UNDERSTANDING REWARD NORMALIZATION IN GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) has become a widely used and effective method for post-training large language models, yet the role of its within-group reward normalization remains underexplored. We identify reward discreteness as a key structure and develop a finite-sample theory for a family of normalized policy-gradient estimators. For binary rewards, we further show that the expected finite-group estimator is exactly collinear with its population counterpart. Our analysis reveals a group-size law that controls finite-group accuracy and motivates an adaptive sampler that assigns more rollouts to unresolved low-variance prompts. We interpret GRPO normalization as prompt reweighting: it strengthens low-variance prompt contributions while controlling aggregate gradient noise and exposes a tradeoff under more aggressive normalization. Finally, we analyze Balanced Aggregation, an existing token-aggregation rule, and prove that for binary rewards it preserves the population policy-gradient direction and obeys an analogous group-size law. Experiments on mathematical reasoning tasks compare normalization schemes and evaluate adaptive sampling.

1 INTRODUCTION

Large language models acquire broad capabilities through pretraining on large-scale corpora, but their reasoning abilities can be further improved during post-training. Preference-based post-training relies on human- or AI-labeled comparisons, using them either to train an explicit reward model for subsequent reinforcement learning (Ouyang et al., 2022; Bai et al., 2022) or to optimize the policy directly through objectives such as Direct Preference Optimization (DPO) (Rafailov et al., 2023). In domains with automatically verifiable outcomes, reinforcement learning with verifiable rewards (RLVR) offers a different approach: it obtains rewards directly from a verifier, for example through answer checking in mathematics or execution-based evaluation in code (Shao et al., 2024; Le et al., 2022). By replacing a learned reward model with an automatic verifier, RLVR simplifies reward construction. However, PPO-style RLVR still requires a learned critic to estimate the value function and construct lower-variance advantage estimates (Schulman et al., 2017; Ouyang et al., 2022). Group Relative Policy Optimization (GRPO) further removes this critic: for each prompt, it samples a group of responses and uses their within-group rewards to construct relative advantages (Shao et al., 2024; Guo et al., 2025).

For a fixed prompt, subtracting the population mean reward is a classical REINFORCE baseline that leaves the expected policy gradient direction unchanged (Williams, 1992). GRPO estimates this baseline using the empirical mean of a group of responses and, crucially, also divides the centered rewards by their empirical within-group standard deviation. Since this normalizer varies across prompts, it reweights their contributions to the aggregate update. This raises two central questions. First, because the normalizer is itself estimated from a random finite group, how accurately does the resulting estimator approximate its population-normalized counterpart? Second, how does the induced prompt reweighting affect optimization across prompts? We study both these statistical and optimization effects of GRPO in this paper.

Several recent analyses illuminate parts of this picture. Population normalization can be interpreted through transformed per-prompt objectives or as a prompt-wise adaptive gradient governed by local curvature (Davis & Recht, 2025; Thrampoulidis et al., 2026; Ge et al., 2026). The unnormalized group-centered estimator admits a U-statistic representation, which explains its oracle-baseline be-

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

havior and supports finite-sample and group-size analyses (Zhou et al., 2026). For binary rewards, Bay & Yearick (2026) prove an exact identity for a realized GRPO group in any policy dimension: the update is the empirical reward standard deviation times a correct-versus-incorrect score contrast. Their subsequent expected attenuation and fidelity calculations specialize to a one-dimensional Bernoulli-logit prompt. These perspectives establish that reward normalization changes both the effective objective and the use of sampled groups, but leave open a unified finite-sample analysis of the empirical normalizer for general policies, general discrete rewards, and normalization strengths beyond standard GRPO.

Our main contributions are:

- We study the σ^α-normalized family for $\alpha \in [0, 2]$. Exploiting reward discreteness, we establish finite-group estimation and bias guarantees for general discrete rewards, and further characterize the collinearity of the bias for binary rewards. The analysis shows that finite-group accuracy relative to a natural signal scale is governed by $K\sigma^2$.
- We formulate this dependence on $K\sigma^2$ as a group-size law and use it to motivate a doubling sampler that allocates additional rollouts to unresolved prompts.
- We interpret normalization as prompt reweighting and analyze its interaction with aggregate gradient noise, providing an optimization perspective on a potential benefit of GRPO ($\alpha = 1$).
- We give a finite-group theory for Balanced Aggregation (Zeng et al., 2026). For binary rewards, we show population collinearity and establish estimation-error and bias bounds with the same group-size dependence as reward normalization.

Experiments on mathematical reasoning are consistent with our theoretical findings. Full training curves for a 0.5B model compare fixed and adaptive normalization, while experiments with a 1.5B model evaluate these methods across three benchmarks.

2 Related Work

Critic-free policy optimization for language models. GRPO is closely related to score-function policy gradients and critic-free, sequence-level methods such as RLOO (Williams, 1992; Kool et al., 2019; Ahmadian et al., 2024). It combines a within-prompt baseline and standard-deviation normalization with a PPO-style surrogate (Schulman et al., 2017; Shao et al., 2024). DAPO and Dr.GRPO modify sampling and aggregation choices around this core (Yu et al., 2025; Liu et al., 2025). Other recent methods primarily change clipping, importance correction, or the policy-improvement objective, including GSPO, CISPO, DPPO, TIC-GRPO, and BPO (Zheng et al., 2025; MiniMax, 2025; Qi et al., 2026; Pang et al., 2025; Song et al., 2026).

Objectives and finite-group normalization. For binary rewards, group-based estimators can be interpreted as optimizing monotone transforms of the per-prompt success probability (Davis & Recht, 2025; Thrampoulidis et al., 2026). Population standard-deviation normalization has also been studied as a prompt-wise adaptive gradient through local curvature (Ge et al., 2026), while alternative advantage maps deliberately alter its prompt weighting (Heckel et al., 2026; Hernandez et al., 2026). Complementary work treats the unnormalized centered estimator as a U-statistic and obtains oracle-baseline and group-size results (Zhou et al., 2026). Empirical ablations further suggest that filtering homogeneous groups can sometimes matter more than normalization itself (Xiong et al., 2025a).

Adaptive sampling and response length. DAPO filters homogeneous groups and samples additional prompts to maintain a fixed number of effective groups, Reinforce-Ada adaptively allocates additional rollouts to unresolved prompts, sGPO sets rollout group sizes using offline estimates of prompt difficulty, and PODS selects informative rollouts from a larger generated pool (Yu et al., 2025; Xiong et al., 2025b; Sudalairaj et al., 2026; Xu et al., 2026). Other methods construct new signals for zero-variance prompts rather than resolving the original normalized estimator (Le et al., 2026). On length aggregation, Dr.GRPO removes the response-dependent denominator used by per-response token averaging, whereas DAPO uses a global token-level denominator (Liu et al., 2025; Yu et al., 2025). Balanced Aggregation averages positive and negative token losses separately to

control sign–length coupling and demonstrates empirical gains (Zeng et al., 2026). An impossibility result shows that trajectory-wise length weights cannot in general be both raw-gradient unbiased and length invariant (Ding et al., 2026).

3 Problem Setup

We study reinforcement learning with verifiable rewards. For a prompt x, a randomized response (or rollout) is a finite token sequence $Y = (Y_1, \ldots, Y_{L(Y)})$ of length $L(Y)$ and is sampled from an autoregressive policy π_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$. For a realized response $y = (y_1, \ldots, y_{L(y)})$,

$$\pi_\theta(y \mid x) = \prod_{t=1}^{L(y)} \pi_\theta(y_t \mid x, y_{<t}). \tag{1}$$

An external verifier $r(\cdot, \cdot)$ assigns the response a scalar reward $r(x, Y)$. We write

$$R := r(x, Y) \tag{2}$$

for the induced reward random variable, suppressing its dependence on x and θ. The verifier is fixed and the reward function has no explicit dependence on θ.

Expected reward and policy gradient. Define the per-prompt expected reward and reward variance as

$$\mu_\theta(x) := \mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[R], \qquad \sigma^2_\theta(x) := \mathrm{Var}_{Y \sim \pi_\theta(\cdot|x)}[R]. \tag{3}$$

The learning objective is to maximize $\mathbb{E}_{x \sim \mathcal{D}}[\mu_\theta(x)]$ for an underlying distribution $\mathcal{D}$. The score function of a realized response y is

$$s_\theta(y \mid x) := \nabla_\theta \log \pi_\theta(y \mid x) = \sum_{t=1}^{L(y)} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}). \tag{4}$$

Under the usual regularity conditions that permit differentiation under the expectation, $\mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[s_\theta(Y \mid x)] = 0$. The score-function identity thus gives the population policy gradient

$$g_\theta(x) := \nabla_\theta \mu_\theta(x) = \mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[R \cdot s_\theta(Y \mid x)] = \mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[(R - \mu_\theta(x)) s_\theta(Y \mid x)]\,. \tag{5}$$

This is the response-level REINFORCE gradient with the population mean reward as a baseline (Williams, 1992).

For $\alpha \in [0, 2]$, we define the σ^α-normalized population direction $g_{\alpha,\theta}(x) := \sigma_\theta(x)^{-\alpha} g_\theta(x)$ if $\sigma_\theta(x) > 0$ and 0 otherwise. We use the same zero-variance convention for empirical directions below and, throughout the paper, set $0^0 := 0$ whenever it is a power of a zero variance. Thus $\alpha = 0$ recovers the mean-centered policy gradient, while $\alpha = 1$ is the population counterpart of the standard-deviation normalization used by GRPO (Shao et al., 2024). We call $g_{\alpha,\theta}(x)$ a direction rather than a gradient because, for general rewards, $\sigma_\theta(x)$ is not determined solely by $\mu_\theta(x)$. Consequently, the parameter-dependent vector field $\sigma_\theta(x)^{-\alpha} \nabla_\theta \mu_\theta(x)$ need not equal $\nabla_\theta F(\theta)$ for any scalar objective F. Binary reward is a special case because $\sigma^2_\theta(x) = \mu_\theta(x)(1 - \mu_\theta(x))$.

Finite-group estimators. Let $K \geq 2$ and draw independent responses $Y_1, \ldots, Y_K \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid x)$. Write $R_j = r(x, Y_j)$ and $s_j = s_\theta(Y_j \mid x)$. The empirical reward mean and the GRPO group variance are

$$\widehat{\mu}_K := \frac{1}{K} \sum_{j=1}^{K} R_j, \qquad \widehat{\sigma}^2_K := \frac{1}{K} \sum_{j=1}^{K} (R_j - \widehat{\mu}_K)^2. \tag{6}$$

Define the empirical centered direction and its σ^α-normalized version

$$\widehat{g}_K := \frac{1}{K} \sum_{j=1}^{K} (R_j - \widehat{\mu}_K) s_j, \qquad \widehat{g}_{\alpha,K} := \begin{cases} \widehat{\sigma}_K^{-\alpha} \widehat{g}_K, & \widehat{\sigma}_K > 0, \\ 0, & \widehat{\sigma}_K = 0. \end{cases} \tag{7}$$

In particular, $\widehat{g}_{1,K}$ is the unclipped, on-policy, response-level GRPO direction.

Scope. Our setup describes response-level on-policy directions. They abstract away clipping, importance ratios, KL regularization, and token-level loss aggregation. Standard GRPO first averages token losses within each response; Balanced Aggregation and its finite-group behavior are treated separately in Section 6.

Assumption 1 (Bounded response score)**.** *There exists $B_s < \infty$ such that, for all parameters under consideration,*

$$\|s_\theta(Y \mid x)\|_2 \le B_s \qquad \textit{almost surely}. \tag{8}$$

Assumption 2 (Discrete rewards)**.** *For some $R_{\max} > 0$, the reward takes values in a finite set $\mathcal{R} \subseteq [0, R_{\max}]$ with minimum spacing $\Delta > 0$: for any distinct $r, r' \in \mathcal{R}$, $|r - r'| \ge \Delta$.*

For integer-valued rewards, we may take $\Delta = 1$; binary verifiable rewards $\mathcal{R} = \{0, 1\}$ are the principal example. This discreteness assumption makes the first absolute centered moment of the reward self-bounding by its variance. It establishes that normalization by powers as large as $\sigma_\theta^2(x)$ can remain controlled.

Proposition 1 (Self-bounding centered moments)**.** *Under Assumption 2,*

$$\mathbb{E}_{Y\sim\pi_\theta(\cdot|x)}|R - \mu_\theta(x)| \le \frac{2\sigma_\theta(x)^2}{\Delta}, \qquad \frac{1}{K}\sum_{j=1}^{K}|R_j - \widehat{\mu}_K| \le \frac{2\widehat{\sigma}_K^2}{\Delta}. \tag{9}$$

Consequently, under Assumption 1, for every $\alpha \in [0, 2]$, whenever the corresponding standard deviation is positive, $\|g_{\alpha,\theta}(x)\|_2 \le 2B_s\sigma_\theta(x)^{2-\alpha}/\Delta$ and $\|\widehat{g}_{\alpha,K}\|_2 \le 2B_s\widehat{\sigma}_K^{2-\alpha}/\Delta$. For integer-valued rewards, standard GRPO therefore obeys $\|g_{1,\theta}(x)\|_2 \le 2B_s\sigma_\theta(x)$, while $\|g_{2,\theta}(x)\|_2 \le 2B_s$. For binary rewards with success probability $\mu_\theta(x)$, the population moment bound is exact: $\mathbb{E}|R - \mu_\theta(x)| = 2\mu_\theta(x)(1 - \mu_\theta(x)) = 2\sigma_\theta(x)^2$. The proof is deferred to Appendix A.

4 Finite-Group Accuracy and Adaptive Sampling

We now quantify how accurately a finite rollout group estimates the σ^α-normalized population direction. Throughout this section we fix a prompt x and parameter θ, suppress their subscripts, and write $\mu = \mu_\theta(x)$, $\sigma = \sigma_\theta(x)$, $g_\alpha = g_{\alpha,\theta}(x)$, and $\widehat{g}_{\alpha,K}$ as in equation 7. All expectations below are over the K i.i.d. rollouts used to construct the estimator. The proofs are deferred to Appendix B. This complements U-statistic analyses of the unnormalized centered estimator (Zhou et al., 2026) and binary finite-group objective analyses (Davis & Recht, 2025; Bay & Yearick, 2026) by allowing a random empirical σ^α denominator and general discrete rewards.

4.1 Estimation error for general discrete rewards

Our first result gives an upper bound on the difference between $\widehat{g}_{\alpha,K}$ and g_α.

Theorem 1 (Finite-group estimation error)**.** *Under Assumptions 1 and 2, for every $K \ge 2$, $\alpha \in [0, 2]$, and $\sigma > 0$, $\mathbb{E}\big[\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2\big] \le A_\alpha B_s^2\sigma^{2-2\alpha}/K$, where $A_\alpha := 2^{\alpha+7}(1 + \alpha^2)\left(1 + \frac{R_{\max}^2}{\Delta^2}\right)$.*

The bound exposes two effects of normalization. For a prompt with small reward standard deviation σ, increasing α enlarges the natural upper scale of both the absolute estimation error $\|\widehat{g}_{\alpha,K} - g_\alpha\|_2$ and the normalized signal $\|g_\alpha\|_2$.

For $\alpha \in [1, 2]$, the bound in Theorem 1 does not vanish as $\sigma \to 0$. This does not reflect the actual behavior of the estimator in the near-degenerate regime. A direct application of the self-bounding property gives a complementary bound that vanishes for every $\alpha < 2$.

Proposition 2 (Near-degenerate estimation error)**.** *Under Assumptions 1 and 2, for every $K \ge 2$, $\alpha \in [1, 2]$, and $\sigma > 0$, $\mathbb{E}\big[\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2\big] \le 16B_s^2\sigma^{4-2\alpha}/\Delta^2$.*

Combining the two results, for $\alpha \in [1, 2]$ the squared estimation error is at most $B_s^2 \min\{A_\alpha\sigma^{2-2\alpha}/K, 16\sigma^{4-2\alpha}/\Delta^2\}$. Theorem 1 gives the smaller branch when $K\sigma^2 \ge A_\alpha\Delta^2/16$, whereas Proposition 2 is smaller when $K\sigma^2 \le A_\alpha\Delta^2/16$. In particular, for $\alpha = 1$ the two branches scale as $1/K$ and σ^2, respectively. When $\alpha \in [1, 2)$, the two-branch bound converges uniformly to 0 over σ as $K \to \infty$. In contrast, at the endpoint $\alpha = 2$, the two-branch bound does not vanish uniformly over σ as $K \to \infty$; it converges to zero only when $K\sigma^2 \to \infty$.

4.2 BIAS FOR GENERAL DISCRETE REWARDS

Normalization by an empirical power of the standard deviation makes $\widehat{g}_{\alpha,K}$ a biased estimator of g_α. We first give an upper bound on the bias for general discrete rewards.

Theorem 2 (Bias of empirical σ^α normalization)**.** *Under Assumptions 1 and 2, for every $K \geq 2$, $\alpha \in [0,2]$, and $\sigma > 0$, $\left\|\mathbb{E}\widehat{g}_{\alpha,K} - g_\alpha\right\|_2 \leq B_s \min\left\{\frac{C_\alpha}{K\sigma^\alpha}, \frac{4\sigma^{2-\alpha}}{\Delta}\right\}$, where $C_\alpha := 2^{\frac{\alpha}{2}+6}(1+\alpha)R_{\max}\left(1+\frac{R_{\max}}{\Delta}\right)$.*

The first branch of the upper bound is smaller than the second whenever $K\sigma^2 \geq C_\alpha\Delta/4$; the second branch is smaller in the complementary near-degenerate regime. To compare the first branch across prompts with different reward variances, recall from Proposition 1 that $B_s\sigma^{2-\alpha}/\Delta$ is an upper scale for the signal norm $\|g_\alpha\|_2$. We therefore use this quantity as a proxy for signal size. Dividing the first branch by this proxy gives

$$\frac{\|\mathbb{E}\widehat{g}_{\alpha,K} - g_\alpha\|_2}{B_s\sigma^{2-\alpha}/\Delta} \leq \frac{C_\alpha\Delta}{K\sigma^2}. \tag{10}$$

Thus the same quantity $K\sigma^2$ controls both finite-group estimation error and proxy-normalized bias, for every $\alpha \in [0,2]$. Because Proposition 1 provides no matching lower bound on $\|g_\alpha\|_2$, equation 10 is not an exact relative-bias guarantee. For binary rewards, however, the bias is collinear with the population direction, which allows the next subsection to derive the exact relative bias.

4.3 EXACT BIAS STRUCTURE FOR BINARY REWARDS

For a binary verifier, the bias has considerably more structure. Let $R \sim \mathrm{Bernoulli}(p)$, $q := 1-p$, and $\sigma^2 = pq$. For standard GRPO, Bay & Yearick (2026) give an exact realized-group identity for arbitrary policy dimension, but their expected attenuation calculation specializes to a scalar Bernoulli-logit prompt. The result below establishes expected collinearity for an arbitrary differentiable policy and every $\alpha \in [0,2]$.

Theorem 3 (Binary collinearity and relative bias)**.** *Let $N := \sum_{j=1}^K R_j$ and $\widehat{p} := N/K$. For every $K \geq 2$, $p \in (0,1)$, and $\alpha \in [0,2]$, $\mathbb{E}\widehat{g}_{\alpha,K} = \phi_{\alpha,K}(p)g_\alpha$, where $\phi_{\alpha,K}(p) := \mathbb{E}\left[(\widehat{p}(1-\widehat{p}))^{1-\alpha/2}\right]/\sigma^{2-\alpha} \in [0,1]$. The relative bias is $1-\phi_{\alpha,K}(p)$, with the first-order expansion*

$$1 - \phi_{\alpha,K}(p) = \frac{c_\alpha(p)}{K} + \rho_{\alpha,K}(p), \qquad c_\alpha(p) := 1 - \frac{\alpha}{2} + \frac{\alpha(2-\alpha)(1-2p)^2}{8\sigma^2}, \tag{11}$$

where the second-order remainder satisfies $|\rho_{\alpha,K}(p)| \leq 6/(K\sigma^2)^2 + 8000/(K\sigma^2)^3$.

This theorem is stronger than a norm bound: empirical σ^α normalization changes only the magnitude of the expected update, not its direction. For standard GRPO ($\alpha = 1$), $c_1(p) = 1/(8\sigma^2)$, recovering the leading relative-bias term $1/(8K\sigma^2)$. More generally, $c_\alpha(p)/K \leq 3/(8K\sigma^2)$, so $0 \leq 1 - \phi_{\alpha,K}(p) \leq 3/(8K\sigma^2) + 6/(K\sigma^2)^2 + 8000/(K\sigma^2)^3$. Thus both the leading term and the remainder are controlled solely by the resolution parameter $K\sigma^2$.

4.4 THE GROUP-SIZE LAW AND ADAPTIVE SAMPLING

The preceding results identify $K\sigma^2$ as the common parameter controlling finite-group accuracy relative to a natural signal scale. Indeed, relative to the squared signal proxy $B_s^2\sigma^{4-2\alpha}/\Delta^2$, the first branch of the estimation-error bound scales as $1/(K\sigma^2)$; the proxy-normalized bias in equation 10 has the same dependence. In the binary case, Theorem 3 shows that the exact relative bias admits the same control. Thus our finite-group bounds improve as $K\sigma^2$ increases for every $\alpha \in [0,2]$. This is the *group-size law*.

Because the population variance is unknown during training, we use $K\widehat{\sigma}_K^2$ as its observable surrogate. Algorithm 1 doubles the group size while this statistic remains below a prescribed threshold. The procedure concentrates additional rollouts on prompts whose observed rewards remain unresolved, while reusing every previously sampled response. The cap $K_{\max}$ controls the worst-case sampling cost. This differs from filtering homogeneous groups in DAPO and from allocation rules

Algorithm 1 Doubling adaptive sampling for one prompt

Require: prompt x, policy π_θ, initial size K_0, maximum size $K_{\max}$, threshold $c > 0$
1: Sample K_0 responses and set $K \leftarrow K_0$
2: Compute the empirical reward variance $\widehat{\sigma}_K^2$
3: **while** $K < K_{\max}$ and $K\widehat{\sigma}_K^2 < c$ **do**
4: $K' \leftarrow \min\{2K, K_{\max}\}$
5: Sample $K' - K$ additional responses and set $K \leftarrow K'$
6: Recompute $\widehat{\sigma}_K^2$ from all K rewards
7: **end while**
8: **return** the K responses, their rewards, and K

based on a chosen nonlinear objective or an offline difficulty profile (Yu et al., 2025; Xiong et al., 2025b; Sudalairaj et al., 2026).

Let $\mathcal{K}$ be the deterministic set of all group sizes that the doubling schedule can examine and let $M := |\mathcal{K}|$. Then $M \le 1 + \lceil \log_2(K_{\max}/K_0) \rceil$. The next proposition gives a uniform guarantee over all these group sizes.

Proposition 3 (Population resolution at adaptive stopping)**.** *Suppose $2 \le K_0 \le K_{\max}$, and the rewards are i.i.d. and satisfy Assumption 2. Define $\bar{c} = \max\{c, (K_0 - 1)\Delta^2/K_0\}$. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$, any returned $K \in \mathcal{K}$ satisfying $K\widehat{\sigma}_K^2 \ge c$ also satisfies*

$$K\sigma^2 \ge \frac{\bar{c}}{e}\left(\frac{\delta}{M}\right)^{R_{\max}^2/\bar{c}}. \tag{12}$$

The proof is deferred to Appendix B.5. We also remark that applying standard concentration inequalities and a union bound over all possible group sizes gives, up to logarithmic factors, the same estimation-error dependence on K and σ as above at every candidate size and hence also at the group size returned by the algorithm.

5 Prompt Reweighting versus Aggregate Gradient Noise

The preceding section establishes when a finite rollout group accurately estimates a normalized direction. We now use those results to examine why normalization can help optimization. Fix prompts $x^1, \dots, x^N$, and write $g_\theta^i := g_\theta(x^i)$, $\sigma_\theta^i := \sigma_\theta(x^i) > 0$, and $g_{\alpha,\theta}^i := g_{\alpha,\theta}(x^i)$. Their aggregate normalized direction is $G_{\alpha,\theta} := \frac{1}{N}\sum_{i=1}^N g_{\alpha,\theta}^i$. This prompt-reweighting view is shared by transformed-objective and local-curvature interpretations of GRPO (Davis & Recht, 2025; Thrampoulidis et al., 2026; Ge et al., 2026); our focus is how that reweighting interacts with aggregate finite-rollout noise. Normalization gives low-variance prompts greater relative weight, so their signals are less easily obscured in the aggregate direction, but it also rescales the corresponding estimation error. To quantify this tradeoff, at a generic parameter θ, construct $\widehat{g}_{\alpha,K}^i(\theta)$ from an independent group of K rollouts for each prompt x^i, and define

$$\widehat{G}_{\alpha,K}(\theta) := \frac{1}{N}\sum_{i=1}^N \widehat{g}_{\alpha,K}^i(\theta), \qquad V_{\alpha,K}(\theta) := \mathbb{E}\left[\left\|\widehat{G}_{\alpha,K}(\theta) - \mathbb{E}[\widehat{G}_{\alpha,K}(\theta)]\right\|_2^2\right]. \tag{13}$$

The prompt set is held fixed, so this quantity measures only rollout noise. Since $\sigma_\theta^i \le R_{\max}/2$, independence and the bounds in Section 4 give

$$V_{\alpha,K}(\theta) \le \begin{cases} \dfrac{A_\alpha B_s^2}{NK}\left(\dfrac{R_{\max}}{2}\right)^{2-2\alpha}, & 0 \le \alpha \le 1, \\ \dfrac{B_s^2}{N}\left(\dfrac{A_\alpha}{K}\right)^{2-\alpha}\left(\dfrac{16}{\Delta^2}\right)^{\alpha-1}, & 1 < \alpha \le 2. \end{cases} \tag{14}$$

Thus normalization up to $\alpha = 1$ strengthens low-variance contributions while retaining the $1/(NK)$ worst-case rollout-noise rate. Beyond $\alpha = 1$, our uniform squared-noise upper bound becomes

$K^{-(2-\alpha)}$. Aggressive normalization with $\alpha \in (1, 2]$ therefore creates a trade-off: it further amplifies low-variance signals, but also worsens the noise bound. Which effect dominates depends on the distribution and directions of the prompt gradients, so the worst-case bounds alone cannot determine whether such aggressive normalization is beneficial or harmful.

Adaptive sampling does not make $\alpha = 2$ safe. A finite adaptive budget cannot rule out extremely low-variance prompts. This risk is attained even for binary rewards: in a Bernoulli-logit example with $p = 1/K$, the $\alpha = 2$ squared estimation error remains $\Theta(1)$ as $K \to \infty$. By contrast, $1 < \alpha < 2$ has the vanishing worst-case squared-error rate $K^{-(2-\alpha)}$, uniformly over σ at every candidate group size. Its noise therefore remains controlled under either fixed or adaptive sampling, making it a safer aggressive normalization.

Below, we restrict the optimization comparison to $\alpha \in \{0, 1\}$, using the same fixed K for all prompts, to reveal a potential benefit of standard GRPO ($\alpha = 1$).

5.1 Optimization of the finite-group objective

We now specialize to binary rewards and compare $\alpha = 0$ with standard GRPO, $\alpha = 1$. Let $p_\theta^i := \mathbb{P}_\theta(R = 1 \mid x^i)$. By Theorem 3,

$$\mathbb{E}\Big[\widehat{G}_{\alpha,K}(\theta)\Big] = \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha,K}(p_\theta^i) g_{\alpha,\theta}^i, \qquad 0 \le \phi_{\alpha,K}(p_\theta^i) \le 1. \tag{15}$$

Thus finite sampling preserves the direction of each prompt's contribution but attenuates its magnitude by a prompt-dependent factor.

For simplicity, we consider full-batch stochastic gradient ascent on these N fixed prompts,

$$\theta_{t+1} = \theta_t + \eta \widehat{G}_{\alpha,K}(\theta_t), \tag{16}$$

using fresh independent groups of K rollouts at every iteration. A similar argument extends to minibatch updates after adding the variance due to prompt subsampling. Throughout this subsection, we additionally assume that the score is differentiable and $\|\nabla_\theta s_\theta(Y \mid x)\|_{\mathrm{op}} \le B_h$ almost surely.

Proposition 4 (Finite-group stationarity)**.** *Under the assumptions above, let $\alpha \in \{0, 1\}$. If $0 < \eta \le 1/(B_h + 3B_s^2)$, then the iterates in equation 16 satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha,K}(p_{\theta_t}^i) g_{\alpha,\theta_t}^i \right\|_2^2 \le \frac{2\pi}{\eta T} + (B_h + 3B_s^2)\eta \frac{A_1 B_s^2}{NK}. \tag{17}$$

The right side is deliberately the same worst-case bound for both methods, where the second term uses a common rollout-noise envelope for $\alpha = 0$ and $\alpha = 1$; the method-specific constants, including a tighter bound for $\alpha = 0$, are given in the appendix. Their essential difference appears on the left: $\alpha = 0$ contains $\phi_{0,K}(p_\theta^i) g_{0,\theta}^i$, whereas $\alpha = 1$ contains $\phi_{1,K}(p_\theta^i) g_{0,\theta}^i / \sigma_\theta^i$. Once a prompt is sufficiently resolved that finite-group attenuation is mild, $\alpha = 1$ therefore magnifies its low-variance gradient contribution on the left-hand side. Thus, under the same upper bound, the $\alpha = 1$ inequality places a stronger constraint on the contributions of sufficiently resolved low-variance prompts than the $\alpha = 0$ inequality does. This conclusion applies only to the sum of the prompt gradients. Because gradients from different prompts may point in different directions and cancel one another, the bound alone does not guarantee improvement on every prompt separately.

6 Finite-Group Analysis of Balanced Aggregation

Token aggregation is distinct from reward normalization. For one prompt x with responses $Y_1, \ldots, Y_K$, standard GRPO's on-policy, unclipped loss is

$$\widehat{\mathcal{L}}_K^{\mathrm{GRPO}}(\theta) := -\frac{1}{K}\sum_{j=1}^{K} \frac{\widehat{A}_j}{L_j} \log \pi_\theta(Y_j \mid x), \qquad \log \pi_\theta(Y_j \mid x) = \sum_{t=1}^{L_j} \log \pi_\theta(Y_{j,t} \mid x, Y_{j,<t}). \tag{18}$$

Here $L_j = L(Y_j)$ and $\widehat{A}_j = (R_j - \widehat{\mu}_K)/\widehat{\sigma}_K$, with $\widehat{A}_j = 0$ when $\widehat{\sigma}_K = 0$. Treating these quantities as stop-gradients gives ascent direction $K^{-1}\sum_j \widehat{A}_j s_\theta(Y_j \mid x)/L_j$. At population level this targets $\sigma^{-1}\mathbb{E}[(R-\mu)s_\theta(Y \mid x)/L(Y)]$, generally not collinear with $g_\theta(x)/\sigma$; response-level length normalization can therefore distort the policy-gradient direction. Balanced Aggregation instead averages positive and negative token losses separately, reducing this sign–length coupling (Zeng et al., 2026). We analyze its finite-group behavior and σ^α-normalized extension.

The length-independent score bound in Assumption 1 is restrictive for variable-length responses because a response score sums token-level scores. We instead assume the length-aware condition $\|s_\theta(Y \mid x)\|_2 \le B_\ell L(Y)$, which allows the score norm to grow linearly with the response length. For binary rewards, write $p = \Pr(R=1)$, $q = 1-p$, $m_a = \mathbb{E}[s_\theta(Y \mid x) \mid R = a]$, $\ell_a = \mathbb{E}[L(Y) \mid R = a]$, and $r_a = m_a/\ell_a$. The population direction corresponding to Balanced Aggregation is

$$g^{\mathrm{BA}}_{\alpha,\theta}(x) := \sigma^{2-\alpha}(r_1 - r_0) = \left(\frac{q}{\ell_1} + \frac{p}{\ell_0}\right)\frac{g_\theta(x)}{\sigma^\alpha}. \tag{19}$$

Thus it is a positive scalar multiple of the population-normalized policy gradient for every differentiable policy, with the length-independent bound $\|g^{\mathrm{BA}}_{\alpha,\theta}(x)\|_2 \le 2B_\ell\sigma^{2-\alpha}$. Balanced Aggregation therefore preserves the direction while placing every prompt's signal under a common worst-case scale that does not depend on its typical response lengths. The remaining question is whether its random class-specific length denominators remain accurate for a finite group.

For a mixed group, let $N_a = \sum_{j=1}^K \mathbf{1}\{R_j = a\}$ and $\widehat{r}_a = (\sum_{j:R_j=a} s_j)/(\sum_{j:R_j=a} L_j)$, and define $\widehat{g}^{\mathrm{BA}}_{\alpha,K} := \widehat{\sigma}_K^{2-\alpha}(\widehat{r}_1 - \widehat{r}_0)$, set to zero for a homogeneous group. The case $\alpha = 1$ is Balanced Aggregation. The following result extends the group-size law to these classwise ratio estimators; the finite, group-size-independent regularity factors $\Lambda^{\mathrm{mse}}_{\mathrm{BA}}$ and $\Lambda^{\mathrm{bias}}_{\mathrm{BA}}$ are defined in Appendix D.

Theorem 4 (Accuracy of Balanced Aggregation)**.** *Suppose rewards are binary, $\mathbb{E}[L(Y)^2 \mid R = a] < \infty$ for $a \in \{0,1\}$, and the length-dependent score bound above holds. For every $K \ge 2$, $\alpha \in [1,2]$, and $\sigma > 0$, there exist constants $C^{\mathrm{mse}}_\alpha, C^{\mathrm{bias}}_\alpha > 0$, depending only on α, such that*

$$\mathbb{E}\|\widehat{g}^{\mathrm{BA}}_{\alpha,K} - g^{\mathrm{BA}}_{\alpha,\theta}(x)\|_2^2 \le C^{\mathrm{mse}}_\alpha(1+\Lambda^{\mathrm{mse}}_{\mathrm{BA}})B_\ell^2 \min\left\{\frac{1}{K\sigma^{2\alpha-2}}, \sigma^{4-2\alpha}\right\}, \tag{20}$$

$$\|\mathbb{E}\widehat{g}^{\mathrm{BA}}_{\alpha,K} - g^{\mathrm{BA}}_{\alpha,\theta}(x)\|_2 \le C^{\mathrm{bias}}_\alpha(1+\Lambda^{\mathrm{bias}}_{\mathrm{BA}})B_\ell \min\left\{\frac{1}{K\sigma^{\alpha}}, \sigma^{2-\alpha}\right\}. \tag{21}$$

7 Experiments

We use binary verifier rewards and report pass@1 as the mean reward of independent responses. Training is purely online, with fresh current-policy samples at every step. Each step uses at most 64 prompts, AdamW (Loshchilov & Hutter, 2019) with zero weight decay, a gradient-norm cap of 1, LoRA (Hu et al., 2022), and Transformers sampling with temperature 0.7, `top_p` $= 1$, `top_k` $= 0$, and repetition penalty 1. Evaluation uses the same decoding parameters. Fixed methods use four rollouts per prompt. Adaptive groups double from four to 64 until the empirical reward variance is positive, reusing previous responses; their mean group size is approximately 20, about five times the fixed rollout budget. Because sampled responses vary in length, we aggregate their losses using the standard per-response length normalization of GRPO. Thus the experiments test the qualitative predictions of our response-level theory within the standard token-level GRPO implementation, rather than exactly reproducing its estimator.

Exact attenuation. Figure 1 evaluates the relative bias in Theorem 3. Each curve fixes $K\sigma^2$ and varies K, confirming that attenuation weakens as $K\sigma^2$ grows rather than with group size alone.

0.5B model. We train Qwen2.5-0.5B-Instruct (Yang et al., 2024a) on GSM8K (Cobbe et al., 2021) for up to eight epochs using LoRA rank 8, a peak learning rate of 10^{-6}, a cosine schedule over 2340 steps with a 0.5% warmup ratio, and a maximum response length of 1024 tokens. We evaluate the full GSM8K test set every quarter epoch using four responses per prompt. Figure 2 shows that fixed σ normalization is usually slightly better than fixed σ^0, particularly late in training, whereas fixed σ^2 gives no consistent gain. In these runs, fixed $\sigma^{3/2}$ is the strongest fixed method over most of training, and both adaptive curves are generally above their fixed counterparts.

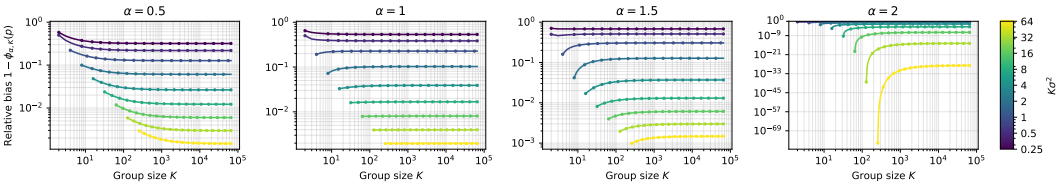

Figure 1: Exact relative bias $1 - \phi_{\alpha,K}$ for binary rewards. Colors index fixed values of $K\sigma^2$.

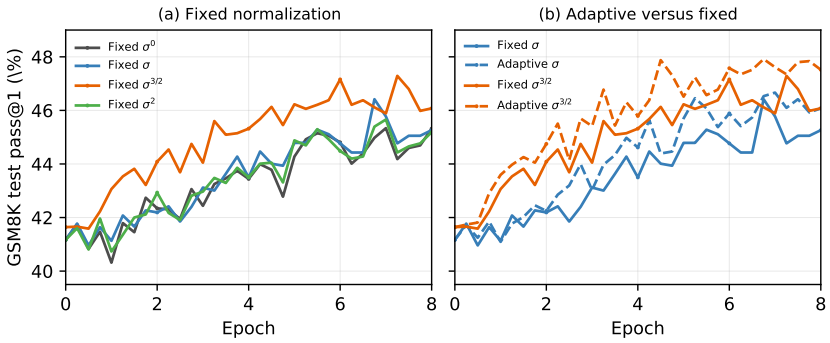

Figure 2: GSM8K test pass@1 during 0.5B training. Left: all fixed σ^α methods. Right: each adaptive method and its fixed counterpart.

1.5B model. We train Qwen2.5-Math-1.5B (Yang et al., 2024b) on DSR-sub (Wang et al., 2025) for ten epochs using LoRA rank 32, a constant learning rate of 10^{-5}, and a maximum response length of 3072 tokens. Table 1 compares the shared initial model with the epoch-10 checkpoints. We use 64 evaluation responses per prompt on MATH-500 (Lightman et al., 2024) and 32 on Minerva Math (Lewkowycz et al., 2022) and the 675-problem English, text-only mathematics split of OlympiadBench (He et al., 2024); the latter is scored with the official math judger. Every method improves substantially over the initial model. In our runs, adaptive sampling improves both σ and $\sigma^{3/2}$ normalization over their fixed counterparts on all three benchmarks. Adaptive $\sigma^{3/2}$ achieves the highest observed values throughout, reaching 61.12%, 22.15%, and 27.73%, respectively.

Table 1: Pass@1 (%) of Qwen2.5-Math-1.5B before training and after ten DSR-sub epochs.

Method	MATH-500	Minerva Math	OlympiadBench
Initial model (epoch 0)	50.68	14.14	24.34
Fixed σ^0	59.93	20.73	27.30
Fixed σ	59.45	20.65	27.21
Fixed $\sigma^{3/2}$	60.18	21.15	27.39
Fixed σ^2	59.96	21.05	27.33
Adaptive σ	60.46	20.96	27.62
Adaptive $\sigma^{3/2}$	**61.12**	**22.15**	**27.73**

8 Conclusion

We developed a finite-group theory of reward normalization in GRPO. We characterized the estimation error and bias of normalized directions and derived a group-size law that motivates adaptive sampling. We further explained normalization through prompt reweighting and aggregate gradient noise, and established analogous finite-group guarantees for Balanced Aggregation. Experiments on mathematical reasoning are consistent with the predicted effects of normalization strength and adaptive sampling. Together, these results clarify how reward normalization, rollout allocation, and token aggregation shape GRPO and its variants. Our experiments focus on binary verifier rewards; a natural future direction is to evaluate reward normalization with more general discrete rewards and extend the theory to continuous-valued verifier rewards.

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for formulating mathematical claims, proving mathematical claims, assisting in the writing of proof, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods and interpreting results. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, assisting with translation, cleaning and reformatting dataset and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, formatting references and proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. We checked LLM-generated mathematical proofs and code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yong Yi Bay and Kathleen A. Yearick. GRPO, Dr. GRPO, and DAPO are three operations on one number: The group-standard-deviation identity. *arXiv preprint arXiv:2607.00152*, 2026.

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

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv preprint arXiv:2510.13651*, 2025.

Fei Ding, Yongkang Zhang, Yuhao Liao, Zijian Zeng, and Huiming Yang. On the impossibility of unbiased and length-invariant policy optimization with outcome rewards. *arXiv preprint arXiv:2607.23364*, 2026.

Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why GRPO needs normalization: A local-curvature perspective on adaptive gradients. *arXiv preprint arXiv:2601.23135*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Reinhard Heckel, Mahdi Soltanolkotabi, and Christos Thrampoulidis. Asymmetric prompt weighting for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2602.11128*, 2026.

Jefferson Hernandez, Jaywon Koo, Zilin Xiao, Chen Wei, and Vicente Ordonez. SoftmaxGRPO: Learning to reason using softmax advantage group estimation. *arXiv preprint arXiv:2608.09271*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR 2019 Deep Reinforcement Learning meets Structured Prediction Workshop*, 2019.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. CodeRL: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

Thanh-Long V Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In *International Conference on Learning Representations*, 2026.

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.

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

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

MiniMax. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

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

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

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

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.

Zhuoqing Song, Haotian Xu, Xikun Zhang, and Lidong Bing. Bellman policy optimization. *arXiv preprint arXiv:2609.15987*, 2026.

Shivchander Sudalairaj, Kai Xu, Akash Srivastava, and Giorgio Giannone. sGPO: Trading inference FLOPs for training efficiency in RLVR. *arXiv preprint arXiv:2606.08854*, 2026.

Christos Thrampoulidis, Sadegh Mahdavi, and Wenlong Deng. Advantage Shaping as Surrogate Reward Maximization: Unifying Pass@K Policy Gradients. *Transactions on Machine Learning Research*, 2026.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

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

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. A minimalist approach to LLM reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*, 2025a.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-Ada: An adaptive sampling framework for REINFORCE-style LLM training. *arXiv preprint arXiv:2510.04996*, 2025b.

Yixuan Even Xu, Yash Savani, Fei Fang, and J Zico Kolter. Not all rollouts are useful: Down-sampling rollouts in LLM reinforcement learning. *Transactions on Machine Learning Research*, 2026.

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. *arXiv preprint arXiv:2412.15115*, 2024a.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2025.

Zhiyuan Zeng, Jiameng Huang, Zhangyue Yin, Jiashuo Liu, Ziniu Li, Bingrui Li, Yuhao Wu, Yining Zheng, Ge Zhang, Wenhao Huang, et al. Balanced aggregation: Understanding and fixing aggregation bias in GRPO. *arXiv preprint arXiv:2605.04077*, 2026.

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

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

A Proof of Proposition 1

Proposition 5 (Self-bounding moments and normalized gradients (Restatement of Proposition 1))**.** *Under Assumption 2,*

$$\mathbb{E}_{Y\sim\pi_\theta(\cdot|x)}|R-\mu_\theta(x)| \le \frac{2\sigma_\theta(x)^2}{\Delta}, \tag{22}$$

$$\frac{1}{K}\sum_{j=1}^{K}|R_j-\widehat{\mu}_K| \le \frac{2\widehat{\sigma}_K^2}{\Delta}. \tag{23}$$

If Assumption 1 also holds, then for every $\alpha\in[0,2]$ and $\sigma_\theta(x)>0$,

$$\|g_{\alpha,\theta}(x)\|_2 \le \frac{2B_s\sigma_\theta(x)^{2-\alpha}}{\Delta}. \tag{24}$$

Similarly, whenever $\widehat{\sigma}_K > 0$,

$$\|\widehat{g}_{\alpha,K}\|_2 \le \frac{2B_s\widehat{\sigma}_K^{2-\alpha}}{\Delta}. \tag{25}$$

Proof. For brevity, write $\mu = \mu_\theta(x)$, $\sigma = \sigma_\theta(x)$, and $s = s_\theta(Y \mid x)$. Let R' be an independent copy of R. Conditional Jensen's inequality and the reward spacing imply

$$\mathbb{E}|R-\mu| = \mathbb{E}\,|\mathbb{E}[R-R' \mid R]| \tag{26}$$

$$\le \mathbb{E}|R-R'| \tag{27}$$

$$\le \frac{1}{\Delta}\mathbb{E}(R-R')^2 = \frac{2\sigma^2}{\Delta}. \tag{28}$$

The penultimate inequality holds because any two values in $\mathcal{R}$ are either equal or at least Δ apart, so $|R-R'| \le (R-R')^2/\Delta$.

The same argument applies to the empirical distribution that assigns mass $1/K$ to each observed reward. Equivalently, directly,

$$\frac{1}{K}\sum_{j=1}^{K}|R_j-\widehat{\mu}_K| \le \frac{1}{K^2}\sum_{j,\ell=1}^{K}|R_j-R_\ell| \tag{29}$$

$$\le \frac{1}{\Delta K^2}\sum_{j,\ell=1}^{K}(R_j-R_\ell)^2 = \frac{2\widehat{\sigma}_K^2}{\Delta}. \tag{30}$$

Here the first inequality is conditional Jensen's inequality under the empirical distribution, and the last equality is the standard pairwise variance identity.

Using equation 5, Assumption 1, and equation 28,

$$\|g_\theta(x)\|_2 \le B_s\mathbb{E}|R-\mu| \le \frac{2B_s\sigma^2}{\Delta}. \tag{31}$$

Dividing by σ^α proves the population normalized-direction bound. Likewise, equation 7 and equation 30 give

$$\|\widehat{g}_K\|_2 \le \frac{B_s}{K}\sum_{j=1}^{K}|R_j-\widehat{\mu}_K| \le \frac{2B_s\widehat{\sigma}_K^2}{\Delta}, \tag{32}$$

and division by $\widehat{\sigma}_K^\alpha$ proves the empirical bound. Finally, if $R \sim \mathrm{Bernoulli}(p)$, then $\mathbb{E}|R-p| = 2p(1-p) = 2\sigma^2$, proving the binary identity stated after the proposition in the main text. □

B Proofs for Finite-Group Accuracy

This appendix restates and proves the results in Section 4. We continue to fix x and θ and suppress them from the notation. Put

$$X := R-\mu, \qquad v := \sigma^2, \qquad a := \frac{\alpha}{2}, \qquad s := s_\theta(Y \mid x). \tag{33}$$

For a group of size K, define

$$\overline{X} := \frac{1}{K}\sum_{j=1}^{K}X_j, \qquad \overline{s} := \frac{1}{K}\sum_{j=1}^{K}s_j, \qquad H := \frac{1}{K}\sum_{j=1}^{K}X_js_j, \qquad \widehat{v} := \widehat{\sigma}_K^2. \tag{34}$$

Then $g = \mathbb{E}H$, $\widehat{g}_K = H - \overline{X}\,\overline{s}$, $g_\alpha = gv^{-a}$, and $\widehat{g}_{\alpha,K} = \widehat{g}_K\widehat{v}^{-a}$ when $\widehat{v} > 0$.

We first record estimates used by both general-reward proofs. Since $|X| \le R_{\max}$, $\|s\|_2 \le B_s$, and $\mathbb{E}s = 0$,

$$\mathbb{E}\|\widehat{g}_K - g\|_2^2 \le \frac{4B_s^2v}{K}, \qquad \mathbb{E}(\widehat{v}-v)^2 \le \frac{2R_{\max}^2v}{K}. \tag{35}$$

For completeness, the first inequality follows from $\|u+w\|_2^2 \le 2\|u\|_2^2 + 2\|w\|_2^2$ applied to $H - g - \overline{X}\,\overline{s}$, together with

$$\mathbb{E}\|H-g\|_2^2 \le \frac{B_s^2 v}{K}, \qquad \mathbb{E}\|\overline{X}\,\overline{s}\|_2^2 \le \frac{B_s^2 v}{K}. \tag{36}$$

For the second inequality, the exact variance formula for the empirical variance with divisor K gives

$$\mathbb{E}(\widehat{v}-v)^2 = \frac{(K-1)^2}{K^3}\mathbb{E}X^4 - \frac{(K-1)(K-3)}{K^3}v^2 + \frac{v^2}{K^2} \le \frac{2R_{\max}^2 v}{K}, \tag{37}$$

where $\mathbb{E}X^4 \le R_{\max}^2 v$ was used in the last step.

B.1 Proof of Theorem 1

Theorem 5 (Finite-group estimation error (Restatement of Theorem 1))**.** *Under Assumptions 1 and 2, for every $K \ge 2$, $\alpha \in [0,2]$, and $\sigma > 0$,*

$$\mathbb{E}\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2 \le \frac{A_\alpha B_s^2 \sigma^{2-2\alpha}}{K}, \qquad A_\alpha = 2^{\alpha+7}(1+\alpha^2)\left(1 + \frac{R_{\max}^2}{\Delta^2}\right). \tag{38}$$

Proof. Assume $v > 0$ and introduce the event

$$\mathcal{E} := \{\widehat{v} \ge v/2\}. \tag{39}$$

By equation 35 and Chebyshev's inequality,

$$\mathbb{P}(\mathcal{E}^c) \le \frac{8R_{\max}^2}{Kv}. \tag{40}$$

On $\mathcal{E}$, the mean-value theorem gives

$$|\widehat{v}^{-a} - v^{-a}| \le a\,2^{a+1} v^{-a-1}|\widehat{v}-v|. \tag{41}$$

Combining equation 41, equation 35, and $\|g\|_2 \le 2B_s v/\Delta$ yields

$$\begin{aligned}&\mathbb{E}\big[\|\widehat{g}_K\widehat{v}^{-a} - g v^{-a}\|_2^2 \mathbf{1}_{\mathcal{E}}\big] \\ &\quad\le \frac{2^{\alpha+3}B_s^2 v^{1-\alpha}}{K} + \frac{2^{\alpha+4}\alpha^2 B_s^2 R_{\max}^2 v^{1-\alpha}}{K\Delta^2}.\end{aligned} \tag{42}$$

It remains to control the event on which the empirical variance is too small. On $\mathcal{E}^c$, Proposition 1 gives

$$\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2 \le \frac{8B_s^2}{\Delta^2}\left(\widehat{v}^{\,2-\alpha} + v^{2-\alpha}\right) \le \frac{16 B_s^2 v^{2-\alpha}}{\Delta^2}. \tag{43}$$

At $\alpha = 2$ and $\widehat{v} = 0$, the empirical term in this display is understood as zero, in accordance with the estimator's definition. Combining this inequality with equation 40 yields

$$\begin{aligned}&\mathbb{E}\big[\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2 \mathbf{1}_{\mathcal{E}^c}\big] \\ &\quad\le \frac{128 B_s^2 R_{\max}^2 v^{1-\alpha}}{K\Delta^2}.\end{aligned} \tag{44}$$

Adding equation 42 and equation 44, and enlarging the numerical constant, proves the claimed inequality because $v^{1-\alpha} = \sigma^{2-2\alpha}$. □

B.2 Proof of Proposition 2

Proposition 6 (Near-degenerate estimation error (Restatement of Proposition 2))**.** *Under the assumptions of Theorem 1, for every $K \ge 2$, $\alpha \in [1,2]$, and $\sigma > 0$,*

$$\mathbb{E}\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2 \le \frac{16 B_s^2 \sigma^{4-2\alpha}}{\Delta^2}. \tag{45}$$

Proof. Write $v = \sigma^2$. By the self-bounding inequalities in Proposition 1,

$$\|\widehat{g}_{\alpha,K}\|_2^2 \le \frac{4B_s^2}{\Delta^2}\widehat{v}^{\,2-\alpha}, \qquad \|g_\alpha\|_2^2 \le \frac{4B_s^2}{\Delta^2}v^{2-\alpha}. \tag{46}$$

For $\alpha < 2$, the map $z \mapsto z^{2-\alpha}$ is concave on $\mathbb{R}_+$, and $\mathbb{E}\widehat{v} = (K-1)v/K$; Jensen's inequality therefore gives

$$\mathbb{E}\widehat{v}^{\,2-\alpha} \le (\mathbb{E}\widehat{v})^{2-\alpha} \le v^{2-\alpha}. \tag{47}$$

At $\alpha = 2$, the left side is the probability that $\widehat{v} > 0$ under the zero-variance convention, so the same inequality holds directly. Therefore, using $\|u-w\|_2^2 \le 2\|u\|_2^2 + 2\|w\|_2^2$,

$$\mathbb{E}\|\widehat{g}_{\alpha,K} - g_\alpha\|_2^2 \le 2\mathbb{E}\|\widehat{g}_{\alpha,K}\|_2^2 + 2\|g_\alpha\|_2^2 \tag{48}$$

$$\le \frac{16B_s^2 v^{2-\alpha}}{\Delta^2} = \frac{16B_s^2\sigma^{4-2\alpha}}{\Delta^2}. \tag{49}$$

□

B.3 Proof of Theorem 2

Theorem 6 (Bias of empirical σ^α normalization (Restatement of Theorem 2))**.** *Under Assumptions 1 and 2, for every $K \ge 2$, $\alpha \in [0,2]$, and $\sigma > 0$,*

$$\|\mathbb{E}\widehat{g}_{\alpha,K} - g_\alpha\|_2 \le B_s \min\left\{\frac{C_\alpha}{K\sigma^\alpha}, \frac{4\sigma^{2-\alpha}}{\Delta}\right\}, \tag{50}$$

where

$$C_\alpha = 2^{\frac{\alpha}{2}+6}(1+\alpha)R_{\max}\left(1 + \frac{R_{\max}}{\Delta}\right). \tag{51}$$

Proof. The second branch follows immediately from the triangle inequality, Proposition 1, and concavity:

$$\|\mathbb{E}\widehat{g}_{\alpha,K} - g_\alpha\|_2 \le \mathbb{E}\|\widehat{g}_{\alpha,K}\|_2 + \|g_\alpha\|_2 \tag{52}$$

$$\le \frac{2B_s}{\Delta}\left(\mathbb{E}\widehat{v}^{\,1-a} + v^{1-a}\right) \le \frac{4B_s v^{1-a}}{\Delta}. \tag{53}$$

We prove the first branch by interpolation between the population moments and their empirical counterparts. Define

$$U := \frac{1}{K}\sum_{j=1}^K (X_j s_j - g), \quad V := \overline{X}, \quad W := \overline{s}, \quad Z := \frac{1}{K}\sum_{j=1}^K (X_j^2 - v), \tag{54}$$

and, for $t \in [0,1]$,

$$n_t := g + tU - t^2 VW, \qquad v_t := v + tZ - t^2 V^2. \tag{55}$$

Whenever $v_t > 0$, define $\Gamma(t) := n_t v_t^{-a}$. At the endpoint $t = 1$, if $v_1 = \widehat{v} = 0$, define $\Gamma(1) := \widehat{g}_{\alpha,K} = 0$, consistently with the definition of the empirical normalized direction. With this convention, $\Gamma(0) = g_\alpha$ and $\Gamma(1) = \widehat{g}_{\alpha,K}$ in all cases. The following moment bounds follow from independence, $|X| \le R_{\max}$, and $\|s\|_2 \le B_s$:

$$\mathbb{E}\|U\|_2^2 \le \frac{B_s^2 v}{K}, \qquad \mathbb{E}V^2 = \frac{v}{K}, \qquad \mathbb{E}\|W\|_2^2 \le \frac{B_s^2}{K},$$

$$\mathbb{E}Z^2 \le \frac{R_{\max}^2 v}{K}, \qquad \mathbb{E}V^4 \le \frac{4R_{\max}^2 v}{K^2}, \qquad \mathbb{E}[V^2\|W\|_2^2] \le \frac{B_s^2 v}{K}. \tag{56}$$

The path in equation 55 has an exact probabilistic interpretation. Let P denote the population distribution of (R,s), $\widehat{P}_K$ its empirical distribution, and $P_t = (1-t)P + t\widehat{P}_K$. Then

$$n_t = \mathrm{Cov}_{P_t}(R,s), \qquad v_t = \mathrm{Var}_{P_t}(R) = (1-t)v + t\widehat{v} + t(1-t)V^2. \tag{57}$$

Thus $v_t \geq v/2$ for every $t \in [0,1]$ on $\mathcal{E} = \{\widehat{v} \geq v/2\}$. Moreover, the reward under P_t has the same finite support and spacing, so

$$\|n_t\|_2 = \|\mathbb{E}_{P_t}[(R - \mathbb{E}_{P_t}[R])s]\,\|_2 \leq B_s \cdot \mathbb{E}_{P_t}[|R - \mathbb{E}_{P_t}[R]|] \leq \frac{2B_s v_t}{\Delta}. \tag{58}$$

A simple calculation gives $\mathbb{E}\Gamma'(0) = \mathbb{E}[v^{-a}U - Zav^{-a-1}g] = 0$. On $\mathcal{E}$, Taylor's formula with integral remainder gives

$$\Gamma(1) - \Gamma(0) = \Gamma'(0) + \int_0^1 (1-t)\Gamma''(t)\,dt. \tag{59}$$

Here

$$\Gamma''(t) = n_t'' v_t^{-a} - 2an_t' v_t^{-a-1} v_t' - an_t v_t^{-a-1} v_t'' \tag{60}$$
$$+ a(a+1) n_t v_t^{-a-2} (v_t')^2, \tag{61}$$

where

$n_t' = U - 2tVW$, $n_t'' = -2VW$, $v_t' = Z - 2tV^2$, and $v_t'' = -2V^2$. Cauchy–Schwarz and equation 56 give

$$\mathbb{E}[|V|\|W\|_2] \leq \frac{B_s\sqrt{v}}{K}, \qquad \mathbb{E}[\|U\|_2|Z|] \leq \frac{B_s R_{\max} v}{K},$$
$$\mathbb{E}[\|U\|_2 V^2] \leq \frac{2B_s R_{\max} v}{K}, \qquad \mathbb{E}[|V|\|W\|_2|Z|] \leq \frac{B_s R_{\max} v}{K},$$
$$\mathbb{E}[|V|^3\|W\|_2] \leq \frac{2B_s R_{\max} v}{K}, \qquad \mathbb{E}[(v_t')^2] \leq \frac{18R_{\max}^2 v}{K}. \tag{62}$$

For example, the third bound uses $\mathbb{E}\|U\|_2 V^2 \leq (\mathbb{E}\|U\|_2^2\,\mathbb{E}V^4)^{1/2}$ and $K^{-3/2} \leq K^{-1}$; the fifth is analogous, using $\|W\|_2 \leq B_s$. The last bound follows from $(Z - 2tV^2)^2 \leq 2Z^2 + 8V^4$ and $K \geq 2$. Consequently, Popoviciu's inequality $2\sqrt{v} \leq R_{\max}$ implies

$$\mathbb{E}\|n_t''\|_2 \leq \frac{B_s R_{\max}}{K}, \qquad \mathbb{E}[\|n_t'\|_2|v_t'|] \leq \frac{15B_s R_{\max} v}{K}. \tag{63}$$

On $\mathcal{E}$, $v_t \geq v/2$, while equation 58 gives $\|n_t\|_2 \leq 2B_s v_t/\Delta$. Applying these bounds term by term in equation 61, and using $v \leq R_{\max}^2/4$ and $a(a+1) \leq 2a$, yields

$$\mathbb{E}\left[\left\|\int_0^1 (1-t)\Gamma''(t)\,dt\right\|_2 \mathbf{1}_{\mathcal{E}}\right] \leq \frac{2^a B_s R_{\max}}{K v^a}\left(1 + 60a + \frac{145aR_{\max}}{\Delta}\right). \tag{64}$$

Equation 40 gives

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{8R_{\max}^2}{Kv}. \tag{65}$$

Furthermore,

$$\mathbb{E}\|\Gamma'(0)\|_2^2 \leq \frac{2B_s^2 v^{1-\alpha}}{K} + \frac{2\alpha^2 B_s^2 R_{\max}^2 v^{1-\alpha}}{K\Delta^2}. \tag{66}$$

Cauchy–Schwarz therefore bounds the truncated linear term by

$$\|\mathbb{E}[\Gamma'(0)\mathbf{1}_{\mathcal{E}^c}]\|_2 \leq \frac{4B_s R_{\max}}{Kv^a}\sqrt{1 + \frac{\alpha^2 R_{\max}^2}{\Delta^2}}. \tag{67}$$

On $\mathcal{E}^c$, the population and empirical self-bounds give

$$\mathbb{E}[\|\Gamma(1) - \Gamma(0)\|_2 \mathbf{1}_{\mathcal{E}^c}] \leq \frac{4B_s v^{1-a}}{\Delta} \cdot \frac{8R_{\max}^2}{Kv}$$
$$= \frac{32B_s R_{\max}^2}{K\Delta v^a}. \tag{68}$$

Writing $r := R_{\max}/\Delta$, the inequalities $\sqrt{1 + 4a^2r^2} \leq 1 + 2ar$ and $2^{-a} \leq 1$ show that the sum of equation 64, equation 67 and equation 68 is at most

$$\frac{2^a B_s R_{\max}}{Kv^a}\left(5 + 60a + \frac{32R_{\max}}{\Delta} + \frac{153aR_{\max}}{\Delta}\right) \tag{69}$$

for every $\alpha \in [0,2]$. Since

$$5 + 60a \leq 64(1+2a), \qquad 32 + 153a \leq 64(1+2a), \tag{70}$$

this is at most $2^{a+6}(1+2a)B_s R_{\max}(1 + R_{\max}/\Delta)/(Kv^a)$. Using $v^a = \sigma^\alpha$ proves the first branch. Taking the smaller of this bound and equation 53 proves the result. □

B.4 Proof of Theorem 3

Theorem 7 (Binary collinearity and relative bias (Restatement of Theorem 3))**.** *Let* $R \sim \mathrm{Bernoulli}(p)$*,* $q = 1 - p$*, and* $N = \sum_{j=1}^{K} R_j$*, where* $\widehat{p} = N/K$ *and* $\sigma^2 = pq$*. For every* $K \geq 2$*,* $p \in (0,1)$*, and* $\alpha \in [0,2]$*,*

$$\mathbb{E}\widehat{g}_{\alpha,K} = \phi_{\alpha,K}(p) g_\alpha, \qquad \phi_{\alpha,K}(p) = \frac{\mathbb{E}\Big[((N/K)(1-N/K))^{1-\alpha/2}\Big]}{\sigma^{2-\alpha}}, \qquad 0 \leq \phi_{\alpha,K}(p) \leq 1. \tag{71}$$

Moreover,

$$1 - \phi_{\alpha,K}(p) = \frac{c_\alpha(p)}{K} + \rho_{\alpha,K}(p), \qquad c_\alpha(p) = 1 - \frac{\alpha}{2} + \frac{\alpha(2-\alpha)(1-2p)^2}{8\sigma^2}, \tag{72}$$

and

$$|\rho_{\alpha,K}(p)| \leq \frac{6}{(K\sigma^2)^2} + \frac{8000}{(K\sigma^2)^3}. \tag{73}$$

Proof. Define

$$m_1 := \mathbb{E}[s_\theta(Y \mid x) \mid R = 1], \qquad m_0 := \mathbb{E}[s_\theta(Y \mid x) \mid R = 0]. \tag{74}$$

The score identity gives $pm_1 + qm_0 = 0$. Condition on $N = k$. Exchangeability then implies

$$\mathbb{E}[\widehat{g}_K \mid N = k] = \frac{k}{K}\left(1 - \frac{k}{K}\right) m_1 - \left(1 - \frac{k}{K}\right)\frac{k}{K} m_0 \tag{75}$$

$$= \frac{k}{K}\left(1 - \frac{k}{K}\right)(m_1 - m_0). \tag{76}$$

Since $\widehat{\sigma}_K = \sqrt{(k/K)(1-k/K)}$,

$$\mathbb{E}[\widehat{g}_{\alpha,K} \mid N = k] = \left[\frac{k}{K}\left(1 - \frac{k}{K}\right)\right]^{1-\alpha/2}(m_1 - m_0). \tag{77}$$

On the other hand, $pm_1 + qm_0 = 0$ gives $g_\alpha = \sigma^{2-\alpha}(m_1 - m_0)$. Averaging over the binomial distribution of N proves the collinearity formula. Under the convention $0^0 = 0$, the function $u \mapsto [u(1-u)]^{1-\alpha/2}$ is concave on $[0,1]$ for every $\alpha \in [0,2]$. Jensen's inequality therefore gives $0 \leq \phi_{\alpha,K}(p) \leq 1$.

We next make the leading term explicit. Let $D = \widehat{p} - p$ and $\mathcal{A} = \{|D| \leq \min\{p,q\}/2\}$. The required binomial moments are

$$\mathbb{E}D = 0, \qquad \mathbb{E}D^2 = \frac{\sigma^2}{K}, \qquad \mathbb{E}D^3 = \frac{\sigma^2(q-p)}{K^2}, \qquad \mathbb{E}D^4 = \frac{3\sigma^4}{K^2} + \frac{\sigma^2(1-6\sigma^2)}{K^3}. \tag{78}$$

Writing $f(u) = [u(1-u)]^{1-\alpha/2}$ on $(0,1)$, direct differentiation at p gives

$$\frac{f''(p)}{f(p)} = -\frac{2-\alpha}{\sigma^2} - \frac{\alpha(2-\alpha)(1-2p)^2}{4\sigma^4}. \tag{79}$$

When $\alpha = 2$, f is understood as the constant function one on the interior interval containing p and $p + D$; the endpoint values are handled by $\mathcal{A}^c$. Let

$$P_3(D) := 1 + \frac{f'(p)}{f(p)} D + \frac{f''(p)}{2f(p)} D^2 + \frac{f'''(p)}{6f(p)} D^3 \tag{80}$$

be the third-order Taylor polynomial of $f(p+D)/f(p)$. Combining equation 79 with $\mathbb{E}D = 0$ and $\mathbb{E}D^2 = \sigma^2/K$ gives

$$\mathbb{E}P_3(D) = 1 - \frac{c_\alpha(p)}{K} + \frac{f'''(p)}{6f(p)}\mathbb{E}D^3. \tag{81}$$

Since $\phi_{\alpha,K}(p) = \mathbb{E}[f(p+D)/f(p)]$, it follows that

$$\rho_{\alpha,K}(p) = \mathbb{E}\left[P_3(D) - \frac{f(p+D)}{f(p)}\right] - \frac{f'''(p)}{6f(p)}\mathbb{E}D^3. \tag{82}$$

Put $b := 1 - \alpha/2 \in [0, 1]$, $h(u) := u(1-u)$, and $z(u) := 1 - 2u$. Direct differentiation gives

$$\frac{f'''(u)}{f(u)} = \frac{b(b-1)(b-2)z(u)^3}{h(u)^3} - \frac{6b(b-1)z(u)}{h(u)^2}, \tag{83}$$

$$\frac{f^{(4)}(u)}{f(u)} = \frac{b(b-1)(b-2)(b-3)z(u)^4}{h(u)^4} - \frac{12b(b-1)(b-2)z(u)^2}{h(u)^3} + \frac{12b(b-1)}{h(u)^2}. \tag{84}$$

Since $b(1-b)(2-b) \le 1/2$, $b(1-b) \le 1/4$, and $\sigma^2 \le 1/4$, the first identity implies

$$\frac{|f'''(p)|}{f(p)} \le \frac{1}{\sigma^6}. \tag{85}$$

On $\mathcal{A}$, $h(u) \ge \sigma^2/2$. Multiplying the three terms in the second identity by $\sigma^8 f(u)/f(p) = \sigma^8[h(u)/\sigma^2]^b$, and using $b(1-b) \le 1/4$, $2 - b \le 2$, and $3 - b \le 3$, bounds them by 24, 12, and $3/4$, respectively. Hence, uniformly over $\alpha \in [0, 2]$,

$$\sup_{u:\,|u-p|\le \min\{p,q\}/2} \frac{|f^{(4)}(u)|}{f(p)} \le \frac{40}{\sigma^8}. \tag{86}$$

Taylor's theorem and equation 78 therefore give

$$\left|\frac{f'''(p)}{6f(p)}\mathbb{E}D^3\right| + \mathbb{E}\left[\left|P_3(D) - \frac{f(p+D)}{f(p)}\right| \mathbf{1}_{\mathcal{A}}\right] \le \frac{6}{K^2\sigma^4} + \frac{2}{K^3\sigma^6}. \tag{87}$$

It remains to control the contribution from $\mathcal{A}^c$. Set $x = K\sigma^2$. By symmetry, suppose that $p \le q$, and define the relative deviation

$$U := \frac{|D|}{p}. \tag{88}$$

Then $\mathcal{A}^c$ implies $U > 1/2$. Checking the cases $D \ge 0$ and $D < 0$ separately gives

$$\frac{(p+D)(q-D)}{pq} \le 1 + U. \tag{89}$$

Because $z^{1-\alpha/2} \le \max\{1, z\}$ for every $z \ge 0$, this implies

$$\frac{f(p+D)}{f(p)} \le 1 + U. \tag{90}$$

Moreover, the derivative formulas above give

$$\left|\frac{f'(p)}{f(p)}D\right| \le \frac{|q-p|}{q}U \le U, \tag{91}$$

$$\left|\frac{f''(p)}{2f(p)}D^2\right| \le \left(\frac{(2-\alpha)p}{2q} + \frac{\alpha(2-\alpha)(q-p)^2}{8q^2}\right)U^2 \le \frac{9}{8}U^2, \tag{92}$$

$$\left|\frac{f'''(p)}{6f(p)}D^3\right| \le \frac{1}{6q^3}U^3 \le \frac{4}{3}U^3, \tag{93}$$

where the last inequalities use $p \le q$, $q \ge 1/2$, and $\alpha(2-\alpha) \le 1$. Thus

$$|P_3(D)| \le 1 + U + \frac{9}{8}U^2 + \frac{4}{3}U^3. \tag{94}$$

On $\mathcal{A}^c$, the inequalities $1 \le 8U^3$, $U \le 4U^3$, and $U^2 \le 2U^3$ consequently give

$$\left|P_3(D) - \frac{f(p+D)}{f(p)}\right| \le 28U^3. \tag{95}$$

We next bound the third moment on this event directly. If $Kp \le 2$, the binomial factorial moments imply

$$\mathbb{E}N^2 \le (Kp)^2 + Kp, \qquad \mathbb{E}N^3 \le (Kp)^3 + 3(Kp)^2 + Kp. \tag{96}$$

Hence

$$\mathbb{E}|KD|^3 \le \mathbb{E}(N + Kp)^3 \le Kp + 6(Kp)^2 + 8(Kp)^3 \le 90. \tag{97}$$

If $Kp \ge 2$, expand the sixth moment of the independent centered variables $R_j - p$. Using $|R_j - p| \le 1$ shows that

$$\mathbb{E}(KD)^6 \le Kpq + 25(Kpq)^2 + 15(Kpq)^3 \le Kp + 25(Kp)^2 + 15(Kp)^3. \tag{98}$$

The coefficients correspond respectively to the partitions 6, $(4, 2)$ together with $(3, 3)$, and $(2, 2, 2)$ of the sixth moment. On $\{U \ge 1/2\}$, one has $|KD| \ge Kp/2$, and therefore

$$\mathbb{E}\big[|KD|^3 \mathbf{1}\{U \ge 1/2\}\big] \le \frac{8}{(Kp)^3}\mathbb{E}(KD)^6 \tag{99}$$

$$\le 8\left(\frac{1}{(Kp)^2} + \frac{25}{Kp} + 15\right) \le 222. \tag{100}$$

Together with the $Kp \le 2$ case, this proves

$$\mathbb{E}\big[U^3 \mathbf{1}\{U \ge 1/2\}\big] \le \frac{222}{(Kp)^3} \le \frac{222}{x^3}, \tag{101}$$

where the last inequality uses $x = Kpq \le Kp$. Equations 95 and 101 bound the bad-event contribution by $6216/x^3$. Combining this with equation 87 gives

$$|\rho_{\alpha,K}(p)| \le \frac{6}{x^2} + \frac{6218}{x^3} \le \frac{6}{x^2} + \frac{8000}{x^3}, \tag{102}$$

as claimed. Finally,

$$8\sigma^2 c_\alpha(p) = 8\sigma^2\left(1 - \frac{\alpha}{2}\right) + \alpha(2 - \alpha)(1 - 4\sigma^2) \le 2 + 1 = 3, \tag{103}$$

which proves the simplified leading-term bound stated after Theorem 3. □

B.5 Proof of Proposition 3

Let $\mathcal{K}$ be the deterministic set of all group sizes that can be examined by Algorithm 1 and let $M := |\mathcal{K}|$. The doubling schedule gives $M \le 1 + \lceil \log_2(K_{\max}/K_0) \rceil$.

Proposition 7 (Population resolution at adaptive stopping (Restatement of Proposition 3))**.** *Suppose that $2 \le K_0 \le K_{\max}$ and that the rewards are i.i.d. and satisfy Assumption 2. Define $\bar{c} := \max\{c, (K_0 - 1)\Delta^2/K_0\}$. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$, any returned group size $K \in \mathcal{K}$ satisfying $K\widehat{\sigma}_K^2 \ge c$ also satisfies*

$$K\sigma^2 \ge \frac{\bar{c}}{e}\left(\frac{\delta}{M}\right)^{R_{\max}^2/\bar{c}}. \tag{104}$$

Proof. Write $\mu = \mathbb{E}R$, $\sigma^2 = \mathrm{Var}(R)$, and $X_j = R_j - \mu$. First, the pairwise representation of the empirical variance gives

$$K\widehat{\sigma}_K^2 = \frac{1}{K}\sum_{1 \le i < j \le K}(R_i - R_j)^2. \tag{105}$$

If the empirical variance is nonzero, at least $K - 1$ pairs in equation 105 contain distinct rewards, and each such pair differs by at least Δ. Therefore, because every examined group size satisfies $K \ge K_0$,

$$K\widehat{\sigma}_K^2 \ge \frac{K - 1}{K}\Delta^2 \ge \frac{K_0 - 1}{K_0}\Delta^2. \tag{106}$$

Consequently, the stopping condition $K\widehat{\sigma}_K^2 \ge c > 0$ implies $K\widehat{\sigma}_K^2 \ge \bar{c}$.

For a fixed $K \in \mathcal{K}$, set $x = K\sigma^2$. Centering at the population mean yields

$$K\widehat{\sigma}_K^2 = \sum_{j=1}^K X_j^2 - K\left(\frac{1}{K}\sum_{j=1}^K X_j\right)^2 \le \sum_{j=1}^K X_j^2. \tag{107}$$

The variables $X_j^2/R_{\max}^2$ belong to $[0,1]$, and the sum of their expectations is $x/R_{\max}^2$. Hence the multiplicative Chernoff bound and equation 107 imply, whenever $x < \bar{c}$,

$$\mathbb{P}\big(K\widehat{\sigma}_K^2 \geq \bar{c}\big) \leq \Big(\frac{ex}{\bar{c}}\Big)^{\bar{c}/R_{\max}^2}. \tag{108}$$

Indeed, this is the usual multiplicative Chernoff inequality for independent random variables in $[0,1]$, applied at threshold $\bar{c}/R_{\max}^2$.

Define

$$L := \frac{\bar{c}}{e}\left(\frac{\delta}{M}\right)^{R_{\max}^2/\bar{c}}. \tag{109}$$

If $x < L$, then equation 108 is strictly smaller than δ/M. A union bound over all $K \in \mathcal{K}$ therefore shows that, with probability at least $1-\delta$, no candidate group size with $K\sigma^2 < L$ can satisfy the stopping condition. This simultaneous event covers the data-dependent group size returned by the algorithm and does not require independence among the statistics at different group sizes. Thus every returned K that meets the empirical threshold satisfies $K\sigma^2 \geq L > 0$, proving equation 12. Finally, the doubling schedule contains at most $1 + \lceil \log_2(K_{\max}/K_0) \rceil$ distinct group sizes. □

C Proof of Proposition 4

Proposition 8 (Finite-group stationarity (Restatement of Proposition 4))**.** *Suppose rewards are binary, Assumption 1 holds, and $\|\nabla_\theta s_\theta(Y \mid x)\|_{\mathrm{op}} \leq B_h$ almost surely. For $\alpha \in \{0,1\}$, if $0 < \eta \leq 1/(B_h + 3B_s^2)$, then the full-batch iterates in equation 16 satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,K}(p_{\theta_t}^i)g_{\alpha,\theta_t}^i\right\|_2^2 \leq \frac{2\pi}{\eta T} + (B_h + 3B_s^2)\eta\frac{A_1 B_s^2}{NK}. \tag{110}$$

Proof. Fix one prompt, suppress its index, and write $p = p(\theta)$, $v = p(1-p)$, $s = s_\theta(Y \mid x)$, and $H_s = \nabla_\theta s_\theta(Y \mid x)$. The score identities are

$$\mathbb{E}s = 0, \qquad \mathbb{E}(H_s + ss^\top) = 0. \tag{111}$$

Since $R \sim \mathrm{Bernoulli}(p)$ and $\mathbb{E}|R-p| = 2v$, differentiation under the expectation gives

$$\begin{aligned} \nabla p &= \mathbb{E}[(R-p)s], & \|\nabla p\|_2 &\leq 2B_s v, \\ \nabla^2 p &= \mathbb{E}[(R-p)(H_s + ss^\top)], & \|\nabla^2 p\|_{\mathrm{op}} &\leq 2v(B_s^2 + B_h). \end{aligned} \tag{112}$$

For $\alpha = 0$, Theorem 3 gives

$$\overline{g}_{0,K}(\theta) = \left(1 - \frac{1}{K}\right)\nabla p(\theta). \tag{113}$$

For $\alpha = 1$, let

$$F_K(p) := \mathbb{E}_{M\sim\mathrm{Binomial}(K,p)}\sqrt{\frac{M}{K}\left(1-\frac{M}{K}\right)}, \qquad w_K(p) := \frac{F_K(p)}{p(1-p)}. \tag{114}$$

The quotient w_K extends continuously to $[0,1]$: after factoring $p(1-p)$ from the binomial sum it is a polynomial of degree $K-2$. Theorem 3 and the score-function identity yield

$$\overline{g}_{1,K}(\theta) = w_K(p(\theta))\nabla p(\theta). \tag{115}$$

The scalar objectives underlying the two finite-group mean directions are therefore

$$\begin{aligned} \mathcal{J}_{0,K}^x(\theta) &:= \left(1 - \frac{1}{K}\right)p(\theta), & \nabla\mathcal{J}_{0,K}^x(\theta) &= \overline{g}_{0,K}(\theta), \\ \mathcal{J}_{1,K}^x(\theta) &:= \int_0^{p(\theta)} w_K(u)\,\mathrm{d}u, & \nabla\mathcal{J}_{1,K}^x(\theta) &= \overline{g}_{1,K}(\theta). \end{aligned} \tag{116}$$

Here the superscript x emphasizes that these are the objectives for the fixed prompt x.

We next derive their smoothness constants. For $\alpha = 0$, equation 112 and $v \le 1/4$ give

$$\|\nabla^2 \mathcal{J}^x_{0,K}(\theta)\|_{\mathrm{op}} \le \frac{1}{2}\left(1 - \frac{1}{K}\right)(B_s^2 + B_h). \tag{117}$$

The function F_K is the Bernstein polynomial of the concave function $u \mapsto \sqrt{u(1-u)}$; hence it is concave. It is also symmetric about $1/2$, vanishes at the endpoints, and satisfies $0 \le F_K \le 1/2$. Concavity gives $pF'_K(p) \le F_K(p)$ for $p \le 1/2$; symmetry gives the corresponding inequality on the other half of the interval. Moreover, $F'_K(p)$ and $1 - 2p$ have the same sign. Therefore

$$\begin{aligned} v|F'_K(p)| &\le F_K(p), \\ v^2|w'_K(p)| &= |vF'_K(p) - F_K(p)(1-2p)| \\ &\le \max\{v|F'_K(p)|, F_K(p)|1-2p|\} \le F_K(p) \le \frac{1}{2}. \end{aligned} \tag{118}$$

Combining equation 112 and equation 118,

$$\begin{aligned} \left\|\nabla_\theta \overline{g}_{1,K}(\theta)\right\|_{\mathrm{op}} &\le w_K(p)\|\nabla^2 p\|_{\mathrm{op}} + |w'_K(p)|\|\nabla p\|_2^2 \\ &\le 2F_K(p)(B_s^2 + B_h) + 2B_s^2 \le B_h + 3B_s^2. \end{aligned} \tag{119}$$

This bound is uniform in both p and K.

We have thus derived the method-specific smoothness constants

$$L_{0,K} := \frac{1}{2}\left(1 - \frac{1}{K}\right)(B_s^2 + B_h), \qquad L_{1,K} := B_h + 3B_s^2. \tag{120}$$

The scalar objectives also have uniformly bounded ranges. The range of $\mathcal{J}^x_{0,K}$ is at most one. For $\alpha = 1$, Jensen's inequality and $\mathbb{E}[(M/K)(1 - M/K)] = (1 - 1/K)v$ give

$$F_K(p) \le \sqrt{\left(1 - \frac{1}{K}\right)v}, \qquad \int_0^1 w_K(p)\,\mathrm{d}p \le \sqrt{1 - \frac{1}{K}}\int_0^1 \frac{\mathrm{d}p}{\sqrt{p(1-p)}} \le \pi. \tag{121}$$

After averaging over prompts, the finite-group objectives retain these range and smoothness bounds.

It remains to bound rollout noise. Independence across prompts and $\mathrm{Var}(Z) \le \mathbb{E}\|Z - z\|_2^2$, followed by Theorem 1, imply

$$\begin{aligned} V_{0,K}(\theta) &\le \frac{A_0 B_s^2}{N^2 K}\sum_{i=1}^N (\sigma^i_\theta)^2 \le \frac{A_0 B_s^2}{4NK}, \\ V_{1,K}(\theta) &\le \frac{A_1 B_s^2}{NK}. \end{aligned} \tag{122}$$

Finally, let $\mathcal{J}_{\alpha,K}$ denote the averaged scalar objective constructed above. Condition on θ_t and abbreviate $\overline{G}_t := N^{-1}\sum_{i=1}^N \phi_{\alpha,K}(p^i_{\theta_t}) g^i_{\alpha,\theta_t}$, $\widehat{G}_t := \widehat{G}_{\alpha,K}(\theta_t)$, and $L := L_{\alpha,K}$. Since $\mathbb{E}_t \widehat{G}_t = \overline{G}_t = \nabla \mathcal{J}_{\alpha,K}(\theta_t)$, smoothness gives

$$\mathbb{E}_t \mathcal{J}_{\alpha,K}(\theta_{t+1}) \ge \mathcal{J}_{\alpha,K}(\theta_t) + \eta\|\overline{G}_t\|_2^2 - \frac{L\eta^2}{2}\mathbb{E}_t\|\widehat{G}_t\|_2^2 \tag{123}$$

$$= \mathcal{J}_{\alpha,K}(\theta_t) + \left(\eta - \frac{L\eta^2}{2}\right)\|\overline{G}_t\|_2^2 - \frac{L\eta^2}{2}V_{\alpha,K}(\theta_t) \tag{124}$$

$$\ge \mathcal{J}_{\alpha,K}(\theta_t) + \frac{\eta}{2}\|\overline{G}_t\|_2^2 - \frac{L\eta^2}{2}V_{\alpha,K}(\theta_t), \tag{125}$$

where the last line uses $\eta \le 1/L$. Take total expectations, sum over t, and substitute the objective-range and noise bounds above. For each $\alpha \in \{0, 1\}$, this first gives the sharper method-specific bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\overline{G}_t\|_2^2 \le \frac{2D_\alpha}{\eta T} + L_{\alpha,K}\eta\nu_{\alpha,K}, \tag{126}$$

where

$$(D_0, \nu_{0,K}) = \left(1, \frac{A_0 B_s^2}{4NK}\right), \qquad (D_1, \nu_{1,K}) = \left(\pi, \frac{A_1 B_s^2}{NK}\right). \tag{127}$$

Finally, $L_{0,K} \le L_{1,K} = B_h + 3B_s^2$, $D_0 \le D_1 = \pi$, and $A_0/4 \le A_1$. Therefore the common step-size condition in the proposition implies $\eta \le 1/L_{\alpha,K}$, and equation 126 yields the stated common bound. □

D Proof of Theorem 4

Lemma 1 (Binomial class-count bounds)**.** *Let $N \sim \mathrm{Binomial}(K, p)$, $q = 1 - p$, and $v = pq > 0$. For $\beta \in [0, 1]$, define*

$$\mathcal{D} = \{1 \le N \le K - 1\}, \qquad \widehat{v} = \frac{N(K-N)}{K^2},$$
$$W = \widehat{v}^{\beta/2} \mathbf{1}_{\mathcal{D}}, \qquad d_N = \left(\frac{1}{N} + \frac{1}{K-N}\right) \mathbf{1}_{\mathcal{D}}.$$

where $W = \mathbf{1}_{\mathcal{D}}$ when $\beta = 0$. There is a constant $C_\beta < \infty$, depending only on β, such that

$$\mathbb{E}|W - v^{\beta/2}|^2 + \mathbb{E}[W^2 d_N] \le C_\beta \min\left\{\frac{1}{Kv^{1-\beta}}, v^\beta\right\}, \tag{128}$$

$$|\mathbb{E}W - v^{\beta/2}| + \mathbb{E}[W d_N] \le C_\beta \min\left\{\frac{1}{Kv^{1-\beta/2}}, v^{\beta/2}\right\}. \tag{129}$$

Proof. We first record an inverse-moment estimate. If $Z \sim \mathrm{Binomial}(K, r)$, $\mu = Kr$, and $\gamma \in [0, 1]$, then

$$\mathbb{E}\big[Z^{-\gamma} \mathbf{1}\{Z \ge 1\}\big] \le C_\gamma \mu^{-\gamma}. \tag{130}$$

For $\mu < 2$, this follows from the trivial bound by one. For $\mu \ge 2$, splitting at $\mu/2$ and using the multiplicative Chernoff bound gives

$$\mathbb{E}[Z^{-\gamma} \mathbf{1}\{Z \ge 1\}] \le (\mu/2)^{-\gamma} + \Pr(Z < \mu/2) \le 2^\gamma \mu^{-\gamma} + e^{-\mu/8},$$

which proves equation 130.

For every $\rho \in [0, 1]$, concavity and $\mathbb{E}\widehat{v} = (K-1)v/K$ imply $\mathbb{E}\widehat{v}^\rho \le v^\rho$, with the endpoint $\rho = 0$ interpreted on $\mathcal{D}$. Since $d_N \le 2$, it follows that

$$\mathbb{E}|W - v^{\beta/2}|^2 + \mathbb{E}[W^2 d_N] \le 6v^\beta,$$
$$|\mathbb{E}W - v^{\beta/2}| + \mathbb{E}[W d_N] \le 4v^{\beta/2}.$$

These are the near-degenerate branches.

We next prove the resolved branches. Suppose first that $\beta > 0$, and write $\rho = \beta/2$. On $\mathcal{G} = \{\widehat{v} \ge v/2\}$, the mean-value theorem gives

$$|\widehat{v}^\rho - v^\rho|^2 \le C_\beta v^{\beta-2} |\widehat{v} - v|^2.$$

By equation 37 and Chebyshev's inequality, we have $\mathbb{E}|\widehat{v} - v|^2 \le 2v/K$ and $\Pr(\mathcal{G}^c) \le 8/(Kv)$. On $\mathcal{G}^c$, both $\widehat{v}^{\beta/2}$ and $v^{\beta/2}$ are at most $v^{\beta/2}$. Consequently,

$$\mathbb{E}|W - v^{\beta/2}|^2 \le \frac{C_\beta}{Kv^{1-\beta}}. \tag{131}$$

For $\beta = 0$, the left side equals $p^K + q^K \le (Kp)^{-1} + (Kq)^{-1} = 1/(Kv)$, so the same conclusion holds.

For the first-moment power error, Theorem 3 with $\alpha = 2 - \beta$ gives $\mathbb{E}W/v^{\beta/2} = \phi_{2-\beta,K}(p)$. Its relative-bias bound therefore yields, when $Kv \ge 1$,

$$|\mathbb{E}W - v^{\beta/2}| \le \frac{C_\beta}{Kv^{1-\beta/2}}. \tag{132}$$

When $Kv < 1$, the near-degenerate bound is no larger than the right-hand side of equation 132; hence that inequality holds for all $Kv > 0$.

It remains to control the factors d_N. Since $\widehat{v} \le N/K$ and $\widehat{v} \le (K-N)/K$, applying equation 130 to N and $K-N$ gives

$$\mathbb{E}[W^2 d_N] \le \frac{C_\beta}{K}\left(p^{-(1-\beta)} + q^{-(1-\beta)}\right) \le \frac{2C_\beta}{Kv^{1-\beta}},$$
$$\mathbb{E}[W d_N] \le \frac{C_\beta}{K}\left(p^{-(1-\beta/2)} + q^{-(1-\beta/2)}\right) \le \frac{2C_\beta}{Kv^{1-\beta/2}}.$$

Combining these estimates with the near-degenerate branches, equation 131, and equation 132 proves both claims. □

Theorem 8 (Accuracy of Balanced Aggregation (Restatement of Theorem 4))**.** *Suppose rewards are binary,* $\mathbb{E}[L(Y)^2 \mid R = a] < \infty$*, and* $\|s_\theta(Y \mid x)\|_2 \le B_\ell L(Y)$*. For every* $K \ge 2$*,* $\alpha \in [1,2]$*, and* $\sigma > 0$*, there exist constants depending only on* α *for which equation 20 and equation 21 hold.*

Proof. Write $p = \Pr(R=1)$, $q = 1-p$, $v = pq = \sigma^2$, and $\beta = 2-\alpha \in [0,1]$. For $a \in \{0,1\}$, let $m_a = \mathbb{E}[s_\theta(Y \mid x) \mid R = a]$, $\ell_a = \mathbb{E}[L(Y) \mid R = a]$, and $r_a = m_a/\ell_a$. The score identity $pm_1 + qm_0 = 0$, together with $g_\theta(x) = pm_1 = -qm_0$, gives

$$v^{\beta/2}(r_1 - r_0) = \left(\frac{q}{\ell_1} + \frac{p}{\ell_0}\right)\frac{g_\theta(x)}{\sigma^\alpha}.$$

Moreover, $\|r_a\|_2 \le B_\ell$, so for $u := r_1 - r_0$ we have $\|u\|_2 \le 2B_\ell$. This proves equation 19 and its stated norm bound.

We first control the two empirical ratios. Under the conditional law given $R = a$, define

$$\zeta_a := s_\theta(Y \mid x) - r_a L(Y), \qquad V_a := \mathbb{E}\|\zeta_a\|_2^2, \qquad \tau_a^2 := \mathrm{Var}(L(Y)).$$

These quantities are finite, $\mathbb{E}\zeta_a = 0$, and $\|r_a\|_2 \le B_\ell$. Set

$$\Lambda_{\mathrm{BA}}^{\mathrm{mse}} := \max_{a\in\{0,1\}}\left\{\frac{4V_a}{B_\ell^2\ell_a^2} + \frac{16\tau_a^2}{\ell_a^2}\right\},$$
$$\Lambda_{\mathrm{BA}}^{\mathrm{bias}} := \max_{a\in\{0,1\}}\left\{\frac{4\sqrt{V_a\tau_a^2}}{B_\ell\ell_a^2} + \frac{8\tau_a^2}{\ell_a^2}\right\}. \tag{133}$$

Consider n independent responses from class a, and write $\widehat{r}_a = \overline{s}_a/\overline{L}_a$. Since $\|\widehat{r}_a\|_2 \le B_\ell$, on $\mathcal{G}_a := \{\overline{L}_a \ge \ell_a/2\}$ we have $\|\widehat{r}_a - r_a\|_2^2 \le 4\|\overline{\zeta}_a\|_2^2/\ell_a^2$, while on its complement the squared error is at most $4B_\ell^2$. Chebyshev's inequality therefore yields

$$\mathbb{E}\|\widehat{r}_a - r_a\|_2^2 \le \frac{4V_a}{n\ell_a^2} + \frac{16B_\ell^2\tau_a^2}{n\ell_a^2}. \tag{134}$$

For the bias, using $\mathbb{E}\overline{\zeta}_a = 0$ gives the exact identity

$$\mathbb{E}(\widehat{r}_a - r_a) = -\mathbb{E}\frac{\overline{\zeta}_a(\overline{L}_a - \ell_a)}{\ell_a\overline{L}_a}.$$

On $\mathcal{G}_a$, Cauchy–Schwarz and $\mathbb{E}\|\overline{\zeta}_a\|_2^2 = V_a/n$, $\mathbb{E}(\overline{L}_a - \ell_a)^2 = \tau_a^2/n$ bound the norm by $2\sqrt{V_a\tau_a^2}/(n\ell_a^2)$. On $\mathcal{G}_a^c$, the pointwise identity

$$-\frac{\overline{\zeta}_a(\overline{L}_a - \ell_a)}{\ell_a\overline{L}_a} = (\widehat{r}_a - r_a) - \frac{\overline{\zeta}_a}{\ell_a}$$

together with $\|\widehat{r}_a - r_a\|_2 \le 2B_\ell$ and $\Pr(\mathcal{G}_a^c) \le 4\tau_a^2/(n\ell_a^2)$ gives

$$\left\|\mathbb{E}\left[(\widehat{r}_a - r_a)\mathbf{1}_{\mathcal{G}_a^c}\right]\right\|_2 \le \frac{8B_\ell\tau_a^2}{n\ell_a^2},$$

while Cauchy–Schwarz gives

$$\frac{1}{\ell_a}\mathbb{E}\big[\|\overline{\zeta}_a\|_2 \mathbf{1}_{\mathcal{G}_a^c}\big] \le \frac{2\sqrt{V_a\tau_a^2}}{n\ell_a^2}.$$

Adding the two events therefore yields

$$\|\mathbb{E}(\widehat{r}_a - r_a)\|_2 \le \frac{4\sqrt{V_a\tau_a^2}}{n\ell_a^2} + \frac{8B_\ell\tau_a^2}{n\ell_a^2}. \tag{135}$$

Let $N = N_1 \sim \mathrm{Binomial}(K, p)$, $\mathcal{D} = \{1 \le N \le K-1\}$, and $d_N = (N^{-1} + (K-N)^{-1})\mathbf{1}_{\mathcal{D}}$. Conditional on N, the two classes are independent. With $\widehat{u} = \widehat{r}_1 - \widehat{r}_0$ on $\mathcal{D}$ and zero otherwise, the bounds in equation 134 and equation 135 imply

$$\mathbb{E}[\|\widehat{u} - u\|_2^2 \mid N] \le 2\Lambda_{\mathrm{BA}}^{\mathrm{mse}} B_\ell^2 d_N, \tag{136}$$

$$\|\mathbb{E}[\widehat{u} - u \mid N]\|_2 \le \Lambda_{\mathrm{BA}}^{\mathrm{bias}} B_\ell d_N \tag{137}$$

on $\mathcal{D}$; also $\|\widehat{u}\|_2 \le 2B_\ell$.

It remains to average over the random class counts. Put $\widehat{v} = N(K-N)/K^2$ and $W = \widehat{v}^{\beta/2}$, with the convention $W = 0$ on $\mathcal{D}^c$. Lemma 1 gives the required power and inverse-class-count estimates equation 128– equation 129.

The empirical Balanced Aggregation direction factorizes as $\widehat{g}_{\alpha,K}^{\mathrm{BA}} = W\widehat{u}$, while $g_{\alpha,\theta}^{\mathrm{BA}}(x) = v^{\beta/2}u$. Therefore

$$\widehat{g}_{\alpha,K}^{\mathrm{BA}} - g_{\alpha,\theta}^{\mathrm{BA}}(x) = (W - v^{\beta/2})u + W(\widehat{u} - u).$$

Using $\|u\|_2 \le 2B_\ell$, conditioning on N, and applying equation 136 and equation 128 yields equation 20 after substituting $\beta = 2 - \alpha$. Taking expectations in the same decomposition and using equation 137 and equation 129 yields equation 21. □