(0, \ (\phi')^{-1}(A^+ - \lambda)\big), & \text{if } \lambda > A^+ - \phi'(1-\epsilon) \quad \textit{(Left Tail)}, \\ (\phi')^{-1}(A^- - \lambda), & \text{if } \lambda < A^- - \phi'(1+\epsilon) \quad \textit{(Right Tail)}, \\ \text{Clip}\big((\phi')^{-1}(A - \lambda), \ [1-\epsilon, 1+\epsilon]\big), & \textit{otherwise} \quad \textit{(Stable Regime)}. \end{cases} \tag{6}$$

See the proof in Appendix.

Proof of Proposition 3.2. The objective function is $L(w) = \mathcal{R}^{\text{clip}}(w, \mathbf{A}) - \phi(w) - \lambda w$, where $r(w) = \mathcal{R}^{\text{clip}}(w, \mathbf{A}) = \min(w, 1+\epsilon)A^+ + \max(w, 1-\epsilon)A^-$ is a concave, piecewise linear function:

$$r(w) = \begin{cases} wA^+ + (1-\epsilon)A^- & \text{if } w < 1-\epsilon \\ wA & \text{if } w \in (1-\epsilon, 1+\epsilon) \\ wA^- + (1+\epsilon)A^+ & \text{if } w > 1+\epsilon, \end{cases}$$

for which we should note that $A^- \le A \le A^+$, and $A^- \le 0 \le A^+$, and $A = A^+ + A^-$. We maximize $L(w)$ subject to $w \ge 0$. Since L is strictly concave, the solution is unique.

We analyze the different cases based on the location of the optimum.

Case 1: Optimum in $[0, 1-\epsilon)$. The maximum of $L(w)$ belongs to the left tail $[0, 1-\epsilon)$ if all the subgradients of $L(w)$ at $w = 1-\epsilon$ is less than zero. This means that

$$\partial L(w) \le A^+ - \phi'(1-\epsilon) - \lambda < 0 \iff \lambda > A^+ - \phi'(1-\epsilon).$$

In this case, the optimization becomes equivalent to:

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

The unconstrained maximum of this sub-problem satisfies $\phi'(w) = A^+ - \lambda$, yielding candidate $w^+_{\text{tail}} = (\phi')^{-1}(A^+ - \lambda)$. Due to the constraint $w \ge 0$, the optimal solution is:

$$w^* = \max(0, w^+_{\text{tail}}) = \max(0, (\phi')^{-1}(A^+ - \lambda)).$$

Case 2: Optimum in $(1+\epsilon, +\infty)$**).** The maximum of $L(w)$ follows into the right regime $(1+\epsilon, +\infty)$, if all the subgradients of $L(w)$ at $w = 1+\epsilon$ are larger than zero:

$$\partial L(w) \ge A^- - \phi'(1+\epsilon) - \lambda > 0 \iff \lambda < A^- - \phi'(1+\epsilon).$$

Here the marginal gain is A^-. The candidate is $w^-_{\text{tail}} = (\phi')^{-1}(A^- - \lambda)$. Since we assume $1+\epsilon > 0$, and this regime implies $w^-_{\text{tail}} > 1+\epsilon$, the constraint $w \ge 0$ is naturally satisfied.

$$w^* = (\phi')^{-1}(A^- - \lambda).$$

Case 3: Optimum in ($[1-\epsilon, 1+\epsilon]$**).** Otherwise, the maximum of $L(w)$ belongs to $[1-\epsilon, 1+\epsilon]$. In this case, the problem reduces to

$$\max_w wA - \phi(w) - \lambda w \quad s.t. \quad w \in [1-\epsilon, 1+\epsilon].$$

The solution is simply clipped to the interval bounds:

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

□

Proof of Theorem 2.3. We solve the optimization in terms of w:

$$\max_w \mathbb{E}_{a\sim\pi^{\text{ref}}}\left[\mathcal{R}^{\text{clip}}(w(s,a), \mathbf{A}(s,a)) - \beta\phi_{\text{KL}}(w(s,a))\right], \qquad w(a,s) \ge 0, \qquad \sum_a \pi^{\text{ref}}(a|s)w(a,s) = 1.$$

where $\phi_{\text{KL}}(w) = w\log w - w + 1$ is the penalty used in KL divergence. Because ϕ_{KL} 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 3.2, for each (s, a), the optimal value of $w(s, a)$ of the Lagrangian is obtained by

$$w^*(s,a) = \text{Clip}\Bigg(w_{mid},\ \underbrace{\max(0,\ \min(1-\epsilon,\ w^+_{\text{tail}}))}_{\text{Dynamic Floor}},\ \underbrace{\max(1+\epsilon,\ w^-_{\text{tail}})}_{\text{Dynamic Ceiling}}\Bigg),$$

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.

□

Proof of Lemma 2.4. We argue in several steps.

1. Continuity. From (3), the function $w^*_\lambda(s,a)$ is obtained by composing min, max, and Clip operations with smooth exponential functions. Since min, max, and Clip are continuous, $w^*_\lambda(s,a)$ is continuous in $\lambda(s)$.

Alternatively, one may inspect the piecewise representation in (4) and verify that the values of the different branches agree at the switching thresholds, which again implies continuity.

2. Monotonicity. Each branch of (4) is monotonically non-increasing in $\lambda(s)$. Since there are no jumps at the transition points, $w^*_\lambda(s,a)$ is globally monotonically non-increasing. Consequently, $Z(\lambda(s))$, being a non-negative weighted sum of such terms, is also monotonically non-increasing.

3. Limit as $\lambda(s) \to +\infty$. If

$$\lambda(s) > \max_a \big(A^+(s,a) - \beta \log(1-\epsilon)\big),$$

then

$$w^*_\lambda(s,a) = \exp\left(\frac{A^+(s,a) - \lambda(s)}{\beta}\right),$$

which implies $Z(\lambda(s)) \to 0$ as $\lambda(s) \to +\infty$.

4. Limit as $\lambda(s) \to -\infty$. If

$$\lambda(s) < \min_a \big(A^-(s,a) - \beta \log(1+\epsilon)\big),$$

then

$$w^*_\lambda(s,a) = \exp\left(\frac{A^-(s,a) - \lambda(s)}{\beta}\right),$$

which implies $Z(\lambda(s)) \to +\infty$ as $\lambda(s) \to -\infty$.

The intermediate value theorem yields existence of a solution to $Z(\lambda(s)) = 1$.

Finally, suppose there exist two solutions $\lambda(s)$ and $\lambda'(s)$ satisfying $Z(\lambda(s)) = Z(\lambda'(s)) = 1$, with $\lambda(s) > \lambda'(s)$. By monotonicity of $w^*_\lambda(s,a)$ in $\lambda(s)$, we have

$$w^*_\lambda(s,a) \le w^*_{\lambda'}(s,a) \qquad \text{for all } a.$$

Since $\pi^{\text{ref}}(a \mid s) > 0$ and

$$\sum_a \pi^{\text{ref}}(a \mid s)\, w^*_\lambda(s,a) = \sum_a \pi^{\text{ref}}(a \mid s)\, w^*_{\lambda'}(s,a) = 1,$$

the above inequality can hold only if

$$w^*_\lambda(s,a) = w^*_{\lambda'}(s,a) \qquad \text{for all } a.$$

Therefore, although the normalization equation may admit multiple solutions $\lambda(s)$, the induced weight $w^*_\lambda(s,a)$ is unique. □

Theorem 3.3. *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:*

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

See the proof in Appendix.

Proof of Theorem 3.3. 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, λ, 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 A^+.
- $w \in (1-\epsilon, 1+\epsilon)$: slope is A.
- $w \in (1+\epsilon, \infty)$: slope is A^-.

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