# FOLLOW THE IMPROVEMENT: GUARANTEED POLICY OPTIMIZATION BY TRACKING PROXIMAL ASCENT TARGETS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern policy optimization approaches are mainly built on two theoretical paradigms: constrained optimization and tracking a target policy. Constrained methods (PPO, TRPO) maximize a surrogate objective within a trust region, but their monotonic improvement guarantees hold only for an idealized procedure, not for its practical implementation. Target-tracking methods (SAC, AWR, MPO) construct a target with improved performance and approach it by minimizing a regression loss; a small tracking loss however does not guarantee improvement. We propose *Follow-the-Improvement RL (FIRL)*, a framework that *unifies* these two main classes of policy optimization paradigms. First, FIRL defines a general proximal ascent target, $\pi^\star = (1 + \epsilon g)\,\pi_0$, where $g$ aligns with the improvement direction and has constrained scale. Second, FIRL characterizes a class of *tracking loss functions* that can bound the performance gap between policies. Then we show that, with a constrained step size, minimizing idealized tracking losses toward the target $\pi^*$, evaluated under the data distribution from $\pi_0$, is sufficient to *guarantee performance improvement* over $\pi_0$. The framework unifies PPO, SPO, and on-policy variants of AWR and MPO, while identifying *new promising instances* motivated by the theory: a clipped-linear target paired with advantage-weighted total variation/hinge (ATV/AH) losses. Across MuJoCo, IsaacLab, Atari, and LLM fine-tuning on mathematical reasoning, we evaluate different design combinations, show that the proposed instances perform competitively, and identify more effective target–loss combinations for each task category.

## 1 INTRODUCTION

Deep reinforcement learning (DRL) has demonstrated success in games, robotics, and language-model post-training (Mnih et al., 2015; Silver et al., 2016; Ouyang et al., 2022). As a classical approach, policy gradient methods learn a policy by updating policy parameters through gradient ascent on the expected return (Williams, 1992; Sutton et al., 1999). To make efficient use of costly interaction data, practical algorithms often perform *multiple optimization steps* on each collected batch (Schulman et al., 2017; Haarnoja et al., 2018). A central question is whether continuing to optimize a training objective on such data leads to a policy with higher expected return. The literature offers two perspectives on this question: i) controlling the range of policy updates directly, and ii) separating the construction of a target policy from fitting a parameterized policy to it.

The first perspective, *constrained optimization*, restricts each update to a trust region around the behavior policy (the policy that collects the data). Representative methods include conservative policy iteration, Trust Region Policy Optimization (TRPO), and Proximal Policy Optimization (PPO) (Kakade & Langford, 2002; Schulman et al., 2015; 2017). However, TRPO's implementation does not follow the objective prescribed by its theoretical analysis, which is too conservative in practice. Further, PPO's clipping alone does not enforce a trust region (Schulman et al., 2015; Wang et al., 2020). Mirror Learning unifies trust region approaches, but its convergence results assume exact policy-space maximization of the objective (Kuba et al., 2022), which practical implementations can only approximate. These gaps motivate guarantees on improving returns while accounting for residual optimization and approximation errors in parameterized policies.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

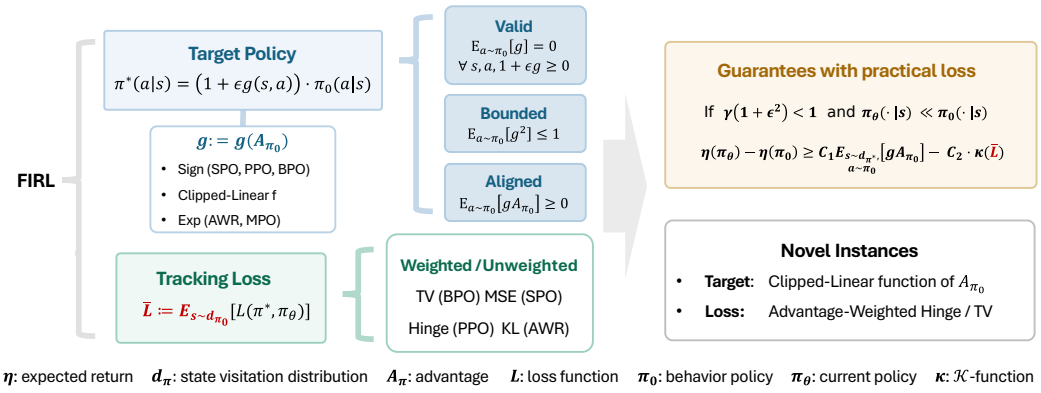



Figure 1: **Overview of the framework.** (1) *Target policies:* we consider proximal ascent perturbations $g$ of the behavior policy, which can have many forms (the sign, a clipped-linear, exponential of the advantage), as long as $g$ satisfies valid, bounded, and aligned conditions. (2) *Tracking loss:* the policy $\pi_\theta$ is trained by minimizing a tracking loss to the analytic target, e.g., Total Variation (TV), $\chi^2$(MSE)/KL divergence, or Hinge loss, evaluated under the behavior data $d_{\pi_0}$. (3) *Guarantees:* under mild assumptions, the policy improvement is lower-bounded by the analytic improvement from the *target* minus a monotonic function of the *tracking loss*. (4) *New instances*: The framework identifies novel target + tracking loss combinations, motivated by our theoretical analysis.

The second perspective, *target tracking*, constructs a target policy and fits a parameterized policy to it. In Soft Actor-Critic (SAC), Advantage-Weighted Regression (AWR), and Maximum a Posteriori Policy Optimisation (MPO), targets are constructed as analytic optimal solutions to intermediate value- or advantage-based objectives with regularization or constraints (Haarnoja et al., 2018; Peng et al., 2019; Abdolmaleki et al., 2018). With exact advantage or action-value estimation, such targets can ensure nondecreasing performance (Schulman et al., 2015; Haarnoja et al., 2018). They then fit the parameterized policies by minimizing a *tracking loss* to the target, such as Kullback–Leibler (KL) divergence. However, errors at rarely visited states can erase the target's gain (Chan et al., 2022). Some analyses address this mismatch through coverage assumptions, but they are difficult to verify in practice (Rashidinejad et al., 2021; Xie et al., 2021).

We propose *follow-the-improvement RL* (FIRL), a unifying framework that connects the two perspectives and addresses these gaps. We construct *a target tilt* from the behavior policy that is *i)* within a trust region and *ii)* aligned with its advantage. We then train a parameterized policy to match this target by minimizing a *tracking loss* on the collected data. We identify explicit conditions on the target and loss that yield guaranteed performance improvement.

Our contributions are:

- **A unifying framework.** FIRL organizes on-policy RL by two design choices: a *proximal ascent target* and a *tracking loss* (Figure 1). These choices connect the two perspectives, and cover existing algorithms, including PPO, Simple Policy Optimization (SPO, Xie et al., 2024), Bounded Policy Optimization (BPO, Ao et al., 2026), and on-policy variants of AWR and MPO.
- **Guarantees after imperfect fitting.** For admissible targets and losses, we lower-bound the learned policy's improvement by the target's gain minus a monotonic function of an idealized tracking loss under *the behavior state distribution*. The guarantee is thus stated in terms of practically accessible training loss and allows imperfect fitting.
- **New target–loss combinations.** Motivated by the theoretical analysis, we identify new promising target–loss combinations: **clipped, standardized advantage** target tilt with **advantage-weighted total variation** and **advantage-weighted hinge** losses.
- **Empirical validation.** Extensive experiments and ablations on MuJoCo, IsaacLab, Atari, and LLM fine-tuning benchmarks evaluate the new instances against variations along the target and loss axes, and identify more effective combinations for each task category.

## 2 PRELIMINARIES

**Markov decision processes.** We consider an infinite-horizon discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, d_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is the reward function, $d_0$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. A stationary policy $\pi(a \mid s)$ maps each state to a distribution over actions. A policy interacts with the MDP to generate a trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$, where $s_0 \sim d_0$, $a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. We write $\mathbb{P}^\pi$ and $\mathbb{E}^\pi$ for the probability law and expectation over trajectories induced by $\pi$. The *(normalized) discounted state occupancy measure* of a policy $\pi$ summarizes how often it visits each state $d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}^\pi(s_t = s)$ which is a probability distribution over $\mathcal{S}$.

**Value functions.** The *state value function* $V_\pi(s) = \mathbb{E}^\pi[\sum_{t\geq 0} \gamma^t r(s_t, a_t) \,|\, s_0 = s]$ is the expected discounted return starting from a state, and the *action value function* $Q_\pi(s, a) = r(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_\pi(s')]$ additionally fixes the first action. The *advantage* $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$ measures the additional value of action $a$ relative to following $\pi$ at state $s$, and satisfies $\mathbb{E}_{a\sim\pi(\cdot|s)}[A_\pi(s, a)] = 0$ at every state.

**Policy performance.** The performance of a policy is its expected return from the initial distribution, $\eta(\pi) = \mathbb{E}_{s\sim d_0}[V_\pi(s)]$. The *performance difference lemma* (Kakade & Langford, 2002) expresses the gap between any two policies $\pi$ and $\pi_0$ exactly as an expected advantage,

$$(1-\gamma)\big(\eta(\pi) - \eta(\pi_0)\big) \;=\; \mathbb{E}_{s\sim d_\pi,\, a\sim\pi(\cdot|s)}[A_{\pi_0}(s, a)]\,. \tag{1}$$

Throughout, $\pi_0$ denotes the behavior policy, $\pi_\theta$ the learned policy with parameter $\theta$, and $\pi^\star$ the target policy that $\pi$ is trained to track; $\rho_\theta$ denotes $\pi_\theta/\pi_0$ and $\rho^*$ denotes $\pi^*/\pi_0$.

## 3 FOLLOW-THE-IMPROVEMENT RL FRAMEWORK

### 3.1 OVERVIEW

The overview of the FIRL framework is illustrated in Figure 1. The framework follows the iterative policy optimization procedure of PPO: within each iteration, it first collects data with the old policy $\pi_0$, then optimizes $\pi_\theta$ based on that data, and finally assigns $\pi_0 \leftarrow \pi_\theta$. To optimize $\pi_\theta$, we first define a general **target policy** $\pi^*$ (Section 3.2) with a small multiplicative tilt $g$ from the behavior policy $\pi^* = (1 + \epsilon g)\pi_0$, where $g$ aligns with the sign of the advantage function and has bounded scale. We then train $\pi_\theta$ by minimizing a general **tracking loss** $\mathcal{L}(\pi^*, \pi_\theta)$ toward this target (Section 3.3), where the loss can upper bound the performance gap between $\pi^*$ and $\pi_\theta$. The resulting framework **guarantees** performance improvement $\eta(\pi_\theta) - \eta(\pi_0) \geq 0$ whenever the idealized training loss, evaluated under the *behavior data* $\mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}(\pi^*, \pi_\theta)]$, is sufficiently low (Section 3.4).

Note that both $g$ and $\mathcal{L}$ can take many forms (Section 3.5). The tilt $g$ can be the sign, a clipped-linear, or exponential function of the advantage. The loss $\mathcal{L}$ may be any member of a certified family, including: weighted or unweighted total variation (TV), $\chi^2$(MSE), hinge, and KL divergence. Specific target–loss pairs recover existing algorithms, including PPO, Simple Policy Optimization (SPO, Xie et al. (2024)), Bounded Policy Optimization (BPO, Ao et al. (2026)), and on-policy variants of AWR and MPO (Section 3.5). FIRL also yields **novel instances**: the clipped-linear target paired with advantage-weighted TV or hinge losses, which are motivated by our theoretical analysis and show strong empirical performance.

### 3.2 GENERAL TARGET POLICY

**General Target.** Given a general direction function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and a step size $\epsilon > 0$, the *target policy* is defined as

$$\pi^\star(a \mid s) \;=\; \big(1 + \epsilon\, g(s, a)\big)\, \pi_0(a \mid s)\,, \tag{2}$$

where $g$ satisfies the following conditions.

**Assumption 1** (Admissible target)**.** *At every state $s$: (i)* centering, $\mathbb{E}_{a\sim\pi_0}[g] = 0$*; (ii)* normalization, $\mathbb{E}_{a\sim\pi_0}[g^2] \leq 1$*; (iii)* alignment, $\mathbb{E}_{a\sim\pi_0}[g\, A_{\pi_0}] \geq 0$*; (iv)* non-negativity, $\forall s, a,\ 1 + \epsilon g \geq 0$.

Centering makes $\sum_a \pi^\star(a|s) = 1$ for all $s$. Normalization upper bounds the scale of $g$, so that targets $\pi^*$ with different shapes have constrained per-step chi-square divergence from $\pi_0$: $\chi^2(\pi^\star\|\pi_0)(s) \le \epsilon^2$. Alignment ensures the target policy $\pi^*$ improves the performance over $\pi_0$, which can be shown by applying equation 1 to $(\pi^\star, \pi_0)$

$$(1-\gamma)\big(\eta(\pi^\star) - \eta(\pi_0)\big) \;=\; \epsilon\, \mathbb{E}_{s\sim d_{\pi^\star},\, a\sim\pi_0}[g(s,a)\, A_{\pi_0}(s,a)] \;\ge\; 0. \tag{3}$$

Finally, non-negativity ensures $\pi^*(a|s)$ a valid distribution: $\forall s, a, \pi^*(a|s) \ge 0$. Specific examples of admissible $g$ are detailed in Section 3.5.

### 3.3 Tracking Loss

**Bounding the performance gap** A suitable tracking loss should upper-bound the performance gap between $\pi_\theta$ and $\pi^*$. Then, by achieving sufficiently low loss, we can guarantee performance improvement since *i)* the target policy improves over $\pi_0$ from equation 3 and *ii)* $\pi_\theta$ has performance close to $\pi^*$ due to its low tracking loss. This is formalized in the following Lemma.

**Lemma 1** (Unified gap bound for tracking losses)**.** *Let Assumption 1 hold, $\pi_\theta \ll \pi_0$. For every tracking loss $\mathcal{L}$ satisfying $C_{\mathcal{L}}\,\mathcal{L}(\pi^*, \pi_\theta; s) \ge \mathbb{E}_{a\sim\pi^*(\cdot|s)}[A_{\pi_0}]$, where $C_{\mathcal{L}}$ is a constant, we have*

$$(1-\gamma)\big(\eta(\pi^\star) - \eta(\pi_\theta)\big) \;\le\; C_{\mathcal{L}}\, \mathbb{E}_{s\sim d_{\pi^\star}}\big[\mathcal{L}(\pi^*, \pi_\theta; s)\big]. \tag{4}$$

The Lemma directly follows by applying equation 1 to $(\pi^*, \pi_\theta)$. Examples of certified loss functions are detailed in Section 3.5. The expectation in equation 4 is still taken under the target policy state visitation $d_{\pi^\star}$. The next section transfers this to the behavior data distribution $d_{\pi_0}$ with an additional condition on the step size.

### 3.4 Guarantee from the practical tracking loss

We first introduce the step-size conditions that constrain the divergence between $d_{\pi^*}$ and $d_{\pi_0}$.

**Assumption 2** (Step size and absolute continuity)**.** *The trust-region radius $\epsilon > 0$ and discount factor $\gamma \in (0,1)$ satisfy*

$$\gamma\,(1+\epsilon^2) \;<\; 1, \qquad \textit{equivalently} \qquad \epsilon \;<\; \sqrt{(1-\gamma)/\gamma}, \tag{5}$$

*and the policy is absolutely continuous on behavior policy, $\pi_\theta(\cdot \mid s) \ll \pi_0(\cdot \mid s)$ for every state $s$.*

The admitted step sizes are compatible with common practice. At discount factors $\gamma = 0.9$, $0.95$, and $0.99$, equation 5 permits $\epsilon$ up to $0.33$, $0.23$, and $0.10$, respectively. Further, the absolute continuity assumption $\pi_\theta(\cdot \mid s) \ll \pi_0(\cdot \mid s)$ holds by construction for standard parameterizations, such as Gaussian or softmax policies. The following theorem, proved in Appendix A.3, attaches performance guarantees directly to a general certified loss in Lemma 1, in expectation under $d_{\pi_0}$.

**Theorem 1** (Improvement from loss minimization)**.** *Let Assumptions 1 and 2 hold. Let $\mathcal{L}$ be any certified loss satisfying equation 4 with constant $C_{\mathcal{L}}$. Then*

$$(1-\gamma)\big(\eta(\pi_\theta) - \eta(\pi_0)\big) \;\ge\; \epsilon\, \mathbb{E}_{\substack{s\sim d_{\pi^\star},\\ a\sim\pi_0}}\big[g\, A_{\pi_0}\big] \;-\; C_{\mathcal{L}} \left( \mathbb{E}_{d_{\pi_0}}\big[\mathcal{L}\big] \;+\; C_\epsilon \sqrt{\mathbb{E}_{d_{\pi_0}}\big[\mathcal{L}^2\big]} \right). \tag{6}$$

*where $C_\epsilon = \sqrt{\gamma\epsilon^2/(1-\gamma(1+\epsilon^2))}$.*

On the right-hand side, the first term is nonnegative from Assumption 1, the on-distribution tracking error enters through the first-order term $\mathcal{L}$, and the state visitation shift correction between $d_{\pi_0}$ and $d_{\pi^*}$ enters through the squared term $\mathcal{L}^2$. In particular, $\pi_\theta$ improves on $\pi_0$ whenever these losses are driven below the improvement from the target.

**Simplification with single loss functions** Minimizing the loss with two terms (first-order and second-order) in equation 6 jointly would require tuning their relative weight. Under one additional boundedness and non-negativity assumption, the guarantee can be achieved by a single loss. This is because for every certified nonnegative and self-bounding loss, $\mathbb{E}_{d_{\pi_0}}[\mathcal{L}^2] \le \mathcal{L}_{\max}\, \mathbb{E}_{d_{\pi_0}}[\mathcal{L}]$, so the second-moment term in Theorem 1 can be absorbed into the first-moment term. Besides, the first-moment term can also be upper-bounded by the second moment: $\sqrt{\mathbb{E}_{d_{\pi_0}}[\mathcal{L}^2]} \ge \mathbb{E}_{d_{\pi_0}}[\mathcal{L}]$.

**Corollary 1** (Single-loss bound)**.** *In addition to the conditions of Theorem 1, assume the state-wise loss $\mathcal{L}$ is non-negative, and $\mathcal{L}_{\max} := \sup_s \mathcal{L}(\theta; s) < \infty$. Then, with $\bar{\mathcal{L}} := \mathbb{E}_{d_{\pi_0}}[\mathcal{L}]$ or $\mathbb{E}_{d_{\pi_0}}[\mathcal{L}^2]$*

$$(1-\gamma)\big(\eta(\pi_\theta) - \eta(\pi_0)\big) \;\geq\; \epsilon\, \mathbb{E}_{\substack{s\sim d_{\pi^\star} \\ a\sim\pi_0}}\big[g\, A_{\pi_0}\big] \;-\; C_{\mathcal{L}}\, \kappa(\bar{\mathcal{L}}), \tag{7}$$

*where $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a strict monotonic function. Specifically, $\kappa(\bar{\mathcal{L}}) := \bar{\mathcal{L}} + C_\epsilon \sqrt{\mathcal{L}_{\max}\, \bar{\mathcal{L}}}$ for $\bar{\mathcal{L}} := \mathbb{E}_{d_{\pi_0}}[\mathcal{L}]$ and $\kappa(\bar{\mathcal{L}}) = (1 + C_\epsilon)\sqrt{\bar{\mathcal{L}}}$ for $\bar{\mathcal{L}} := \mathbb{E}_{d_{\pi_0}}[\mathcal{L}^2]$.*

Here, zero $\bar{\mathcal{L}}$ recovers the target improvement in equation 3. For finite-horizon MDPs, we also established a similar guarantee *without step size constraints (Assumption 2)*, as detailed in Appendix A.4.

**Comparison with existing bounds.** The bounds in equation 6-equation 7 differ from prior improvement and tracking guarantees in three ways. First, it contains *no irreducible worst-case constants* in the regularizer. Prior bounds penalize worst-case divergences over all states (Kakade & Langford, 2002; Schulman et al., 2015; Chan et al., 2022) or carry sup-norm regularization coefficients (Pirotta et al., 2013); either way, the certified steps become impractically conservative. In our bound, the worst-case constant is instead multiplied by the single tracking loss itself, which can be driven down by training. Second, it needs *no uncheckable coverage assumption*. Prior tracking guarantees depend on assumptions on how well the data covers the states the target visits, which the data itself cannot certify (Rashidinejad et al., 2021; Xie et al., 2021; Russo, 2026). We replace this with the checkable conditions in Assumption 2 and normalization in Assumption 1. Third, the guarantee is *more aligned with the practical procedure*. It certifies the standard rollout-and-training loop, rather than an idealized variant with exact subproblem solutions or oracle updates (Kakade & Langford, 2002; Schulman et al., 2015; Kuba et al., 2022; Huang et al., 2024).

### 3.5 Examples of target ratios and tracking losses

**Examples of admissible $g$.** Table 1 lists three example target designs of $g$, using **sign**, **clipped-linear**, and **exponential** functions of the advantage, in uncentered form $\hat{g}$; centering is obtained by subtracting a constant, $g := \hat{g} - \mathbb{E}_{\pi_0}[\hat{g}]$, which leaves the analytic improvement $\mathbb{E}_{\pi_0}[gA_{\pi_0}]$ unchanged since $\mathbb{E}_{\pi_0}[A_{\pi_0}] = 0$. The sign and clipped-linear shapes satisfy Assumption 1 by construction; the exponential shape requires a specific temperature design $\lambda$. Full verification of Assumption 1 and detailed design for each target are in Appendix A.1.

Table 1: Examples of uncentered target directions and the existing algorithms they relate to. The last column states the mechanism ensuring $\mathbb{E}_{\pi_0}[g^2] \leq 1$, as detailed in Appendix A.1. We write $Z := \mathbb{E}_{\pi_0}[\exp(A_{\pi_0}/\lambda)]$.

| Uncentered $\hat{g}$ | Related works | $\mathbb{E}_{\pi_0}[g^2] \leq 1$ via |
|---|---|---|
| $\mathrm{sign}(A_{\pi_0})$ | PPO (Schulman et al., 2017), SPO (Xie et al., 2024) | bounded shape |
| $\mathrm{clip}\big(\frac{A_{\pi_0}}{\mathrm{std}_{\pi_0}(A_{\pi_0})}, -\frac{1}{2\epsilon}, \frac{1}{2\epsilon}\big)$ | success conditioning (Russo, 2026) | standardization |
| $\frac{1}{\epsilon}\big(\exp(A_{\pi_0}/\lambda)/Z - 1\big)$ | AWR (Peng et al., 2019), MPO (Abdolmaleki et al., 2018) | temperature $\lambda$ |

**Unconstrained optimum for target policy.** A natural question is whether there exists a direction with the largest improvement $\mathbb{E}_{\pi_0}[g\, A_{\pi_0}]$. By the Cauchy–Schwarz inequality and $\mathbb{E}_{\pi_0}[g^2] \leq 1$,

$$\mathbb{E}_{\pi_0}[g\, A_{\pi_0}] \;\leq\; \sqrt{\mathbb{E}_{\pi_0}[A_{\pi_0}^2]\, \mathbb{E}_{\pi_0}[g^2]} \;\leq\; \mathrm{std}_{\pi_0}(A_{\pi_0}), \tag{8}$$

where the last step uses $\mathbb{E}_{\pi_0}[A_{\pi_0}] = 0$. Equality holds if and only if $g^\star = A_{\pi_0}/\mathrm{std}_{\pi_0}(A_{\pi_0})$. Here, the maximizer $g^\star$ is centered automatically, but it may violate the non-negativity constraint (Assumption 1 (iv)) when $A_{\pi_0} < -\mathrm{std}_{\pi_0}(A_{\pi_0})/(2\epsilon)$. Nevertheless, our **clipped-linear target** (second row of Table 1) is close to $g^*$ if the probability of being clipped vanishes with small $\epsilon$.

**Examples of certified loss functions** All $\mathcal{L}$ in Table 2 satisfy equation 4 with $C_{\mathcal{L}}$ listed in the Table 2. The proof is detailed in Appendix A.2. Note that for the MSE and KL losses, the certified quantity $\mathcal{L}$ in equation 4 is the square root of the loss. This square root can be removed in the final guarantee by using the definition $\bar{\mathcal{L}} := \mathbb{E}_{d_{\pi_0}}[\mathcal{L}^2]$ in Corollary 1. Most losses can also add a per-action weight $w$ with $w \geq |A_{\pi_\theta}|$ pointwise. However, $A_{\pi_\theta}$ is not available from the behavior data in practice. We therefore address this with the following remarks.

Table 2: Certified tracking losses: the specific loss $\mathcal{L}$ entering the unified bound of Lemma 1, and its constant $C_{\mathcal{L}}$. We denote the Hinge, Total Variation (TV), Mean Squared Error (MSE) and KL losses with $\mathcal{L}_H, \mathcal{L}_{TV}, \mathcal{L}_{MSE}$, and $\mathcal{L}_{KL}$. We denote the weighted counterpart of any loss $\mathcal{L}_X$ by $\mathcal{L}_X^w$. Here $w \geq |A_{\pi_\theta}|$, $\delta := \rho_\theta - \rho^\star$, $\mathrm{sg} := \mathrm{sign}(A_{\pi_\theta})$, $\bar{A} := \sup_s \frac{1}{2} \operatorname{span}_a A_{\pi_\theta}$, $A_\infty := \sup_{s,a} |A_{\pi_\theta}|$, $w_{\max} := \sup_s \mathbb{E}_{a\sim\pi^\star}[w]$. KL has no weighted counterpart since $\rho^\star \log \frac{\rho}{\rho^*}$ is not pointwise positive.

| | Advantage-weighted | | | Unweighted ($w \equiv 1$) | | |
|---|---|---|---|---|---|---|
| Loss | Definition | $\mathcal{L}$ in (4) | $C_{\mathcal{L}}$ | Definition | $\mathcal{L}$ in (4) | $C_{\mathcal{L}}$ |
| Hinge | $\mathbb{E}_{\pi_0}[w \max\{-\delta \,\mathrm{sg}, 0\}]$ (PPO) | $\mathcal{L}_{\mathrm{H}}^w$ | 1 | $\mathbb{E}_{\pi_0}[\max\{-\delta \,\mathrm{sg}, 0\}]$ | $\mathcal{L}_{\mathrm{H}}$ | $A_\infty$ |
| TV | $\mathbb{E}_{\pi_0}[w \,\lvert\delta\rvert]$ (BPO) | $\mathcal{L}_{\mathrm{TV}}^w$ | 1 | $\mathbb{E}_{\pi_0}[\,\lvert\delta\rvert\,]$ | $\mathcal{L}_{\mathrm{TV}}$ | $\bar{A}$ |
| MSE | $\mathbb{E}_{\pi_0}[w \,\delta^2/\rho^\star]$ (SPO) | $\sqrt{\mathcal{L}_{\mathrm{MSE}}^w}$ | $\sqrt{w_{\max}}$ | $\mathbb{E}_{\pi_0}[\delta^2/\rho^\star]$ | $\sqrt{\mathcal{L}_{\mathrm{MSE}}}$ | $\bar{A}$ |
| KL | — | — | — | $\mathbb{E}_{\pi_0}[\rho^\star \log \frac{\rho}{\rho^*}]$ | $\sqrt{\mathcal{L}_{\mathrm{KL}}}$ | $\bar{A}$ |

**Remark 1** (Practical weights)**.** *The weight requires $w \geq |A_{\pi_\theta}|$, which is satisfied by $w = |A_{\pi_0}| + (1+\gamma)\sigma_V$ with $\sigma_V := \max_s |V_{\pi_\theta}(s) - V_{\pi_0}(s)|$ (proof in Appendix A.2). In practice, we replace $(1+\gamma)\sigma_V$ by a constant and use $w = |A_{\pi_0}| + \sigma$. The substitution preserves $w \geq |A_{\pi_\theta}|$ whenever $(1+\gamma)\sigma_V \leq \sigma$. For $(1+\gamma)\sigma_V > \sigma$, we still have $|A_{\pi_\theta}| \leq |A_{\pi_0}| + (1+\gamma)\sigma_V < ((1+\gamma)\sigma_V/\sigma)\, w$, so the losses remain certified with $C_{\mathcal{L}}$ inflated by $(1+\gamma)\sigma_V/\sigma$, for any $\sigma > 0$.*

**Remark 2** (Sign of $A_{\pi_\theta}$ for Hinge loss)**.** *Hinge loss requires $sg := sign(A_{\pi_\theta})$. In practice, we substitute sg with $sign(A_{\pi_0})$, and the resulting weighted Hinge loss is denoted by $L_{H0}^w$. Here, the shift between $L_{H0}^w$ and $L_H^w$ can be upper bounded by $L_{TV}^w$: $L_H^w \leq L_{H0}^w + \mathbb{E}_{\pi_0}[w|\delta| \cdot \mathbf{1}(sign(A_{\pi_0}) \neq sign(A_{\pi_\theta}))] \leq L_{H0}^w + L_{TV}^w$. Therefore, any weighted loss $\mathcal{L} := L_{H0}^w + \sigma_2 L_{TV}^w \geq \min\{1, \sigma_2\} L_H^w$ is a certified loss with $C_{\mathcal{L}}$ inflated by $\max\{1, 1/\sigma_2\}$.*

**Order of losses** While every loss in Table 2 upper bounds the performance gap and satisfy Lemma 1, they are not equally tight. Per state, the losses can be ordered as

$$\mathcal{L}_{\mathrm{H}}^w \;\leq\; \mathcal{L}_{\mathrm{TV}}^w \;\leq\; \sqrt{\mathbb{E}_{a\sim\pi^\star}[w]\; \mathcal{L}_{\mathrm{MSE}}^w}, \qquad \mathcal{L}_{\mathrm{H}} \;\leq\; \mathcal{L}_{\mathrm{TV}} \;\leq\; \sqrt{\mathcal{L}_{\mathrm{MSE}}}, \qquad \mathcal{L}_{\mathrm{TV}} \;\leq\; \sqrt{\mathcal{L}_{\mathrm{KL}}} \quad (9)$$

and each weighted loss is dominated by its unweighted counterpart up to the weight scale, $\mathcal{L}_X^w \leq \|w\|_\infty \mathcal{L}_X$ for $X \in \{\mathrm{H}, \mathrm{TV}, \mathrm{MSE}\}$.

**Connection to existing algorithms.** Our framework unifies several existing algorithms through specific choices of targets and losses. First, pairing an **exponential tilt** with a KL loss recovers on-policy variants of AWR and MPO. Second, the **clipped-linear target** approximately recovers the success-conditioning tilt $1 + A_{\pi_0}/V_{\pi_0}$ of Russo (2026) when $\epsilon$ is set to $1/V_{\pi_0}$ and the loss is set to KL. Corollary 1 equips these methods with a coverage-free improvement guarantee, which is missing from existing works (Chan et al., 2022). Third, combining the **sign target** with an $A_{\pi_0}$**-weighted MSE** or $A_{\pi_0}$**-weighted TV** loss recovers SPO (Xie et al., 2024) or BPO (Ao et al., 2026) with mean advantages. Note that PPO is also covered in our framework: its clipped objective is equivalent to tracking the **sign target** under an $A_{\pi_0}$**-weighted hinge** loss (Huang et al., 2024).

**Proposed novel instances from the analysis.** Our theory also points to instances not yet explored. The inequality (8) identifies the **clipped-linear** target as the admissible design closest to the unconstrained optimum, and inequality (9) shows that the $A_{\pi_0}$**-weighted TV** and $A_{\pi_0}$**-weighted hinge** losses carry tighter certificates than the weighted MSE of SPO (Xie et al., 2024) and the KL divergence of success conditioning (Russo, 2026). The $A_{\pi_0}$**-weighted hinge** loss has already been validated at scale in PPO. The combination of these tighter losses with the clipped-linear target, however, corresponds to no existing algorithm. Our experiments in Section 4 therefore center on these new instances, compared against their counterparts with varying losses and targets.

### 3.6 Practical Algorithms

Algorithm 1 instantiates the framework with the standard on-policy loop. We parameterize the policy $\pi_\theta$ and the value function $V_\phi$ with neural networks. Each iteration collects a batch under the behavior policy $\pi_0$, estimates advantages $\hat{A}$ by GAE (Schulman et al., 2016), forms the target policy

**Algorithm 1** Follow-the-Improvement RL (FIRL)

**Require:** step size $\epsilon$, target shape $\tilde{g}$, loss $\ell$, epochs $K$, initial $\theta$, $\phi$

**for** iteration $= 1, 2, \ldots$ **do**
  $\pi_0 \leftarrow \pi_\theta$; collect batch $\mathcal{B}$ under $\pi_0$, storing $\log \pi_0(a_i \mid s_i)$
  Estimate return $\hat{R}_i$ and $\hat{A}_i$ by GAE with $V_\phi$, or by group statistics over $G$ actions per state
  Form $\hat{g}_i$ from $\tilde{g}(\hat{A}_i)$ with batch- or group-estimated moments; $\rho_i^\star \leftarrow 1 + \epsilon \hat{g}_i$
  **for** epoch $= 1, \ldots, K$ **do**
    Update $\phi$ by minimizing the MSE loss between $V_\phi$ and $\hat{R}_i$.
    Update $\theta$ by minibatch gradient descent on $J(\theta)$ in Section 3.6.
  **end for**
**end for**

$\pi^\star = (1 + \epsilon \hat{g})\pi_0$, and takes $K$ epochs of minibatch gradient steps on $J(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell\big(\pi_\theta, \pi_i^\star\big)$, where $\tilde{g}$ can be any admissible direction from Table 1, and $\ell$ any certified loss from Table 2.

**Batch-based estimation** The moments defining each target (e.g., $\mathbb{E}[g], \mathrm{std}(A_{\pi_0})$) are estimated from the rollout batch. This matches the advantage-normalization practice standard in PPO implementations, though it enforces the moment conditions of Assumption 1 only marginally over the visited states rather than per state. Similarly, the GRPO-style variant (Shao et al., 2024) uses group statistics over $G$ actions to estimate these moments.

## 4 Experiments

We conduct experiments to answer the following questions. **(Q1)** Do the new instances identified by the theory, clipped-linear target with tighter certified losses (Section 3.5), dominate some alternatives? Rather than exhaustively evaluating all target–loss combinations, we start from the instances that our theory deems promising: the clipped-linear target with $A_{\pi_0}$-weighted TV/hinge losses, and vary the design along target and loss axes to isolate their influence. We also consider the variations that induce strong baselines like PPO and SPO. **(Q2)** Are the more effective target-loss design choices regime-dependent? **(Q3)** Are the theoretically inspired components of the algorithm, like step-size conditions, normalization, and the weight offset $\sigma$, effective for performance?

### 4.1 Setup

**Methods.** We denote instances by {target}_{loss}. `CLinear_ATV` and `CLinear_AH` are the new instances: the clipped-linear target with the advantage-weighted TV / Hinge loss. `CLinear_TV/MSE` pairs the clipped-linear target with the best *unweighted* loss from TV, MSE and TV + MSE. `Sign_ATV/AMSE` and `Exp_ATV/AMSE` pair the sign and exponential targets with the best advantage-weighted loss from ATV, AMSE, and ATV + AMSE. The sign target with the AMSE loss covers SPO-style updates, and the exponential target corresponds to the tilt tracked by AWR and MPO. PPO and GRPO (Shao et al., 2024) are equivalent to `Sign_AH` (Huang et al., 2024).

**Benchmarks.** We evaluate on four domains: MuJoCo (Todorov et al., 2012) locomotion (Ant, Hopper, Swimmer), IsaacLab (Mittal et al., 2025) robot learning with massive parallel environments (G1-rough, Anymal-C-rough), Atari (Mnih et al., 2013) with discrete actions and pixel observations (Pong, Breakout, Asterix), and LLM fine-tuning for mathematical reasoning, trained with verifiable rewards and evaluated on AIME and AMC (mean@16 accuracy).

**Protocol.** All methods are tuned under the same search budget per environment; search ranges and selected hyperparameters are listed in Appendix A.5. For the grouped instances (`Sign_ATV/AMSE`, `Exp_ATV/AMSE`, `CLinear_TV/MSE`), the choice of the loss is part of the search. For the clipped-linear target, we find the ATV loss always matches or outperforms AMSE or AMSE+ATV. On Atari, we share hyperparameters across the three games. We report mean $\pm$ std over the last 100 (20 for LLM fine-tuning) logged evaluations of each run, pooled across 5 seeds.

Table 3: Final episode reward on the MuJoCo and IsaacLab benchmarks. Mean ± std over the last 100 logged evaluations of each run, pooled across 5 random seeds. Per environment, the top three ranks are shaded, lightening with rank.

| | MuJoCo | | | IsaacLab | |
|---|---|---|---|---|---|
| Method | Ant | Hopper | Swimmer | G1-rough | Anymal-C-rough |
| **CLinear_AH** | $5681 \pm 384$ | $3105 \pm 316$ | $362 \pm 0$ | $46.7 \pm 1.0$ | $17.3 \pm 0.6$ |
| **CLinear_ATV** | $5603 \pm 357$ | $3215 \pm 164$ | $362 \pm 0$ | $43.5 \pm 2.2$ | $14.7 \pm 0.3$ |
| CLinear_TV/MSE | $4562 \pm 455$ | $2936 \pm 171$ | $363 \pm 1$ | $42.9 \pm 2.3$ | $16.4 \pm 0.2$ |
| Exp_ATV/AMSE | $5015 \pm 161$ | $3023 \pm 275$ | $362 \pm 0$ | $15.6 \pm 6.5$ | $9.0 \pm 5.2$ |
| Sign_ATV/AMSE | $4821 \pm 327$ | $3207 \pm 263$ | $362 \pm 0$ | $42.7 \pm 0.9$ | $12.0 \pm 0.6$ |
| PPO (Sign_AH) | $5335 \pm 396$ | $3030 \pm 242$ | $362 \pm 0$ | $45.6 \pm 1.1$ | $16.8 \pm 0.3$ |

Table 4: Final performance on the Atari (left) and LLM fine-tuning (right) benchmarks. Left is episode reward; right is accuracy (mean@16). For Atari, mean ± std over the last 100 logged evaluations of each, pooled across 5 random seeds. For LLM, the numbers are the mean over the last 20 logged evaluations. Per environment, the top three ranks are shaded, lightening with rank.

| | Atari | | |
|---|---|---|---|
| Method | Pong | Breakout | Asterix |
| **CLinear_AH** | $20.5 \pm 0.4$ | $381 \pm 14$ | $4969 \pm 495$ |
| **CLinear_ATV** | $20.5 \pm 0.3$ | $414 \pm 20$ | $15198 \pm 9390$ |
| CLinear_TV/MSE | $20.7 \pm 0.1$ | $395 \pm 13$ | $7344 \pm 1080$ |
| Exp_ATV/AMSE | $20.4 \pm 0.3$ | $371 \pm 17$ | $9029 \pm 1959$ |
| Sign_ATV/AMSE | $20.4 \pm 0.3$ | $361 \pm 17$ | $7301 \pm 1103$ |
| PPO (Sign_AH) | $20.6 \pm 0.2$ | $361 \pm 15$ | $4497 \pm 1122$ |

| | LLM ft. (acc.) | |
|---|---|---|
| Method | AMC | AIME |
| **CLinear_AH** | 0.506 | 0.251 |
| **CLinear_ATV** | 0.497 | 0.270 |
| Sign_ATV | 0.506 | 0.238 |
| GRPO (Sign_AH) | 0.508 | 0.260 |

### 4.2 Main Results

**Overall comparisons.** Tables 3 and 4 report the main comparison between family members **(Q1)**. `CLinear_ATV` outperforms `CLinear_TV/MSE` on five of the eight benchmarks where both are evaluated (Ant, Hopper, G1-rough, Breakout, and Asterix), matches it on Pong and Swimmer, and trails only on Anymal-C-rough. Since `CLinear_TV/MSE` uses the best unweighted loss per environment, the comparison shows that advantage weighting is preferable in most evaluated tasks. To isolate the effect of the ratio, we can treat the loss type as a hyperparameter and select the best loss for each ratio in each environment. For this, as shown in Figure 2 and Table 3- 4, `CLinear_AH/ATV` achieves the highest averaged rank across all benchmarks, surpassing `Sign_AH/ATV`, `Sign_ATV/AMSE`, and `Exp_ATV/AMSE`. This shows clipped-linear as the preferred target choice, as advocated in Section 3.5.

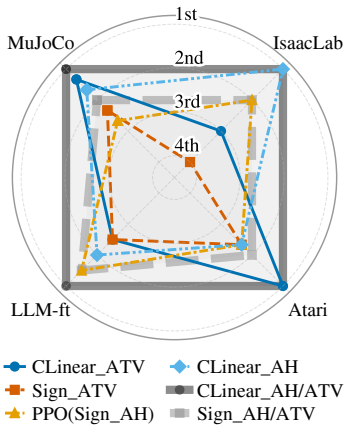



Figure 2: Mean rank per benchmark suite of the displayed methods. *CLinear_AH/ATV* achieves the highest overall ranking across all regimes.

**Benchmark-specific comparisons (Q2).** Holding the loss fixed, as shown in Figure 2, `CLinear_ATV` achieves higher averaged rank than `Sign_ATV` on all four suites, and `CLinear_AH` achieves higher averaged rank than `Sign_AH` on MuJoCo, IsaacLab, and Atari. The one exception is LLM fine-tuning, where GRPO (`Sign_AH`) attains the best averaged ranks. Holding the target fixed, the loss choice splits by regime: the advantage-weighted hinge loss is clearly preferable on IsaacLab, while the ATV loss is clearly superior on Atari. On the remaining suites, the two losses are comparable.

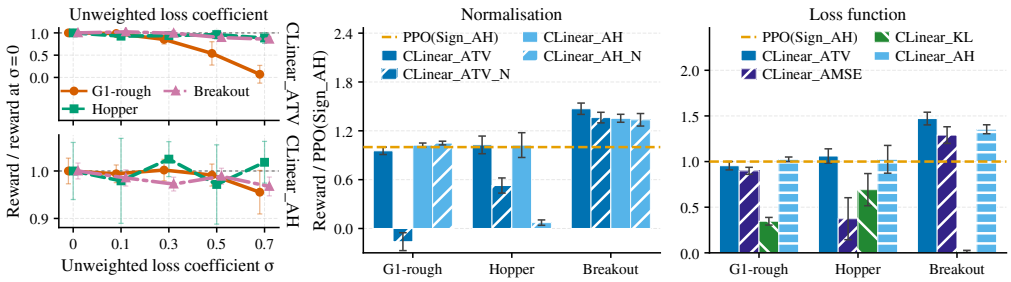



Figure 3: **Ablations of the $\sigma$ (Remark 1), normalization, and loss functions**. **Left:** Increasing $\sigma$ does not generally improve performance. Rewards are normalized w.r.t. $\sigma = 0$. **Middle:** Removing normalization causes a large performance drop on some tasks. Here *_N* denotes without normalization. **Right:** ATV/AH losses achieve the highest performance among all advantage-weighted losses.

### 4.3 Ablation studies (Q3)

**Step size $\epsilon$ with Assumption 2.** For Swimmer ($\gamma = 0.9999$) and Hopper ($\gamma = 0.999$), the commonly used step-size (clip ratio in PPO) search range $(0.1, 0.2, 0.3)$ violates Assumption 2. We therefore include smaller step sizes: 0.03 for Hopper and 0.01 for Swimmer, as listed in Table 5. Figure 4 shows the total reward ratio between the aforementioned step size and the best step size within $(0.1, 0.2, 0.3)$. These smaller step sizes retain comparable or improved performance for the reported methods. This suggests that the design principles motivated by Assumption 2 remain effective in practice.

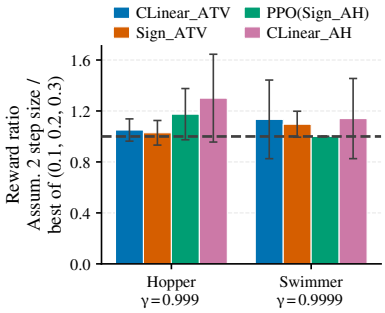



Figure 4: Ablation of step size (clip ratio). Enforcing step size to satisfy Assumption 2 increases performance in Hopper and Swimmer.

**Weight offset $\sigma$.** As shown in Figure 3, for `CLinear_ATV`, performance is flat or decreases on Breakout, Hopper and G1-rough as $\sigma$ grows. For `CLinear_AH`, only large $\sigma$ clearly decreases the performance for G1. Therefore, we use $\sigma = 0$ as default in benchmarking experiments.

**Normalization.** For `CLinear_ATV`, removing the batch normalization of the direction degrades performance severely on all three environments. For `CLinear_AH`, performance decreases or remains similar. This is the empirical footprint of the normalization conditions in Assumption 1.

**Loss function.** Fixing the clipped-linear target and varying the loss, ATV and AH are the most effective choices in all three environments. We attribute this to their proximity to the bound of Theorem 1: among the certified losses, the ATV and AH are the tightest, as mentioned in Section 3.3.

## 5 Conclusion

We proposed Follow-the-Improvement RL, a framework that unifies the constrained-optimization and target-tracking paradigms of policy optimization. For the entire class of admissible targets and certified losses, the practical training loss under the behavior distribution provably bounds the performance gap. The framework casts existing algorithms, including PPO and SPO, as instances, and identifies new algorithms with strong performance: clipped-linear targets with advantage-weighted TV / Hinge losses. Empirically, these new instances achieve top rankings from robotic control to Atari games, as well as comparable performance in LLM fine-tuning with GRPO. Our study also maps how design choices align with task regimes.

**Limitations and future work.** Gaps remain between theory and practice: the moment conditions are enforced by batch rather than per-state statistics, the certified weight is replaced by a constant offset, and advantage estimation error is not modeled. Moreover, we did not identify a single instance that dominates in every benchmark. Future directions include investigation of broader designs, extensions to off-policy training, and sample-efficient exploration.

### AI USE STATEMENT

In this work, we used generative AI tools for refining hypotheses, helping formulate and prove mathematical claims, and refining implementation. We have not used generative AI tools for proposing hypotheses, developing the core conceptual framework, designing the experiments, supporting qualitative and thematic data analysis, and interpreting results. Generating synthetic data sets, assisting with translation, cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature and brainstorming. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. LLM-assisted proofs were verified by two authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

The specific tuned hyperparameters and hyperparameter tuning range are detailed in the appendix. The learning curves of benchmarking experiments are detailed in the appendix. The assumptions are clearly stated in the main text, and the step-by-step complete proof is included in the appendix.

## REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations (ICLR)*, 2018.

Yunke Ao, Le Chen, Bruce D Lee, Assefa S Wahd, Aline Czarnobai, Philipp Fürnstahl, Bernhard Schölkopf, and Andreas Krause. Bounded ratio reinforcement learning. *arXiv preprint arXiv:2604.18578*, 2026.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse KL divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1861–1870, 2018.

Nai-Chieh Huang, Ping-Chun Hsieh, Kuo-Hao Ho, and I-Chen Wu. Ppo-clip attains global optimality: Towards deeper understandings of clipping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12600–12607, 2024.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, pp. 267–274, 2002.

Jakub Grudzien Kuba, Christian Schroeder de Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 7825–7844, 2022.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *International conference on machine learning*, pp. 307–315. PMLR, 2013.

Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Daniel J Russo. Success conditioning as policy improvement: The optimization problem solved by imitating success. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.18175.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Clemens Schwarke, Mayank Mittal, Nikita Rudin, David Hoeller, and Marco Hutter. Rsl-rl: A learning library for robotics research. *arXiv preprint arXiv:2509.10771*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12:1057–1063, 1999.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 113–122, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hu, and Renjing Xing. Simple policy optimization. *arXiv preprint arXiv:2401.16025*, 2024.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

# A Appendix

## A.1 Conditions satisfaction for the target ratios

Throughout this section, fix a state $s$ and let all expectations, variances, and the constants $Z$, $A_{\max}$, $\mathrm{std}_{\pi_0}(A_{\pi_0})$ be taken conditionally under $\pi_0(\cdot \mid s)$; we suppress the dependence on $s$. Recall $\mathbb{E}_{\pi_0}[A_{\pi_0}] = 0$ and $|A_{\pi_0}| \le A_{\max}$.

**The sign target.** This target only requires centering: $g = \mathrm{sign}(A_{\pi_0}) - \mathbb{E}_{\pi_0}[\mathrm{sign}(A_{\pi_0})]$. Since the sign function is bounded by 1, its variance cannot exceed 1, so $\mathbb{E}_{\pi_0}[g^2] \le 1$ holds automatically. The improvement rate can be computed as $\mathbb{E}_{\pi_0}[gA_{\pi_0}] = \mathbb{E}_{\pi_0}[\mathrm{sign}(A_{\pi_0})A_{\pi_0}] = \mathbb{E}_{\pi_0}[|A_{\pi_0}|] \ge 0$. The non-negativity holds for any $\epsilon < 1$.

**The clipped-linear target.** We first normalize the advantage, then clip it for boundedness: $\tilde{g} = \mathrm{clip}(\frac{A_{\pi_0}}{\mathrm{std}_{\pi_0}(A_{\pi_0})}, -\frac{1}{2\epsilon}, \frac{1}{2\epsilon})$. Then we center it by subtracting the mean: $g = \tilde{g} - \mathbb{E}_{\pi_0}[\tilde{g}]$. This also yields the second momentum condition

$$\mathbb{E}_{\pi_0}[g^2] = \mathbb{E}_{\pi_0}[\tilde{g}^2] - \mathbb{E}_{\pi_0}[\tilde{g}]^2 \le \mathbb{E}_{\pi_0}[A_{\pi_0}^2/\mathrm{var}_{\pi_0}(A_{\pi_0})] = 1.$$

and the non-negativity condition

$$\epsilon g = \epsilon \cdot (\tilde{g} - \mathbb{E}_{\pi_0}[\tilde{g}]) \ge \epsilon \cdot (-\frac{1}{2\epsilon} - \frac{1}{2\epsilon}) = -1$$

improvement rate satisfies

$$\mathbb{E}_{\pi_0}[gA_{\pi_0}] = \mathbb{E}_{\pi_0}[\tilde{g}A_{\pi_0}] = \mathbb{E}_{\pi_0}[|\tilde{g}| \cdot |A_{\pi_0}|] \ge \mathbb{E}_{\pi_0}[\tilde{g}^2] \cdot \mathrm{std}_{\pi_0}(A_{\pi_0}) \ge 0,$$

given that $\tilde{g}$ always has the same sign as $A_{\pi_0}$, and $|A_{\pi_0}| \ge |\tilde{g}| \cdot \mathrm{std}_{\pi_0}(A_{\pi_0})$.

**The exponential target.** We use $Z$ to denote the normalizing factor $\mathbb{E}_{\pi_0}[\exp(A_{\pi_0}/\lambda)]$. Then $g = \frac{1}{\epsilon}\big(\exp(A_{\pi_0}/\lambda)/Z - 1\big)$ is centered by construction, as the tilt has unit mean under $\pi_0$; the second moment is instead controlled through the temperature. Setting $\lambda = \max\{4A_{\max}/3,\ 2\,\mathrm{std}_{\pi_0}(A_{\pi_0})/\epsilon\}$ ensures $\mathbb{E}_{\pi_0}[g^2] \le 1$. This follows from the Lemma below.

**Lemma 2** (Temperature selection)**.** *Let* $Z = \mathbb{E}_{\pi_0}\big[\exp(A_{\pi_0}/\lambda)\big]$ *and* $g = \frac{1}{\epsilon}\big(\exp(A_{\pi_0}/\lambda)/Z - 1\big)$. *If*

$$\lambda \ \ge\ \max\left\{\tfrac{4}{3}\,A_{\max},\ \frac{2\,\mathrm{std}_{\pi_0}(A_{\pi_0})}{\epsilon}\right\},$$

*then* $\mathbb{E}_{\pi_0}[g] = 0$, $\mathbb{E}_{\pi_0}[g^2] \le 1$, *and* $1 + \epsilon g \ge 0$.

*Proof.* Write $X = A_{\pi_0}/\lambda$, so that $\mathbb{E}_{\pi_0}[X] = 0$ and, by the first branch of the temperature condition, $|2X| \le 2A_{\max}/\lambda \le \frac{3}{2}$.

*Centering and nonnegativity.* By definition of $Z$, $\mathbb{E}_{\pi_0}[e^X/Z] = 1$, hence $\mathbb{E}_{\pi_0}[g] = 0$. Moreover $1 + \epsilon g = e^X/Z > 0$, so the induced ratio is nonnegative without any condition on $\lambda$.

*Second moment.* Since $\mathbb{E}_{\pi_0}[g^2] = \mathrm{Var}_{\pi_0}\big(e^X/Z\big)/\epsilon^2$, it suffices to show $\mathrm{Var}_{\pi_0}(e^X/Z) \le \epsilon^2$. By Jensen's inequality, $Z = \mathbb{E}_{\pi_0}[e^X] \ge e^{\mathbb{E}_{\pi_0}[X]} = 1$, and therefore

$$\mathrm{Var}_{\pi_0}\big(e^X/Z\big) \ =\ \frac{\mathbb{E}_{\pi_0}\big[e^{2X}\big]}{Z^2} - 1 \ \le\ \mathbb{E}_{\pi_0}\big[e^{2X}\big] - 1. \tag{10}$$

We use the elementary inequality

$$e^u \;\le\; 1 + u + u^2 \qquad \text{for all } u \le \tfrac{3}{2}. \tag{11}$$

To verify equation 11, let $h(u) = 1 + u + u^2 - e^u$, so $h(0) = h'(0) = 0$ and $h''(u) = 2 - e^u$. On $u \le \log 2$, $h'' \ge 0$, so $h$ is convex with a global minimum $h(0) = 0$ there, giving $h \ge 0$. On $[\log 2, \frac{3}{2}]$, $h$ is concave, and $h(\log 2) > 0$, $h(\frac{3}{2}) = \frac{19}{4} - e^{3/2} > 0$, so $h \ge 0$ on the whole interval by concavity.

Applying equation 11 with $u = 2X$ (admissible since $|2X| \le \frac{3}{2}$) and taking expectations,

$$\mathbb{E}_{\pi_0}\big[e^{2X}\big] \;\le\; 1 + 2\,\mathbb{E}_{\pi_0}[X] + 4\,\mathbb{E}_{\pi_0}\big[X^2\big] \;=\; 1 + \frac{4\,\mathrm{Var}_{\pi_0}(A_{\pi_0})}{\lambda^2}.$$

Combining with equation 10 and the second branch of the temperature condition, $\lambda \ge 2\,\mathrm{std}_{\pi_0}(A_{\pi_0})/\epsilon$,

$$\mathrm{Var}_{\pi_0}\big(e^X/Z\big) \;\le\; \frac{4\,\mathrm{Var}_{\pi_0}(A_{\pi_0})}{\lambda^2} \;\le\; \epsilon^2,$$

which yields $\mathbb{E}_{\pi_0}[g^2] \le 1$. □

**Lemma 3** (Nonnegative improvement rate)**.** *Under the setting of Lemma 2,*

$$\mathbb{E}_{\pi_0}\big[g\,A_{\pi_0}\big] \;=\; \frac{\mathrm{Cov}_{\pi_0}\big(e^{A_{\pi_0}/\lambda},\, A_{\pi_0}\big)}{\epsilon\, Z} \;\ge\; 0,$$

*with equality if and only if $A_{\pi_0}$ is $\pi_0$-almost-surely constant (and hence zero).*

*Proof.* Using $\mathbb{E}_{\pi_0}[A_{\pi_0}] = 0$,

$$\mathbb{E}_{\pi_0}\big[g\,A_{\pi_0}\big] = \frac{1}{\epsilon Z}\,\mathbb{E}_{\pi_0}\big[e^{A_{\pi_0}/\lambda}A_{\pi_0}\big] = \frac{1}{\epsilon Z}\,\mathrm{Cov}_{\pi_0}\big(e^{A_{\pi_0}/\lambda},\, A_{\pi_0}\big).$$

For nonnegativity, let $A$ and $A'$ be independent copies of $A_{\pi_0}$ under $\pi_0$. Since $x \mapsto e^{x/\lambda}$ is increasing,

$$\big(e^{A/\lambda} - e^{A'/\lambda}\big)\big(A - A'\big) \;\ge\; 0 \qquad \text{pointwise},$$

and taking expectations gives $2\,\mathrm{Cov}_{\pi_0}(e^{A_{\pi_0}/\lambda}, A_{\pi_0}) \ge 0$ (Chebyshev's association inequality). Since $e^{x/\lambda}$ is strictly increasing, the product above is strictly positive whenever $A \ne A'$, so the covariance vanishes only if $A_{\pi_0}$ is almost surely constant; a constant advantage with zero mean is zero. □

### A.2 Proof of certified losses in Table 2

*Proof.* By the performance difference lemma and the change of measure $a \sim \pi^\star \to a \sim \pi_0$, $(1-\gamma)(\eta(\pi^\star) - \eta(\pi_\theta)) = \mathbb{E}_{s\sim d_{\pi^\star}}[\Delta(s)]$ with $\Delta(s) = \mathbb{E}_{a\sim\pi^*}[A_{\pi_\theta}] = \mathbb{E}_{a\sim\pi_0}[(\rho^\star - \rho_\theta)\,A_{\pi_\theta}]$, where we used $\mathbb{E}_{a\sim\pi_\theta}[A_{\pi_\theta}] = 0$. It suffices to bound $\Delta(s)$ by each $C_{\mathcal{L}}\,\mathcal{L}(\theta; s)$ pointwise; the claim follows by integrating over $d_{\pi^\star}$.

**Hinge.** Pointwise $x \le \max\{x, 0\}$ with $x = (\rho^\star - \rho_\theta)\,\mathrm{sign}(A_{\pi_\theta})$ gives $\Delta(s) \le \mathbb{E}_{\pi_0}[\max\{x,0\}\,|A_{\pi_\theta}|] \le \mathcal{L}^w_{\mathrm{Hinge}}(\theta; s)$ using $w \ge |A_{\pi_\theta}|$; the unweighted case follows from $|A_{\pi_\theta}| \le A_\infty$.

**TV.** Dropping the truncation, $\Delta(s) \le \mathbb{E}_{\pi_0}[\,|\rho_\theta - \rho^\star|\,|A_{\pi_\theta}|\,] \le \mathcal{L}^w_{\mathrm{TV}}(\theta; s)$. For the unweighted case, since $\mathbb{E}_{\pi_0}[\rho^\star - \rho_\theta] = 0$, any baseline $b(s)$ may be subtracted: $\Delta(s) = \mathbb{E}_{\pi_0}[(\rho^\star - \rho_\theta)(A_{\pi_\theta} - b)] \le \bar{A}\,\mathcal{L}_{\mathrm{TV}}(\theta; s)$ with $b$ the midpoint of the action range.

**MSE.** By the Cauchy–Schwarz inequality and $\mathbb{E}_{\pi_0}[w\rho^\star] = \mathbb{E}_{\pi^\star}[w]$,

$$\mathcal{L}^w_{\mathrm{TV}}(\theta; s) = \mathbb{E}_{\pi_0}\Big[\sqrt{w\rho^\star}\cdot\tfrac{\sqrt{w}\,|\rho_\theta - \rho^\star|}{\sqrt{\rho^\star}}\Big] \le \sqrt{\mathbb{E}_{a\sim\pi^\star}[w]}\,\sqrt{\mathcal{L}^w_{\mathrm{MSE}}(\theta; s)} \le \sqrt{w_{\max}}\,\sqrt{\mathcal{L}^w_{\mathrm{MSE}}(\theta; s)},$$

and the unweighted case sets $w \equiv 1$, for which $\mathbb{E}_{\pi^\star}[w] = 1$, then applies the TV step.

**KL.** Pinsker's inequality gives $\mathcal{L}_{\mathrm{TV}}(\theta; s) = \sum_a |\pi_\theta - \pi^\star| \le \sqrt{2\,\mathcal{L}_{\mathrm{KL}}(\theta; s)}$, and the TV step concludes. □

**Proof and more details on Remark 1.** Writing $\Delta V = V_{\pi_\theta} - V_{\pi_0}$, the identity

$$A_{\pi_\theta}(s,a) - A_{\pi_0}(s,a) = \gamma\,\mathbb{E}[\Delta V(s') \mid s,a] - \Delta V(s)$$

gives the pointwise bound

$$|A_{\pi_\theta}(s,a)| \le |A_{\pi_0}(s,a)| + (1+\gamma)\|\Delta V\|_\infty.$$

Alternatively, for ATV we can retain a span-based correction by bounding the performance-gap integrand directly. At a fixed state, define $\Delta(s) = \mathbb{E}_{\pi_0}[(\rho^\star - \rho_\theta)A_{\pi_\theta}]$. Since $\mathbb{E}_{\pi_0}[\rho^\star - \rho_\theta] = 0$, action-independent terms cancel. Subtracting the midpoint of the statewise range of $\Delta V$, we can obtain

$$\Delta(s) \le \mathbb{E}_{\pi_0}\left[\left(|A_{\pi_0}| + \frac{\gamma}{2}\,\mathrm{span}(\Delta V)\right)|\rho^\star - \rho_\theta|\right],$$

where expectations are over actions at the fixed state and the span is over all states. Notably, here the constant $\frac{\gamma}{2}\,\mathrm{span}(\Delta V)$ is much tighter than $(1+\gamma)\|\Delta V\|_\infty$ for the ATV loss.

### A.3 Proof of Theorem 1

We start by proving the Hammersley–Chapman–Robbins (HCR) bound:

**Lemma 4** (MSE transfer)**.** *Let $P$ and $Q$ be probability measures on a measurable space $(\mathcal{S}, \Sigma)$ with $Q \ll P$, and let $f\colon \mathcal{S} \to \mathbb{R}$ be measurable with $\mathbb{E}_P[f^2] < \infty$. Then*

$$\left|\mathbb{E}_Q[f] - \mathbb{E}_P[f]\right| \;\le\; \sqrt{\chi^2(Q\,\|\,P)}\,\sqrt{\mathrm{Var}_P(f)},$$

*where $\chi^2(Q\|P) = \mathbb{E}_P\big[\big(\frac{dQ}{dP} - 1\big)^2\big]$. In particular, $\mathbb{E}_Q[f] \le \mathbb{E}_P[f] + \sqrt{\chi^2(Q\|P)}\,\sqrt{\mathrm{Var}_P(f)}$. Equality holds in the two-sided bound if and only if $f$ is a $P$-almost-sure affine function of the density ratio $\frac{dQ}{dP}$.*

*Proof.* Write $\rho = \frac{dQ}{dP}$ and note $\mathbb{E}_P[\rho] = 1$. Since $\mathbb{E}_P[f^2] < \infty$ and $\mathbb{E}_P[(\rho-1)^2] = \chi^2(Q\|P)$ (both sides being infinite if the latter diverges, in which case the bound is trivial), the change of measure gives

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] = \mathbb{E}_P\big[(\rho-1)\,f\big] = \mathbb{E}_P\big[(\rho-1)\,\big(f - \mathbb{E}_P[f]\big)\big],$$

where the second equality uses $\mathbb{E}_P[\rho - 1] = 0$. The Cauchy–Schwarz inequality then yields

$$\left|\mathbb{E}_Q[f] - \mathbb{E}_P[f]\right| \;\le\; \sqrt{\mathbb{E}_P\big[(\rho-1)^2\big]}\,\sqrt{\mathbb{E}_P\big[\big(f - \mathbb{E}_P[f]\big)^2\big]} \;=\; \sqrt{\chi^2(Q\|P)}\,\sqrt{\mathrm{Var}_P(f)},$$

with equality if and only if $f - \mathbb{E}_P[f]$ and $\rho - 1$ are $P$-a.s. proportional. □

Then, we upper bound $\chi^2(d_{\pi^*}||d_{\pi_0})$ under condition 5:

**Lemma 5** (Bounded MSE divergence of occupancy)**.** *Let $\epsilon$ and $\gamma$ satisfy condition 5. Let Assumption 1 hold for $g$. Then the occupancy maps $d_{\pi^*}$ and $d_{\pi_0}$ with $\pi^*(a|s) := (1+\epsilon g)\pi_0(a|s)$ has bounded $\chi^2(d_{\pi^*}||d_{\pi_0}) \le \gamma\epsilon^2/(1-\gamma(1+\epsilon^2))$.*

*Proof.* Throughout, write $\rho(a \mid s) = \pi^*(a \mid s)/\pi_0(a \mid s) = 1 + \epsilon g(s,a)$. Under Assumption 1, $\mathbb{E}_{a\sim\pi_0(\cdot|s)}[g] = 0$ and $1 + \epsilon g \ge 0$, so $\pi^*$ is a valid policy, and $\mathbb{E}_{a\sim\pi_0(\cdot|s)}[g^2] \le 1$ gives the per-step bound

$$\mathbb{E}_{a\sim\pi_0(\cdot|s)}\big[\rho(a \mid s)^2\big] = 1 + \epsilon^2\,\mathbb{E}_{a\sim\pi_0(\cdot|s)}\big[g^2\big] \le 1 + \epsilon^2 \qquad \text{for every } s. \tag{12}$$

*Step 1 (time-$t$ marginals).* Let $p_t^*$ and $p_t^0$ denote the state distributions at time $t$ under $\pi^*$ and $\pi_0$, respectively, from the common initial distribution $\mu$, and let $m_t = \mathbb{E}_{s\sim p_t^0}\big[\big(p_t^*(s)/p_t^0(s)\big)^2\big]$. We claim $m_t \le (1+\epsilon^2)^t$, by induction; $m_0 = 1$ since $p_0^* = p_0^0 = \mu$. For the step, define the joint state–action distributions $q_t^*(s,a) = p_t^*(s)\,\pi^*(a \mid s)$ and $q_t^0(s,a) = p_t^0(s)\,\pi_0(a \mid s)$. Their second moment factorizes and is controlled by equation 12:

$$\mathbb{E}_{q_t^0}\left[\left(\tfrac{q_t^*}{q_t^0}\right)^2\right] = \mathbb{E}_{s\sim p_t^0}\left[\left(\tfrac{p_t^*(s)}{p_t^0(s)}\right)^2 \mathbb{E}_{a\sim\pi_0(\cdot|s)}\big[\rho^2\big]\right] \le (1+\epsilon^2)\,m_t.$$

Both $p^*_{t+1}$ and $p^0_{t+1}$ arise from $q^*_t$ and $q^0_t$ by the *same* transition kernel $P(s' \mid s, a)$, and the second moment is nonincreasing under a common kernel: by the Cauchy–Schwarz inequality,

$$p^*_{t+1}(s')^2 = \left( \sum_{s,a} \frac{q^*_t(s,a)}{\sqrt{q^0_t(s,a)}} \sqrt{q^0_t(s,a)}\, P(s' \mid s, a) \right)^2 \le \left( \sum_{s,a} \frac{q^*_t(s,a)^2}{q^0_t(s,a)}\, P(s' \mid s, a) \right) p^0_{t+1}(s'),$$

and dividing by $p^0_{t+1}(s')$ and summing over $s'$ yields $m_{t+1} \le \mathbb{E}_{q^0_t}\big[(q^*_t/q^0_t)^2\big] \le (1+\epsilon^2)\, m_t$, completing the induction.

*Step 2 (from marginals to occupancies).* The occupancies are the geometric mixtures $d_{\pi^*} = \sum_{t\ge 0} \lambda_t\, p^*_t$ and $d_{\pi_0} = \sum_{t\ge 0} \lambda_t\, p^0_t$ with $\lambda_t = (1-\gamma)\gamma^t$. Applying the Cauchy–Schwarz inequality to the mixture, for every $s$,

$$d_{\pi^*}(s)^2 = \left( \sum_t \frac{\sqrt{\lambda_t}\, p^*_t(s)}{\sqrt{p^0_t(s)}} \cdot \sqrt{\lambda_t\, p^0_t(s)} \right)^2 \le \left( \sum_t \lambda_t\, \frac{p^*_t(s)^2}{p^0_t(s)} \right) d_{\pi_0}(s).$$

Dividing by $d_{\pi_0}(s)$, summing over $s$, and invoking Step 1,

$$\mathbb{E}_{s\sim d_{\pi_0}} \left[ \left( \tfrac{d_{\pi^*}(s)}{d_{\pi_0}(s)} \right)^2 \right] \le \sum_{t\ge 0} \lambda_t\, m_t \le (1-\gamma) \sum_{t\ge 0} \gamma^t (1+\epsilon^2)^t = \frac{1-\gamma}{1-\gamma(1+\epsilon^2)},$$

where the series converges by condition 5, i.e., $\gamma(1+\epsilon^2) < 1$. Since $\chi^2(d_{\pi^*} \| d_{\pi_0}) = \mathbb{E}_{d_{\pi_0}}\big[(d_{\pi^*}/d_{\pi_0})^2\big] - 1$, this gives

$$\chi^2(d_{\pi^*} \,\|\, d_{\pi_0}) \;\le\; \frac{1-\gamma}{1-\gamma(1+\epsilon^2)} - 1 \;=\; \frac{\gamma\,\epsilon^2}{1-\gamma(1+\epsilon^2)}. \qquad \square$$

Now we prove Theorem 1:

*Proof.* Applying Lemma 4 to Lemma 1 with $f := \mathcal{L}$, $Q := d_{\pi^*}$ and $P := d_{\pi_0}$, we have:

$$\begin{aligned} \mathbb{E}_{s\sim d_{\pi^*}}[\mathcal{L}] \le & \mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}] + \sqrt{\chi^2(d_{\pi^*}||d_{\pi_0})} \sqrt{\mathrm{Var}_{d_{\pi_0}}(\mathcal{L})} \\ \le & \mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}] + \sqrt{\chi^2(d_{\pi^*}||d_{\pi_0})} \sqrt{\mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}^2]}, \end{aligned} \tag{13}$$

because $\mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}^2] = \mathrm{Var}_{d_{\pi_0}}(\mathcal{L}) + \mathbb{E}_{s\sim d_{\pi_0}}[\mathcal{L}]^2$. Finally, applying Lemma 5 to bound $\chi^2(d_{\pi^*}||d_{\pi_0})$ finishes the proof.

$\square$

### A.4 Finite-Horizon Analysis

We extend Theorem 1 to finite-horizon MDPs. The step-size condition equation 5 of Assumption 2 is not needed. Its only role in the discounted analysis is the convergence of the geometric series in Lemma 5, which a finite horizon renders finite for every $\epsilon$.

**Setting.** Consider a finite-horizon MDP $(\mathcal{S}, \mathcal{A}, \{P_t\}, r, d_0, H)$ with horizon $H$ and (possibly time-dependent) kernels $P_t(s' \mid s, a)$. A policy is a collection $\pi = (\pi_t)_{t=0}^{H-1}$; a stationary parameterization shares $\theta$ across $t$. The return is $\eta(\pi) = \mathbb{E}^\pi\big[\sum_{t=0}^{H-1} r(s_t, a_t)\big]$, with time-indexed values $V^\pi_t$, $Q^\pi_t$, advantages $A^\pi_t = Q^\pi_t - V^\pi_t$, and the convention $V^\pi_H \equiv 0$. As before, $\mathbb{E}_{a\sim\pi_t(\cdot|s)}[A^\pi_t(s,a)] = 0$. The occupancy is a probability distribution over state–time pairs,

$$d^\pi(s,t) \;=\; \tfrac{1}{H}\, \mathbb{P}^\pi(s_t = s), \qquad (s,t) \in \mathcal{S} \times \{0, \dots, H-1\},$$

with time-$t$ slices $p^\pi_t(s) = \mathbb{P}^\pi(s_t = s)$. Assumption 1 is imposed per $(s,t)$ on directions $g_t$, defining $\pi^\star_t = (1+\epsilon g_t)\, \pi_{0,t}$, and we retain only the absolute-continuity part of Assumption 2: $\pi_{\theta,t}(\cdot \mid s) \ll \pi_{0,t}(\cdot \mid s)$ for all $s, t$.

**Lemma 6** (Finite-horizon performance difference)**.** *For any policies $\pi, \pi'$,*

$$\tfrac{1}{H}\big(\eta(\pi) - \eta(\pi')\big) = \mathbb{E}_{(s,t)\sim d^\pi,\, a\sim\pi_t(\cdot|s)}\big[A^{\pi'}_t(s,a)\big].$$

*Proof.* With $V_H^{\pi'} \equiv 0$, the sum $\sum_{t=0}^{H-1}\big(r(s_t,a_t) + V_{t+1}^{\pi'}(s_{t+1}) - V_t^{\pi'}(s_t)\big)$ telescopes to $\sum_t r(s_t,a_t) - V_0^{\pi'}(s_0)$, and $\mathbb{E}_{s_0\sim d_0}[V_0^{\pi'}(s_0)] = \eta(\pi')$, so its expectation under $\pi$ equals $\eta(\pi) - \eta(\pi')$. Taking the conditional expectation over $s_{t+1} \sim P_t(\cdot \mid s_t, a_t)$ turns each summand into $Q_t^{\pi'}(s_t,a_t) - V_t^{\pi'}(s_t) = A_t^{\pi'}(s_t,a_t)$. Averaging over $t$ with weight $1/H$ gives the occupancy form. □

Applying Lemma 6 to $(\pi^\star, \pi_0)$ and changing the action measure $a \sim \pi_t^\star \to a \sim \pi_{0,t}$ gives the analytic improvement

$$\tfrac{1}{H}\big(\eta(\pi^\star) - \eta(\pi_0)\big) = \epsilon\, \mathbb{E}_{(s,t)\sim d^{\pi^\star},\, a\sim\pi_{0,t}}\big[g_t\, A_t^{\pi_0}\big] \;\geq\; 0. \tag{14}$$

Replacing $s$ by $(s,t)$ in Lemma 1 (with the constants $A_\infty$, $\bar{A}$, $w_{\max}$ of Table 2 now taken over $(s,a,t)$) yields

$$\tfrac{1}{H}\big(\eta(\pi^\star) - \eta(\pi_\theta)\big) \leq C_L\, \mathbb{E}_{(s,t)\sim d^{\pi^\star}}[L(\theta; s,t)]$$

for every certified loss. Now it makes no use of the discount.

Lemma 4 is stated for arbitrary probability measures and applies to $\mathcal{S} \times \{0,\dots,H-1\}$ without changes. The only statement requiring a new proof is the occupancy bound, as detailed below.

**Lemma 7** (Bounded MSE divergence of occupancies, finite horizon)**.** *Let Assumption 1 hold per $(s,t)$. Then, for every $\epsilon > 0$ and $H \geq 1$,*

$$\chi^2\big(d^{\pi^\star} \,\big\|\, d^{\pi_0}\big) \;\leq\; C_H^2 \;:=\; \frac{(1+\epsilon^2)^H - 1}{H\epsilon^2} \;-\; 1,$$

*where $C_1 = 0$.*

*Proof.* Write $\rho_t(a \mid s) = 1 + \epsilon g_t(s,a)$ and $m_t = \mathbb{E}_{s\sim p_t^0}\big[(p_t^\star(s)/p_t^0(s))^2\big]$. Centering and normalization give, at every $(s,t)$,

$$\mathbb{E}_{a\sim\pi_{0,t}(\cdot|s)}\big[\rho_t(a \mid s)^2\big] = 1 + \epsilon^2\, \mathbb{E}_{a\sim\pi_{0,t}}[g_t^2] \;\leq\; 1 + \epsilon^2. \tag{15}$$

*Step 1* of the proof of Lemma 5 applies without modification, with $P$ replaced by $P_t$: for the joint laws $q_t^\star(s,a) = p_t^\star(s)\,\pi_t^\star(a \mid s)$ and $q_t^0(s,a) = p_t^0(s)\,\pi_{0,t}(a \mid s)$, inequality equation 15 gives $\mathbb{E}_{q_t^0}[(q_t^\star/q_t^0)^2] \leq (1+\epsilon^2)\, m_t$, and the Cauchy–Schwarz contraction under the common kernel $P_t$ gives $m_{t+1} \leq (1+\epsilon^2)\, m_t$. Since $m_0 = 1$ ($p_0^\star = p_0^0 = d_0$), induction yields $m_t \leq (1+\epsilon^2)^t$, with no condition on $\epsilon$.

*Step 2* simplifies relative to the discounted case: because the time index is part of the augmented state, the occupancies are not mixtures over $t$ but couple the uniform law on $t$ with the time-$t$ marginals, $d^{\pi^\star}(s,t) = \frac{1}{H}p_t^\star(s)$ and $d^{\pi_0}(s,t) = \frac{1}{H}p_t^0(s)$. Hence, with no further inequality,

$$\mathbb{E}_{(s,t)\sim d^{\pi_0}}\left[\left(\tfrac{d^{\pi^\star}(s,t)}{d^{\pi_0}(s,t)}\right)^2\right] = \frac{1}{H}\sum_{t=0}^{H-1} m_t \;\leq\; \frac{1}{H}\sum_{t=0}^{H-1}(1+\epsilon^2)^t = \frac{(1+\epsilon^2)^H - 1}{H\epsilon^2},$$

and subtracting 1 gives the bound. For $H = 1$ the right-hand side is 1, so $C_1 = 0$. □

**Theorem 2** (Finite-horizon improvement from loss minimization)**.** *Let Assumption 1 hold per $(s,t)$, let $\pi_{\theta,t} \ll \pi_{0,t}$ at every state, and for weighted losses let $w \geq |A_t^{\pi_\theta}|$ pointwise. For any certified state-wise loss $L$ from Table 2 with constant $C_L$, and for every $\epsilon > 0$ and $H \geq 1$,*

$$\tfrac{1}{H}\big(\eta(\pi_\theta) - \eta(\pi_0)\big) \;\geq\; \epsilon\, \mathbb{E}_{(s,t)\sim d^{\pi^\star},\, a\sim\pi_0}\big[g\, A^{\pi_0}\big] \;-\; C_L\Big(\mathbb{E}_{d^{\pi_0}}[L] + C_H\sqrt{\mathbb{E}_{d^{\pi_0}}[L^2]}\Big),$$

*with $C_H$ as in Lemma 7.*

*Proof.* Decompose $\eta(\pi_\theta) - \eta(\pi_0) = \big(\eta(\pi^\star) - \eta(\pi_0)\big) - \big(\eta(\pi^\star) - \eta(\pi_\theta)\big)$. The first difference is equation 14; the second is bounded by the finite-horizon form of Lemma 1 by $C_L\, \mathbb{E}_{d^{\pi^\star}}[L]$. Applying Lemma 4 with $f := L$, $Q := d^{\pi^\star}$, $P := d^{\pi_0}$, together with $\mathrm{Var}_{d^{\pi_0}}(L) \leq \mathbb{E}_{d^{\pi_0}}[L^2]$,

$$\mathbb{E}_{d^{\pi^\star}}[L] \leq \mathbb{E}_{d^{\pi_0}}[L] + \sqrt{\chi^2\big(d^{\pi^\star}\|d^{\pi_0}\big)}\sqrt{\mathbb{E}_{d^{\pi_0}}[L^2]},$$

and Lemma 7 bounds the divergence by $C_H^2$. □

**Corollary 2** (Single-loss bound, finite horizon)**.** *Under the conditions of Theorem 2, assume additionally* $L_{\max} := \sup_{s,t} L(\theta; s, t) < \infty$*. Then, with* $\bar{L} := \mathbb{E}_{d^{\pi_0}}[L]$ *or* $\bar{L} := \mathbb{E}_{d^{\pi_0}}[L^2]$*,*

$$\tfrac{1}{H}\big(\eta(\pi_\theta) - \eta(\pi_0)\big) \;\geq\; \epsilon\, \mathbb{E}_{d^{\pi^\star},\, \pi_0}\big[g\, A^{\pi_0}\big] - C_L\, \kappa(\bar{L}),$$

*where* $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is a strict monotonic function. Specifically,* $\kappa(\bar{L}) := \bar{L} + C_H \sqrt{L_{\max}\, \bar{L}}$ *for* $\bar{L} := \mathbb{E}_{d^{\pi_0}}[L]$ *and* $\kappa(\bar{L}) = (1 + C_\epsilon)\sqrt{\bar{L}}$ *for* $\bar{L} := \mathbb{E}_{d^{\pi_0}}[L^2]$*.*

*Proof.* Nonnegativity of $L$ gives $\mathbb{E}_{d^{\pi_0}}[L^2] \leq L_{\max} \bar{L}$. Then we substitute it into Theorem 2. □

Here, the Corollary has no step-size condition. Both sides are expectations under the behavior data distribution. This covers sequence-level language-model fine-tuning with verifiable terminal rewards, where the prompt is the state and the full response the action.

### A.5 Implementation Details and hyperparameter search

The algorithms are implemented with Stable Baselines3 Raffin et al. (2021) for MuJoCo and Atari, RSL-RL Schwarke et al. (2025) for IsaacLab, and Test Time RL (TTRL Zuo et al. (2026)) for LLM fine-tuning. The hyperparameter search range per environment is summarized in Table 5. The resulting optimal hyperparameters are detailed in Table 6. We tune each method in each environment with the same random seed (1), with the grid search based on the final performance. We use the default network architecture in all benchmarks except Hopper, for which we choose the hidden size $256, 256$. We use GAE $\lambda$ 0.95 except for Swimmer, for which we take $0.98$ following RL-Zoo Raffin (2020). The rollout buffer sizes are 2048, 2048, 4096, 1024, 98304, and 98304 for Ant, Hopper, Swimmer, Atari, G1-rough, and Anymal-C-rough. MuJoCo, Atari, and IsaacLab experiments are conducted on CPUs, NVIDIA RTX 2080 Ti GPUs, and NVIDIA RTX 3090 GPUs. For the LLM fine-tuning implementation, we set the discount factor $\gamma = 1$ and use a mini-batch size of 1. We set the group size to 32 and the maximum sequence length to 4,096 tokens. All other hyperparameters follow the original training scripts provided by TTRL (Zuo et al., 2026). All LLM fine-tuning experiments are conducted on $4\times$ NVIDIA H100 GPUs.

Table 5: Hyperparameter search ranges per environment. All methods, including PPO, are tuned over the same grid; the loss column applies to FIRL instances (weighted variants for `_ATV/AMSE` methods, unweighted for `_TV/MSE`). For LLM fine-tuning, all remaining hyperparameters follow the GRPO recipe and only $\epsilon$ and the number of epochs are searched.

| Environment | $\epsilon$ | $\gamma$ | Epochs | Minibatches | Learning rate | Loss |
|---|---|---|---|---|---|---|
| G1-rough | $0.05, 0.1, 0.2, 0.3$ | $0.91, 0.96, 0.99$ | $5, 10$ | $4$ | $10^{-3}$ (Adaptive) | TV, MSE, MSE + TV |
| Anymal-C-rough | $0.1, 0.2, 0.3$ | $0.91, 0.96, 0.99$ | $5, 10$ | $4$ | $10^{-3}$ (Adaptive) | TV, MSE, MSE + TV |
| Ant | $0.1, 0.2, 0.3$ | $0.91, 0.96, 0.99$ | $5, 10$ | $4, 8$ | $1, 3 \times 10^{-4}$ | TV, MSE, MSE + TV |
| Hopper | $0.03, 0.1, 0.2, 0.3$ | $0.96, 0.99, 0.999$ | $5, 10$ | $32, 64$ | $1, 3 \times 10^{-4}$ | TV, MSE, MSE + TV |
| Swimmer | $0.01, 0.1, 0.2, 0.3$ | $0.96, 0.99, 0.9999$ | $5, 10$ | $16, 32$ | $1, 3 \times 10^{-4}$ | TV, MSE, MSE + TV |
| Atari (shared) | $0.1, 0.2, 0.3$ | $0.91, 0.96, 0.99$ | $5, 10$ | $4, 8$ | $2.5 \times 10^{-4}$ | TV, MSE, MSE + TV |
| AMC | $0.1, 0.2$ | — | $5, 10$ | — | — | TV |
| AIME | $0.05, 0.1, 0.2$ | — | $5, 10, 20$ | — | — | TV |

### A.6 Additional Ablation Study

**Centering.** Removing the centering step results in nearly the same performance for both *CLinear_AH* and *CLinear_ATV* (Figure 5). This is expected: the uncentered $\tilde{g}$ is a clipped *standardized* advantage, so its mean deviates from zero only through rare clipping.

**ATV coefficient for AH loss.** Augmenting AH with a weighted-TV term ($\mathcal{L} = L^w_{H_0} + \sigma_2 L^w_{TV}$, Remark 2) has regime-dependent effects: any $\sigma_2 > 0$ reduces return on G1-rough, while larger $\sigma_2$ helps on Hopper and Breakout, where *CLinear_ATV* ranks highest (Table 3-4). Given no clear improvement, we use $\sigma_2 = 0$ for all benchmarking experiments.

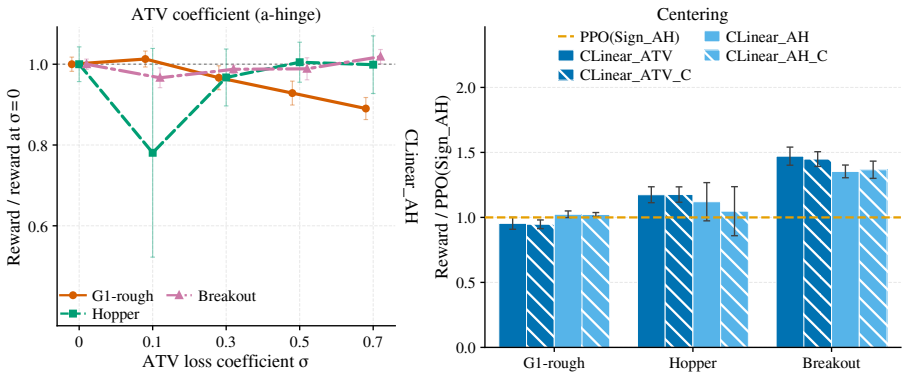



Figure 5: **Ablations of ATV loss coefficient for advantage-weighted Hinge loss and centering**. **Left:** ablation of the ATV loss coefficient $\sigma_2$ to the AH objective, motivated from Remark 2. Rewards are normalized w.r.t. the reward when $\sigma_2 = 0$. **Right:** ablation of centering. Here $_C$ denotes adding centering.

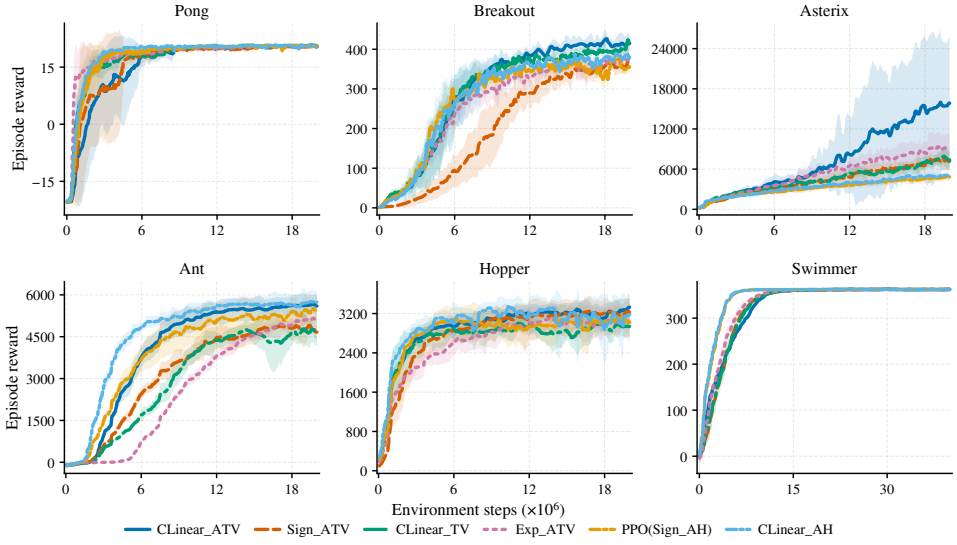



Figure 6: **Training curves** on Atari (top) and MuJoCo (bottom); mean episode reward over seeds, shaded bands one standard deviation. Final statistics in Tables 3–4.

### A.7 Learning Curves

The learning curves for MuJoCo + Atari experiments are detailed in Figure 6. The learning curves for IsaacLab + LLM are detailed in Figure 7. Learning curves for the ablation study on unweighted coef, normalization, loss functions, TV coefficent of AHing loss, and centering are detailed in Figure 8, 9, 10, 12 and 11.

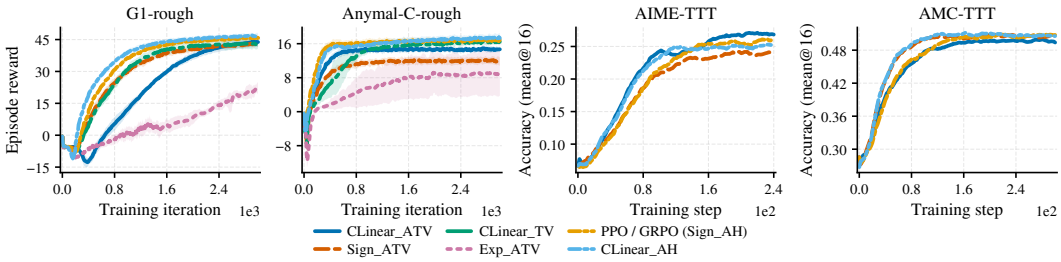



Figure 7: **Training curves** on IsaacLab (episode reward) and AIME (accuracy, mean@16); means over seeds, bands one standard deviation.

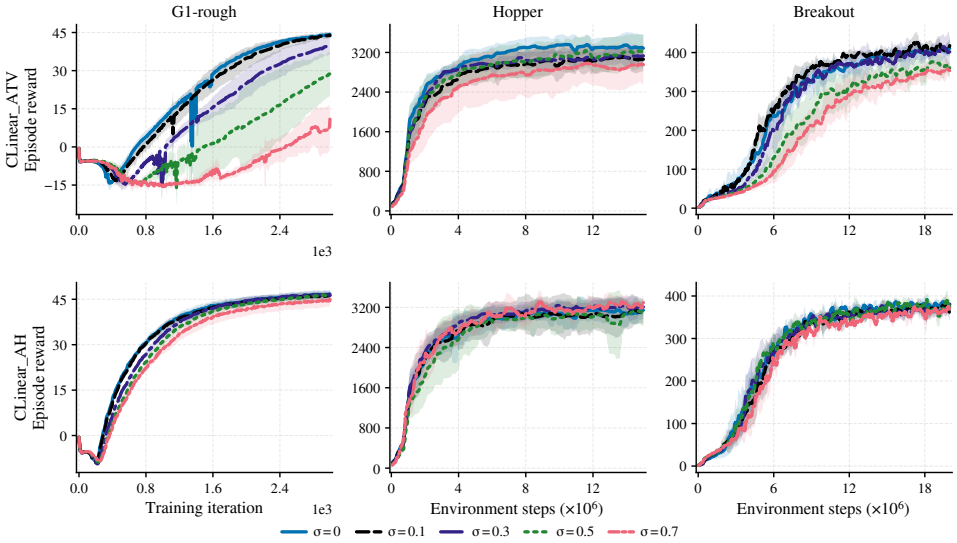



Figure 8: **Unweighted loss coefficient sweep**: `CLinear_ATV/AH` training curves with an added unweighted loss term of weight $\sigma$, other hyperparameters fixed.

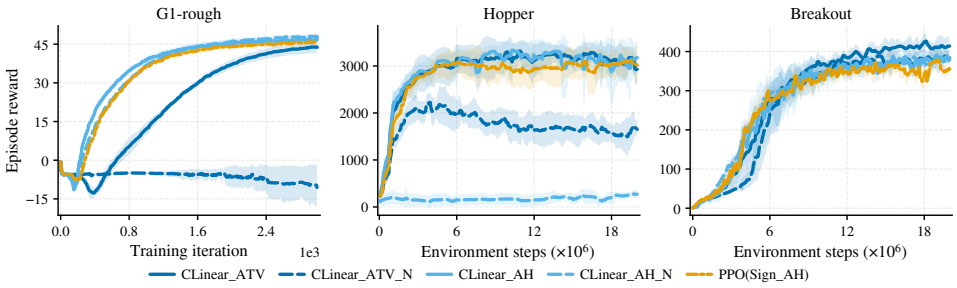



Figure 9: **Normalization ablation**: `CLinear_ATV` with and without batch normalization of the direction (`_N`), PPO as reference; means over seeds, bands one standard deviation.

Table 6: Complete selected hyperparameters for every method and environment. For LLM fine-tuning, only $\epsilon$ and epochs are searched; all other settings follow the default GRPO settings in TTRL.

| Method | $\epsilon$ | $\gamma$ | Epochs | Minibatches | LR | Loss |
|---|---|---|---|---|---|---|
| *G1-rough (IsaacLab)* | | | | | | |
| CLinear_ATV | 0.05 | 0.96 | 10 | 4 | $10^{-3}$ | ATV |
| Sign_ATV/AMSE | 0.2 | 0.96 | 5 | 4 | $10^{-3}$ | AMSE |
| CLinear_TV/MSE | 0.3 | 0.96 | 10 | 4 | $10^{-3}$ | TV |
| Exp_ATV/AMSE | 0.1 | 0.96 | 10 | 4 | $10^{-3}$ | AMSE |
| PPO (Sign_AH) | 0.05 | 0.96 | 10 | 4 | $10^{-3}$ | AH |
| CLinear_AH | 0.05 | 0.96 | 10 | 4 | $10^{-3}$ | AH |
| CLinear_AMSE | 0.1 | 0.96 | 10 | 4 | $10^{-3}$ | AMSE |
| CLinear_KL | 0.3 | 0.96 | 5 | 4 | $10^{-3}$ | KL |
| CLinear_ATV_N | 0.3 | 0.99 | 5 | 4 | $10^{-3}$ | ATV |
| *Anymal-C-rough (IsaacLab)* | | | | | | |
| CLinear_ATV | 0.2 | 0.99 | 5 | 4 | $10^{-3}$ | ATV |
| Sign_ATV/AMSE | 0.3 | 0.99 | 10 | 4 | $10^{-3}$ | AMSE |
| CLinear_TV/MSE | 0.3 | 0.99 | 5 | 4 | $10^{-3}$ | MSE |
| Exp_ATV/AMSE | 0.1 | 0.96 | 5 | 4 | $10^{-3}$ | ATV |
| PPO (Sign_AH) | 0.2 | 0.99 | 10 | 4 | $10^{-3}$ | AH |
| CLinear_AH | 0.1 | 0.99 | 5 | 4 | $10^{-3}$ | AH |
| *Ant (MuJoCo)* | | | | | | |
| CLinear_ATV | 0.1 | 0.99 | 10 | 4 | $10^{-4}$ | ATV |
| Sign_ATV/AMSE | 0.1 | 0.99 | 5 | 8 | $10^{-4}$ | ATV |
| CLinear_TV/MSE | 0.2 | 0.99 | 5 | 4 | $10^{-4}$ | TV |
| Exp_ATV/AMSE | 0.3 | 0.99 | 5 | 4 | $10^{-4}$ | AMSE |
| PPO (Sign_AH) | 0.2 | 0.99 | 10 | 4 | $10^{-4}$ | AH |
| CLinear_AH | 0.1 | 0.99 | 10 | 8 | $10^{-4}$ | AH |
| *Hopper (MuJoCo)* | | | | | | |
| CLinear_ATV | 0.03 | 0.999 | 10 | 64 | $10^{-4}$ | ATV |
| Sign_ATV/AMSE | 0.1 | 0.999 | 10 | 64 | $10^{-4}$ | ATV |
| CLinear_TV/MSE | 0.1 | 0.999 | 10 | 64 | $3\times10^{-4}$ | TV |
| Exp_ATV/AMSE | 0.03 | 0.999 | 5 | 64 | $10^{-4}$ | ATV |
| PPO (Sign_AH) | 0.03 | 0.999 | 10 | 64 | $10^{-4}$ | AH |
| CLinear_AH | 0.03 | 0.999 | 10 | 64 | $10^{-4}$ | AH |
| CLinear_AMSE | 0.1 | 0.99 | 5 | 64 | $10^{-4}$ | AMSE |
| CLinear_KL | 0.3 | 0.999 | 5 | 64 | $10^{-4}$ | KL |
| CLinear_ATV_N | 0.1 | 0.99 | 5 | 32 | $10^{-4}$ | ATV |
| *Swimmer (MuJoCo)* | | | | | | |
| CLinear_ATV | 0.01 | 0.9999 | 10 | 16 | $10^{-4}$ | ATV |
| Sign_ATV/AMSE | 0.01 | 0.9999 | 10 | 32 | $3 \times 10^{-4}$ | A(MSE + TV) |
| CLinear_TV/MSE | 0.01 | 0.9999 | 5 | 32 | $3 \times 10^{-4}$ | TV |
| Exp_ATV/AMSE | 0.01 | 0.9999 | 10 | 32 | $3 \times 10^{-4}$ | A(MSE + TV) |
| PPO (Sign_AH) | 0.01 | 0.9999 | 10 | 32 | $3\times10^{-4}$ | AH |
| CLinear_AH | 0.01 | 0.9999 | 5 | 32 | $3 \times 10^{-4}$ | AH |
| *Atari (shared across Pong, Breakout, Asterix)* | | | | | | |
| CLinear_ATV | 0.2 | 0.99 | 5 | 4 | $2.5\times10^{-4}$ | ATV |
| Sign_ATV/AMSE | 0.2 | 0.96 | 5 | 4 | $2.5\times10^{-4}$ | AMSE |
| CLinear_TV/MSE | 0.3 | 0.96 | 5 | 8 | $2.5\times10^{-4}$ | MSE |
| Exp_ATV/AMSE | 0.3 | 0.96 | 5 | 4 | $2.5\times10^{-4}$ | ATV |
| PPO (Sign_AH) | 0.1 | 0.96 | 5 | 4 | $2.5\times10^{-4}$ | AH |
| CLinear_AH | 0.1 | 0.96 | 5 | 4 | $2.5 \times 10^{-4}$ | AH |
| CLinear_AMSE | 0.2 | 0.99 | 5 | 4 | $2.5\times10^{-4}$ | AMSE |
| CLinear_KL | 0.3 | 0.96 | 10 | 4 | $2.5\times10^{-4}$ | KL |
| CLinear_ATV_N | 0.3 | 0.99 | 5 | 4 | $2.5\times10^{-4}$ | ATV |
| *LLM fine-tuning (AMC)* | | | | | | |
| CLinear_ATV | 0.1 | 1.0 | 10 | — | — | ATV |
| Sign_ATV | 0.2 | 1.0 | 10 | — | — | ATV |
| GRPO (Sign_AH) | 0.1 | 1.0 | 5 | — | — | AH |
| CLinear_AH | 0.1 | 1.0 | 10 | — | — | AH |
| *LLM fine-tuning (AIME)* | | | | | | |
| CLinear_ATV | 0.1 | 1.0 | 20 | — | — | ATV |
| Sign_ATV | 0.1 | 1.0 | 10 | — | — | ATV |
| GRPO (Sign_AH) | 0.1 | 1.0 | 5 | — | — | AH |
| CLinear_AH | 0.05 | 1.0 | 10 | — | — | AH |

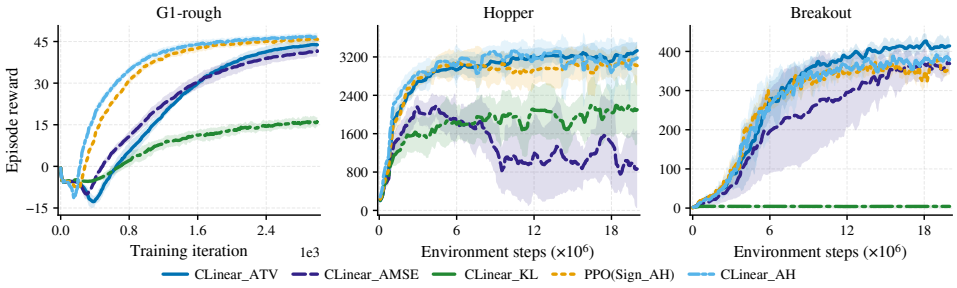



Figure 10: **Loss ablation**: clipped-linear target with ATV, AMSE, AH, and KL losses; PPO as reference; means over seeds, bands one standard deviation.

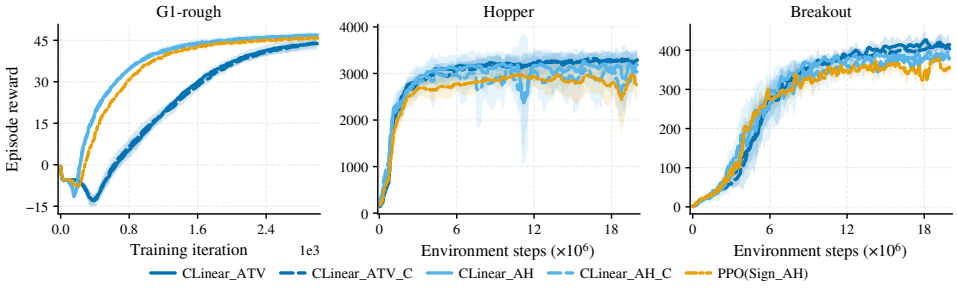



Figure 11: **Centering ablation**: clipped-linear target with ATV/AH KL losses, with or without centering; PPO as reference; means over seeds, bands one standard deviation.

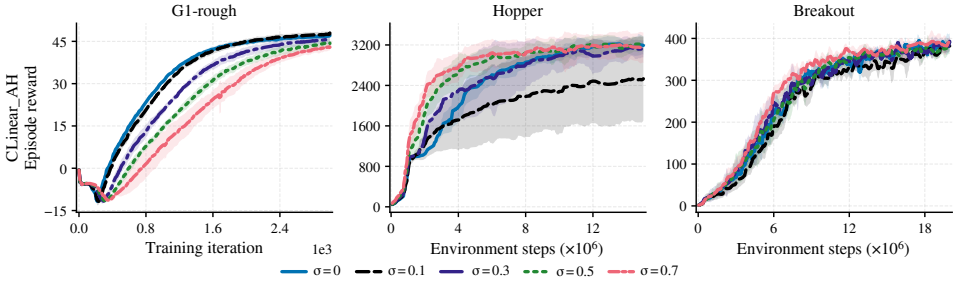



Figure 12: **Ablation on ATV coefficient of advantage-weighted hinge loss**: clipped-linear target with AH loss, with varying ATV coefficients; means over seeds, bands one standard deviation.