TOWARDS BETTER TRAINING SIGNAL: ADVANTAGE CLIPPED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a cornerstone for improving the reasoning capabilities of large language models (LLMs), but the need for on-policy data substantially limits training efficiency. Reusing off-policy data through importance sampling (IS) can improve efficiency but introduce considerable instability. Hence, algorithms such as PPO and GRPO widely adopt IS-ratio clipping to stabilize training. However, training stability and gradient estimate are mainly determined by the product of IS ratio and advantage. To further stabilize training, we propose ACPO, which clips the product of the IS ratio and the advantage, leading to more stable gradient estimates. We also establish a connection between ACPO and gradient clipping in policy mirror descent (PMD), which is a standard technique to stabilize optimization process, and prove the convergence of clipped-PMD under the standard RL setting. Experiments on widely used mathematical reasoning benchmarks show that ACPO consistently outperforms PPO and GRPO in both accuracy and training efficiency, delivering 4-6% points gains on standard math benchmarks, with Qwen3-8B+PPO. Hence, ACPO is a practical and effective alternative to conventional IS-ratio clipping for RL post-training of LLMs.

1 INTRODUCTION

Large language models (LLMs) now serve as a core part for modern natural language understanding and generation, driven by rapid growth in model scale and training data (Brown et al., 2020; Ouyang et al., 2022). However, strong performance on challenging tasks requires more than retrieval and reproducing knowledge; models must also decompose complex problems, integrate intermediate conclusions, generalize across contexts, and generate logically coherent solutions. Reinforcement learning with objectively assessable feedback has recently shown considerable potential for cultivating such abilities, particularly with verifiable, outcome-based rewards (Guo et al., 2025; Lambert et al., 2025). Under this framework, the training objective is to maximize the expected reward of model outputs under a given prompt distribution. The primary challenge lies in achieving efficient and stable training rather than formulating the objective. The need for on-policy data substantially limits training efficiency: on-policy trajectories require generation after each policy update, which is particularly costly for hard reasoning tasks. Reusing off-policy data through importance sampling (IS) can reduce this cost, but the resulting importance weights lead to unstable gradient estimate.

A central approach to stabilizing reinforcement learning for LLM post-training is the actor–critic framework, in which the critic estimates a baseline for the actor's policy-gradient updates, reducing the variance of gradient estimation (Barto et al., 1983). And to address the training instability caused by the partially off-policy trajectories, PPO introduces a clipping mechanism on the IS ratio, thereby constraining excessive policy updates and improving training stability (Schulman et al., 2017).

Building on the clipping mechanism introduced in PPO, numerous studies have proposed alternative clipping strategies centered on the importance-sampling (IS) ratio (MiniMax et al., 2025; Yue et al., 2025; Yu et al., 2025; Ye et al., 2020; Liu & Chen, 2026). Despite this growing body of work, our understanding of ratio clipping remains limited, even for the standard clipping rule used in PPO. However, a principled understanding of when and why this mechanism improves training stability is still lacking. Although the IS ratio measures the deviation of the current policy from the rollout policy, policy update and gradient estimate are mainly determined by the product of IS ratio and advantage, where the advantage determines both the direction and the magnitude of the update.

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.

Our work starts by directly looking at the gradient estimator. Then we try to control the updating magnitude by controlling the product of the importance-sampling (IS) ratio and the advantage, which determines the magnitude and direction of each sample's policy-gradient contribution. We propose **Advantage Clipped Policy Optimization** (ACPO), which clips the product of IS ratio and advantage to $[-\alpha, \alpha]$. Unlike conventional PPO-style methods that clip only the IS ratio, ACPO directly controls the scalar coefficient of each policy-gradient contribution and can be readily integrated into both PPO and GRPO. Under explicit assumptions, we show that ACPO has no larger gradient second moment than PPO and provides a variance-reduction guarantee under an additional assumption.

Another starting point of our work is the policy mirror descent in optimization theory. Clipping is widely used over gradient in optimization, and we want ask is there any connection between gradient clipping and clipping in policy optimization. Our work connects ACPO clipping mechanism to gradient clipping in policy mirror descent and establishes convergence guarantees for the corresponding gradient-clipped PMD in standard RL setting. It is a novel connection between these clipping designs, where gradient clipping has been widely use in both theory and practical optimizer to stabilize optimization process. Our result shows the ACPO clipping principle derived from gradient-clipped PMD also stabilizes the policy optimization process.

1.1 Our Contribution

Algorithm. We propose Advantage Clipped Policy Optimization (ACPO), which applies clipping to the product of the importance-sampling ratio and advantage. The resulting objective retains sample gradient contributions within the clipping range and suppresses those outside it.

Theory. We characterize the second moment of the ACPO gradient estimator and establish a lower second moment than PPO under mild assumptions. Under additional assumption, this comparison yields a variance-reduction guarantee. We further establish a connection to policy mirror descent with gradient clipping, deriving convergence bounds for the associated PMD algorithm in standard RL settings. Together, these results provide an optimization perspective on ACPO.

Experiments. We evaluate ACPO on four mathematical reasoning benchmarks using Qwen2.5-Math-7B and Qwen3 models(1.7B, 4B, and 8B). ACPO improves the best observed weighted accuracy in all eight matched comparisons, yielding gains of 1.6–4.0 percentage points over PPO and 0.6–1.5 percentage points over GRPO, while achieving each baseline's best observed performance in substantially fewer training steps (speedup 6.25x over PPO and 1.7x over GRPO). Moreover, PPO-AC achieves competitive performance relative to GRPO while using only one rollout per prompt.

2 Preliminaries

We first introduce some basic notation of Markov decision process(MDP), which we will use in following paper. We consider a γ-discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $r(s, a)$ denotes the reward function (or $c(s, a)$ the cost function when minimizing cost), $\gamma \in [0, 1)$ is the discount factor, and ρ is the initial-state distribution. A stochastic policy π specifies an action distribution $\pi(\cdot \mid s)$ at each state, and induces trajectories according to $s_0 \sim \rho$, $a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. The performance of π is measured by the expected discounted return $J_\rho(\pi) = \mathbb{E}_{\rho,\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. We denote by $d_\rho^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_{\rho,\pi}(s_t = s)$ the normalized discounted frequency of visiting state s, and by $\mu_\rho^\pi(s, a) = d_\rho^\pi(s)\pi(a \mid s)$ the corresponding discounted state-action visitation distribution. The state-value and action-value functions of π are defined as $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s]$ and $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a]$, respectively. Then, the advantage function is given by the difference of $Q^\pi(s, a)$ and $V^\pi(s)$: $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$, which measures the relative benefit of choosing action a instead of following the policy's average action at state s.

Now we can consider more general Q function that $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \mid s_0 = s, a_0 = a]$ in minimizing cost case. Here h^π is a closed convex function w.r.t. the policy π, i.e., there exists $\mu \geq 0$ satisfying

$$h^\pi(s) - \left[h^{\pi'}(s) + \left\langle (\partial h)^{\pi'}(s, \cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle\right] \geq \mu D_{\pi'}^\pi(s), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product over $\mathcal{A}$, $(\partial h)^{\pi'}(s, \cdot)$ is a subgradient of $\pi \mapsto h^\pi(s)$ at π', and $D^\pi_{\pi'}(s)$ is a Bregman divergence. The Bregman divergence induced by a differentiable and strongly convex mirror map ψ is defined as $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$. Setting $h^\pi = 0$ recovers the standard action-value function; setting $h^\pi(s) = \mu D^\pi_{\pi_0}(s)$ with $\mu > 0$ and D equals to D_{KL}, then Q^π reduces to the so-called entropy regularized action-value function.

2.1 Reinforcement Learning for LLMs

In this section, we briefly review the most popular reinforcement learning methods for large language models. We formulate the generation process of a large language model as a MDP. Let $x \sim \mathcal{D}$ denote an input prompt and $y = (y_1, \ldots, y_T)$ denote the tokens generated by the model. At generation step t, the state is defined as the prompt together with the previously generated tokens $s_t = (x, y_{<t})$, and the action $a_t = y_t$ corresponds to selecting the next token. The policy $\pi_\theta(a_t \mid s_t)$ is parameterized by the policy model, and the transition is deterministic: $s_{t+1} = (x, y_{\le t})$. An episode terminates when an end-of-sequence token is generated or when the maximum generation length is reached.

Proximal Policy Optimization (PPO). PPO (Schulman et al., 2017) stabilizes policy optimization by introducing a clipped surrogate objective. Let $\pi_{\theta_{\text{old}}}$ denote the policy of training trajectories, and define the importance sampling ratio as $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. The clipped PPO objective is

$$\mathcal{J}^{\text{PPO}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\left[\min\left(r_t(\theta)\widehat{A}_t,\ \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\widehat{A}_t\right)\right],$$

where ϵ is the clipping parameter and $\widehat{A}_t$ is an estimate of the advantage function. In practice, the advantage is often computed using generalized advantage estimation (GAE; Schulman et al., 2017): $\widehat{A}_t = \sum_{l=0}^{T-t-1}(\lambda)^l \delta_{t+l}$, and $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$, where V_ϕ is a learned value function. In policy optimization, PPO first introduce the clipping technique and achieve empirical success.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) is a critic-free method that, in practical implementations for LLMs like DeepSeek-R1(Guo et al., 2025), samples G responses $o^{(1)}, \ldots, o^{(G)}$ for each prompt x and computes advantages by normalizing rewards within each prompt group. In the MDP notation above, this corresponds to: $o^{(i)} = \left(a_1^{(i)}, a_2^{(i)}, \ldots, a_{|o^{(i)}|}^{(i)}\right)$ and $s_t^{(i)} = \left(x, a_{<t}^{(i)}\right)$, where $x \sim \mathcal{D}$. Then, the advantage for the i-th response $o^{(i)}$ is computed as: $\widehat{A}^{(i)} = \frac{r\left(x,o^{(i)}\right) - \text{mean}\left(\left\{r\left(x,o^{(1)}\right),\ldots,r\left(x,o^{(G)}\right)\right\}\right)}{\text{std}\left(\left\{r\left(x,o^{(1)}\right),\ldots,r\left(x,o^{(G)}\right)\right\}\right)}$.

This response-level advantage $\widehat{A}^{(i)}$ is then used to replace the step-wise advantage function $\widehat{A}_h(s_h, a_h)$ in the PPO objective $\mathcal{J}^{\text{PPO}}$. The GRPO objective also includes a KL-regularization term:

$$\mathcal{J}^{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{o^{(i)}\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^{(i)}|}\sum_{t=1}^{|o^{(i)}|}\left\{\min\left[r_t^{(i)}(\theta)\widehat{A}^{(i)}, \text{clip}\left(r_t^{(i)}, 1-\varepsilon, 1+\varepsilon\right)\widehat{A}^{(i)}\right] - \beta D_{\text{KL}}\left(\pi \,\|\, \pi_{\text{ref}}\right)\right\}\right],$$

where $r_t^{(i)}(\theta) = \frac{\pi_\theta\left(a_t^{(i)}|s_t^{(i)}\right)}{\pi_{\theta_{\text{old}}}\left(a_t^{(i)}|s_t^{(i)}\right)}$ and GRPO uses the same IS ratio clipping design as PPO.

2.2 Mirror Descent and Policy Mirror Descent

Mirror Descent. Mirror descent (MD) is a first-order optimization method for constrained convex optimization that generalizes gradient descent to non-Euclidean geometries (Beck & Teboulle, 2003). Consider the constrained optimization problem $\min_{x\in\mathcal{X}} f(x)$, where $\mathcal{X}$ is a convex set and $f : \mathcal{X} \to \mathbb{R}$ is a differentiable convex function. Given the current iterate x_k and a gradient estimate g_k, the mirror descent update with step size $\eta_k > 0$ is

$$x_{k+1} = \arg\min_{x\in\mathcal{X}}\left\{\langle g_k, x\rangle + \frac{1}{\eta_k} D_\psi(x, x_k)\right\}.$$

For optimization over the probability simplex $\Delta(\mathcal{A}) = \left\{p \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_{a\in\mathcal{A}} p^a = 1\right\}$, a common choice is the negative entropy $\psi(p) = \sum_{a\in\mathcal{A}} p^a \log p^a$. Then, associated Bregman divergence is the Kullback–Leibler divergence, $D_\psi(p, q) = D_{\mathrm{KL}}(p\|q) = \sum_{a\in\mathcal{A}} p^a \log \frac{p^a}{q^a}$. Under this geometry, mirror descent yields a exponentiated gradient descent update:

$$x_{k+1}^a = \frac{x_k^a \exp(-\eta_k g_k^a)}{\sum_{a'\in\mathcal{A}} x_k^{a'} \exp(-\eta_k g_k^{a'})},$$

where x_{k+1}^a and g_k^a denote the a-th coordinate of x_{k+1} and g_k. This update naturally preserves the simplex constraint and is therefore particularly suitable for optimizing probability distributions.

Policy Mirror Descent. Policy mirror descent (PMD) extends mirror descent to policy optimization in MDP. The main objective in RL to find an optimal policy minimizing the total cost with general Q function can be formulate as an optimization problem with any ρ satisfying $\rho(s) > 0, \forall s$ and $\sum_s \rho(s) = 1$:

$$\min_\pi f(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)] \quad \text{s.t. } \pi(.\mid s) \in \Delta(\mathcal{A}), \forall s \in \mathcal{S}.$$

Then applying MD to this optimization problem gives us the update (see Lan (2023)):

$$\pi_{k+1}(\cdot \mid s) = \arg\min_{p\in\Delta(\mathcal{A})} \left\{ \langle Q^{\pi_k}(s,\cdot), p + h^p(s)\rangle + \frac{1}{\eta_k} D_\psi\left(p, \pi_k(\cdot \mid s)\right) \right\}, \tag{2}$$

where $\eta_k > 0$ is the step size, h^π is a closed convex function w.r.t p and D_ψ is the Bregman divergence induced by a mirror map ψ. For convenience, we also denote $D_\psi(p, \pi_k(\cdot \mid s))$ as $D_{\pi_k}^p(s)$. The first term encourages the policy to assign more probability to actions with high estimated values, while the divergence term constrains the new policy to remain close to the current policy. The action-value function can equivalently be replaced by the advantage function A^{π_k}, since subtracting a state-dependent baseline does not change the solution.

When the negative entropy is used as the mirror map and $h^p = 0$, the update has the closed-form expression $\pi_{k+1}(a \mid s) = \frac{\pi_k(a|s)\exp(\eta_k A^{\pi_k}(s,a))}{\sum_{a'\in\mathcal{A}} \pi_k(a'|s)\exp(\eta_k A^{\pi_k}(s,a'))}$. Thus, PMD exponentially increases the probabilities of actions with positive advantages and decreases those with negative advantages.

3 Advantage Clipped Policy Optimization

TRPO Schulman et al. (2015) maximizes a "surrogate" objective with the trust region method

$$\mathcal{J}^{\mathrm{CPI}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\widehat{A}_t\right] = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\left[r_t(\theta)\widehat{A}_t\right], \tag{3}$$

where CPI refers to conservative policy iteration (Kakade & Langford, 2002). And PPO introduces clipping to this CPI objective to constrain the update of policy during maximizing this objective. Similar to policy gradient theorem (cf. Sutton & Barto, 2018), we have

$$\nabla_\theta J^{\mathrm{CPI}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\left[r_t(\theta)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\widehat{A}_t\right]. \tag{4}$$

A natural way to control the magnitude of policy updates is to control the gradient contribution $r_t(\theta)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\widehat{A}_t$ of each token. However, explicitly computing per-token gradients $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ is computationally expensive and impossible. We therefore examine whether regulating the scalar coefficient $r_t(\theta)\widehat{A}_t$ is sufficient.

Following this idea we design our ACPO objective:

$$\mathcal{J}^{\mathrm{ACPO}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\Big[\mathrm{clip}\big(r_t(\theta)\widehat{A}_t, -\alpha, \alpha\big)\Big], \tag{5}$$

where $\widehat{A}_t$ can be calculated by PPO-type advantage estimate or GRPO-type. We defer the design of the clipping range later. We next present Proposition 1, which establishes that, under mild assumptions, the gradient of the ACPO (G_{ACPO}) has a smaller second moment than that of PPO (G_{PPO}), indicating smaller update magnitudes under gradient sense. The proof can be found in Section A.

Proposition 1. *Let us denote* $\mathbf{1}_{\mathrm{PPO}} = \mathbf{1}_{\{\widehat{A}>0, r(\theta)\le 1+\epsilon \text{ or } \widehat{A}<0, r(\theta)\ge 1-\epsilon\}}$, $\mathbf{1}_{\mathrm{ACPO}} = \mathbf{1}_{\{|r(\theta)\widehat{A}|\le\alpha\}}$. *Suppose that the two clipping rules retain samples with equal probability,* $\mathbb{E}[\mathbf{1}_{\mathrm{PPO}}] = \mathbb{E}[\mathbf{1}_{\mathrm{ACPO}}]$, *and that* $\mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \,\middle|\, r(\theta), \widehat{A}\right] = C$ *almost surely for some finite constant* $C \ge 0$. *Assuming that the second moments below are finite, we have*

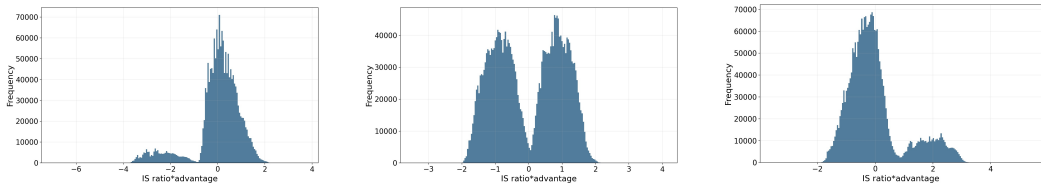
$$\mathbb{E}\|G_{\mathrm{ACPO}}\|^2 = \mathbb{E}\left[\left\|r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{ACPO}}\right\|^2\right] \le \mathbb{E}\left[\left\|r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{PPO}}\right\|^2\right] = \mathbb{E}\|G_{\mathrm{PPO}}\|^2.$$

Remark 1. *The proof does not rely on the specific form of the PPO clipping rule. The same comparison holds for any binary retention rule measurable with respect to* $(r(\theta), \widehat{A})$ *that has the same retention probability as ACPO. Furthermore, if we assume that the gradient estimator after clipping has similar(same) norm* $\|\mathbb{E}[r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{ACPO}}]\| \approx (=)\|\mathbb{E}[r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{PPO}}]\|$, *for example when both are unbiased estimators of the same target gradient, then Proposition 1 implies*

$$\mathrm{Var}(G_{\mathrm{ACPO}}) \le \mathrm{Var}(G_{\mathrm{PPO}}), \text{by } \mathrm{Var}(G) = \mathbb{E}\big[\|G - \mathbb{E}[G]\|^2\big] = \mathbb{E}\big[\|G\|^2\big] - \|\mathbb{E}[G]\|^2.$$

Hence Proposition 1 with additional assumption provides a variance-reduction guarantee of ACPO.

The quality and difficulty of training data play an important role in RL post-training. Prior studies have shown that prompts of intermediate difficulty provide more effective training signals than prompts that are either too easy or too difficult (Zhang et al., 2026; Bae et al., 2026; Gao et al., 2026). However recent work only focuses on prompt selection, for example, Gao et al. (2026); Zhang et al. (2026) introduce adaptive prompt selection and curriculum learning to improve training efficiency. How to account for prompt difficulty directly in the policy optimization algorithm remains less explored. One mechanism linking prompt difficulty to policy optimization is its influence on advantage estimates. Motivated by this connection, we track the distribution of the token-level gradient coefficient $r_t(\theta)\widehat{A}_t$ across prompts of varying difficulty throughout training in Figure 1.



(a) Easy prompt (b) Medium prompt (c) Hard prompt

Figure 1: $r_t(\theta)\widehat{A}_t$ across prompts

As shown in Figure 1, a clearer pattern emerges when comparing the tails of the token-level gradient coefficient $r_t(\theta)\widehat{A}_t$ across difficulty levels. For hard prompts (accuracy 0–30%), the distribution exhibits a pronounced positive tail, indicating that a subset of tokens receives relatively large positive gradient coefficients and therefore contributes disproportionately to reinforcing the current policy. In contrast, for easy prompts (accuracy 70–100%), the distribution develops a substantially heavier negative tail, suggesting that easy prompts are more likely to produce large negative updates for a small fraction of tokens. Meanwhile, medium-difficulty prompts (accuracy 30–70%) show a more concentrated distribution with substantially lighter tails, implying that their token-level updates are comparatively more moderate and stable. Following the principle that **prompts of intermediate difficulty provide more effective training signals**, we introduce a symmetric clipping range $[-\alpha, \alpha]$ for $r_t(\theta)\widehat{A}_t$ to mitigate the influence of the long tail induced by overly easy or overly difficult prompts. This clipping encourages token-level updates to remain within a moderate regime, making their optimization behavior more similar to that observed for prompts of intermediate difficulty.

4 Connection to Gradient Clipping in Policy Mirror Descent

Clipping is widely used in RL post-training, with numerous variants proposed to stabilize policy optimization, while gradient clipping in optimization algorithms likewise plays a central role in stabilizing large-model training. These observations raise a natural question: is there a connection between these two forms of clipping? We establish this connection through policy mirror descent, showing that applying gradient clipping within this framework yields the clipping objective of ACPO. This derivation links gradient clipping in optimization to clipping in post-training.

Algorithm 1 The policy mirror descent (PMD) method with gradient clipping

1: **Input:** initial points π_0, $\pi_0(a \mid s) = 1/|\mathcal{A}|$, clipping threshold α and stepsizes $\eta_k \geq 0$.
2: **for** $k = 0, 1, \ldots$ **do**
3:

$$\pi_{k+1}(\cdot \mid s) = \underset{p(\cdot|s)\in\Delta(\mathcal{A})}{\arg\min} \left\{ \eta_k \frac{\alpha}{\max\{\alpha, \|A^{\pi_k}(s,\cdot)\|\}} \left[\langle A^{\pi_k}(s,\cdot), p(\cdot \mid s)\rangle + h^p(s)\right] + D^p_{\pi_k}(s) \right\}, \forall s \in \mathcal{S}$$

4: **end for**

Introducing gradient clipping in (2) gives us Algorithm 1, because $A^{\pi_k}(s,\cdot)$ serves as gradient in the update. Then starting from PMD with gradient clipping and $h^p = 0$, we first apply importance sampling to express the advantage term using samples from the old policy, introducing the importance ratio $r_\theta(s,a) = \frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}$. Then we restrict the optimization to a parameterized policy class, approximately solve the resulting subproblem using multiple stochastic gradient steps (see Section B for details). Finally, we remove the explicit KL proximity penalty arising from the mirror map, obtaining ACPO objective 5. Omitting the KL penalty isolates the role of clipping and enables us to investigate whether ACPO can maintain stable training without explicit KL regularization. Moreover, removing the KL penalty is consistent with the experimental settings in recent papers (Yu et al., 2025; Xiong et al., 2025).

This construction connects ACPO to the clipping principle underlying PMD while introducing a different clipping mechanism. During the translation of the clipping principle into a sample-based surrogate, we replace PMD's statewise advantage rescaling with direct clipping of the importance-weighted advantage $r_t(\theta)\widehat{A}_t$. Differentiating this surrogate preserves each sample's gradient contribution within the clipping interval and suppresses it outside the interval.

Now, we present convergence results of Algorithm 1 and proofs are presented in Section C. Here, we take the Bregman Divergence induced by negative entropy that $D^{\pi'}_{\pi} = D_{\mathrm{KL}}(\pi'||\pi)$. Following Lan (2023), let ν^* be the stationary state distribution induced by the optimal policy π^*, we consider the convergence criterion of the following optimization problem with specific distribution ν^*:

$$\min_{\pi} f(\pi) := \mathbb{E}_{s\sim\nu^*}[V^\pi(s)] \quad \text{s.t. } \pi(\cdot \mid s) \in \Delta(\mathcal{A}), \forall s \in \mathcal{S},$$

then we define $F_k = f(\pi_k) - f(\pi^*) = E_{s\sim\nu^*}[V_{\pi_k}(s)] - E_{s\sim\nu^*}[V_{\pi^*}(s)]$ and $\mathcal{D}_k = E_{s\sim\nu^*}[D^{\pi^*}_{\pi_k}(s)]$. We analyze convergence under constant and adaptive step sizes, both with strongly convex regularization and without regularization. We show that Algorithm 1 can achieve a linear convergence rate for solving RL problems with strongly convex regularizers and a sublinear rate without it. For convenience, we define

$$\lambda_k(s) := \frac{\alpha}{\max\{\alpha, \|A^{\pi_k}(s,\cdot)\|\}}, \qquad \kappa_0 := \min_{s\in\mathcal{S}} \lambda_0(s).$$

Theorem 1 (Convergence with strongly convex regularizer)**.** *Suppose h^p is μ-strongly convex with $\mu > 0$, $D^{\pi'}_{\pi} = D_{\mathrm{KL}}(\pi'||\pi)$, cost function $c(s,a)$ and h^p are bounded. Then the following bounds hold for all $N \geq 1$:*

1. ***Constant step sizes.*** *There exists a constant $1 \geq \kappa > 0$, for $\eta_k = \eta$ for all $k \geq 0$, we have*

$$F_N + (\mu + \frac{1}{\eta})\mathcal{D}_N \leq \rho^N \left[F_0 + (\mu + \frac{1}{\eta}) \log|\mathcal{A}| \right],$$

   where $\rho = \max\left\{\gamma, \frac{1}{\kappa(1+\eta\mu)}\right\}$. If $\eta\mu > 1/\kappa - 1$, then $\rho < 1$.

2. ***Adaptive step sizes.*** *For $\eta_0 > 0$, choose $\eta_{k+1} = \eta_k \max\left\{1, \max_{s\in\mathcal{S}} \frac{\lambda_k(s)}{\lambda_{k+1}(s)}\right\}$. Then*

$$F_N \leq \rho_0^N \left[F_0 + \left(\mu + \frac{1}{\eta_0\kappa_0}\right) \log|\mathcal{A}| \right],$$

   where $\rho_0 = \max\left\{\gamma, \frac{1}{1+\mu\eta_0\kappa_0}\right\} < 1$.

Theorem 2 (Convergence without regularizer). *Suppose $h^p = 0$, $D_\pi^{\pi'} = D_{\mathrm{KL}}(\pi'||\pi)$, cost function $c(s,a)$ and h^p are bounded. Then the following bounds hold for all $N \geq 1$:*

1. ***Constant step sizes.*** *There exists a constant $C_0 > 0$, for $\eta_k = \eta$ for all $k \geq 0$, we have*

$$F_N \leq \frac{C_0}{(1-\gamma)N}\left[\gamma F_0 + \frac{\log|\mathcal{A}|}{\eta\kappa_0}\right].$$

2. ***Adaptive step sizes.*** *For $\eta_0 > 0$, choose $\eta_{k+1} = \eta_k \max\left\{1, \max_{s\in\mathcal{S}} \frac{\lambda_k(s)}{\lambda_{k+1}(s)}\right\}$. Then*

$$F_N \leq \frac{1}{(1-\gamma)N}\left[\gamma F_0 + \frac{\log|\mathcal{A}|}{\eta_0\kappa_0}\right].$$

5 Experiments

5.1 Experimental Setup.

In this section, we present experiments to evaluate the performance of ACPO on reasoning tasks. We incorporate PPO and GRPO with ACPO and denote it as PPO-AC and GRPO-AC. We compare GRPO-AC and PPO-AC against two standard IS ratio clipping methods: GRPO and PPO. **Datasets and models.** We train our models using the training set from DAPO (Yu et al., 2025), which comprises approximately 17.4k math problems sourced from the AoPS website and official competition homepages. We employ the Math-Verify tool (Kydlíček, 2025) for automatic solution correctness verification. To demonstrate the generality of our method across model scales, we conduct experiments using **Qwen2.5-Math-7B** and **Qwen3(1.7B, 4B and 8B)**.

Evaluation. We assess the models' reasoning abilities on several standard mathematical reasoning benchmarks: **MATH500** (Hendrycks et al., 2021), **Minerva Math** (Lewkowycz et al., 2022), **OlympiadBench** (He et al., 2024) and **AIME-like** (Xiong et al., 2025), which consists of 230 problems from recent competitions: AIME24, AIME25, HMMT24, HMMT25, BRUMO25, AMC23, and CMIMC25. All evaluations report the Pass@1 accuracy averaged over 16 or 32 samples, generated with a temperature of 1.0, top-p=1 and a max token limit of 8196.

RL Training Details. We conduct all experiments using the VERL reinforcement learning framework (Sheng et al., 2025). In PPO, we generate one response per prompt ($n = 1$), for GRPO we generate four responses per prompt ($n = 4$) for Qwen3 and eight responses per prompt ($n = 8$) for Qwen2.5-Math-7B. Following prior work (Schulman et al., 2017; Yu et al., 2025), we set the standard clipping range to 0.2 for PPO and $[0.2, 0.28]$ for GRPO. For PPO-AC, we set $\alpha = 3$ in all experiments. For GRPO-AC, we set $\alpha = 3$ for Qwen2.5-Math-7B and $\alpha = 2$ for Qwen3. More details of our training setting can be found in Appendix D.1.

5.2 Main Results

We conduct experiments on Qwen2.5-Math-7B and Qwen3 (1.7B, 4B and 8B). We consider the weighted accuracy(weighted by problem amount on each benchmark).The evaluation performance curves and best evaluation results throughout training are presented in Table 1 and Figure 2, respectively. The effectiveness of our method is demonstrated in Table 1. Across four model configurations and four standard mathematical reasoning benchmarks, incorporating ACPO consistently improves both PPO and GRPO. In particular, PPO-AC increases the weighted accuracy over PPO by 1.6, 2.4, 2.7, and 4.0 percentage points on Qwen2.5-7B-Math, Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively, demonstrating increasingly pronounced gains on stronger models. Similar improvements are observed for GRPO-AC, which consistently outperforms its corresponding GRPO baseline in weighted accuracy. The gains are especially substantial on the more challenging OLYMPIAD and AIME-like benchmarks. For example, on Qwen3-8B, PPO-AC improves OLYMPIAD accuracy from 63.7% to 70.5% and AIME-like accuracy from 43.6% to 48.8%. These results demonstrate that ACPO provides robust and broadly applicable improvements across different policy optimization algorithms, model families, and model scales.

Figure 2 illustrates the step-wise weighted accuracy averaged across all benchmarks for ACPO and the baseline methods. Although all approaches exhibit improved reasoning performance as training

progresses, ACPO consistently converges faster and achieves higher performance than baselines. In particular, GRPO-AC maintains the best performance throughout the entire training process. Moreover, PPO-AC consistently outperforms PPO and achieves performance comparable to GRPO, while GRPO requires multiple rollouts per prompt.

Table 1: Performance of methods on math benchmarks, where we choose the best weighted accuracy result during training. Bold marks the better result within each matched PPO or GRPO pair.

Method	Weighted Acc.	MATH500	MINERVA	OLYMPIAD	AIME-like
Qwen2.5-7B-Math (avg@32)					
PPO	48.4	79.0	33.0	41.2	21.4
PPO-AC	**50.0**	**80.5**	**35.3**	**42.8**	**21.9**
GRPO	51.2	82.0	37.2	43.2	24.1
GRPO-AC	**52.6**	**82.8**	**37.8**	**45.6**	**24.9**
Qwen3-1.7B (avg@16)					
PPO	57.3	86.2	39.2	52.8	28.8
PPO-AC	**59.7**	**88.1**	**39.7**	**56.9**	**30.1**
GRPO	60.6	88.2	39.5	58.9	30.8
GRPO-AC	**62.1**	**89.4**	**40.3**	**60.3**	**34.3**
Qwen3-4B (avg@16)					
PPO	66.7	93.3	45.4	63.8	43.3
PPO-AC	**69.4**	**94.0**	**46.4**	**68.3**	**46.3**
GRPO	70.0	94.7	46.6	69.2	46.4
GRPO-AC	**70.6**	**94.8**	**46.8**	**69.8**	**48.5**
Qwen3-8B (avg@16)					
PPO	67.3	93.1	48.6	63.7	43.6
PPO-AC	**71.3**	**94.8**	**48.9**	**70.5**	**48.8**
GRPO	70.6	**95.0**	48.7	69.1	47.3
GRPO-AC	**71.4**	**95.0**	**49.4**	**70.4**	**49.0**

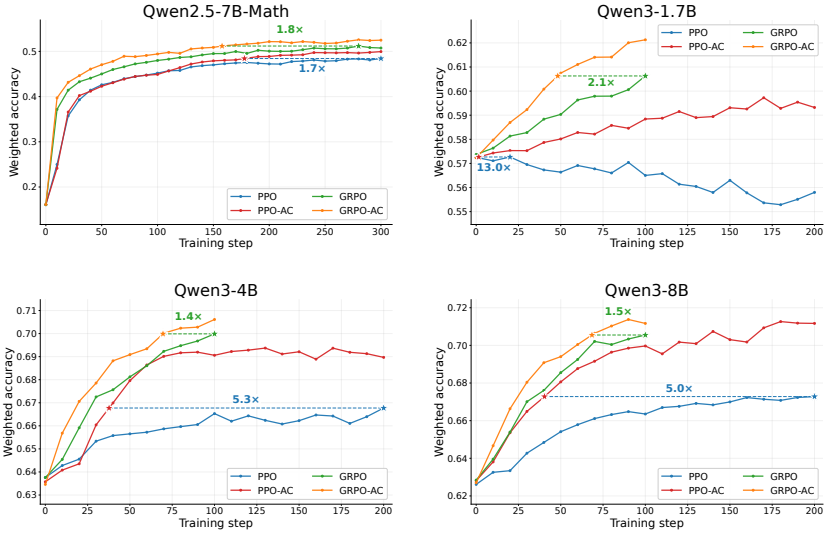


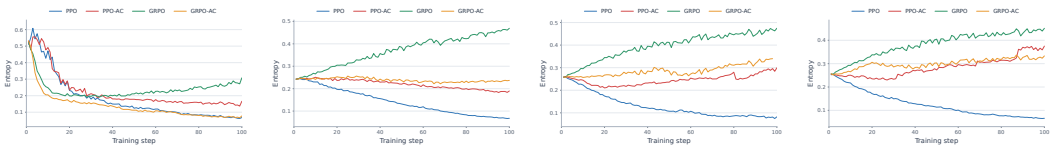
Figure 2: Accuracy during Training Process

5.3 Training Dynamics and Analysis

Training efficiency. Our clipping mechanism in ACPO stabilizes policy optimization from both policy-gradient and mirror-descent perspectives while making more effective use of the training signal. As shown in Figure 2, ACPO achieves average training-step speedups of 6.25× over PPO and 1.7× over GRPO, measured by the steps required to match each baseline's best accuracy. For

Qwen3-4B and Qwen3-8B, PPO-AC matches the best accuracy attained by PPO over 200 steps in approximately 40 steps (5.3× and 5.0×). Especially on Qwen3-1.7B, PPO yields little initial improvement and its accuracy declines after 20 steps, whereas PPO-AC continues to improve model accuracy, suggesting greater robustness in this setting. These results confirm that ACPO improves accuracy while accelerating training speed, with larger efficiency gains over PPO than over GRPO. This difference arises from how the two methods estimate advantages: PPO uses token-level advantages derived from a value model, whereas GRPO assigns the same advantage to all tokens within a sequence. This finer-grained signal allows PPO to benefit more from ACPO's clipping mechanism.

Entropy. Vanilla PPO and GRPO can exhibit rapid entropy decay, limiting exploration and prematurely concentrating the policy on a narrow set of reasoning trajectories (Yu et al., 2025; Yue et al., 2025; Cui et al., 2025). The standard clipping mechanism in PPO-style objectives further exacerbates this issue by constraining probability increases for low-probability exploratory tokens. As shown in Figure 3, ACPO substantially alleviates entropy collapse during training. The policy entropy of GRPO-AC remains relatively stable and fluctuates around its initial value, whereas that of GRPO increases rapidly with the use of the Clip-Higher technique(Yu et al., 2025). Although the entropy of PPO-AC still declines during training, its decrease is considerably slower than that of PPO. Our clipping mechanism in ACPO preserves the contributions of tokens with relatively small advantages, which helps sustain exploration and policy diversity.



(a) Qwen2.5-7B-Math (b) Qwen3-1.7B (c) Qwen3-4B (d) Qwen3-8B

Figure 3: Policy entropy over the first 100 training steps across four models.

6 Related Work

Mirror Descent for Policy Optimization. Mirror descent provides a principled framework for constrained optimization and has been widely applied to policy optimization. Tomar et al. (2022) propose mirror descent policy optimization (MDPO), which updates the policy by approximately solving a trust-region subproblem. Song et al. (2026) use the Bellman equation to reformulate policy mirror descent (PMD) as a trajectory-level objective and derive a practical BPO loss through approximation. While these studies derive practical objectives from mirror descent, our work takes a novel perspective: using PMD to understand and motivate clipping mechanisms in policy optimization.

Clipping Mechanisms. Clipping mechanisms have received considerable attention in policy optimization and post-training (Yu et al., 2025; MiniMax et al., 2025; Jia et al., 2026; Zheng et al., 2025). These methods primarily modify how importance ratios are clipped. DAPO introduces Clip-Higher, using clipping range $[0.2, 0.28]$ to encourage exploration (Yu et al., 2025). CISPO clips the importance-sampling weights while retaining the corresponding gradient contributions (MiniMax et al., 2025). From token level to sequence level, Zheng et al. (2025) design clipping over trajectory, while GAPO adapts the IS clipping range to trajectory advantages in GRPO (Jia et al., 2026).

7 Conclusion

We introduced Advantage Clipped Policy Optimization (ACPO), which directly clips the product of IS ratio and advantage to stabilize policy optimization. ACPO can be readily integrated into PPO and GRPO, admits an interpretation through gradient-clipped policy mirror descent, and consistently improves accuracy and training efficiency across multiple Qwen models and mathematical-reasoning benchmarks. These results demonstrate that advantage clipping is a simple and effective alternative to conventional IS-ratio clipping for reinforcement-learning post-training of LLMs.

AI USE STATEMENT

In this paper, we employed Large Language Models (LLMs) solely for polishing the writing. And all the outputs from LLMs were treated as references or drafts, subject to human review, modification, and validation. We do not use any LLMs to generate the technical content, experimental results, or conclusions.

ETHICS STATEMENT

This research follows the ICLR Code of Ethics. We have carefully considered the ethical implications of the datasets, methods, and models used in our experiments. Our work focuses on theoretical and empirical advances in policy optimization algorithm and reasoning models, and does not involve sensitive personal data or content likely to contribute to societal bias, harm, or discrimination.

REPRODUCIBILITY STATEMENT

We present the experimental settings and hyperparameters in Section 5, with additional details in Section D.1. Unless otherwise stated, no additional hyperparameter tuning is performed. All datasets and models used in our experiments are open sourced. For theoretical results, we provide detailed proof in Section A and C, and clarify all the assumptions and notations we used.

REFERENCES

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning, 2026. URL https://arxiv.org/abs/2504.03380.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, May 2003. ISSN 0167-6377. doi: 10.1016/S0167-6377(02)00231-6. URL https://doi.org/10.1016/S0167-6377(02)00231-6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

Zhaolin Gao, Joongwon Kim, Wen Sun, Thorsten Joachims, Sid Wang, Richard Yuanzhe Pang, and Liang Tan. Prompt curriculum learning for efficient LLM post-training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zqOCacBD3P.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng,

Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Sheng Jia, Xiao Wang, Shiva Prasad Kasiviswanathan, and Rein Houthooft. Group adaptive clipping policy optimization, 2026. URL https://arxiv.org/abs/2609.00444.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Hynek Kydlíček. Math-verify: Math verification library. https://github.com/huggingface/Math-Verify, 2025. Accessed: 2026-07-16.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Guanghui Lan. Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198(1):1059–1106, 2023. doi: 10.1007/s10107-022-01816-5. URL https://doi.org/10.1007/s10107-022-01816-5.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam

Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Chi Liu and Xin Chen. Adaptive-boundary-clipping grpo: Ensuring bounded ratios for stable and generalizable training, 2026. URL https://arxiv.org/abs/2601.03895.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiguang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jiaren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Linge Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujie Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songquan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, Yonghong Duan, Yongxiang Fu, Yongyi Hu, Yu Gao, Yuanxiang Fan, Yufeng Yang, Yuhao Li, Yulin Hu, Yunan Huang, Yunji Li, Yunzhi Xu, Yuxin Mao, Yuxuan Shi, Yuze Wenren, Zehan Li, Zelin Li, Zhanxu Tian, Zhengmao Zhu, Zhenhua Fan, Zhenzhen Wu, Zhichao Xu, Zhihang Yu, Zhiheng Lyu, Zhuo Jiang, Zibo Gao, Zijia Wu, Zijian Song, and Zijun Sun. Minimax-m1: Scaling test-time compute efficiently with lightning attention, 2025. URL https://arxiv.org/abs/2506.13585.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Zhuoqing Song, Haotian Xu, Xikun Zhang, and Lidong Bing. Bellman policy optimization, 2026. URL https://arxiv.org/abs/2609.15987.

R.S. Sutton and A.G. Barto. *Reinforcement Learning, second edition: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 9780262039246. URL https://books.google.com.hk/books?id=sWV0DwAAQBAJ.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBO5SvgSt1.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework under non-linear rl objectives, 2025. URL https://arxiv.org/abs/2510.04996.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. Mastering complex control in moba games with deep reinforcement learning, 2020. URL https://arxiv.org/abs/1912.09729.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks, 2025. URL https://arxiv.org/abs/2504.05118.

Ruiqi Zhang, Daman Arora, Song Mei, and Andrea Zanette. Speed-rl: Faster training of reasoning models via online curriculum learning, 2026. URL https://arxiv.org/abs/2506.09016.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Appendix

In Section A and C, we provide detailed proofs of the proposition and theorem in our main pages. Section B discusses how to derive our ACPO objective from PMD with gradient clipping step by step. And in Section D, we present more experiment results and settings.

A Proof of Proposition 1

Proposition A.1 (cf. Proposition 1). *Let we denote* $\mathbf{1}_{\mathrm{PPO}} = \mathbf{1}_{\{\widehat{A}>0, r(\theta)\le 1+\epsilon \text{ or } \widehat{A}<0, r(\theta)\ge 1-\epsilon\}}$, $\mathbf{1}_{\mathrm{ACPO}} = \mathbf{1}_{\{|r(\theta)\widehat{A}|\le\alpha\}}$. *Suppose that the two clipping rules retain samples with equal probability,* $\mathbb{E}[\mathbf{1}_{\mathrm{PPO}}] = \mathbb{E}[\mathbf{1}_{\mathrm{ACPO}}]$ *, and that* $\mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \,\middle|\, r(\theta), \widehat{A}\right] = C$ *almost surely for some finite constant* $C \ge 0$. *Assuming that the second moments below are finite, we have*

$$\mathbb{E}\left[\left\|r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{ACPO}}\right\|^2\right] \le \mathbb{E}\left[\left\|r(\theta)\widehat{A}\nabla_\theta \log \pi_\theta(a \mid s)\mathbf{1}_{\mathrm{PPO}}\right\|^2\right].$$

Proof. Note that $(|r(\theta)\widehat{A}|^2 - \alpha^2)(\mathbf{1}_{\mathrm{PPO}} - \mathbf{1}_{\mathrm{ACPO}}) \geq 0$ for any (s, a) by the definition of $\mathbf{1}_{\mathrm{ACPO}}$ and denote the gradient estimator as G_{PPO}, G_{ACPO}.

$$\begin{aligned}
&\mathbb{E}\|G_{\mathrm{PPO}}\|^2 - \mathbb{E}\|G_{\mathrm{ACPO}}\|^2 \\
=&\mathbb{E}[|r(\theta\widehat{A})|^2\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2(\mathbf{1}_{\mathrm{PPO}} - \mathbf{1}_{\mathrm{ACPO}})] \\
\geq&\mathbb{E}[\alpha^2\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2(\mathbf{1}_{\mathrm{PPO}} - \mathbf{1}_{\mathrm{ACPO}})] \\
=&\alpha^2\mathbb{E}\left[(\mathbf{1}_{\mathrm{PPO}} - \mathbf{1}_{\mathrm{ACPO}})\mathbb{E}[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \mid r(\theta), \widehat{A}]\right] \\
=&\alpha^2 C\left(\mathbb{E}[\mathbf{1}_{\mathrm{PPO}}] - \mathbb{E}[\mathbf{1}_{\mathrm{ACPO}}]\right) \\
=&0.
\end{aligned}$$

This proves the claimed inequality. □

B TRANSFORMATION OF PMD WITH GRADIENT CLIPPING

In this section we discuss how to transfer PMD with gradient clipping Algorithm 1 to our ACPO objective 5.

First, we look at the PMD update 2 to solve minimizing total cost problem. The update takes following form:

$$\pi_{k+1}(\cdot \mid s) = \underset{p \in \Delta(\mathcal{A})}{\arg\min} \left\{ \langle Q^{\pi_k}(s, \cdot), p + h^p(s) \rangle + \frac{1}{\eta_k} D_\psi\left(p, \pi_k(\cdot \mid s)\right) \right\}, \forall s \in \mathcal{S}.$$

Here we replace $Q^{\pi_k}(s, \cdot)$ with $A^{\pi_k}(s, \cdot)$. Note this replacement does not the subproblem or its optimal solution. Then in the update, the term $A^{\pi_k}(s, \cdot)$ serves as gradient. Combining the gradient clipping technique, we have

$$\pi_{k+1}(\cdot \mid s) = \underset{p \in \Delta(\mathcal{A})}{\arg\min} \left\{ \frac{\alpha}{\max\{\alpha, \|A^{\pi_k}(s, \cdot)\|\}} \langle A^{\pi_k}(s, \cdot), p + h^p(s) \rangle + \frac{1}{\eta_k} D_\psi\left(p, \pi_k(\cdot \mid s)\right) \right\}, \forall s \in \mathcal{S},$$

which is equivalent to Algorithm 1.

Next, we take $h^p = 0$ and $D_\psi\left(p, \pi_k(\cdot \mid s)\right) = D_{\mathrm{KL}}(p, \pi_k)$. To match the objective of PPO, which maximizes the total reward, we consider the maximizing problem formulation. Then, we come to

$$\pi_{k+1}(\cdot \mid s) = \underset{p \in \Delta(\mathcal{A})}{\arg\max} \left\{ \frac{\alpha}{\max\{\alpha, \|A^{\pi_k}(s, \cdot)\|\}} \langle A^{\pi_k}(s, \cdot), p \rangle - \frac{1}{\eta_k} D_{\mathrm{KL}}\left(p, \pi_k(\cdot \mid s)\right) \right\}, \forall s \in \mathcal{S},$$

which is equivalent to

$$\pi_{k+1}(\cdot \mid s) \leftarrow \underset{\pi \in \Pi}{\arg\max}\, \mathbb{E}_{s \sim \rho_{\pi_k}, a \sim \pi}\left[\mathrm{clip}(A^{\pi_k}(s, \cdot), \alpha)\right] - \frac{1}{t_k} D_{\mathrm{KL}}(\pi, \pi_k),$$

if ρ_{π_k} has full support.

Here we move the clipping to each $A^{\pi_k}(s, a)$, because Under a parameterized policy, the gradient contribution comes from each $A^{\pi_k}(s, a)$, then obtain

$$\pi_{k+1}(\cdot \mid s) \leftarrow \underset{\theta}{\arg\max}\, \mathbb{E}_{(s,a) \sim \pi_{\theta_k}}\left[\mathrm{clip}(A^{\pi_{\theta_k}}(s, a), -\alpha, \alpha) - \frac{1}{t_k} D_{\mathrm{KL}}(\pi, \pi_k)\right].$$

Finally, we solve this subproblem by several optimization steps and then update the θ_k. Because we need to use the estimated advantage from trajectory generated by old policy and integrate importance sampling, the objective is

$$\max_\theta\, \mathbb{E}_{(s,a) \sim \pi_{\mathrm{old}}}\left[\mathrm{clip}(r(\theta)\widehat{A}(s, a), -\alpha, \alpha) - \frac{1}{t_k} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{old}})\right].$$

After removing the D_{KL} term, it comes to our ACPO objective 5.

C Proof of Theorem 1 and 2

This section gives a convergence analysis for policy mirror descent (PMD) with gradient clipping. Under relative strong convexity this objective-preserving update converges linearly. In the unregularized case, the same update reduces to advantage clipping and enjoys a last-iterate $O(1/T)$ guarantee, including for every fixed constant stepsize.

C.1 Notation

Let ψ be a differentiable and strongly convex mirror map on the relative interior of $\Delta(\mathcal{A})$. We use $D_\psi(x, y)$ to denote the Bregman divergence of ψ between x and y:

$$D_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle . \tag{6}$$

In particular, for policies π and π', we use the following notation:

$$D^\pi_{\pi'}(s) := D_\psi\big(\pi(\cdot \mid s), \pi'(\cdot \mid s)\big). \tag{7}$$

We also write $(\partial h)^{\pi'}(s, \cdot)$ for a subgradient of $\pi \mapsto h^\pi(s)$ at π'.

Then we consider a γ-discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $c(s, a)$ denotes the cost function (or $r(s, a)$ the reward function when maximizing reward), $\gamma \in [0, 1)$ is the discount factor, and ρ is the initial-state distribution. A stochastic policy π specifies an action distribution $\pi(\cdot \mid s)$ at each state, and induces trajectories according to $s_0 \sim \rho$, $a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Additionally, we consider a state-separable policy regularizer $h^\pi(s)$, and we say that h is μ-strongly convex relative to policy, if for every π, π',

$$h^\pi(s) \geq h^{\pi'}(s) + \Big\langle (\partial h)^{\pi'}(s, \cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s)\Big\rangle + \mu D^\pi_{\pi'}(s), \qquad \mu \geq 0. \tag{8}$$

The state-value and action-value functions of π are defined as

$$V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t (c(s_t, a_t) + h^\pi(s_t)) \mid s_0 = s], \tag{9}$$

and

$$Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t (c(s_t, a_t) + h^\pi(s_t)) \mid s_0 = s, a_0 = a] = c(s, a) + h^\pi(s) + \gamma \sum_{s'} P(s' \mid s, a) V^\pi(s'). \tag{10}$$

Then, advantage function is given by the difference of $Q^\pi(s, a)$ and $V^\pi(s)$:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \tag{11}$$

Consider an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^{|\mathcal{A}|}$ and a clipping threshold $\alpha > 0$. At iteration k, set

$$\lambda_k(s) := \frac{\alpha}{\max\{\alpha, \|A^{\pi_k}(s, \cdot)\|\}} \tag{12}$$

The stepsize $\eta_k > 0$ is a single scalar independent of states. The objective-preserving clipped PMD update is

$$\pi_{k+1}(\cdot \mid s) \in \arg\min_{p \in \Delta(\mathcal{A})} \Big\{ \eta_k \lambda_k(s) \big[\langle A^{\pi_k}(s, \cdot), p\rangle + h_s(p)\big] + D^p_{\pi_k(\cdot|s)}(s) \Big\}. \tag{13}$$

Let π^* be an optimal stationary policy, so $V^{\pi^*}(s) \leq V^\pi(s)$ for all s and all stationary π. Let ν^* be the distribution of the optimal policy, and define

$$e_k(s) := V^{\pi_k}(s) - V^{\pi^*}(s), \qquad F_k := \mathbb{E}_{s\sim\nu^*}[e_k(s)], \tag{14}$$

$$D_k(s) := D^{\pi^*(\cdot|s)}_{\pi_k(\cdot|s)}(s), \qquad E_k(s) := D^{\pi_{k+1}(\cdot|s)}_{\pi_k(\cdot|s)}(s). \tag{15}$$

In particular, $e_k(s) \geq 0$ and $F_k \geq 0$. For the KL divergence we take the uniform initialization $\pi_0(a \mid s) = 1/|\mathcal{A}|$, which gives

$$D_0(s) = D_{\mathrm{KL}}\big(\pi^*(\cdot \mid s) \| \pi_0(\cdot \mid s)\big) \leq \log |\mathcal{A}|. \tag{16}$$

C.2 USEFUL LEMMAS

For policies π and π', define the local improvement quantity

$$g_{\pi,\pi'}(s) := \langle A^{\pi}(s,\cdot), \pi'(\cdot \mid s) - \pi(\cdot \mid s)\rangle + h^{\pi'}(s) - h^{\pi}(s). \tag{17}$$

For the iterates, abbreviate $g_k := g_{\pi_k,\pi_{k+1}}$ and we also define the comparison to the optimal policy

$$b_k(s) := \langle A^{\pi_k}(s,\cdot), \pi_k(\cdot \mid s) - \pi^*(\cdot \mid s)\rangle + h^{\pi_k}(s) - h^{\pi^*}(s). \tag{18}$$

Lemma A.1 (Performance difference)**.** *For any feasible policies π and π',*

$$V^{\pi'} - V^{\pi} = (I - \gamma P^{\pi'})^{-1} g_{\pi,\pi'}. \tag{19}$$

Equivalently, we have

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi'}}[g_{\pi,\pi'}(s')], \tag{20}$$

where

$$d_s^{\pi'} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t e_s^{\top}(P^{\pi'})^t. \tag{21}$$

Proof. Without loss of generality, we assume that the policies are stationary. Fix a state s, averaging the definition of Q^{π} under $\pi'(\cdot \mid s)$ and subtracting $V^{\pi}(s)$ gives

$$\langle A^{\pi}(s,\cdot), \pi'(\cdot \mid s)\rangle + h^{\pi'}(s) - h^{\pi}(s) = \langle c(s,\cdot), \pi'(\cdot \mid s)\rangle + h^{\pi'}(s) + \gamma(P^{\pi'}V^{\pi})(s) - V^{\pi}(s).$$

Because $\langle A^{\pi}(s,\cdot), \pi(\cdot \mid s)\rangle = 0$, the left-hand side is exactly $g_{\pi,\pi'}(s)$. The Bellman equation for π' therefore implies

$$V^{\pi}(s) = \sum_a \pi(a \mid s)\left[c(s,a) + h^{\pi}(s) + \gamma\sum_{s'} P(s' \mid s,a)V^{\pi}(s')\right],$$

which is equivalent to

$$V^{\pi'}(s) = \langle c(s,\cdot), \pi'(\cdot \mid s)\rangle + h^{\pi'}(s) + \gamma(P^{\pi'}V^{\pi'})(s).$$

Combining above results gives

$$V^{\pi'}(s) - V^{\pi}(s) = g_{\pi,\pi'}(s) + \gamma\big(P^{\pi'}(V^{\pi'} - V^{\pi})\big)(s).$$

Then we stack this identity over all states yields

$$(I - \gamma P^{\pi'})(V^{\pi'} - V^{\pi}) = g_{\pi,\pi'},$$

where since $P^{\pi'}$ is stochastic and $\gamma < 1$, the Neumann series converges:

$$(I - \gamma P^{\pi'})^{-1} = \sum_{t=0}^{\infty}\gamma^t(P^{\pi'})^t.$$

This proves equation 19. Left-multiplying by $e_s^{\top}$, inserting the factor $(1-\gamma)/(1-\gamma)$, and using equation 21 gives

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_{t=0}^{\infty}\gamma^t e_s^{\top}(P^{\pi'})^t g_{\pi,\pi'} = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi'}}[g_{\pi,\pi'}(s')],$$

giving us the desired result. □

Lemma A.2 (Optimal-policy identity)**.** *The vector b_k satisfies*

$$b_k = (I - \gamma P^{\pi^*})e_k, \tag{22}$$

and consequently

$$\mathbb{E}_{\nu^*}[b_k(s)] = (1-\gamma)F_k. \tag{23}$$

Proof. For each state s, expand the inner product in equation 18:

$$\begin{aligned} b_k(s) &= V^{\pi_k}(s) - \sum_a \pi^*(a \mid s) Q^{\pi_k}(s,a) + h^{\pi_k}(s) - h^{\pi^*}(s) \\ &= V^{\pi_k}(s) - \langle c(s,\cdot), \pi^*(\cdot \mid s)\rangle - h^{\pi^*}(s) - \gamma(P^{\pi^*} V^{\pi_k})(s). \end{aligned}$$

The Bellman equation $V^{\pi^*}(s) = \langle c(s,\cdot), \pi^*(\cdot \mid s)\rangle + h^{\pi^*}(s) + \gamma(P^{\pi^*} V^{\pi^*})(s)$ then gives

$$\begin{aligned} b_k(s) &= V^{\pi_k}(s) - V^{\pi^*}(s) - \gamma\big(P^{\pi^*}(V^{\pi_k} - V^{\pi^*})\big)(s) \\ &= \big((I - \gamma P^{\pi^*}) e_k\big)(s), \end{aligned}$$

which proves equation 22. Finally,

$$\begin{aligned} \mathbb{E}_{\nu^*}[b_k] &= (\nu^*)^\top (I - \gamma P^{\pi^*}) e_k \\ &= (1-\gamma)(\nu^*)^\top e_k = (1-\gamma) F_k, \end{aligned}$$

where stationarity gives $(\nu^*)^\top P^{\pi^*} = (\nu^*)^\top$. □

Lemma A.3. *For every $s \in \mathcal{S}$ and every feasible $p \in \Delta(\mathcal{A})$, update equation 13 satisfies*

$$\eta_k \lambda_k(s) \Big[\langle A^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot \mid s) - p\rangle + h^{\pi_{k+1}}(s) - h_s(p) \Big] + E_k(s) \le D^p_{\pi_k(\cdot|s)}(s) - \big(1 + \mu\eta_k\lambda_k(s)\big) D^p_{\pi_{k+1}(\cdot|s)}(s). \tag{24}$$

Proof. Fix s and take $x = \pi_k(\cdot \mid s)$, $y = \pi_{k+1}(\cdot \mid s)$ and $A = A^{\pi_k}(s,\cdot)$. The objective minimized at this state is

$$p \longmapsto \eta_k \lambda_k(s) \big[\langle A, p\rangle + h_s(p)\big] + D^p_x(s).$$

Since the feasible set $\Delta(\mathcal{A})$ is convex, the first-order optimality condition at y, with $(\partial h)^{\pi_{k+1}}(s,\cdot) \in \partial h_s(y)$, gives

$$\big\langle \eta_k \lambda_k(s) \big(A + (\partial h)^{\pi_{k+1}}(s,\cdot)\big) + \nabla\psi(y) - \nabla\psi(x), p - y \big\rangle \ge 0.$$

Rearranging equation C.2 yields

$$\eta_k \lambda_k(s) \langle A, y - p\rangle \le \eta_k \lambda_k(s) \langle (\partial h)^{\pi_{k+1}}(s,\cdot), p - y\rangle + \langle \nabla\psi(y) - \nabla\psi(x), p - y\rangle .$$

Since the regularizer h^p is μ-strongly convex w.r.t. policy, equation 8 gives

$$\langle (\partial h)^{\pi_{k+1}}(s,\cdot), p - y\rangle \le h_s(p) - h_s(y) - \mu D^p_y(s).$$

Substituting this bound into the preceding inequality and moving the regularizer difference to the left gives

$$\begin{aligned} &\eta_k \lambda_k(s) \big[\langle A, y - p\rangle + h_s(y) - h_s(p)\big] \\ &\quad \le \langle \nabla\psi(y) - \nabla\psi(x), p - y\rangle - \mu\eta_k\lambda_k(s) D^p_y(s). \end{aligned}$$

Finally, direct expansion of the three Bregman divergences gives the three-point identity

$$\langle \nabla\psi(y) - \nabla\psi(x), p - y\rangle = D^p_x(s) - D^y_x(s) - D^p_y(s)$$

and $D^y_x(s) = E_k(s)$. Combining these identities proves equation 24. □

Lemma A.4 (Monotonic improvement)**.** *For every state,*

$$g_k(s) \le -\frac{E_k(s)}{\eta_k\lambda_k(s)} - \left(\frac{1}{\eta_k\lambda_k(s)} + \mu\right) D^{\pi_k}_{\pi_{k+1}}(s) \le 0, \tag{25}$$

and

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \le g_k(s) \le 0, \tag{26}$$

leading to the conclusion that $V^{\pi_k}(s)$ and F_k are nonincreasing in k.

Proof. Take $p = \pi_k(\cdot \mid s)$ in Equation 24 and divide by $\eta_k \lambda_k(s) > 0$. The bracket on the left becomes

$$\langle A^{\pi_k}(s,\cdot), \pi_{k+1}(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) = g_k(s),$$

whereas the two divergences on the right are $D^{\pi_k}_{\pi_k}(s) = 0$ and $D^{\pi_k}_{\pi_{k+1}}(s)$. This gives the first inequality in equation 25. Both remaining terms are nonpositive because Bregman divergences are nonnegative, so $g_k(s) \le 0$.

Apply Lemma A.1 with $(\pi, \pi') = (\pi_k, \pi_{k+1})$:

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi_{k+1}}(s') g_k(s').$$

The $t = 0$ term in equation 21 shows that $d_s^{\pi_{k+1}}(s) \ge 1 - \gamma$. Since every $g_k(s') \le 0$, discarding all terms with $s' \ne s$ can only increase the sum. Hence

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) \le \frac{d_s^{\pi_{k+1}}(s)}{1-\gamma} g_k(s) \le g_k(s),$$

where the second inequality uses both $d_s^{\pi_{k+1}}(s)/(1-\gamma) \ge 1$ and $g_k(s) \le 0$. This proves equation 26. Subtracting the fixed value V^{π^*} and averaging under ν^* also shows that F_k is nonincreasing. □

Lemma A.5 (Fundamental recursion)**.** *The iterates of equation 13 satisfy the following recursion for every k:*

$$F_{k+1} + \mathbb{E}_{\nu^*}\left[\left(\frac{1}{\eta_k \lambda_k(s)} + \mu\right) D_{k+1}(s)\right] + \mathbb{E}_{\nu^*}\left[\frac{E_k(s)}{\eta_k \lambda_k(s)}\right] \le \gamma F_k + \mathbb{E}_{\nu^*}\left[\frac{D_k(s)}{\eta_k \lambda_k(s)}\right]. \tag{27}$$

Proof. Set $p = \pi^*(\cdot \mid s)$ in Equation 24 and divide by $\eta_k \lambda_k(s)$. To identify the bracket, add and subtract $\pi_k(\cdot \mid s)$ and $h^{\pi_k}(s)$:

$$\begin{aligned}\langle A^{\pi_k}, \pi_{k+1} - \pi^*\rangle + h^{\pi_{k+1}} - h^{\pi^*} &= \left(\langle A^{\pi_k}, \pi_{k+1} - \pi_k\rangle + h^{\pi_{k+1}} - h^{\pi_k}\right) + \left(\langle A^{\pi_k}, \pi_k - \pi^*\rangle + h^{\pi_k} - h^{\pi^*}\right) \\ &= g_k(s) + b_k(s),\end{aligned}$$

where all policy arguments are evaluated at any fixed state s. Therefore

$$b_k(s) + g_k(s) + \frac{E_k(s)}{\eta_k \lambda_k(s)} \le \frac{D_k(s)}{\eta_k \lambda_k(s)} - \left(\frac{1}{\eta_k \lambda_k(s)} + \mu\right) D_{k+1}(s). \tag{28}$$

By Lemma A.4, $V^{\pi_{k+1}} - V^{\pi_k} \le g_k$ statewise. Thus replacing $g_k(s)$ on the left of equation 28 by $V^{\pi_{k+1}}(s) - V^{\pi_k}(s)$ preserves the inequality. Averaging with respect to ν^* gives

$$\mathbb{E}_{\nu^*}[b_k] + F_{k+1} - F_k + \mathbb{E}_{\nu^*}\left[\frac{E_k(s)}{\eta_k \lambda_k(s)}\right] \le \mathbb{E}_{\nu^*}\left[\frac{D_k(s)}{\eta_k \lambda_k(s)}\right] - \mathbb{E}_{\nu^*}\left[\left(\frac{1}{\eta_k \lambda_k(s)} + \mu\right) D_{k+1}(s)\right].$$

Now use $\mathbb{E}_{\nu^*}[b_k] = (1-\gamma)F_k$ in equation 23, so the first two scalar terms on the left combine as $(1-\gamma)F_k + F_{k+1} - F_k = F_{k+1} - \gamma F_k$. Reranging $-\gamma F_k$ and the D_{k+1} term yields equation 27. □

Lemma A.6 (Uniform lower bound on clipping)**.** *Assume the cost function $c(s,a)$ and regularizer $h_s(p)$ are bounded. Then there exists a constant ℓ such that following definitions are well-defined and finite:*

$$\begin{aligned} K_{\text{norm}} &:= \max_{\|x\|_\infty \le 1} \|x\|, \\ M_0 &:= \max_s V^{\pi_0}(s), \\ C_{\text{osc}} &:= \max_s \left(\max_a c(s,a) - \min_a c(s,a)\right), \\ B_A &:= K_{\text{norm}}\left[C_{\text{osc}} + \gamma\left(M_0 - \frac{\ell}{1-\gamma}\right)\right]. \end{aligned}$$

Then $\|A^{\pi_k}(s,\cdot)\| \le B_A$ for every k, s, and hence

$$\lambda_k(s) \ge \kappa := \min\{1, \alpha/B_A\} > 0, \tag{29}$$

with the convention $\kappa = 1$ if $B_A = 0$.

Proof. For any policy π, there exists a finite ℓ such that

$$c(s,a) + h_s(p) \geq \ell \quad \text{for all } s, a \text{ and all } p \in \operatorname{dom} h_s. \tag{30}$$

Hence

$$V^{\pi}(s) \geq \sum_{t=0}^{\infty} \gamma^t \ell = \frac{\ell}{1-\gamma}.$$

Optimality of π^* and Lemma A.4 therefore give, for every k and s,

$$\frac{\ell}{1-\gamma} \leq V^{\pi^*}(s) \leq V^{\pi_k}(s) \leq M_0.$$

For any two actions a, b, the definition of Q yields

$$|Q^{\pi_k}(s,a) - Q^{\pi_k}(s,b)| \leq |c(s,a) - c(s,b)| + \gamma \left| \sum_{s'} \big(P(s' \mid s,a) - P(s' \mid s,b)\big) V^{\pi_k}(s') \right|.$$

The two transition expectations both lie in the interval $[\ell/(1-\gamma), M_0]$. Their difference is therefore at most the length of this interval, and consequently

$$\max_a Q^{\pi_k}(s,a) - \min_a Q^{\pi_k}(s,a) \leq C_{\mathrm{osc}} + \gamma \left(M_0 - \frac{\ell}{1-\gamma} \right).$$

Each advantage coordinate is obtained by subtracting from $Q^{\pi_k}(s,a)$ a convex combination of the coordinates of $Q^{\pi_k}(s,\cdot)$, because

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - \sum_b \pi(b \mid s) Q^{\pi}(s,b).$$

Hence we have

$$\|A^{\pi_k}(s,\cdot)\| \leq K_{\mathrm{norm}} \|A^{\pi_k}(s,\cdot)\|_{\infty} \leq B_A.$$

If $B_A > 0$, the definition of equation 12 gives

$$\lambda_k(s) \geq \min\left\{1, \frac{\alpha}{B_A}\right\} = \kappa.$$

If $B_A = 0$, every advantage vector is zero, so we have $\lambda_k(s) = 1 = \kappa$. □

The following lemma is needed only for a fixed stepsize in the unregularized case.

Lemma A.7 (Finite relative variation of clipping weights)**.** *Suppose $h_s \equiv 0$ and $\eta_k \equiv \eta > 0$. Let*

$$r_k(s) := \frac{1}{\lambda_k(s)} = \max\left\{1, \frac{\|A^{\pi_k}(s,\cdot)\|}{\alpha}\right\}, \qquad \theta_k := \max\left\{1, \max_s \frac{r_{k+1}(s)}{r_k(s)}\right\}. \tag{31}$$

Then

$$\sum_{k=0}^{\infty} (\theta_k - 1) \leq T_0 := \frac{K_{\mathrm{norm}}}{\alpha} \sum_s e_0(s), \qquad R_N := \prod_{k=0}^{N-1} \theta_k \leq e^{T_0}. \tag{32}$$

Proof. Set $u_k(s) := V^{\pi_k}(s) - V^{\pi_{k+1}}(s) \geq 0$ and $U_k := \|u_k\|_{\infty}$. When $h \equiv 0$, equation 10 and equation 11 give

$$A^{\pi}(s,a) = c(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi}(s') - V^{\pi}(s).$$

Subtracting this identity at π_k from the one at π_{k+1} proves

$$A^{\pi_{k+1}}(s,a) - A^{\pi_k}(s,a) = u_k(s) - \gamma \sum_{s'} P(s' \mid s,a) u_k(s').$$

Because $0 \leq u_k(s') \leq U_k$, the second term on the right belongs to $[-\gamma U_k, 0]$ and the first to $[0, U_k]$. Thus the entire right-hand side lies in $[-\gamma U_k, U_k]$, and in particular

$$\max_s \|A^{\pi_{k+1}}(s,\cdot) - A^{\pi_k}(s,\cdot)\| \leq K_{\mathrm{norm}} U_k.$$

The triangle inequality gives

$$|\|A^{\pi_{k+1}}(s,\cdot)\| - \|A^{\pi_k}(s,\cdot)\|| \le \|A^{\pi_{k+1}}(s,\cdot) - A^{\pi_k}(s,\cdot)\|.$$

Since $x \mapsto \max\{1, x/\alpha\}$ is $1/\alpha$-Lipschitz,

$$|r_{k+1}(s) - r_k(s)| \le \frac{K_{\text{norm}}}{\alpha} U_k.$$

Moreover, when $r_{k+1}(s) > r_k(s)$,

$$\frac{r_{k+1}(s)}{r_k(s)} - 1 = \frac{r_{k+1}(s) - r_k(s)}{r_k(s)} \le |r_{k+1}(s) - r_k(s)|.$$

Taking the maximum over states yields

$$\theta_k - 1 \le \max_s |r_{k+1}(s) - r_k(s)| \le \frac{K_{\text{norm}}}{\alpha} U_k.$$

For every N,

$$\sum_{k=0}^{N-1} U_k \le \sum_s \sum_{k=0}^{N-1} u_k(s) = \sum_s \big(V^{\pi_0}(s) - V^{\pi_N}(s)\big) \le \sum_s e_0(s).$$

Here the first inequality uses $U_k \le \sum_s u_k(s)$, the equality telescopes, and the last inequality uses $V^{\pi_N}(s) \ge V^{\pi^*}(s)$. Combining the last two displays and letting $N \to \infty$ proves the first claim. Finally, $\log x \le x - 1$ for $x > 0$, so

$$\log R_N = \sum_{k=0}^{N-1} \log \theta_k \le \sum_{k=0}^{N-1} (\theta_k - 1) \le T_0.$$

□

C.3 Convergence theorem and proof

Theorem A.1 (Convergence of clipped PMD, cf. Theorem 1 and 2)**.** *Assume the setting of Section C.1, use the KL divergence and uniform initialization, and let κ be the clipping lower bound in Lemma A.6.*

(i) ***Strongly convex regularizer, fixed stepsize.*** *Suppose $\mu > 0$ and $\eta_k \equiv \eta > 0$. Define*

$$\rho := \max\left\{\gamma, \frac{1}{\kappa(1+\eta\mu)}\right\}.$$

If $\eta\mu > 1/\kappa - 1$, then $\rho < 1$ and, for every $N \ge 0$, we have

$$F_N + (\mu + \frac{1}{\eta})\mathbb{E}_{\nu^*}[D_N(s)] \le \rho^N \left(F_0 + (\mu + \frac{1}{\eta})\mathbb{E}_{\nu^*}[D_0(s)]\right) \le \rho^N \left(F_0 + (\mu + \frac{1}{\eta}) \log|\mathcal{A}|\right).$$

In particular, if $\eta\mu \ge \frac{1}{\gamma\kappa} - 1$, then $\rho = \gamma$.

(ii) ***Strongly convex regularizer, adaptive stepsize.*** *Suppose $\mu > 0$ and choose*

$$\eta_{k+1} := \eta_k \max\left\{1, \max_{s \in \mathcal{S}} \frac{\lambda_k(s)}{\lambda_{k+1}(s)}\right\}. \tag{33}$$

Let $\kappa_0 := \min_s \lambda_0(s)$ and $\rho_0 := \max\left\{\gamma, \frac{1}{1+\mu\eta_0\kappa_0}\right\} < 1$. Then

$$F_N \le \rho_0^N \left[F_0 + \left(\mu + \frac{1}{\eta_0\kappa_0}\right) \log|\mathcal{A}|\right].$$

(iii) ***No regularizer, adaptive stepsize.*** *Suppose $h_s \equiv 0$, and choose stepsizes as in equation 33. Then, for every $N \ge 1$,*

$$F_N \le \frac{\gamma F_0 + \mathbb{E}_{\nu^*}[D_0(s)/(\eta_0\lambda_0(s))]}{(1-\gamma)N} \le \frac{\gamma F_0 + \dfrac{\log|\mathcal{A}|}{\eta_0\kappa_0}}{(1-\gamma)N}.$$

(iv) ***No regularizer, fixed stepsize.*** *Suppose $h_s \equiv 0$ and $\eta_k \equiv \eta > 0$. With R_N and T_0 from Lemma A.7,*

$$F_N \le \frac{R_N\left[\gamma F_0 + \mathbb{E}_{\nu^*}[D_0(s)/(\eta\lambda_0(s))]\right]}{(1-\gamma)N} \le \frac{\exp\left(\frac{K_{\mathrm{norm}}}{\alpha}\sum_s e_0(s)\right)\left[\gamma F_0 + \frac{\log|\mathcal{A}|}{\eta\kappa_0}\right]}{(1-\gamma)N}.$$

Proof. **Part (i).** For a fixed stepsize and $\lambda_k(s) \in [\kappa, 1]$,

$$\frac{1}{\eta} \le \frac{1}{\eta\lambda_k(s)} \le \frac{1}{\eta\kappa}.$$

Drop the nonnegative E_k term in equation 27 and apply these bounds. On the left we use the lower bound $1/\eta$, while on the right we use the upper bound $1/(\eta\kappa)$, obtaining

$$F_{k+1} + \left(\mu + \frac{1}{\eta}\right)\mathbb{E}_{\nu^*}[D_{k+1}] \le \gamma F_k + \frac{1}{\eta\kappa}\mathbb{E}_{\nu^*}[D_k].$$

By the choice of ρ, we have

$$\gamma \le \rho, \qquad \frac{1}{\eta\kappa} \le \rho\left(\mu + \frac{1}{\eta}\right).$$

Hence the potential $\Psi_k := F_k + (\mu + \frac{1}{\eta})\mathbb{E}_{\nu^*}[D_k]$ has following property:

$$\Psi_{k+1} \le \gamma F_k + \frac{1}{\eta\kappa}\mathbb{E}_{\nu^*}[D_k] \le \rho F_k + \rho(\mu + \frac{1}{\eta})\mathbb{E}_{\nu^*}[D_k] = \rho\Psi_k.$$

Induction gives $\Psi_N \le \rho^N \Psi_0$. Since the uniform initialization satisfies $D_0(s) \le \log|\mathcal{A}|$ for every state, this is precisely desired inequality. Furthermore,

$$\frac{1}{\kappa(1+\eta\mu)} < 1 \quad \Longleftrightarrow \quad \eta\mu > \frac{1}{\kappa} - 1,$$

so the stated condition makes both entries in the maximum defining ρ strictly smaller than one.

Part (ii). The stepsizes scheduled in equation 33 implies

$$\eta_{k+1}\lambda_{k+1}(s) \ge \eta_k\lambda_k(s) \quad \text{for all } s, \qquad \frac{1}{\eta_{k+1}\lambda_{k+1}(s)} \le \frac{1}{\eta_k\lambda_k(s)}. \tag{34}$$

Indeed, the maximum in equation 33 is at least $\lambda_k(s)/\lambda_{k+1}(s)$ for each fixed s; multiplication by $\eta_k\lambda_{k+1}(s)$ proves the first inequality, and taking reciprocals proves the second. Define the time-varying potential

$$\Phi_k := F_k + \mathbb{E}_{\nu^*}\left[\left(\frac{1}{\eta_k\lambda_k(s)} + \mu\right)D_k(s)\right].$$

Using equation 34 and then equation 27,

$$\Phi_{k+1} \le F_{k+1} + \mathbb{E}_{\nu^*}\left[\left(\frac{1}{\eta_k\lambda_k(s)} + \mu\right)D_{k+1}(s)\right] \le \gamma F_k + \mathbb{E}_{\nu^*}\left[\frac{D_k(s)}{\eta_k\lambda_k(s)}\right].$$

Moreover, $\eta_k\lambda_k(s) \ge \eta_0\lambda_0(s) \ge \eta_0\kappa_0$, and therefore

$$\frac{1/(\eta_k\lambda_k(s))}{1/(\eta_k\lambda_k(s)) + \mu} = \frac{1}{1 + \mu\eta_k\lambda_k(s)} \le \frac{1}{1 + \mu\eta_0\kappa_0} \le \rho_0.$$

Equivalently, for every state,

$$\frac{1}{\eta_k\lambda_k(s)} \le \rho_0\left(\frac{1}{\eta_k\lambda_k(s)} + \mu\right).$$

Together with $\gamma \le \rho_0$, this bounds the last line in the display for Φ_{k+1} by $\rho_0 \Phi_k$. Hence $\Phi_N \le \rho_0^N \Phi_0$. Dropping the nonnegative divergence term from Φ_N and using

$$\Phi_0 \le F_0 + \left(\mu + \frac{1}{\eta_0 \kappa_0}\right) \mathbb{E}_{\nu^*}[D_0(s)] \le F_0 + \left(\mu + \frac{1}{\eta_0 \kappa_0}\right) \log |\mathcal{A}|,$$

we have the desired result.

Part (iii). With $\mu = 0$, define

$$W_k := \mathbb{E}_{\nu^*}\left[\frac{D_k(s)}{\eta_k \lambda_k(s)}\right], \qquad Y_k := \gamma F_k + W_k.$$

Since equation 34 makes the reciprocal coefficients nonincreasing, equation 27 gives

$$F_{k+1} + W_{k+1} \le \gamma F_k + W_k = Y_k.$$

Indeed, the D_{k+1} coefficient defining W_{k+1} is no larger than the corresponding coefficient on the left of equation 27; the omitted E_k term is nonnegative. Equivalently,

$$Y_{k+1} + (1-\gamma) F_{k+1} \le Y_k. \tag{35}$$

Summing equation 35 from $k = 0$ to $N - 1$ telescopes the Y_k terms:

$$Y_N + (1-\gamma) \sum_{j=1}^{N} F_j \le Y_0.$$

Since $Y_N \ge 0$ and $F_j \ge F_N$ for $1 \le j \le N$ by Lemma A.4,

$$(1-\gamma) N F_N \le (1-\gamma) \sum_{j=1}^{N} F_j \le Y_0.$$

This proves the first inequality in equation A.1. The second follows from

$$\mathbb{E}_{\nu^*}\left[\frac{D_0(s)}{\eta_0 \lambda_0(s)}\right] \le \frac{1}{\eta_0 \kappa_0} \mathbb{E}_{\nu^*}[D_0(s)] \le \frac{\log |\mathcal{A}|}{\eta_0 \kappa_0}.$$

Part (iv). For a fixed stepsize, Lemma A.7 gives $1/(\eta \lambda_{k+1}(s)) \le \theta_k / (\eta \lambda_k(s))$, because $r_{k+1}(s) \le \theta_k r_k(s)$. From equation 27 with $\mu = 0$,

$$\mathbb{E}_{\nu^*}\left[\frac{D_{k+1}(s)}{\eta \lambda_k(s)}\right] \le Y_k - F_{k+1}.$$

Consequently,

$$Y_{k+1} = \gamma F_{k+1} + W_{k+1} \le \gamma F_{k+1} + \theta_k \mathbb{E}_{\nu^*}\left[\frac{D_{k+1}(s)}{\eta \lambda_k(s)}\right] \le \theta_k Y_k - (\theta_k - \gamma) F_{k+1}.$$

The first inequality uses the preceding comparison of reciprocal clipping factors; the second substitutes the bound from equation 27. Set $R_0 = 1$ and $R_{k+1} = \theta_k R_k$. Dividing above inequality by R_{k+1} gives

$$\frac{Y_{k+1}}{R_{k+1}} \le \frac{Y_k}{R_k} - \frac{(\theta_k - \gamma) F_{k+1}}{R_{k+1}}.$$

Summing from $k = 0$ to $N - 1$ and using $\theta_k - \gamma \ge 1 - \gamma$ gives

$$(1-\gamma) \sum_{j=1}^{N} \frac{F_j}{R_j} \le Y_0 - \frac{Y_N}{R_N} \le Y_0,$$

where the last inequality follows from $Y_N \ge 0$. Since $F_j \ge F_N$ by monotonicity and $R_j \le R_N$ because each $\theta_k \ge 1$,

$$\sum_{j=1}^{N} \frac{F_j}{R_j} \ge \frac{N F_N}{R_N}.$$

Combining the last two displays yields

$$F_N \le \frac{R_N Y_0}{(1-\gamma)N}.$$

Finally, using $R_N \le e^{T_0}$ in Lemma A.7, equation 16 and the definition of T_0, we obtain

$$\mathbb{E}_{\nu^*}\left[\frac{D_0(s)}{\eta\lambda_0(s)}\right] \le \frac{\log|\mathcal{A}|}{\eta\kappa_0},$$

□

Remark 1 (From ν^*-weighted error to all-state error). *If $\nu^*_{\min} := \min_s \nu^*(s) > 0$, then every bound on F_N immediately implies*

$$\left\|V^{\pi_N} - V^{\pi^*}\right\|_\infty \le \frac{F_N}{\nu^*_{\min}}.$$

Without full support, one may instead choose any full-support distribution ρ and set

$$\widetilde{\nu}^\top := (1-\gamma)\rho^\top (I - \gamma P^{\pi^*})^{-1}.$$

Then $\widetilde{\nu}(s) > 0$ and we have

$$\mathbb{E}_{\widetilde{\nu}}[b_k] = (1-\gamma)\mathbb{E}_\rho[e_k] \ge \delta \mathbb{E}_{\widetilde{\nu}}[e_k].$$

Now we define $\delta := (1-\gamma)\min_s \frac{\rho(s)}{\widetilde{\nu}(s)} > 0$, then repeat the proof with $\widetilde{\nu}$. This replaces the contraction coefficient γ in the fundamental recursion Lemma A.5 by $1-\delta < 1$ and yields similar convergence results on all states.

D Additional Experimental Details

D.1 Training Details

All experiments are conducted on a single node with parameter and optimizer offloading enabled. We use eight NVIDIA H100 GPUs for Qwen2.5-Math-7B and Qwen3-8B, four for Qwen3-4B, and two for Qwen3-1.7B.

We post-train models with our algorithm using VERL framework (Sheng et al., 2025). And all the hyperparameters are presented in Table 2.

D.2 Detailed Benchmark Results

In this section, we report results on four benchmarks: **MATH500** (Hendrycks et al., 2021), **Minerva Math** (Lewkowycz et al., 2022), **OlympiadBench** (He et al., 2024), and **AIME-like** (Xiong et al., 2025). AIME-like comprises 230 problems from recent competitions: AIME24, AIME25, HMMT24, HMMT25, BRUMO25, AMC23, and CMIMC25. We estimate pass@1 accuracy by averaging correctness over 32 sampled responses per problem for Qwen2.5-7B-Math (Avg@32) and 16 for Qwen3 (Avg@16). All responses are generated with a temperature of 1.0, top-$p = 1$, and a maximum generation length of 8,196 tokens.

As shown in Figures 4, 5,6 and 7, ACPO consistently outperforms the PPO and GRPO baselines, particularly on challenging benchmarks such as **OlympiadBench** (He et al., 2024) and **AIME-like** (Xiong et al., 2025). Beyond improving performance, ACPO substantially accelerates training, with particularly pronounced gains for PPO-AC.

D.3 Additional Results for Qwen3-1.7B with PPO-Type Learning Rate 10^{-6}

Here we report the post-training result of Qwen3-1.7B with lr=1e-6 for PPO-type algorithms (Figure 8 and 9). On Qwen3-1.7B, with the learning rate set to 10^{-6} for both PPO and PPO-AC, PPO-AC achieves a peak weighted accuracy of 60.55% at step 110, outperforming PPO's best accuracy of

Table 2: RL training hyperparameters. Qwen3 includes the 1.7B, 4B, and 8B models. Ratio clipping thresholds apply to the baselines, whereas α applies to ACPO variants. All sampling parameters refer to training rollouts.

	PPO-type		GRPO-type	
Hyperparameter	Qwen2.5-Math-7B[a]	Qwen3	Qwen2.5-Math-7B[a]	Qwen3
Data and rollout				
Training dataset	DAPO	DAPO	DAPO	DAPO
Prompt batch size	512	512	512	256
Rollouts per prompt	1	1	8	4
Maximum prompt length	1024	1024	1024	1024
Maximum response length	8196	8196	8196	8196
Sampling temperature	1.0	1.0	1.0	1.0
Top-p	1.0	1.0	1.0	1.0
Top-k	–	–	–	–
Policy optimization				
Optimizer	AdamW	AdamW	AdamW	AdamW
Actor learning rate	10^{-6}	10^{-6}[b]	10^{-6}	10^{-6}
Learning-rate schedule	Constant	Constant	Constant	Constant
Actor mini-batch size	128	128	128	64
Gradient clipping norm	1.0	1.0	1.0	1.0
Entropy coefficient	0	0	0	0
KL penalty in reward	No	No	No	No
KL loss coefficient	0	0	0	0
Baseline ϵ_{low}	0.2	0.2	0.2	0.2
Baseline ϵ_{high}	0.2	0.2	0.28	0.28
ACPO threshold α	3	3	3	2
Critic learning rate	10^{-5}	10^{-5}	–	–
Total training steps	300	300	200	100
Total optimizer steps	1200	1200	800	400

[a] The maximum context length of Qwen2.5-Math-7B is 4096 tokens.
[b] The actor learning rate is 5×10^{-7} for Qwen3-1.7B in PPO-type experiments.

57.37% by 3.18 percentage points within 200 steps training. Notably, PPO attains its best performance at initialization, whereas PPO-AC improves beyond this baseline, demonstrating more effective policy optimization in this setting. Hence we tune the learning rate such that PPO do not keep decrease during the training. We search for 5 learning rates, and finally choose lr=5e-7 for Qwen3-1.7B. Overall, the results on Qwen3-1.7B with a learning rate of 10^{-6} demonstrate the substantial performance gains achieved by ACPO over PPO.

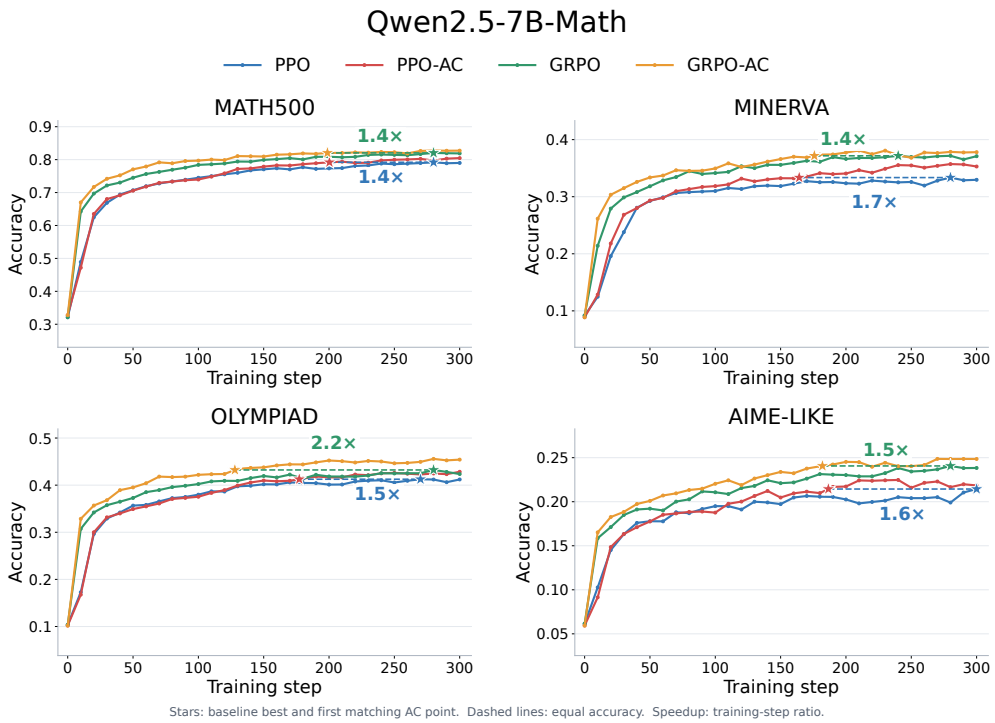


Figure 4: Qwen2.5-7B-Math on benchmarks

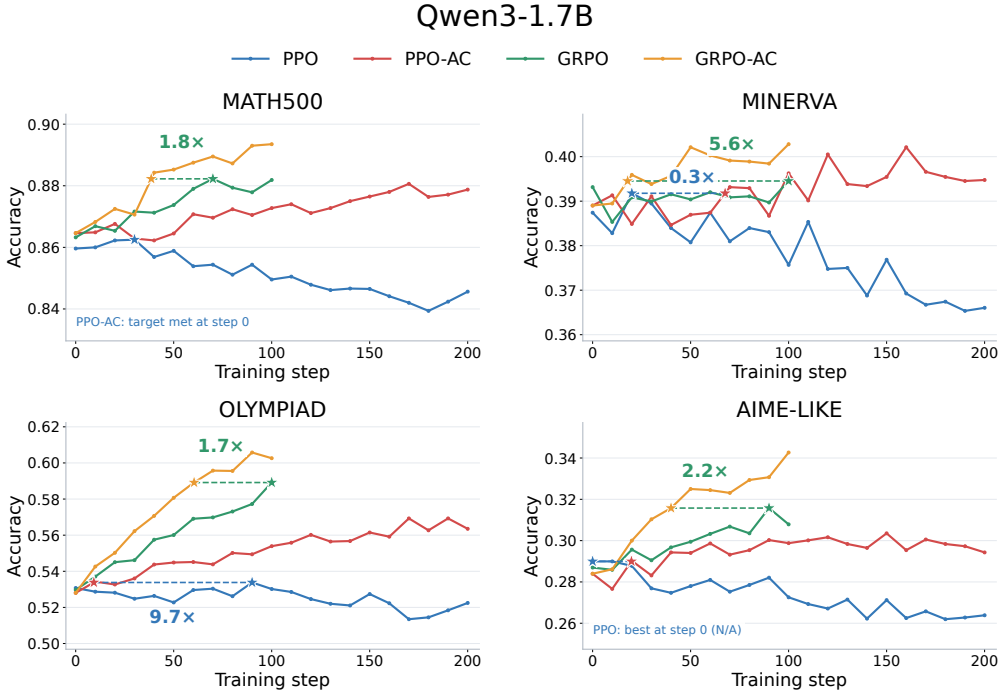


Figure 5: Qwen3-1.7B on benchmarks

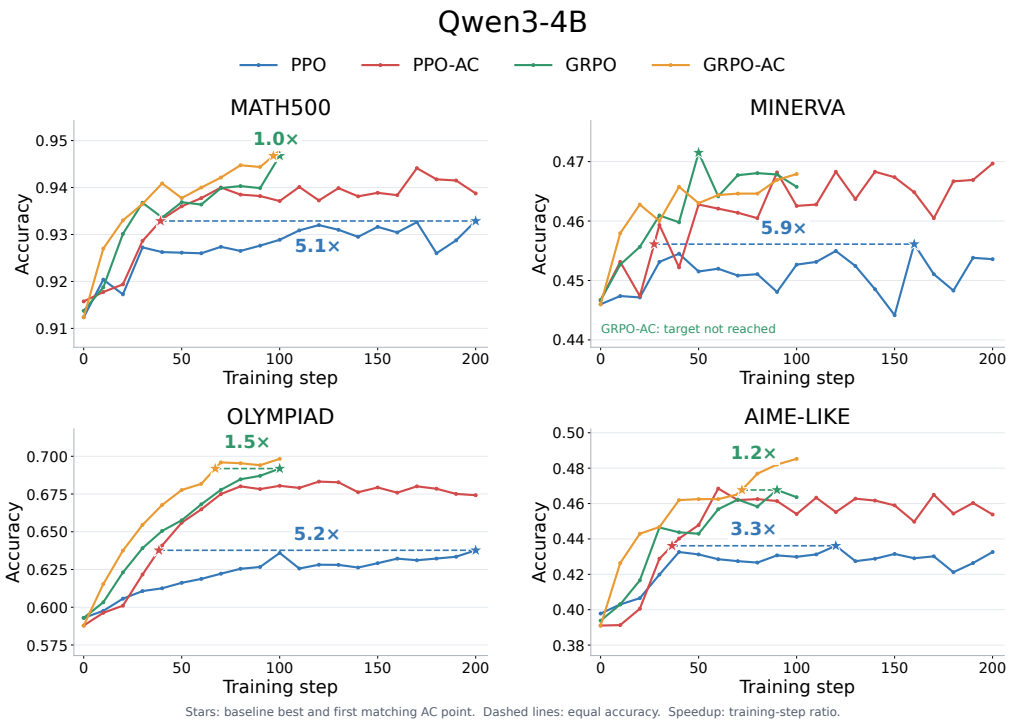


Figure 6: Qwen3-4B on benchmarks

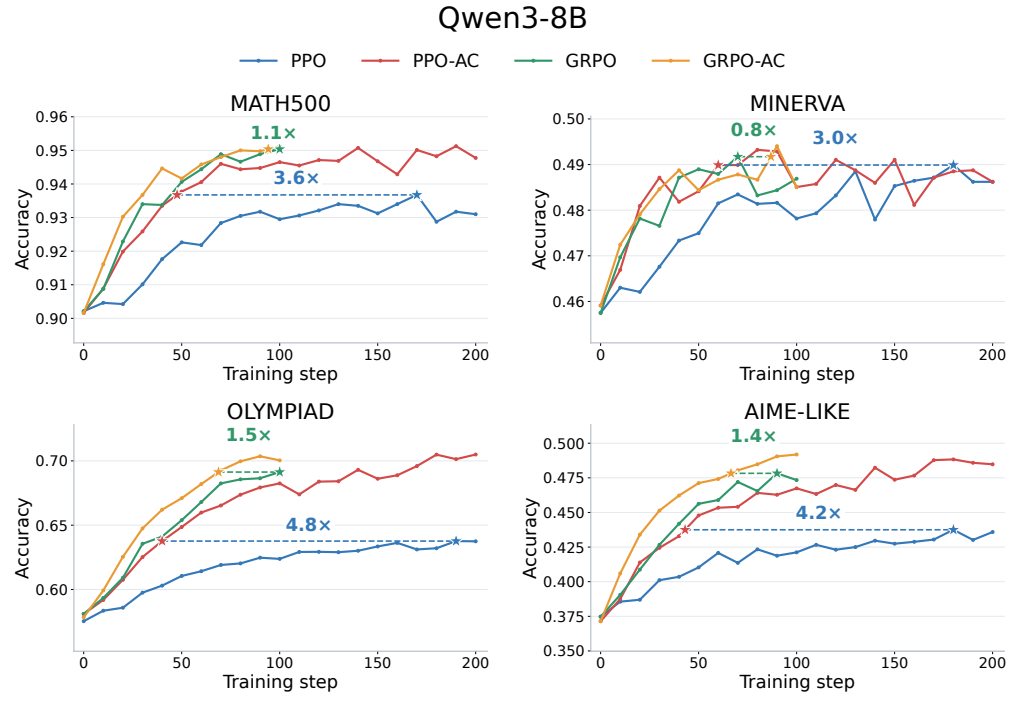


Figure 7: Qwen3-8B on benchmarks

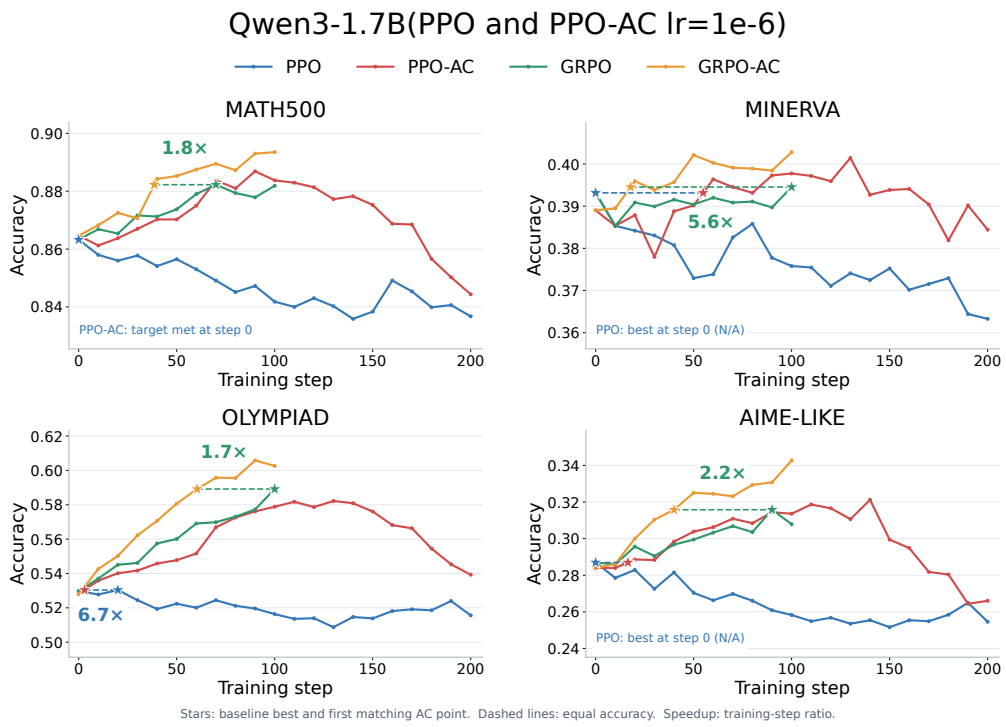


Figure 8: Qwen3-1.7B with PPO-Type lr=1e-6 on benchmarks

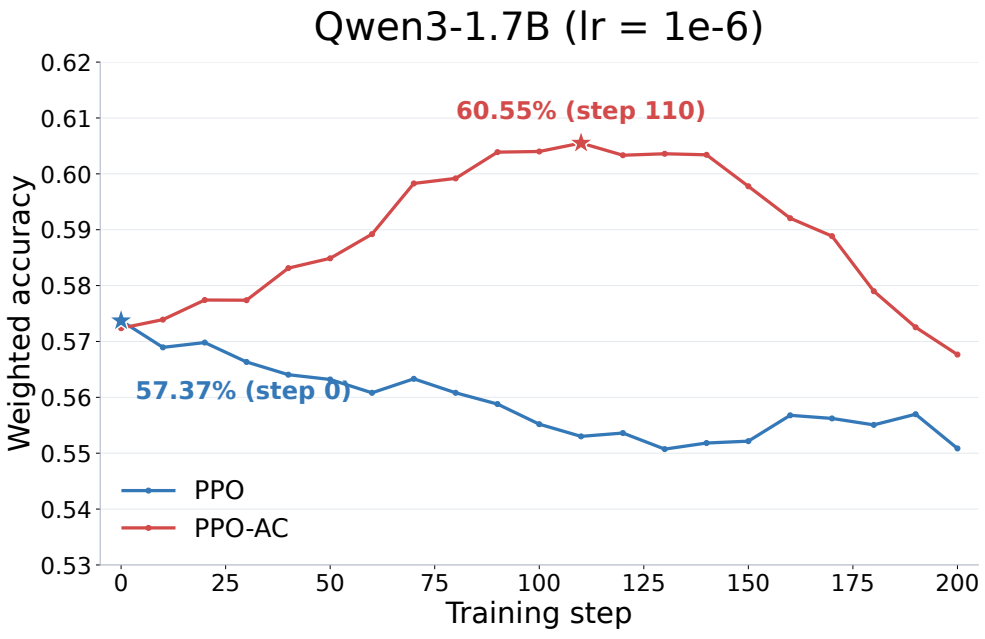


Figure 9: Weighted Accuracy of Qwen3-1.7B with PPO-Type lr=1e-6