Hölder Policy Optimisation

Yuxiang Chen[1,*] **Dingli Liang**[1,*] **Yihang Chen**[1,*] **Ziqin Gong**[3] **Chenyang Le**[2] **Zhaokai Wang**[2]
Jiachen Zhu[2] **Lingyu Yang**[2] **Jianghao Lin**[2] **Weinan Zhang**[2] **Jun Wang**[1,†]

Abstract

Group Relative Policy Optimisation (GRPO) enhances large language models by estimating advantages across a group of sampled trajectories. However, mapping these trajectory-level advantages to policy updates requires aggregating token-level probabilities within each sequence. Relying on a fixed aggregation mechanism for this step fundamentally limits the algorithm's adaptability. Empirically, we observe a critical trade-off: certain fixed aggregations frequently suffer from training collapse, while others fail to yield satisfactory performance. To resolve this, we propose **HölderPO**, a generalised policy optimisation framework unifying token-level probability aggregation via the Hölder mean. By explicitly modulating the parameter p, our framework provides continuous control over the trade-off between gradient concentration and variance bounds. Theoretically, we prove that a larger p concentrates the gradient to amplify sparse learning signals, whereas a smaller p strictly bounds gradient variance. Because no static configuration can universally resolve this concentration-stability trade-off, we instantiate the framework with a dynamic annealing algorithm that progressively schedules p across the training lifecycle. Extensive evaluations demonstrate superior stability and convergence over existing baselines. Specifically, our approach achieves a state-of-the-art average accuracy of 54.9% across multiple mathematical benchmarks, yielding a substantial 7.2% relative gain over standard GRPO and secures an exceptional 93.8% success rate on ALFWorld.

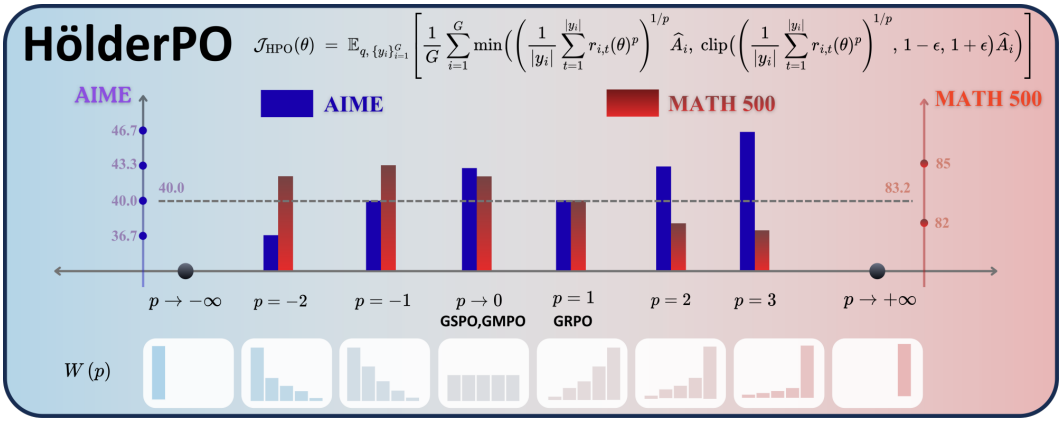

Figure 1: **HölderPO unifies token-level aggregation under a single parameter p.** The objective at the top generalises GRPO by replacing its arithmetic mean over token-level importance ratios with the Hölder mean of order $p \in \mathbb{R}$, recovering GRPO ($p = 1$) and GMPO/GSPO ($p \to 0$) as special cases. The bar chart reports accuracy on AIME24 (blue, sparse signal) and MATH500 (red, dense signal), with dashed lines marking GRPO baselines. Bottom: the token weight distribution $W(p)$, with each panel ordering tokens from small (left) to large (right) importance ratio.

[*]Equal contribution. [†]Corresponding author: `jun.wang@cs.ucl.ac.uk`. Code available at `https://github.com/YihangChen9/HolderPO`. [1]University College London, London, United Kingdom. [2]Shanghai Jiao Tong University, Shanghai, China. [3]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.

Preprint.

1 Introduction

Reinforcement Learning (RL) has emerged as a key technique for advancing the alignment and complex reasoning capabilities of Large Language Models (LLMs) [Ouyang et al., 2022, Schulman et al., 2017]. Recently, Group Relative Policy Optimisation (GRPO) has emerged as a highly effective and compute-efficient algorithm, largely driving the success of reasoning models like DeepSeek-R1 [Shao et al., 2024]. GRPO operates by estimating advantages across a group of sampled trajectories, substantially reducing training overhead by eliminating the need for an external critic model. However, mapping these trajectory-level advantages to policy updates requires aggregating token-level probabilities within each sequence. As the demand for solving long-horizon reasoning tasks grows, the fundamental mechanics of this fixed aggregation step have come under scrutiny [Liu et al., 2025]. Existing algorithms rigidly rely on static aggregation functions: standard GRPO ($p = 1$) defaults to the Arithmetic Mean, while recent variants like GMPO [Zhao et al., 2025] and GSPO [Zheng et al., 2025] ($p \to 0$) attempt to mitigate variance by employing the Geometric Mean.

Despite their empirical success, these fixed aggregation mechanisms implicitly impose a static optimisation landscape, limiting their adaptability across long-horizon reasoning tasks of varying signal density — the regime in which the trade-off we identify becomes acute. Through empirical investigation, we observe a critical trade-off: certain fixed aggregations frequently suffer from training collapse, while others fail to yield satisfactory performance. Specifically, on **dense-signal tasks** (where supervision is distributed across many tokens, e.g., MATH [Hendrycks et al., 2021]), standard GRPO ($p = 1$) disproportionately over-weights minor token-level errors, inducing high-variance gradient updates that can lead to training collapse. Conversely, on **sparse-signal tasks** (where correct reasoning is concentrated in rare, high-magnitude tokens, e.g., AIME [Jia et al., 2024]), GSPO ($p \to 0$) overly smooths the probability ratios, suppressing the effective use of these rare "aha moments". Figure 1 visualises this divergence: AIME24 accuracy peaks at $p = 3$ while MATH500 peaks at $p = -1$, with the bottom row showing how the underlying token weight distribution $W(p)$ deforms across the p-axis. Essentially, there is no "silver bullet" among static mean functions; the optimal probability aggregation is not a constant, but rather a function of task signal density and the model's training progression.

To address these fundamental limitations, we propose **HölderPO**, a generalised policy optimisation framework unifying token-level probability aggregation via the adaptable Hölder mean (p-norm). By explicitly modulating the parameter p, the framework provides continuous control over the trade-off between gradient concentration and variance bounds. Theoretically, we prove a two-sided trade-off in p: a larger p concentrates the gradient weight distribution on a small subset of tokens, amplifying the effective use of rare informative learning signals at the cost of looser variance bounds. Conversely, a smaller p strictly tightens the variance of the policy gradient estimator, ensuring training stability at the cost of weakening the response to those same sparse signals. Because no static configuration can simultaneously realise both endpoint advantages, we instantiate the HölderPO framework with a dynamic annealing algorithm. By progressively scheduling p from a higher positive value to a negative value during training, this algorithm seamlessly transitions the model from aggressive signal amplification in the early stages to variance-controlled convergence in the later stages.

Extensive empirical evaluations across a comprehensive suite of complex reasoning and decision-making benchmarks strongly validate our claims. Built upon the **Qwen2.5-Math-7B** base [Yang et al., 2024], our ablation studies first confirm the task-specific sensitivity of p: sparse-signal tasks strictly favour higher p values for aggressive signal amplification, whereas dense-signal tasks benefit from lower (possibly negative) p values for gradient stability. Crucially, when explicitly setting $p = 3$, our approach effectively breaks the existing performance ceiling on the highly challenging AIME benchmark, surpassing the previous 43.3% accuracy record to achieve 46.7%. Building on these insights, by employing our dynamic annealing algorithm, HölderPO unifies these advantages without incurring additional computational overhead. Consequently, our approach achieves a state-of-the-art average accuracy of **54.9%** across five mathematical benchmarks (AIME, AMC, MATH, Minerva [Lewkowycz et al., 2022], and OlympiadBench [He et al., 2024]), a 7.2% relative gain over standard GRPO that surpasses concurrent token-aggregation methods including PMPO [Zhao et al., 2026]. Beyond mathematical reasoning, this dynamic adaptability extends to open-world agentic tasks, securing an exceptional **93.8%** success rate on the ALFWorld benchmark [Shridhar et al., 2020], a 28.8% relative gain over GRPO (72.8%).

In summary, our main contributions are as follows:

- **The HölderPO Framework:** We propose HölderPO, a generalised policy optimisation framework that dynamically unifies various mean-based probability aggregations through the adaptable Hölder parameter p.
- **Theoretical Foundation:** We theoretically characterise the two-sided role of p in long-horizon reasoning: a larger p concentrates gradient weight to amplify sparse learning signals, whereas a smaller p strictly bounds gradient variance to ensure training stability. No fixed p realises both endpoint advantages simultaneously, motivating dynamic scheduling.
- **Empirical Breakthroughs and SOTA Performance:** Empirically, explicitly employing a large $p = 3$ breaks the existing performance ceiling on the highly challenging AIME benchmark. Furthermore, instantiating the framework with a dynamic p-annealing algorithm achieves state-of-the-art results, securing a 54.9% average accuracy across five mathematical benchmarks and an exceptional 93.8% success rate on ALFWorld agentic tasks.

2 Related Work

Reinforcement Learning for Complex Reasoning. Reinforcement Learning (RL) has become the cornerstone of LLM post-training. While foundational work used RLHF for behavioural alignment [Ouyang et al., 2022, Stiennon et al., 2020], recent advances focus on complex reasoning via RLVR [Wen et al., 2025], pioneered by OpenAI o-series [Jaech et al., 2024] and DeepSeek-R1 [Guo et al., 2025, Shao et al., 2024], inspiring both proprietary [Comanici et al., 2025, Yang et al., 2025a] and open-source successors. GRPO [Shao et al., 2024] has emerged as the dominant algorithm; its broader ecosystem of refinements is surveyed in Appendix A.

Token-Level Aggregation. The aggregation operator that maps token-level importance ratios to a sequence-level signal is the most direct analogue of our framework. GRPO uses the arithmetic mean, while GMPO [Zhao et al., 2025] and GSPO [Zheng et al., 2025] adopt the geometric mean to mitigate outlier variance. Concurrent PMPO [Zhao et al., 2026] parameterises a power-mean exponent $p \in [0, 1]$, adapted per-trajectory via clip-aware ESS matching. Our framework differs in two key respects: (i) we extend p to the *full real range*, identifying $p < 0$ as a qualitatively distinct *inverse-concentration* phase unexplored by prior work; and (ii) we adapt p along the *temporal* axis (across training steps) rather than per trajectory, enabling complementary roles for early-stage signal amplification and late-stage variance contraction.

Token Reweighting via Auxiliary Signals. A parallel line reweights tokens within each rollout using signals *external* to the importance ratio: token entropy [Wang et al., 2025a, Yu and Li, 2026, Simoni et al., 2025], token probability [Yang et al., 2025b], hidden contributions to response confidence [Deng et al., 2025], or selective KL masking [Lin et al., 2025]. These approaches are orthogonal to ours and could in principle be combined with HölderPO's power-mean aggregation.

3 HölderPO: A Generalised Aggregation Framework

When adapting PPO for LLMs, particularly for training long-horizon reasoning tasks, group-based variants like GRPO [Shao et al., 2024] formulate the unclipped objective as

$$\mathcal{J}(\theta) = \mathbb{E}_{x,\{y_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \rho_i(\theta) \widehat{A}_i \right].$$

Here, $\rho_i(\theta)$ is the sequence-level surrogate term, which can be regarded as an *aggregation operator*—a functional projection that compresses the full sequence of token-level importance ratios $\{r_{i,t}(\theta)\}_{t=1}^{|y_i|}$ into a well-behaved sequence-level scalar. While GRPO uses the arithmetic mean, GMPO [Zhao et al., 2025] and GSPO [Zheng et al., 2025] use geometric mean. However, these methods only represent static, isolated points within a broader, continuous spectrum of aggregation operators.

In section 3.1, we propose **Hölder Policy Optimisation**, a generalised framework that parameterises the aggregation operators by a single scalar $p \in \mathbb{R}$ via the Hölder mean. Pivotally, the single parameter p governs a trade-off between **gradient concentration** (defined in Section 3.2), which selectively amplifies targeted learning signals, and the **variance bound** (analysed in Section 3.3), which ensures training stability. Finally, the interplay between these two competing properties motivates our **dynamic scheduling strategy** in Section 3.4.

3.1 Aggregation via the Hölder Mean

Given a prompt context x and a rollout y_i sampled from $\pi_{\theta_{\text{old}}}$, the token-level importance ratio for t-th token is $r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x,y_{i,<t})}$. Rather than relying on a fixed operator, HölderPO generalises the token-level aggregation by the Hölder mean of order p:

$$\rho_{i,p}(\theta) = \begin{cases} \left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p\right)^{1/p}, & \text{if } p \neq 0, \\ \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log r_{i,t}(\theta)\right), & \text{if } p = 0. \end{cases} \tag{1}$$

Due to the limit for $p \to 0$, we take the geometric mean for $p = 0$ branch (see Appendix G.4). The HölderPO objective then takes the standard PPO-style form with sequence-level clipping:

$$\mathcal{J}_{H_s}(\theta) = \mathbb{E}_{x,\,\{y_i\}_{i=1}^{G}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_{i,p}(\theta)\widehat{A}_i,\ \text{clip}\big(\rho_{i,p}(\theta),\,1-\epsilon,\,1+\epsilon\big)\widehat{A}_i\Big)\right]. \tag{2}$$

Here $\widehat{A}_i$ is the advantage estimator and ϵ is the clipping threshold. The reason we choose sequence-level clipping is to control gradient variance (see Appendix D and I.2). Specifically, $p = 1$ recovers GRPO (Appendix G.2), while $p = 0$ recovers GSPO (Appendix G.3). To analyse how p shapes the optimisation, we study $\nabla_\theta \rho_{i,p}(\theta)$, which governs the direction of the policy gradients (see Eq. (9), (13), (16)). A direct calculation (Appendix G.1) yields

$$\nabla_\theta \rho_{i,p}(\theta) = \rho_{i,p}(\theta)\sum_{t=1}^{|y_i|} W_{i,t}(p)\cdot\nabla_\theta \log \pi_\theta(y_{i,t}\mid x, y_{i,<t}) \quad W_{i,t}(p) \coloneqq \frac{r_{i,t}(\theta)^p}{\sum_{k=1}^{|y_i|} r_{i,k}(\theta)^p}, \tag{3}$$

where the per-token gradient weights $W_{i,t}(p)$ form a probability distribution denoted by W_i^p. Crucially, varying p does not alter the per-token log-gradient directions; instead, it solely reweights the directions and modulates the weight distribution.

3.2 Distributional Deformation and Gradient Concentration

We formalise the gradient concentration by analysing W_i^p through two complementary lenses. Locally, Theorem 5 (Appendix H.1) shows an increasingly strict token-level weight allocation: as p grows, maximal-ratio tokens monotonically dominate. Non-maximal ones may briefly gain weight before strictly decaying to zero once the rising threshold $\mu_i(p)$ surpasses their log-ratios. Globally, our next result (Appendix H.2) captures the dispersion of the entire weight distribution by Shannon entropy.

Theorem 1. *Assume the sequence y_i contains at least two tokens with distinct importance ratios. Then Shannon entropy of the weight distribution attains its global maximum at $p = 0$, where $W_i^0 = \frac{1}{|y_i|}\text{Unif}$, and strictly decreases as $|p|$ increases. Moreover, as $p \to \pm\infty$, W_i^p concentrates uniformly on the subset $\mathcal{T}^+ = \arg\max_t r_{i,t}(\theta)$ and $\mathcal{T}^- = \arg\min_t r_{i,t}(\theta)$, respectively.*

Together, these dual perspectives formally characterise gradient concentration—the skewing of the weight distribution toward a specific subset of tokens. By governing the intensity and target of this skew, p shapes the gradient contributions in three distinct regimes:

Upward Concentration ($p > 0$). A positive p drives the gradient concentration toward tokens with relatively high importance ratios. A prevailing view suggests that RL for reasoning primarily acts to sharpen the pre-existing knowledge distribution of the base model (e.g., Zhou et al. [2023], Li et al. [2024], Yue et al. [2025]). Under this view, an importance ratio > 1 serves as a confidence signal that, ideally, highlights the critical bottleneck tokens within reasoning steps. In long-horizon tasks, where such high-confidence tokens are sparse [Zelikman et al., 2022, Lightman et al., 2023, Yao et al., 2023], setting $p > 0$ explicitly amplifies their weight to prevent their gradients from being diluted.

Uniform Dispersion ($p \to 0$). As p decreases, the specific contributions of individual tokens are increasingly flatten out. At $p = 0$, every token contributes equally.

Downward Concentration ($p < 0$). A negative p inverts the gradient allocation, aggressively upweighting tokens with importance ratios < 1, which signal current model's hesitation and pinpoint

unconventional yet effective decision points in successful trajectories. Consequently, a moderately negative p promotes reasoning diversity by forcing the model to consolidate alternative pathways. More details about the relation between our gradient concentration mechanism and exploration-exploitation trade-off can be found in Appendix H.3.

3.3 Policy Gradient Variance Bound

Next, we analyse the variance of the policy gradient estimator induced by (2). In long-horizon reasoning, while concentration enables the amplification of targeted signals, it risks magnifying gradient variance. The next theorem (proof is in Appendix I.2) shows that such selectivity can destabilise convergence if left uncontrolled.

Theorem 2. *Let $\widehat{\nabla}_\theta \mathcal{J}_{H_s}$ (Eq. (17)) denote the unbiased mini-batch estimator induced by (2). Assume $\|\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\| \le M$ for all tokens within the batch, the variance admits the bound*

$$\|\mathrm{Var}(\widehat{\nabla}_\theta \mathcal{J}_{H_s})\| \;\le\; \frac{M^2}{B}\,\mathbb{E}\Big[\widehat{A}_i^2\,\rho_{i,p}^2(\theta)\Big], \tag{4}$$

which is monotonically increasing in p for all $p \in \mathbb{R}$, where B is the batch size.

In addition, if we assume approximate orthogonality of gradients of tokens within sequences (Assumption 1), we prove the variance itself has a global minimum at some $p^* \le 0$. (Theorem 7).

Trade-off with concentration. Theorems 1 and 2 highlight a structural trade-off controlled by the scalar p: driving p upward isolates targeted pivotal signals, but incurs the cost of a looser variance bound. While shifting p downward strictly tightens this bound, it dilutes these critical signals or redirects the concentration entirely. In long-horizon reasoning, this trade-off becomes a bottleneck: we must amplify sparse signals without letting variance scale uncontrollably across the entire trajectory. Therefore, **no fixed p can be uniformly optimal**, since the optimal balance between these two requirements varies depending on the specific task and training stage.

3.4 A Dynamic p-Scheduling Strategy

The trade-off above motivates a dynamic schedule for long-horizon reasoning tasks that monotonically decays p from a positive initial value p_{high} to a low (possibly negative) terminal value p_{low} over the course of training: $p(0) = p_{\text{high}}, p(T) = p_{\text{low}}$, and $p(t_1) \ge p(t_2) \;\; \forall\, 0 \le t_1 < t_2 \le T$. The early phase leverages positive concentration to amplify sparse, high-magnitude signals signals crucial for initial policy improvement. In the late phase, the schedule focuses on contracting the variance bound to guarantee stable convergence. Where $p_{\text{low}} < 0$, the algorithm utilises inverse concentration, moderately redirecting the gradient towards underemphasised tokens to foster reasoning diversity.

Theorem 3. *Let $V(p) := \mathbb{E}[\widehat{A}_i^2\,\rho_{i,p}^2(\theta)]$ denote the term in the bound in Eq.(4), and let $p_{stat} \in [p_{low}, p_{high}]$ be any fixed parameter. Given a y_i of length n, the dynamic schedule satisfies:*

1. Early-phase signal amplification: *If y_i has a high-ratio token t^* with $r_{i,t^*} \gg 1$, while the other tokens have constant-bounded ratios. Under the pre-saturation condition $r_{i,t^*}^{p_{high}} \ll n - 1$, shifting from p_{stat} to p_{high} exponentially amplifies its gradient weight: there exists a constant $C > 0$ such that*

$$\frac{W_{i,t^*}(p_{high})}{W_{i,t^*}(p_{stat})} \;\ge\; C \cdot r_{i,t}^{\,p_{high} - p_{stat}}. \tag{5}$$

2. Late-phase variance contraction: *The terminal variance bound is strictly contracted:*

$$V(p_{low}) \;<\; V(p_{stat}). \tag{6}$$

This theorem (proof in Appendix J) reveals that any static parameter p_{stat}, the standard paradigm in current GRPO-based methods, is a compromise for long-horizon reasoning tasks: it must sacrifice either early-stage signal amplification (if p is low) or late-stage variance control (if p is high). Our schedule bypasses the dilemma, dynamically allocating required mechanism to each training phase.

Figure 2 provides direct visual support for this choice: the per-step ratio envelopes under static $p \in \{+2, 0, -2\}$ illustrate how decreasing p monotonically tightens the gap between the largest and smallest token-level ratios, and our linear schedule inherits the early-stage concentration of $p = +2$ while converging to the controlled regime of $p = -2$.

4 Experiment

To empirically validate the effectiveness of HölderPO, we evaluate our method against state-of-the-art policy optimisation baselines on mathematical reasoning and agentic benchmarks. Our experiments are designed to follow a clear logical progression: (1) revealing the task-specific sensitivity of the p parameter on distinct benchmarks, (2) demonstrating how dynamic scheduling resolves the concentration–stability trade-off identified in Section 3, and (3) comparing our overall performance against established baselines.

4.1 Implementation Details

Model. We evaluate our framework on two task families: mathematical reasoning and agentic decision-making. For mathematical reasoning, following Dr.GRPO [Liu et al., 2025], we cover a broad spectrum of base models ranging from 1.5B to 8B parameters, including the Qwen2.5-Math series (1.5B and 7B) [Yang et al., 2024], DeepSeek-R1-Distill-Qwen-7B [Guo et al., 2025], and the Qwen3 series (4B and 8B) [Yang et al., 2025a]. For agentic tasks, we adopt Qwen2.5-1.5B-Instruct [Qwen et al., 2025] as the policy backbone.

Training. Our training pipeline follows two established protocols depending on the task. For mathematical reasoning, we adopt the recipe of Dr.GRPO [Liu et al., 2025]: training data consists of 8,523 problems from MATH [Hendrycks et al., 2021] (Levels 3–5), and each prompt is paired with 8 sampled rollouts capped at 3,000 tokens. Within each RL round, $\pi_{\theta_{\text{old}}}$ produces 1,024 trajectories, after which the current policy π_θ is refreshed 8 times using a mini-batch size of 128. For agentic tasks, we adhere to the GiGPO protocol [Feng et al., 2025] for both training and evaluation on ALFWorld. In terms of compute, all models are trained on 4×H100 GPUs. We primarily compare HölderPO against GRPO [Shao et al., 2024], Dr.GRPO [Liu et al., 2025], and GMPO [Zhao et al., 2025] under matched configurations.

Evaluation. We report mathematical performance on five benchmarks that span a wide difficulty range. AIME24 contains 30 olympiad-level problems drawn from the 2024 American Invitational Mathematics Examination, while AMC provides 83 competition problems of intermediate difficulty. MATH500 is a 500-problem subset of MATH covering algebra, geometry, and number theory. Minerva [Lewkowycz et al., 2022] consists of 272 graduate-level problems that demand multi-step derivations, and OlympiadBench (Oly.) [He et al., 2024] collects 675 high-difficulty olympiad problems. For agentic evaluation, we use the six ALFWorld [Shridhar et al., 2020] sub-task categories, namely Pick, Look, Clean, Heat, Cool, and Pick Two. Following Dr.GRPO [Liu et al., 2025], we adopt Pass@1 as the primary metric for mathematical tasks and decode greedily with temperature 0.0, generating one sample per question. For ALFWorld, we report task success rate under the given standard evaluation protocol.

4.2 Task-Specific Sensitivity of p

A fundamental premise of our work is that a static aggregation function cannot optimally solve all tasks. To illustrate this, we isolate the performance of HölderPO across different static p values on two benchmarks with distinct signal-density profiles: AIME24, where correct reasoning is concentrated in a small number of rare, high-magnitude tokens (sparse-signal regime), and MATH500, where supervision is more densely distributed across many tokens (dense-signal regime).

As detailed in Table 1 and visually summarised by the diverging performance curves in Figure 1, the optimal p value diverges significantly across the two regimes.

Sparse-signal tasks favour high p. On AIME24, where correct reasoning traces (i.e., pivotal reasoning steps) are exceptionally sparse, larger positive values of p (e.g., $p \geq 2$) yield the highest accuracy. This empirically confirms Theorem 1: in the positive-concentration regime, the gradient weight distribution concentrates on tokens with the largest importance ratios (as visually depicted by the right-skewed $W(p)$ distributions at the bottom of Figure 1), allowing the rare, high-quality reasoning steps to drive the update rather than being averaged out by the bulk of unremarkable tokens.

Dense-signal tasks favour low p. Conversely, on MATH500, where supervision is distributed across many tokens, lower values of p (e.g., $p \leq 0$) perform better. This is consistent with Theorem 2: decreasing p tightens the variance bound on the policy gradient estimator, preventing the high-

Training Objectives	AIME24 (Sparse-Signal)	MATH500 (Dense-Signal)
GRPO ($p = 1$)	40.0	83.4
GMPO ($p \to 0$)	43.3	82.0
HölderPO ($p = -2$)	36.7	84.6
HölderPO ($p = -1$)	36.7	**85.0**
HölderPO ($p \to 0$)	43.3	84.6
HölderPO ($p = 1$)	40.0	83.2
HölderPO ($p = 2$)	43.3	82.0
HölderPO ($p = 3$)	**46.7**	81.8

Table 1: Performance on benchmarks with distinct signal-density profiles. On AIME24, higher p amplifies rare high-magnitude signals for complex reasoning. Conversely, on MATH500, a lower p strictly tightens the gradient variance bound to ensure training stability, yielding superior performance on simpler tasks.

variance updates that occur when relative-magnitude differences among many comparable tokens get over-weighted. This mechanism directly corresponds to the flatter, left-leaning W_i^p distributions shown in Figure 1, which systematically redistribute credit to underemphasised steps.

4.3 Main Performance and Dynamic Scheduling

The empirical observation that no single static p yields optimal performance universally directly motivates our dynamic scheduling approach. We hypothesise that any reasoning task effectively functions as a sparse-signal task during the early stages of training. At this point, the model has yet to internalise the correct reasoning patterns, thus requiring a high p for signal amplification. As the model masters the underlying logic, the task gradually transitions into a dense-signal regime, necessitating a low p to ensure stable convergence.

To validate this, we evaluate our dynamic annealing scheduler (employing a linear decay of p from 2 to -2) alongside the best static configuration and existing state-of-the-art baselines across a diverse suite of benchmarks.

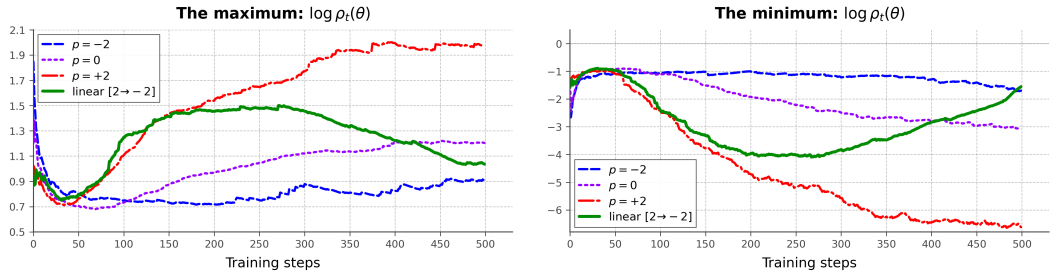

Figure 2: Token-level importance ratio $\log \rho_t(\theta)$ during training. **Left** and **Right** track the per-step upper and lower envelopes respectively. As p decreases, the upper envelope drops and the lower envelope rises, tightening the gap monotonically. Our decaying schedule $p: 2 \to -2$ (solid green) thus enables aggressive updates in the early stage and progressively converges to stable optimization in the later stage. Constant-p baselines ($p \in \{+2, 0, -2\}$) shown as dashed/dotted/dash-dotted.

Table 2 summarises the overall performance. While our best static configuration ($p \to 0$) achieves highly competitive average scores, it remains a single-point compromise on the concentration–stability trade-off. The dynamic p-scheduling approach achieves state-of-the-art results across the board: by progressively annealing p, the model leverages the early-stage signal amplification provided by $p = 2$, while benefiting from the strict variance contraction of $p = -2$ during the final convergence phase. The advantage is *distributional* rather than pointwise: tasks whose optimal p^* lies near a single endpoint may remain best served by the corresponding static configuration. For example, AIME24 favours static $p = 3$ and MATH500 favours $p = -1$. But the schedule strictly outperforms every static setting on the overall task average.

Training Objectives	AIME24	AMC	MATH500	Minerva	Oly.	Avg.
1.5B Models						
Base & Instruct Models						
Qwen2.5-Math-1.5B [Yang et al., 2024]	16.7	43.4	61.8	15.1	28.4	33.1
Qwen2.5-Math-1.5B-Instruct [Yang et al., 2024]	10.0	48.2	74.2	26.5	40.2	39.8
RL Post-Trained Models						
Oat-Zero-1.5B [Liu et al., 2025]	20.0	53.0	74.2	25.7	37.6	42.1
GMPO-1.5B [Zhao et al., 2025]	20.0	53.0	77.6	30.1	38.7	43.9
HölderPO-1.5B (Ours)	30.0	48.1	77.0	27.9	39.1	**44.5**
7B Models						
Base Models						
Qwen2.5-Math-7B [Yang et al., 2024]	16.7	38.6	50.6	9.9	16.6	26.5
RL Post-Trained Models						
SimpleRL-Zero-7B [Zeng et al., 2025]	26.7	60.2	78.2	27.6	40.3	46.6
PRIME-Zero-7B [Cui et al., 2025]	16.7	62.7	83.8	36.0	40.9	48.0
OpenReasoner-Zero-7B @ 3k [Hu et al., 2025]	13.3	47.0	79.2	31.6	44.0	43.0
OpenReasoner-Zero-7B @ 8k [Hu et al., 2025]	13.3	54.2	82.4	31.6	47.9	45.9
Eurus-7B [Yuan et al., 2024]	16.7	62.7	83.8	36.0	40.9	48.0
GPG-7B [Chu et al., 2025]	33.3	65.0	80.0	34.2	42.4	51.0
Oat-Zero-7B [Liu et al., 2025]	43.3	62.7	80.0	30.1	41.0	51.4
GRPO ($p = 1$) [Shao et al., 2024]	40.0	59.0	83.4	32.4	41.3	51.2
GMPO ($p \to 0$) [Zhao et al., 2025]	43.3	61.4	82.0	33.5	43.6	52.7
PMPO [Zhao et al., 2026]	36.7	68.7	83.8	34.9	**46.7**	54.2
HölderPO (Ours)						
HölderPO ($p = -2$)	36.7	53.0	84.6	33.5	44.7	50.5
HölderPO ($p = -1$)	40.0	59.0	**85.0**	33.8	42.1	52.0
HölderPO ($p \to 0$)	43.3	57.8	84.6	31.6	45.5	52.6
HölderPO ($p = 1$)	40.0	57.8	83.2	30.9	44.9	51.4
HölderPO ($p = 2$)	43.3	55.4	82.0	31.2	46.5	51.7
HölderPO ($p = 3$)	**46.7**	61.4	81.8	32.4	40.9	52.6
HölderPO (Linear Des: $2 \to -2$**)**	43.3	**68.7**	82.2	**34.9**	45.3	**54.9**
R1-Distill-Qwen-7B						
RL Post-Trained Models						
GRPO ($p = 1$) [Shao et al., 2024]	43.3	67.5	89.0	39.7	56.7	59.3
Dr.GRPO [Liu et al., 2025]	50.0	74.7	89.6	37.5	55.7	61.5
GMPO ($p \to 0$) [Zhao et al., 2025]	46.7	78.3	91.4	37.9	62.5	63.4
PMPO [Zhao et al., 2026]	46.7	79.5	93.4	39.3	64.2	64.6
HölderPO (Linear Des: $2 \to -2$**, Ours)**	53.3	79.5	92.6	42.3	64.1	**66.4**

Table 2: Comprehensive comparison of HölderPO against state-of-the-art baselines across different model scales and base architectures. The static rows report fixed p settings, while the dynamic row employs our linear annealing scheduler, which progressively decays p from an initial value of 2 to a terminal value of -2 over the course of training.

4.4 Selecting the Schedule Range

Theorem 3 establishes the benefit of dynamic scheduling but leaves the endpoints $[p_{\text{low}}, p_{\text{high}}]$ open. We select this range based on three considerations.

Empirical performance is concentrated in a moderate range. The static sweep in Section 4.2 shows that the strongest configurations across benchmarks fall within $p \in [-2, 2]$, with performance degrading smoothly outside this interval.

The lower bound is constrained by optimisation stability. Corollary 7 refines Theorem 2: under mild gradient-orthogonality, the second moment is minimised at some $p^* \leq 0$ rather than $p \to -\infty$, since weight concentration grows exponentially and counteracts the Hölder-mean decrease.

The optimal range is task-dependent. We adopt $[2, -2]$ as the default for mathematical reasoning, where Qwen2.5-Math's strong pre-training tolerates the aggressive upper bound for early-phase signal amplification. The endpoints are not universal: for ALFWorld (Section 4.5), where the base model lacks domain-specific pre-training, a more conservative $[1, -1]$ outperforms $[2, -2]$, suggesting both endpoints should be calibrated to the base model's reasoning maturity and the task's signal-density profile.

4.5 Generalisation to Agentic Reasoning

To demonstrate that the advantages of HölderPO extend beyond pure mathematical domains to broader sequential decision-making scenarios, we evaluate our framework on the ALFWorld benchmark [Shridhar et al., 2020]. ALFWorld is a challenging embodied agentic environment that requires models to complete multi-step, open-ended tasks (e.g., finding, cleaning, or heating objects) based entirely on textual observations and actions. Unlike mathematical proofs, where reasoning is largely self-contained, agentic tasks suffer from compounding errors over long horizons, making stable policy optimisation crucial for success. Following established setups, we employ Qwen2.5-Instruct-1.5B as our base model for this agentic reasoning task. Table 3 presents the success rates across the six distinct sub-task categories in the ALFWorld evaluation suite.

Training Objectives	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick Two**	**Avg.**
Baselines (Base Model: Qwen2.5-Instruct-1.5B)							
GRPO ($p = 1$) [Shao et al., 2024]	85.3	53.7	84.5	78.2	59.7	53.5	72.8
GMPO ($p \to 0$) [Zhao et al., 2025]	93.1	78.6	81.0	88.2	82.1	**89.5**	85.9
GiGPO [Feng et al., 2025]	94.4	67.5	94.8	94.4	79.8	76.4	86.7
HölderPO (Ours)							
HölderPO (Linear Dec: $2 \to -2$**)**	**97.2**	85.7	87.5	91.7	79.2	81.5	87.5
HölderPO (Linear Dec: $1 \to -1$**)**	96.9	**100.0**	**100.0**	**100.0**	**85.7**	84.5	**93.8**

Table 3: Success rates (%) on the ALFWorld agentic reasoning benchmark. HölderPO demonstrates strong generalisation to open-ended, multi-step decision-making tasks.

Consistent with our findings in the mathematical domain, HölderPO yields substantial performance gains in agentic environments. The dynamic scheduling of p proves particularly well-suited for the compounding challenges of ALFWorld. During the early stages of training, a positive initial p amplifies the sparse, high-magnitude signals associated with rare successful trajectories, effectively exploiting the positive-concentration regime (Theorem 1). In the later stages, annealing to a negative p tightens the gradient variance bound (Theorem 2), protecting the policy from being derailed by spurious environmental feedback or minor missteps.

Notably, because our base model (Qwen2.5-Instruct-1.5B) lacks the extensive domain-specific pretraining seen in the mathematical models, it does not initially possess strong, reliable intuitions for embodied environments. Consequently, an overly aggressive initial parameter (e.g., $p = 2$) risks over-amplifying early, noisy exploration. Instead, a more conservative schedule (decaying from 1 to -1) proves optimal. By providing a steadier phase of signal amplification before transitioning into variance contraction, this tailored schedule achieves an exceptional average success rate of 93.8%, substantially outperforming all baselines. This careful calibration of the concentration–stability trade-off yields a highly robust policy for long-horizon planning.

5 Conclusion

We introduced **Hölder Policy Optimisation (HölderPO)**, a generalised framework that resolves the concentration–stability trade-off inherent in static policy optimisation methods like GRPO. By unifying importance-ratio aggregation through the Hölder mean, the parameter p serves as a continuous dial: larger p amplifies sparse high-magnitude learning signals, while smaller p tightens the gradient variance bound. Built on this principle, our dynamic p-annealing scheduler achieves state-of-the-art performance across mathematical and agentic benchmarks, securing 54.9% on five mathematical reasoning benchmarks and 93.8% on ALFWorld.

Limitations. Two limitations stand out. First, the schedule introduces hyperparameters (p_{high}, p_{low}, decay shape) that require empirical tuning per task; while linear decay performed best in our setup, we provide no theoretical characterisation of the optimal shape. Second, the positive-concentration regime amplifies tokens with high importance ratios, making HölderPO more susceptible to reward hacking when the verifier provides false-positive signals.

Future Work. A primary direction is an *adaptive* scheduler that adjusts p from real-time metrics (e.g., batch-level gradient variance or token-ratio dispersion), removing the need for manual tuning.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Colby Denison, Amanda Askell, Robert Lasenby, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Wenlong Deng, Yi Ren, Yushu Li, Boying Gong, Danica J Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. Token hidden reward: Steering exploration-exploitation in group relative deep reinforcement learning. *arXiv preprint arXiv:2510.03669*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

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.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy rl with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025.

Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25548–25560. Association for Computational Linguistics, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL `https://aclanthology.org/2024.acl-long.211/`.

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.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

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.

LI Jia, Beeching Edward, Tunstall Lewis, Lipkin Ben, Soletskyi Roman, Huang Shengyi Costa, Rasul Kashif, Yu Longhui, Jiang Albert, Shen Ziju, Qin Zihan, Dong Bin, Zhou Li, Fleureau Yann, Lample Guillaume, and Polu Stanislas. Numinamath, 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.

Yuhui Li, Fangyun Fang, Huan Liu, et al. Rain: Your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xingyu Lin, Yilin Wen, En Wang, Du Su, Wenbin Liu, Chenfu Bao, and Zhonghou Lv. Token-level policy optimization: Linking group-level rewards to token-level aggregation via markov likelihood. *arXiv preprint arXiv:2510.09369*, 2025.

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.

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.

Yichen Ouyang, Lu Wang, Fangkai Yang, Pu Zhao, Chenghua Huang, Jianfeng Liu, Bochen Pang, Yaming Yang, Yuefeng Zhan, Hao Sun, et al. Token-level proximal policy optimization for query generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31184–31198, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Luke Zettlemoyer, et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 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.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Marco Simoni, Aleksandar Fontana, Giulio Rossolini, Andrea Saracino, and Paolo Mori. Gtpo: Stabilizing group relative policy optimization via gradient and entropy control. *arXiv preprint arXiv:2508.03772*, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Yining Wang, Jinman Zhao, Chuangxin Zhao, Shuhao Guan, Gerald Penn, and Shinan Liu. $lambda$-grpo: Unifying the grpo frameworks with learnable token preferences. *arXiv preprint arXiv:2510.06870*, 2025b.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. Bnpo: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025.

Jian Xiong, Jingbo Zhou, Jingyong Ye, Qiang Huang, and Dejing Dou. Aapo: Enhancing the reasoning capabilities of llms with advantage momentum. *arXiv preprint arXiv:2505.14264*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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*, 2025a.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2023.

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. *arXiv preprint arXiv:2503.14476*, 2025.

Song Yu and Li Li. Erpo: Token-level entropy-regulated policy optimization for large reasoning models. *arXiv preprint arXiv:2603.28204*, 2026.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Weisong Zhao, Tong Wang, Zichang Tan, Te Yang, Siran Peng, Haoyuan Zhang, Tianshuo Zhang, Haichao Shi, Meng Meng, Yang Yang, et al. One ring to rule them all: Unifying group-based rl via dynamic power-mean geometry. *arXiv preprint arXiv:2601.22521*, 2026.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2023.

Appendix Outline

- Appendix A: Extended related work.
- Appendix B: Training dynamics (entropy and gradient-norm) under different p.
- Appendix C: Pseudocode for the HölderPO loss in log-space.
- Appendix D: HölderPO results under token-level clipping.
- Appendix E: Schedule-shape ablation across linear, square, cube, and sinusoidal interpolations.
- Appendix F: Generalisation to Qwen3-4B-Base and Qwen3-8B-Base.
- Appendix G: Formulas and gradient derivations under three clipping regimes.
- Appendix H: Proofs of Theorem 5 and Theorem 1 (distribution deformation).
- Appendix I: Proof of Theorem 2 and Corollary 7 (variance behaviours).
- Appendix J: Proof of Theorem 3 (advantage of dynamic scheduling).
- Appendix K: Broader Impacts

A Extended Related Work

We expand on the broader GRPO ecosystem briefly mentioned in Section 2. The variants below address aspects of the RL pipeline orthogonal to token-level aggregation.

Surrogate Loss and Critic-Free Variants. GPG [Chu et al., 2025] simplifies the GRPO objective by removing surrogate losses entirely, while DAPO [Yu et al., 2025] introduces dynamic sampling and decoupled clipping bounds. Dr.GRPO [Liu et al., 2025] mitigates length bias by removing the per-sequence length normalisation, λ-GRPO [Wang et al., 2025b] learns the length preference via a trainable parameter. These methods modify the loss normalisation rather than the aggregation function and are complementary to our framework.

Advantage Estimation and Reward Shaping. AAPO [Xiong et al., 2025] introduces advantage momentum to mitigate zero-gradient situations; BNPO [Xiao et al., 2025] adaptively normalises rewards via a Beta distribution; OPO [Hao et al., 2025] provides a variance-minimising baseline; and Seed-GRPO [Chen et al., 2025] scales updates by question difficulty. These contributions modify the advantage signal rather than how token-level ratios are aggregated.

Value-Model-Based Variants. To circumvent GRPO's variance issues, some approaches revert to PPO with pre-trained value models, including VC-PPO [Yuan et al., 2025] and T-PPO [Ouyang et al., 2025]. While effective, the external value model introduces confounding factors and computational overhead that the critic-free GRPO family, including ours, deliberately avoids.

Data-Centric and Curriculum-Based Approaches. Open-Reasoner-Zero [Hu et al., 2025], PRIME [Cui et al., 2025], and SimpleRL-Zero [Zeng et al., 2025] democratise scalable RL training through curated training data, curriculum learning, and clean base models. These contributions are at the data and pipeline level, complementary to algorithmic refinements such as ours.

B Training Dynamics: Entropy and Gradient-Norm

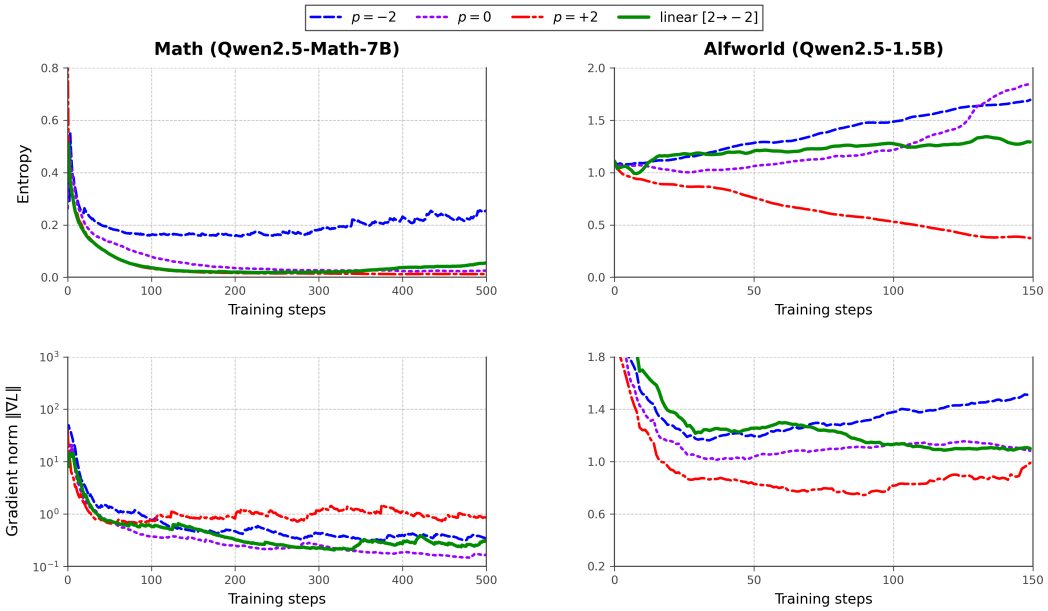

Figure 3: **Entropy and gradient-norm dynamics under different Hölder exponents p. Columns:** Math (Qwen2.5-Math-7B on MATH-12k) and Alfworld (Qwen2.5-1.5B). **Rows:** per-step policy entropy and gradient norm $\|\nabla\mathcal{L}\|$ (log scale on Math, linear on Alfworld). Constant-p baselines ($p \in \{+2, 0, -2\}$, dashed/dotted/dash-dotted) are compared with our linearly-decaying schedule $p{:}\,2 \to -2$ (solid green). Positive p concentrates mass on high-likelihood tokens and pushes entropy down; negative p disperses mass and pushes it up. The schedule inherits both regimes in sequence and keeps the gradient norm in a tighter band than any constant choice.

C Pseudocode

HölderPO is a single-line modification of the GRPO loss. To preserve numerical stability for large $|p|$, all power operations are evaluated in log-space via the log-sum-exp identity. Algorithm 1 summarises the full computation. The aggregation operator is applicable at any granularity: in our experiments we use a single sequence-level $\rho_{i,p}$, but token-level or block-level aggregation can be substituted without changing the algorithm or theory.

Algorithm 1 HölderPO Loss Computation

Require: current policy π_θ, reference policy $\pi_{\theta_{\text{old}}}$
Input: sequence y of length T, valid-token mask M, advantage $\widehat{A}$, parameter p, clip range ϵ
// Step 1: log-ratio computation
$\Delta_t \leftarrow \log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi_{\theta_{\text{old}}}(y_t \mid x, y_{<t})$
$|y| \leftarrow \sum_{t=1}^{T} M_t$
// Step 2: Hölder-mean aggregation in log-space
if $|p| < 10^{-6}$ **then**
  $\rho \leftarrow \exp\Big(\frac{1}{|y|} \sum_t M_t \, \Delta_t\Big)$ // limit $p \to 0$ (geometric mean)
else
  $\rho \leftarrow \Big(\frac{1}{|y|} \sum_t M_t \, \exp(p \, \Delta_t)\Big)^{1/p}$
end if
// Step 3: PPO-style clipping and loss
$\rho_{\text{clip}} \leftarrow \text{clip}(\rho, \, 1-\epsilon, \, 1+\epsilon)$
$\mathcal{L}_{\text{unclip}} \leftarrow -\widehat{A} \cdot \rho$
$\mathcal{L}_{\text{clip}} \leftarrow -\widehat{A} \cdot \rho_{\text{clip}}$
$\mathcal{L}_{\text{HPO}} \leftarrow \max\big(\mathcal{L}_{\text{unclip}}, \, \mathcal{L}_{\text{clip}}\big)$ // minimised via SGD
return $\mathcal{L}_{\text{HPO}}$

D Token-Level Clipping of HölderPO

Training Objectives	AIME24	AMC	MATH500	Minerva	Oly.	Avg.
Token-Level Clip HölderPO						
HölderPO ($p = -2$)	36.7	61.4	81.6	35.7	43.6	51.8
HölderPO ($p = -1$)	40.0	62.7	81.6	33.5	44.5	52.5
HölderPO ($p \to 0$)	43.3	61.3	81.6	34.6	42.7	52.7
HölderPO ($p = 1$)	43.3	63.9	80.8	31.6	42.8	52.5
HölderPO ($p = 2$)	43.3	60.2	81.2	33.5	44.9	52.6

Table 4: HölderPO under *token-level* clipping (Eq. 11) on Qwen2.5-Math-7B across five mathematical benchmarks. Compared with the sequence-level clipping setting reported in Table 2, the token-level variant produces a noticeably narrower performance spread across p, consistent with our discussion in Section 3.3: token-level clipping breaks the algebraic structure underlying the variance bound's monotonicity in p, weakening the controlled trade-off that motivates dynamic scheduling.

E Schedule-Shape Ablation

Training Objectives	AIME24	AMC	MATH500	Minerva	Oly.	Avg.
Dynamic HölderPO Variants						
HölderPO (Square Asc: $-2 \to 2$)	33.3	62.7	82.6	33.8	44.4	51.4
HölderPO (Cube Asc: $-2 \to 2$)	36.7	62.7	82.8	32.0	43.1	51.4
HölderPO (Sin Asc: $-2 \to 2$)	43.3	59.0	81.6	35.7	45.8	**53.1**
HölderPO (Linear Asc: $-2 \to 2$)	40.0	**66.3**	81.2	32.7	42.7	52.6
HölderPO (Square Des: $2 \to -2$)	36.7	61.4	81.6	33.1	44.0	51.3
HölderPO (Cube Des: $2 \to -2$)	36.7	60.2	80.8	31.6	**46.1**	51.1
HölderPO (Sin Des: $2 \to -2$)	36.7	61.4	81.6	35.3	46.8	52.4

Table 5: Comparison of alternative annealing shapes for the dynamic schedule on Qwen2.5-Math-7B. We sweep four monotonic interpolation families (linear, square, cube, sinusoidal) in both ascending ($-2 \to 2$) and descending ($2 \to -2$) directions, holding the endpoints fixed at $\{-2, +2\}$. Among the seven variants listed here, none surpasses the descending linear schedule of 54.9% reported in Table 2, supporting our choice of linear decay as the default.

F Generalisation to Qwen3 Base Models

To verify that HölderPO transfers beyond the Qwen2.5-Math-7B setting, we additionally evaluate on the Qwen3-Base series and compare against three strong token-aggregation baselines (GRPO, GSPO, DAPO) under matched configurations.

Training Objectives	MATH500	AIME25†	AMC23	Minerva	Oly.	Avg.
Base Model						
Qwen3-4B-Base [Yang et al., 2025a]	58.2	7.4	45.0	14.0	28.6	30.6
RL Post-Trained Models						
GRPO [Shao et al., 2024]	79.3	18.5	60.0	21.0	40.5	43.9
GSPO [Zheng et al., 2025]	78.5	18.5	62.5	23.2	39.8	44.5
DAPO [Yu et al., 2025]	81.3	22.2	**65.0**	21.7	**41.8**	46.4
HölderPO (Linear Des: $2 \to -2$, Ours)	**88.0**	**30.0**	60.2	**39.0**	40.6	**50.9**

Table 6: Results on **Qwen3-4B-Base**. We report Pass@1 accuracy (%) for MATH500, AMC23, Minerva, and Olympiad, and Pass@8 accuracy (%) for AIME25 (†). HölderPO with our default linear annealing schedule (p: $2 \to -2$) achieves 50.9% average accuracy, a 4.5-point absolute gain over the strongest aggregation baseline DAPO (46.4%) and 7.0-point gain over GRPO (43.9%).

Training Objectives	MATH500	AIME25†	AMC23	Minerva	Oly.	Avg.
Base Model						
Qwen3-8B-Base [Yang et al., 2025a]	65.0	11.1	45.0	17.3	31.1	33.9
RL Post-Trained Models						
GRPO [Shao et al., 2024]	80.1	22.2	67.5	27.6	42.4	48.0
GSPO [Zheng et al., 2025]	81.7	22.2	67.5	26.8	45.5	48.7
DAPO [Yu et al., 2025]	85.3	25.9	75.0	27.9	48.7	52.6
HölderPO (Linear Des: $2 \to -2$, Ours)	**88.4**	**33.3**	**75.1**	**37.5**	**50.3**	**56.9**

Table 7: Results on **Qwen3-8B-Base**. Same evaluation protocol as Table 6. The advantage of HölderPO grows with model scale: at 8B, our method reaches 56.9% average accuracy, a 4.3-point gain over DAPO (52.6%) and 8.9-point gain over GRPO (48.0%). Notably, the relative improvement on Minerva ($+9.6$ over DAPO) and Olympiad ($+1.6$) confirms that the gains generalise across both standard and competition-level mathematical reasoning.

G Formulas and Derivation

In this section, we derive the formulas involved in our theory. These formulas can be divided into three parts according to different clipping mechanisms: the original unclipped, token-level clipping, and sequence-level clipping. Finally, we discuss the definition of Hölder p-norm when $p = 0$.

We define some notation used throughout this chapter. Let $\mathcal{D}$ denote the dataset of input prompts, from which a query q (or context x) is sampled. For each prompt, we sample a group of G responses, denoted as $\{y_i\}_{i=1}^{G}$, from the reference or old policy $\pi_{\theta_{old}}$. For the i-th response y_i, let $|y_i|$ represent its total token length, where $y_{i,t}$ is the t-th token and $y_{i,<t}$ denotes the prefix. The current policy parameterised by θ is denoted as π_θ. Finally, $\widehat{A}_i$ represents the estimated advantage for the i-th response, defined as

$$\widehat{A}_i = \frac{r\left(x, y_i\right) - \operatorname{mean}\left(\left\{r\left(x, y_i\right)\right\}_{i=1}^{G}\right)}{\operatorname{std}\left(\left\{r\left(x, y_i\right)\right\}_{i=1}^{G}\right)},$$

where $r(x, y_i)$ denotes the absolute reward score assigned to the i-th generated response y_i conditioned on the input x. Here $\operatorname{mean}(\cdot)$ and $\operatorname{std}(\cdot)$ represent the arithmetic mean and standard deviation. For simplicity, in this section we omit the KL regularisation term from all PPO-style objective function formulas.

G.1 No Clipping Formulas

As we know, the simplest unclipped GRPO objective function formula is

$$\mathbb{E}_{q\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x,y_{i,<t})}\right)\widehat{A}_i\right],$$

which can be regarded as a special case of the objective function given by Schulman et al. [2015] with the surrogate term $\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x,y_{i,<t})}$. Obviously it is the arithmetic mean value of importance sampling ratios $r_{i,t} = \frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x,y_{i,<t})}$ with respect to tokens in the sequence y_i. By Hölder-p norm, we extend the arithmetic mean value of ratios to $\rho_{i,p}(\theta)$, which is defined as

$$\rho_{i,p}(\theta) \;=\; \left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p\right)^{1/p}, \qquad p \in \mathbb{R}\setminus\{0\}. \tag{1}$$

Later we discuss the case for $p = 0$ in G.4. The unclipped objective function of Hölder-MPO is

$$\mathcal{J}_{\mathrm{H}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{\mathrm{G}}\rho_{i,p}(\theta)\widehat{A}_i\right]. \tag{7}$$

To calculate the policy gradient of this objective function, we first prove

$$\nabla_\theta\rho_{i,p}(\theta) = \frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x, y_{i,<t}\right). \tag{8}$$

Starting from the definition of the Hölder mean in (1),

$$\log\rho_{i,p}(\theta) \;=\; \frac{1}{p}\log\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p\right).$$

Differentiating both sides with respect to θ:

$$\frac{\nabla_\theta\rho_{i,p}(\theta)}{\rho_{i,p}(\theta)} \;=\; \frac{1}{p}\cdot\frac{\nabla_\theta\sum_t r_{i,t}^p(\theta)}{\sum_t r_{i,t}^p(\theta)}.$$

Since π_{old} does not depend on θ, the chain rule gives

$$\nabla_\theta r_{i,t}(\theta) \;=\; \frac{\nabla_\theta \pi_\theta(y_{i,t}\,|\,x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}\,|\,x, y_{i,<t})} \;=\; r_{i,t}(\theta)\cdot \nabla_\theta \log \pi_\theta(y_{i,t}\,|\,x, y_{i,<t}),$$

and consequently $\nabla_\theta r_{i,t}^p(\theta) = p\, r_{i,t}^p(\theta)\, \nabla_\theta \log \pi_\theta(y_{i,t}\,|\,\cdot)$. Substituting and cancelling the factor of p:

$$\frac{\nabla_\theta \rho_{i,p}(\theta)}{\rho_{i,p}(\theta)} \;=\; \frac{\sum_t r_{i,t}^p(\theta)\, \nabla_\theta \log \pi_\theta(y_{i,t}\,|\,\cdot)}{\sum_t r_{i,t}^p(\theta)} \;=\; \sum_t W_{i,t}(p)\, \nabla_\theta \log \pi_\theta(y_{i,t}\,|\,\cdot),$$

where $W_{i,t}(p)$ is defined in (3). Multiplying through by $\rho_{i,p}(\theta)$ recovers Eq. (3).

Then the policy gradient of the unclipped Hölder-MPO is

$$\nabla_\theta \mathcal{J}_{\mathrm{H}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{\mathrm{G}} \nabla_\theta \rho_{i,p}(\theta)\widehat{A}_i\right], \tag{9}$$

whose unbiased mini-batch estimator is denoted by $\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}}(\theta)$

$$\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{G}\sum_{i=1}^{\mathrm{G}} \nabla_\theta \rho_{i,p}(\theta)\widehat{A}_i\right].$$

where B denotes the batch size. By Eq. (8), we know

$$\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{G}\sum_{i=1}^{\mathrm{G}}\left(\widehat{A}_i \frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|}\right)\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p \nabla_\theta \log \pi_\theta(y_{i,t}\mid x, y_{i,<t}). \tag{10}$$

G.2 Token-Level Clipping Formulas

There are many PPO extensions adopting token-level clipping mechanisms to ensure training stability and prevent policy collapse. For instance, Group Relative Policy Optimisation (GRPO), Geometric-Mean Policy Optimisation (GMPO) [Zhao et al., 2025] and Dynamic Sampling Policy Optimisation (DAPO) [Yu et al., 2025]. With token-level clipping, the objective function (7) of our Hölder-MPO becomes

$$\mathcal{J}_{\mathrm{H}_t}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\left(\sum_{\widehat{A}_i>0} C_{i,p}\widehat{A}_i + \sum_{\widehat{A}_i<0} D_{i,p}\widehat{A}_i\right)\right], \tag{11}$$

$$C_{i,p} = \left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(r_{i,t}(\theta), \mathrm{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\right)^p\right)^{1/p},$$

$$D_{i,p} = \left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\max\left(r_{i,t}(\theta), \mathrm{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\right)^p\right)^{1/p},$$

where the clipping function is defined by

$$\mathrm{clip}(x, 1-\epsilon, 1+\epsilon) := \begin{cases} 1-\epsilon, & \text{if } x < 1-\epsilon \\ x, & \text{if } 1-\epsilon \le x \le 1+\epsilon \\ 1+\epsilon, & \text{if } x > 1+\epsilon. \end{cases} \tag{12}$$

To deduce this formula, we firstly recall the token-level clipping GRPO objective function in [Shao et al., 2024]

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min(r_{i,t}\widehat{A}_{i,t}, \mathrm{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_{i,t})\right],$$

where $\widehat{A}_{i,t} = \widehat{A}_i$ is the estimator of sequence-level reward. According to the sign of $\widehat{A}_i$, the content inside the expectation of GRPO objective function should equal to

$$\frac{1}{G}\left(\left(\sum_{\widehat{A}_i>0}+\sum_{\widehat{A}_i<0}\right)\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min(r_{i,t}\widehat{A}_i,\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)\widehat{A}_i)\right).$$

For $\widehat{A}_i > 0$, it is obvious that

$$\min(r_{i,t}\widehat{A}_i,\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)\widehat{A}_i) = \min(r_{i,t},\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon))\widehat{A}_i.$$

For $\widehat{A}_i < 0$, it is obvious that

$$\min(r_{i,t}\widehat{A}_i,\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)\widehat{A}_i) = \max(r_{i,t},\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon))\widehat{A}_i.$$

Therefore, the positive and negative part of the content inside the expectation of GRPO objective function should be expressed as

$$\frac{1}{G}\sum_{\widehat{A}_i>0}\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min(r_{i,t},\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon))\right)\widehat{A}_i,$$

$$\frac{1}{G}\sum_{\widehat{A}_i<0}\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\max(r_{i,t},\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon))\right)\widehat{A}_i,$$

which are special cases of $C_{i,p}$ and $D_{i,p}$ when $p = 1$. Later in G.4 we will show the objective function of GMPO is a special case of (11) when $p = 0$.

Next we deduce the policy gradient formula of token-level clipping objective function. It is obvious that

$$\nabla_\theta \mathcal{J}_{\mathrm{H}_t}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left[\frac{1}{G}\left(\sum_{\widehat{A}_i>0}\nabla_\theta C_{i,p}\widehat{A}_i+\sum_{\widehat{A}_i<0}\nabla_\theta D_{i,p}\widehat{A}_i\right)\right].$$

The derivatives of $C_{i,p}$ and $D_{i,p}$ depend on the value taken by the clipping function. When $r_{i,t} \le 1+\epsilon$, the smaller one of $r_{i,t}$ and $\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)$ is $r_{i,t}$, whereas the smaller value becomes $1+\epsilon$ when $r_{i,t} > 1+\epsilon$. In the former case, the contribution of token t to $\nabla_\theta C_{i,p}$ is

$$\frac{C_{i,p}(\theta)^{1-p}}{|y_i|}r_{i,t}(\theta)^p\cdot\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right).$$

In the latter case, this token's partial derivative contribution to $\nabla_\theta C_{i,p}$ is zero. Similarly, when $r_{i,t} \ge 1-\epsilon$, the larger one of $r_{i,t}$ and $\mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)$ is $r_{i,t}$, whereas the larger value becomes $1-\epsilon$ when $r_{i,t} < 1-\epsilon$. In the former case, the contribution of token t to $\nabla_\theta D_{i,p}$ is

$$\frac{D_{i,p}(\theta)^{1-p}}{|y_i|}r_{i,t}(\theta)^p\cdot\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right).$$

In the latter case, this token's partial derivative contribution is zero. We summarise all the cases in the following formula

$$\nabla_\theta \mathcal{J}_{\mathrm{H}_t}(\theta) = \mathbb{E}_{x,\{y_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\cdot\frac{H_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|}\mathbb{I}_{i,t}(\theta)\cdot r_{i,t}(\theta)^p\cdot\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right)\right], \tag{13}$$

$$H_{i,p}(\theta)=\begin{cases}C_{i,p}, & \text{if } \widehat{A}_i\ge 0\\ D_{i,p}, & \text{if } \widehat{A}_i<0,\end{cases}\qquad \mathbb{I}_{i,t}(\theta)=\begin{cases}0, & \text{if } \widehat{A}_i>0 \text{ and } r_{i,t}(\theta)>1+\epsilon, \text{ or, if } \widehat{A}_i<0 \text{ and } r_{i,t}(\theta)<1-\epsilon\\ 1, & \text{otherwise.}\end{cases}$$

The unbiased mini-batch estimator is

$$\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}_t}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{G}\sum_{i=1}^{G}\widehat{A}_i\cdot\frac{H_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|}\mathbb{I}_{i,t}(\theta)\cdot r_{i,t}(\theta)^p\cdot\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right). \tag{14}$$

G.3 Sequence-Level Clipping Formulas

Notably, alongside the widespread adoption of token-level clipping, several recent studies have shifted towards sequence-level clipping strategies, such as Group Sequence Policy Optimisation (GSPO) [Zheng et al., 2025], whose objective function is

$$\mathcal{J}_{\mathrm{GSPO}}(\theta)=\mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(s_i(\theta)\widehat{A}_i,\mathrm{clip}\left(s_i(\theta),1-\varepsilon,1+\varepsilon\right)\widehat{A}_i\right)\right],$$

where $s_i(\theta)=\left(\prod_{t=1}^{|y_i|}r_{i,t}\right)^{\frac{1}{|y_i|}}=\exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t}|x,y_{i,<t})}\right)$ is the geometric mean of ratio of each token. Actually it is a special case of our Hölder-MPO objective function with sequence-level clipping

$$\mathcal{J}_{\mathrm{H}_s}(\theta)=\mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}\left\{\frac{1}{G}\sum_{i=1}^{G}\min\left[\rho_{i,p}(\theta)\widehat{A}_i,\mathrm{clip}\left(\rho_{i,p}(\theta),1-\epsilon,1+\epsilon\right)\widehat{A}_i\right]\right\}. \tag{15}$$

In Lemma 1, we will show $s_i(\theta)$ in GSPO is equal to $\rho_{i,0}(\theta)$. By a similar discussion for $\rho_{i,p}$, we can obtain the policy gradient

$$\nabla_\theta\mathcal{J}_{\mathrm{H}_s}(\theta)=\mathbb{E}_{x,\{y_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\mathbb{I}_i(\theta)\cdot\widehat{A}_i\cdot\frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|}r_{i,t}(\theta)^p\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right)\right], \tag{16}$$

$$\mathbb{I}_i(\theta)=\begin{cases}0, & \text{if } \widehat{A}_i>0 \text{ and } \rho_{i,p}(\theta)>1+\epsilon, \text{ or, if } \widehat{A}_i<0 \text{ and } \rho_{i,p}(\theta)<1-\epsilon\\ 1, & \text{otherwise.}\end{cases}$$

The unbiased mini-batch estimator is

$$\widehat{\nabla}_\theta\mathcal{J}_{\mathrm{H}_s}(\theta)=\frac{1}{B}\sum_{b=1}^{B}\frac{1}{G}\sum_{i=1}^{G}\mathbb{I}_i(\theta)\cdot\widehat{A}_i\cdot\frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|}r_{i,t}(\theta)^p\nabla_\theta\log\pi_\theta\left(y_{i,t}\mid x,y_{i,<t}\right). \tag{17}$$

G.4 $p=0$ Formulas

In this section, we extend the three kinds of formulas to $p=0$. By functional analysis, the mean value given by Hölder p-norm for a sequence of positive real numbers $x_1,\ldots,x_n$ is

$$M_p=\left(\tfrac{1}{n}\right)^{\frac{1}{p}}\left(x_1^p+\cdots+x_n^p\right)^{\frac{1}{p}}.$$

The following lemma shows that when $p\to 0$, the limit of the mean value given by Hölder p-norm is the geometric mean value.

Lemma 1. *For any sequence of positive real numbers $x_1,\ldots,x_n$, the Hölder mean M_p converges to the geometric mean as p approaches 0.*

Proof. To evaluate the limit of M_p as $p\to 0$, we first take the natural logarithm of M_p:

$$\ln(M_p)=\frac{\ln\left(\frac{1}{n}\sum_{i=1}^{n}x_i^p\right)}{p}.$$

Let us define an auxiliary function $f(p)=\ln\left(\frac{1}{n}\sum_{i=1}^{n}x_i^p\right)$. Notice that at $p=0$, $f(0)=\ln\left(\frac{1}{n}\sum_{i=1}^{n}1\right)=\ln(1)=0$.

Therefore, the limit as $p\to 0$ is precisely the definition of the derivative of $f(p)$ evaluated at $p=0$:

$$\lim_{p\to 0}\ln(M_p)=\lim_{p\to 0}\frac{f(p)-f(0)}{p-0}=f'(0).$$

We can explicitly compute the derivative $f'(p)$ using the chain rule:

$$f'(p)=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}x_i^p}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}x_i^p\ln(x_i)\right).$$

Evaluating this derivative at $p = 0$ gives:

$$f'(0) = \frac{1}{\frac{1}{n}(n)} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} 1 \cdot \ln(x_i) \right) = \frac{1}{n} \sum_{i=1}^{n} \ln(x_i).$$

Finally, exponentiating both sides recovers the limit for the original expression M_p:

$$\lim_{p \to 0} M_p = \lim_{p \to 0} e^{\ln(M_p)} = e^{f'(0)} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln(x_i)} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}.$$

This recovers exactly the geometric mean of the sequence, completing the proof. □

Naturally, we can define all of our objective functions by the geometric mean value for $p = 0$. Hence we can see the GSPO (resp. GMPO) objective function is the $p = 0$ special case of our sequence-level (resp. token-level) clipping objective function.

For the policy gradient calculations, we need to discuss the commutativity of operators $\lim_{p \to 0}$ and ∇_θ. We first define the concept of class C^1 multi-variable function $f(p, \theta)$.

Definition 1 (Class C^1)**.** *Let $U \subseteq \mathbb{R} \times \mathbb{R}^d$ be an open set. A function $f : U \to \mathbb{R}$ is said to be jointly continuously differentiable, or of class C^1 on U (denoted as $f \in C^1(U)$), if it satisfies that all first-order partial derivatives of f, namely $\frac{\partial f}{\partial p}$ and the gradient vector $\nabla_\theta f$, exist at every point $(p, \theta) \in U$ and are jointly continuous on U.*

The next theorem, whose study object is C^1-function, can be utilised to guarantee the commutativity of the two operators in the no-clipping case.

Theorem 4. *Let $f(p, \theta)$ be a parameterised function defined on $(I \setminus \{0\}) \times U$ ($0 \in I$), where $p \in I \subset \mathbb{R}$ and $\theta \in U \subset \mathbb{R}^d$. Suppose the singularity at $p = 0$ is removable, such that the extended function defined as*

$$\tilde{f}(p, \theta) = \begin{cases} f(p, \theta), & \text{if } p \neq 0, \\ \lim_{p \to 0} f(p, \theta), & \text{if } p = 0, \end{cases}$$

is of class C^1 on the joint neighbourhood $I \times U$. Then the differential operator commutes with the limit operator as $p \to 0$:

$$\lim_{p \to 0} \nabla_\theta \tilde{f}(p, \theta) = \nabla_\theta \tilde{f}(0, \theta) = \nabla_\theta \left(\lim_{p \to 0} f(p, \theta) \right).$$

Proof. By hypothesis, the extended objective function $\tilde{f}(p, \theta)$ is of class C^1 on the joint domain $I \times U$. According to Thm. 9.21 in Rudin [1976], the partial derivative operator with respect to θ, denoted as $\nabla_\theta \tilde{f}(p, \theta)$, forms a continuous mapping from $I \times U$ to $\mathbb{R}^d$. Then for any fixed parameter $\theta \in U$, the mapping $p \mapsto \nabla_\theta \tilde{f}(p, \theta)$ is continuous at $p = 0$. Thus we obtain the result. □

For the no-clipping case (7), the function inside the expectation is $L(p, \theta) = \frac{1}{G} \sum_{i=1}^{G} \rho_{i,p}(\theta) \widehat{A}_i$. Because the group size G and the advantage estimates $\widehat{A}_i$ are scalars, the function $L(p, \theta)$ is of class C^1 if and only if $\rho_{i,p}(\theta)$ is of class C^1. This holds true based on the two properties below.

Firstly, following standard assumptions in deep reinforcement learning, the neural network π_θ utilising smooth activation functions (e.g., Swish, GeLU) and linear transformations is continuously differentiable (C^1) with respect to its weights θ. By the chain rule, the strictly positive composite function $r_{i,t}(\theta)$ identically inherits this C^1 property. For widely adopted Lipschitz continuous activation functions that are not strictly C^1 globally (e.g., ReLU), Rademacher's theorem guarantees that they are differentiable almost everywhere. In the context of stochastic optimisation over continuous parameter spaces, the set of points where the derivative is undefined has Lebesgue measure zero. Consequently, they admit generalised gradients (e.g., Clarke subdifferentials) and are conventionally treated within the C^1 framework without loss of theoretical generality.

Secondly, for any $p \neq 0$, $\rho_{i,p}(\theta)$ is a composition of smooth elementary functions and is inherently C^1. At the singularity $p = 0$, we evaluate the extended function through its logarithmic form:

$$\ln \rho_{i,p}(\theta) = \frac{1}{p} \ln \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} e^{p \ln r_{i,t}(\theta)} \right). \tag{18}$$

By expanding the inner exponential term via its Taylor series around $p = 0$, we obtain $\frac{1}{|y_i|} \sum (1 + p \ln r_{i,t} + \mathcal{O}(p^2)) = 1 + p(\frac{1}{|y_i|} \sum \ln r_{i,t}) + \mathcal{O}(p^2)$. Applying the first-order Taylor expansion to the outer logarithm $\ln(1 + z) \approx z$ yields:

$$\ln \rho_{i,p}(\theta) = \frac{1}{p} \left[p \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \ln r_{i,t}(\theta) \right) + \mathcal{O}(p^2) \right]. \tag{19}$$

The parameter p in the denominator perfectly cancels the leading p in the numerator. Because the singularity is analytically removed through this cancellation, the extended function $\rho_{i,p}(\theta)$ and its partial derivatives ($\frac{\partial}{\partial p}$ and ∇_θ) exhibit no discontinuities or undefined behaviour at $p = 0$.

Therefore, the inner objective function $L(p, \theta)$ is mathematically guaranteed to be jointly C^1 on the neighbourhood encompassing $p = 0$. This fulfils the prerequisites of Theorem 4, justifying the unconditional exchange of the limit $\lim_{p \to 0}$ and the policy gradient ∇_θ.

H Distribution Deformation

This appendix supplements Section 3.2 by providing formal proofs and broader theoretical contexts for our gradient concentration mechanism. Specifically, H.1 and H.2 present the proofs for the local weight allocation (Theorem 5) and the global distributional deformation (Theorem 1), respectively. Furthermore, H.3 discusses the profound connection between our three gradient concentration regimes and the traditional exploration-exploitation trade-off in reinforcement learning.

H.1 Local Property

In this section, we prove the following theorem, which is mentioned in Section 3.2 as the local property of the token-level weight allocation $W_{i,t}(p)$ induced by the aggregation parameter p.

Theorem 5. *Given an initial parameter state p_0. Let $\mathcal{T}^* = \{t \mid r_{i,t} = \max_k r_{i,k}(\theta)\}$ denote the set of strictly optimal tokens. As $p \to \infty$, $W_{i,t^*}(p)$ increases monotonically and converges to $1/|\mathcal{T}^*|$. For any $t \notin \mathcal{T}^*$, there exists a critical p-value $p_t > p_0$ such that $W_{i,t}(p)$ reaches its maximum at p_t, and strictly decays to zero thereafter as $p \to \infty$.*

To prove Theorem 5, we first establish two fundamental lemmas regarding the dynamic weight allocation mechanism controlled by p. Let $\mu_{y_i}(p)$ denote the $W_{i,t}$-weighted mean of the log-ratios across the sequence:

$$\mu_{y_i}(p) := \sum_{t=1}^{|y_i|} W_{i,t}(p) \log r_{i,t}(\theta). \tag{20}$$

Lemma 2. *The partial derivative of the gradient weight $W_{i,t}(p)$ with respect to the Hölder parameter p is strictly governed by its log-ratio relative to the sequence mean $\mu_{y_i}(p)$:*

$$\frac{\partial W_{i,t}(p)}{\partial p} = W_{i,t}(p) \Big(\log r_{i,t}(\theta) - \mu_{y_i}(p) \Big). \tag{21}$$

Proof. To provide a complete calculation, we begin by rewriting the gradient weight definition in its exponential form. By expanding the base $r_{i,t}(\theta)^p$, the weight can be expressed as

$$W_{i,t}(p) = \frac{\exp(p \log r_{i,t}(\theta))}{\sum_{k=1}^{|y_i|} \exp(p \log r_{i,k}(\theta))}.$$

Let $u = \exp(p \log r_{i,t}(\theta))$ and $v = \sum_{k=1}^{|y_i|} \exp(p \log r_{i,k}(\theta))$. Using the chain rule, the derivative of the numerator is simply

$$\frac{\partial u}{\partial p} = \exp(p \log r_{i,t}(\theta)) \log r_{i,t}(\theta),$$

while the derivative of the denominator is

$$\frac{\partial v}{\partial p} = \sum_{k=1}^{|y_i|} \exp(p \log r_{i,k}(\theta)) \log r_{i,k}(\theta).$$

The quotient rule formula is

$$\frac{d}{dp}\left[\frac{u}{v}\right] = \frac{u'v - uv'}{v^2}.$$

Now, we substitute these components back into the quotient rule formula

$$\frac{\partial W_{i,t}(p)}{\partial p} = \frac{\exp(p \log r_{i,t}(\theta)) \log r_{i,t}(\theta)}{\sum_k \exp(p \log r_{i,k}(\theta))} - \frac{\exp(p \log r_{i,t}(\theta)) \left[\sum_k \exp(p \log r_{i,k}(\theta)) \log r_{i,k}(\theta)\right]}{\left[\sum_k \exp(p \log r_{i,k}(\theta))\right]^2}.$$

Looking closely at the first term, we can isolate the definition of the original weight $W_{i,t}(p)$, leaving us with $W_{i,t}(p) \log r_{i,t}(\theta)$. For the second term, we can factor the fraction into the product of two separate fractions. The first fraction is exactly $W_{i,t}(p)$, and the second fraction represents the weighted sum over all tokens

$$\frac{\partial W_{i,t}(p)}{\partial p} = W_{i,t}(p) \log r_{i,t}(\theta) - W_{i,t}(p) \left[\sum_{k=1}^{|y_i|} \frac{\exp(p \log r_{i,k}(\theta))}{\sum_j \exp(p \log r_{i,j}(\theta))} \log r_{i,k}(\theta)\right].$$

We know that the term inside the summation is simply $W_{i,k}(p) \log r_{i,k}(\theta)$, and the entire bracketed sum represents $\mu_{y_i}(p) = \sum_k W_{i,k}(p) \log r_{i,k}(\theta)$. Substituting this notation into our equation gives

$$\frac{\partial W_{i,t}(p)}{\partial p} = W_{i,t}(p) \log r_{i,t}(\theta) - W_{i,t}(p)\mu_{y_i}(p) = W_{i,t}(p)\Big(\log r_{i,t}(\theta) - \mu_{y_i}(p)\Big).$$

□

Lemma 3. *Assuming the sequence contains at least two tokens with differing importance ratios, the weighted sequence mean $\mu_{y_i}(p)$ is strictly monotonically increasing with respect to p.*

Proof. Taking the derivative of $\mu_{y_i}(p)$ with respect to p, we have

$$\frac{\partial \mu_{y_i}(p)}{\partial p} = \sum_{t=1}^{|y_i|} \frac{\partial W_{i,t}(p)}{\partial p} \log r_{i,t} = \sum_{t=1}^{|y_i|} W_{i,t}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big) \log r_{i,t}.$$

Since $\sum_{t=1}^{|y_i|} W_{i,t}(p) = 1$ and by the definition of the mean $\mu_{y_i}(p) = \sum_{t=1}^{|y_i|} W_{i,t}(p) \log r_{i,t}$, we have

$$\sum_{t=1}^{|y_i|} W_{i,t}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big) = \mu_{y_i}(p) - \mu_{y_i}(p) = 0.$$

Multiplying this entire zero-sum by the constant $\mu_{y_i}(p)$ yields

$$\sum_{t=1}^{|y_i|} W_{i,t}(p)\mu_{y_i}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big) = 0.$$

We can subtract this identically zero term from our derivative equation without changing its value. By grouping the common factor $W_{i,t}(p)\big(\log r_{i,t} - \mu_{y_i}(p)\big)$, the equation collapses into a squared difference

$$\begin{aligned}\frac{\partial \mu_{y_i}(p)}{\partial p} &= \sum_{t=1}^{|y_i|} W_{i,t}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big) \log r_{i,t} - \sum_{t=1}^{|y_i|} W_{i,t}(p)\mu_{y_i}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big) \\ &= \sum_{t=1}^{|y_i|} W_{i,t}(p)\Big(\log r_{i,t} - \mu_{y_i}(p)\Big)^2.\end{aligned}$$

This final expression is exactly the definition of the variance of $\log r_{i,t}$ under the weight distribution W_i^p, denoted as $\mathrm{Var}_{W_i^p}(\log r_{i,t})$. Given our assumption that the sequence contains at least two tokens with differing importance ratios, this variance is strictly positive. □

Proof of Theorem 5. By definition, $\mathcal{T}^*$ contains the tokens with the strictly maximum importance ratio. Since the weighted sequence mean $\mu_{y_i}(p)$ is a convex combination of all token log-ratios (with $W_{i,k}(p) > 0$ for all finite p), it must be strictly bounded by the maximum value: $\mu_{y_i}(p) < \log r_{i,t^*}$ for any finite p. By Lemma 2, the derivative of the weight is governed by its deviation from this mean: $\frac{\partial W_{i,t^*}(p)}{\partial p} = W_{i,t^*}(p)(\log r_{i,t^*} - \mu_{y_i}(p))$. Because both the weight and the deviation are strictly positive, the weight of any optimal token increases monotonically as p grows.

Furthermore, Lemma 3 establishes that $\mu_{y_i}(p)$ is strictly monotonically increasing with p. Since it is continuously increasing and bounded above by the maximum log-ratio, it is convergent as $p \to +\infty$. By dividing the numerator and the denominator of $W_{i,t}(p)$ by r_{i,t^*}^p, where r_{i,t^*} is the maximum ratio, we rewrite the weight as

$$W_{i,t}(p) = \frac{\left(\frac{r_{i,t}}{r_{i,t^*}}\right)^p}{\sum_{k=1}^{|y_i|}\left(\frac{r_{i,k}}{r_{i,t^*}}\right)^p}.$$

For any optimal token $t^* \in \mathcal{T}^*$, the base is 1. For any sub-optimal token $k \notin \mathcal{T}^*$, the base is strictly less than 1, causing $(r_{i,k}/r_{i,t^*})^p \to 0$ as $p \to \infty$. Consequently, the denominator converges exactly to $|\mathcal{T}^*|$, the total number of optimal tokens. Thus, the weight distribution concentrates entirely on the optimal subset: $\lim_{p\to\infty} W_{i,t^*}(p) = 1/|\mathcal{T}^*|$ and $\lim_{p\to\infty} W_{i,k}(p) = 0$. Since the sequence mean is defined as $\mu_{y_i}(p) = \sum_t W_{i,t}(p)\log r_{i,t}$, taking the limit yields:

$$\lim_{p\to\infty}\mu_{y_i}(p) = \sum_{t\in\mathcal{T}^*}\left(\frac{1}{|\mathcal{T}^*|}\log r_{i,t^*}\right) + \sum_{k\notin\mathcal{T}^*}(0\cdot\log r_{i,k}) = \log r_{i,t^*}.$$

Combining this exact limit with Lemma 3, we establish that $\mu_{y_i}(p)$ strictly and monotonically approaches the maximum log-ratio $\log r_{i,t^*}$.

For any sub-optimal token $k \notin \mathcal{T}^*$, its log-ratio is strictly less than the maximum ($\log r_{i,k} < \log r_{i,t^*}$). Because $\mu_{y_i}(p)$ continuously sweeps upward towards $\log r_{i,t^*}$, there must exist a critical point p_t where the rising mean exactly crosses the token's log-ratio, yielding $\mu_{y_i}(p_t) = \log r_{i,k}$. For all $p > p_t$, the sequence mean surpasses the token's log-ratio ($\mu_{y_i}(p) > \log r_{i,k}$), which flips the sign of its derivative $\frac{\partial W_{i,k}(p)}{\partial p}$ to negative. Consequently, $W_{i,k}(p)$ reaches its peak at p_t and strictly decays thereafter. As $p \to \infty$, the exponential growth of the optimal tokens' weights strictly dominates the denominator, forcing the weight of all sub-optimal tokens to decay exactly to 0, and leaving the probability mass uniformly distributed exclusively among the optimal subset with weight $1/|\mathcal{T}^*|$. □

H.2 Global Property

In this section, we prove the following Theorem 1, which is mentioned in Section 3.2 as the global property of the sequence-level distributional deformation induced by the aggregation parameter p.

Theorem. *Assume the sequence y_i contains at least two tokens with distinct importance ratios. Then the Shannon entropy of the weight distribution attains its global maximum at $p = 0$, where $W_i^0 = \frac{1}{|y_i|}\mathrm{Unif}$, and strictly decreases as $|p|$ increases. Moreover, as $p \to \pm\infty$, W_i^p concentrates uniformly on the subset $\mathcal{T}^+ = \arg\max_t r_{i,t}(\theta)$ and $\mathcal{T}^- = \arg\min_t r_{i,t}(\theta)$, respectively.*

Proof. The Shannon entropy of the weight distribution is defined as

$$\mathcal{H}(W_i^p) := -\sum_t W_{i,t}(p)\ln W_{i,t}(p).$$

To analyse its monotonicity, we compute the derivative of $\mathcal{H}$ with respect to p. First, we compute the derivative of the token weight $W_{i,t}$. Let

$$\mathbb{E}_{W_i^p}[\ln r_{i,t}] := \sum_{k=1}^{|y_i|} W_{i,k}\ln r_{i,k}$$

denote the expected log-ratio under the current weight distribution. The derivative of the weight is given by

$$\frac{\partial W_{i,t}}{\partial p} = W_{i,t}\left(\ln r_{i,t} - \sum_k W_{i,k}\ln r_{i,k}\right) = W_{i,t}(\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]).$$

Next, taking the derivative of the entropy yields

$$\frac{\partial \mathcal{H}}{\partial p} = -\sum_t \left(\frac{\partial W_{i,t}}{\partial p}\ln W_{i,t} + W_{i,t}\frac{\partial \ln W_{i,t}}{\partial p}\right).$$

Notice that

$$\sum_t W_{i,t}\frac{\partial \ln W_{i,t}}{\partial p} = \sum_t \frac{\partial W_{i,t}}{\partial p} = \frac{\partial}{\partial p}\left(\sum_t W_{i,t}\right) = 0.$$

Thus, the entropy derivative simplifies to

$$\frac{\partial \mathcal{H}}{\partial p} = -\sum_t \frac{\partial W_{i,t}}{\partial p}\ln W_{i,t}.$$

To proceed, we explicitly write out $\ln W_{i,t}$. Let $Z(p) \coloneqq \sum_k r_{i,k}^p$. Since $W_{i,t} = r_{i,t}^p/Z(p)$, taking the natural logarithm gives

$$\ln W_{i,t} = p\ln r_{i,t} - \ln\sum_k r_{i,k}^p = p\ln r_{i,t} - \ln Z(p).$$

Substituting both $\frac{\partial W_{i,t}}{\partial p}$ and $\ln W_{i,t}$ into the simplified entropy derivative, we obtain

$$\frac{\partial \mathcal{H}}{\partial p} = -\sum_t \Big[W_{i,t}\big(\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\big)\Big]\Big[p\ln r_{i,t} - \ln Z(p)\Big].$$

We can expand this product into two separate sums. Notice that the expected deviation from the mean is identically zero. Specifically, because $\mathbb{E}_{W_i^p}[\ln r_{i,t}]$ is a constant with respect to the summation index t, we have

$$\sum_t W_{i,t}\big(\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\big) = \sum_t W_{i,t}\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\sum_t W_{i,t} = 0.$$

Because this term is zero, any constant multiplier distributed into it vanishes. When we distribute the expanded brackets, the term multiplied by the constant $\ln Z(p)$ completely drops out

$$\sum_t W_{i,t}\big(\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\big)\ln Z(p) = 0.$$

This leaves only the term

$$\frac{\partial \mathcal{H}}{\partial p} = -p\sum_t W_{i,t}\big(\ln r_{i,t} - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\big)\ln r_{i,t}.$$

Finally, we expand the remaining summation by distributing $\ln r_{i,t}$

$$\frac{\partial \mathcal{H}}{\partial p} = -p\left(\sum_t W_{i,t}(\ln r_{i,t})^2 - \mathbb{E}_{W_i^p}[\ln r_{i,t}]\sum_t W_{i,t}\ln r_{i,t}\right).$$

Recognising that $\sum_t W_{i,t}\ln r_{i,t}$ is exactly $\mathbb{E}_{W_i^p}[\ln r_{i,t}]$, this equation collapses into the definition of variance

$$\frac{\partial \mathcal{H}}{\partial p} = -p\left(\mathbb{E}_{W_i^p}[(\ln r_{i,t})^2] - \big(\mathbb{E}_{W_i^p}[\ln r_{i,t}]\big)^2\right) = -p\cdot \mathrm{Var}_{W_i^p}(\ln r_{i,t}).$$

Since the sequence contains non-uniform importance ratios, the variance is strictly positive ($\mathrm{Var}_{W_i^p}(\ln r_{i,t}) > 0$). Therefore for $p > 0$, $\frac{\partial \mathcal{H}}{\partial p} < 0$, meaning $\mathcal{H}$ strictly decreases. For $p < 0$, $\frac{\partial \mathcal{H}}{\partial p} > 0$, meaning $\mathcal{H}$ strictly increases towards $p = 0$ (or decreases as $p \to -\infty$). At $p = 0$, W_i^0

becomes a uniform distribution where each token is assigned an identical weight of $1/|y_i|$, and $\mathcal{H}$ reaches its global maximum $\ln |y_i|$.

Finally, we evaluate the limits as $p \to \pm\infty$. Let $r_{\max} = \max_t r_{i,t}$ and $\mathcal{M}^* = \{k \mid r_{i,k} = r_{\max}\}$. We can rewrite $W_{i,t}(p)$ by dividing the numerator and denominator by $r_{\max}^p$:

$$W_{i,t}(p) = \frac{(r_{i,t}/r_{\max})^p}{\sum_{k\in\mathcal{M}^*} 1 + \sum_{j\notin\mathcal{M}^*}(r_{i,j}/r_{\max})^p}.$$

For $j \notin \mathcal{M}^*$, $r_{i,j}/r_{\max} < 1$, so $(r_{i,j}/r_{\max})^p \to 0$ as $p \to \infty$. Thus, the denominator converges to $|\mathcal{M}^*|$. For the numerator, if $t \in \mathcal{M}^*$, $(r_{i,t}/r_{\max})^p = 1$. If $t \notin \mathcal{M}^*$, $(r_{i,t}/r_{\max})^p \to 0$. Therefore, $\lim_{p\to\infty} W_{i,t}(p) = \frac{1}{|\mathcal{M}^*|}$ for $t \in \mathcal{M}^*$, and 0 otherwise.

Let $r_{\min} = \min_t r_{i,t}$ and $\mathcal{M}_{\min} = \{k \mid r_{i,k} = r_{\min}\}$. Let $q = -p$, so $q \to \infty$. We can rewrite the weight as:

$$W_{i,t}(-q) = \frac{(1/r_{i,t})^q}{\sum_k (1/r_{i,k})^q} = \frac{(r_{\min}/r_{i,t})^q}{\sum_{k\in\mathcal{M}_{\min}} 1 + \sum_{j\notin\mathcal{M}_{\min}}(r_{\min}/r_{i,j})^q}.$$

Since $r_{\min}/r_{i,j} < 1$ for $j \notin \mathcal{M}_{\min}$, the term $(r_{\min}/r_{i,j})^q \to 0$ as $q \to \infty$. Following the exact same logic, the distribution collapses to a uniform distribution over $\mathcal{M}_{\min}$. □

H.3 Gradient Concentration vs. Exploration-Exploitation Trade-off

As established in Section 3.2, the aggregation parameter p induces three distinct gradient concentration regimes: upward concentration ($p > 0$) strictly allocates the gradient weights onto high-ratio tokens, uniform dispersion ($p \to 0$) distributes the gradient equally across all tokens, and downward concentration ($p < 0$) upweights the gradient contributions of low-ratio, hesitant tokens. Conceptually, dynamically shifting between these regimes closely mirrors the classical exploration-exploitation dilemma in reinforcement learning [Sutton and Barto, 2018]. However, the exact nature of this mechanism in the context of LLMs requires careful theoretical contextualisation.

Is a large p considered "Exploitation"? When $p \gg 0$, the algorithm hyper-focuses the gradient updates on tokens where the current policy has already shown the most aggressive improvement relative to the reference policy (i.e., maximal importance ratios). In traditional RL, exploitation implies a behavioural shift—acting greedily according to the current value function during environmental interaction. In our framework, however, a large p acts as a form of post-hoc exploitation that precisely targets two urgent algorithmic crises recently identified in LLM reasoning: the distribution sharpening trap and spurious rewards.

Recent work by [He et al., 2025] reveals that standard GRPO is fundamentally constrained by a distribution sharpening effect, predominantly rewarding tokens that are already likely while failing to amplify sparse, unlikely, yet correct reasoning leaps. Furthermore, [Shao et al., 2025] demonstrate that Reinforcement Learning with Verifiable Rewards (RLVR) is heavily plagued by spurious signals, where flawed intermediate logic coincidentally yields a correct final answer. When sequence-level advantages are distributed uniformly across the entire trajectory, the optimiser inevitably reinforces this uninformative or even toxic pseudo-logic.

Our upward concentration mechanism ($p > 0$) provides a mathematically principled resolution to both vulnerabilities. It does not exploit by altering the sampling trajectory, but by aggressively filtering the learning signal. By exponentially amplifying the gradient weights of rare, high-ratio tokens, it bypasses the sharpening trap to successfully obtain genuine "aha moments" [He et al., 2025]. Simultaneously, by starving the gradient from the bulk of unremarkable tokens, it naturally defends the policy against the integration of spurious background noise [Shao et al., 2025]. This theoretical intuition is vividly corroborated by our empirical results in Table 1: on the AIME24 benchmark, where correct reasoning steps are exceptionally sparse, a highly aggressive static configuration of $p = 3$ achieves the peak accuracy of 46.7%. Furthermore, as demonstrated in Figure 3, setting $p = +2$ rapidly drives the policy entropy down during the early training stages, visually confirming this intense knowledge-sharpening and mass-concentration effect.

Is a negative p considered "Exploration"? Conversely, when $p < 0$, the gradient concentration shifts toward tokens where the model exhibits hesitation or deviation from previously confident

paths. In standard continuous-control RL, exploration is typically enforced via entropy bonuses in the objective [Haarnoja et al., 2018, Schulman et al., 2017] or noise injection during sampling. While $p < 0$ empirically preserves reasoning diversity, it is not an exploration mechanism in the active sense. Instead, it serves as retrospective diversity preservation. By upweighting less-confident tokens within successful trajectories, it forces the optimiser to consolidate alternative, unconventional reasoning pathways rather than collapsing into a single, greedy solution. We observe this exact dynamic in Figure 3, where a static $p = -2$ sustains significantly higher entropy levels across the entire training trajectory compared to positive p values, actively resisting mode collapse. Moreover, Figure 2 illustrates that decreasing p systematically tightens the gap between the upper and lower envelopes of token-level ratios, redistributing credit to underemphasised tokens. This variance-controlling mechanism proves exceptionally beneficial for dense-signal tasks like MATH500, which strictly favours lower p values (peaking at $p = -1$) to maintain stable optimisation.

To formalise this critical boundary between our gradient mechanisms and traditional RL terminology, we state the following remark.

Remark 1. *While our concentration mechanism conceptually echoes the exploration-exploitation tradeoff, with $p < 0$ preserving diversity and $p > 0$ sharpening known knowledge, it must not be conflated with classical exploration. In standard RL, exploration refers to actively altering the trajectory sampling distribution (the behavioural policy) to visit unseen states. In contrast, our parameter p operates entirely on the hindsight aggregation of already-sampled trajectories. It functions strictly as a gradient reweighting mechanism, reshaping how the optimisation priority is distributed across a fixed rollout without intervening in how the rollouts are generated.*

I Variance Behaviours

This appendix provides analysis of the policy gradient variance under the HölderPO framework. We first establish a monotonic upper bound for the variance of the unclipped estimator in Section I.1, then immediately extend it to the sequence-level clipping case and formalise the structural necessity of sequence-level clipping in Section I.2. Subsequently, we derive a more refined variance expression in Section I.4 under the assumption of token-gradient orthogonality stated in Section I.3.

I.1 An upper bound with monotonicity

In this section, we prove another version of Theorem 2 for the unclipped gradient estimator (10).

Theorem 6. *Let $\widehat{\nabla}_\theta \mathcal{J}_H$ (Eq. (10)) denote the estimator. Assume $\|\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\| \le M$ for all tokens within the batch, the variance admits the bound*

$$\|\mathrm{Var}(\widehat{\nabla}_\theta \mathcal{J}_H)\| \;\le\; \frac{M^2}{B}\,\mathbb{E}\Big[\widehat{A}_i^2\,\rho_{i,p}^2(\theta)\Big], \tag{22}$$

which is monotonically increasing in p for all $p \in \mathbb{R}$, where B is the batch size.

Proof. We compute the unbiased estimator of the policy gradient by sampling a mini-batch of size B, denoted as $\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}}$. For a mini-batch containing B i.i.d. sampled trajectories, we have

$$\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}} = \frac{1}{B}\sum_{i=1}^{B} \hat{g}_i(p),$$

where the unclipped single-step stochastic gradient $\hat{g}_i(p)$ is

$$\hat{g}_i(p) = \widehat{A}_i \cdot \nabla_\theta \rho_{i,p}(\theta) = \widehat{A}_i \cdot \left[\frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|}\sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right].$$

Since every rollout in the mini-batch is independent, we obtain

$$\mathrm{Var}(\widehat{\nabla}_\theta \mathcal{J}_{\mathrm{H}}) = \mathrm{Var}\left(\frac{1}{B}\sum_{i=1}^{B}\hat{g}_i(p)\right) = \frac{1}{B^2}\sum_{i=1}^{B}\mathrm{Var}(\hat{g}_i(p)) = \frac{1}{B}\mathrm{Var}(\hat{g}_1(p)).$$

For any stochastic gradient, its variance $\text{Var}(\hat{g}_i(p))$ is strictly bounded by its second moment

$$\text{Var}(\hat{g}_i(p)) = \mathbb{E}[||\hat{g}_i(p)||^2] - ||\mathbb{E}[\hat{g}_i(p)]||^2 \leq \mathbb{E}[||\hat{g}_i(p)||^2].$$

By applying the triangle inequality and $||\nabla_\theta \log \pi_\theta(y_{i,t})|| \leq M$, we can obtain the upper bound

$$\begin{aligned}
\|\nabla_\theta \rho_{i,p}(\theta)\| &= \left\| \frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|} \sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \right\| \\
&\leq \frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|} \sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p \|\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\| \\
&\leq M \cdot \frac{\rho_{i,p}(\theta)^{1-p}}{|y_i|} \sum_{t=1}^{|y_i|} r_{i,t}(\theta)^p = M \cdot \rho_{i,p}(\theta).
\end{aligned}$$

Thus we can bound the squared norm of $\hat{g}_i(p)$ by

$$||\hat{g}_i(p)||^2 = \widehat{A}_i^2 \cdot ||\nabla_\theta \rho_{i,p}(\theta)||^2 \leq \widehat{A}_i^2 \cdot \left(M \cdot \rho_{i,p}(\theta)\right)^2 = M^2 \cdot \widehat{A}_i^2 \cdot \rho_{i,p}(\theta)^2,$$

which implies the upper bound of variance of $\widehat{\nabla}_\theta \mathcal{J}_{\text{H}}$

$$\text{Var}(\widehat{\nabla}_\theta \mathcal{J}_{\text{H}}) \leq \frac{1}{B} \mathbb{E}[||\hat{g}_i(p)||^2] \leq \frac{M^2}{B} \cdot \mathbb{E}_{q,\{y_i\}} \left[\widehat{A}_i^2 \cdot \rho_{i,p}(\theta)^2 \right].$$

According to the Generalised Mean Inequality, for any non-uniform sequence of importance ratios, the p-mean $\rho_{i,p}(\theta)$ is a strictly monotonically increasing function of p. Thus we obtain the conclusion. □

I.2 Variance and Sequence-level Clipping

To ensure training stability, policy optimisation algorithms typically employ clipping mechanisms to prevent destructively large updates. In the HölderPO framework, we explicitly adopt a *sequence-level clipping* strategy rather than the standard token-level clipping. In this section, we formalise the mathematical necessity of clipping itself, and justify why sequence-level clipping is structurally required to preserve the variance properties established in Theorem 6.

Why Clipping is Necessary. To understand why the unclipped case is susceptible to exponential explosion, we can analyse its gradient dynamics through an ordinary differential equation. Recall

$$\nabla_\theta \rho_{i,p}(\theta) \;=\; \rho_{i,p}(\theta) \sum_{t=1}^{|y_i|} W_{i,t}(p) \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}).$$

By abstracting the weighted sum of token-level log-gradients into a single vector $g_w(\theta)$, the gradient equation simplifies to a form

$$\nabla_\theta \rho_{i,p}(\theta) = \rho_{i,p}(\theta) g_w(\theta).$$

During neural network training, the parameters θ are not static; they continuously evolve along an optimisation trajectory parameterised by a continuous virtual time τ. Applying the chain rule, the rate of change of the ratio with respect to this optimisation time is given by the derivative

$$\frac{d\rho_{i,p}}{d\tau} = \left(\nabla_\theta \rho_{i,p}(\theta)\right)^T \frac{d\theta}{d\tau}.$$

Assuming a standard gradient ascent step aimed at maximising a trajectory with a positive advantage, the parameter update direction $\frac{d\theta}{d\tau}$ naturally aligns with the gradient. Substituting our simplified gradient expression into the derivative yields

$$\frac{d\rho_{i,p}}{d\tau} = \left(\rho_{i,p} g_w(\theta)\right)^T \frac{d\theta}{d\tau} = \rho_{i,p} \left(g_w(\theta)^T \frac{d\theta}{d\tau} \right).$$

Because the optimiser attempts to increase the likelihood of these correct tokens, the update direction $\frac{d\theta}{d\tau}$ forms an acute angle with the composite gradient direction $g_w(\theta)$. This implies that their inner

product is a strictly positive scalar, which we can denote as $k(\tau) > 0$. Substituting this scalar reduces the complex optimisation dynamics into a canonical ODE for exponential growth

$$\frac{d\rho_{i,p}}{d\tau} = k(\tau)\rho_{i,p}.$$

Integrating this differential equation over the time interval $[0, \tau]$ provides the exact analytical solution

$$\rho_{i,p}(\tau) = \rho_{i,p}(0) \exp\left(\int_0^\tau k(t)\, dt\right).$$

This mathematically dictates that without an explicit clipping mechanism to interrupt the ODE, the scaling factor will inevitably suffer from an unbounded exponential explosion. Therefore, explicitly bounding the surrogate ratio via a clipping mechanism is an absolute prerequisite.

Proof of Theorem 2. The clipping operator acts as a binary sequence-level mask $\mathbb{I}_i(\theta) \in \{0, 1\}$ applied directly to the aggregated ratio $\rho_{i,p}(\theta)$ and its advantage $\widehat{A}_i$ (see Eq. (16)). Consequently, the clipped stochastic gradient $\hat{g}_i^{\text{clip}}(p)$ is either preserved in full (when $\mathbb{I}_i = 1$) or completely zeroed out (when $\mathbb{I}_i = 0$). Mathematically, this guarantees that the squared norm of the clipped gradient is universally bounded by the unclipped one:

$$\|\hat{g}_i^{\text{clip}}(p)\|^2 \;=\; \mathbb{I}_i(\theta) \cdot \|\hat{g}_i(p)\|^2 \;\le\; \|\hat{g}_i(p)\|^2.$$

Because the variance is bounded by the second moment, the upper bound we derived in Theorem 6 carries over:

$$\|\mathrm{Var}(\widehat{\nabla}_\theta \mathcal{J}_{H_s})\| \;\le\; \frac{1}{B}\mathbb{E}\big[\|\hat{g}_i^{\text{clip}}(p)\|^2\big] \;\le\; \frac{M^2}{B}\,\mathbb{E}\Big[\widehat{A}_i^2\,\rho_{i,p}^2(\theta)\Big].$$

Thus, the monotonic relationship between the variance bound and the parameter p remains intact. □

In contrast, *token-level clipping* applies the clipping operator *inside* the summation over individual tokens. It unpredictably alters specific token ratios, destroying the correspondence between the outer multiplier $\rho_{i,p}(\theta)^{1-p}$ and the inner weights $W_{i,t}(p)$ (see Eq. (14)). This structural fracture voids the monotonic upper bound derived above, making the variance highly uncontrolled. Our empirical results in Appendix D (Table 4) corroborate this: token-level clipping narrows the performance spread across p, confirming that the parameter p loses its tight, predictable control over gradient variance.

I.3 Approximate orthogonality of policy gradients

Assumption 1. *In long-horizon reasoning tasks, we assume that within a given sequence y_i, the policy gradients with respect to tokens at different positions are approximately orthogonal, i.e. $\mathbb{E}[g_t^T g_k] \approx 0$ ($g_t = \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$) for any two distinct tokens $t \neq k$ in y_i.*

This assumption is practical and well-founded in the context of LLMs due to two factors: the blessing of dimensionality and linguistic feature decoupling. First, geometrically, in a parameter space with billions of dimensions, any two distinct gradient vectors are statistically bound to be nearly orthogonal [Vershynin, 2018]. Second, from a linguistic and mechanistic perspective, tokens at different positions within a long sequence typically serve distinct semantic and syntactic functions (e.g., predicting a generic preposition versus a complex domain-specific entity). Recent advances in mechanistic interpretability reveal that Transformer feed-forward layers operate as sparse key-value memories, where distinct neurons exclusively fire for specific linguistic patterns [Geva et al., 2021, Bricken et al., 2023]. Consequently, the subset of parameters responsible for encoding and predicting these distinct tokens are largely disjoint. This functional specialisation ensures that the back-propagated learning signals for distinct tokens are routed to different parameter subspaces, naturally leading to approximately uncorrelated, orthogonal gradient directions.

I.4 Monotonicity of Variance

While Assumption 1 and Remark 2 establish that the token-level gradients are approximately orthogonal and uniformly bounded in practice, analysing the exact variance dynamics requires a formal mathematical model. To achieve this, we adopt a standard theoretical abstraction: we transition from the empirical approximations to an idealised setting where these conditions hold exactly. This formal idealisation allows us to decouple the intrinsic sequence-level aggregation behaviour from token-specific optimisation noise, paving the way for the analysis presented in Theorem 7.

Theorem 7. *Under the idealised assumption of exact token-level gradient orthogonality ($\mathbb{E}[g_t^T g_k] = 0$ for $t \neq k$) and a uniformly bounded expected gradient norm ($\mathbb{E}[||g_t||^2] = M^2$), the exact second moment (and proportionally, the variance) of the Hölder-aggregated policy gradient estimator $\hat{g}_i$ simplifies to:*

$$\mathbb{E}[||\hat{g}_i||^2] = \widehat{A}_i^2 M^2 \rho_{i,p}(\theta)^2 \sum_{t=1}^{|y_i|} W_{i,t}(p)^2.$$

Consequently, we have the following properties:

1. *As p decays from $+\infty$ to 0, the variance strictly decreases.*
2. *As $p \to -\infty$, the weight concentration index (HHI), defined as $\sum_t W_{i,t}(p)^2$, grows exponentially and collapses to 1, counteracting the decrease in the Hölder mean $\rho_{i,p}(\theta)^2$.*
3. *There exists a $p^* \leq 0$ that strictly minimises the variance.*

Proof of Theorem 7. Keeping symbols from the proof of Theorem 6 and Assumption 1, the Hölder-aggregated policy gradient for a single trajectory y_i is:

$$\hat{g}_i(p) = \widehat{A}_i \rho_{i,p}(\theta) \sum_{t=1}^{|y_i|} W_{i,t}(p) g_t.$$

We expand the squared L_2-norm of this estimator:

$$||\hat{g}_i(p)||^2 = \widehat{A}_i^2 \rho_{i,p}(\theta)^2 \left(\sum_{t=1}^{|y_i|} \sum_{k=1}^{|y_i|} W_{i,t}(p) W_{i,k}(p) g_t^T g_k \right).$$

Taking the expectation with respect to the trajectory distribution, and applying the idealised token-level gradient orthogonality ($\mathbb{E}[g_t^T g_k] = 0$ for $t \neq k$), all cross-terms vanish exactly. Using the idealised uniform expected magnitude assumption ($\mathbb{E}[||g_t||^2] = M^2$), we obtain the exact second moment:

$$\mathbb{E}[||\hat{g}_i(p)||^2] = \widehat{A}_i^2 \rho_{i,p}(\theta)^2 \sum_{t=1}^{|y_i|} W_{i,t}(p)^2 \mathbb{E}[||g_t||^2] = \widehat{A}_i^2 M^2 \rho_{i,p}(\theta)^2 \sum_{t=1}^{|y_i|} W_{i,t}(p)^2.$$

Recognising that $\sum_t W_{i,t}(p)^2$ is exactly the Herfindahl-Hirschman Index (HHI) of the weight distribution, denoted as $\mathcal{H}_{HHI}(p)$, we analyse the exact variance dynamics based on this factorisation $V(p) \propto \rho_{i,p}(\theta)^2 \cdot \mathcal{H}_{HHI}(p)$:

Proof of Property 1. As established in Lemma 1, the Hölder mean $\rho_{i,p}(\theta)$ is strictly monotonically increasing with respect to p. Concurrently, for $p > 0$, the weight distribution gradually disperses from a strict one-hot distribution (at $p \to +\infty$) towards a uniform distribution (at $p \to 0$). Because the uniform distribution globally minimises the HHI (where $\mathcal{H}_{HHI}(0) = 1/|y_i|$), $\mathcal{H}_{HHI}(p)$ is strictly monotonically decreasing as p decays from $+\infty$ to 0. Since both $\rho_{i,p}(\theta)^2$ and $\mathcal{H}_{HHI}(p)$ are strictly decreasing as p decreases in $(0, +\infty)$, their product $V(p)$ must strictly decrease.

Proof of Property 2. As $p \to -\infty$, the Hölder mechanism heavily upweights the minimum elements. The weight distribution collapses into a one-hot distribution centred exclusively on the token(s) with the minimum importance ratio. Consequently, $\lim_{p \to -\infty} W_{i,t_{\min}}(p) = 1$, which drives the concentration index $\mathcal{H}_{HHI}(p)$ to grow exponentially back to its maximum possible value of 1. This sharp exponential growth of the HHI acts as a strong regulariser, counteracting the continuing decay of the Hölder mean $\rho_{i,p}(\theta)^2$.

Proof of Property 3. Let $V(p) = \rho_{i,p}^2 \cdot \mathcal{H}_{HHI}(p)$ represent the variance objective. From Property 1, $V(p)$ is strictly monotonically increasing for all $p \in (0, +\infty)$, implying that the global minimum of $V(p)$ cannot exist in the positive domain. At $p = 0$, the variance evaluates to $V(0) = \rho_{i,0}^2 \cdot (1/|y_i|)$. As p decreases into the negative domain ($p < 0$), $\rho_{i,p}^2$ continues to decay, but $\mathcal{H}_{HHI}(p)$ begins to increase towards 1 (Property 2). Because $V(p)$ is a continuous function bounded from below (by 0) defined on the closed interval $[-\infty, 0]$, by the Extreme Value Theorem, it must attain a minimum. Since it strictly increases for $p > 0$, this global variance-minimising point p^* is strictly guaranteed to satisfy $p^* \leq 0$. □

Remark 2. *The assumption that $\mathbb{E}[||g_t||^2] \approx M^2$ (i.e., token-level policy gradients have homogeneously expected magnitudes) is both a standard simplification and practically well-founded for modern LLMs for two reasons.*

1. Architectural Normalisation: Modern LLMs heavily rely on RMSNorm or LayerNorm before the final classification head. This strictly bounds the magnitude of the hidden states, thereby stabilising the scale of the back-propagated log-probability gradients across different token positions.

2. Statistical Stationarity over Long Horizons: While specific tokens might incur momentary gradient spikes, the expected squared norm over the data distribution tends toward a stationary value M^2 because all tokens share the same underlying language modelling head and projection matrices.

J Quantitative Advantage of Dynamic Scheduling

Remark 3. *In Theorem 3, the exponential amplification of the sparse reward signal relies on the* pre-saturation condition $r_{i,t^*}^{p_{high}} \ll n-1$. *This inequality is not merely a mathematical artefact, but rather a formalisation of the early-phase training dynamics in long-horizon LLM reasoning. We clarify its physical meaning and empirical validity as follows.*

1. Mathematically, the term $r_{i,t^*}^{p_{high}}$ represents the amplified signal of the single correct reasoning token, while $n-1$ represents the aggregated background mass of the remaining unremarkable tokens in the sequence. When $r_{i,t^*}^{p_{high}} \gg n-1$, the weight W_{i,t^*} saturates to 1, meaning the model has already become overwhelmingly confident in this step, and the gradient is completely monopolised. Therefore, the pre-saturation condition ($\ll n-1$) defines the critical case where the policy has discovered a high-reward token but is not yet absolutely confident. It is precisely in this window that the model desperately needs the exponential gradient boost provided by p_{high} to escape the noise.

2. This condition is exceptionally easy to satisfy in modern LLM reasoning tasks (e.g., AIME or MATH) due to two structural factors:

- ***Massive Sequence Length** (n): Chain-of-Thought (CoT) trajectories are inherently long, often spanning hundreds or thousands of tokens ($n \sim 10^3$). Consequently, the background mass $n-1$ provides a massive buffer.*
- ***Early-Phase Low Confidence** (r_{i,t^*}): In the early stages of RLVR, finding the correct reasoning path is a rare event. Even when the model stumbles upon the correct logic, its generation probability π_θ is only marginally higher than the reference π_{ref}. Thus, the initial ratio r_{i,t^*} is moderately greater than 1, but absolutely not large enough to let its p-th power immediately overpower thousands of background tokens.*

3. Crucially, the pre-saturation condition justifies our dynamic scheduling design. As training progresses, the model fits the correct trajectory, and r_{i,t^*} grows. Eventually, the condition $r_{i,t^*}^{p_{high}} \ll n-1$ will be violated (saturation occurs), rendering p_{high} mathematically ineffective at further isolating the signal. Exactly at this point, our dynamic schedule seamlessly decays $p \to p_{low} \leq 0$, shifting the algorithmic focus from signal amplification to variance contraction (Theorem 3, Part 2).

Proof of Theorem 3. Let $R \coloneqq r_{i,t^*} \gg 1$. For the remaining tokens $t \neq t^*$, since their ratios are constant-bounded, we can denote their sum of p-th powers as $S(p) \coloneqq \sum_{t \neq t^*} r_{i,t}^p = \Theta(n-1)$, which holds uniformly for any p in a bounded interval $[p_{\text{low}}, p_{\text{high}}]$. By definition, the weight of the high-ratio token is:

$$W_{i,t^*}(p) \;=\; \frac{R^p}{R^p + S(p)}.$$

Therefore, the relative amplification of the gradient weight when shifting from p_{stat} to p_{high} is given by:

$$\begin{aligned}\frac{W_{i,t^*}(p_{\text{high}})}{W_{i,t^*}(p_{\text{stat}})} \;&=\; \frac{R^{p_{\text{high}}}}{R^{p_{\text{high}}} + S(p_{\text{high}})} \cdot \frac{R^{p_{\text{stat}}} + S(p_{\text{stat}})}{R^{p_{\text{stat}}}} \\ &=\; R^{\,p_{\text{high}} - p_{\text{stat}}} \cdot \frac{R^{p_{\text{stat}}} + S(p_{\text{stat}})}{R^{p_{\text{high}}} + S(p_{\text{high}})}.\end{aligned}$$

Under the pre-saturation condition $R^{p_{\text{high}}} \ll n-1$, the term $R^{p_{\text{high}}}$ is asymptotically dominated by the denominator's background sum $S(p_{\text{high}}) = \Theta(n-1)$. Since $p_{\text{stat}} < p_{\text{high}}$, we naturally also have $R^{p_{\text{stat}}} \ll n-1$. Consequently, the fractional multiplier is bounded from below by a strictly positive constant $C = \Theta(1)$:

$$\frac{R^{p_{\text{stat}}} + S(p_{\text{stat}})}{R^{p_{\text{high}}} + S(p_{\text{high}})} \;\geq\; \frac{S(p_{\text{stat}})}{R^{p_{\text{high}}} + S(p_{\text{high}})} \;\geq\; C \;>\; 0.$$

Substituting $R = r_{i,t^*}$ back into the expression yields the desired exponential lower bound for the signal amplification:

$$\frac{W_{i,t^*}(p_{\text{high}})}{W_{i,t^*}(p_{\text{stat}})} \;\geq\; C \cdot r_{i,t^*}^{\,p_{\text{high}} - p_{\text{stat}}}.$$

By the definition provided in the theorem, the variance bound term is exactly $V(p) \coloneqq \mathbb{E}[\widehat{A}_i^2\, \rho_{i,p}^2(\theta)]$. Assuming the importance ratios within the sequence are non-degenerate (i.e., not all tokens share the exact same ratio), the generalised mean inequality guarantees that the Hölder mean $\rho_{i,p}(\theta)$ is strictly monotonically increasing with respect to p. Thus, for any $p_{\text{low}} < p_{\text{stat}}$, we have $\rho_{i,p_{\text{low}}}(\theta) < \rho_{i,p_{\text{stat}}}(\theta)$ pointwise for every sequence y_i. Since the squared advantage $\widehat{A}_i^2 \geq 0$ (and is strictly positive for meaningful updates), squaring the strictly positive Hölder means yields the following pointwise inequality for the random variables:

$$\widehat{A}_i^2\, \rho_{i,p_{\text{low}}}^2(\theta) \;<\; \widehat{A}_i^2\, \rho_{i,p_{\text{stat}}}^2(\theta).$$

Taking the expectation over the trajectory distribution strictly preserves this inequality, yielding:

$$\mathbb{E}\Big[\widehat{A}_i^2\, \rho_{i,p_{\text{low}}}^2(\theta)\Big] \;<\; \mathbb{E}\Big[\widehat{A}_i^2\, \rho_{i,p_{\text{stat}}}^2(\theta)\Big],$$

which directly concludes that $V(p_{\text{low}}) < V(p_{\text{stat}})$. □

K Broader Impacts

By improving the efficiency and stability of RL post-training, HölderPO can reduce the compute required to reach competitive performance on complex reasoning benchmarks, lowering the barrier for researchers and practitioners to develop capable reasoning models. Like any policy optimisation method, it inherits the standard dual-use risks of strong LLMs, including potential misuse for misinformation or automated content generation. A concern more specific to our framework is that the gradient amplification in the positive-pp p regime can intensify reward hacking when learning signals are misspecified, a limitation we discuss explicitly in Section 5.