ASYNCHRONOUS LLM POST-TRAINING: GROUP-MASS CAPPING AND CONVERGENCE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Asynchronous reinforcement learning (RL) improves the efficiency of large language model post-training but introduces stale rollouts generated by earlier policies. Theoretical understanding of how this staleness affects convergence and how to mitigate its impact remains limited. We derive a convergence bound for GRPO-style algorithms that explicitly characterizes the tradeoff between the gradient estimator's second moment and bias. For trajectory-level importance-weighted estimators, our analysis shows that once the second moment is uniformly controlled, delay enters the bound through the bias introduced by clipping or rescaling. Guided by this insight, we propose a novel group mass capping GRPO (GMC-GRPO) method, which minimizes a ratio-based bias bound within a class of weighted estimators sharing a common second-moment guarantee. We establish convergence guarantees for asynchronous GMC-GRPO and show that, compared with TIC-GRPO, it improves the threshold dependence of the fourth-order delay term from $O(\epsilon^{-4})$ to $O(\epsilon^{-2})$ as $\epsilon \to 0$, where $1+\epsilon$ is the ratio threshold. Under local policy overlap, the delay-dependent term decreases as $G^{-2/5}$ after tuning the step size, where G is the group size. For fixed behavior and current policies, the bias introduced by group rescaling also vanishes as $G \to \infty$, whereas the bias from trajectory-wise clipping can persist. Experiments across Qwen3 models and reasoning benchmarks demonstrate improved robustness to stale rollouts, with GMC-GRPO achieving the best performance among stable baselines under large rollout delays.

1 INTRODUCTION

Reinforcement learning (RL) has become a widely used paradigm for post-training large language models (LLMs), enabling them to improve from outcome-level feedback rather than relying solely on static demonstrations (Ouyang et al., 2022; Lambert et al., 2024). Given a prompt, RL allows an LLM to generate thinking tokens before producing a final answer, and reinforces the generated trajectory based on its final quality (Jaech et al., 2024; Guo et al., 2025). Models post-trained with algorithms such as GRPO (Shao et al., 2024) and its variants (Yu et al., 2026; Liu et al., 2025; Zheng et al., 2026) have shown strong capabilities on challenging reasoning problems, such as math (Cobbe et al., 2021; He et al., 2024), coding (Jain et al., 2025), and agentic tasks (Yao et al., 2024).

To reduce the inefficiency of repeatedly alternating between rollout generation and policy optimization, modern RL training pipelines increasingly adopt **asynchronous** execution (Zhong et al., 2024; Sheng et al., 2025). Such systems decouple rollout generation from policy optimization: rollout workers continuously generate responses using their local policy versions, while the learner updates the policy as soon as sufficient completed rollouts have accumulated and publishes the updated parameters without blocking generation (Fu et al., 2026; Hu et al., 2026; Noukhovitch et al., 2025). Together with optimized training engines and workload scheduling, this design allows inference and training to proceed concurrently, improving resource utilization and substantially increasing the training throughput of GRPO-style methods (Wang et al., 2025).

The decoupling, however, introduces *data staleness*. Since rollout workers and the learner progress independently, the learner may update the policy several times before a generated rollout is consumed. Consequently, a training batch can mix trajectories collected under multiple earlier policy versions, creating a mismatch between the behavior policies that generated the data and the cur-

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.

rent policy being optimized (Fu et al., 2026). As the delay grows, these trajectories can become increasingly off-policy, making policy optimization more difficult and reducing training efficiency (Tian et al., 2026; Zheng et al., 2026; Xu et al., 2026). Despite the empirical success of delay-aware algorithms (Roux et al., 2025; Xi et al., 2026; Huang et al., 2026), how this staleness affects the convergence of GRPO-style algorithms, and how to mitigate its impact through principled algorithm design, remain poorly understood. This raises a natural but fundamental research question:

Can we theoretically characterize how data staleness affects the convergence of GRPO-style algorithms, and use this to design RL algorithms better suited to asynchronous training systems?

We answer this question by first identifying what any importance-weighted algorithm must control in order to converge, and then designing the correction that best meets this requirement under delay. Our key contributions are as follows:

- **Second moment and bias tradeoff.** We derive a convergence bound for GRPO-style algorithms which explicitly captures the tradeoff between the second moment of gradient estimators and the bias error. With trajectory-level importance sampling, stale rollouts can produce large importance ratios and enlarge the gradient estimator's second moment. Clipping these ratios controls the second moment but can introduce bias. When the resulting second-moment bound is independent of delay, delay enters the convergence bound through this clipping bias. This motivates algorithms that minimize a bias bound while maintaining a common second-moment guarantee.
- **GMC-GRPO.** Guided by this principle, we introduce a class of weighted gradient estimators that satisfy a common second-moment bound. This class includes several GRPO-style methods as special cases. Within this class, we derive group mass capping (GMC)-GRPO by minimizing a bias bound and selecting a solution that preserves each group's gradient direction, while retaining the same second-moment guarantee. This makes GMC-GRPO particularly well suited to asynchronous training, where its principled weighting rule effectively limits the bias.
- **Convergence of asynchronous GMC-GRPO.** We establish a convergence guarantee for asynchronous GMC-GRPO with explicit dependence on stochastic rollout delays. Compared with TIC-GRPO (Pang et al., 2025), GMC-GRPO yields a strictly smaller delay-dependent penalty: in the tight-clipping regime, the penalty improves from $O(\epsilon^{-4})$ to $O(\epsilon^{-2})$, where $1+\epsilon$ is the threshold for individual importance ratios in TIC-GRPO and their group average in GMC-GRPO. Moreover, because GMC-GRPO rescales a group only when its average importance ratio exceeds the stability cap, its correction bias vanishes as G grows and the group average concentrates. In contrast, trajectory-wise clipping acts separately on every ratio, leaving a nonzero clipping bias even in the large-group limit.
- **Experiments.** Finally, we evaluate GMC-GRPO under both controlled fixed delays and in a simulated asynchronous pipeline where rollout staleness is correlated with response length. Across Qwen3 models and reasoning benchmarks, GMC-GRPO achieves the best performance among vanilla GRPO-style algorithms and other baselines designed for asynchronous training. These results demonstrate effectiveness of our theory-guided design under asynchronous rollout delays.

Related Work. We defer the comprehensive discussion of related work to Appendix A. Our paper is most closely related to off-policy correction for stale LLM rollouts in asynchronous RL systems (Noukhovitch et al., 2025; Fu et al., 2026; Hu et al., 2026). Typical strategies include adaptive clipping thresholds (Roux et al., 2025; Xi et al., 2026; Zhao et al., 2026; Li et al., 2025), variance-aware step-size control (Zheng et al., 2026; Huang et al., 2026; Xu et al., 2026), and reuse of stale experience through staged training with selective filtering or replay buffers (Tian et al., 2026; Arnal et al., 2026; Ritter et al., 2026). Despite the empirical success of these methods, few works have studied GRPO-style algorithms from a theoretical perspective, especially under asynchronous rollout delays. Zhou et al. (2026) characterize the GRPO gradient as a U-statistic and analyze its finite-sample error and dependence on group size, while Pang et al. (2025) introduce TIC-GRPO and provide the first rigorous convergence analysis of GRPO-style methods. Nevertheless, these studies are not under asynchronous training. More recently, Xu et al. (2026) analyze asynchronous GRPO through a generic biased-gradient framework, but leave the staleness-induced bias and higher-order remainder terms as abstract quantities in the final bound. **To the best of our knowledge, we are the first to provide an explicit delay-dependent convergence analysis of GRPO-style algorithms in asynchronous RL systems.**

2 PRELIMINARIES

2.1 RL POST-TRAINING

We denote π_θ as an autoregressive language model parameterized by $\theta \in \mathbb{R}^d$. Given a prompt x drawn from a training distribution $\mathcal{D}$ over the prompt space $\mathcal{X}$, the likelihood of a response $y = (y_1, \ldots, y_T)$ is $\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$, where $\mathcal{A}$ denotes the token vocabulary, $y_{<t} := (y_1, \ldots, y_{t-1})$ is the generated prefix, and T is the maximum response length. A verifier $r_\phi : \mathcal{X} \times \mathcal{A}^T \to [0, R]$ scores each completed response.

The goal of post-training is to maximize the expected verifier score. Following standard practice in reinforcement learning, we divide each global batch into mini-batches and update the parameters sequentially. The rollout samples used by an update may therefore be generated by a behavior policy $\pi_{\theta_{\text{behav}}}$ that differs from the current policy. Using trajectory-level importance sampling to account for this mismatch, we write the objective as

$$J(\theta) := \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[r_\phi(x,y)\big] = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\theta_{\text{behav}}}(\cdot|x)}\big[\rho\, r_\phi(x,y)\big], \tag{1}$$

where $\rho := \pi_\theta(y \mid x)/\pi_{\theta_{\text{behav}}}(y \mid x)$. We further denote the optimal value by $J^\star := \sup_{\theta\in\mathbb{R}^d} J(\theta)$.

To find an optimal solution, at learner iteration $n \in \{1, \ldots, N\}$, where $N \geq 1$ is the total number of updates, GRPO style algorithms draws B prompts $x^{(1)}, \ldots, x^{(B)}$ from $\mathcal{D}$ and a group of $G \geq 2$ responses $y^{(b,1)}, \ldots, y^{(b,G)} \sim \pi_{\theta_{\text{behav}}}(\cdot \mid x^{(b)})$ per prompt, and takes the ascent step $\theta_{n+1} = \theta_n + \eta\, g_n^{\text{IS}}$ with step size $\eta > 0$, starting from θ_0 and ending at θ_N. Here g_n^{IS} is the stochastic policy gradient of eq. (1) on the returned mini-batch, normalized by the response length T,

$$g_n^{\text{IS}} := \frac{1}{TBG}\sum_{b=1}^{B}\sum_{i=1}^{G}\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}\big(y^{(b,i)} \mid x^{(b)}\big)\,\widehat{A}^{(b,i)}, \tag{2}$$

where $\rho_n^{(b,i)} := \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})/\pi_{\theta_{\text{behav}}}(y^{(b,i)} \mid x^{(b)})$ is the importance ratio evaluated at θ_n, and $\widehat{A}^{(b,i)} := r^{(b,i)} - \frac{1}{G-1}\sum_{j\neq i} r^{(b,j)}$ with $r^{(b,i)} := r_\phi(x^{(b)}, y^{(b,i)})$ being the leave-one-out group advantage. Both the ratio and the score function $\nabla_\theta \log \pi_\theta(y \mid x)$ act on the whole trajectory, so a single scalar advantage is shared by all tokens of $y^{(b,i)}$.

The policy gradient in eq. (2) depends on the policy through its score function and the importance sampling ratio. Our later analysis will be based on the following token-level regularity and coverage assumptions.

Assumption 1 (Token-level Lipschitz score)**.** *There is $L > 0$ such that, for every token context and all θ, θ',*

$$\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}) - \nabla_\theta \log \pi_{\theta'}(y_t \mid x, y_{<t})\|_2 \leq L\,\|\theta - \theta'\|_2\,.$$

Assumption 2 (Token-level coverage)**.** *There exists $\mathcal{C} \geq 1$ such that $\pi_\theta(y_t \mid x, y_{<t}) \geq 1/\mathcal{C}$ for every token, context, and learner or behavior policy encountered during training.*

Although the constant $\mathcal{C}$ can be large in LLM settings, it appears only through $\log \mathcal{C}$ in our analytical bounds, which justifies it as a controlled constant.

2.2 ASYNCHRONOUS RL SYSTEMS AND DATA STALENESS

As illustrated in Figure 1, we consider an asynchronous RL system in which multiple rollout workers generate responses continuously while a single learner updates the policy (Fu et al., 2026; Hu et al., 2026; Noukhovitch et al., 2025). Each worker draws a prompt $x \sim \mathcal{D}$ and generates a group of G responses under the policy snapshot it holds at launch. Once all responses have been scored, the worker deposits the group and its behavior-policy version into a shared buffer. At iteration n, the learner takes B completed groups from the buffer, performs the update in eq. (2) at θ_n, and publishes θ_{n+1} while rollout generation continues. Since groups start under different snapshots and take different amounts of time to generate, score, and wait in the buffer, a single mini-batch can mix several degrees of staleness.

Formally, let $s_n^{(b)} \in \{0, \dots, n\}$ denote the parameter-version index used to generate group b consumed at iteration n, and define its delay as $\tau_n^{(b)} := n - s_n^{(b)}$, the lag between the behavior-policy version and the current learner version, measured in learner updates. All G responses in a group share one delay, whereas groups within a mini-batch need not. For example, a batch used at θ_8 may contain groups generated under θ_5 and θ_7, with delays 3 and 1. The behavior policy of Section 2.1 is thus group dependent, $y^{(b,i)} \sim \pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})$, and the ratio in eq. (2) becomes

$$\rho_n^{(b,i)} = \frac{\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})}{\pi_{\theta_{n-\tau_n^{(b)}}}(y^{(b,i)} \mid x^{(b)})}. \tag{3}$$

To state the sampling conditions for eq. (3), let $\mathcal{F}_n^-$ collect the randomness available before the mini-batch is returned, so that $\{\theta_m\}_{m=0}^n$ is $\mathcal{F}_n^-$-measurable while the prompts and delays carried by the current batch need not be, and let $\mathcal{H}_n := \sigma(\mathcal{F}_n^-, \{x^{(b)}, \tau_n^{(b)}\}_{b=1}^B)$. Every behavior parameter $\theta_{n-\tau_n^{(b)}}$ is $\mathcal{H}_n$-measurable, so eq. (3) can be analyzed with the behavior policy held fixed. For our analysis, we impose the following regularity assumption on asynchronous sampling.

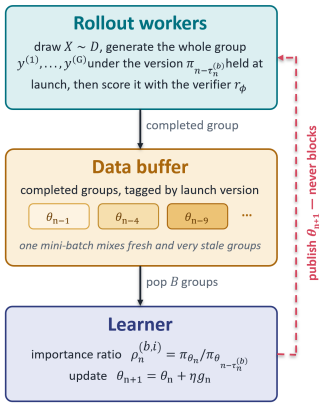


Figure 1: Asynchronous rollout. Each worker generates a group using its launch-time policy, a batch may contain groups with different delays.

Assumption 3 (Asynchronous sampling regularity)**.** *Conditional on $\mathcal{H}_n$, the responses within each returned group are independent draws from $\pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})$. The returned mini-batch preserves conditional unbiasedness of the raw importance-sampling update: $\mathbb{E}[g_n^{\mathrm{IS}} \mid \mathcal{F}_n^-] = T^{-1}\nabla J(\theta_n)$. Queue admission and mini-batch selection may otherwise depend on launch and completion times. For the sharper delay-moment conversion only, we additionally require the delay-regularity condition $\mathbb{E}[\|\theta_n - \theta_{n-t}\|_2^2 \mid \tau_n^{(b)} = t] = \mathbb{E}\|\theta_n - \theta_{n-t}\|_2^2$ whenever $\Pr(\tau_n^{(b)} = t) > 0$.*

Assumption 3 is a natural abstraction for stationary, large-scale RL training. With a fixed prompt stream, lossless rollout collection, and many concurrent workers, the delay of any individual rollout is approximately decoupled from both the distribution of returned samples and the aggregate learner trajectory. Thus, delay affects only the staleness of a rollout, without systematically biasing the returned data or the associated learner drift. We exclude RL training systems that explicitly filter rollouts by completion time, such as timeouts or maximum-delay gating (Fu et al., 2026; Zeng et al., 2026), as they can favor shorter responses and introduce length bias (Wang et al., 2025).

3 Second Moment-Bias Tradeoff and Group-Mass-Capping GRPO

How we design gradient estimators plays an important role in convergence, particularly in asynchronous training. In this section, we first establish a general convergence bound for a broad class of gradient estimators, which shows how their second moments and bias affect the convergence. This bound will then guide us toward a principled gradient estimator that minimizes a bias upper bound while maintaining a common second-moment guarantee, thereby reducing sensitivity to asynchronous rollout delays.

Lemma 1. *Suppose that $\theta_{n+1} = \theta_n + \eta g_n$ and $\mathbb{E}[g_n \mid \mathcal{F}_n^-] = T^{-1}\nabla J(\theta_n) + b_n$. Under Assumption 1-2, for every $\eta > 0$ and $N \geq 1$,*

$$\frac{\sum_{n=0}^{N-1} \mathbb{E}\|\nabla J(\theta_n)\|_2^2}{N} \leq \frac{2T\big(J^\star - \mathbb{E}[J(\theta_0)]\big)}{\eta N} + \frac{T^2}{N}\sum_{n=0}^{N-1}\Big[\eta RL(2\log|\mathcal{A}|+1)\mathbb{E}\|g_n\|_2^2 + \mathbb{E}\|b_n\|_2^2\Big].$$

The proof of Lemma 1 is deferred to Appendix D. Lemma 1 characterizes the convergence bound into three terms. The first is the standard decay term $O(1/(\eta N))$, which vanishes as the number N of iterations becomes large. The other two depend on the **gradient estimator's second moment** $\mathbb{E}\|g_n\|_2^2$ and its **squared bias** relative to the current policy gradient $\mathbb{E}\|b_n\|_2^2$. A good gradient estimator should therefore keep both quantities small. A common way to prevent gradient's second

moment become large is to *clip* individual samples whose importance ratios exceed a prescribed threshold, as in GRPO and its variants. However, clipping can also enlarge bias and slow convergence (Zheng et al., 2026). This tradeoff becomes especially important in asynchronous post-training, where delayed rollouts increase the mismatch between the behavior and learner policies. As this mismatch grows, more importance ratios may exceed the clipping threshold, increasing the bias introduced by clipping. Although raising the clipping threshold allows more samples to contribute, it weakens control over the gradient estimator's second moment and can lead to unstable training or even collapse (Zheng et al., 2026; Huang et al., 2026; Tian et al., 2026).

This tradeoff motivates us to consider a class of **weighted** gradient estimators, in which each sample receives an additional weight to adjust its contribution. We require the average weighted importance ratio within each group to stay below a fixed threshold, while allowing individual weighted ratios to exceed it. As we show below, this constraint gives all estimators in the class a common second-moment guarantee. We can then compare different weighting rules and choose one that minimizes the bias bound.

Definition 1 (Weighted gradient estimator class)**.** *Define the weighted gradient estimator as*

$$g_n^w := \frac{1}{B}\sum_{b=1}^{B}\frac{1}{TG}\sum_{i=1}^{G} \boldsymbol{w}_n^{(b,i)}\rho_n^{(b,i)}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}, \tag{4}$$

where the weights $w_n^{(b,i)}$ should satisfy, for every realized group,

$$0 \le w_n^{(b,i)} \le 1 \quad \textit{for all } i \in [G], \qquad \tfrac{1}{G}\textstyle\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)} \le 1+\epsilon. \tag{5}$$

Remark 1. In eq. (5), the first condition prevents the correction from amplifying any trajectory, and the second precisely controls the group's average gradient second moment. Such a class is broad, and includes many existing algorithms as special cases. For example $w_n^{(b,i)} = \mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}$ corresponds to TIC-GRPO (Pang et al., 2025); $w_n^{(b,i)} = \min\{1, c/\rho_n^{(b,i)}\}$ corresponds to VCPO (Huang et al., 2026) which applies sequence-level truncated importance sampling with a detached weight $\min\{\rho, c\}$; and CISPO (Chen et al., 2025) and SAPO (Gao et al., 2025) can be viewed as token-level analogues. However, the vanilla GRPO (Shao et al., 2024) is not in this class, because for negative advantages, it retains the gradient of tokens with arbitrarily large ratios, so that its second moment scales as $O(\mathcal{C}^2)$ in the worst case (Pang et al., 2025).

The following proposition shows that all gradient estimators in this class satisfy a common second-moment bound. This allows us to design a principled bias-minimal weighting rule.

Proposition 1. *Under Assumptions 1-3, the gradient estimators in the class given in Definition 1 share the following common bound*

$$\mathbb{E}[\|g_n^w\|_2^2 \mid \mathcal{F}_n^-] \le \tfrac{8}{T}(1+\epsilon)LR^2\log|\mathcal{A}|. \tag{6}$$

Further, their group bias $b_{n,b}^w := -\frac{1}{TG}\sum_{i=1}^{G}(1-w_n^{(b,i)})\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}$ satisfies

$$\left\|\mathbb{E}[b_{n,b}^w \mid \mathcal{H}_n]\right\|_2 \le 2R\sqrt{2L\log\mathcal{C}}\,\mathbb{E}\left[\tfrac{1}{G}\textstyle\sum_{i=1}^{G}(1-w_n^{(b,i)})\rho_n^{(b,i)} \mid \mathcal{H}_n\right]. \tag{7}$$

Proposition 1 shows that all weighting rules in this class share a common second-moment guarantee. With this guarantee in place, we focus on choosing weights that reduce the bias. Ideally, we would minimize the exact bias $\|\mathbb{E}[b_{n,b}^w \mid \mathcal{H}_n]\|_2$. However, this quantity depends on the directions of individual gradient contributions and how they cancel, which cannot generally be determined from the importance ratios alone. We therefore minimize the upper bound in eq. (7), which depends only on the importance ratios and the chosen weights. This bound is proportional to the expected reduction in the group's average importance ratio caused by weighting. Since the unweighted average $\frac{1}{G}\sum_{i=1}^{G}\rho_n^{(b,i)}$ is fixed for each realized group, minimizing this reduction is equivalent to maximizing the average weighted importance ratio subject to the class constraints:

$$\max_{0\le w_n^{(b,i)}\le 1} \tfrac{1}{G}\textstyle\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)} \quad \text{subject to} \quad \tfrac{1}{G}\textstyle\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)} \le 1+\epsilon. \tag{8}$$

The optimal weights are not unique in general. When the unweighted average importance ratio exceeds $1+\epsilon$, any feasible weighting rule whose average weighted importance ratio equals $1+\epsilon$

attains the optimum. Different optimal rules can nevertheless produce different gradient directions within the same group. To select among these rules, we restrict attention to *direction-preserving* rules that assign a common positive weight to all responses in a group: $w_n^{(b,i)} = \alpha_n^{(b)} > 0$ for all $i \in [G]$. This scales the original importance-weighted group gradient estimator by a single factor, preserving the relative contributions of the responses and hence the direction of the group gradient estimate. The following proposition gives the optimal weighting rule under this restriction.

Proposition 2. *Fix a realized group and let* $\overline{\rho}_n^{(b)} := G^{-1}\sum_{i=1}^{G}\rho_n^{(b,i)}$. *Among weighted direction-preserving corrections, the unique solution to eq.* (8) *is* $\alpha_n^{(b),\star} = \min\left\{1, (1+\epsilon)/\overline{\rho}_n^{(b)}\right\}$.

We call this weighting rule *group mass capping* (GMC), which caps the group's average importance ratio by replacing each raw ratio with $\widehat{\rho}_n^{(b,i)} := \alpha_n^{(b),\star}\rho_n^{(b,i)}$, yielding the following gradient estimator:

$$g_n^{\mathrm{GMC}} := \frac{1}{TBG}\sum_{b=1}^{B}\sum_{i=1}^{G}\frac{\rho_n^{(b,i)}}{\max\{1, \bar{\rho}_n^{(b)}/(1+\epsilon)\}}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}. \tag{9}$$

We provide a comprehensive pseudo code of asynchronous GMC-GRPO in Appendix C.

4 Convergence Analysis

4.1 Convergence Guarantees for GMC-GRPO

We now apply Lemma 1 to establish convergence guarantees for asynchronous GMC-GRPO. Since Proposition 1 already controls the gradient second moment, it remains to bound the correction bias in terms of rollout delays. The resulting bound depends on the averaged fourth moment of the delay defined as $\overline{\tau}_4 := \frac{1}{NB}\sum_{n=0}^{N-1}\sum_{b=1}^{B}\mathbb{E}[(\tau_n^{(b)})^4]$.

Theorem 1 (Convergence of asynchronous GMC-GRPO)**.** *Under Assumptions 1–3, for every* $\eta > 0$ *and* $N \geq 1$*, the convergence rate of GMC-GRPO satisfies*

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla J(\theta_n)\right\|_2^2 = O\left(\frac{1}{\eta N} + (1+\epsilon)\eta + \frac{(1+\epsilon)^3}{\log^2(1+2\epsilon)}\eta^4\overline{\tau}_4\right).$$

For $\overline{\tau}_4 > 0$*, taking* $\eta^\star \asymp \min\{1/\sqrt{(1+\epsilon)N}, (\log^2(1+2\epsilon)/((1+\epsilon)^3\overline{\tau}_4 N))^{1/5}\}$ *gives*

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\|\nabla J(\theta_n)\|_2^2 = O\left(\sqrt{\frac{1+\epsilon}{N}} + \left(\frac{(1+\epsilon)^3\overline{\tau}_4}{\log^2(1+2\epsilon)N^4}\right)^{1/5}\right).$$

When $\overline{\tau}_4 = 0$*, the convergence recovers the standard* $O(N^{-1/2})$ *rate for fixed* ϵ*.*

Remark 2 (Effect of rollout delay on convergence rate). We use Theorem 1 to characterize how much rollout delay can be tolerated without degrading the convergence rate. Suppose $\overline{\tau}_4 > 0$. If $\overline{\tau}_4 = O(N^{3/2})$, choosing $\epsilon \asymp \min\{1, \overline{\tau}_4^{1/7}N^{-3/14}\}$ and $\eta \asymp N^{-1/2}$ preserves the no-delay $O(N^{-1/2})$ convergence guarantee; the delay contribution is lower order if $\overline{\tau}_4 = o(N^{3/2})$. For larger scale delays, where $\overline{\tau}_4 = \Theta(N^p)$ with $3/2 < p < 4$, the rate-optimal choices are $\epsilon = \Theta(1)$ and $\eta \asymp N^{-(p+1)/5}$, which yield the slower convergence guarantee $O(N^{-(4-p)/5})$. Thus, the bound tolerates fourth delay moments up to $N^{3/2}$ without degrading the standard rate. Under polynomial scaling, moments below N^4 still yield a vanishing convergence guarantee, but at a slower convergence rate.

Corollary 1 (Asymptotically optimal GMC threshold)**.** *Under the conditions of Theorem 1, assume that* $\overline{\tau}_4 > 0$ *does not scale with* N*. Then as* $N \to \infty$*, any minimizer of the leading* ϵ*-dependent terms over* $\epsilon > 0$ *in the derived GMC upper bound satisfies* $\epsilon^\star_{\mathrm{GMC}} = \Theta\left(\overline{\tau}_4^{1/7}N^{-3/14}\right)$.

Corollary 1 reflects a simple intuition: larger delays favor a higher threshold to avoid reducing samples' contributions too aggressively, while more training iterations favor a lower threshold to keep the gradient estimator's second moment small.

Corollary 2. *Under the conditions of Theorem 1, further suppose that for some finite* $\delta \geq 0$, $|\rho_n^{(b,i)} - 1| \leq \delta$ *almost surely for all* n, b, i*. For* $\overline{\tau}_4 > 0$*, taking* $\eta^\star \asymp \min\{1/\sqrt{(1+\epsilon)N}, (G^2\epsilon^2/((1+$

$\delta)^2(1+\epsilon)^3\overline{\tau}_4 N))^{1/5}\}$, *then the group-concentration branch in the GMC bound gives*

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\|\nabla J(\theta_n)\|_2^2 = O\left(\sqrt{\frac{1+\epsilon}{N}} + \left(\frac{(1+\delta)^2(1+\epsilon)^3\overline{\tau}_4}{G^2\epsilon^2 N^4}\right)^{1/5}\right).$$

Corollary 2 shows that the delay-dependent term decreases as $G^{-2/5}$, so larger groups reduce the effect of rollout delays on the convergence bound.

Remark 3. Corollary 2 reflects a simple distinction between group-mass capping (GMC) and individual clipping. Importance ratios have conditional mean one under the behavior policy, and GMC intervenes only when their empirical group average is unusually large. As the group size grows, this average concentrates around one, so this excess as well as the bias introduced by correcting it vanishes. In contrast, individual clipping makes a separate decision for every trajectory (Pang et al., 2025; Huang et al., 2026); averaging more such decisions does not remove the clipping bias already introduced at the trajectory level.

The condition $|\rho_n^{(b,i)} - 1| \le \delta$ in Corollary 2 identifies a near-policy regime in which this concentration benefit carries over to delayed training. When the behavior and current policies assign comparable likelihoods to the sampled trajectories, the importance ratios remain bounded, and the delay term decays with the group size. As this overlap deteriorates, the constant $(1+\delta)^2$ grows and the refinement loses its advantage, although Theorem 1 continues to apply. The group-size gain should be interpreted as a local benefit of aggregating trajectories with controlled likelihood ratios, rather than a universal claim that larger groups eliminate bias under arbitrary distribution shift.

4.2 Proof Outline and the Role of Rollout Delay

We outline the proof of Theorem 1, which further explains how rollout delay affects convergence of asynchronous training. Theorem 1 follows from Lemma 1 by bounding the second moment and bias term separately. Proposition 1 controls the second-moment term uniformly, independently of delay, which yields the $O(\eta)$ term. Delay therefore enters only through the correction bias, via a three-step chain. First, Proposition 1 bounds the bias of each group by the importance ratio it removes, which for GMC equals the group average's excess over the cap: $\mathbb{E}[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ \mid \mathcal{H}_n]$. Second, a convexity argument bounds this excess by $D_n^{(b)}/\log(1+2\epsilon)$, where $D_n^{(b)} := D_{\mathrm{KL}}(\pi_{\theta_n}(\cdot \mid x^{(b)}) \,\|\, \pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)}))$ measures the mismatch between the current and behavior policies. Third, Assumption 1 gives $D_n^{(b)} \le \frac{TL}{2}\|\theta_n - \theta_{n-\tau_n^{(b)}}\|_2^2$, and the uniform second-moment bound limits the parameter drift accumulated over $\tau_n^{(b)}$ updates to $O(\eta^2(\tau_n^{(b)})^2)$. Squaring the bias thus produces the term of order $\eta^4\overline{\tau}_4/\log^2(1+2\epsilon)$. The second-moment control therefore plays a dual role: it stabilizes each update, and it also limits how far the policy can drift within a delay window.

Remark 4 (Comparison with asynchronous SGD). Delay plays a different role in Theorem 1 than in classical asynchronous stochastic gradient descent (ASGD) (Tan et al., 2023; Maranjyan et al., 2025; Koloskova et al., 2022; Mishchenko et al., 2022). In ASGD, an update applied at iteration n is conditionally unbiased for the gradient at the stale iterate $\theta_{n-\tau_n}$ rather than at the current iterate θ_n, so staleness enters the descent argument directly through the discrepancy between the two gradients. In our setting, the trajectory-level ratio instead changes measure from the behavior policy $\pi_{\theta_{n-\tau_n^{(b)}}}$ to the current policy π_{θ_n}, so the uncorrected estimator remains unbiased for $\nabla J(\theta_n)$ regardless of delay. Staleness instead enlarges the spread of the importance ratios, which makes a second-moment correction necessary; delay then affects the bound only through the bias of that correction as explained above. This difference also changes the appropriate remedy. ASGD commonly bounds staleness by discarding updates whose delay exceeds a threshold (Tan et al., 2023; Maranjyan et al., 2025), but the analogous gating in LLM post-training preferentially discards long responses and thus induces a length bias (Wang et al., 2025). Improving the delay dependence therefore calls for RL algorithms that minimize the bias under the same second-moment control, which is precisely what group-mass-capping is designed to do.

4.3 Comparison with TIC-GRPO

We now compare GMC-GRPO with TIC-GRPO (Pang et al., 2025), which replaces the token-level ratios of GRPO with a single trajectory-level ratio and clips it from above, corresponding to

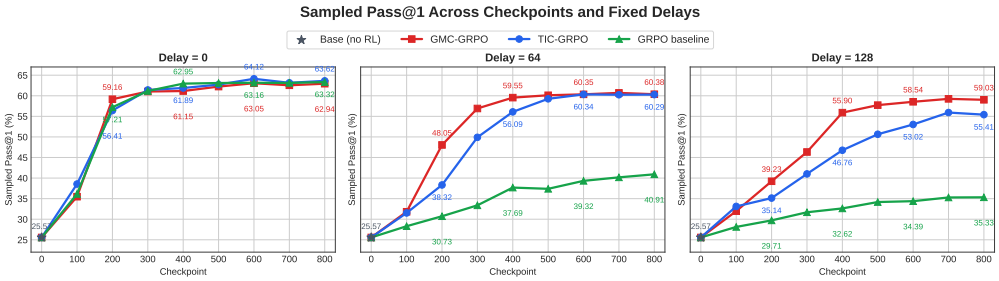


Figure 2: Sampled Pass@1 on MATH500 across training checkpoints under fixed rollout delays. GMC-GRPO remains robust as staleness increases, substantially outperforming vanilla GRPO and competitive with or exceeding TIC-GRPO.

$w_n^{(b,i)} = \mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}$ in eq. (4). TIC-GRPO is a natural baseline for two reasons. First, it provably converges faster than GRPO (Pang et al., 2025). Second, it belongs to the weighted gradient estimator class and therefore shares the second-moment bound of Proposition 1 with GMC-GRPO. The following theorem establishes the convergence of TIC-GRPO under stochastic delays.

Theorem 2 (Convergence of asynchronous TIC-GRPO)**.** *Let* $c_\epsilon := \log(1+\epsilon) - \epsilon/(1+\epsilon)$. *Under Assumptions 1–3, for every* $\eta > 0$ *and* $N \ge 1$, *TIC-GRPO satisfies the bound of Theorem 1 with* $\log^2(1+2\epsilon)$ *replaced by* c_ϵ^2. *Consequently, for* $\overline{\tau}_4 > 0$, *taking* $\eta^\star \asymp \min\{1/\sqrt{(1+\epsilon)N}, (c_\epsilon^2/((1+\epsilon)^3\overline{\tau}_4 N))^{1/5}\}$ *gives*

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\|\nabla J(\theta_n)\|_2^2 = O\left(\sqrt{\frac{1+\epsilon}{N}} + \left(\frac{(1+\epsilon)^3\overline{\tau}_4}{c_\epsilon^2 N^4}\right)^{1/5}\right).$$

Remark 5 (Convergence bound improvement). Theorems 1 and 2 differ only in the fourth-order delay coefficient: TIC-GRPO replaces $\log^2(1+2\epsilon)$ in the GMC-GRPO bound with c_ϵ^2. Since $c_\epsilon < \log(1+2\epsilon)$ for every $\epsilon > 0$, the GMC coefficient is strictly smaller, and after optimizing the step size the delay-dependent rate improves by the factor $(c_\epsilon/\log(1+2\epsilon))^{2/5} < 1$. This distinction is substantial for a tight threshold. As $\epsilon \to 0$, $c_\epsilon = \Theta(\epsilon^2)$ while $\log(1+2\epsilon) = \Theta(\epsilon)$; hence the TIC penalty scales as ϵ^{-4}, whereas the GMC penalty scales as ϵ^{-2}. GMC obtains this improvement without enlarging the conditional second-moment upper bound.

This improvement arises from how much importance weight each algorithm removes. Following the discussion in Section 4.2, GMC-GRPO removes only the amount by which the group average exceeds $1+\epsilon$, whereas TIC-GRPO removes the full contribution of every trajectory above the threshold. The KL divergence between the current and behavior policies bounds these two quantities with denominators $\log(1+2\epsilon)$ and c_ϵ, respectively, producing the coefficient difference above.

Finally, the delay term in Theorem 2 does not improve with the group size, while GMC-GRPO benefits from larger groups under local overlap (Remark 3). Section 5 examines whether these theoretical gaps translate into faster training in practice.

5 Experiments

In this section, we evaluate GMC-GRPO against several baselines: (i) vanilla GRPO (Shao et al., 2024); (ii) GAC, which controls stale-aligned update directions through gradient projection (Xu et al., 2026); (iii) VCPO, which combines effective-sample-size-guided step-size scaling with an off-policy minimum-variance baseline (Huang et al., 2026); and (iv) TIC-GRPO (Pang et al., 2025). Among these methods, GAC and VCPO are designed specifically for asynchronous training. We first evaluate robustness under controlled fixed delays and then construct a simulated asynchronous system in which rollout delay is correlated with response length. Detailed parameter settings and implementation details are deferred to Appendix B.1. We also discuss related approaches that selectively clip a fraction of importance ratios (Fu et al., 2026; Li et al., 2025) and report an additional comparison with A-3PO (Li et al., 2025) in Appendix B.2.

Table 1: Performance under simulated asynchronous delays. Each benchmark cell is greedy/avg@n, with $n = 16$ for MATH500, Minerva, and AMC23, and $n = 8$ for GSM8K, Gaokao, and Olympiad-Bench. *Avg. perf.* is the average over the six benchmarks, reported as mean greedy/mean avg@n.

Method	Avg. Delay	MATH500	GSM8K	Gaokao	Olympiad	Minerva	AMC23	Avg. perf.
				Qwen3-1.7B				
GRPO	158.70	48.20 / 31.04	72.18 / 47.90	39.74 / 26.27	14.69 / 6.77	16.18 / 8.32	22.50 / 9.53	35.58 / 21.64
GAC	160.02	47.00 / 31.03	72.18 / 47.09	40.26 / 25.16	15.73 / 6.70	15.81 / 8.18	20.00 / 11.09	35.16 / 21.54
VCPO	157.87	65.20 / 54.55	80.21 / 67.76	55.06 / 45.06	27.60 / 20.49	26.47 / 18.31	27.50 / 26.72	47.01 / 38.82
TIC-GRPO	159.60	62.20 / 53.61	**82.34** / 67.14	53.25 / 43.02	27.60 / 19.51	**26.84** / 17.53	32.50 / 26.25	47.45 / 37.84
GMC-GRPO	159.75	**66.00 / 57.29**	81.65 / **72.66**	**55.32 / 47.73**	**28.19 / 22.61**	26.47 / **20.11**	**35.00 / 28.44**	**48.77 / 41.47**
				Qwen3-4B				
GRPO	137.92	57.40 / 43.26	74.30 / 51.56	48.57 / 34.35	22.11 / 9.66	18.01 / 11.99	35.00 / 17.19	42.57 / 28.00
GAC	138.85	58.00 / 42.39	77.79 / 55.32	50.39 / 33.64	21.81 / 10.55	16.18 / 12.22	35.00 / 18.91	43.19 / 28.84
VCPO	139.50	61.00 / 48.39	76.88 / 64.30	46.75 / 36.30	29.38 / 13.61	26.84 / 16.98	37.50 / 21.88	46.39 / 33.58
TIC-GRPO	138.87	76.40 / 68.31	90.75 / 81.70	**65.19** / 55.68	42.88 / 33.33	**36.03** / 27.55	**57.50** / 43.91	**61.46** / 51.75
GMC-GRPO	141.01	**77.60 / 72.54**	**91.13 / 87.03**	**65.19 / 60.91**	**43.18 / 37.72**	34.19 / **30.10**	52.50 / **48.91**	60.63 / **56.20**
				Qwen3-8B				
GRPO	140.90	61.80 / 49.41	88.70 / 72.74	55.06 / 38.83	27.30 / 16.49	21.69 / 16.96	25.00 / 22.03	46.59 / 36.08
GAC	141.7	61.20 / 48.34	88.10 / 72.09	54.55 / 39.29	26.41 / 16.26	22.79 / 15.99	27.50 / 21.41	46.76 / 35.56
VCPO	142.7	67.40 / 58.88	88.93 / 78.79	57.40 / 47.50	33.38 / 21.88	31.25 / 23.09	47.50 / 31.56	54.31 / 43.62
TIC-GRPO	141.6	**79.20** / 75.14	92.87 / 87.92	68.31 / 61.20	**45.10** / 38.41	**40.07 / 34.81**	**60.00** / 50.47	64.26 / 57.99
GMC-GRPO	141.1	78.80 / **75.22**	**93.33 / 91.37**	**69.09 / 63.64**	44.96 / **40.34**	**40.07** / 34.35	**60.00 / 52.81**	**64.37 / 59.62**

Controlled Fixed Delays. To isolate the effect of staleness, we evaluate GRPO, TIC-GRPO, and GMC-GRPO on Qwen3-1.7B under fixed rollout delays. Figure 2 reports sampled Pass@1 on MATH500 for delays of $D = 0, 64, 128$ learner updates; see Appendix B.1. The full sweep over $D \in \{0, 32, 64, 128, 256, 512\}$ is deferred to Appendix B.2. Without delay, all three methods attain similar final performance. At delay 64, GMC-GRPO learns faster than TIC-GRPO and reaches comparable final Pass@1, with both outperforming vanilla GRPO. At delay 128, GMC-GRPO achieves 59.03% Pass@1, compared with 55.41% for TIC-GRPO and 35.33% for vanilla GRPO. These results show that GMC supports faster learning and stronger final performance as staleness increases.

Simulated Asynchronous Delays. The fixed-delay study isolates staleness, but real rollout workers finish at different times: longer responses typically take longer to generate and thus reach the learner under older policies. To capture this coupling, we simulate an asynchronous pipeline in which rollout delay depends on generation length, producing heterogeneous delays averaging approximately 160 learner updates for Qwen3-1.7B and 140 for Qwen3-4B and Qwen3-8B. This setting tests whether GMC-GRPO remains robust when staleness varies across rollouts. Among the methods reported in Table 1, GMC-GRPO achieves the highest mean avg@n scores across all three model sizes (41.47, 56.20, and 59.62) and the highest mean greedy scores on Qwen3-1.7B and Qwen3-8B (48.77 and 64.37). On Qwen3-4B, its mean greedy score of 60.63 is 0.83 points below TIC-GRPO. Notably, GMC-GRPO also outperforms the two of the most recent asynchronous-RL stabilization methods: GAC, which controls stale-aligned update directions via gradient projection, and VCPO, which combines ESS-guided step-size scaling with a minimum-variance baseline. Both improve over vanilla GRPO under delay but remain below TIC-GRPO-style trajectory correction at 4B and 8B. These results indicate that the robustness of group-mass capping is not restricted to fixed delays, but holds under heterogeneous, length-dependent staleness.

6 Conclusions and Limitations

This work studies how rollout staleness affects the convergence of GRPO-style algorithms in asynchronous LLM training and introduces GMC-GRPO to improve learning from delayed rollouts. Our theory and experiments show that group-level capping improves robustness to stale rollouts while maintaining performance on fresh data. Within the proposed class of weighted gradient estimators, our analysis provides a foundation for understanding and improving delay tolerance. Future work can extend this framework to more general weighting rules, such as those that use information beyond importance ratios, to further improve the balance between second-moment control and bias.

AI USE STATEMENT

We used generative AI tools to improve the clarity of our writing, assist with checking proofs, and write code for experiments. We did not use these tools to originate the core research ideas or formulate the theorem statements. All AI-assisted content was reviewed by the authors. In particular, AI-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

We provide the information needed to examine and reproduce our theoretical and empirical results. Assumptions 1–3 are stated in Section 2. GMC-GRPO is defined in Section 3, and Appendix C gives pseudocode for the synchronous and asynchronous algorithms. Complete proofs of all theoretical results appear in Appendices D–H. Appendix B.1 documents the detailed experiment setup, while Appendix B.2 reports supplemental experiment results. We also provide the code used for our experiments as a ZIP archive in the supplementary materials.

REFERENCES

Charles Arnal, Vivien Cabannes, Taco Cohen, Julia Kempe, and Remi Munos. Efficient rl training for llms with experience replay. *arXiv preprint arXiv:2604.08706*, 2026.

Art of Problem Solving. AMC problems and solutions. `https://artofproblemsolving.com/wiki/index.php/AMC_Problems_and_Solutions`, 2024. Accessed: 2026-08-29.

Brian Bartoldson, Siddarth Venkatraman, James Diffenderfer, Moksh Jain, Tal Ben-Nun, Seanie Lee, Minsu Kim, Johan Obando Ceron, Yoshua Bengio, and Bhavya Kailkhura. Trajectory balance with asynchrony: Decoupling exploration and learning for fast, scalable llm post-training. *Advances in Neural Information Processing Systems*, 38:113901–113931, 2026.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

Madeleine Dwyer, Adam Sobey, and Adriane Chapman. It's not you, it's clipping: A soft trust-region via probability smoothing for llm rl. *arXiv preprint arXiv:2509.21282*, 2025.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *Advances in Neural Information Processing Systems*, 38:36256–36282, 2026.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Tianhao Hu, Xiangcheng Liu, Yuchun Miao, Youshao Xiao, Hongyu Zang, Yang Zheng, Xuan Huang, Jinrui Ding, Yufei Zhang, Yu Yang, et al. Dora: A scalable asynchronous reinforcement learning system for language model training. *arXiv preprint arXiv:2604.26256*, 2026.

Audrey Huang, Adam Block, Dylan Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *International Conference on Learning Representations*, volume 2025, pp. 76687–76739, 2025a.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason Lee, Wen Sun, Akshay Krishnamurthy, and Dylan Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 92647–92701, 2025b.

Luke J Huang, Zhuoyang Zhang, Qinghao Hu, Shang Yang, and Song Han. Stable asynchrony: Variance-controlled off-policy rl for llms. *arXiv preprint arXiv:2602.17616*, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International conference on machine learning*, pp. 1453–1461. PMLR, 2013.

Anastasiia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous sgd for distributed and federated learning. *Advances in Neural Information Processing Systems*, 35:17202–17215, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Xiaocan Li, Shiliang Wu, and Zheng Shen. A-3po: Accelerating asynchronous llm training with staleness-aware proximal policy approximation. *arXiv preprint arXiv:2512.06547*, 2025.

Minpeng Liao, Chengxi Li, Wei Luo, Wu Jing, and Kai Fan. Mario: Math reasoning with code interpreter output-a reproducible pipeline. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 905–924, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Artavazd Maranjyan, Alexander Tyurin, and Peter Richtárik. Ringmaster asgd: The first asynchronous sgd with optimal time complexity. *arXiv preprint arXiv:2501.16168*, 2025.

Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake E Woodworth. Asynchronous sgd beats minibatch sgd under arbitrary delays. *Advances in Neural Information Processing Systems*, 35:420–433, 2022.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous rlhf: Faster and more efficient off-policy rl for language models. In *International Conference on Learning Representations*, volume 2025, pp. 4003–4029, 2025.

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.

Lei Pang, Jun Luo, and Ruinan Jin. Tic-grpo: Provable and efficient optimization for reinforcement learning from human feedback. *arXiv preprint arXiv:2508.02833*, 2025.

Daniel Ritter, Owen Oertell, Bradley Guo, Jonathan Chang, Kianté Brantley, and Wen Sun. Llms can learn to reason via off-policy rl. *arXiv preprint arXiv:2602.19362*, 2026.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. *arXiv preprint arXiv:2503.14286*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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*, pp. 1279–1297, 2025.

Qiao Tan, Feng Zhu, and Jingjing Zhang. Abs: Adaptive bounded staleness converges faster and communicates less. *arXiv preprint arXiv:2301.08895*, 2023.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can grpo be? mu-grpo for efficient llm reinforcement learning. *arXiv preprint arXiv:2605.17570*, 2026.

Alexander Tyurin, Andrei Spiridonov, and Varvara Rudenko. Asynchronous policy gradient aggregation for efficient distributed reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8826–8857, 2026.

Liangyu Wang, Huanyi Xie, Xinhai Wang, Tianjin Huang, Mengdi Li, and Di Wang. Infinite sampling: Efficient and stable grouped rl training for large language models. *arXiv preprint arXiv:2506.22950*, 2025.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Xun Deng, Zhihao Zhang, et al. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. In *International Conference on Learning Representations*, volume 2026, pp. 126204–126228, 2026.

Tengyang Xie, Dylan Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed H Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. In *International Conference on Learning Representations*, volume 2025, pp. 43632–43669, 2025.

Haofeng Xu, Junwei Su, Yukun Tian, Lansong Diao, Zhengping Qian, and Chuan Wu. Gac: Stabilizing asynchronous rl training for llms via gradient alignment control. *arXiv preprint arXiv:2603.01501*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Guanqun Zhao, Zijun Xie, Binbin Zheng, Enlei Gong, Jiafeng Lu, Yehan Yang, Aoqi Hu, and Zeyu Chen. Deconstructing off-policy ratios: Entropy-scaled trust regions for asynchronous reinforcement learning. *arXiv preprint arXiv:2607.22186*, 2026.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms? In *International Conference on Learning Representations*, volume 2026, pp. 82657–82679, 2026.

Yinmin Zhong, Zili Zhang, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, et al. Rlhfuse: Efficient rlhf training for large language models with inter-and intra-stage fusion. *arXiv preprint arXiv:2409.13221*, 1, 2024.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a u-statistic. *arXiv preprint arXiv:2603.01162*, 2026.

Appendix Contents

A. Related Work 14

B. Experiment Details and Supplemental Experiments 15

B.1. Detailed Experiment Setup 15

B.2. Supplemental Experiment Results 18

C. Mixed-Delay Setting, GMC-GRPO and TIC-GRPO Algorithms 21

D. Technical Preliminaries 23

E. Optimization over a Second-Moment-Controlled Weighting Class 26

F. Convergence Analysis of GMC-GRPO 28

G. Convergence Analysis of TIC-GRPO 34

H. From the Appendix Bounds to the Main-Text Rates 38

Table 2: Summary of the main notation used in the paper.

Notation	Description
$\mathcal{X}, \mathcal{D}, \mathcal{A}$	Prompt space, training prompt distribution, and token vocabulary.
$x, y, y_t, y_{<t}$	Prompt, response, token at position t, and prefix $(y_1, \dots, y_{t-1})$.
T	Maximum response length; also the normalization factor in each gradient update.
$\theta, \pi_\theta(y \mid x)$	Policy parameter in $\mathbb{R}^d$ and trajectory likelihood $\prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$.
$r_\phi(x, y), R$	Verifier score and its upper bound: $r_\phi(x, y) \in [0, R]$.
$J_x(\theta), J(\theta), J^\star$	Expected reward for prompt x, its average over $x \sim \mathcal{D}$, and $\sup_\theta J(\theta)$; see eq. (1).
n, N, η	Learner update index $n \in \{0, \dots, N-1\}$, N total updates, and step size $\eta > 0$.
B, G, b, i	Groups per mini-batch, responses per group ($G \geq 2$), group index $b \in [B]$, and response index $i \in [G]$.
$\mathcal{M}_n$	Mini-batch of B completed prompt groups consumed at iteration n.
$\theta_n, \theta_{n-\tau_n^{(b)}}$	Current learner parameter and behavior parameter used to generate group b.
$\tau_n^{(b)}$	Number of learner updates between the launch and consumption of group b; $\tau_n^{(b)} \in \{0, \dots, n\}$.
$\overline{\tau}_k$	Averaged kth delay moment: $\frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}[\frac{1}{B}\sum_{b=1}^{B} (\tau_n^{(b)})^k]$, $k \geq 1$.
$\widehat{A}^{(b,i)}$	Leave-one-out advantage $r^{(b,i)} - \frac{1}{G-1}\sum_{j \neq i} r^{(b,j)}$, where $r^{(b,i)} = r_\phi(x^{(b)}, y^{(b,i)})$.
$\rho_n^{(b,i)}$	Current-to-behavior trajectory likelihood ratio $\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)}) / \pi_{\theta_{n-\tau_n^{(b)}}}(y^{(b,i)} \mid x^{(b)})$.
$\overline{\rho}_n^{(b)}$	Average raw importance ratio in group b: $G^{-1}\sum_{i=1}^{G} \rho_n^{(b,i)}$.
ϵ, c_ϵ	Threshold parameter $\epsilon > 0$ (the cap is $1+\epsilon$), and $c_\epsilon = \log(1+\epsilon) - \epsilon/(1+\epsilon)$.
$w_n^{(b,i)}$	Correction weight, held fixed during differentiation, satisfying $0 \leq w_n^{(b,i)} \leq 1$ and $G^{-1}\sum_i w_n^{(b,i)} \rho_n^{(b,i)} \leq 1 + \epsilon$.
$\alpha_n^{(b),\star}$	Common GMC scaling factor $\min\{1, (1+\epsilon)/\overline{\rho}_n^{(b)}\}$.
$\widehat{\rho}_n^{(b,i)}$	GMC effective importance ratio $\alpha_n^{(b),\star} \rho_n^{(b,i)}$.
$g_n^{\mathrm{GMC}}, g_n^{\mathrm{TIC}}, g_n^{w}$	Update estimators for GMC, TIC, and a general weighted correction; see eq. (9), eq. (35), and eq. (4).
$\mathcal{F}_n^-, \mathcal{H}_n$	Information before observing the returned mini-batch, and $\mathcal{H}_n = \sigma(\mathcal{F}_n^-, \{x^{(b)}, \tau_n^{(b)}\}_{b=1}^{B})$ after revealing its prompts and delays.
b_n	Conditional update bias $\mathbb{E}[g_n \mid \mathcal{F}_n^-] - T^{-1}\nabla J(\theta_n)$ for the estimator g_n under study.
$L, \mathcal{C}$	Token-score Lipschitz constant and token-coverage constant $(\pi_\theta(y_t \mid x, y_{<t}) \geq 1/\mathcal{C})$; see Assumptions 1 and 2.
$D_n^{(b)}, \chi_n^{2,(b)}$	Trajectory KL and chi-square divergences from $\pi_{\theta_n}(\cdot \mid x^{(b)})$ to $\pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})$; see Appendix C.
δ	Local-overlap bound: $\lvert\rho_n^{(b,i)} - 1\rvert \leq \delta$ almost surely, for the refinement in Corollary 2.

A Related Work

Asynchronous RL Systems for LLM Post-Training. LLM RL pipelines typically alternate between autoregressive rollout generation and policy optimization, incurring idle time from variable response lengths and repeated switches between inference and training (Zhong et al., 2024; Sheng et al., 2025). Recent work improves this pipeline along two main directions. Asynchronous RLHF and LLM RL systems decouple rollout generation from policy optimization so that the two stages can proceed concurrently (Noukhovitch et al., 2025; Fu et al., 2026; Hu et al., 2026). Complementary approaches improve rollout-side utilization through continuous grouped sampling or scale exploration and learning through asynchronous replay (Wang et al., 2025; Bartoldson et al., 2026). This decoupling, however, causes the learner to consume trajectories generated by lagged behavior policies, making rollout staleness an intrinsic off-policy optimization issue. Rather than proposing another systems architecture, we aim at finding robust algorithms against data staleness in asyn-

chronous RL systems: we analyze GRPO-style optimization under stochastic rollout delays and design an importance correction that controls the resulting bias without sacrificing stability.

Off-Policy Correction for Stale LLM Rollouts. Stale rollouts can induce heavy-tailed importance ratios, allowing a small number of trajectories to dominate the gradient estimate and destabilize training. Learning from such data therefore requires a careful choice of how the update is corrected. Existing corrections differ mainly in where they intervene. A common strategy acts locally on individual ratios: TIC-GRPO clips a trajectory once its ratio exceeds the clipping threshold (Pang et al., 2025), while related methods replace fixed clipping with tapered weights, adaptive or asymmetric thresholds, and entropy-scaled trust regions (Roux et al., 2025; Xi et al., 2026; Zhao et al., 2026; Li et al., 2025). A second strategy controls batch- or update-level statistics through second-moment constraints, variance-aware step sizes, or gradient-alignment control (Zheng et al., 2026; Huang et al., 2026; Xu et al., 2026). A third changes how stale experience is reused through staged training with selective filtering, replay buffers, or objectives designed for lagged inference policies (Tian et al., 2026; Arnal et al., 2026; Ritter et al., 2026). Together, these methods show that stale data can remain informative when its effect on the update is controlled. Our work instead moves from individual correction to group-level control. We identify a cap on corrected importance mass that preserves gradient second-moment bound and derive a uniform rescaling rule that reduces delay-induced bias without selectively clipping trajectories.

Theory of GRPO-Style Optimization under Delay. The convergence theory of asynchronous optimization is well developed, with rigorous and tight convergence analysis that matches the lower bound under various settings (Joulani et al., 2013; Koloskova et al., 2022; Mishchenko et al., 2022; Tyurin et al., 2026). In parallel, a growing line of work establishes theoretical guarantees for RL-based LLM post-training (Xie et al., 2025; Huang et al., 2025b; Foster et al., 2025; Huang et al., 2025a). Despite this progress, most existing RL-for-LLM theories are developed within the natural policy gradient and RLHF frameworks, and rigorous convergence guarantees for widely used GRPO-style algorithms remain limited, particularly under asynchronous rollout generation. Zhou et al. (2026) characterize the GRPO gradient as a U-statistic and analyze its finite-sample error and dependence on group size, while Pang et al. (2025) introduce TIC-GRPO and provide the first rigorous convergence analysis of GRPO-style methods. More recently, Xu et al. (2026) analyze asynchronous GRPO through a generic biased-gradient framework, but leave the staleness-induced bias and higher-order remainder terms as abstract quantities in the final bound. **To the best of our knowledge, we are the first to provide an explicit delay-dependent convergence analysis of GRPO-style algorithms in asynchronous RL systems.**

B Experiment Details and supplemental Experiments

B.1 Detailed Experiment Setup

Models and training data. We run the experiments with the Qwen3-1.7B-Base, Qwen3-4B-Base and Qwen3-8B-Base models (Yang et al., 2025). All methods update the full actor; we use neither LoRA adapters nor an explicit KL-divergence penalty. The training set is constructed from the seven subject-specific training splits of MATH (Hendrycks et al., 2021): algebra, counting and probability, geometry, intermediate algebra, number theory, prealgebra, and precalculus. After removing examples for which the reference answer cannot be extracted, the resulting training set contains 7,498 problems. Training and evaluation use the same two-shot prompt, which instructs the model to reason step by step and place its final answer in `\boxed{}`.

Shared training configuration. We generate rollouts with vLLM (Kwon et al., 2023) in BF16, using temperature 1.0 and top-$p = 1.0$. Prompts and completions are truncated at 1,024 and 4,096 tokens, respectively. Each rollout mini-batch contains eight prompts and eight responses per prompt, for a total of 64 trajectories. Except for the zero-delay runs in the fixed-delay suite described below, we set `inner_steps=1`, so these 64 trajectories produce one actor update.

We optimize with 8-bit AdamW (Dettmers et al., 2021), a learning rate of 10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and zero weight decay. The learning rate follows a cosine schedule with a 3% warmup, and the gradient norm is clipped at 1.0. We use a mix of NVIDIA RTX 6000 Ada Generation GPUs and NVIDIA A100 Tensor Core GPUs for training, depending on resource availability.

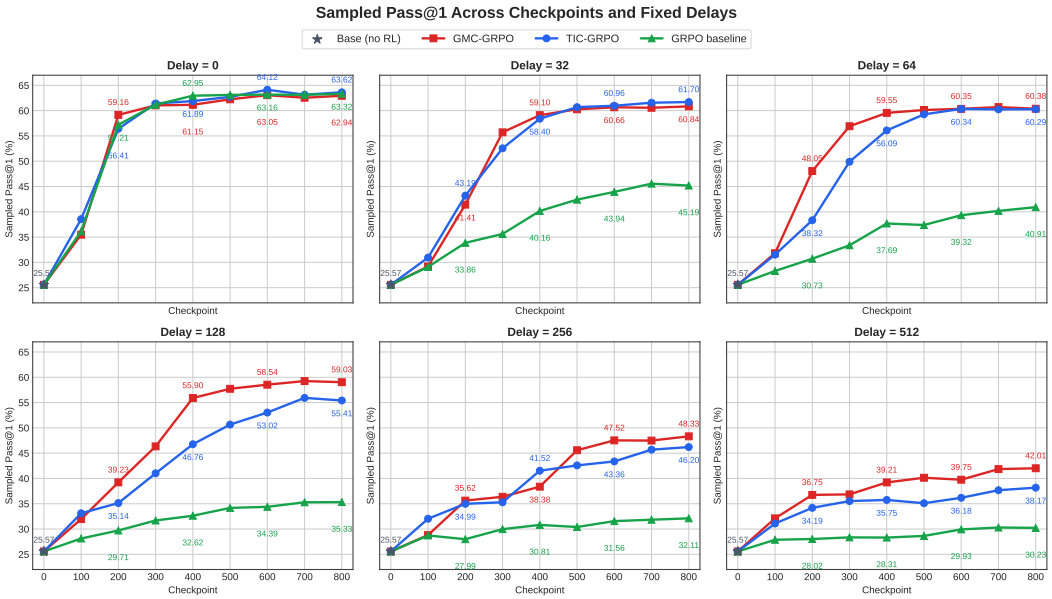


Figure 3: A full version of Figure 2 with delay=0, 32, 64, 128, 256, 512.

Controlled fixed delays. The fixed-delay study uses Qwen3-1.7B-Base and sweeps the delay over $D \in \{0, 32, 64, 128, 256, 512\}$ optimizer updates. For $D = 0$, we use `inner_steps=4`, so each rollout batch comprises four 64-trajectory optimizer mini-batches, for 256 responses in total. For every $D > 0$, we use `inner_steps=1` and maintain a FIFO queue of D 64-trajectory rollout batches. The queue is initialized with batches generated by the initial policy. In these delayed runs, at each learner iteration, the oldest batch is consumed in one optimizer update and a new batch from the current policy is appended, yielding a steady-state behavior-policy lag of D optimizer updates after the initial fill transition. All remaining optimization, rollout, and sequence-length settings match the shared configuration above.

Simulated asynchronous delays. The second suite models the coupling between response length and asynchronous rollout latency. We maintain a hidden in-flight buffer and a virtual clock τ. Launching a 64-trajectory rollout batch advances the clock by $C_{\text{gen}} = 85$. If the batch has mean completion length L, it becomes visible at

$$\tau_{\text{ready}} = \tau_{\text{launch}} + \lambda L, \qquad \lambda = 28.$$

Each optimizer update advances the clock by $C_{\text{update}} = 1$, and ready batches are exposed in completion-time order. Thus, longer batches tend to be consumed under older learner policies. A batch generated at optimizer step g and consumed at step t has realized delay $d = t - g$; its approximate steady-state mean is

$$\mathbb{E}[d] \simeq \frac{\lambda\, \mathbb{E}[L]}{C_{\text{gen}} + C_{\text{update}}} = \frac{28\, \mathbb{E}[L]}{86}.$$

Consequently, λ controls the delay distribution, while the measured mean delay—approximately 160 updates in the reported experiments—also depends on the model's response-length dynamics.

At initialization, we launch rollouts from policy 0 until the first batch is visible. If the visible queue later becomes empty, we launch the minimum number of additional batches needed to expose the next sample; both warmup and refill use a safety cap of 400 launches. For Qwen3-1.7B, rollout generation and actor updates execute sequentially on one GPU. For Qwen3-4B, one GPU serves vLLM rollouts and a second GPU performs actor optimization, with rollout generation overlapped with learner updates. This systems-level overlap does not alter the sample order or the simulated-delay schedule.

Remark 6 (**Scope of the asynchronous-sampling assumption**). The fixed-delay suite satisfies Assumption 3 exactly, since every group's delay is fixed in advance and independent of its responses.

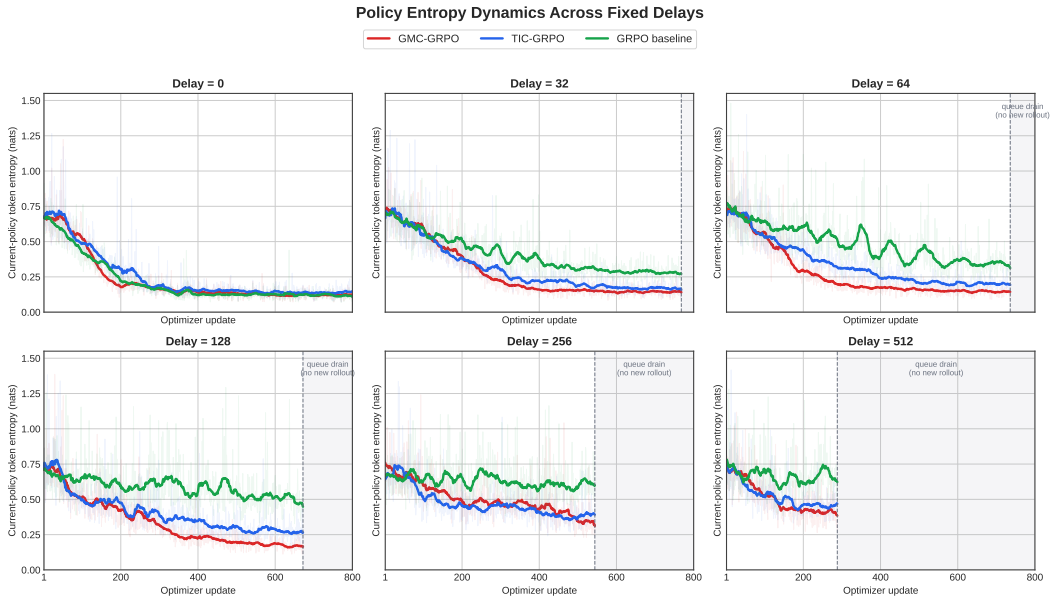


Figure 4: Current-policy token entropy across optimizer updates under fixed rollout delays. Faint curves show the per-update measurements, and solid curves show centered rolling means with a window of 25 updates. Dashed vertical lines mark queue drain, after which no new rollouts are generated. GMC-GRPO generally reaches a stable low-entropy regime earlier as the delay increases, while all methods behave similarly without delay.

The simulated suite intentionally relaxes the conditional-independence part of the assumption: because delay is determined by response length, conditioning on a long delay tilts the returned responses toward longer completions. Our theorems therefore do not formally cover this setting. The key mechanism behind Assumption 3 is nevertheless preserved. The pipeline is lossless, so response length changes only *when* a rollout is consumed, not *whether* it is consumed, and each importance ratio uses the exact behavior policy that generated the rollout. As a result, the length–delay coupling largely averages out over the course of training, unlike completion-time filtering, which systematically removes long responses. Moreover, this coupling makes the problem harder rather than easier: long responses are both more stale and more prone to extreme importance ratios. We therefore view the simulated suite as a **stress test** beyond the idealized assumption, and GMC-GRPO's robustness there suggests that its advantage does not hinge on this idealization.

Method-specific settings. Vanilla GRPO (Shao et al., 2024) uses the two-sided clipping range $\epsilon = 0.2$ and the unnormalized group-centered advantage $r_i - \overline{r}_g$. VCPO (Huang et al., 2026) uses effective-sample-size-guided step-size scaling and its closed-form off-policy minimum-variance baseline. GAC (Xu et al., 2026) uses cosine-similarity thresholds 0.05 and 0.30. TIC-GRPO (Pang et al., 2025) and GMC-GRPO both use leave-one-out advantages, $\epsilon_{\text{high}} = 0.28$. All other training, rollout, data-order, and delay parameters are held fixed across methods within each suite.

Evaluation. For the simulated asynchronous study, we evaluate the step-800 checkpoint on six mathematical-reasoning benchmarks: MATH-500 (500 problems) (Hendrycks et al., 2021; Lightman et al., 2024), GSM8K (1,319) (Cobbe et al., 2021), Gaokao2023-Math-En (385) (Liao et al., 2024), the `OE_TO_maths_en_COMP` subset of OlympiadBench (674) (He et al., 2024), Minerva Math (272) (Lewkowycz et al., 2022), and AMC23 (40) (Art of Problem Solving, 2024). Greedy accuracy uses temperature zero and one output per problem. Sampled evaluation uses temperature 1.0, top-$p = 1.0$. We draw 16 outputs per problem for MATH-500, Minerva Math, and AMC23, and eight for GSM8K, Gaokao2023-Math-En, and OlympiadBench. Each main-table entry is reported as *greedy/avg*@n, where avg@n is the mean correctness across all n sampled outputs rather than pass@any. The final column is the unweighted macro-average over the six benchmarks, computed separately for greedy accuracy and avg@n.

Table 3: Performance on MATH500 under different fixed behavior-policy delays (Qwen3-1.7B-Base, checkpoint 800).

Delay	Method	Greedy	pass@1	pass@2	pass@4	pass@8	pass@16
–	Base (no RL)	44.40	25.57	37.52	50.00	61.65	72.40
0	GMC-GRPO	66.40	62.94	72.39	79.25	84.64	89.00
	TIC-GRPO	66.80	63.62	72.59	79.47	84.90	89.00
	GRPO	66.20	63.32	72.41	79.25	84.84	88.80
32	GMC-GRPO	64.80	60.84	70.47	77.85	83.89	88.60
	TIC-GRPO	66.40	61.70	71.09	78.38	84.24	88.80
	GRPO	56.20	45.19	58.01	69.04	77.62	83.60
64	GMC-GRPO	65.20	60.38	69.97	77.30	83.30	88.00
	TIC-GRPO	65.40	60.29	70.21	77.42	82.92	86.80
	GRPO	52.00	40.91	53.93	65.43	74.47	81.20
128	GMC-GRPO	68.40	59.03	69.57	77.34	83.45	88.00
	TIC-GRPO	63.20	55.41	66.86	75.64	82.59	87.80
	GRPO	51.40	35.33	47.89	59.87	70.46	78.60
256	GMC-GRPO	59.60	48.33	61.64	72.02	79.85	85.40
	TIC-GRPO	56.00	46.20	59.78	71.12	80.05	86.20
	GRPO	47.60	32.11	44.78	57.10	68.17	77.60
512	GMC-GRPO	56.40	42.01	55.97	67.78	76.91	83.40
	TIC-GRPO	53.80	38.17	51.07	62.59	72.47	80.60
	GRPO	46.80	30.23	42.73	55.03	66.16	76.00

Greedy uses temperature 0. pass@k uses the unbiased estimator from $n = 16$ samples per question (temperature 1.0, top-p 1.0), averaged over 500 questions.

The fixed-delay study evaluates GRPO, TIC-GRPO, and GMC-GRPO on MATH-500 at checkpoint 800. We use greedy decoding and 16 sampled responses per problem to compute the unbiased pass@k estimators for $k \in \{1, 2, 4, 8, 16\}$. Both suites use the same 1,024-token prompt limit, 4,096-token generation limit, two-shot prompt, and last-complete-`\boxed{}` answer scorer as in training.

B.2 Supplemental Experiment Results

B.2.1 Additional experiment data and training dynamics

Our theoretical results concern optimization convergence rather than only final task accuracy. To connect these guarantees to observable training dynamics, Figure 4 reports current-policy token entropy, which is the most direct diagnostic among our logged statistics of how quickly the policy distribution stabilizes. Without delay, all three methods exhibit similar entropy decay, consistent with their nearly identical Pass@1 trajectories in Figure 2. As the delay increases, GMC-GRPO typically enters a stable low-entropy regime earlier than TIC-GRPO and vanilla GRPO. Although entropy is not the formal stationarity measure used in our theory, its faster decay and earlier stabilization, together with the Pass@1 curves, provide empirical evidence consistent with the predicted faster convergence of GMC-GRPO under stale rollouts.

Table 3 complements Figure 2 by reporting the complete checkpoint-800 results, including greedy accuracy and pass@k for $k \in \{1, 2, 4, 8, 16\}$. The three methods remain nearly indistinguishable at zero delay. Under stale rollouts, GMC-GRPO consistently improves over vanilla GRPO and is generally competitive with or stronger than TIC-GRPO, with the clearest gains at delays 128 and 512. Thus, the trend in the main-text Pass@1 curves extends to greedy decoding and a broader range of sampling budgets.

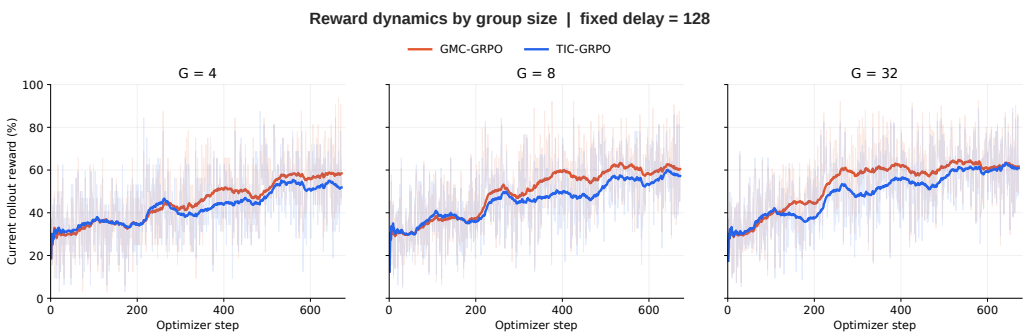


Figure 5: Training-reward dynamics for GMC-GRPO and TIC-GRPO with group sizes $G \in \{4, 8, 32\}$ under the fixed-delay setting with delay 128. All other experimental settings are unchanged. Faint curves show individual measurements, and solid curves show smoothed trends. GMC-GRPO establishes a reward advantage earlier as the group size increases, while the gap narrows toward the end of training, particularly for $G = 32$.

B.2.2 GMC-GRPO BENEFITS FROM LARGER GROUP SIZES

Corollary 2 shows that larger groups reduce the delay-dependent term in the GMC-GRPO convergence bound. To examine this benefit empirically, we vary the group size over $G \in \{4, 8, 32\}$ while keeping all other experimental settings identical to the fixed-delay experiment with delay 128. Figure 5 shows that GMC-GRPO establishes a reward advantage over TIC-GRPO progressively earlier as the group size increases: the curves begin to separate around step 300 for $G = 4$, around step 200 for $G = 8$, and around step 100 for $G = 32$. This earlier separation indicates that GMC-GRPO benefits more promptly from larger groups, providing empirical evidence consistent with the faster convergence predicted by Corollary 2.

The training-reward gap also becomes smaller at larger group sizes toward the end of training, with the two curves nearly meeting for $G = 32$. This pattern is consistent with both methods approaching a reward plateau: GMC-GRPO makes faster progress early on, while TIC-GRPO catches up as training continues. Thus, the benefit of larger groups is most apparent in the learning dynamics, even when the final training rewards are similar.

Table 4 complements these training curves with MATH500 evaluations at checkpoints 200, 300, 400, and 800. At checkpoint 200, GMC-GRPO's sampled Pass@1 advantage over TIC-GRPO increases from -1.11 percentage points at $G = 4$ to 3.79 at $G = 8$ and 7.10 at $G = 32$, showing that the earlier training gains also translate into better evaluation performance. At the final checkpoint, GMC-GRPO retains higher sampled Pass@1 at all three group sizes and matches or exceeds TIC-GRPO in greedy accuracy. In particular, at $G = 32$, both methods achieve 66.00% greedy accuracy, but GMC-GRPO still achieves higher sampled Pass@1 (61.18% versus 59.80%).

B.2.3 COMPARISON AGAINST A-3PO.

In this section, we present an independent comparison between GMC-GRPO and A-3PO (Li et al., 2025), which achieves the best average performance on Qwen3-1.7B but suffers training collapse on Qwen3-8B in our experiments. We therefore report this comparison separately and discuss the observed instability in detail.

To make use of stale rollouts without overly restricting policy updates, A-3PO follows decoupled policy optimization (Fu et al., 2026): it uses a proximal anchor for clipping while retaining importance correction relative to the behavior policy. Suppressing token arguments $(y_t \mid x, y_{<t})$, the decoupled surrogate is

$$L^{\mathrm{CLIP}}_{\mathrm{decoupled}}(\theta) = \mathbb{E}\left[\frac{\pi_{\mathrm{prox}}}{\pi_{\theta_{\mathrm{behav}}}} \min\left\{\frac{\pi_\theta}{\pi_{\mathrm{prox}}}\widehat{A},\ \mathrm{clip}\left(\frac{\pi_\theta}{\pi_{\mathrm{prox}}}, 1-\epsilon, 1+\epsilon\right)\widehat{A}\right\}\right], \tag{10}$$

Table 4: MATH500 evaluation of Qwen3-1.7B-Base at checkpoints 200, 300, 400, and 800 under the fixed-delay setting with delay 128. All settings except group size match the fixed-delay experiment. Each method entry reports greedy accuracy / sampled Pass@1 (%), with sampled Pass@1 computed as mean correctness over 16 responses per problem (avg@16). Δ denotes the sampled Pass@1 of GMC-GRPO minus that of TIC-GRPO, measured in percentage points (pp).

Checkpoint	G	GMC-GRPO	TIC-GRPO	Δ Pass@1 (pp)
200	4	51.40 / 33.70	51.60 / 34.81	-1.11
200	8	55.40 / 39.23	50.20 / 35.44	$+3.79$
200	32	62.80 / 46.51	53.60 / 39.41	$+7.10$
300	4	57.00 / 39.02	54.80 / 36.05	$+2.97$
300	8	61.40 / 46.34	55.20 / 41.01	$+5.33$
300	32	65.40 / 56.34	56.00 / 42.10	$+14.24$
400	4	62.60 / 48.52	58.60 / 42.29	$+6.23$
400	8	67.00 / 55.90	60.00 / 46.76	$+9.14$
400	32	64.00 / 58.55	62.80 / 51.41	$+7.14$
800	4	67.00 / 56.17	63.00 / 53.91	$+2.26$
800	8	68.40 / 59.03	63.20 / 55.41	$+3.62$
800	32	66.00 / 61.18	66.00 / 59.80	$+1.38$

Table 5: Comparison of A-3PO and GMC-GRPO under simulated asynchronous delays. Each benchmark cell is greedy/avg@n, with $n = 16$ for MATH500, Minerva, and AMC23, and $n = 8$ for GSM8K, Gaokao, and OlympiadBench. *Avg. perf.* is the average over the six benchmarks, reported as mean greedy/mean avg@n. For Qwen3-1.7B and Qwen3-4B, the better result for each metric is bolded. A-3PO exhibited training collapse in our Qwen3-8B experiment.

Method	**Avg. Delay**	**MATH500**	**GSM8K**	**Gaokao**	**Olympiad**	**Minerva**	**AMC23**	**Avg. perf.**
				Qwen3-1.7B				
A-3PO	160.80	**67.40 / 61.26**	**81.88 / 76.42**	53.77 / **51.69**	**30.71 / 25.56**	**28.68 / 24.20**	32.50 / **32.50**	**49.16 / 45.27**
GMC-GRPO	159.75	66.00 / 57.29	81.65 / 72.66	**55.32** / 47.73	28.19 / 22.61	26.47 / 20.11	**35.00** / 28.44	48.77 / 41.47
				Qwen3-4B				
A-3PO	138.34	70.80 / 69.16	90.14 / **88.55**	59.48 / 58.38	36.80 / 35.16	31.62 / 29.78	50.00 / 46.88	56.47 / 54.65
GMC-GRPO	141.01	**77.60 / 72.54**	**91.13** / 87.03	**65.19 / 60.91**	**43.18 / 37.72**	**34.19 / 30.10**	**52.50 / 48.91**	**60.63 / 56.20**
				Qwen3-8B				
A-3PO	-			*Training collapsed*				
GMC-GRPO	141.1	78.80 / 75.22	93.33 / 91.37	69.09 / 63.64	44.96 / 40.34	40.07 / 34.35	60.00 / 52.81	64.37 / 59.62

where the expectation is over sampled rollout tokens and $\widehat{A}$ is the group-based advantage. A-3PO approximates the proximal anchor by

$$\log \pi_{\text{prox}} = \alpha \log \pi_{\theta_{\text{behav}}} + (1-\alpha) \log \pi_\theta, \qquad \alpha = \frac{1}{\max\{1, \tau\}},$$

This design is reasonable when delays are small: the policy shift is limited, and retaining more learning signals from stale rollouts can accelerate learning. Under large delays, however, α approaches zero, so the ratio used for clipping can remain close to one and the update can resemble *unclipped GRPO*. Since the outer importance factor $\pi_{\text{prox}}/\pi_{\theta_{\text{behav}}}$ remains unclipped, large importance weights can destabilize updates and even lead to training collapse in extreme cases.

This is also demonstrated in our experiments: Table 5 shows that A-3PO outperforms GMC-GRPO on Qwen3-1.7B in mean greedy/mean avg@n accuracy (49.16/45.27 versus 48.77/41.47). This advantage is consistent with prior evidence that hard clipping can discard useful gradients and restrict policy improvement (Dwyer et al., 2025; Zheng et al., 2026). Retaining more of these signals can therefore improve performance when extreme importance ratios are rare.

The Qwen3-8B experiment illustrates the accompanying stability challenge. In Figure 6, A-3PO initially learns faster but its training reward collapses around step 430, whereas GMC-GRPO continues to improve and reaches mean greedy/mean avg@n accuracy of 64.37/59.62. This trajectory

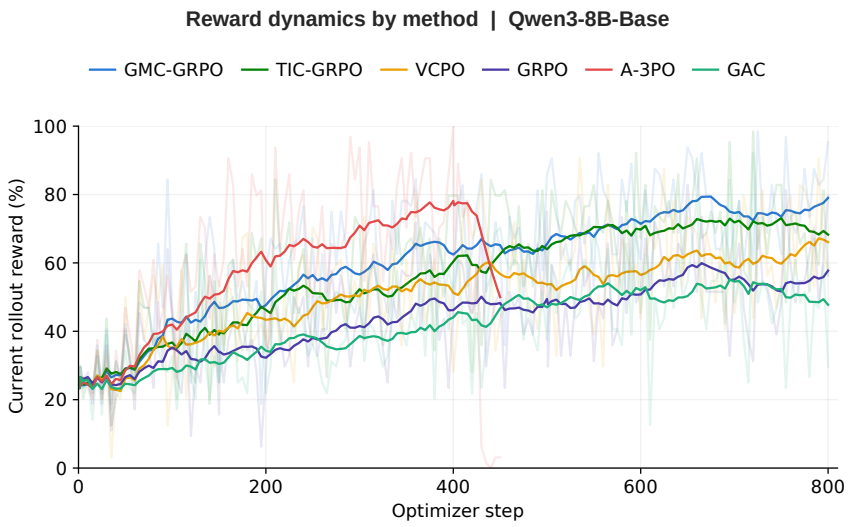


Figure 6: Training reward on Qwen3-8B under simulated asynchronous delays. Faint curves show individual measurements, and solid curves show smoothed trends. A-3PO initially improves rapidly, but its recorded reward falls to near zero around steps 430–440. GMC-GRPO continues to improve over the remaining training steps.

resembles the prosperity-before-collapse behavior reported by Zheng et al. (2026), where removing the trust region exposes useful stale-data signals but also permits destabilizing updates. The comparison highlights the motivation for GMC: preserving learning signals must be coupled with control of update magnitude. GMC caps the average importance mass within each group and preserves the relative contributions of its responses, controlling the update second moment under the assumptions of Proposition 1.

C Mixed-Delay Setting, GMC-GRPO and TIC-GRPO Algorithms

TIC-GRPO was originally proposed for synchronous training (Pang et al., 2025), so Algorithm 1 first restates this original synchronous form for reference. Algorithm 2 then gives the asynchronous versions of GMC-GRPO and TIC-GRPO analyzed in this paper. The two asynchronous methods share the same rollout and learner schedule and differ only in their importance correction. Each completed group stores its prompt, responses, rewards, behavior-policy version s, and behavior log-probabilities $\ell^{(i)} = \log \pi_{\theta_s}(y^{(i)} \mid x)$. For a group b consumed at iteration n, both algorithms compute

$$\rho_n^{(b,i)} = \exp\Big(\log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)}) - \ell^{(b,i)}\Big).$$

For GMC-GRPO, the learner forms g_n^{GMC} in eq. (9) using the group-rescaled ratios $\widehat{\rho}_n^{(b,i)}$. For TIC-GRPO, it instead uses the individually truncated ratios $\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}$ to form g_n^{TIC}, as defined in eqs. (34) and (35). In either case, the learner takes the ascent step $\theta_{n+1} = \theta_n + \eta g_n$. Both methods use the leave-one-out advantages defined in Section 2.1. Behavior log-probabilities and advantages are held fixed during each update.

In the synchronous schedule, K mini-batches share one behavior-policy snapshot: rollout collection finishes before optimization begins, and the next global batch is generated only after these updates finish. Thus, synchronous training can still use stale data during successive mini-batch updates. In the asynchronous schedule, rollout workers and the learner run concurrently. The learner needs only B completed groups to update; it does not wait for all workers or require a common behavior-policy version.

We use the asynchronous sampling model and notation introduced in Section 2.2. In particular, $\tau_n^{(b)}$ is the delay of group b, $\rho_n^{(b,i)}$ is the delayed importance ratio in eq. (3), and $\widehat{A}^{(b,i)}$ is the leave-one-out

Algorithm 1 TIC-GRPO with synchronous rollouts

Require: initial parameter θ_0; step size η; total updates N; group size $G \geq 2$; mini-batch size B; mini-batches per global batch $K \geq 1$; clipping threshold $1+\epsilon$; prompt distribution $\mathcal{D}$; reward r_ϕ.

$n \leftarrow 0$.
while $n < N$ **do**
  freeze the behavior-policy snapshot $\theta_s \leftarrow \theta_n$ with $s \leftarrow n$; set $K_n \leftarrow \min\{K, N-n\}$.
  draw $K_n B$ prompts from $\mathcal{D}$ and generate G responses per prompt under the fixed policy π_{θ_s}.
  wait until all $K_n B$ groups are generated and scored by r_ϕ; store their behavior log-probabilities and version s.
  partition the completed global batch into K_n mini-batches of B whole groups each.
  for each mini-batch in the partition **do**
    denote its groups by $b \in [B]$ and set $s_n^{(b)} \leftarrow s$, $\tau_n^{(b)} \leftarrow n - s$.
    form $\widehat{A}^{(b,i)}$ from the rewards of all G responses in each group.
    compute $\rho_n^{(b,i)}$ from the current policy and stored behavior log-probabilities for all $b \in [B]$, $i \in [G]$.
    form g_n^{TIC} using eq. (35) and update $\theta_{n+1} \leftarrow \theta_n + \eta g_n^{\mathrm{TIC}}$.
    $n \leftarrow n+1$.
  end for
end while

advantage defined in Section 2.1. Let $\mathcal{M}_n$ denote the mini-batch of B completed groups returned at learner iteration n. To aggregate the mixed delays over the training horizon, for any integer $k \geq 1$, define the averaged kth delay moment

$$\overline{\tau}_k := \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left[\frac{1}{B} \sum_{b=1}^{B} \left(\tau_n^{(b)}\right)^k \right]. \tag{11}$$

Thus, $\overline{\tau}_2$ and $\overline{\tau}_4$ denote the averaged second and fourth delay moments, respectively. The second moment controls intermediate parameter-drift terms, while the fourth moment appears in the final convergence bounds after the delay-induced bias is squared.

For the appendix proofs, we additionally write the average importance ratio in group b as $\overline{\rho}_n^{(b)} := G^{-1} \sum_{i=1}^{G} \rho_n^{(b,i)}$. Softmax policies have full support, so every ratio is finite and $\overline{\rho}_n^{(b)} > 0$ almost surely. Using the common scaling factor from Proposition 2, the GMC effective ratio admits the equivalent form

$$\widehat{\rho}_n^{(b,i)} := \frac{\rho_n^{(b,i)}}{\max\{1, \overline{\rho}_n^{(b)}/(1+\epsilon)\}}. \tag{12}$$

This form is used in the groupwise decomposition and moment bounds below.

The GMC branch of Algorithm 2 applies this groupwise rescaling within the shared asynchronous training loop. The behavior log-probabilities stored with each completed group are sufficient to compute every delayed importance ratio at consumption time.

Remark 7 (Autodifferentiation implementation). Equation (9) defines the update directly. In a surrogate-loss implementation, $\max\{1, \overline{\rho}_n^{(b)}/(1+\epsilon)\}$ and $\widehat{A}^{(b,i)}$ must be treated as constants using stop-gradient. The numerator ratio supplies the policy gradient. Differentiating through the common group denominator defines a different estimator and is not covered by the theorems below.

Proof-specific conditioning and divergence notation. The filtration pair $\mathcal{F}_n^- \subseteq \mathcal{H}_n$ is defined in Section 2.2, and Assumption 3 specifies the corresponding conditional sampling properties. For each returned group, define the current-to-behavior divergences

$$D_n^{(b)} := D_{\mathrm{KL}}\Big(\pi_{\theta_n}(\cdot \mid x^{(b)}) \,\|\, \pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})\Big),$$
$$\chi_n^{2,(b)} := \chi^2\Big(\pi_{\theta_n}(\cdot \mid x^{(b)}) \,\|\, \pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})\Big).$$

Algorithm 2 TIC-GRPO and GMC-GRPO with asynchronous rollouts

Require: method $\mathsf{method} \in \{\mathrm{TIC}, \mathrm{GMC}\}$; initial parameter θ_0; step size η; total updates N; group size $G \geq 2$; mini-batch size B; threshold $1+\epsilon$; rollout workers over $\mathcal{D}$ with reward r_ϕ; shared buffer $\mathcal{Q}$.

initialize $\mathcal{Q} \leftarrow \varnothing$ and publish the snapshot $(0, \theta_0)$.
Run the following worker and learner loops concurrently.
Each rollout worker:
while training is active **do**
  read a published snapshot (s, θ_s) and draw $x \sim \mathcal{D}$.
  generate $y^{(1)}, \ldots, y^{(G)} \sim \pi_{\theta_s}(\cdot \mid x)$, keeping θ_s fixed throughout this group.
  compute $r^{(i)} \leftarrow r_\phi(x, y^{(i)})$ and store $\ell^{(i)} \leftarrow \log \pi_{\theta_s}(y^{(i)} \mid x)$ for all $i \in [G]$.
  append $(x, \{y^{(i)}, r^{(i)}, \ell^{(i)}\}_{i=1}^{G}, s)$ to $\mathcal{Q}$ as one completed group.
end while
Learner:
for $n = 0, 1, \ldots, N-1$ **do**
  once $|\mathcal{Q}| \geq B$, pop B completed groups $\mathcal{M}_n$; their behavior-policy versions need not match.
  denote each group's stored version by $s_n^{(b)}$ and set $\tau_n^{(b)} \leftarrow n - s_n^{(b)}$.
  form $\widehat{A}^{(b,i)}$ from the rewards of all G responses in each group.
  compute $\rho_n^{(b,i)}$ from the current policy and stored behavior log-probabilities for all $b \in [B]$, $i \in [G]$.
  if $\mathsf{method} = \mathrm{TIC}$ **then**
    form $g_n \leftarrow g_n^{\mathrm{TIC}}$ using eq. (35).
  else ▷ GMC-GRPO
    for each group $b \in [B]$ **do**
      compute $\overline{\rho}_n^{(b)} \leftarrow G^{-1} \sum_{i=1}^{G} \rho_n^{(b,i)}$ using log-sum-exp.
      set $\widehat{\rho}_n^{(b,i)} \leftarrow \rho_n^{(b,i)} / \max\{1, \overline{\rho}_n^{(b)}/(1+\epsilon)\}$ for all $i \in [G]$.
    end for
    form $g_n \leftarrow g_n^{\mathrm{GMC}}$ using eq. (9).
  end if
  update $\theta_{n+1} \leftarrow \theta_n + \eta g_n$.
  publish $(n+1, \theta_{n+1})$ without waiting for ongoing rollouts; in-progress groups retain their launch snapshots.
end for
signal the rollout workers to stop.

Both divergences, as well as $J_{x^{(b)}}(\theta_n)$, are $\mathcal{H}_n$-measurable and hence fixed inside conditional expectations given $\mathcal{H}_n$. Because $\mathcal{F}_n^- \subseteq \mathcal{H}_n$, every integrable random variable Z satisfies

$$\mathbb{E}[Z \mid \mathcal{F}_n^-] = \mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{H}_n] \mid \mathcal{F}_n^-\big].$$

The proofs first condition on $\mathcal{H}_n$ to hold the prompt, delay, and behavior policy fixed, and then average back to $\mathcal{F}_n^-$ to recover a statement about the learner iterate.

With the algorithm and conditioning convention fixed, the next section collects the analytic tools shared by the GMC-GRPO and TIC-GRPO proofs.

D TECHNICAL PRELIMINARIES

The convergence proofs for both algorithms rely on a common set of analytic tools. Starting from Assumption 1, we establish trajectory-score bounds, control trajectory KL divergence by the squared parameter distance, and prove smoothness of the reward objective. We then use this smoothness bound to prove Lemma 1, which reduces convergence analysis to controlling the update second moment and the squared conditional bias. The following sections supply these controls for GMC-GRPO and TIC-GRPO.

Lemma 2 (Trajectory score bounds)**.** *Under Assumption 1, for every prompt x, parameter θ, and length-T response y,*

$$\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \|\nabla \log \pi_\theta(y \mid x)\|_2^2 \le 2LT \log|\mathcal{A}|, \tag{13}$$

$$\|\nabla \log \pi_\theta(y \mid x)\|_2^2 \le -2LT \log \pi_\theta(y \mid x). \tag{14}$$

If Assumption 2 also holds, then

$$\|\nabla \log \pi_\theta(y \mid x)\|_2 \le T\sqrt{2L \log \mathcal{C}}. \tag{15}$$

Proof. Fix a token position t and its prefix $(x, y_{<t})$, and write $f_t(\theta) := -\log \pi_\theta(y_t \mid x, y_{<t})$. Assumption 1 states that ∇f_t is L-Lipschitz. The descent lemma therefore gives, for every vector v,

$$f_t(\theta + v) \le f_t(\theta) + \langle \nabla f_t(\theta),\, v\rangle + \frac{L}{2}\|v\|_2^2 .$$

Choose $v = -L^{-1}\nabla f_t(\theta) = L^{-1}\nabla \log \pi_\theta(y_t \mid x, y_{<t})$. Since f_t is nonnegative,

$$\begin{aligned} 0 &\le -\log \pi_{\theta + L^{-1}\nabla \log \pi_\theta(y_t|x,y_{<t})}(y_t \mid x, y_{<t}) \\ &\le -\log \pi_\theta(y_t \mid x, y_{<t}) - \frac{1}{2L}\|\nabla \log \pi_\theta(y_t \mid x, y_{<t})\|_2^2 , \end{aligned}$$

Rearranging the last inequality shows that each token score satisfies

$$\|\nabla \log \pi_\theta(y_t \mid x, y_{<t})\|_2^2 \le -2L \log \pi_\theta(y_t \mid x, y_{<t}). \tag{16}$$

Now let $s_t := \nabla \log \pi_\theta(y_t \mid x, y_{<t})$. The autoregressive factorization gives $\nabla \log \pi_\theta(y \mid x) = \sum_{t=1}^T s_t$, so Cauchy–Schwarz and eq. (16) yield

$$\|\nabla \log \pi_\theta(y \mid x)\|_2^2 \le T\sum_{t=1}^T \|s_t\|_2^2 \le -2LT\sum_{t=1}^T \log \pi_\theta(y_t \mid x, y_{<t}) = -2LT \log \pi_\theta(y \mid x),$$

which proves eq. (14).

For the expected bound, first expand the conditional score identity:

$$\mathbb{E}[s_t \mid x, y_{<t}] = \sum_{a\in\mathcal{A}} \pi_\theta(a \mid x, y_{<t})\nabla \log \pi_\theta(a \mid x, y_{<t}) = \sum_{a\in\mathcal{A}} \nabla \pi_\theta(a \mid x, y_{<t}) = \nabla 1 = 0.$$

If $s < t$, then s_s is determined by $(x, y_{<t})$. The tower property therefore gives $\mathbb{E}\langle s_s,\, s_t\rangle = \mathbb{E}\langle s_s,\, \mathbb{E}[s_t \mid x, y_{<t}]\rangle = 0$; hence distinct token scores are orthogonal in expectation. Expanding the square and applying eq. (16),

$$\begin{aligned} \mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \|\nabla \log \pi_\theta(y \mid x)\|_2^2 &= \mathbb{E}\left\|\sum_{t=1}^T s_t\right\|_2^2 = \sum_{t=1}^T \mathbb{E}\,\|s_t\|_2^2 \\ &\le 2L\sum_{t=1}^T \mathbb{E}[-\log \pi_\theta(y_t \mid x, y_{<t})] \le 2LT \log|\mathcal{A}|, \end{aligned}$$

where the last inequality uses

$$\mathbb{E}[-\log \pi_\theta(y_t \mid x, y_{<t})] = \mathbb{E}_{y_{<t}}\big[H\big(\pi_\theta(\cdot \mid x, y_{<t})\big)\big] \le \log|\mathcal{A}|.$$

This proves eq. (13). Finally, Assumption 2 implies $-\log \pi_\theta(y_t \mid x, y_{<t}) \le \log \mathcal{C}$. Combining this with eq. (16) and the triangle inequality,

$$\|\nabla \log \pi_\theta(y \mid x)\|_2 \le \sum_{t=1}^T \|s_t\|_2 \le T\sqrt{2L \log \mathcal{C}},$$

which is eq. (15). □

Lemma 3 (Trajectory KL is quadratically controlled)**.** *Under Assumption 1, for every prompt x and all θ, θ',*

$$D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\theta'}(\cdot \mid x)\big) \le \frac{TL}{2}\|\theta - \theta'\|_2^2 . \tag{17}$$

Proof. For a fixed prefix $h = (x, y_{<t})$ and token a, the L-Lipschitz score implies the two-sided smoothness bound

$$\log \pi_\theta(a \mid h) - \log \pi_{\theta'}(a \mid h) \le \langle \nabla \log \pi_\theta(a \mid h),\, \theta - \theta' \rangle + \frac{L}{2} \|\theta - \theta'\|_2^2 .$$

The autoregressive KL chain rule sums this inequality over token positions:

$$\begin{aligned} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\theta'}(\cdot \mid x)\big) &= \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \sum_{t=1}^{T} \log \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\theta'}(y_t \mid x, y_{<t})} \\ &\le \sum_{t=1}^{T} \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[\langle \nabla \log \pi_\theta(y_t \mid x, y_{<t}),\, \theta - \theta' \rangle + \frac{L}{2} \|\theta - \theta'\|_2^2 \right] \\ &= \frac{TL}{2} \|\theta - \theta'\|_2^2 . \end{aligned}$$

For the last equality, condition on $(x, y_{<t})$ in each linear term. The score identity proved above gives

$$\mathbb{E}[\langle \nabla \log \pi_\theta(y_t \mid x, y_{<t}),\, \theta - \theta' \rangle \,|\, x, y_{<t}] = \left\langle \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a \mid x, y_{<t}),\, \theta - \theta' \right\rangle = 0.$$

Thus only the $L \|\theta - \theta'\|_2^2 / 2$ remainder remains at each of the T positions, proving eq. (17). □

Lemma 4 (Smoothness of the reward objective)**.** *Under Assumption 1, the objective has $TRL(2 \log |\mathcal{A}| + 1)$-Lipschitz gradient. Consequently, any update $\theta_{n+1} = \theta_n + \eta g_n$ satisfies*

$$J(\theta_{n+1}) \ge J(\theta_n) + \eta \,\langle \nabla J(\theta_n),\, g_n \rangle - \frac{TRL(2 \log |\mathcal{A}| + 1)\eta^2}{2} \|g_n\|_2^2 . \tag{18}$$

Proof. Fix a prompt–response pair (x, y) and write $S_\theta(y \mid x) := \nabla \log \pi_\theta(y \mid x)$. The autoregressive decomposition and Assumption 1 imply

$$\|S_\theta(y \mid x) - S_{\theta'}(y \mid x)\|_2 \le TL \|\theta - \theta'\|_2 .$$

Let $\Delta := \theta - \theta'$ and $\gamma(s) := \theta' + s\Delta$. The restriction $s \mapsto S_{\gamma(s)}(y \mid x)$ is $TL \|\Delta\|_2$-Lipschitz, hence absolutely continuous and differentiable for almost every s, with

$$\left\| \frac{\mathrm{d}}{\mathrm{d}s} S_{\gamma(s)}(y \mid x) \right\|_2 \le TL \|\Delta\|_2 .$$

Moreover, $\nabla \pi_\theta(y \mid x) = \pi_\theta(y \mid x) S_\theta(y \mid x)$ and $\frac{\mathrm{d}}{\mathrm{d}s} \pi_{\gamma(s)}(y \mid x) = \pi_{\gamma(s)}(y \mid x) \left\langle S_{\gamma(s)}(y \mid x),\, \Delta \right\rangle$. Both factors are bounded on the compact segment, so their product is absolutely continuous. At almost every s, the one-dimensional product rule gives

$$\frac{\mathrm{d}}{\mathrm{d}s} \nabla \pi_{\gamma(s)}(y \mid x) = \pi_{\gamma(s)}(y \mid x) \left\{ \left\langle S_{\gamma(s)}(y \mid x),\, \Delta \right\rangle S_{\gamma(s)}(y \mid x) + \frac{\mathrm{d}}{\mathrm{d}s} S_{\gamma(s)}(y \mid x) \right\}.$$

Lemma 2 now allows us to sum the derivative bound over the finite trajectory space:

$$\begin{aligned} \sum_y \left\| \frac{\mathrm{d}}{\mathrm{d}s} \nabla \pi_{\gamma(s)}(y \mid x) \right\|_2 &\le \|\Delta\|_2 \, \mathbb{E}_{y \sim \pi_{\gamma(s)}(\cdot|x)} \left[\left\| S_{\gamma(s)}(y \mid x) \right\|_2^2 + TL \right] \\ &\le TL(2 \log |\mathcal{A}| + 1) \|\Delta\|_2 . \end{aligned}$$

The fundamental theorem of calculus for absolutely continuous functions now yields the promptwise estimate

$$\begin{aligned} \sum_y \|\nabla \pi_\theta(y \mid x) - \nabla \pi_{\theta'}(y \mid x)\|_2 &\le \int_0^1 \sum_y \left\| \frac{\mathrm{d}}{\mathrm{d}s} \nabla \pi_{\gamma(s)}(y \mid x) \right\|_2 \mathrm{d}s \\ &\le TL(2 \log |\mathcal{A}| + 1) \|\theta - \theta'\|_2 . \end{aligned}$$

It remains to pass from a fixed prompt to the objective. First,

$$\sum_y \|\nabla \pi_\theta(y \mid x)\|_2 = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \|S_\theta(y \mid x)\|_2 \le \sqrt{2LT \log|\mathcal{A}|},$$

where Cauchy–Schwarz and eq. (13) give the last inequality. This uniform bound, together with $|r_\phi| \le R$, dominates the first difference quotients of J. Therefore $\nabla J(\theta) = \mathbb{E}_x \sum_y r_\phi(x,y)\nabla\pi_\theta(y \mid x)$, and the promptwise estimate above gives

$$\begin{aligned}\|\nabla J(\theta) - \nabla J(\theta')\|_2 &\le R\mathbb{E}_x \sum_y \|\nabla\pi_\theta(y \mid x) - \nabla\pi_{\theta'}(y \mid x)\|_2 \\ &\le TRL(2\log|\mathcal{A}|+1)\,\|\theta-\theta'\|_2 .\end{aligned}$$

To obtain the claimed one-step inequality, integrate once more along $\theta_n + s\eta g_n$:

$$\begin{aligned}J(\theta_{n+1}) - J(\theta_n) - \eta\,\langle \nabla J(\theta_n),\, g_n\rangle &= \eta \int_0^1 \langle \nabla J(\theta_n + s\eta g_n) - \nabla J(\theta_n),\, g_n\rangle \,\mathrm{d}s \\ &\ge -TRL(2\log|\mathcal{A}|+1)\eta^2 \|g_n\|_2^2 \int_0^1 s\,\mathrm{d}s \\ &= -\frac{TRL(2\log|\mathcal{A}|+1)\eta^2}{2}\|g_n\|_2^2 .\end{aligned}$$

Rearranging proves eq. (18). □

Proof of Lemma 1. Conditioning eq. (18) on $\mathcal{F}_n^-$, substituting $\mathbb{E}[g_n \mid \mathcal{F}_n^-] = T^{-1}\nabla J(\theta_n) + b_n$, and using that θ_n is $\mathcal{F}_n^-$-measurable gives the first line below. For the second line, apply Young's inequality in the explicit form

$$\langle u,\, v\rangle \ge -\frac{1}{2T}\|u\|_2^2 - \frac{T}{2}\|v\|_2^2 \qquad (u, v \in \mathbb{R}^d)$$

with $u = \nabla J(\theta_n)$ and $v = b_n$:

$$\begin{aligned}\mathbb{E}[J(\theta_{n+1}) \mid \mathcal{F}_n^-] - J(\theta_n) &\ge \frac{\eta}{T}\|\nabla J(\theta_n)\|_2^2 + \eta\,\langle \nabla J(\theta_n),\, b_n\rangle - \frac{TRL(2\log|\mathcal{A}|+1)\eta^2}{2}\mathbb{E}[\|g_n\|_2^2 \mid \mathcal{F}_n^-] \\ &\ge \frac{\eta}{2T}\|\nabla J(\theta_n)\|_2^2 - \frac{\eta T}{2}\|b_n\|_2^2 - \frac{TRL(2\log|\mathcal{A}|+1)\eta^2}{2}\mathbb{E}[\|g_n\|_2^2 \mid \mathcal{F}_n^-].\end{aligned}$$

After taking expectations, move the objective increment to the right and sum from $n = 0$ to $N-1$. The increments telescope according to

$$\sum_{n=0}^{N-1} \mathbb{E}[J(\theta_{n+1}) - J(\theta_n)] = \mathbb{E}[J(\theta_N)] - \mathbb{E}[J(\theta_0)] \le J^\star - \mathbb{E}[J(\theta_0)].$$

Consequently,

$$\frac{\eta}{2T}\sum_{n=0}^{N-1}\mathbb{E}\,\|\nabla J(\theta_n)\|_2^2 \le J^\star - \mathbb{E}[J(\theta_0)] + \frac{\eta T}{2}\sum_{n=0}^{N-1}\mathbb{E}\,\|b_n\|_2^2 + \frac{TRL(2\log|\mathcal{A}|+1)\eta^2}{2}\sum_{n=0}^{N-1}\mathbb{E}\,\|g_n\|_2^2 .$$

Dividing by N and multiplying by $2T/\eta$ gives exactly Lemma 1. □

E Optimization over a Second-Moment-Controlled Weighting Class

This section justifies the weighting rule underlying GMC-GRPO. We first prove that the admissible mass constraint gives a uniform second-moment bound and, under coverage, bounds the correction bias by the removed importance mass. We then show that GMC uniquely maximizes retained mass among admissible direction-preserving corrections. Comparing the mass removed by GMC and TIC establishes the corresponding ordering of their certified bias envelopes, providing the basis for the convergence analysis that follows.

Proof of Proposition 1. For brevity, denote the contribution of group b to eq. (4) by

$$U^w_{n,b} := \frac{1}{TG}\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}, \qquad g^w_n = \frac{1}{B}\sum_{b=1}^{B} U^w_{n,b}.$$

For one group, weighted Cauchy–Schwarz and eq. (5) give the pathwise bound

$$\begin{aligned}\left\|U^w_{n,b}\right\|_2^2 &\le \left(\frac{1}{G}\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)}\right)\times\left(\frac{1}{G}\sum_{i=1}^{G} w_n^{(b,i)}\rho_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2\right)\\ &\le \frac{1+\epsilon}{G}\sum_{i=1}^{G}\rho_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2.\end{aligned}$$

The first factor is at most $1+\epsilon$, while $w_n^{(b,i)} \le 1$ removes the weights from the second factor. Taking conditional expectation, using $|\widehat{A}^{(b,i)}| \le 2R$, and changing measure from the behavior policy to the current policy yield

$$\begin{aligned}\mathbb{E}[\left\|U^w_{n,b}\right\|_2^2 \mid \mathcal{H}_n] &\le \frac{4R^2(1+\epsilon)}{T^2G}\sum_{i=1}^{G}\mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2\,\middle|\,\mathcal{H}_n\right]\\ &= \frac{4R^2(1+\epsilon)}{T^2}\mathbb{E}_{y\sim\pi_{\theta_n}(\cdot|x^{(b)})}\left\|\nabla \log \pi_{\theta_n}(y \mid x^{(b)})\right\|_2^2\\ &\le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T},\end{aligned}$$

where the last line is eq. (13). Finally, convexity of the squared norm and the tower property give

$$\begin{aligned}\mathbb{E}[\|g^w_n\|_2^2 \mid \mathcal{F}_n^-] &\le \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\left\|U^w_{n,b}\right\|_2^2\,\middle|\,\mathcal{F}_n^-\right]\\ &= \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}[\left\|U^w_{n,b}\right\|_2^2 \mid \mathcal{H}_n]\,\middle|\,\mathcal{F}_n^-\right] \le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}.\end{aligned}$$

For the bias envelope, $w_n^{(b,i)} \le 1$ implies $(1-w_n^{(b,i)})\rho_n^{(b,i)} \ge 0$. Conditional Jensen's inequality, the triangle inequality, $|\widehat{A}^{(b,i)}| \le 2R$, and eq. (15) then give

$$\begin{aligned}\left\|\mathbb{E}[b^w_{n,b} \mid \mathcal{H}_n]\right\|_2 &\le \frac{2R}{TG}\sum_{i=1}^{G}\mathbb{E}\Big[(1-w_n^{(b,i)})\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2\,\Big|\,\mathcal{H}_n\Big]\\ &\le 2R\sqrt{2L\log\mathcal{C}}\,\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}(1-w_n^{(b,i)})\rho_n^{(b,i)}\,\middle|\,\mathcal{H}_n\right],\end{aligned}$$

which is eq. (7). □

Proof of Proposition 2. For a direction-preserving correction with common weight α, the retained mass is $\alpha\overline{\rho}_n^{(b)}$. Since $\overline{\rho}_n^{(b)} > 0$, maximizing the retained mass—or, equivalently, minimizing the removed mass—amounts to choosing the largest α satisfying $0 \le \alpha \le 1$ and $\alpha\overline{\rho}_n^{(b)} \le 1+\epsilon$. The unique optimizer is therefore the stated $\alpha_n^{(b),\star}$, which is exactly the common scaling in eq. (12). □

The same calculation also yields a comparison with TIC-GRPO. GMC removes importance mass $(\overline{\rho}_n^{(b)} - (1+\epsilon))_+$. This quantity is zero when $\overline{\rho}_n^{(b)} \le 1+\epsilon$. Otherwise,

$$\overline{\rho}_n^{(b)} - (1+\epsilon) = \frac{1}{G}\sum_{i:\,\rho_n^{(b,i)}>1+\epsilon}\rho_n^{(b,i)} + \frac{1}{G}\sum_{i:\,\rho_n^{(b,i)}\le 1+\epsilon}\rho_n^{(b,i)} - (1+\epsilon) \le \frac{1}{G}\sum_{i:\,\rho_n^{(b,i)}>1+\epsilon}\rho_n^{(b,i)},$$

because the second normalized sum is at most $1+\epsilon$. Thus, in either case, GMC removes no more importance mass than TIC-GRPO. Applying eq. (7) and taking conditional expectations gives the corresponding comparison between their certified bias envelopes.

The scope of Proposition 2 is deliberate. It does not claim that GMC minimizes the exact vector bias over arbitrary sample-specific rules: equal retained mass can still lead to different cancellations and hence different biases. Rather, Propositions 1 and 2 identify GMC as the unique direction-preserving rule that retains the shared second-moment guarantee while minimizing a certified bias envelope. TIC-GRPO enforces the same mass cap by deleting individual high-ratio responses; GMC instead applies the smallest uniform attenuation needed to satisfy the cap. The next section uses this groupwise structure to derive the GMC conditional-bias and convergence bounds.

F CONVERGENCE ANALYSIS OF GMC-GRPO

This section establishes the convergence guarantee for GMC-GRPO under mixed rollout delays. By Lemma 1, it suffices to control the update second moment and the squared conditional bias. We first decompose the GMC update into the current-policy gradient and an explicit groupwise correction term, and show that the group-mass cap bounds the update second moment independently of the delay. We then bound the correction bias through trajectory divergences between the current and behavior policies, and relate these divergences to the parameter drift accumulated within each delay window. Combining these estimates with Lemma 1 yields the delay-explicit guarantee in Theorem 3. A final local-overlap argument converts the chi-square branch into the group-size refinement used in the main text. Throughout, conditional statements rely on the sampling and filtration conditions in Assumption 3.

For each group, define the raw importance-sampling, self-normalized, and group-mass-capped estimators

$$U_{n,b}^{\mathrm{IS}} := \frac{1}{TG}\sum_{i=1}^{G}\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)}\mid x^{(b)})\widehat{A}^{(b,i)}, \tag{19}$$

$$U_{n,b}^{\mathrm{SN}} := \frac{1}{TG}\sum_{i=1}^{G}\frac{\rho_n^{(b,i)}}{\overline{\rho}_n^{(b)}}\nabla\log\pi_{\theta_n}(y^{(b,i)}\mid x^{(b)})\widehat{A}^{(b,i)},$$

$$U_{n,b}^{\mathrm{GMC}} := \frac{1}{TG}\sum_{i=1}^{G}\widehat{\rho}_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)}\mid x^{(b)})\widehat{A}^{(b,i)}.$$

Proposition 3 (Conditional gradient decomposition)**.** *Under Assumption 1 and the sampling conditions in Assumption 3, for every n, b, the following pathwise identities hold:*

$$\frac{1}{G}\sum_{i=1}^{G}\widehat{\rho}_n^{(b,i)} = \min\{\overline{\rho}_n^{(b)}, 1+\epsilon\},$$

$$U_{n,b}^{\mathrm{GMC}} = \begin{cases} U_{n,b}^{\mathrm{IS}}, & \overline{\rho}_n^{(b)} \le 1+\epsilon, \\ (1+\epsilon)U_{n,b}^{\mathrm{SN}}, & \overline{\rho}_n^{(b)} > 1+\epsilon, \end{cases} \tag{20}$$

$$U_{n,b}^{\mathrm{GMC}} - U_{n,b}^{\mathrm{IS}} = -(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ U_{n,b}^{\mathrm{SN}}. \tag{21}$$

Moreover, $0 \le \widehat{\rho}_n^{(b,i)} \le \rho_n^{(b,i)}$ *for every i. Define*

$$b_{n,b} := -\mathbb{E}\left[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ U_{n,b}^{\mathrm{SN}} \,\middle|\, \mathcal{H}_n\right], \tag{22}$$

$$b_n := \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} b_{n,b} \,\middle|\, \mathcal{F}_n^-\right]. \tag{23}$$

Then

$$\mathbb{E}[U_{n,b}^{\mathrm{GMC}} \mid \mathcal{H}_n] = \frac{1}{T}\nabla J_{x^{(b)}}(\theta_n) + b_{n,b}, \tag{24}$$

$$\mathbb{E}[g_n^{\mathrm{GMC}} \mid \mathcal{F}_n^-] = \frac{1}{T}\nabla J(\theta_n) + b_n, \tag{25}$$

Thus every subsequent bias lemma only needs to control the explicit correction term in eq. (22)*.*

Proof. The definition eq. (12) gives the pathwise chain

$$\frac{1}{G}\sum_{i=1}^{G}\widehat{\rho}_n^{(b,i)} = \frac{\overline{\rho}_n^{(b)}}{\max\{1, \overline{\rho}_n^{(b)}/(1+\epsilon)\}} = \min\{\overline{\rho}_n^{(b)}, 1+\epsilon\},$$
$$U_{n,b}^{\mathrm{GMC}} = \min\{\overline{\rho}_n^{(b)}, 1+\epsilon\}U_{n,b}^{\mathrm{SN}}, \qquad U_{n,b}^{\mathrm{IS}} = \overline{\rho}_n^{(b)}U_{n,b}^{\mathrm{SN}},$$
$$U_{n,b}^{\mathrm{GMC}} - U_{n,b}^{\mathrm{IS}} = -(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ U_{n,b}^{\mathrm{SN}}.$$

This proves eq. (20)–eq. (21); the denominator in eq. (12) is at least one, so also $0 \le \widehat{\rho}_n^{(b,i)} \le \rho_n^{(b,i)}$.

It remains to compute the mean of the raw estimator. Since $\rho_n^{(b,i)}$ is the current-to-behavior likelihood ratio, it cancels the behavior-policy density conditional on $\mathcal{H}_n$. Hence, for the reward paired with the ith score,

$$\mathbb{E}\Big[\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})r^{(b,i)} \mid \mathcal{H}_n\Big] = \mathbb{E}_{y\sim\pi_{\theta_n}(\cdot|x^{(b)})}\Big[\nabla\log\pi_{\theta_n}(y \mid x^{(b)})r_\phi(x^{(b)}, y)\Big]$$
$$= \sum_y \pi_{\theta_n}(y \mid x^{(b)})\nabla\log\pi_{\theta_n}(y \mid x^{(b)})r_\phi(x^{(b)}, y) = \nabla J_{x^{(b)}}(\theta_n).$$

For a leave-one-out reward $r^{(b,j)}$ with $j \ne i$, conditional independence separates the two sampled responses. The score factor has conditional mean zero:

$$\mathbb{E}\Big[\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)}) \mid \mathcal{H}_n\Big] = \sum_y \pi_{\theta_n}(y \mid x^{(b)})\nabla\log\pi_{\theta_n}(y \mid x^{(b)})$$
$$= \sum_y \nabla\pi_{\theta_n}(y \mid x^{(b)}) = 0,$$
$$\mathbb{E}\Big[\rho_n^{(b,i)}\nabla\log\pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})r^{(b,j)} \mid \mathcal{H}_n\Big] = 0 \cdot \mathbb{E}[r^{(b,j)} \mid \mathcal{H}_n] = 0.$$

Substitution of the leave-one-out advantage therefore yields

$$\mathbb{E}[U_{n,b}^{\mathrm{IS}} \mid \mathcal{H}_n] = \frac{1}{TG}\sum_{i=1}^{G}\left(\nabla J_{x^{(b)}}(\theta_n) - \frac{1}{G-1}\sum_{j\ne i}0\right) = \frac{1}{T}\nabla J_{x^{(b)}}(\theta_n),$$
$$\mathbb{E}[U_{n,b}^{\mathrm{GMC}} \mid \mathcal{H}_n] = \mathbb{E}[U_{n,b}^{\mathrm{IS}} \mid \mathcal{H}_n] - \mathbb{E}\Big[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ U_{n,b}^{\mathrm{SN}} \mid \mathcal{H}_n\Big] = \frac{1}{T}\nabla J_{x^{(b)}}(\theta_n) + b_{n,b},$$
$$\mathbb{E}[g_n^{\mathrm{GMC}} \mid \mathcal{F}_n^-] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}[U_{n,b}^{\mathrm{GMC}} \mid \mathcal{H}_n]\,\middle|\,\mathcal{F}_n^-\right]$$
$$= \frac{1}{T}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\nabla J_{x^{(b)}}(\theta_n)\,\middle|\,\mathcal{F}_n^-\right] + b_n = \frac{1}{T}\nabla J(\theta_n) + b_n.$$

The final equality is the conditional prompt-unbiasedness clause in Assumption 3. These are eq. (24) and eq. (25); bounded rewards and eq. (13) ensure integrability by Cauchy–Schwarz. □

Lemma 5 (Expected update second moment)**.** *Under Assumption 1 and the sampling conditions in Assumption 3,*

$$\mathbb{E}[\|g_n^{\mathrm{GMC}}\|_2^2 \mid \mathcal{F}_n^-] \le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}. \tag{26}$$

Proof. We use the weighted Cauchy–Schwarz inequality

$$\left\|\sum_i a_i v_i\right\|_2^2 \le \left(\sum_i a_i\right)\left(\sum_i a_i\|v_i\|_2^2\right), \qquad a_i \ge 0.$$

Apply it with $a_i = \widehat{\rho}_n^{(b,i)}/G$ and $v_i = T^{-1}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}$. Proposition 3 identifies the sum of the weights, while $\widehat{\rho}_n^{(b,i)} \le \rho_n^{(b,i)}$ enlarges the second factor. Hence

$$\begin{aligned}
\left\|U_{n,b}^{\mathrm{GMC}}\right\|_2^2 &= \left\|\frac{1}{TG}\sum_i \widehat{\rho}_n^{(b,i)}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2 \\
&\le \left(\frac{1}{G}\sum_i \widehat{\rho}_n^{(b,i)}\right)\left(\frac{1}{G}\sum_i \widehat{\rho}_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2\right) \\
&\le \frac{\min\{1+\epsilon, \overline{\rho}_n^{(b)}\}}{G}\sum_i \rho_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2 \\
&\le \frac{1+\epsilon}{G}\sum_i \rho_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2.
\end{aligned}$$

The second inequality uses $G^{-1}\sum_i \widehat{\rho}_n^{(b,i)} = \min\{1+\epsilon, \overline{\rho}_n^{(b)}\}$ and $\widehat{\rho}_n^{(b,i)} \le \rho_n^{(b,i)}$; the third uses $\min\{1+\epsilon, \overline{\rho}_n^{(b)}\} \le 1+\epsilon$. Taking conditional expectation and using $|\widehat{A}^{(b,i)}| \le 2R$ gives

$$\begin{aligned}
\mathbb{E}[\left\|U_{n,b}^{\mathrm{GMC}}\right\|_2^2 \mid \mathcal{H}_n] &\le \frac{4R^2(1+\epsilon)}{T^2 G}\sum_{i=1}^G \mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2 \,\middle|\, \mathcal{H}_n\right] \\
&= \frac{4R^2(1+\epsilon)}{T^2}\mathbb{E}_{y\sim\pi_{\theta_n}(\cdot|x^{(b)})}\left\|\nabla \log \pi_{\theta_n}(y \mid x^{(b)})\right\|_2^2 \\
&\le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}.
\end{aligned}$$

The equality changes measure from the behavior policy to the current policy, and the final inequality is eq. (13). Finally, convexity gives the pathwise batch bound

$$\left\|g_n^{\mathrm{GMC}}\right\|_2^2 = \left\|\frac{1}{B}\sum_{b=1}^B U_{n,b}^{\mathrm{GMC}}\right\|_2^2 \le \frac{1}{B}\sum_{b=1}^B \left\|U_{n,b}^{\mathrm{GMC}}\right\|_2^2.$$

Conditioning first on $\mathcal{H}_n$ and then on $\mathcal{F}_n^-$ proves eq. (26). □

Lemma 6 (Bias controls)**.** *Under Assumptions 1 and 2 and the sampling conditions in Assumption 3, for every n, b,*

$$\|b_{n,b}\|_2^2 \le \min\left\{\frac{8R^2L\log|\mathcal{A}|}{T}, \frac{8R^2L\log|\mathcal{A}|}{Tc_\epsilon}D_n^{(b)}, \frac{8R^2L\log\mathcal{C}}{\log^2(1+2\epsilon)}(D_n^{(b)})^2, \frac{R^2L\log\mathcal{C}}{2G^2\epsilon^2}(\chi_n^{2,(b)})^2\right\}. \tag{27}$$

Proof. Let $\varphi(u) := u\log u - u + 1$ and abbreviate $E_b := \{\overline{\rho}_n^{(b)} > 1+\epsilon\}$. Expanding the self-normalized update in eq. (22),

$$b_{n,b} = -\frac{1}{T}\mathbb{E}\left[\frac{(\overline{\rho}_n^{(b)} - (1+\epsilon))_+}{\overline{\rho}_n^{(b)}}\frac{1}{G}\sum_{i=1}^G \rho_n^{(b,i)}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)} \,\middle|\, \mathcal{H}_n\right].$$

The scalar prefactor is at most $\mathbf{1}_{E_b}$. Conditional Jensen and $|\widehat{A}^{(b,i)}| \le 2R$ therefore imply

$$\|b_{n,b}\|_2 \le \frac{2R}{T}\frac{1}{G}\sum_{i=1}^G \mathbb{E}\left[\mathbf{1}_{E_b}\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2 \,\middle|\, \mathcal{H}_n\right].$$

For each i, conditional Cauchy–Schwarz bounds the expectation inside the sum as

$$\mathbb{E}\left[\mathbf{1}_{E_b}\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2 \,\middle|\, \mathcal{H}_n\right] \le \mathbb{E}[\mathbf{1}_{E_b}\rho_n^{(b,i)} \mid \mathcal{H}_n]^{1/2} \times \mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2 \,\middle|\, \mathcal{H}_n\right]^{1/2}.$$

A second Cauchy–Schwarz inequality over i multiplies the averages of the two right-hand factors. The first average is $\mathbb{E}[\overline{\rho}_n^{(b)} \mathbf{1}_{E_b} \mid \mathcal{H}_n]$, so we obtain

$$\|b_{n,b}\|_2^2 \le \frac{4R^2}{T^2}\mathbb{E}\Big[\overline{\rho}_n^{(b)}\mathbf{1}\{\overline{\rho}_n^{(b)} > 1+\epsilon\} \mid \mathcal{H}_n\Big] \frac{1}{G}\sum_{i=1}^{G}\mathbb{E}\Big[\rho_n^{(b,i)} \Big\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\Big\|_2^2 \mid \mathcal{H}_n\Big]$$
$$\le \frac{8R^2 L \log|\mathcal{A}|}{T}\mathbb{E}\Big[\overline{\rho}_n^{(b)}\mathbf{1}\{\overline{\rho}_n^{(b)} > 1+\epsilon\} \mid \mathcal{H}_n\Big].$$

In the last line, change of measure turns each second conditional expectation in the preceding display into the current-policy score moment, which is at most $2LT\log|\mathcal{A}|$ by eq. (13). Moreover, $\mathbb{E}[\overline{\rho}_n^{(b)} \mid \mathcal{H}_n] = 1$ because each importance ratio has conditional mean one. Dropping $\mathbf{1}_{E_b}$ proves the uniform branch.

For the KL branch, $\varphi(u)/u = \log u - 1 + u^{-1}$ has derivative $(u-1)/u^2 > 0$ for $u > 1$. Its value at $u = 1+\epsilon$ is $c_\epsilon = \log(1+\epsilon) - \epsilon/(1+\epsilon)$, so the pointwise inequality

$$c_\epsilon u \mathbf{1}\{u > 1+\epsilon\} \le \varphi(u), \qquad u \ge 0,$$

holds. Apply it to $u = \overline{\rho}_n^{(b)}$, then use convexity of φ:

$$c_\epsilon\,\mathbb{E}\Big[\overline{\rho}_n^{(b)}\mathbf{1}\{\overline{\rho}_n^{(b)} > 1+\epsilon\} \mid \mathcal{H}_n\Big] \le \mathbb{E}[\varphi(\overline{\rho}_n^{(b)}) \mid \mathcal{H}_n] \le \frac{1}{G}\sum_{i=1}^{G}\mathbb{E}[\varphi(\rho_n^{(b,i)}) \mid \mathcal{H}_n] = D_n^{(b)}.$$

For the last equality, change of measure gives $\mathbb{E}[\varphi(\rho_n^{(b,i)}) \mid \mathcal{H}_n] = D_{\mathrm{KL}}(\pi_{\theta_n}\|\pi_{\theta_{n-\tau_n^{(b)}}}) = D_n^{(b)}$. Substitution in the preceding bound proves the KL branch.

The first two branches do not use Assumption 2. For the remaining branches, coverage and eq. (15) give $\big\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\big\|_2 \le T\sqrt{2L\log \mathcal{C}}$ for every sampled response. Since the normalized weights sum to one, the triangle inequality and $|\widehat{A}^{(b,i)}| \le 2R$ yield the pathwise bound

$$\big\|U_{n,b}^{\mathrm{SN}}\big\|_2 \le \frac{2R}{TG}\sum_{i=1}^{G}\frac{\rho_n^{(b,i)}}{\overline{\rho}_n^{(b)}}\Big\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\Big\|_2 \le 2R\sqrt{2L\log \mathcal{C}}.$$

Using this inside eq. (22) and applying conditional Jensen gives

$$\|b_{n,b}\|_2 \le \mathbb{E}\Big[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ \big\|U_{n,b}^{\mathrm{SN}}\big\|_2 \,\Big|\, \mathcal{H}_n\Big] \le 2R\sqrt{2L\log \mathcal{C}}\,\mathbb{E}[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ \mid \mathcal{H}_n]. \quad (28)$$

For $u \ge 1+\epsilon$, convexity and the tangent to φ at $1+2\epsilon$ give

$$\begin{aligned}\varphi(u) &\ge \varphi(1+2\epsilon) + \varphi'(1+2\epsilon)(u-1-2\epsilon)\\ &= \log(1+2\epsilon)u - 2\epsilon\\ &\ge \log(1+2\epsilon)(u-(1+\epsilon)).\end{aligned}$$

The last inequality is equivalent to $(1+\epsilon)\log(1+2\epsilon) \ge 2\epsilon$, which follows from $\log v \ge 2(v-1)/(v+1)$ with $v = 1+2\epsilon$. Applying the resulting pointwise bound to $u = \overline{\rho}_n^{(b)}$, followed by Jensen and change of measure, gives

$$\log(1+2\epsilon)\,\mathbb{E}\left[(\overline{\rho}_n^{(b)} - (1+\epsilon))_+ \,\middle|\, \mathcal{H}_n\right] \le \mathbb{E}[\varphi(\overline{\rho}_n^{(b)}) \mid \mathcal{H}_n] \le \frac{1}{G}\sum_{i=1}^{G}\mathbb{E}[\varphi(\rho_n^{(b,i)}) \mid \mathcal{H}_n] = D_n^{(b)}.$$

Substitution in eq. (28) proves the quadratic-KL branch.

For the chi-square branch, the scalar inequality

$$(u-1-\epsilon)_+ \le \frac{(u-1)^2}{4\epsilon}, \qquad u \ge 0,$$

is immediate when $u \le 1+\epsilon$; when $u > 1+\epsilon$, it is equivalent to $(u-1-2\epsilon)^2 \ge 0$. Next expand the conditional variance of the group mean. Because the responses are conditionally independent and each importance ratio has conditional mean one, for $i \ne j$,

$$\mathbb{E}[(\rho_n^{(b,i)} - 1)(\rho_n^{(b,j)} - 1) \mid \mathcal{H}_n] = 0.$$

Therefore

$$\mathbb{E}\left[(\overline{\rho}_n^{(b)}-1)^2 \,\middle|\, \mathcal{H}_n\right] = \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}[(\rho_n^{(b,i)}-1)^2 \mid \mathcal{H}_n] = \frac{\chi_n^{2,(b)}}{G},$$

$$\mathbb{E}[(\overline{\rho}_n^{(b)}-(1+\epsilon))_+ \mid \mathcal{H}_n] \le \frac{1}{4\epsilon}\mathbb{E}[(\overline{\rho}_n^{(b)}-1)^2 \mid \mathcal{H}_n] = \frac{\chi_n^{2,(b)}}{4G\epsilon}.$$

Substitution in eq. (28) proves the last branch. □

Lemma 7 (KL–delay conversions)**.** *Suppose the iterates are generated by the GMC update eq.* (9)*, Assumption 1 holds, and both the sampling and delay-regularity conditions in Assumption 3 hold.*

$$\mathbb{E}[D_n^{(b)} \mid \tau_n^{(b)}] \le 4(1+\epsilon)L^2R^2\log|\mathcal{A}|\,\eta^2(\tau_n^{(b)})^2. \tag{29}$$

If Assumption 2 also holds, then

$$\begin{aligned} D_n^{(b)} &\le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2(\tau_n^{(b)})^2 \quad \textit{almost surely},\\ \mathbb{E}[(D_n^{(b)})^2 \mid \tau_n^{(b)}] &\le 16(1+\epsilon)^3TL^4R^4\log|\mathcal{A}|\log\mathcal{C}\,\eta^4(\tau_n^{(b)})^4. \end{aligned} \tag{30}$$

Proof. Fix $t \ge 0$ with $\Pr(\tau_n^{(b)}=t)>0$. The learner recursion gives

$$\theta_n-\theta_{n-t} = \eta\sum_{s=n-t}^{n-1} g_s^{\mathrm{GMC}}, \qquad \|\theta_n-\theta_{n-t}\|_2^2 \le \eta^2 t\sum_{s=n-t}^{n-1}\left\|g_s^{\mathrm{GMC}}\right\|_2^2,$$

where the inequality is Cauchy–Schwarz over the t learner updates. The delay-regularity condition in Assumption 3 removes the conditioning on the realized delay in the first line below. The tower property and Lemma 5 then bound every update moment:

$$\mathbb{E}[\|\theta_n-\theta_{n-t}\|_2^2 \mid \tau_n^{(b)}=t] = \mathbb{E}\,\|\theta_n-\theta_{n-t}\|_2^2 \le \eta^2 t\sum_{s=n-t}^{n-1}\mathbb{E}\left\|g_s^{\mathrm{GMC}}\right\|_2^2 \le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}\eta^2t^2,$$

$$\mathbb{E}[D_n^{(b)} \mid \tau_n^{(b)}=t] \le \frac{TL}{2}\mathbb{E}[\|\theta_n-\theta_{n-t}\|_2^2 \mid \tau_n^{(b)}=t] \le 4(1+\epsilon)L^2R^2\log|\mathcal{A}|\,\eta^2t^2.$$

The second line applies the trajectory-KL bound eq. (17); this proves eq. (29).

Under Assumption 2, the capped group mass and eq. (15) give, for every group,

$$\left\|U_{s,b}^{\mathrm{GMC}}\right\|_2 \le \frac{1}{TG}\sum_{i=1}^{G}\widehat{\rho}_s^{(b,i)}\left\|\nabla\log\pi_{\theta_s}(y^{(b,i)} \mid x^{(b)})\right\|_2|\widehat{A}^{(b,i)}| \le 2R(1+\epsilon)\sqrt{2L\log\mathcal{C}}.$$

The same bound holds for g_s^{GMC}, because it is the average of the group updates. The triangle inequality along the learner path therefore yields

$$\|\theta_n-\theta_{n-t}\|_2 \le \eta\sum_{s=n-t}^{n-1}\left\|g_s^{\mathrm{GMC}}\right\|_2 \le 2R(1+\epsilon)\sqrt{2L\log\mathcal{C}}\,\eta t.$$

Combining this pathwise estimate with eq. (17) gives

$$\begin{aligned} D_n^{(b)} &\le \frac{TL}{2}\|\theta_n-\theta_{n-t}\|_2^2 \le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2t^2 =: D_{\max},\\ \mathbb{E}[(D_n^{(b)})^2 \mid \tau_n^{(b)}=t] &\le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2t^2\mathbb{E}[D_n^{(b)} \mid \tau_n^{(b)}=t]\\ &\le 16(1+\epsilon)^3TL^4R^4\log|\mathcal{A}|\log\mathcal{C}\,\eta^4t^4, \end{aligned}$$

where the middle inequality uses $0 \le D_n^{(b)} \le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2t^2$, hence $(D_n^{(b)})^2 \le D_{\max}D_n^{(b)}$. This proves eq. (30). □

Define the local coverage penalty

$$\Gamma_{n,b}^{\mathrm{GMC}} := \min\left\{ 8LTR^2 \log|\mathcal{A}|, \frac{8TR^2L\log|\mathcal{A}|}{c_\epsilon} D_n^{(b)}, \frac{8T^2R^2L\log\mathcal{C}}{\log^2(1+2\epsilon)} (D_n^{(b)})^2, \frac{T^2R^2L\log\mathcal{C}}{2G^2\epsilon^2} (\chi_n^{2,(b)})^2 \right\}. \tag{31}$$

Since $D_n^{(b)}$ and $\chi_n^{2,(b)}$ are $\mathcal{H}_n$-measurable, $\Gamma_{n,b}^{\mathrm{GMC}}$ is a nonnegative $\mathcal{H}_n$-measurable random variable bounded by $8LTR^2 \log|\mathcal{A}|$.

Theorem 3 (Delay-explicit convergence guarantee for GMC-GRPO)**.** *Suppose the iterates are generated by the GMC update eq.* (9)*, Assumptions 1 and 2 hold, and the sampling conditions in Assumption 3 hold. Then, for every $\eta > 0$ and $N \geq 1$,*

$$\frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\,\|\nabla J(\theta_n)\|_2^2 \leq \frac{2T\big(J^\star - \mathbb{E}[J(\theta_0)]\big)}{\eta N} + 8T(1+\epsilon)L^2R^3 \log|\mathcal{A}|\,(2\log|\mathcal{A}|+1)\eta + \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \Gamma_{n,b}^{\mathrm{GMC}}\right]. \tag{32}$$

If the delay-regularity condition in Assumption 3 also holds, the final term is at most

$$\min\left\{ 8LTR^2\log|\mathcal{A}|, \frac{32(1+\epsilon)L^3R^4T(\log|\mathcal{A}|)^2}{c_\epsilon}\eta^2\overline{\tau}_2, \frac{128(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{\log^2(1+2\epsilon)}\eta^4\overline{\tau}_4, \frac{T^2R^2L\log\mathcal{C}}{2G^2\epsilon^2}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(\chi_n^{2,(b)})^2\right] \right\}, \tag{33}$$

where a branch equal to $+\infty$ is simply inactive in the minimum.

Each branch of eq. (33) is a valid bound on its own. The uniform branch shows that the correction bias always contributes a bounded amount. The second- and fourth-order branches make the dependence on rollout delay explicit through $\overline{\tau}_2$ and $\overline{\tau}_4$. The chi-square branch yields the group-size refinement in Corollary 2. Theorem 1 in the main text retains the fourth-order branch.

Proof. The proof has three steps: we verify the hypothesis of Lemma 1, bound its update-moment and bias terms, and finally convert the bias penalty into delay moments.

Step 1: conditional-mean structure. By eq. (25), $\mathbb{E}[g_n^{\mathrm{GMC}} \mid \mathcal{F}_n^-] = T^{-1}\nabla J(\theta_n) + b_n$, where b_n is given by eq. (23) and is $\mathcal{F}_n^-$-measurable. The second-moment bound in Lemma 5 and the uniform branch in eq. (27) show that g_n^{GMC} and b_n are square integrable. Hence Lemma 1 applies and gives

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\,\|\nabla J(\theta_n)\|_2^2 \leq \frac{2T(J^\star - \mathbb{E}[J(\theta_0)])}{\eta N} + \frac{T^2RL(2\log|\mathcal{A}|+1)\eta}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|g_n^{\mathrm{GMC}}\right\|_2^2 + \frac{T^2}{N}\sum_{n=0}^{N-1}\mathbb{E}\,\|b_n\|_2^2.$$

Step 2: update moment and bias. Taking expectations in Lemma 5 gives $\mathbb{E}\left\|g_n^{\mathrm{GMC}}\right\|_2^2 \leq 8(1+\epsilon)LR^2\log|\mathcal{A}|/T$ for every n. Therefore

$$\frac{T^2RL(2\log|\mathcal{A}|+1)\eta}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|g_n^{\mathrm{GMC}}\right\|_2^2 \leq T^2RL(2\log|\mathcal{A}|+1)\eta\cdot\frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T} = 8T(1+\epsilon)L^2R^3\log|\mathcal{A}|(2\log|\mathcal{A}|+1)\eta.$$

For the bias term, substitute eq. (23). Conditional Jensen for the squared norm, followed by convexity of the squared norm over the uniform average of the B groups, gives

$$\|b_n\|_2^2 = \left\|\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} b_{n,b}\,\middle|\,\mathcal{F}_n^-\right]\right\|_2^2 \leq \mathbb{E}\left[\left\|\frac{1}{B}\sum_{b=1}^{B} b_{n,b}\right\|_2^2\,\middle|\,\mathcal{F}_n^-\right] \leq \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\|b_{n,b}\|_2^2\,\middle|\,\mathcal{F}_n^-\right].$$

The bound eq. (27) in Lemma 6 holds pathwise for every (n, b), and $\Gamma_{n,b}^{\mathrm{GMC}}$ is exactly T^2 times its four-branch minimum. Taking expectations therefore yields

$$T^2\mathbb{E}\,\|b_n\|_2^2 \le T^2\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\|b_{n,b}\|_2^2\right] \le \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\Gamma_{n,b}^{\mathrm{GMC}}\right].$$

Substituting the moment and bias bounds into the inequality of Step 1 proves eq. (32).

Step 3: conversion to delay moments. Suppose now that the delay-regularity condition in Assumption 3 also holds. For nonnegative random variables $X_1, \dots, X_4$, $\mathbb{E}[\min_k X_k] \le \min_k \mathbb{E}[X_k]$. Hence the final term of eq. (32) is at most the minimum of the four averaged branches, and it suffices to bound each branch separately.

The uniform branch is the constant $8LTR^2 \log|\mathcal{A}|$. For the linear KL branch, the tower property and eq. (29) give $\mathbb{E}[D_n^{(b)}] = \mathbb{E}\big[\mathbb{E}[D_n^{(b)} \mid \tau_n^{(b)}]\big] \le 4(1+\epsilon)L^2R^2\log|\mathcal{A}|\,\eta^2\,\mathbb{E}[(\tau_n^{(b)})^2]$. Averaging over b and n and using the definition eq. (11) of $\overline{\tau}_2$,

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\frac{8TR^2L\log|\mathcal{A}|}{c_\epsilon}D_n^{(b)}\right] \le \frac{32(1+\epsilon)L^3R^4T(\log|\mathcal{A}|)^2}{c_\epsilon}\eta^2\overline{\tau}_2.$$

For the quadratic-KL branch, the tower property and eq. (30) give $\mathbb{E}[(D_n^{(b)})^2] \le 16(1+\epsilon)^3TL^4R^4\log|\mathcal{A}|\log\mathcal{C}\,\eta^4\,\mathbb{E}[(\tau_n^{(b)})^4]$. Averaging in the same way,

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\frac{8T^2R^2L\log\mathcal{C}}{\log^2(1+2\epsilon)}(D_n^{(b)})^2\right] \le \frac{128(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{\log^2(1+2\epsilon)}\eta^4\overline{\tau}_4.$$

The chi-square branch is kept in its averaged form

$$\frac{T^2R^2L\log\mathcal{C}}{2G^2\epsilon^2}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(\chi_n^{2,(b)})^2\right].$$

Taking the minimum of these four upper bounds proves eq. (33). □

Corollary 3 (Local overlap converts chi-square to KL)**.** *Suppose additionally that, for some finite $\delta \ge 0$, the trajectory likelihood ratio satisfies $|\rho_n^{(b,i)} - 1| \le \delta$ almost surely for all n, b, i. Then $\chi_n^{2,(b)} \le 2(1+\delta)D_n^{(b)}$ almost surely for every n, b.*

Proof. Let $\varphi(u) = u\log u - u + 1$. Softmax full support implies $\rho_n^{(b,i)} > 0$ almost surely. For any $u > 0$ satisfying $|u-1| \le \delta$ and every $s \in [0,1]$, $0 < 1 + s(u-1) \le \max\{1, u\} \le 1+\delta$. Hence $\varphi''(1+s(u-1)) \ge (1+\delta)^{-1}$. Since $\varphi(1) = \varphi'(1) = 0$, Taylor's formula with integral remainder gives

$$\varphi(u) = (u-1)^2\int_0^1 (1-s)\varphi''(1+s(u-1))\,\mathrm{d}s \ge \frac{(u-1)^2}{2(1+\delta)}.$$

Apply this pointwise to $u = \rho_n^{(b,i)}$ and take expectation under the behavior policy:

$$D_n^{(b)} = \mathbb{E}[\varphi(\rho_n^{(b,i)}) \mid \mathcal{H}_n] \ge \frac{1}{2(1+\delta)}\mathbb{E}[(\rho_n^{(b,i)} - 1)^2 \mid \mathcal{H}_n] = \frac{\chi_n^{2,(b)}}{2(1+\delta)},$$

which proves the claim. □

G Convergence Analysis of TIC-GRPO

The TIC-GRPO analysis follows the same proof structure, making the effect of individual ratio truncation explicit. We decompose its update into the raw importance-sampling estimator and the discarded contribution, establish the shared second-moment bound, and bound the clipping bias using trajectory divergences. Combining these estimates with the parameter-drift argument and

Lemma 1 yields the TIC convergence bound. The resulting coefficients allow a direct comparison with GMC under the same assumptions.

For every group, define the individually up-clipped estimator

$$U_{n,b}^{\mathrm{TIC}} := \frac{1}{TG}\sum_{i=1}^{G}\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}, \tag{34}$$

and the full mini-batch update

$$g_n^{\mathrm{TIC}} := \frac{1}{B}\sum_{b=1}^{B} U_{n,b}^{\mathrm{TIC}}, \qquad \theta_{n+1} = \theta_n + \eta g_n^{\mathrm{TIC}}. \tag{35}$$

Proposition 4 (TIC-GRPO conditional gradient decomposition)**.** *Under Assumption 1 and the sampling conditions in Assumption 3, let $U_{n,b}^{\mathrm{IS}}$ be the raw estimator in eq.* (19)*. Pathwise,*

$$U_{n,b}^{\mathrm{TIC}} - U_{n,b}^{\mathrm{IS}} = -\frac{1}{TG}\sum_{i=1}^{G}\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}. \tag{36}$$

Set $b_{n,b}^{\mathrm{TIC}} := \mathbb{E}[U_{n,b}^{\mathrm{TIC}} - U_{n,b}^{\mathrm{IS}} \mid \mathcal{H}_n]$. Then

$$\mathbb{E}[U_{n,b}^{\mathrm{TIC}} \mid \mathcal{H}_n] = \frac{1}{T}\nabla J_{x^{(b)}}(\theta_n) + b_{n,b}^{\mathrm{TIC}}, \tag{37}$$

$$\mathbb{E}[g_n^{\mathrm{TIC}} \mid \mathcal{F}_n^-] = \frac{1}{T}\nabla J(\theta_n) + b_n^{\mathrm{TIC}}, \tag{38}$$

where $b_n^{\mathrm{TIC}} := \mathbb{E}[B^{-1}\sum_{b=1}^{B} b_{n,b}^{\mathrm{TIC}} \mid \mathcal{F}_n^-]$.

Proof. Subtracting eq. (19) from eq. (34) gives eq. (36). Proposition 3 already proves the mean identity for the raw estimator, including cancellation of the leave-one-out baseline. Hence

$$\mathbb{E}[U_{n,b}^{\mathrm{TIC}} \mid \mathcal{H}_n] = \mathbb{E}[U_{n,b}^{\mathrm{IS}} \mid \mathcal{H}_n] + b_{n,b}^{\mathrm{TIC}} = \frac{1}{T}\nabla J_{x^{(b)}}(\theta_n) + b_{n,b}^{\mathrm{TIC}},$$

$$\mathbb{E}[g_n^{\mathrm{TIC}} \mid \mathcal{F}_n^-] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}[U_{n,b}^{\mathrm{TIC}} \mid \mathcal{H}_n]\,\middle|\,\mathcal{F}_n^-\right] = \frac{1}{T}\nabla J(\theta_n) + b_n^{\mathrm{TIC}}.$$

This proves eq. (37)–eq. (38). □

Remark 8 (The TIC bias does not depend on G). Fix $\mathcal{H}_n$ and write $\mathbb{E}_{\mathrm{b}}$ for expectation under the behavior policy $\pi_{\theta_{n-\tau_n^{(b)}}}(\cdot \mid x^{(b)})$. Since the responses in a group are conditionally i.i.d. and $r^{(b,j)}$ is independent of $y^{(b,i)}$ for $j \neq i$, substituting the leave-one-out advantage into eq. (36) gives

$$b_{n,b}^{\mathrm{TIC}} = -\frac{1}{T}\,\mathbb{E}_{\mathrm{b}}\Big[\rho\,\mathbf{1}\{\rho > 1+\epsilon\}\,\nabla \log \pi_{\theta_n}(y \mid x^{(b)})\big(r_\phi(x^{(b)},y) - \mathbb{E}_{\mathrm{b}}[r_\phi(x^{(b)},y)]\big)\Big],$$

where ρ is the current-to-behavior ratio of y. The right-hand side does not depend on G and is nonzero in general. In contrast, the chi-square branch of eq. (27) shows that the GMC bias is $O\big(\chi_n^{2,(b)}/(G\epsilon)\big)$ and vanishes as $G \to \infty$ for fixed behavior and current policies.

Lemma 8 (Expected TIC-GRPO update second moment)**.** *Under Assumption 1 and the sampling conditions in Assumption 3,*

$$\mathbb{E}[\|g_n^{\mathrm{TIC}}\|_2^2 \mid \mathcal{F}_n^-] \le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}. \tag{39}$$

The GMC counterpart in Lemma 5 has the same constant.

Proof. We use the weighted Cauchy–Schwarz inequality

$$\left\|\sum_i a_i v_i\right\|_2^2 \le \left(\sum_i a_i\right)\left(\sum_i a_i \|v_i\|_2^2\right), \qquad a_i \ge 0.$$

Apply it with $a_i = G^{-1}\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}$ and $v_i = T^{-1}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}$. Since each retained ratio is at most $1+\epsilon$,

$$\left\|U_{n,b}^{\mathrm{TIC}}\right\|_2^2 \le \left(\frac{1}{G}\sum_{i=1}^{G}\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}\right) \times \left(\frac{1}{G}\sum_{i=1}^{G}\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} \le 1+\epsilon\}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2\right)$$
$$\le \frac{1+\epsilon}{G}\sum_{i=1}^{G}\rho_n^{(b,i)}\left\|\frac{1}{T}\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\widehat{A}^{(b,i)}\right\|_2^2.$$

Taking conditional expectation and using $|\widehat{A}^{(b,i)}| \le 2R$ gives

$$\mathbb{E}[\left\|U_{n,b}^{\mathrm{TIC}}\right\|_2^2 \mid \mathcal{H}_n] \le \frac{4R^2(1+\epsilon)}{T^2G}\sum_{i=1}^{G}\mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2 \,\middle|\, \mathcal{H}_n\right]$$
$$= \frac{4R^2(1+\epsilon)}{T^2}\mathbb{E}_{y\sim\pi_{\theta_n}(\cdot|x^{(b)})}\left\|\nabla \log \pi_{\theta_n}(y \mid x^{(b)})\right\|_2^2$$
$$\le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}.$$

The equality changes measure from the behavior policy to the current policy, and the last inequality is eq. (13). Finally, convexity gives

$$\left\|g_n^{\mathrm{TIC}}\right\|_2^2 = \left\|\frac{1}{B}\sum_{b=1}^{B}U_{n,b}^{\mathrm{TIC}}\right\|_2^2 \le \frac{1}{B}\sum_{b=1}^{B}\left\|U_{n,b}^{\mathrm{TIC}}\right\|_2^2.$$

Conditioning first on $\mathcal{H}_n$ and then on $\mathcal{F}_n^-$ proves eq. (39). □

Lemma 9 (TIC-GRPO clipping-error bounds)**.** *Under Assumptions 1 and 2 and the sampling conditions in Assumption 3, for every n, b,*

$$\left\|b_{n,b}^{\mathrm{TIC}}\right\|_2^2 \le \min\left\{\frac{8R^2L\log|\mathcal{A}|}{T}, \frac{8R^2L\log|\mathcal{A}|}{Tc_\epsilon}D_n^{(b)}, \frac{8R^2L\log\mathcal{C}}{c_\epsilon^2}(D_n^{(b)})^2, \frac{8R^2L\log\mathcal{C}(1+\epsilon)^2}{\epsilon^4}(\chi_n^{2,(b)})^2\right\}. \tag{40}$$

Proof. Let $\varphi(u) := u\log u - u + 1$. Expanding eq. (36), applying conditional Jensen, and using $|\widehat{A}^{(b,i)}| \le 2R$ give

$$\left\|b_{n,b}^{\mathrm{TIC}}\right\|_2 \le \frac{2R}{TG}\sum_{i=1}^{G}\mathbb{E}\left[\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2 \,\middle|\, \mathcal{H}_n\right].$$

For each i, conditional Cauchy–Schwarz gives

$$\mathbb{E}\left[\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2 \,\middle|\, \mathcal{H}_n\right]^2$$
$$\le \mathbb{E}\left[\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\} \,\middle|\, \mathcal{H}_n\right]\mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2 \,\middle|\, \mathcal{H}_n\right].$$

All G responses have the same conditional law. Squaring the preceding bound, changing measure in the score-moment factor, and applying eq. (13) therefore yield

$$\left\|b_{n,b}^{\mathrm{TIC}}\right\|_2^2 \le \frac{4R^2}{T^2}\mathbb{E}[\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\} \mid \mathcal{H}_n]\mathbb{E}\left[\rho_n^{(b,i)}\left\|\nabla \log \pi_{\theta_n}(y^{(b,i)} \mid x^{(b)})\right\|_2^2 \mid \mathcal{H}_n\right]$$
$$\le \frac{8R^2L\log|\mathcal{A}|}{T}\mathbb{E}[\rho_n^{(b,i)}\mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\} \mid \mathcal{H}_n].$$

The tail mass is at most $\mathbb{E}[\rho_n^{(b,i)} \mid \mathcal{H}_n] = 1$, which proves the uniform branch. For the KL branch, $\varphi(u)/u = \log u - 1 + u^{-1}$ is increasing for $u > 1$, and its value at $u = 1+\epsilon$ is c_ϵ. Hence

$$c_\epsilon u\mathbf{1}\{u > 1+\epsilon\} \le \varphi(u), \qquad u \ge 0.$$

Applying this inequality to $u = \rho_n^{(b,i)}$ and changing measure gives

$$c_\epsilon \, \mathbb{E}[\rho_n^{(b,i)} \mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\} \mid \mathcal{H}_n] \le \mathbb{E}[\varphi(\rho_n^{(b,i)}) \mid \mathcal{H}_n] = D_n^{(b)}.$$

Substitution proves the KL branch.

The first two branches do not use Assumption 2. For the remaining branches, eq. (15) bounds the score norm in the first display uniformly, so

$$\begin{aligned} \left\| b_{n,b}^{\mathrm{TIC}} \right\|_2 &\le 2R\sqrt{2L\log \mathcal{C}}\, \mathbb{E}[\rho_n^{(b,i)} \mathbf{1}\{\rho_n^{(b,i)} > 1+\epsilon\} \mid \mathcal{H}_n] \\ &\le 2R\sqrt{2L\log \mathcal{C}}\, \min\left\{ \frac{D_n^{(b)}}{c_\epsilon}, \frac{1+\epsilon}{\epsilon^2} \chi_n^{2,(b)} \right\}, \end{aligned}$$

where the first alternative is the KL tail bound above. For the second, use the pointwise inequality

$$u\mathbf{1}\{u > 1+\epsilon\} \le \frac{1+\epsilon}{\epsilon^2}(u-1)^2.$$

Indeed, $u/(u-1)^2$ is decreasing for $u > 1$, so on the event $u > 1+\epsilon$ it is at most $(1+\epsilon)/\epsilon^2$. Taking behavior-policy expectation turns $(u-1)^2$ into $\chi_n^{2,(b)}$. Squaring the two alternatives proves the additional branches in eq. (40). □

For each realized group, define the local TIC penalty

$$\Gamma_{n,b}^{\mathrm{TIC}} := \min\left\{ 8LTR^2 \log|\mathcal{A}|, \frac{8TR^2L\log|\mathcal{A}|}{c_\epsilon} D_n^{(b)}, \frac{8T^2R^2L\log\mathcal{C}}{c_\epsilon^2}(D_n^{(b)})^2, \frac{8T^2R^2L\log\mathcal{C}(1+\epsilon)^2}{\epsilon^4}(\chi_n^{2,(b)})^2 \right\}. \tag{41}$$

Theorem 4 (Delay-explicit convergence guarantee for TIC-GRPO)**.** *Suppose the iterates are generated by the TIC update eq.* (35)*, Assumptions 1 and 2 hold, and the sampling conditions in Assumption 3 hold. Then, for every $\eta > 0$ and $N \ge 1$,*

$$\begin{aligned} \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\,\|\nabla J(\theta_n)\|_2^2 \le \frac{2T\big(J^\star - \mathbb{E}[J(\theta_0)]\big)}{\eta N} &+ 8T(1+\epsilon)L^2R^3\log|\mathcal{A}|\,(2\log|\mathcal{A}|+1)\eta \\ &+ \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \Gamma_{n,b}^{\mathrm{TIC}}\right], \end{aligned} \tag{42}$$

where $\Gamma_{n,b}^{\mathrm{TIC}}$ is defined in eq. (41)*. If the delay-regularity condition in Assumption 3 also holds, the final term is at most*

$$\begin{aligned} \min\Bigg\{ &8LTR^2\log|\mathcal{A}|, \frac{32(1+\epsilon)L^3R^4T(\log|\mathcal{A}|)^2}{c_\epsilon}\eta^2\overline{\tau}_2, \frac{128(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{c_\epsilon^2}\eta^4\overline{\tau}_4, \\ &\frac{8T^2R^2L\log\mathcal{C}(1+\epsilon)^2}{\epsilon^4}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(\chi_n^{2,(b)})^2\right]\Bigg\}. \end{aligned} \tag{43}$$

Proof. Equation (38) verifies the conditional-mean hypothesis of Lemma 1. The update-moment bound eq. (39) contributes

$$T^2RL(2\log|\mathcal{A}|+1)\eta \cdot \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T} = 8T(1+\epsilon)L^2R^3\log|\mathcal{A}|(2\log|\mathcal{A}|+1)\eta.$$

For the bias term, conditional Jensen and eq. (40) give

$$T^2\mathbb{E}\left\|b_n^{\mathrm{TIC}}\right\|_2^2 \le T^2\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\left\|b_{n,b}^{\mathrm{TIC}}\right\|_2^2\right] \le \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\Gamma_{n,b}^{\mathrm{TIC}}\right].$$

Substituting both estimates into Lemma 1 gives

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla J(\theta_n)\right\|_2^2 \le \frac{2T(J^\star-\mathbb{E}[J(\theta_0)])}{\eta N} + 8T(1+\epsilon)L^2R^3\log|\mathcal{A}|(2\log|\mathcal{A}|+1)\eta + \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\Gamma_{n,b}^{\mathrm{TIC}}\right].$$

This proves eq. (42).

Now additionally impose the delay-regularity condition in Assumption 3. Fix $t \ge 0$ with $\Pr(\tau_n^{(b)} = t) > 0$. The learner recursion and Cauchy–Schwarz over the t updates give

$$\theta_n - \theta_{n-t} = \eta\sum_{s=n-t}^{n-1} g_s^{\mathrm{TIC}}, \qquad \|\theta_n-\theta_{n-t}\|_2^2 \le \eta^2 t\sum_{s=n-t}^{n-1}\left\|g_s^{\mathrm{TIC}}\right\|_2^2.$$

This condition removes the conditioning on the realized delay, and eq. (39) bounds every update moment. Therefore

$$\mathbb{E}[\|\theta_n-\theta_{n-t}\|_2^2 \mid \tau_n^{(b)}=t] = \mathbb{E}\|\theta_n-\theta_{n-t}\|_2^2 \le \frac{8(1+\epsilon)LR^2\log|\mathcal{A}|}{T}\eta^2t^2,$$
$$\mathbb{E}[D_n^{(b)} \mid \tau_n^{(b)}=t] \le \frac{TL}{2}\mathbb{E}[\|\theta_n-\theta_{n-t}\|_2^2 \mid \tau_n^{(b)}=t] \le 4(1+\epsilon)L^2R^2\log|\mathcal{A}|\,\eta^2t^2.$$

Under coverage, every retained ratio is at most $1+\epsilon$, and eq. (15) gives, for every group,

$$\left\|U_{s,b}^{\mathrm{TIC}}\right\|_2 \le \frac{2R}{TG}\sum_{i=1}^{G}\rho_s^{(b,i)}\mathbf{1}\{\rho_s^{(b,i)}\le 1+\epsilon\}\left\|\nabla\log\pi_{\theta_s}(y^{(b,i)}\mid x^{(b)})\right\|_2 \le 2R(1+\epsilon)\sqrt{2L\log\mathcal{C}}.$$

The same bound holds for g_s^{TIC}, because it is the average of the group updates. The triangle inequality along the learner path thus yields

$$\|\theta_n-\theta_{n-t}\|_2 \le \eta\sum_{s=n-t}^{n-1}\left\|g_s^{\mathrm{TIC}}\right\|_2 \le 2R(1+\epsilon)\sqrt{2L\log\mathcal{C}}\,\eta t.$$

Combining this pathwise estimate with eq. (17) gives

$$D_n^{(b)} \le \frac{TL}{2}\|\theta_n-\theta_{n-t}\|_2^2 \le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2t^2 =: D_{\max},$$
$$\mathbb{E}[(D_n^{(b)})^2 \mid \tau_n^{(b)}=t] \le 4(1+\epsilon)^2TR^2L^2\log\mathcal{C}\,\eta^2t^2\mathbb{E}[D_n^{(b)}\mid\tau_n^{(b)}=t] \le 16(1+\epsilon)^3TL^4R^4\log|\mathcal{A}|\log\mathcal{C}\,\eta^4t^4.$$

The middle inequality uses the almost-sure upper bound on $D_n^{(b)}$, namely $(D_n^{(b)})^2 \le D_{\max}D_n^{(b)}$. For example, substituting the last line into the quadratic-KL branch gives

$$\frac{8T^2R^2L\log\mathcal{C}}{c_\epsilon^2}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(D_n^{(b)})^2\right] \le \frac{128(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{c_\epsilon^2}\eta^4\overline{\tau}_4.$$

The expected-drift line gives the second branch of eq. (43); the uniform branch is constant, and the empirical chi-square branch is left in its averaged form. The average pointwise minimum is no larger than any of these four branchwise averages, so their minimum proves eq. (43).

□

H From the appendix bounds to the main-text rates

This section derives the main-text rates from the detailed bounds above. We first prove a step-size balancing lemma for the optimization, stochastic, and fourth-order delay terms, then apply it to the GMC bound. We next optimize the GMC threshold asymptotically and use the local-overlap estimate to obtain the group-size refinement. Applying the same balancing lemma to TIC completes the main-text guarantees and makes their different delay-dependent coefficients explicit.

Lemma 10 (Balancing optimization, stochastic, and delay terms)**.** *Let $A, B > 0$, $C \geq 0$, and $M \geq 1$. Suppose that, for every $\eta > 0$,*

$$F_M(\eta) \leq \frac{A}{M\eta} + B\eta + C\eta^4, \tag{44}$$

then, with $(A/(CM))^{1/5} := +\infty$ when $C = 0$, the choice

$$\eta^\star := \min\left\{\sqrt{\frac{A}{BM}}, \left(\frac{A}{CM}\right)^{1/5}\right\}, \tag{45}$$

satisfies

$$F_M(\eta^\star) \leq 2\sqrt{\frac{AB}{M}} + 2\left(\frac{A^4C}{M^4}\right)^{1/5}. \tag{46}$$

Proof. If $C = 0$, substituting $\eta^\star = \sqrt{A/(BM)}$ directly gives $F_M(\eta^\star) \leq 2\sqrt{AB/M}$. Suppose henceforth that $C > 0$. The definition eq. (45) implies

$$\frac{A}{M\eta^\star} = \max\left\{\sqrt{\frac{AB}{M}}, \left(\frac{A^4C}{M^4}\right)^{1/5}\right\} \leq \sqrt{\frac{AB}{M}} + \left(\frac{A^4C}{M^4}\right)^{1/5},$$
$$B\eta^\star \leq \sqrt{\frac{AB}{M}}, \qquad C(\eta^\star)^4 \leq \left(\frac{A^4C}{M^4}\right)^{1/5},$$

where the equality uses $1/\min\{u, v\} = \max\{1/u, 1/v\}$ and the final inequality uses $\max\{a, b\} \leq a + b$ for $a, b \geq 0$. Summing the three displayed bounds proves eq. (46). □

Proof of Theorem 1. Retain the quadratic-KL branch in eq. (33). With $M = N$, the result has the form eq. (44) with

$$A = 2T\big(J^\star - \mathbb{E}[J(\theta_0)]\big),$$
$$B = 8T(1+\epsilon)L^2R^3 \log|\mathcal{A}|(2\log|\mathcal{A}| + 1),$$
$$C = \frac{128(1+\epsilon)^3L^5R^6T^3 \log|\mathcal{A}|(\log\mathcal{C})^2}{\log^2(1+2\epsilon)}\overline{\tau}_4.$$

This is the first bound in Theorem 1. Assume first that $A > 0$ and $\overline{\tau}_4 > 0$. Lemma 10 gives

$$\eta^\star = \min\left\{\sqrt{\frac{A}{BN}}, \left(\frac{A}{CN}\right)^{1/5}\right\},$$
$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla J(\theta_n)\right\|_2^2 \leq 2\sqrt{\frac{AB}{N}} + 2\left(\frac{A^4C}{N^4}\right)^{1/5}.$$

Substituting C and suppressing the fixed problem constants gives

$$\eta^\star \asymp \min\left\{\frac{1}{\sqrt{(1+\epsilon)N}}, \left(\frac{\log^2(1+2\epsilon)}{(1+\epsilon)^3\overline{\tau}_4 N}\right)^{1/5}\right\},$$

and the preceding optimized bound gives the second claim of Theorem 1. When $\overline{\tau}_4 = 0$ and $A > 0$, the $C = 0$ branch of Lemma 10 instead gives $\eta^\star = \sqrt{A/(BN)}$ and the first term of the optimized bound. If $A = 0$, choose any positive step size of the main-text order and substitute it directly into eq. (44); the optimization term is then zero and the same stated upper order follows. □

Proof of Corollary 1. Suppressing fixed positive problem-dependent coefficients, the leading ϵ-dependent terms in the optimized GMC bound have the form

$$\mathcal{B}_N(\epsilon) = aN^{-1/2}\sqrt{1+\epsilon} + b\overline{\tau}_4^{1/5}N^{-4/5}(1+\epsilon)^{3/5}\log^{-2/5}(1+2\epsilon),$$

for constants $a, b > 0$ independent of N, $\overline{\tau}_4$, and ϵ. The first term increases with ϵ, while the second diverges as $\epsilon \downarrow 0$ and captures the correction-bias penalty. For each N, $\mathcal{B}_N$ is continuous on $(0, \infty)$ and diverges at both endpoints, so it attains a global minimum in $(0, \infty)$.

Let $h(\epsilon) := (1+\epsilon)^{3/5} \log^{-2/5}(1+2\epsilon)$. Since $\log(1+2\epsilon) \sim 2\epsilon$ as $\epsilon \downarrow 0$,

$$-h'(\epsilon) = \frac{h(\epsilon)}{5}\left(\frac{4}{(1+2\epsilon)\log(1+2\epsilon)} - \frac{3}{1+\epsilon}\right) \asymp \epsilon^{-7/5} \qquad (\epsilon \downarrow 0).$$

With $\overline{\tau}_4 > 0$ fixed, every minimizer tends to zero: at $\epsilon = N^{-1/5}$, $\mathcal{B}_N(\epsilon) = aN^{-1/2} + o(N^{-1/2})$, whereas for any fixed $\epsilon_0 > 0$ and $\epsilon \geq \epsilon_0$, $\mathcal{B}_N(\epsilon) \geq a\sqrt{1+\epsilon_0}\, N^{-1/2}$. At any such minimizer, the first-order condition and $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\sqrt{1+\epsilon} \asymp 1$ then give

$$N^{-1/2} \asymp \overline{\tau}_4^{1/5} N^{-4/5} (\epsilon^\star_{\mathrm{GMC}})^{-7/5},$$

or, equivalently,

$$(\epsilon^\star_{\mathrm{GMC}})^{7/5} \asymp \overline{\tau}_4^{1/5} N^{-3/10}.$$

Taking both sides to the power $5/7$ yields

$$\epsilon^\star_{\mathrm{GMC}} = \Theta\left(\overline{\tau}_4^{1/7} N^{-3/14}\right).$$

□

Proof of Corollary 2. Corollary 3 gives $(\chi_n^{2,(b)})^2 \leq 4(1+\delta)^2 (D_n^{(b)})^2$. Therefore the chi-square branch in eq. (31) and eq. (30) imply

$$\begin{aligned}\frac{T^2R^2L\log\mathcal{C}}{2G^2\epsilon^2}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(\chi_n^{2,(b)})^2\right] &\leq \frac{2(1+\delta)^2T^2R^2L\log\mathcal{C}}{G^2\epsilon^2}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(D_n^{(b)})^2\right] \\ &\leq \frac{32(1+\delta)^2(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{G^2\epsilon^2}\eta^4\overline{\tau}_4.\end{aligned}$$

Substitution in eq. (32) proves the first displayed bound in the corollary. Lemma 10, with C equal to the complete coefficient of η^4 on the last line, gives the stated step size and optimized rate. □

Proof of Theorem 2. Retain the quadratic-KL branch in eq. (43). With $M = N$, the result has the form eq. (44) with the same A and B as in the proof of Theorem 1, and

$$C = \frac{128(1+\epsilon)^3L^5R^6T^3\log|\mathcal{A}|(\log\mathcal{C})^2}{c_\epsilon^2}\overline{\tau}_4.$$

This is the first bound in Theorem 2. Assume first that $A > 0$ and $\overline{\tau}_4 > 0$. Lemma 10 gives

$$\eta^\star = \min\left\{\sqrt{\frac{A}{BN}}, \left(\frac{A}{CN}\right)^{1/5}\right\},$$

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla J(\theta_n)\right\|_2^2 \leq 2\sqrt{\frac{AB}{N}} + 2\left(\frac{A^4C}{N^4}\right)^{1/5}.$$

Treating the objective gap and $T, L, R, |\mathcal{A}|, \mathcal{C}$ as fixed constants reduces the first line to

$$\eta^\star \asymp \min\left\{\frac{1}{\sqrt{(1+\epsilon)N}}, \left(\frac{c_\epsilon^2}{(1+\epsilon)^3\overline{\tau}_4 N}\right)^{1/5}\right\},$$

and the second line is exactly the optimized order stated in the theorem. When $\overline{\tau}_4 = 0$ and $A > 0$, the $C = 0$ branch of Lemma 10 instead gives $\eta^\star = \sqrt{A/(BN)}$ and the first term of the optimized bound. If $A = 0$, choose any positive step size of the main-text order and substitute it directly into eq. (44); the optimization term is then zero and the same stated upper order follows. □