PROMPT-LEVEL ADVANTAGE SHAPING AND α-GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning methods such as PPO and its variants are widely used to improve pretrained LLMs. GRPO successfully removes the PPO critic network by replacing it with group-based estimates, but introduces several effects that remain insufficiently understood. The algorithm yields weak gradient signals for both simple and hard prompts and uses heuristic modifications such as prompt-level advantage shaping. Shaping gives biased gradient estimators that need not converge to stationary points of the mean-reward objective. This paper contributes to the theoretical understanding of GRPO. Based on recent Bernstein polynomial computations for the optimization bias, we analyze the expected gradient change caused by advantage shaping. Explicit formulas allow to understand how different prompts contribute to the expected gradient. We introduce α-GRPO, a family of GRPO variants with a possibly time-varying degree of prompt-level advantage shaping. By annealing this weighting over training, α-GRPO can retain the learning signal provided by prompt-level shaping early in training while recovering asymptotic stationarity guarantees for the original mean-reward objective. The mathematical feasibility of α-GRPO is shown using biased SGD arguments.

1 INTRODUCTION

Reinforcement learning from verifiable rewards (RLVR) has become a central post-training paradigm for improving the reasoning abilities of large language models. Group-based methods such as GRPO are especially attractive because they avoid a learned critic and construct relative advantages from multiple completions of the same prompt (Shao et al., 2024; Guo et al., 2025). GRPO and its variants involve several modifications to the basic policy-gradient update, including response-level length normalization, batch-level reweighting, prompt filtering, clipping, and prompt-level advantage shaping, e.g. Yu et al. (2025); Liu et al. (2025); Heckel et al. (2026); Dai et al. (2026). In this work, we focus on the latter: prompt-level advantage shaping changes the relative contribution of prompts and can strengthen otherwise weak learning signals, but also biases the expected gradient away from the original mean-reward objective. Recent work has begun developing mathematical theory for GRPO-style RLVR (Davis & Recht, 2025; Thrampoulidis et al., 2026; Suk & Duan, 2026). Building on this line of work, we characterize how prompt-level advantage shaping changes the underlying optimization objective and the contribution of prompts with different success probabilities. We then introduce α-GRPO, a family that controls the strength of this shaping and allows it to be annealed during training, retaining stronger learning signals early while recovering stationarity bounds for the mean-reward objective along randomly selected iterates.

1.1 SETUP AND REQUIRED BACKGROUND FOR PROMPT-LEVEL ADVANTAGE SHAPING

Setup. We study RLVR for autoregressive language models. Let $\mathcal{Q}$ be a finite prompt space, $\mathcal{V}$ a finite token vocabulary, and $\mathcal{O} = \bigcup_{T=1}^{T_{\max}} \mathcal{V}^T$ the response space. Given a prompt q and parameter $\theta \in \mathbb{R}^d$, a policy π_θ generates a response $o = (o_1, \ldots, o_{T(o)}) \in \mathcal{O}$ according to $\pi_\theta(o \,|\, q) = \prod_{t=1}^{T(o)} \pi_\theta(o_t \,|\, q, o_{<t})$, where $T(o)$ is the response length. Here, we use π_θ for both the induced response distribution and its autoregressive token conditionals. For all (q, o) we assume $\pi_\theta(o \,|\, q) > 0$ and that $\theta \mapsto \pi_\theta(o \,|\, q)$ is differentiable. A verifier assigns a binary reward $r(q, o) \in \{0, 1\}$. The true RLVR optimization objective is

$$J(\theta) := \mathbb{E}_{Q,O}[r(Q, O)] = \mathbb{E}_Q[\rho_Q(\theta)], \qquad \rho_q(\theta) := \mathbb{E}_O[r(q, O)]. \tag{1}$$

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.

The quantity $\rho_q(\theta)$ is called the pass probability, i.e. the probability of producing a correct response to prompt q. Maximizing J means maximizing the pass probability averaged over all prompts according to the prompt distribution. We sample prompts from the uniform law $\mathbb{P}_Q$; no prompt is prioritized. To optimize J a sample-based approximation of the gradient

$$\nabla_\theta J(\theta) = \mathbb{E}_Q[\nabla_\theta \rho_Q(\theta)] \tag{2}$$

is needed. Applying the policy gradient theorem with value-baseline to the gradient of $\rho_Q(\theta)$ yields

$$\nabla J(\theta) = \mathbb{E}_{Q,O}\left[A^{\pi_\theta}(Q,O)\nabla_\theta \log \pi_\theta(O \mid Q)\right], \qquad A^{\pi_\theta}(q,o) := r(q,o) - \rho_q(\theta), \tag{3}$$

where $\nabla_\theta \log \pi_\theta(o \mid q) = \sum_{t=1}^{T(o)} \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t})$. The baseline $\rho_q(\theta)$ may be replaced by any response-independent baseline. In this contextual-bandit formulation, $\rho_q(\theta)$ is the classical RL state-value function and A^{π_θ} the advantage. Actor-critic methods such as PPO (Schulman et al., 2017) estimate the value function with a learned critic. GRPO (Shao et al., 2024) instead estimates the baseline from a group of responses, avoiding a critic network. For a batch size B and group size $G \geq 2$, prompts $Q_1, \ldots, Q_B \sim \mathbb{P}_Q$ are sampled independently. For a prompt b, $O_{b,1}, \ldots, O_{b,G} \sim \pi_\theta(\cdot \mid Q_b)$ with rewards $R_{b,i} = r(Q_b, O_{b,i})$. The mean-centered advantage estimator is

$$\widehat{A}^{\text{group}}_{b,i} := R_{b,i} - \widehat{\rho}_b, \qquad \widehat{\rho}_b := \frac{1}{G}\sum_{j=1}^{G} R_{b,j},$$

Because $\widehat{\rho}_b$ contains $R_{b,i}$, using $\widehat{A}^{\text{group}}$ gives a biased gradient estimator. RLOO (Ahmadian et al., 2024; Kool et al., 2019) removes this bias: $\widehat{A}^{\text{RLOO}}_{b,i} = R_{b,i} - \frac{1}{G-1}\sum_{j\neq i} R_{b,j} = \frac{G}{G-1}\big(R_{b,i} - \widehat{\rho}_b\big)$, and the resulting constant factor $\frac{G}{G-1}$ can be absorbed into the learning rate.

A key limitation of group-based baselines is *gradient collapse*. If all responses in a group receive the same reward (hard and easy prompts), then $\widehat{A}^{\text{group}}_{b,i} = 0$, so the prompt contributes no gradient. Only mixed reward groups produce learning signals, underrepresenting both easy and hard prompts.

Proposition 1 (Mixed group probability)**.** *The probability of mixed responses to a prompt with a pass probability ρ is $p_{mix}(\rho) = 1 - \rho^G - (1-\rho)^G$.*

Proof. For a fixed prompt, the rewards are i.i.d. Bernoulli(ρ). A group is mixed unless all rewards equal 0 or all equal 1, which occur with probabilities $(1-\rho)^G$ and ρ^G, respectively. □

Several methods mitigate gradient collapse at the cost of introducing gradient bias. DAPO (Yu et al., 2025) uses dynamic sampling, while GRPO (Shao et al., 2024) applies prompt-level advantage shaping by increasing gradient magnitudes for hard and easy prompts. Its shaped advantage is

$$\widehat{A}^{\text{GRPO}}_{b,i} := \frac{R_{b,i} - \widehat{\rho}_b}{\widehat{\sigma}_b}, \qquad \widehat{\sigma}_b^2 := \frac{1}{G}\sum_{j=1}^{G}(R_{b,j} - \widehat{\rho}_b)^2 \overset{\text{binary rewards}}{=} \widehat{\rho}_b(1-\widehat{\rho}_b).$$

The advantage is defined as 0 whenever $\widehat{\sigma}_b = 0$; implementations typically stabilize the denominator with a small constant $\epsilon > 0$. Despite the empirical success of GRPO, the role of prompt-level shaping remains poorly understood. Dr. GRPO (Liu et al., 2025) removes standard-deviation scaling, while Heckel et al. (2026); Dai et al. (2026) propose shaping rules that emphasize harder prompts.

General advantage shaping. Motivated by variants of GRPO, we study general advantage shaping. Let $f : [0,1] \to \mathbb{R}_+$ be a shaping function, and define the shaped advantage estimator as

$$\widehat{A}^f_{b,i} := f(\widehat{\rho}_b)\big(R_{b,i} - \widehat{\rho}_b\big).$$

We study SGD $\theta_{n+1} = \theta_n + \eta_n\, \widehat{g}^f_{B,G}(\theta_n)$, with group-based batch gradient estimator

$$\widehat{g}^f_{B,G}(\theta) = \frac{1}{BG}\sum_{b=1}^{B}\sum_{i=1}^{G} \widehat{A}^f_{b,i}\, \nabla_\theta \log \pi_\theta(O_{b,i} \mid Q_b). \tag{4}$$

Its expectation $g^f_G(\theta) := \mathbb{E}_{Q,O}[\widehat{g}^f_{B,G}(\theta)]$ defines the expected gradient field induced by the finite-group estimator. It would be desirable to have $g^f_G(\theta) = \nabla_\theta J(\theta) = \mathbb{E}_Q[\nabla_\theta \rho_Q(\theta)]$; however, that is generally not the case. To separate advantage shaping from group estimation, we also consider the infinite-group gradient field obtained by replacing the empirical pass probability with the exact pass probability:

$$g^f(\theta) := \mathbb{E}_{Q,O}\left[f(\rho_Q(\theta))(r(Q,O) - \rho_Q(\theta))\nabla_\theta \log \pi_\theta(O \mid Q)\right]. \tag{5}$$

It holds that $g^f = \lim_{G\to\infty} g^f_G$. We call g^f the exact-critic (or infinite-group) gradient field.

The true GRPO optimization objective and related work (more in Appendix A). Several recent methods differ mainly in their advantage shaping; Table 1 summarizes the variants considered here. Up to the factor $\frac{G}{G-1}$, RLOO (Ahmadian et al., 2024; Kool et al., 2019) coincides with the unnormalized centering of Dr. GRPO (Liu et al., 2025), corresponding to $f \equiv 1$. For binary rewards, GRPO (Shao et al., 2024; Guo et al., 2025) uses $f(\rho) = (\rho(1-\rho))^{-1/2}$, while Uniform-R and Linear-R (Heckel et al., 2026) apply stronger (asymmetric) shaping. The present paper continues ideas of Davis & Recht (2025) and Thrampoulidis et al. (2026). By the policy gradient identity, the exact-critic expected gradient from Equation 5 can be written as

$$g^f(\theta) = \mathbb{E}_Q\left[f\big(\rho_Q(\theta)\big)\nabla_\theta \rho_Q(\theta)\right].$$

Hence, if $F' = f$ for a differentiable transformation F, the chain rule gives

$$g^f(\theta) = \nabla J_F(\theta), \qquad J_F(\theta) := \mathbb{E}_Q[F\big(\rho_Q(\theta)\big)]. \tag{6}$$

Thus, prompt-level shaping replaces the original mean-reward RLVR objective $J(\theta) = \mathbb{E}_Q[\rho_Q(\theta)]$ with the transformed objective J_F. The resulting updates are therefore biased with respect to $\nabla J(\theta)$, changing the relative contributions of prompts with different pass probabilities. The infinite-group surrogates F for the principal shaping rules are listed in Table 1. For finite groups, Davis & Recht (2025) showed that the expected finite group gradients induced by a shaping function f can also be represented as gradients of transformed objectives with a surrogate transformation F_G:

$$g_G^f(\theta) = \nabla J_{F_G}(\theta), \qquad J_{F_G}(\theta) = \mathbb{E}_Q[F_G(\rho_Q(\theta))]. \tag{7}$$

Equation 7 identifies the optimization objective behind GRPO. For gradient-based algorithm design, however, the primitive F_G is not itself the operational quantity. The expected update is governed by its derivative F_G'. An objective-level representation alone, therefore, does not answer which pass probability regimes are amplified or attenuated, how the allocation changes with group size and shaping strength, or how shaping can be varied over training while recovering stationarity for the original mean-reward objective. *This paper works out theoretical insights in these directions.*

1.2 From finite-group surrogate objectives to gradient-allocation dynamics

Davis & Recht (2025) identified the optimization objective behind GRPO. Our contribution is to characterize the marginal prompt weights governing the expected update and to analyze their consequences for prompt allocation and stationarity. To this end, we identify, normalize, and analyze the marginal prompt weights F_G' that govern the expected update. This turns the existence of a surrogate objective into quantitative statements about gradient allocation, amplification thresholds, effective prompt sampling, and motivates annealing schedules that asymptotically remove the shaping bias. The distinction is important: F_G describes the accumulated surrogate utility assigned to a pass probability, whereas F_G' describes the marginal optimization effort assigned to a prompt at the current iterate. Our analysis focuses on the latter. The chain-rule representation

$$g_G^f(\theta) = \mathbb{E}_Q\big[\underbrace{F_G'\left(\rho_Q(\theta)\right)}_{\text{additional group factor}} \cdot \underbrace{\nabla_\theta \rho_Q(\theta)}_{\text{policy gradient direction}}\big] \neq \mathbb{E}_Q[\nabla_\theta \rho_Q(\theta)] = \nabla_\theta J(\theta).$$

of the expected gradient dynamics can be interpreted in two ways:

(i) **Prompt-dependent length adjustment**: The group estimator averages standard policy gradient directions $\nabla_\theta \rho_Q(\theta)$, with prompt-dependent length adjustments $F_G'(\rho_Q(\theta))$. If $F_G'(\rho_Q(\theta)) > 1$, signals are intensified. If $F_G'(\rho_Q(\theta)) < 1$, signals are weakened.

(ii) **Change of sampling:** Changing measures yields $g_G^f(\theta) = Z_G^f(\theta)\mathbb{E}_{Q\sim\mathbb{P}_\theta}\left[\nabla_\theta \rho_Q(\theta)\right]$ with the new prompt sampling distribution $\mathbb{P}_\theta^f(q) = \frac{F_G'(\rho_q(\theta))}{Z_G^f(\theta)}\mathbb{P}_Q(q)$, where $Z_G^f(\theta)$ normalizes $F_G'(\rho_q(\theta))$ to a probability. Hence, the effect of advantage shaping on the expected gradient is equivalent to computing policy gradient updates $\nabla_\theta \rho_Q(\theta)$ with prompts sampled from changing distributions $\mathbb{P}_\theta^f$ rather than uniformly, and local change of learning-rate.

These interpretations are only interesting if F_G' can be made explicit. While Davis & Recht (2025) derive a Bernstein-polynomial representation of the finite-group objective for standard GRPO, the same can be done in greater generality. The formulas are presented in Proposition 2, while the two interpretations (i) and (ii) are discussed in Sections 2.2 and 2.3, respectively. In fact, Proposition

Table 1: Prompt-level advantage shaping and induced objectives for α-GRPO and related methods.

Method	$f(\rho) = F'(\rho)$	$F(\rho)$
RLOO / Dr. GRPO	1	ρ
GRPO	$\frac{1}{\sqrt{\rho(1-\rho)}}$	$2\arcsin(\sqrt{\rho})$
Uniform-R	$\frac{1}{\rho(1-\rho)}$	$\log\left(\frac{\rho}{1-\rho}\right)$
α-GRPO (ours) $\alpha \in (0,1]$	$\frac{1}{(\rho(1-\rho))^{1-\alpha}}$	$B_\rho(\alpha,\alpha)$

For $\alpha \in (0,1]$, the corresponding optimization objective involves the incomplete Beta function

$$B_\rho(\alpha,\alpha) := \int_0^\rho u^{\alpha-1}(1-u)^{\alpha-1}\,du,$$

with the $\alpha = 0$ endpoint understood through the logarithmic limit. The parameter α can be kept fixed or annealed toward 1, thereby recovering stationarity bounds for the mean-reward objective.

2 shows that the functions F'_G are explicit polynomials, also for a more general class of shaping functions. For this purpose, we introduce α**-GRPO**, a one-parameter family of prompt-level shaping functions indexed by $\alpha \in [0,1]$, $f_\alpha(\rho) := \big(\rho(1-\rho)\big)^{\alpha-1}$, with corresponding advantage estimators $\widehat{A}^{\alpha}_{b,i} = \frac{R_{b,i}-\widehat{\rho}_b}{(\widehat{\rho}_b(1-\widehat{\rho}_b))^{1-\alpha}}$. For collapsed groups, the advantage is set to zero, which is consistent because $R_i - \widehat{\rho} = 0$; hence $f_\alpha(0)$ and $f_\alpha(1)$ need not be defined. The special cases $\alpha \in \{0, \frac{1}{2}, 1\}$ recover Uniform-R, GRPO, and the unnormalized RLOO/Dr. GRPO rule. The parameter α controls the prompt-dependent scaling factor $\big(\widehat{\rho}_b(1-\widehat{\rho}_b)\big)^{\alpha-1}$ applied to each prompt's contribution to the batch gradient. Quantitative effects of α will be analyzed in the next sections. Since this determines how optimization effort is allocated across prompts, we refer to α as the *prompt-allocation parameter*.

In a nutshell, our **main contributions** are the following.

- We show that, for the group-based shaping rules considered here, the expected update admits interpretations as prompt-level gradient amplification and adaptive prompt sampling.
- We introduce a natural interpolating class of shaping functions, α-GRPO, that includes GRPO for $\alpha = \frac{1}{2}$, for which all required quantities can be computed explicitly.
- We show that DAPO-style dynamic filtering preserves the GRPO population-update direction and acts only as a state-dependent stepsize rescaling.
- We rigorously prove uniform smoothness and stationarity guarantees for the normalized α-GRPO finite-group objectives, including annealing that asymptotically removes the mean-reward bias.

2 Group-size and prompt-allocation effects on α-GRPO

2.1 Expected Gradient Structure of α-GRPO

Motivated by the discussion above, we first identify the pass-probability transformation $F_{G,\alpha}$ underlying the objective $J_{G,\alpha}(\theta) = \mathbb{E}_Q[F_{G,\alpha}(\rho_Q(\theta))]$, together with its derivative $F'_{G,\alpha}$. This extends the Bernstein-polynomial representation of standard GRPO from Davis & Recht (2025) to the entire α-GRPO family. The following sections then use $F'_{G,\alpha}$ to characterize how group size and advantage shaping affect the expected gradient.

Proposition 2 (Finite-group gradient weights for α-GRPO)**.** *Let $G \geq 2$ and $\alpha \in [0,1]$, and define $I_\rho(a,b) := \frac{B_\rho(a,b)}{B(a,b)}$. Then*

$$F_{G,\alpha}(\rho) = \frac{1}{G}\sum_{k=0}^{G-2}\left(\frac{k+1}{G}\left(1-\frac{k+1}{G}\right)\right)^{\alpha-1} I_\rho(k+1, G-1-k). \tag{8}$$

In particular, the crucial derivative takes the polynomial form

$$F'_{G,\alpha}(\rho) = \frac{G-1}{G}\sum_{k=0}^{G-2}\left(\frac{k+1}{G}\left(1-\frac{k+1}{G}\right)\right)^{\alpha-1}\binom{G-2}{k}\rho^k(1-\rho)^{G-2-k}. \tag{9}$$

The proof is given in Appendix C. The derivative $F'_{G,\alpha}$ also admits a useful probabilistic representation for studying the infinite-group limit. Let $S_G \sim \mathrm{Bin}(G-2,\rho)$. Then

$$F'_{G,\alpha}(\rho) = \frac{G-1}{G}\,\mathbb{E}\Big[\Big(\frac{S_G+1}{G}\Big(1-\frac{S_G+1}{G}\Big)\Big)^{\alpha-1}\Big].$$

For fixed $\rho \in (0,1)$, the law of large numbers implies $\frac{S_G+1}{G} \to \rho$ a.s. Hence, $\lim_{G\to\infty} F'_{G,\alpha}(\rho) = (\rho(1-\rho))^{\alpha-1} = f_\alpha(\rho)$. A similar argument shows that $\lim_{G\to\infty} F_{G,\alpha}(\rho) = B_\rho(\alpha,\alpha)$.

To facilitate comparisons we normalize $F_{G,\alpha}$ to $\widetilde{F}_{G,\alpha}$ so that $\widetilde{F}_{G,\alpha}(0) = 0$ and $\widetilde{F}_{G,\alpha}(1) = 1$:

$$\widetilde{F}_{G,\alpha}(\rho) := \frac{F_{G,\alpha}(\rho) - F_{G,\alpha}(0)}{F_{G,\alpha}(1) - F_{G,\alpha}(0)}, \quad \rho \in [0,1].$$

Since constant shifts do not affect stationary points and global scaling can be absorbed into the learning rate one can work without loss of generality with $\widetilde{F}_{G,\alpha}$. Examples of $\widetilde{F}_{G,\alpha}$ are provided in Figure 1. Mean-reward ($\alpha = 1$) corresponds to the identity function. Decreasing α results in larger deviations from mean-reward.

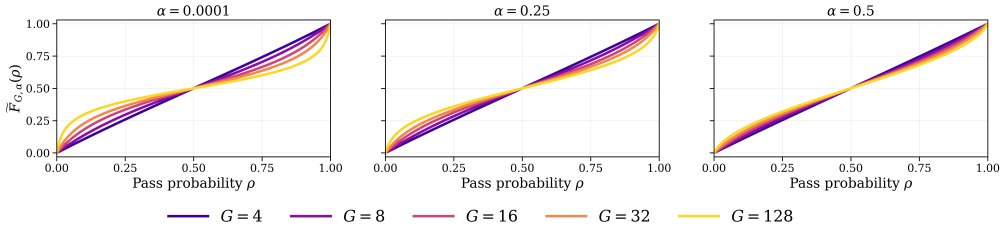

Figure 1: $\widetilde{F}_{G,\alpha}$ for different α and group-size G. The case $\alpha = 0$ is approximated by $\alpha = 10^{-4}$.

2.2 Prompt-level weighting as adaptive gradient amplification

The expected α-GRPO gradient optimizes the normalized objective $\widetilde{J}_{G,\alpha}(\theta) := \mathbb{E}_Q\big[\widetilde{F}_{G,\alpha}(\rho_Q(\theta))\big]$, which is a transformed version of the mean-reward objective $J(\theta) = \mathbb{E}_Q[\rho_Q(\theta)]$. Prior surrogate-objective results identify the scalar function optimized in expectation. We now study its derivative, which determines the marginal weight assigned to each prompt's policy-gradient direction. Figure 1 motivates the following decomposition of $\widetilde{F}_{G,\alpha}$ into the average mean-reward contribution (identity function) and a remainder (sine-type function that we can compute):

$$\widetilde{F}_{G,\alpha}(\rho) = \rho + \mathcal{R}_{G,\alpha}(\rho),$$

where we call $\mathcal{R}_{G,\alpha}(\rho) := \widetilde{F}_{G,\alpha}(\rho) - \rho$ the *prompt-allocation potential*. The left-hand side of Figure 2 illustrates the potentials for $G = 4$ and $G = 64$. Consequently,

$$\nabla_\theta \widetilde{J}_{G,\alpha}(\theta) = \mathbb{E}_Q[\nabla_\theta \widetilde{F}_{G,\alpha}(\rho_Q(\theta))] = \mathbb{E}_Q[\omega_{G,\alpha}(\rho_Q(\theta))\nabla_\theta \rho_Q(\theta)],$$

where $\omega_{G,\alpha}(\rho) = \widetilde{F}'_{G,\alpha}(\rho) = 1 + \mathcal{R}'_{G,\alpha}(\rho)$. Thus, the expected α-GRPO update can be viewed as the mean-reward policy gradient with a prompt-dependent multiplicative weight $\omega_{G,\alpha}$. Since $\omega_{G,1} \equiv 1$, RLOO (resp. Dr.GRPO) recovers the mean-reward gradient, whereas $\alpha < 1$ introduces prompt-dependent bias that becomes more pronounced as α decreases; see Figure 2.

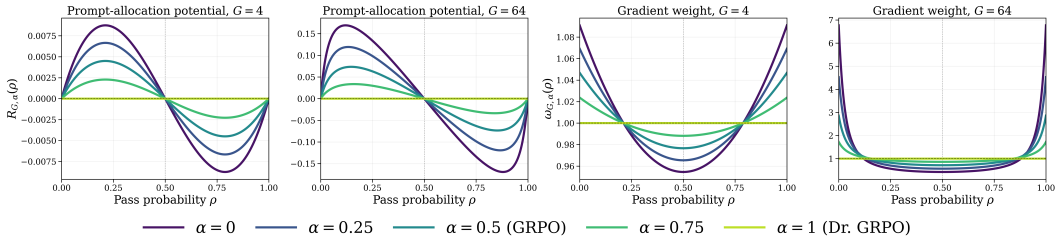

Figure 2: Prompt allocation potential $\mathcal{R}_{G,\alpha}(\rho)$ (left) and gradient weights $\omega_{G,\alpha}(\rho)$ (right)

The explicit form of $\omega_{G,\alpha}$ reveals which prompts are amplified or attenuated. For prompts with $\omega_{G,\alpha}(\rho_q(\theta)) > 1$ the original policy gradient signal $\nabla_\theta \rho_q(\theta)$ is strengthened, while for $\omega_{G,\alpha}(\rho_q(\theta)) < 1$ the gradient signal is weakened. The key to our interpretation is that the function

$\omega_{G,\alpha}$ is explicitly known and the threshold values with $\omega_{G,\alpha}(\rho) = 1$ can be determined (compare right panel of Figure 2).

Let us briefly look into the polynomial structure of the prompt allocation:

Proposition 3 (α-GRPO prompt-allocation potential)**.** *For every $G \geq 4$ and $\alpha \in [0,1]$ it holds that*

$$\mathcal{R}_{G,\alpha}(\rho) = (2\rho - 1) \sum_{j=1}^{\lfloor G/2 \rfloor - 1} c_{G,\alpha,j} \big(\rho(1-\rho)\big)^j .$$

In a simplified form, there exists a polynomial $Q_{G,\alpha}$ of degree at most $\lfloor G/2 \rfloor - 2$ such that

$$\mathcal{R}_{G,\alpha}(\rho) = \rho(1-\rho)(2\rho-1) Q_{G,\alpha}(\rho(1-\rho)) .$$

For fixed G, the polynomial basis functions are independent of α, with all dependence contained in the coefficients $c_{G,\alpha,j}$. Furthermore, $\mathcal{R}_{G,\alpha}$ is antisymmetric in ρ, i.e. $\mathcal{R}_{G,\alpha}(1-\rho) = -\mathcal{R}_{G,\alpha}(\rho)$.

Proof sketch. For integer beta parameters, $I_\rho(k, G-k)$ is a polynomial of degree at most $G-1$. Hence $\widetilde{F}_{G,\alpha}$, and therefore $\mathcal{R}_{G,\alpha}$, is a polynomial of degree at most $G-1$. Symmetry of the Bernstein weights implies $\widetilde{F}_{G,\alpha}(1-\rho) = 1 - \widetilde{F}_{G,\alpha}(\rho)$, which yields $\mathcal{R}_{G,\alpha}(1-\rho) = -\mathcal{R}_{G,\alpha}(\rho)$. Since the remainder vanishes at 0, $1/2$, and 1, it is divisible by $\rho(1-\rho)(2\rho-1)$. The quotient is symmetric under $\rho \mapsto 1-\rho$, hence a polynomial in $\rho(1-\rho)$. The degree bound follows instantly. □

For $G = 4$ the situation is particularly simple: $\omega_{G,\alpha}(\rho) > 1$ holds if and only if

$$\rho < \frac{1}{2}\Big(1 - \frac{1}{\sqrt{3}}\Big) \approx 0.2113 \qquad \text{or} \qquad \rho > \frac{1}{2}\Big(1 + \frac{1}{\sqrt{3}}\Big) \approx 0.7887$$

independently of α; see Example 10 in Appendix D for the derivation. For general group sizes $G > 5$ the threshold is not α-independent. Here is an abstract theorem on threshold values:

Proposition 4 (Prompt-allocation)**.** *Let $G \geq 4$ and $\alpha \in (0,1)$. There exists a unique $r_{G,\alpha} \in (0, \frac{1}{2})$ such that*

$$\omega_{G,\alpha}(\rho) > 1 \quad \Longleftrightarrow \quad \rho < r_{G,\alpha} \text{ or } \rho > 1 - r_{G,\alpha}.$$

The threshold $r_{G,\alpha}$ is the unique solution of a polynomial equation of degree at most $\lfloor \frac{G-2}{2} \rfloor$. Moreover, in the infinite group-size limits, the threshold values converge:

$$\lim_{G\to\infty} r_{G,\alpha} = \frac{1}{2}\Big(1 - \sqrt{1 - 4B(\alpha,\alpha)^{-1/(1-\alpha)}}\Big).$$

Sketch, full proof in Appendix D. The coefficients of $\omega_{G,\alpha}$ are symmetric and strictly decrease toward the center. Hence, $\omega_{G,\alpha}$ is symmetric and strictly U-shaped, implying a unique threshold. Since $\omega_{G,\alpha}(\rho) - 1$ is symmetric, it is a polynomial in $u = \rho(1-\rho)$, yielding the stated degree bound. For fixed $\alpha \in (0,1)$, the finite-group weights converge locally uniformly on $(0,1)$ to $\omega_{\infty,\alpha}(\rho) = \frac{[\rho(1-\rho)]^{-(1-\alpha)}}{B(\alpha,\alpha)}$. Strict monotonicity on $(0, \frac{1}{2})$ gives convergence of the unique roots. Now solve $\omega_{\infty,\alpha}(\rho) = 1$. □

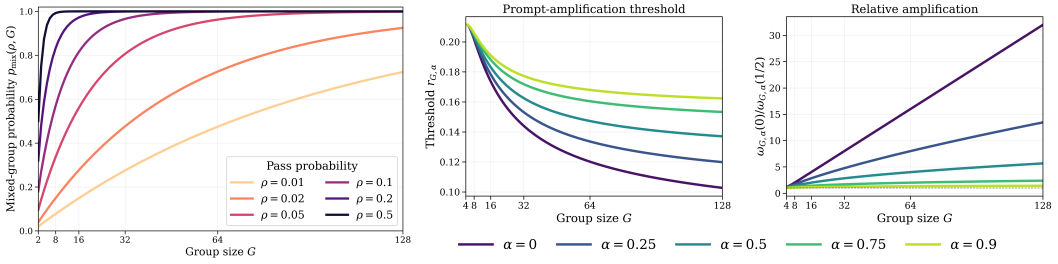

Figure 3: Left: Mixed-group probability only depends on G (independently of α). Middle: Threshold values are shifted when changing G. Right: Maximal amplification of extreme prompts relative to medium prompts increases with G and decreases with α.

Closed-form solutions for $r_{G,\alpha}$ are generally unavailable for arbitrary G (requires finding roots for polynomials of arbitrary order), but they can be derived for small group sizes; see Example 11 for the cases $G = 6, 7$. Using numerical solvers for the polynomial roots, Figure 3 shows numerical approximations of $r_{G,\alpha}$ together with the corresponding gradient amplification factors.

2.3 PROMPT-LEVEL WEIGHTING: THE VIEW AS IMPLICIT ADAPTIVE PROMPT SAMPLING

The finite-group gradient from normalized $\widetilde{F}_{G,\alpha}$ is $\widetilde{g}_{G,\alpha}(\theta) = \mathbb{E}_Q\left[\omega_{G,\alpha}\big(\rho_Q(\theta)\big)\nabla_\theta \rho_Q(\theta)\right]$ which admits an interpretation as reweighted prompt sampling:

Proposition 5. *Define* $Z_{G,\alpha}(\theta) := \mathbb{E}_Q\left[\omega_{G,\alpha}\big(\rho_Q(\theta)\big)\right]$. *Whenever* $Z_{G,\alpha}(\theta) > 0$,

$$\widetilde{g}_{G,\alpha}(\theta) = \underbrace{Z_{G,\alpha}(\theta)}_{\textit{adaptive step-size}} \cdot \underbrace{\mathbb{E}_{Q\sim\mathbb{P}_\theta^{G,\alpha}}\left[\nabla_\theta \rho_Q(\theta)\right]}_{\textit{non-uniform prompt sampling}}, \tag{10}$$

with non-uniform prompt sampling distribution $\mathbb{P}_\theta^{G,\alpha}(q) := \frac{\omega_{G,\alpha}(\rho_q(\theta))}{Z_{G,\alpha}(\theta)}\,\mathbb{P}_Q(q)$.

Proof. The measure-change is well-defined because $\int_Q \frac{\omega_{G,\alpha}\left(\rho_q(\theta)\right)}{Z_{G,\alpha}(\theta)}\,\mathrm{d}\mathbb{P}_Q(q) = \frac{Z_{G,\alpha}(\theta)}{Z_{G,\alpha}(\theta)} = 1$. The claim then follows from the one-line computation

$$\widetilde{g}_{G,\alpha}(\theta) = \mathbb{E}_Q\left[\omega_{G,\alpha}\big(\rho_Q(\theta)\big)\nabla_\theta \rho_Q(\theta)\right] = Z_{G,\alpha}(\theta)\,\mathbb{E}_{Q\sim\mathbb{P}_\theta^{G,\alpha}}\left[\nabla_\theta \rho_Q(\theta)\right]. \qquad \square$$

Equation 10 yields two complementary interpretations compared to the standard policy gradient:

- The expected α-GRPO gradient can be interpreted as changing the prompt sampling distribution from uniform sampling to a distribution proportional to $\omega_{G,\alpha}\big(\rho_q(\theta)\big)$, compare Figure 2. A prompt is upweighted relative to the original prompt distribution if and only if $\omega_{G,\alpha}\big(\rho_q(\theta)\big) > Z_{G,\alpha}(\theta)$. Since $\omega_{G,\alpha}$ is U-shaped, this sampling favors sufficiently hard and easy prompts relative to prompts with intermediate pass probabilities.
- The expected α-GRPO gradient locally changes the learning rate to an effective learning rate $\eta_{\text{eff}}(\theta) = \eta\, Z_{G,\alpha}(\theta)$. More hard/easy prompts lead to larger learning rates.

Remark 6 (Relating GRPO and DAPO-style dynamic sampling, see Appendix E for details). The attentive reader might recognize a relation to the dynamic sampling mechanism of DAPO Yu et al. (2025). Dynamic sampling in the spirit of DAPO oversamples prompt–response groups and retains only mixed-reward groups. Let $M_G(\theta) := \mathbb{E}_Q\big[p_{\text{mix}}(\rho_Q(\theta))\big]$ denote the average probability of mixed groups. By Bayes' rule, conditioning on acceptance induces the prompt distribution

$$\mathbb{P}_\theta^{\text{DS}}(q) = \frac{p_{\text{mix}}(\rho_q(\theta))}{M_G(\theta)}\,\mathbb{P}_Q(q), \qquad \mathbb{P}_Q = \text{Uniform}(\mathcal{Q}).$$

If this filtering is applied to an otherwise unchanged α-GRPO estimator and rejected groups are replaced by fresh prompt–group proposals, then

$$g_{G,\alpha}^{\text{DS}}(\theta) = \frac{1}{M_G(\theta)}\,\nabla J_{G,\alpha}(\theta).$$

The dynamic-sampling component of DAPO yields the same expected gradient direction as GRPO, up to an additional state-dependent rescaling of the learning rate. Although the accepted-prompt distribution favors prompts with larger mixed-group probability, the conditional update magnitude contributes the corresponding inverse factor. Consequently, the net expected update differs only by a global state-dependent scale. We emphasize that this statement concerns only the dynamic-filtering component in *isolation*, rather than all details of the DAPO algorithm.

2.4 α-ANNEALING: DECOUPLING THE EFFECTS.

We introduced α-GRPO as an interpolation between GRPO ($\alpha = \frac{1}{2}$) and RLOO ($\alpha = 1$), with even more extreme prompt reweighting for small α. Let us summarize the effects (compare Figure 3):

gradient activity	:	mixed-group probabilities p_{mix} increase as G increases,
threshold values	:	$r_{G,\alpha}$ can be shifted by changing α,
prompt weighting	:	$\omega_{G,\alpha}(\rho)$ becomes less tail-focused as α increases,
mean-reward bias	:	gradient bias relative to $\nabla_\theta J(\theta)$ decreases as α increases.

This suggests the following interpretation of the available design parameters:

- G controls the probability p_{mix} of observing minority outcomes,
- α controls the strength of prompt reweighting: smaller α more strongly counteracts weak gradient signals, but also introduces greater bias relative to the mean-reward objective.

If one wishes to keep a fixed threshold for prompt amplification while increasing the group size to obtain sufficient gradient signals, one can increase G while slightly increasing α to compensate. If, instead, stronger gradient signals are desirable early in training while the final optimization target should remain the mean-reward objective, we propose annealing α toward 1 over training. We call this approach **α-annealing** and show in the next section that annealed α-GRPO admits stationarity bounds for the mean-reward objective.

Algorithm 1 Annealed α-GRPO

Require: schedule (α_t) with $\alpha_t \uparrow 1$
1: **for** $t = 0, \dots, T-1$ **do**
2: Sample a GRPO batch.
3: Compute group rewards and $\widehat{\rho}_b$.
4: Form shaped advantages

$$\widehat{A}_{b,i}^{\alpha_t} = (\widehat{\rho}_b(1-\widehat{\rho}_b))^{\alpha_t - 1}(R_{b,i} - \widehat{\rho}_b).$$

5: Perform the usual GRPO gradient update.
6: **end for**

3 Stochastic optimization analysis

The above characterized how advantage shaping modifies expected gradients. This raises two optimization questions. Does the strong amplification of hard and easy prompts deteriorate the regularity of the induced surrogate objective? Can the resulting prompt-allocation bias be removed by annealing α towards 1? We next prove that normalized finite-group objectives remain uniformly smooth over G and α, and that annealing can drive the prompt-dependent weighting back to the constant mean-reward weighting. This gives guarantees relative to the *original* mean-reward objective.

Assumption 7. There exist constants $\Pi_*, L_s > 0$ such that $\sup_{q\in\mathcal{Q},\, o\in\mathcal{O}} \|\nabla_\theta \log \pi_\theta(o \mid q)\| \le \Pi_*$ and $\sup_{q\in\mathcal{Q},\, o\in\mathcal{O}} \|\nabla_\theta \log \pi_\theta(o \mid q) - \nabla_\theta \log \pi_{\theta'}(o \mid q)\| \le L_s\|\theta - \theta'\|$ for all $\theta, \theta' \in \mathbb{R}^d$.

The assumption is standard in policy-gradient theory and holds, for example, on compact parameter domains for smooth policy parameterizations. For simplicity, we assume it globally. Next, define $C_{G,\alpha} := F_{G,\alpha}(1) - F_{G,\alpha}(0)$ and recall that $\nabla J_{G,\alpha} = C_{G,\alpha}\nabla \widetilde{J}_{G,\alpha}$. For arbitrary $G \ge 2,\ \alpha \in [0,1]$, define the normalized stochastic gradient $\widetilde{g}^{\alpha}_{B,G} := C_{G,\alpha}^{-1}\widehat{g}^{\alpha}_{B,G}$. For a schedule $(\alpha_t)_{t\ge 0} \subset [0,1]$ and $\eta > 0$, let $\gamma_t := \eta C_{G,\alpha_t}$ and consider the update

$$\theta_{t+1} = \theta_t + \eta\widehat{g}^{\alpha_t}_{B,G}(\theta_t) = \theta_t + \gamma_t\widetilde{g}^{\alpha_t}_{B,G}(\theta_t), \qquad \alpha_t \in [0,1],\ t \ge 0\,. \tag{11}$$

Let $(\mathcal{F}_t)_{t\ge 0}$ denote the filtration generated before the samples at iteration t are drawn. Under fresh on-policy sampling, $\mathbb{E}[\widetilde{g}^{\alpha}_{B,G}(\theta_t) \mid \mathcal{F}_t] = \nabla\widetilde{J}_{G,\alpha}(\theta_t)$ and $\mathbb{E}[\|\widetilde{g}^{\alpha}_{B,G}(\theta_t) - \nabla\widetilde{J}_{G,\alpha}(\theta_t)\|^2 \mid \mathcal{F}_t] \le 16\Pi_*^2/B$. Appendix F.2 shows that all $\widetilde{J}_{G,\alpha}$ are L-smooth with the same $L = 4L_s + 20\Pi_*^2$.

Theorem 8 (Uniform stationarity of the normalized objective)**.** *Suppose that Assumption 7 holds. Fix $G \ge 2$, $\alpha_t = \alpha \in [0,1]$ and let $\eta \in (0, \frac{1}{C_{G,\alpha}L}]$, $\gamma := \eta C_{G,\alpha}$. Consider the iterates $(\theta_t)_{t\ge 0}$ generated by Equation 11. For $T \ge 2$, let $R \sim \mathrm{Unif}\{0,\dots,T-1\}$ independent of $\mathcal{F}_{T-1}$. Then it holds*

$$\mathbb{E}[\|\nabla\widetilde{J}_{G,\alpha}(\theta_R)\|^2] \le \frac{2}{\gamma T} + \frac{16L\Pi_*^2\gamma}{B}.$$

The proof is given in Appendix F.3.2. This is a standard nonconvex stochastic-gradient stationarity behavior: the optimization term decays as $(\gamma T)^{-1}$, while a constant stepsize leaves a variance-controlled error floor. See, e.g., Bottou et al. (2018, Sec. 4.3). The key point is that, after removing the global scale $C_{G,\alpha}$, neither the smoothness constant nor the stochastic-gradient bound deteriorates with G or α. Thus, the prompt amplification identified in Section 2 changes the surrogate objective and the gradient directions, but not the worst-case optimization regularity captured by these bounds. The link to Section 2 is explicit. Since $F'_{G,\alpha} = C_{G,\alpha}\omega_{G,\alpha}$ and $F'_{G,1} \equiv c_G$, the bias of the update is

$$e_t := \nabla J_{G,\alpha_t}(\theta_t) - c_G\nabla J(\theta_t) = \mathbb{E}_Q\big[(C_{G,\alpha_t}\omega_{G,\alpha_t}(\rho_Q(\theta_t)) - c_G)\nabla\rho_Q(\theta_t)\big].$$

Appendix F.4.1 gives the uniform bound $\|e_t\| \le \beta_t := \Pi_* c_G[((G-1)/G^2)^{\alpha_t - 1} - 1]$, which vanishes for fixed G whenever $\alpha_t \to 1$. Moreover, Appendix F.3.1 gives a uniform variance bound $\Gamma := 2\Pi_*$, valid for all $\alpha_t \in [0,1]$ and Appendix F.2 shows smoothness of J with $L_J = L_s + \Pi_*^2$.

Theorem 9 (Stationarity bound relative to the RLVR objective)**.** *Suppose that Assumption 7 holds and let $J^* := \sup_\theta J(\theta)$. Fix $G \ge 2$, let $(\alpha_t)_{t\ge 0}$ satisfy $\alpha_t \in [0,1]$ for all $t \ge 0$ and $\lim_{t\to\infty}\alpha_t = 1$, and let $\eta \in (0, \frac{1}{2L_J c_G}]$. For $T \ge 2$, let $R \sim \mathrm{Unif}\{0,\dots,T-1\}$ independent of $\mathcal{F}_{T-1}$. Then we have*

$$\mathbb{E}\big[\big\|\nabla J(\theta_R)\big\|^2\big] \le \frac{4\big(J^* - J(\theta_0)\big)}{c_G\eta T} + \frac{8}{c_G^2 T}\sum_{t=0}^{T-1}\beta_t^2 + \frac{2L_J\eta\Gamma^2}{c_G B}.$$

The proof is given in Appendix F.4.2. The three terms correspond to optimization error, prompt-shaping bias, and stochastic variance. The middle term is precisely the cumulative cost of the prompt-dependent reweighting characterized in Section 2. If $\alpha_t \to 1$, then $\beta_t \to 0$ and hence $T^{-1}\sum_{t<T}\beta_t^2 \to 0$. Annealing therefore removes the shaping-specific bias term. With constant stepsize, the stochastic variance term remains. Appendix F.4.2 shows that Robbins-Monro stepsizes yield vanishing mean-reward stationarity under the corresponding weighted-bias condition.

4 Experiments

The insights of this paper are of theoretical nature. We still present initial experiments to show that advantage shaping $\frac{1}{(\rho(1-\rho))^{1-\alpha}}$ with constant or annealed α can have practical benefits. Since theory for AdamW optimizer is out of scope, our theory results of Section 3 hold for SGD only. We thus follow best practice from learning-rate scheduling, where SGD results also suggest Robbins-Monro type annealing, while practical implementations often use linear or cosine schedules.

We train Qwen2.5-3B, Qwen2.5-1.5B, and Qwen2.5-Math-1.5B on MATH and vary only α, while keeping the remaining training configuration, including response-level length normalization and nominal learning rate, fixed. Unlike the normalized analysis above, the implementation does not compensate for the α-dependent global scale of the raw gradient estimator; gradients are subsequently optimized with AdamW. For further results and implementation details see Appendix B.

Table 2: Average performance across six math benchmarks. For mean@k, we use $k = 64$ for AIME24, AIME25, and AMC23, and $k = 16$ for MATH500, Minerva Math, and OlympiadBench.

	Qwen2.5-3B		Qwen2.5-1.5B		Qwen2.5-Math-1.5B	
Method	mean@k	pass@8	mean@k	pass@8	mean@k	pass@8
GRPO ($\alpha = 0.5$)	25.09	39.09	20.08	35.03	27.87	45.52
Dr. GRPO ($\alpha = 1$)	25.50	41.34	**20.60**	35.42	**28.44**	**46.45**
$\alpha = 0$	25.21	41.03	20.39	35.63	27.57	45.85
$\alpha = 0.25$	25.58	40.34	19.77	34.12	28.10	46.32
$0 \nearrow 1$ linear	**25.82**	**41.91**	19.73	35.65	27.17	44.99
$0 \nearrow 1$ cosine	25.20	40.76	20.41	**36.13**	27.22	45.98

No shaping rule consistently dominates across model initializations. Linear $0 \nearrow 1$ annealing performs best on average for Qwen2.5-3B, whereas Dr. GRPO is strongest for Qwen2.5-Math-1.5B, and the ordering differs again for Qwen2.5-1.5B. Since differences are generally modest, we do not interpret these results as evidence for a universally superior α. Rather, they indicate that prompt-level shaping is an empirically relevant but setting-dependent design choice, consistent with our theory in which its effect depends on the policy-induced pass probability distribution.

5 Limitations and future work

Our contributions range from new interpretation of the expected gradient as prompt-level gradient amplification and dynamically changing prompt sampling, over the introduction of a mathematically tractable class of shaping functions (α-GRPO), to a rigorous SGD stationary point analysis under the usual RL theory assumptions. In particular, we prove that optimization bias obtained in policy gradients with shaped advantages can be turned down to zero if the prompt-allocation parameter α is annealed towards 1. For future work it would be most interesting to consider more flexible advantage scaling profiles f (e.g. asymmetric) and go beyond a first moment analysis. How to quantitatively connect the new expected gradient interpretations and the mixed-group probabilities for vanishing gradients (pathwise or at least variances)? Can adaptive α-tuning give good practical algorithms?

Limitations: (i) This article is primarily a theory paper, with a few experiments highlighting that different advantage scaling and annealing can be helpful. Training LLMs is expensive, so we had to stick to one seed and models of modest size. For illustrative purposes the small experiments do their job, future work should contain a significantly extended experimental study. (ii) The article considers purely advantage scaling, no further GRPO details such as clipping or length normalization. Length normalization makes gradients non-conservative, therefore hard to analyze mathematically.

AI USE STATEMENT

We used LLMs to assist in polishing the writing of this paper, improving grammar, clarity, and style. LLMs were also used during the development of the theoretical analysis to explore derivation steps, verify intermediate calculations, and discuss alternative proof strategies. All mathematical arguments, proofs, and theoretical results were critically verified by the authors, who determined the final proof structure and take full responsibility for their correctness. All core intellectual contributions, including the proposed methods, theoretical framework, experimental design, and analysis, were conceived and executed by the human authors.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, our implementation is based on the Open-R1 codebase (Hugging Face, 2025). Comprehensive details regarding the experimental setup, including model configurations and all hyperparameters, are described in Section 4 and further elaborated in Appendix B.

REFERENCES

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Yanqi Dai, Yuxiang Ji, Xiao Zhang, Yong Wang, Xiangxiang Chu, and Zhiwu Lu. Harder is better: Boosting mathematical reasoning via difficulty-aware GRPO and multi-aspect question reformulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nfURupkdRJ.

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning?, 2025. URL https://arxiv.org/abs/2510.13651.

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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

Reinhard Heckel, Mahdi Soltanolkotabi, and Christos Thrampoulidis. Asymmetric prompt weighting for reinforcement learning with verifiable rewards. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nYwP9hcr38.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Jefferson Hernandez, Jaywon Koo, Zilin Xiao, Chen Wei, and Vicente Ordonez. SoftmaxGRPO: Learning to reason using softmax advantage group estimation. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=DblODk8V9B.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1. GitHub repository, 2025. URL https://github.com/huggingface/open-r1. Accessed 2026-07-13.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

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 *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, Daria Korotyshova, and Daniil Gavrilov. F-grpo: Don't let your policy learn the obvious and forget the rare, 2026. URL https://arxiv.org/abs/2602.06717.

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

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

Joe Suk and Yaqi Duan. On the optimization dynamics of rlvr: Gradient gap and step size thresholds, 2026. URL https://arxiv.org/abs/2510.08539.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 1999.

Christos Thrampoulidis, Sadegh Mahdavi, and Wenlong Deng. Advantage shaping as surrogate reward maximization: Unifying pass@k policy gradients. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=R1RhBFUk8t.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. Bnpo: Beta normalization policy optimization, 2025. URL https://arxiv.org/abs/2506.02864.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

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. *CoRR*, abs/2507.18071, July 2025. URL https://doi.org/10.48550/arXiv.2507.18071.

Yifu Zheng. Rl2ml: Finite-rollout surrogate objectives from reinforcement learning to maximum likelihood, 2026. URL https://arxiv.org/abs/2605.30154.

A MORE RELATED WORK

Policy-gradient methods. Policy-gradient methods optimize a parameterized policy directly using score-function estimators of the expected return (Williams, 1992; Sutton et al., 1999). For the use of non-linear optimization techniques that we also use in Section 3 we refer for instance to Agarwal et al. (2021). Most importantly, the (strong) assumptions of our Assumption 7 are standard in the RL community. In language-model post-training, policy-gradient methods are typically applied to sequence-level rewards, with PPO historically relying on a learned critic. More recent critic-free methods, including RLOO and GRPO, estimate baselines from multiple sampled completions instead (Kool et al., 2019; Ahmadian et al., 2024; Shao et al., 2024). Our work focuses on how the resulting group-based advantage normalization changes the expected policy-gradient field and its underlying optimization objective.

GRPO-style estimators and prompt weighting. Several works modify different components of group-relative policy optimization. RLOO and Dr. GRPO remove within-group standard-deviation normalization (Ahmadian et al., 2024; Liu et al., 2025), while DAPO and GSPO modify sampling, clipping, or importance weighting (Yu et al., 2025; Zheng et al., 2025). Other methods directly alter how learning signal is allocated across prompts: DGPO, F-GRPO, and Heckel et al. (2026) introduce difficulty-aware or explicit prompt weighting (Dai et al., 2026; Plyusov et al., 2026; Heckel et al., 2026), while BNPO adaptively normalizes rewards based on the distribution of prompt success probabilities (Xiao et al., 2025). SoftmaxGRPO instead replaces z-score normalization with softmax-based group advantages and analyzes the resulting prompt-difficulty weighting (Hernandez et al., 2026). In contrast, we study the symmetric α-GRPO shaping family and characterize how (G, α) determine its finite-group prompt allocation and its evolution under annealing.

Our analysis deliberately isolates prompt-level advantage shaping from other components of practical GRPO implementations. In particular, response-length normalization, clipping, and stopped-gradient importance weighting can alter the population vector field and may destroy its representation as the gradient of a scalar objective. Extending the present convergence analysis to such non-conservative update fields is an interesting direction for future work.

Surrogate objectives and prompt-allocation views. Davis & Recht (2025) show that several binary-reward learning rules optimize monotone transformations of prompt-level success probability and derive a Bernstein-polynomial representation of the corresponding finite-group objectives; in particular, idealized GRPO yields an arcsine-square-root transformation. Thrampoulidis et al. (2026) similarly interpret advantage shaping through surrogate reward objectives and connect the resulting transformations to Pass@K-type policy gradients. More recently, Zheng (2026) derive finite-rollout surrogate objectives and corresponding estimators for an asymmetric power-weighting family that increasingly emphasizes low-success prompts. In contrast, we study the surrogate objectives induced by symmetric GRPO-style shaping and use their derivatives to characterize finite-group prompt allocation. SoftmaxGRPO likewise derives the finite-group population objective induced by an alternative group-advantage estimator which uses temperature-scaled softmax advantages (Hernandez et al., 2026).

Building on this surrogate-objective perspective, we use the finite-group Bernstein analysis for the symmetric α-GRPO shaping family $(\widehat{\rho}(1-\widehat{\rho}))^{\alpha-1}$ and study $F'_{G,\alpha}$, the marginal weight multiplying each prompt's mean-reward policy-gradient direction. Unlike the asymmetric hard-prompt weighting considered in RL2ML, α-GRPO reweights hard and easy prompts symmetrically. The resulting finite-group weights let us characterize amplification thresholds and their dependence on (G, α), reinterpret the expected field as adaptive prompt sampling, and analyze annealing schedules that remove the shaping bias relative to the original mean-reward field.

Our DAPO-style filtering analysis should also be distinguished from the rejection-sampling estimators studied by Davis & Recht (2025, Sec. 5): their updates are formed from accepted correct responses and induce different transformed objectives, whereas we filter otherwise unchanged prompt-group updates that already vanish on rejected homogeneous groups. Under fresh prompt-group proposals, this filtering therefore only rescales the existing population field.

B Supplemental Material on Experiments

Our experiments are intended as a controlled empirical study of the prompt-level advantage-shaping mechanism analyzed in the main text, rather than as a large-scale comparison of complete RLVR training recipes. In particular, we vary the prompt-level shaping parameter α while keeping the remaining training setup fixed. This allows us to study whether changing this theoretically identified degree of freedom alone can affect downstream reasoning performance. We do not expect a universally optimal value of α: according to our analysis, its effect depends on the distribution of prompt pass probabilities induced by the current policy.

B.1 Experimental Setup

Models and training data. We consider three model initializations: Qwen2.5-3B, Qwen2.5-1.5B, and Qwen2.5-Math-1.5B. The first two are general-purpose base models of different sizes, whereas Qwen2.5-Math-1.5B is specialized toward mathematical reasoning. This provides settings with substantially different initial mathematical capabilities.

All models are trained on a 12k-example subset of the MATH dataset (Hendrycks et al., 2021). We use only a rule-based accuracy reward and no explicit KL penalty, i.e., $\beta = 0$. Rollouts are sampled on-policy from the current model. For every experiment, we use the system prompt *"Please reason step by step, and put your final answer within \boxed{}."*

Training is performed for 500 optimization steps on four NVIDIA A100 GPUs with 80GB memory each. Each training configuration uses one fixed training seed; consequently, the experiments should be interpreted as controlled exploratory comparisons rather than precise estimates of expected performance across training seeds.

Compared shaping rules. We compare standard GRPO ($\alpha = 0.5$), Dr. GRPO ($\alpha = 1$), stronger fixed prompt-level shaping with $\alpha \in \{0, 0.25\}$, and linear and cosine schedules that anneal α from 0 to 1 over training. These schedules are simple practical choices rather than direct implementations of the stochastic-approximation conditions in Section 3; the nominal learning rate is held fixed across all configurations.

All methods use the same rollout configuration, optimizer, reward function, response-level length normalization, and nominal learning rate. Thus, unlike comparisons between complete RLVR recipes, the methods considered here differ only in their prompt-level advantage shaping.

Our normalized theoretical analysis in Section 2.2 factors the finite-group objective as

$$\nabla J_{G,\alpha}(\theta) = C_{G,\alpha} \nabla \widetilde{J}_{G,\alpha}(\theta),$$

where normalization by $C_{G,\alpha}$ removes an α-dependent global scale and isolates the relative allocation of gradient weight across prompts. In the experiments, we implement the raw α-GRPO advantages and keep the nominal learning rate fixed across methods, so this global scaling is not explicitly compensated for.

The theoretical analysis characterizes the stochastic gradient estimator and its population expectation, whereas the experiments use AdamW to transform these gradients into parameter updates. Consequently, the factor $C_{G,\alpha}$ should be interpreted as an α-dependent scale of the gradient signal supplied to the optimizer, not as an exact multiplicative change in parameter-update magnitude. AdamW's momentum and adaptive second-moment normalization can attenuate or otherwise transform such global scaling. The experiments therefore compare the practical raw shaping rules rather than isolating the normalized prompt-allocation effect alone.

Training hyperparameters. Our implementation is based on the Open-R1 codebase (Hugging Face, 2025). Since the experiments use TRL version 0.18.0, the modifications required for the different shaping rules were implemented directly in the GRPO trainer. Table 3 summarizes the model-specific training configurations.

The batching configuration is shared across models. With four GPUs, `per_device_train_batch_size=` 16, `gradient_accumulation_steps=` 4, and `steps_per_generation=` 4, each optimizer step uses $4 \cdot 16 \cdot 4 = 256$ generated completions.

Table 3: Training hyperparameters for the three model initializations.

Hyperparameter	Qwen2.5-3B	Qwen2.5-1.5B	Qwen2.5-Math-1.5B
Training steps	500	500	500
Optimizer	AdamW	AdamW	AdamW
Precision	bfloat16	bfloat16	bfloat16
Learning rate	1.0×10^{-6}	1.0×10^{-6}	5.0×10^{-7}
LR scheduler	constant	constant	constant
Warmup ratio	0.03	0.03	0.03
Per-device train batch size	16	16	16
Gradient accumulation steps	4	4	4
Steps per generation	4	4	4
Generations per prompt G	8	8	8
Effective prompt batch size	32	32	32
Completions per optimizer step	256	256	256
Max prompt length	1024	1024	1024
Max completion length	2048	2048	2048
Rollout temperature	1.0	1.0	1.0
Rollout top-p	1.0	1.0	1.0
KL coefficient	0	0	0
Reward function	Accuracy	Accuracy	Accuracy

Since $G = 8$, this corresponds to 32 unique prompts per optimization step. Over 500 steps, each run therefore processes $16{,}000$ prompt instances, or approximately 1.33 prompt-level epochs over the 12k-example training set.

Implementation of α-GRPO. For prompt b, let $r_{b,i}$ denote the reward of completion i, and let $\widehat{\rho}_b$ and $\widehat{\sigma}_b$ denote the mean and standard deviation of rewards within its group. We implement α-GRPO by modifying the reward-scaling step in the GRPO advantage computation:

$$\widehat{A}^{\alpha}_{b,i} = \frac{r_{b,i} - \widehat{\rho}_b}{(\widehat{\sigma}_b + \epsilon)^{2-2\alpha}},$$

where $\epsilon = 10^{-4}$ is used for numerical stability. Since the rewards are binary, $\widehat{\sigma}_b^2 = \widehat{\rho}_b(1 - \widehat{\rho}_b)$, up to the numerical stabilization used in the implementation. Standard GRPO is recovered at $\alpha = 0.5$,

$$\widehat{A}^{0.5}_{b,i} = \frac{r_{b,i} - \widehat{\rho}_b}{\widehat{\sigma}_b + \epsilon},$$

whereas $\alpha = 1$ removes reward-standard-deviation scaling. Smaller values of α apply stronger prompt-level shaping.

For annealed runs, α is varied over optimization steps according to the specified schedule while all other training hyperparameters remain unchanged.

B.2 Evaluation

We evaluate all trained models on six mathematical reasoning benchmarks: AMC23, AIME24, AIME25 (Zhang & Math-AI, 2025), MATH500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). For AIME24, AIME25, and AMC23, we generate $k = 64$ completions per problem; for MATH500, Minerva Math, and OlympiadBench we generate $k = 16$.

Evaluation uses temperature 0.6, top-$p = 0.95$, and a maximum completion length of 2048 tokens. Correctness is determined using the standard Open-R1 verifier, which extracts and compares final answers across supported answer formats.

We report two complementary metrics. First, $\mathrm{mean@}k$ is the fraction of sampled completions that are correct, averaged over problems, and therefore estimates the success probability of an individual sample. Second, $\mathrm{pass@}8$ measures whether at least one response among eight sampled completions is correct. The former is more directly related to the mean-reward objective considered theoretically, whereas the latter captures the ability of the policy to produce at least one successful solution under repeated sampling.

B.3 MEAN-SUCCESS RESULTS

Table 4 reports $\mathrm{mean@}k$. No single shaping rule dominates across all model initializations or benchmarks.

Table 4: $\mathrm{mean@}k$ performance, in percent. We use $k = 64$ for AIME24, AIME25, and AMC23 and $k = 16$ otherwise. Bold values denote the best trained method within each model block. The final column is the unweighted average over all six benchmarks.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-3B	2.76	1.52	22.11	44.16	10.80	7.73	14.85
+ GRPO ($\alpha = 0.5$)	4.68	1.56	38.20	67.23	25.96	12.92	25.09
+ Dr.GRPO ($\alpha = 1$)	4.79	2.39	39.64	**67.30**	25.48	**13.39**	25.50
+ $\alpha = 0$	**6.97**	1.92	37.18	67.15	25.51	12.54	25.21
+ $\alpha = 0.25$	5.88	2.60	38.71	66.41	**27.32**	12.58	25.58
+ $0 \nearrow 1$ linear	6.77	**3.54**	**40.55**	65.87	25.48	12.70	**25.82**
+ $0 \nearrow 1$ cosine	6.04	1.56	38.67	65.78	26.72	12.44	25.20
Qwen2.5-1.5B	0.21	0.00	1.17	2.13	0.83	0.31	0.77
+ GRPO ($\alpha = 0.5$)	3.13	1.35	30.47	57.49	18.61	9.44	20.08
+ Dr.GRPO ($\alpha = 1$)	2.81	1.09	**31.64**	58.18	**20.20**	9.67	**20.60**
+ $\alpha = 0$	3.18	**1.56**	30.70	57.99	19.67	9.23	20.39
+ $\alpha = 0.25$	2.76	0.89	28.05	**58.24**	19.37	9.32	19.77
+ $0 \nearrow 1$ linear	3.54	1.04	28.24	57.58	18.84	9.13	19.73
+ $0 \nearrow 1$ cosine	**4.01**	1.51	29.96	57.48	19.81	**9.69**	20.41
Qwen2.5-Math-1.5B	6.72	3.23	31.72	31.71	7.81	9.58	15.13
+ GRPO ($\alpha = 0.5$)	10.21	4.90	48.91	66.55	23.02	13.62	27.87
+ Dr.GRPO ($\alpha = 1$)	**10.89**	**5.68**	49.45	**67.66**	**23.53**	13.42	**28.44**
+ $\alpha = 0$	10.42	4.95	48.55	65.84	22.22	13.44	27.57
+ $\alpha = 0.25$	9.74	5.52	**49.80**	66.47	23.35	**13.70**	28.10
+ $0 \nearrow 1$ linear	9.84	5.47	48.09	65.15	21.35	13.13	27.17
+ $0 \nearrow 1$ cosine	9.58	5.63	47.70	65.22	21.74	13.46	27.22

For Qwen2.5-3B, the linear $0 \nearrow 1$ schedule obtains the highest average $\mathrm{mean@}k$, although the differences among trained methods are modest. For Qwen2.5-1.5B, Dr. GRPO has the highest average by a small margin, with the cosine schedule and $\alpha = 0$ close behind. For the math-specialized Qwen2.5-Math-1.5B initialization, Dr. GRPO performs best on average, while $\alpha = 0.25$ remains competitive.

Thus, the results do not identify a universally preferred amount of prompt-level shaping. Rather, changing only this component of the training rule can change downstream performance, and the relative ordering of shaping rules varies across model initializations.

B.4 PASS@8 RESULTS

Table 5 reports $\mathrm{pass@8}$ for the same trained models. The qualitative picture is similar to the $\mathrm{mean@}k$ results: the different shaping variants remain close in aggregate performance, and the best-performing rule depends on the model initialization.

For Qwen2.5-3B, linear annealing again obtains the highest average score and performs strongly on several individual benchmarks. On Qwen2.5-1.5B, the cosine schedule has the highest average $\mathrm{pass@8}$, whereas on Qwen2.5-Math-1.5B Dr. GRPO has the highest average. Hence the pass@8 results do not change the central conclusion obtained from mean@k: no shaping rule consistently dominates across model initializations.

B.5 INTERPRETATION AND LIMITATIONS

The experiments support two deliberately limited conclusions. First, prompt-level advantage shaping is an empirically relevant design choice. The compared methods share the same training configuration and differ only in their prompt-level shaping, yet their downstream performance is not identical. Second, the preferred amount of shaping appears to depend on the training setting. This is

Table 5: pass@8 performance, in percent. Bold values denote the best trained method within each model block. The final column is the unweighted average over the six benchmarks.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-3B	12.31	8.58	61.78	79.11	34.07	18.24	35.68
+ GRPO ($\alpha = 0.5$)	12.76	9.10	68.44	83.12	41.38	19.72	39.09
+ Dr.GRPO ($\alpha = 1$)	16.41	12.76	72.47	83.91	41.90	20.58	41.34
+ $\alpha = 0$	**18.64**	11.11	69.89	84.29	42.37	19.89	41.03
+ $\alpha = 0.25$	15.41	11.26	69.20	83.16	41.85	**21.15**	40.34
+ $0 \nearrow 1$ linear	16.76	13.20	**73.41**	**84.56**	**42.84**	20.70	**41.91**
+ $0 \nearrow 1$ cosine	16.06	9.60	72.69	83.59	42.13	20.50	40.76
Qwen2.5-1.5B	1.53	0.00	8.57	13.00	5.28	2.11	5.08
+ GRPO ($\alpha = 0.5$)	9.42	7.89	59.13	**80.55**	35.60	17.61	35.03
+ Dr.GRPO ($\alpha = 1$)	9.76	7.01	61.03	80.03	**36.63**	18.04	35.42
+ $\alpha = 0$	11.29	7.79	**61.23**	79.28	36.45	17.71	35.63
+ $\alpha = 0.25$	9.35	4.72	57.68	80.09	35.73	17.15	34.12
+ $0 \nearrow 1$ linear	13.28	6.66	59.28	80.00	36.43	**18.27**	35.65
+ $0 \nearrow 1$ cosine	**14.66**	**8.84**	60.56	79.59	35.08	18.07	**36.13**
Qwen2.5-Math-1.5B	23.69	14.61	70.94	74.37	27.15	19.45	38.37
+ GRPO ($\alpha = 0.5$)	28.74	16.73	79.42	86.81	40.37	21.04	45.52
+ Dr.GRPO ($\alpha = 1$)	29.40	19.27	79.68	**87.94**	**41.48**	20.89	**46.45**
+ $\alpha = 0$	**29.93**	18.35	78.53	86.68	40.54	21.04	45.85
+ $\alpha = 0.25$	27.96	20.10	79.77	87.56	40.44	**22.07**	46.32
+ $0 \nearrow 1$ linear	27.66	18.65	77.72	86.27	38.30	21.34	44.99
+ $0 \nearrow 1$ cosine	28.15	**20.12**	**81.06**	85.16	39.86	21.51	45.98

qualitatively compatible with our theoretical characterization: the weight $\omega_{G,\alpha}(\rho)$ acts on the pass probabilities of the current policy, so changing the policy initialization changes which prompts fall into different regions of the shaping profile.

We do not interpret these experiments as evidence that a particular fixed α or annealing schedule is universally superior to GRPO or Dr. GRPO. Differences between trained variants are often small, and each configuration was trained with a single seed. Moreover, our experiments use AdamW rather than plain stochastic gradient ascent, and keep the nominal learning rate fixed across shaping rules. The experiments consequently combine the relative prompt-allocation effects isolated by the normalized theory with the α-dependent global scale of the raw gradient estimator, after which AdamW further transforms the update through momentum and adaptive preconditioning.

A more extensive empirical study could separately control for the global gradient scale, evaluate several training seeds and larger model families, and directly track the evolution of prompt pass probabilities during training. Our present experiments are instead intended as an initial controlled demonstration that the prompt-allocation degree of freedom studied theoretically can have measurable, but setting-dependent, consequences in realistic RLVR training.

C Proof of Proposition 2

We first derive the result for a general shaping function $f : (0, 1) \to \mathbb{R}$, and then specialize to $f = f_\alpha$. For collapsed groups, the advantage is set to zero, which is consistent because $R_i - \widehat{\rho} = 0$; hence the endpoint values $f(0)$ and $f(1)$ are not needed.

For $a, b > 0$, recall that the beta function is

$$B(a, b) = \int_0^1 u^{a-1}(1 - u)^{b-1}\, du.$$

If a and b are positive integers, then

$$B(a, b) = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!}.$$

The incomplete and regularized incomplete beta functions are

$$B_\rho(a,b) = \int_0^\rho u^{a-1}(1-u)^{b-1}\,du, \qquad I_\rho(a,b) = \frac{B_\rho(a,b)}{B(a,b)}.$$

For $G \geq 2$, define the finite-group transformation

$$F_G(\rho) := \frac{1}{G}\sum_{s=0}^{G-2} f\Big(\frac{s+1}{G}\Big) I_\rho(s+1, G-1-s), \qquad \rho \in [0,1]. \tag{12}$$

Since f is evaluated only at the interior grid points $\frac{1}{G}, \ldots, \frac{G-1}{G}$, this expression is well-defined without specifying $f(0)$ or $f(1)$.

Derivative representation. By the fundamental theorem of calculus,

$$\frac{d}{d\rho} I_\rho(a,b) = \frac{\rho^{a-1}(1-\rho)^{b-1}}{B(a,b)}.$$

For $a = s+1$ and $b = G-1-s$, the integer-argument formula gives

$$B(s+1, G-1-s) = \frac{s!(G-2-s)!}{(G-1)!},$$

and therefore

$$\frac{1}{B(s+1, G-1-s)} = (G-1)\binom{G-2}{s}.$$

Differentiating Equation 12 term by term yields

$$\begin{aligned} F_G'(\rho) &= \frac{1}{G}\sum_{s=0}^{G-2} f\left(\frac{s+1}{G}\right) \frac{d}{d\rho} I_\rho(s+1, G-1-s) \\ &= \frac{G-1}{G}\sum_{s=0}^{G-2} f\left(\frac{s+1}{G}\right) \binom{G-2}{s} \rho^s (1-\rho)^{G-2-s}. \end{aligned} \tag{13}$$

Thus F_G' is a degree-$(G-2)$ Bernstein polynomial whose coefficients are values of f on the interior grid $\left\{\frac{1}{G}, \ldots, \frac{G-1}{G}\right\}$.

Binomial representation. Let $X \sim \mathrm{Bin}(G, \rho)$ with $\rho \in (0,1)$. Then

$$\begin{aligned} &\mathbb{E}_X\Big[\mathbb{1}_{\{1\leq X\leq G-1\}} f\Big(\frac{X}{G}\Big) \frac{(X/G)(1-X/G)}{\rho(1-\rho)}\Big] \\ &\quad = \frac{1}{\rho(1-\rho)}\sum_{x=1}^{G-1} f\Big(\frac{x}{G}\Big) \frac{x(G-x)}{G^2} \binom{G}{x} \rho^x (1-\rho)^{G-x}. \end{aligned} \tag{14}$$

Using

$$x(G-x)\binom{G}{x} = G(G-1)\binom{G-2}{x-1}$$

and substituting $s = x-1$, we obtain

$$\begin{aligned} &\mathbb{E}_X\left[\mathbb{1}_{\{1\leq X\leq G-1\}} f\left(\frac{X}{G}\right) \frac{(X/G)(1-X/G)}{\rho(1-\rho)}\right] \\ &\quad = \frac{G-1}{G}\sum_{s=0}^{G-2} f\left(\frac{s+1}{G}\right) \binom{G-2}{s} \rho^s (1-\rho)^{G-2-s}. \end{aligned}$$

Comparing with Equation 13 gives

$$F_G'(\rho) = \mathbb{E}_X\Big[\mathbb{1}_{\{1\leq X\leq G-1\}} f\Big(\frac{X}{G}\Big) \frac{(X/G)(1-X/G)}{\rho(1-\rho)}\Big], \qquad X \sim \mathrm{Bin}(G, \rho). \tag{15}$$

Proof of Proposition 2. We first prove the surrogate-gradient identity for a general shaping function f, and then specialize to f_α. Fix a prompt $q \in \mathcal{Q}$, and write $\rho := \rho_q(\theta)$, $R_i := r(q, O_i)$, $S_i := \nabla_\theta \log \pi_\theta(O_i \mid q)$, where $O_1, \dots, O_G \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid q)$. Let $\widehat{\rho} := \frac{1}{G}\sum_{i=1}^{G} R_i$. For a general shaping function $f : (0,1) \to \mathbb{R}$, define

$$\widehat{A}_i^f := \begin{cases} f(\widehat{\rho})(R_i - \widehat{\rho}), & \widehat{\rho} \in (0,1), \\ 0, & \widehat{\rho} \in \{0,1\}, \end{cases}$$

and the corresponding fixed-prompt estimator

$$\widehat{g}_G^f(q,\theta) := \frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i^f S_i.$$

If $\rho \in \{0,1\}$, all sampled rewards are almost surely identical, hence $\widehat{\rho} = \rho$ and $\widehat{g}_G^f(q,\theta) = 0$ almost surely. Moreover, since the differentiable function $\rho_q : \mathbb{R}^d \to [0,1]$ attains a global minimum or maximum at θ, $\nabla_\theta \rho_q(\theta) = 0$. Consequently,

$$\nabla_\theta F_G(\rho_q(\theta)) = F_G'(\rho_q(\theta))\nabla_\theta \rho_q(\theta) = 0.$$

Thus the desired identity holds at the endpoints. In the remainder, assume $\rho \in (0,1)$.

The score-function identities give $\mathbb{E}[S_i] = 0$ and $\mathbb{E}[R_i S_i] = \nabla_\theta \rho_q(\theta)$. Since $R_i \in \{0,1\}$,

$$\mathbb{E}[S_i \mid R_i = 1] = \frac{1}{\rho}\nabla_\theta \rho_q(\theta),$$

while, using $\mathbb{E}[S_i] = 0$,

$$\mathbb{E}[S_i \mid R_i = 0] = -\frac{1}{1-\rho}\nabla_\theta \rho_q(\theta).$$

Equivalently,

$$\mathbb{E}[S_i \mid R_i] = \left(\frac{R_i}{\rho} - \frac{1-R_i}{1-\rho}\right)\nabla_\theta \rho_q(\theta). \tag{16}$$

Let $X := \sum_{i=1}^{G} R_i$, so that $\widehat{\rho} = X/G$. Conditioning on the complete reward vector $R_{1:G} := (R_1, \dots, R_G)$, and using the conditional independence of S_i from the other rewards given R_i, we obtain

$$\begin{aligned} \mathbb{E}\left[\widehat{g}_G^f(q,\theta) \,\middle|\, R_{1:G}\right] &= \mathbb{1}_{\{0<\widehat{\rho}<1\}} \frac{f(\widehat{\rho})}{G}\sum_{i=1}^{G}(R_i - \widehat{\rho})\mathbb{E}[S_i \mid R_i] \\ &= \mathbb{1}_{\{0<\widehat{\rho}<1\}} \frac{f(\widehat{\rho})}{G}\sum_{i=1}^{G}(R_i - \widehat{\rho})\left(\frac{R_i}{\rho} - \frac{1-R_i}{1-\rho}\right)\nabla_\theta \rho_q(\theta). \end{aligned}$$

For $R_i \in \{0,1\}$,

$$(R_i - \widehat{\rho})\Big(\frac{R_i}{\rho} - \frac{1-R_i}{1-\rho}\Big) = \frac{1-\widehat{\rho}}{\rho}\,\mathbb{1}_{\{R_i=1\}} + \frac{\widehat{\rho}}{1-\rho}\,\mathbb{1}_{\{R_i=0\}}.$$

Since there are $G\widehat{\rho}$ successful and $G(1-\widehat{\rho})$ unsuccessful responses,

$$\frac{1}{G}\sum_{i=1}^{G}(R_i - \widehat{\rho})\Big(\frac{R_i}{\rho} - \frac{1-R_i}{1-\rho}\Big) = \widehat{\rho}(1-\widehat{\rho})\Big(\frac{1}{\rho} + \frac{1}{1-\rho}\Big) = \frac{\widehat{\rho}(1-\widehat{\rho})}{\rho(1-\rho)}.$$

Hence

$$\mathbb{E}\Big[\widehat{g}_G^f(q,\theta) \mid |R_{1:G}\Big] = \mathbb{1}_{\{0<\widehat{\rho}<1\}} f(\widehat{\rho})\frac{\widehat{\rho}(1-\widehat{\rho})}{\rho(1-\rho)}\nabla_\theta \rho_q(\theta). \tag{17}$$

Taking expectation over the rewards, with

$$X = G\widehat{\rho} \sim \mathrm{Bin}(G, \rho),$$

and applying Equation 15, gives

$$\begin{aligned}\mathbb{E}\big[\widehat{g}^f_G(q,\theta)\big] &= \mathbb{E}_{X\sim\mathrm{Bin}(G,\rho)}\Big[\mathbb{1}_{\{1\le X\le G-1\}} f\Big(\frac{X}{G}\Big)\frac{(X/G)(1-X/G)}{\rho(1-\rho)}\Big]\nabla_\theta\rho_q(\theta)\\ &= F'_G(\rho_q(\theta))\nabla_\theta\rho_q(\theta)\\ &= \nabla_\theta F_G(\rho_q(\theta)).\end{aligned}$$

The full batch estimator is the average of the fixed-prompt estimators,

$$\widehat{g}^f_{B,G}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{g}^f_G(Q_b,\theta).$$

Since the prompt groups are identically distributed,

$$\mathbb{E}\big[\widehat{g}^f_{B,G}(\theta)\big] = \mathbb{E}_Q[\nabla_\theta F_G(\rho_Q(\theta))] = \nabla_\theta\mathbb{E}_Q[F_G(\rho_Q(\theta))],$$

where the interchange of differentiation and expectation is immediate because $\mathcal{Q}$ is finite. Thus, defining

$$J_{F_G}(\theta) := \mathbb{E}_Q[F_G(\rho_Q(\theta))],$$

we obtain

$$\mathbb{E}[\widehat{g}^f_{B,G}(\theta)] = \nabla_\theta J_{F_G}(\theta).$$

It remains to specialize the construction to α-GRPO. For $\alpha\in[0,1]$, let

$$f_\alpha(\rho) = \big(\rho(1-\rho)\big)^{\alpha-1}, \qquad \rho\in(0,1).$$

Substituting $f=f_\alpha$ into Equation 12 gives

$$\begin{aligned}F_G(\rho) &= \frac{1}{G}\sum_{s=0}^{G-2} f_\alpha\Big(\frac{s+1}{G}\Big) I_\rho(s+1,G-1-s)\\ &= \frac{1}{G}\sum_{s=0}^{G-2}\Big[\frac{s+1}{G}\Big(1-\frac{s+1}{G}\Big)\Big]^{\alpha-1} I_\rho(s+1,G-1-s).\end{aligned}$$

Hence $F_G = F_{G,\alpha}$ with $F_{G,\alpha}$ as defined in Equation 8. Similarly, Equation 13 yields

$$F'_{G,\alpha}(\rho) = \frac{G-1}{G}\sum_{s=0}^{G-2}\Big[\frac{s+1}{G}\Big(1-\frac{s+1}{G}\Big)\Big]^{\alpha-1}\binom{G-2}{s}\rho^s(1-\rho)^{G-2-s},$$

which is Equation 9. Therefore,

$$g^\alpha_G(\theta) = \nabla_\theta\mathbb{E}_Q\left[F_{G,\alpha}(\rho_Q(\theta))\right] = \nabla_\theta J_{G,\alpha}(\theta),$$

which proves the proposition. □

D Additional details for Section 2.2

Proof of Proposition 4. Fix $G\ge 4$ and $\alpha\in(0,1)$, and write

$$n := G-2, \qquad b_{s,n}(\rho) := \binom{n}{s}\rho^s(1-\rho)^{n-s}.$$

By Theorem 2 and the normalization of $F_{G,\alpha}$,

$$\omega_{G,\alpha}(\rho) = \frac{F'_{G,\alpha}(\rho)}{C_{G,\alpha}} = \kappa_{G,\alpha}\sum_{s=0}^{n} a_s b_{s,n}(\rho),$$

where

$$a_s := \Big[\frac{s+1}{G}\Big(1-\frac{s+1}{G}\Big)\Big]^{\alpha-1}, \qquad \kappa_{G,\alpha} := \frac{G-1}{GC_{G,\alpha}} > 0.$$

The coefficients are symmetric:

$$a_{n-s} = a_s.$$

Moreover, because $\alpha - 1 < 0$ and $x(1-x)$ is strictly increasing on $(0, 1/2)$, the sequence $(a_s)_{s=0}^{n}$ strictly decreases toward the center and then increases symmetrically.

We first show that $\omega_{G,\alpha}$ is strictly decreasing on $(0, 1/2)$. The Bernstein derivative identity gives

$$\omega'_{G,\alpha}(\rho) = \kappa_{G,\alpha} n \sum_{s=0}^{n-1} (a_{s+1} - a_s) b_{s,n-1}(\rho).$$

Set $\Delta_s := a_{s+1} - a_s$. Symmetry of (a_s) implies

$$\Delta_{n-1-s} = -\Delta_s.$$

Pairing the terms indexed by s and $n-1-s$ therefore gives

$$\frac{\omega'_{G,\alpha}(\rho)}{\kappa_{G,\alpha} n} = \sum_{s<(n-1)/2} \Delta_s \big(b_{s,n-1}(\rho) - b_{n-1-s,n-1}(\rho)\big),$$

where, if $n-1$ is even, the unpaired central term vanishes because $\Delta_{(n-1)/2} = 0$.

For $s < (n-1)/2$, we have $\Delta_s < 0$. Furthermore, for $\rho \in (0, 1/2)$,

$$\frac{b_{s,n-1}(\rho)}{b_{n-1-s,n-1}(\rho)} = \Big(\frac{\rho}{1-\rho}\Big)^{2s-(n-1)} > 1,$$

because $\rho/(1-\rho) < 1$ and $2s - (n-1) < 0$. Every paired term is therefore strictly negative, and hence

$$\omega'_{G,\alpha}(\rho) < 0, \qquad \rho \in (0, 1/2).$$

By symmetry,

$$\omega_{G,\alpha}(1-\rho) = \omega_{G,\alpha}(\rho),$$

so $\omega_{G,\alpha}$ is strictly increasing on $(1/2, 1)$. Thus it is strictly U-shaped.

The normalization of $\widetilde{F}_{G,\alpha}$ gives

$$\int_0^1 \omega_{G,\alpha}(\rho)\, d\rho = \widetilde{F}_{G,\alpha}(1) - \widetilde{F}_{G,\alpha}(0) = 1.$$

Since $\omega_{G,\alpha}$ is nonconstant and strictly decreasing on $(0, 1/2)$, this implies

$$\omega_{G,\alpha}(0) > 1, \qquad \omega_{G,\alpha}(1/2) < 1.$$

Indeed, if $\omega_{G,\alpha}(0) \le 1$, then $\omega_{G,\alpha} \le 1$ everywhere, with a strict inequality on a set of positive measure, contradicting that its integral is one. The argument for $\omega_{G,\alpha}(1/2) < 1$ is analogous.

Continuity and strict monotonicity now imply that there is a unique $r_{G,\alpha} \in (0, 1/2)$ satisfying

$$\omega_{G,\alpha}(r_{G,\alpha}) = 1.$$

By symmetry,

$$\omega_{G,\alpha}(1 - r_{G,\alpha}) = 1,$$

and the strict U-shape yields

$$\omega_{G,\alpha}(\rho) > 1 \quad \Longleftrightarrow \quad \rho < r_{G,\alpha} \text{ or } \rho > 1 - r_{G,\alpha}.$$

It remains to establish the polynomial and asymptotic claims. Because $\omega_{G,\alpha}$ is a polynomial of degree at most $G-2$ and satisfies

$$\omega_{G,\alpha}(1-\rho) = \omega_{G,\alpha}(\rho),$$

there exists a polynomial $P_{G,\alpha}$ of degree at most $\lfloor (G-2)/2 \rfloor$ such that

$$\omega_{G,\alpha}(\rho) - 1 = P_{G,\alpha}\big(\rho(1-\rho)\big).$$

To see this, write $\rho = 1/2 + z$. Symmetry makes the resulting polynomial even in z, and hence a polynomial in

$$z^2 = \frac{1}{4} - \rho(1-\rho).$$

Consequently, $u_{G,\alpha} := r_{G,\alpha}(1 - r_{G,\alpha})$ is the unique root in $(0, 1/4)$ of a polynomial equation of degree at most $\lfloor (G-2)/2 \rfloor$, and

$$r_{G,\alpha} = \frac{1}{2}\left(1 - \sqrt{1 - 4u_{G,\alpha}}\right).$$

Finally, for every compact $K \subset (0,1)$, the Bernstein representation in Theorem 2 gives the locally uniform limit

$$F'_{G,\alpha}(\rho) \longrightarrow [\rho(1-\rho)]^{\alpha-1}, \qquad \rho \in K.$$

Moreover,

$$C_{G,\alpha} = \frac{1}{G}\sum_{k=1}^{G-1}\Big[\frac{k}{G}\Big(1 - \frac{k}{G}\Big)\Big]^{\alpha-1} \longrightarrow \int_0^1 [u(1-u)]^{\alpha-1}\,du = B(\alpha,\alpha).$$

Therefore, locally uniformly on $(0,1)$,

$$\omega_{G,\alpha}(\rho) \longrightarrow \omega_{\infty,\alpha}(\rho) := \frac{[\rho(1-\rho)]^{\alpha-1}}{B(\alpha,\alpha)}.$$

The limiting function is strictly decreasing on $(0, 1/2)$. It has a unique crossing of level one there, since

$$B(\alpha,\alpha) = \int_0^1 [u(1-u)]^{\alpha-1}\,du > 4^{1-\alpha},$$

and hence $\omega_{\infty,\alpha}(1/2) < 1$, while $\omega_{\infty,\alpha}(\rho) \to \infty$ as $\rho \downarrow 0$. Solving $\omega_{\infty,\alpha}(\rho) = 1$ gives

$$\rho(1-\rho) = B(\alpha,\alpha)^{-1/(1-\alpha)}.$$

Thus its left crossing is

$$r_{\infty,\alpha} = \frac{1}{2}\Big(1 - \sqrt{1 - 4B(\alpha,\alpha)^{-1/(1-\alpha)}}\Big).$$

To conclude, fix $\varepsilon > 0$ sufficiently small. Strict monotonicity of $\omega_{\infty,\alpha}$ gives

$$\omega_{\infty,\alpha}(r_{\infty,\alpha} - \varepsilon) > 1, \qquad \omega_{\infty,\alpha}(r_{\infty,\alpha} + \varepsilon) < 1.$$

Local uniform convergence implies that the same inequalities hold for $\omega_{G,\alpha}$ for all sufficiently large G. By uniqueness of its crossing,

$$r_{\infty,\alpha} - \varepsilon < r_{G,\alpha} < r_{\infty,\alpha} + \varepsilon.$$

Therefore,

$$\lim_{G\to\infty} r_{G,\alpha} = \frac{1}{2}\Big(1 - \sqrt{1 - 4B(\alpha,\alpha)^{-1/(1-\alpha)}}\Big).$$

□

Example 10 (Explicit threshold value for $G = 4$)**.** For $G = 4$, the prompt-allocation potential consists of a single universal (in α) cubic. Define $t_\alpha := \left(\frac{3}{4}\right)^{1-\alpha}$ and $\kappa_{4,\alpha} := \frac{1-t_\alpha}{2+t_\alpha}$, then

$$\mathcal{R}_{4,\alpha}(\rho) = \kappa_{4,\alpha}\rho(1-\rho)(1-2\rho).$$

Standard GRPO corresponds to $\kappa_{4,1/2} = \frac{1-\sqrt{3}/2}{2+\sqrt{3}/2}$. Differentiating gives $\omega_{4,\alpha}(\rho) = 1 + \kappa_{4,\alpha}\big(1 - 6\rho(1-\rho)\big)$. It follows that prompt amplification $\omega_{G,\alpha}(\rho) > 1$ occurs for batch prompts if and only if

$$\rho < \frac{1}{2}\left(1 - \frac{1}{\sqrt{3}}\right) \approx 0.2113 \quad \text{or} \quad \rho > \frac{1}{2}\left(1 + \frac{1}{\sqrt{3}}\right) \approx 0.7887.$$

Since the α-dependence for $G = 4$ is multiplicative, for $G = 4$ the same prompts are amplified.

Example 11 (Explicit promotion threshold for $G = 6$)**.** Let

$$s := 1 - \alpha > 0, \qquad a := 5^{-s}, \qquad b := 8^{-s}, \qquad c := 9^{-s},$$

and put $\overline{a} := \frac{2a+2b+c}{5}$. Writing $u = \rho(1-\rho)$, the marginal prompt weight is

$$\omega_{6,\alpha}(\rho) = \frac{a + 4(b-a)u + (2a - 8b + 6c)u^2}{\overline{a}}.$$

Consequently, the equation $\omega_{6,\alpha}(\rho) = 1$ is equivalent to $Au^2 + Bu + C = 0$, where

$$A := 2a - 8b + 6c, \qquad B := 4(b-a), \qquad C := a - \overline{a} = \frac{3a - 2b - c}{5}.$$

Its unique root in $(0, 1/4)$ is

$$u_{6,\alpha} = \frac{-B - \sqrt{B^2 - 4AC}}{2A}.$$

Hence the left promotion threshold is explicitly

$$r_{6,\alpha} = \frac{1}{2}\left(1 - \sqrt{1 - 4u_{6,\alpha}}\right).$$

E Dynamic sampling as state-dependent rescaling

We isolate the dynamic sampling component of DAPO. Specifically, we consider the setting in which fresh prompt–response groups are repeatedly proposed and groups with homogeneous rewards are discarded. We show that, for an otherwise unchanged group estimator, this filtering changes the population update only by a state-dependent scalar factor.

Let

$$Q \sim \mathbb{P}_Q, \qquad O_1, \ldots, O_G \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid Q), \qquad R_i := r(Q, O_i),$$

and define the mixed-group event

$$A := \Big\{ 0 < \sum_{i=1}^{G} R_i < G \Big\}.$$

By Proposition 1,

$$\mathbb{P}_\theta(A \mid Q = q) = 1 - \rho_q(\theta)^G - (1 - \rho_q(\theta))^G =: p_{\text{mix}}(\rho_q(\theta)).$$

Hence, the population acceptance probability is given by

$$M_G(\theta) := \mathbb{P}_\theta(A) = \mathbb{E}_Q[p_{\text{mix}}(\rho_Q(\theta))].$$

Whenever $M_G(\theta) > 0$, Bayes' rule gives the prompt distribution conditional on acceptance:

$$\mathbb{P}_\theta^{\text{DS}}(q) := \mathbb{P}_\theta(Q = q \mid A) = \frac{p_{\text{mix}}(\rho_q(\theta))}{M_G(\theta)}\, \mathbb{P}_Q(q)\,. \tag{18}$$

Thus, conditioning on mixed groups changes the prompt distribution toward prompts with larger mixed-group probability.

Proposition 12 (Dynamic sampling rescaling)**.** *Given $\theta \in \Theta$ let $H_\theta(q, o_{1:G})$ denote a deterministic group-level update satisfying $H_\theta(q, o_{1:G}) = 0$ for every homogeneous group. Assume that dynamic sampling repeatedly draws independent fresh prompt–group proposals until the mixed-group event A occurs. Define*

$$g(\theta) := \mathbb{E}_{Q, O_{1:G}}\left[H_\theta(Q, O_{1:G})\right], \qquad g^{\text{DS}}(\theta) := \mathbb{E}_{Q, O_{1:G}}\left[H_\theta(Q, O_{1:G}) \mid A\right].$$

Then, whenever $M_G(\theta) := \mathbb{P}_\theta(A) > 0$, we have

$$g^{\text{DS}}(\theta) = \frac{1}{M_G(\theta)}\, g(\theta).$$

In particular, taking H_θ to be the α-GRPO group update, with homogeneous groups defined to contribute zero, gives

$$g_{G,\alpha}^{\text{DS}}(\theta) = \frac{1}{M_G(\theta)}\, g_{G,\alpha}(\theta).$$

Proof. Since the group-level update vanishes on homogeneous groups, we have

$$H_\theta(Q, O_{1:G}) = H_\theta(Q, O_{1:G}) \mathbb{1}_A,$$

implying that,

$$g(\theta) = \mathbb{E}_{Q,O_{1:G}}[H_\theta(Q, O_{1:G})\mathbb{1}_A] = \mathbb{P}_\theta(A)\, \mathbb{E}_{Q,O_{1:G}}[H_\theta(Q, O_{1:G}) \mid A] = M_G(\theta)\, g^{\mathrm{DS}}(\theta)\,.$$

Since $M_G(\theta) > 0$, rearranging yields the claim,

$$g^{\mathrm{DS}}(\theta) = \frac{1}{M_G(\theta)}\, g(\theta)\,.$$

□

The Bayes reweighting from Equation 18 might suggest that dynamic sampling changes the relative contributions of prompts. Proposition 12 shows that this is not the case at the population-update level. Indeed, for each prompt with positive mixed-group probability, conditioning on acceptance amplifies its expected group contribution by the factor $p_{\mathrm{mix}}(\rho_q(\theta))^{-1}$. This exactly cancels the prompt-dependent $p_{\mathrm{mix}}(\rho_q(\theta))$ factor introduced by the change of prompt measure. What remains is only the global rescaling $1/M_G(\theta)$.

Proposition 12 shows that this change of measure has a particularly simple effect on the population update. Although conditioning on A changes the prompt marginal from $\mathbb{P}_Q$ to $\mathbb{P}_\theta^{\mathrm{DS}}$, the resulting expected update differs from the unfiltered population update only by the scalar factor $1/M_G(\theta)$. This can also be seen directly at the prompt level. For every q with $p_{\mathrm{mix}}(\rho_q(\theta)) > 0$,

$$\mathbb{E}_{O_{1:G}}[H_\theta(Q, O_{1:G}) \mid Q = q, A] = \frac{\mathbb{E}_{O_{1:G}}[H_\theta(Q, O_{1:G}) \mid Q = q]}{p_{\mathrm{mix}}(\rho_q(\theta))},$$

because the group update vanishes outside A. Combining this with the Bayes reweighting from Equation 18 shows that the two prompt-dependent factors cancel:

$$\mathbb{P}_\theta^{\mathrm{DS}}(q)\, \mathbb{E}_{O_{1:G}}[H_\theta(Q, O_{1:G}) \mid Q = q, A] = \frac{\mathbb{P}_Q(q)}{M_G(\theta)} \mathbb{E}_{O_{1:G}}[H_\theta(Q, O_{1:G}) \mid Q = q].$$

Hence dynamic sampling changes the measure over accepted prompts, but this measure change does not alter the population update direction. It only induces the state-dependent rescaling

$$g^{\mathrm{DS}}(\theta) = \frac{1}{M_G(\theta)} g(\theta).$$

Consequently, whenever $M_G(\theta) > 0$, dynamic sampling preserves the population update direction and its stationary points. For a nominal stepsize η, it corresponds at the mean-field level to the state-dependent effective stepsize

$$\eta_{\mathrm{eff}}(\theta) = \frac{\eta}{M_G(\theta)}.$$

This connection is per optimizer step and does not imply equal sampling cost or variance. The argument assumes that rejected groups are replaced by fresh prompt–group proposals. If instead a prompt is sampled once and only its responses are repeatedly resampled until acceptance, the prompt marginal remains $\mathbb{P}_Q$ and the measure-change interpretation above no longer applies. Finally, it is important to note that the result isolates the dynamic-filtering component only and does not establish an equivalence between full DAPO and standard GRPO.

F Convergence Analysis Details

This appendix provides the technical details for the convergence analysis in Section 3. We first establish regularity properties of the finite-group transformation $F_{G,\alpha}$, the prompt success probability $\rho_q(\theta)$, and its gradient. These results yield a smoothness bound for the normalized finite-group objective $\widetilde{J}_{G,\alpha}$ that is uniform over group size G and shaping parameter α, together with a uniform variance bound for the normalized stochastic α-GRPO estimator. We then prove the finite-group

stochastic-ascent result underlying Theorem 8, which gives a stationarity bound for the normalized surrogate objective $\widetilde{J}_{G,\alpha}$ with constants independent of G and α.

Finally, we analyze annealed α-GRPO relative to the original mean-reward objective J. We bound the bias between the expected finite-group update and the scaled mean-reward gradient $c_G \nabla J$, show that this bias vanishes as $\alpha \to 1$, and use a biased stochastic-gradient argument to prove Theorem 9. We additionally give a variable-stepsize extension showing vanishing mean-reward stationarity under Robbins–Monro stepsizes and the corresponding stepsize-weighted bias condition.

Throughout, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, while $\|\cdot\|_\infty$ denotes the maximum norm on $\mathbb{R}^d$. Supremum norms over prompts and outputs are written explicitly.

F.1 Preliminary results

Recall from Proposition 2 that we use the canonical representative

$$F_{G,\alpha}(\rho) = \frac{1}{G} \sum_{s=0}^{G-2} \Big[\frac{s+1}{G} \Big(1 - \frac{s+1}{G} \Big) \Big]^{\alpha-1} I_\rho(s+1, G-1-s),$$

which satisfies $F_{G,\alpha}(0) = 0$. Its derivative is $f_{G,\alpha}(\rho) := F'_{G,\alpha}(\rho)$. Moreover, recall that

$$C_{G,\alpha} := F_{G,\alpha}(1) - F_{G,\alpha}(0) = F_{G,\alpha}(1) = \frac{1}{G} \sum_{k=1}^{G-1} \Big[\frac{k}{G} \big(1 - \frac{k}{G} \big) \Big]^{\alpha-1}.$$

We further recall the normalized transformation

$$\widetilde{F}_{G,\alpha}(\rho) := \frac{F_{G,\alpha}(\rho)}{C_{G,\alpha}},$$

and its marginal weight

$$\omega_{G,\alpha}(\rho) := \widetilde{F}'_{G,\alpha}(\rho) = \frac{f_{G,\alpha}(\rho)}{C_{G,\alpha}}.$$

Since $f_{G,\alpha} \geq 0$, the map $F_{G,\alpha}$ is nondecreasing. Therefore, we have

$$0 \leq F_{G,\alpha}(\rho) \leq C_{G,\alpha}, \qquad \rho \in [0,1]$$

or, equivalently, $0 \leq \widetilde{F}_{G,\alpha}(\rho) \leq 1$. This implies that

$$0 \leq J_{G,\alpha}(\theta) \leq C_{G,\alpha}, \qquad 0 \leq \widetilde{J}_{G,\alpha}(\theta) \leq 1,$$

and in particular

$$J^\star_{G,\alpha} := \sup_\theta J_{G,\alpha}(\theta) < \infty.$$

Moreover, since $0 < \frac{k}{G}\left(1 - \frac{k}{G}\right) < 1$ and $\alpha - 1 \leq 0$,

$$C_{G,\alpha} \geq C_{G,1} = \frac{G-1}{G} \geq \frac{1}{2}.$$

Weighted regularity of the normalized marginal weights. The key point is that the normalized marginal prompt weights $\omega_{G,\alpha}$ are not uniformly Lipschitz in G and α: for $\alpha < 1$ they develop increasingly steep boundary layers as G grows. However, the policy sensitivity of the pass probability vanishes at precisely these boundaries. The following weighted bounds make this cancellation explicit. Throughout, let

$$m(\rho) := \rho(1-\rho), \qquad d(\rho) := \min\{\rho, 1-\rho\}.$$

Lemma 13. *For every $G \geq 2$, $\alpha \in [0,1]$, and $\rho \in [0,1]$, we have*

$$m(\rho)\omega_{G,\alpha}(\rho) \leq 2 \qquad \textit{and} \qquad m(\rho)^2 \left|\omega'_{G,\alpha}(\rho)\right| \leq 4.$$

Proof. Recall that

$$\omega_{G,\alpha}(\rho) = \frac{f_{G,\alpha}(\rho)}{C_{G,\alpha}}.$$

Let $X \sim \mathrm{Bin}(G, \rho)$ and define

$$h_{G,\alpha}(k) := \begin{cases} \left[\frac{k}{G}\left(1 - \frac{k}{G}\right)\right]^{\alpha}, & k = 1, \ldots, G-1, \\ 0, & k \in \{0, G\}. \end{cases}$$

At $\alpha = 0$, the interior values are understood as $h_{G,0}(k) = 1$, consistent with the convention that homogeneous groups contribute zero.

By Proposition 2 we have, for $\rho \in (0, 1)$

$$\begin{aligned} \rho(1-\rho) f_{G,\alpha}(\rho) &= \sum_{k=1}^{G-1} \left[\frac{k}{G}\left(1 - \frac{k}{G}\right)\right]^{\alpha-1} \frac{G-1}{G} \binom{G-2}{k-1} \rho^k (1-\rho)^{G-k} \\ &= \sum_{k=1}^{G} \left[\frac{k}{G}\left(1 - \frac{k}{G}\right)\right]^{\alpha} \binom{G}{k} \rho^k (1-\rho)^{G-k} \\ &= \mathbb{E}[h_{G,\alpha}(X)] =: N_{G,\alpha}(\rho)\,, \end{aligned}$$

where we have used

$$\frac{G-1}{G}\binom{G-2}{k-1} = \frac{(G-1)!}{G(k-1)!(G-k-1)!} = \frac{G!}{k!(G-k)!}\frac{k(G-k)}{G^2} = \left[\frac{k}{G}\left(1 - \frac{k}{G}\right)\right]\binom{G}{k}.$$

Since $0 \le h_{G,\alpha}(k) \le 1$, we have

$$0 \le N_{G,\alpha}(\rho) \le 1\,, \tag{19}$$

and together with $C_{G,\alpha} \ge \frac{1}{2}$, we obtain

$$m(\rho)\omega_{G,\alpha}(\rho) = \frac{N_{G,\alpha}(\rho)}{C_{G,\alpha}} \le 2\,.$$

It remains to control the derivative. Write

$$b_{k,n}(\rho) := \binom{n}{k} \rho^k (1-\rho)^{n-k}$$

for the Bernstein basis, then

$$N_{G,\alpha}(\rho) = \sum_{k=0}^{G} h_{G,\alpha}(k) b_{k,G}(\rho),$$

and the Bernstein derivative identity gives

$$N'_{G,\alpha}(\rho) = G \sum_{k=0}^{G-1} \left(h_{G,\alpha}(k+1) - h_{G,\alpha}(k)\right) b_{k,G-1}(\rho).$$

Multiplying by $m(\rho) = \rho(1-\rho)$ and using the identity

$$G m(\rho) b_{k,G-1}(\rho) = \frac{(k+1)(G-k)}{G+1} b_{k+1,G+1}(\rho)$$

gives

$$m(\rho) N'_{G,\alpha}(\rho) = \sum_{k=0}^{G-1} \gamma_k \left(h_{G,\alpha}(k+1) - h_{G,\alpha}(k)\right) b_{k+1,G+1}(\rho),$$

where

$$\gamma_k := \frac{(k+1)(G-k)}{G+1}. \tag{20}$$

For $\alpha \in (0, 1]$, define

$$\phi_\alpha(x) := [x(1-x)]^\alpha, \qquad x \in [0, 1].$$

The function ϕ_α is nonnegative, symmetric about $1/2$, increasing on $[0, 1/2]$, and concave on $[0, 1]$. Indeed, $x \mapsto x(1-x)$ is nonnegative and concave, while $x \mapsto x^\alpha$ is nondecreasing and concave on $\mathbb{R}_+$.

Consider first a grid interval contained in the increasing half, i.e. $(k+1)/G \le 1/2$. Since

$$\frac{k}{G} = \frac{k}{k+1}\frac{k+1}{G} + \frac{1}{k+1}\,0,$$

concavity and $\phi_\alpha(0) = 0$ imply

$$h_{G,\alpha}(k) = \phi_\alpha(k/G) \ge \frac{k}{k+1}\phi_\alpha((k+1)/G) = \frac{k}{k+1}h_{G,\alpha}(k+1).$$

Since ϕ_α is increasing on this interval,

$$0 \le h_{G,\alpha}(k+1) - h_{G,\alpha}(k) \le \frac{h_{G,\alpha}(k+1)}{k+1}.$$

Consequently,

$$\frac{(k+1)(G-k)}{G+1}|h_{G,\alpha}(k+1) - h_{G,\alpha}(k)| \le \frac{G-k}{G+1}h_{G,\alpha}(k+1) \le 1.$$

For grid intervals in the decreasing half, symmetry gives

$$|h_{G,\alpha}(k+1) - h_{G,\alpha}(k)| \le \frac{h_{G,\alpha}(k)}{G-k},$$

and therefore

$$\frac{(k+1)(G-k)}{G+1}|h_{G,\alpha}(k+1) - h_{G,\alpha}(k)| \le \frac{k+1}{G+1}h_{G,\alpha}(k) \le 1.$$

If G is odd, the central pair has equal values by symmetry, so its difference is zero.

Consequently,

$$\gamma_k\,|h_{G,\alpha}(k+1) - h_{G,\alpha}(k)| \le \frac{G-k}{G+1}h_{G,\alpha}(k+1) \le 1.$$

The decreasing half follows by symmetry. At $\alpha = 0$, the coefficient sequence is

$$0, 1, \ldots, 1, 0,$$

so only the two endpoint differences are nonzero, and the same bound holds directly. Therefore,

$$\gamma_k\,|h_{G,\alpha}(k+1) - h_{G,\alpha}(k)| \le 1 \qquad \text{for all } k.$$

Since the Bernstein basis is nonnegative and

$$\sum_{k=0}^{G-1} b_{k+1,G+1}(\rho) \le 1,$$

we obtain

$$\left|m(\rho)N'_{G,\alpha}(\rho)\right| \le 1. \tag{21}$$

Recall that

$$f_{G,\alpha}(\rho) = \frac{N_{G,\alpha}(\rho)}{m(\rho)},$$

which implies

$$m(\rho)^2 f'_{G,\alpha}(\rho) = m(\rho)N'_{G,\alpha}(\rho) - m'(\rho)N_{G,\alpha}(\rho).$$

Using $|m'(\rho)| = |1 - 2\rho| \le 1$, together with Equation 19 and Equation 21, yields

$$m(\rho)^2|f'_{G,\alpha}(\rho)| \le 2.$$

Finally, dividing by $C_{G,\alpha} \ge 1/2$ gives

$$m(\rho)^2|\omega'_{G,\alpha}(\rho)| \le 4.$$

The endpoint cases follow by continuity, since $f_{G,\alpha}$ and $\omega_{G,\alpha}$ are finite-degree polynomials for every fixed G. □

Boundary-adaptive regularity of the pass probability. The preceding lemma allows the normalized marginal weights to become large or steep near $\rho \in \{0, 1\}$. The next result shows that the policy sensitivity of ρ_q vanishes sufficiently fast in exactly these regions.

Lemma 14. *Suppose that Assumption 7 is in place. Then, for every prompt q,*

$$\|\nabla \rho_q(\theta)\| \le \Pi_* d(\rho_q(\theta)) \le 2\Pi_* m(\rho_q(\theta)). \tag{22}$$

Moreover, for any $\theta, \theta' \in \mathbb{R}^d$, let

$$\theta_t := \theta + t(\theta' - \theta), \qquad t \in [0, 1].$$

Then the map $t \mapsto \nabla \rho_q(\theta_t)$ is absolutely continuous and, for almost every $t \in [0, 1]$,

$$\big\| \frac{d}{dt} \nabla \rho_q(\theta_t) \big\| \le 2(L_s + \Pi_*^2) m(\rho_q(\theta_t)) \|\theta' - \theta\|.$$

In particular,

$$\|\nabla \rho_q(\theta) - \nabla \rho_q(\theta')\| \le (L_s + \Pi_*^2) \|\theta - \theta'\|.$$

Proof. Define $S_\theta(q, o) := \nabla_\theta \log \pi_\theta(o \mid q)$ and, for brevity, write

$$\rho := \rho_q(\theta), \qquad R := r(q, O), \qquad S := S_\theta(q, O).$$

The policy-gradient identity gives $\nabla \rho_q(\theta) = \mathbb{E}[RS]$. Since $R \in \{0, 1\}$ and $\|S\| \le \Pi_*$,

$$\|\nabla \rho_q(\theta)\| \le \rho \Pi_*.$$

On the other hand, the score has zero expectation, $\mathbb{E}[S] = 0$, and therefore

$$\nabla \rho_q(\theta) = -\mathbb{E}[(1 - R)S].$$

Hence

$$\|\nabla \rho_q(\theta)\| \le (1 - \rho) \Pi_*.$$

Taking the smaller of the two estimates gives

$$\|\nabla \rho_q(\theta)\| \le \Pi_* \min\{\rho, 1 - \rho\}.$$

Since $d(\rho) \le 2\rho(1 - \rho) = 2m(\rho)$, Equation 22 follows.

We next control the variation of $\nabla \rho_q$ along line segments. Fix $\theta, \theta' \in \mathbb{R}^d$, write

$$\Delta := \theta' - \theta, \qquad \theta_t := \theta + t\Delta,$$

and abbreviate

$$S_t(o) := S_{\theta_t}(q, o), \qquad \pi_t(o) := \pi_{\theta_t}(o \mid q), \qquad \rho_t := \rho_q(\theta_t).$$

Since $S_\theta(q, o)$ is L_s-Lipschitz in θ, the map $t \mapsto S_t(o)$ is $L_s\|\Delta\|$-Lipschitz and therefore absolutely continuous. Hence, for almost every t,

$$\|\dot{S}_t(o)\| \le L_s \|\Delta\|.$$

Moreover, we have $\dot{\pi}_t(o) = \pi_t(o)\langle S_t(o), \Delta \rangle$. Using

$$\nabla \rho_q(\theta_t) = \sum_{o \in \mathcal{O}} r(q, o) \pi_t(o) S_t(o),$$

and the finiteness of $\mathcal{O}$, it follows that $t \mapsto \nabla \rho_q(\theta_t)$ is absolutely continuous. At almost every t,

$$\frac{d}{dt} \nabla \rho_q(\theta_t) = \sum_{o \in \mathcal{O}} r(q, o) \pi_t(o) \left[\dot{S}_t(o) + S_t(o) S_t(o)^\top \Delta \right].$$

Define

$$T_t^\Delta(o) := \dot{S}_t(o) + S_t(o) S_t(o)^\top \Delta.$$

which, by Assumption 7, satisfies

$$\|T_t^\Delta(o)\| \le (L_s + \Pi_*^2) \|\Delta\|.$$

Therefore,

$$\Big\| \frac{d}{dt} \nabla \rho_q(\theta_t) \Big\| \le \rho_t (L_s + \Pi_*^2) \|\Delta\|.$$

On the other hand, the zero-score identity gives

$$\sum_{o \in \mathcal{O}} \pi_t(o) S_t(o) = 0.$$

Differentiating this absolutely continuous identity along the line segment yields, for almost every t,

$$\sum_{o \in \mathcal{O}} \pi_t(o) T_t^{\Delta}(o) = 0.$$

Consequently,

$$\frac{d}{dt} \nabla \rho_q(\theta_t) = - \sum_{o \in \mathcal{O}} (1 - r(q,o)) \pi_t(o) T_t^{\Delta}(o),$$

and hence

$$\Big\| \frac{d}{dt} \nabla \rho_q(\theta_t) \Big\| \le (1 - \rho_t)(L_s + \Pi_*^2) \|\Delta\|.$$

Taking the smaller of the two bounds gives

$$\Big\| \frac{d}{dt} \nabla \rho_q(\theta_t) \Big\| \le (L_s + \Pi_*^2) d(\rho_t) \|\Delta\| \le 2 (L_s + \Pi_*^2) m(\rho_t) \|\Delta\|.$$

Finally, absolute continuity and integration over $t \in [0,1]$ give

$$\|\nabla \rho_q(\theta') - \nabla \rho_q(\theta)\| \le (L_s + \Pi_*^2) \|\theta' - \theta\|.$$

□

F.2 Uniform smoothness of the normalized finite-group objective

Theorem 15 (Smoothness of the finite-group objective)**.** *Suppose Assumption 7 holds and define $L := 4L_s + 20\Pi_*^2$. Then, simultaneously for every $G \ge 2$ and $\alpha \in [0,1]$, the normalized finite-group objective $\widetilde{J}_{G,\alpha}$ is L-smooth. Consequently, the raw finite-group objective $J_{G,\alpha}$ is $L_{G,\alpha}$-smooth with the valid choice $L_{G,\alpha} = C_{G,\alpha} L$.*

Proof. Fix $G \ge 2$, $\alpha \in [0,1]$, and a prompt $q \in \mathcal{Q}$. Define the normalized per-prompt surrogate $\widetilde{\Phi}_{G,\alpha,q}(\theta) := \widetilde{F}_{G,\alpha}(\rho_q(\theta))$. By the chain rule, we have

$$\nabla \widetilde{\Phi}_{G,\alpha,q}(\theta) = \omega_{G,\alpha}(\rho_q(\theta)) \nabla \rho_q(\theta).$$

Fix $\theta, \theta' \in \mathbb{R}^d$, and write

$$\Delta := \theta' - \theta, \qquad \theta_t := \theta + t\Delta, \qquad \rho_t := \rho_q(\theta_t), \qquad m_t := m(\rho_t).$$

By Lemma 14, the map $t \mapsto \nabla \rho_q(\theta_t)$ is absolutely continuous. Since $\omega_{G,\alpha}$ is a polynomial for fixed G, it follows that

$$t \mapsto \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta_t) = \omega_{G,\alpha}(\rho_t) \nabla \rho_q(\theta_t)$$

is absolutely continuous as well. For almost every $t \in [0,1]$, we have

$$\frac{d}{dt} \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta_t) = \omega'_{G,\alpha}(\rho_t) \dot{\rho}_t \nabla \rho_q(\theta_t) + \omega_{G,\alpha}(\rho_t) \frac{d}{dt} \nabla \rho_q(\theta_t),$$

where $\dot{\rho}_t = \langle \nabla \rho_q(\theta_t), \Delta \rangle$. For the first term, Lemma 14 gives $\|\nabla \rho_q(\theta_t)\| \le 2\Pi_* m_t$, and therefore

$$|\dot{\rho}_t| \le \|\nabla \rho_q(\theta_t)\| \|\Delta\| \le 2\Pi_* m_t \|\Delta\|.$$

Using Lemma 13, $m_t^2 |\omega'_{G,\alpha}(\rho_t)| \le 4$, we obtain

$$\|\omega'_{G,\alpha}(\rho_t) \dot{\rho}_t \nabla \rho_q(\theta_t)\| \le |\omega'_{G,\alpha}(\rho_t)| \|\nabla \rho_q(\theta_t)\|^2 \|\Delta\| \le 4\Pi_*^2 m_t^2 |\omega'_{G,\alpha}(\rho_t)| \|\Delta\| \le 16 \Pi_*^2 \|\Delta\|.$$

For the second term, Lemma 14 gives, almost everywhere,

$$\Big\| \frac{d}{dt} \nabla \rho_q(\theta_t) \Big\| \le 2(L_s + \Pi_*^2) m_t \|\Delta\|.$$

Together with the weighted bound from Lemma 13, $m_t \omega_{G,\alpha}(\rho_t) \le 2$, this yields

$$\Big\| \omega_{G,\alpha}(\rho_t) \frac{d}{dt} \nabla \rho_q(\theta_t) \Big\| \le 2(L_s + \Pi_*^2) m_t \omega_{G,\alpha}(\rho_t) \|\Delta\| \le 4(L_s + \Pi_*^2) \|\Delta\|.$$

Consequently, for almost every $t \in [0,1]$,

$$\Big\| \frac{d}{dt} \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta_t) \Big\| \le (16\Pi_*^2 + 4(L_s + \Pi_*^2)) \|\Delta\|.$$

Define $L := 4L_s + 20\Pi_*^2$, then

$$\Big\| \frac{d}{dt} \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta_t) \Big\| \le L \|\Delta\|$$

almost everywhere. By absolute continuity, we have

$$\big\| \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta') - \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta) \big\| \le \int_0^1 \Big\| \frac{d}{dt} \nabla \widetilde{\Phi}_{G,\alpha,q}(\theta_t) \Big\| dt \le L \|\theta' - \theta\|,$$

where the constant L is independent of q, G, and α. Averaging over prompts therefore gives

$$\begin{aligned} \big\| \nabla \widetilde{J}_{G,\alpha}(\theta') - \nabla \widetilde{J}_{G,\alpha}(\theta) \big\| &= \big\| \mathbb{E}_Q \big[\nabla \widetilde{\Phi}_{G,\alpha,Q}(\theta') - \nabla \widetilde{\Phi}_{G,\alpha,Q}(\theta) \big] \big\| \\ &\le \mathbb{E}_Q \big[\big\| \nabla \widetilde{\Phi}_{G,\alpha,Q}(\theta') - \nabla \widetilde{\Phi}_{G,\alpha,Q}(\theta) \big\| \big] \le L \|\theta' - \theta\|. \end{aligned}$$

This shows that $\widetilde{J}_{G,\alpha}$ is L-smooth simultaneously for every $G \ge 2$ and $\alpha \in [0,1]$.

Finally, since $J_{G,\alpha}(\theta) = C_{G,\alpha} \widetilde{J}_{G,\alpha}(\theta)$ and $C_{G,\alpha}$ is independent of θ,

$$\nabla J_{G,\alpha}(\theta) = C_{G,\alpha} \nabla \widetilde{J}_{G,\alpha}(\theta).$$

Therefore, we obtain

$$\|\nabla J_{G,\alpha}(\theta') - \nabla J_{G,\alpha}(\theta)\| \le C_{G,\alpha} L \|\theta' - \theta\|,$$

meaning that $J_{G,\alpha}$ is $L_{G,\alpha}$-smooth with the valid choice $L_{G,\alpha} = C_{G,\alpha}(4L_s + 20\Pi_*^2)$. □

At $\alpha = 1$, the finite-group shaping function is constant:

$$f_{G,1}(\rho) = \frac{G-1}{G} =: c_G.$$

Since $F_{G,1}(0) = 0$, it follows that

$$F_{G,1}(\rho) = c_G \rho, \qquad J_{G,1}(\theta) = c_G J(\theta).$$

Equivalently,

$$C_{G,1} = c_G, \qquad \widetilde{F}_{G,1}(\rho) = \rho, \qquad \widetilde{J}_{G,1}(\theta) = J(\theta).$$

Moreover, Lemma 14 implies that the original mean-reward objective J is L_J-smooth with the valid choice

$$L_J := L_s + \Pi_*^2.$$

Consequently, $J_{G,1} = c_G J$ is $c_G L_J$-smooth.

F.3 Unbiased Stochastic Gradient Ascent on $J_{G,\alpha}$

F.3.1 Variance Bound for the Finite-Group Estimator

We next bound the variance of the empirical α-GRPO estimator. Since $\widehat{g}^{\alpha}_{B,G}(\theta)$ is unbiased for $\nabla J_{G,\alpha}(\theta)$, the relevant stochastic error is

$$\widehat{g}^{\alpha}_{B,G}(\theta) - \nabla J_{G,\alpha}(\theta).$$

Lemma 16 (Variance bound for the finite-group estimator)**.** *Let $G \geq 2$ and $\alpha \in [0,1]$, and define*

$$\Gamma_{G,\alpha}^2 := 4\Pi_*^2 \sup_{\rho\in[0,1]} \mathbb{E}_{X\sim\mathrm{Bin}(G,\rho)}\Big[\Big(\frac{X}{G}\Big(1-\frac{X}{G}\Big)\Big)^{2\alpha}\Big], \tag{23}$$

where, for $\alpha = 0$, the integrand is defined to be zero when $X \in \{0, G\}$, consistently with the convention that homogeneous groups contribute zero. Under Assumption 7, for every fixed $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\|\widehat{g}_{B,G}^{\alpha}(\theta) - \nabla J_{G,\alpha}(\theta)\|^2\Big] \leq \frac{\Gamma_{G,\alpha}^2}{B}.$$

Consequently,

$$\mathbb{E}\Big[\|\widehat{g}_{B,G}^{\alpha}(\theta)\|^2\Big] \leq \|\nabla J_{G,\alpha}(\theta)\|^2 + \frac{\Gamma_{G,\alpha}^2}{B}.$$

Moreover, $\Gamma_{G,\alpha} \leq 2\Pi_*(\frac{1}{4})^{\alpha} \leq 2\Pi_*$. Thus, uniformly over $\alpha \in [0,1]$, one may take $\Gamma := 2\Pi_*$.

Proof. Fix a prompt $q \in \mathcal{Q}$ and write

$$R_i := r(q, O_i), \qquad S_i := S_\theta(q, O_i) = \nabla_\theta \log \pi_\theta(O_i \mid q), \qquad \widehat{\rho} := \frac{1}{G}\sum_{i=1}^{G} R_i.$$

The corresponding fixed-prompt estimator is

$$\widehat{g}_G^{\alpha}(q,\theta) := \frac{1}{G}\sum_{i=1}^{G} f_\alpha(\widehat{\rho})(R_i - \widehat{\rho})S_i,$$

where homogeneous groups contribute zero without requiring the evaluation of f_α at the endpoints. Conditional on q, the rewards R_i are independent Bernoulli random variables with parameter $\rho_q(\theta)$. Let

$$X := \sum_{i=1}^{G} R_i.$$

Then $X \sim \mathrm{Bin}(G, \rho_q(\theta))$ and $\widehat{\rho} = X/G$. Since there are X successful and $G - X$ unsuccessful responses,

$$\frac{1}{G}\sum_{i=1}^{G} |R_i - \widehat{\rho}| = \frac{X}{G}(1-\widehat{\rho}) + \frac{G-X}{G}\widehat{\rho} = 2\widehat{\rho}(1-\widehat{\rho}).$$

Using Assumption 7,

$$\|\widehat{g}_G^{\alpha}(q,\theta)\| \leq |f_\alpha(\widehat{\rho})|\,\frac{1}{G}\sum_{i=1}^{G} |R_i - \widehat{\rho}|\|S_i\| \leq 2\Pi_*\,|f_\alpha(\widehat{\rho})|\,\widehat{\rho}(1-\widehat{\rho}).$$

For an interior empirical success rate,

$$f_\alpha(\widehat{\rho}) = \big(\widehat{\rho}(1-\widehat{\rho})\big)^{\alpha-1},$$

and therefore

$$\|\widehat{g}_G^{\alpha}(q,\theta)\| \leq 2\Pi_*\big(\widehat{\rho}(1-\widehat{\rho})\big)^{\alpha}.$$

The same inequality holds for homogeneous groups because their estimator is zero. Squaring and taking conditional expectation gives

$$\mathbb{E}\big[\|\widehat{g}_G^{\alpha}(q,\theta)\|^2 \mid q\big] \leq 4\Pi_*^2\mathbb{E}_{X\sim\mathrm{Bin}(G,\rho_q(\theta))}\Big[\Big(\frac{X}{G}\Big(1-\frac{X}{G}\Big)\Big)^{2\alpha}\Big] \leq \Gamma_{G,\alpha}^2.$$

Write the minibatch estimator as

$$\widehat{g}_{B,G}^{\alpha}(\theta) = \frac{1}{B}\sum_{b=1}^{B} Y_b, \qquad Y_b := \widehat{g}_G^{\alpha}(Q_b, \theta).$$

The random vectors $Y_1, \dots, Y_B$ are independent and identically distributed. By Proposition 2,

$$\mathbb{E}[Y_b] = \nabla J_{G,\alpha}(\theta).$$

Hence,

$$\mathbb{E}\big[\|\widehat{g}^{\alpha}_{B,G}(\theta) - \nabla J_{G,\alpha}(\theta)\|^2\big] = \frac{1}{B}\mathbb{E}\big[\|Y_1 - \mathbb{E}[Y_1]\|^2\big] \le \frac{1}{B}\mathbb{E}[\|Y_1\|^2] \le \frac{\Gamma^2_{G,\alpha}}{B},$$

where the second-moment identity follows from unbiasedness:

$$\mathbb{E}\big[\|\widehat{g}^{\alpha}_{B,G}(\theta)\|^2\big] = \|\nabla J_{G,\alpha}(\theta)\|^2 + \mathbb{E}\big[\|\widehat{g}^{\alpha}_{B,G}(\theta) - \nabla J_{G,\alpha}(\theta)\|^2\big].$$

Finally, under the preceding endpoint convention and since $x(1-x) \le 1/4$ on $[0,1]$,

$$(x(1-x))^{\alpha} \le \left(\frac{1}{4}\right)^{\alpha}$$

for every $\alpha \in [0,1]$. Thus,

$$\Gamma_{G,\alpha} \le 2\Pi_* \left(\frac{1}{4}\right)^{\alpha} \le 2\Pi_*,$$

which proves the uniform claim. □

The refined constant $\Gamma_{G,\alpha}$ may still depend on G through the distribution of the empirical success rate X/G, whereas the uniform choice $\Gamma := 2\Pi_*$ is independent of both G and α and is therefore sufficient for the analysis of a varying schedule $(\alpha_t)_{t\ge 0}$. The sharper bound decreases with α. Smaller values of α assign larger magnitude to groups with empirical success rates near the endpoints, while increasing α attenuates these contributions. At $\alpha = 0$, the product of the shaping factor and the centered empirical-success factor is constant on non-homogeneous groups.

F.3.2 Proof of Stationarity Bounds

We now prove stationarity of the random iterate selection. Let $(\mathcal{F}_t)_{t\ge 0}$ denote the filtration generated by the algorithm before the samples at iteration t are drawn. Under fresh on-policy sampling, θ_t is $\mathcal{F}_t$-measurable and the samples used to construct $\widehat{g}^{\alpha}_{B,G}(\theta_t)$ are conditionally independent draws from the on-policy distribution determined by θ_t. Therefore, Proposition 2 and Lemma 16 imply

$$\mathbb{E}\big[\widehat{g}^{\alpha}_{B,G}(\theta_t) \mid \mathcal{F}_t\big] = \nabla J_{G,\alpha}(\theta_t) \qquad \text{a.s.}, \tag{24}$$

and

$$\mathbb{E}\big[\|\widehat{g}^{\alpha}_{B,G}(\theta_t) - \nabla J_{G,\alpha}(\theta_t)\|^2 \mid \mathcal{F}_t\big] \le \frac{\Gamma^2_{G,\alpha}}{B} \qquad \text{a.s.} \tag{25}$$

By Theorem 15, $J_{G,\alpha}$ is $L_{G,\alpha}$-smooth with the valid choice $L_{G,\alpha} = C_{G,\alpha}L$ and $L = 4L_s + 20\Pi^2_*$. We are now ready to formulate the main result of this section.

Theorem 17 (Stationarity guarantee for the finite-group objective)**.** *Suppose that Assumption 7 is in place. Fix $G \ge 2$ and $\alpha \in [0,1]$, and let $J^*_{G,\alpha} := \sup_{\theta\in\mathbb{R}^d} J_{G,\alpha}(\theta)$. Let $L = 4L_s + 20\Pi^2_*$ and $0 < \eta \le \frac{1}{C_{G,\alpha}L}$, consider the iterates $(\theta_t)_{t\ge 0}$ generated by $\theta_{t+1} = \theta_t + \eta\widehat{g}^{\alpha}_{B,G}(\theta_t)$. Let $R \sim \mathrm{Unif}\{0,\dots,T-1\}$ independent of $\mathcal{F}_{T-1}$, then*

$$\mathbb{E}[\|\nabla J_{G,\alpha}(\theta_R)\|^2] \le \frac{2(J^\star_{G,\alpha} - J_{G,\alpha}(\theta_0))}{\eta T} + \frac{C_{G,\alpha}L\,\eta\Gamma^2_{G,\alpha}}{B}.$$

Proof of Theorem 17. By $L_{G,\alpha}$-smoothness of $J_{G,\alpha}$ and the update

$$\theta_{t+1} = \theta_t + \eta\widehat{g}^{\alpha}_{B,G}(\theta_t),$$

using the ascent version of the common descent lemma we have

$$J_{G,\alpha}(\theta_{t+1}) \ge J_{G,\alpha}(\theta_t) + \eta\left\langle \nabla J_{G,\alpha}(\theta_t), \widehat{g}^{\alpha}_{B,G}(\theta_t)\right\rangle - \frac{L_{G,\alpha}\eta^2}{2}\left\|\widehat{g}^{\alpha}_{B,G}(\theta_t)\right\|^2.$$

Taking conditional expectation given $\mathcal{F}_t$ and using Equation 24,

$$\mathbb{E}\left[\left\langle \nabla J_{G,\alpha}(\theta_t), \widehat{g}^{\alpha}_{B,G}(\theta_t)\right\rangle \,\middle|\, \mathcal{F}_t\right] = \left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2.$$

Moreover, by Equation 25,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}^{\alpha}_{B,G}(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] &= \left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2 + \mathbb{E}\left[\left\|\widehat{g}^{\alpha}_{B,G}(\theta_t) - \nabla J_{G,\alpha}(\theta_t)\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\ &\le \left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2 + \frac{\Gamma^2_{G,\alpha}}{B}.\end{aligned}$$

It follows that

$$\mathbb{E}\left[J_{G,\alpha}(\theta_{t+1}) - J_{G,\alpha}(\theta_t) \,|\, \mathcal{F}_t\right] \ge \left(\eta - \frac{L_{G,\alpha}\eta^2}{2}\right)\left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2 - \frac{L_{G,\alpha}\eta^2\Gamma^2_{G,\alpha}}{2B}.$$

Since $0 < \eta \le 1/L_{G,\alpha}$,

$$\eta - \frac{L_{G,\alpha}\eta^2}{2} \ge \frac{\eta}{2}.$$

Therefore,

$$\mathbb{E}\left[J_{G,\alpha}(\theta_{t+1}) - J_{G,\alpha}(\theta_t) \,|\, \mathcal{F}_t\right] \ge \frac{\eta}{2}\left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2 - \frac{L_{G,\alpha}\eta^2\Gamma^2_{G,\alpha}}{2B}.$$

Rearranging and taking total expectation yields

$$\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2\right] \le \frac{2}{\eta}\mathbb{E}\left[J_{G,\alpha}(\theta_{t+1}) - J_{G,\alpha}(\theta_t)\right] + \frac{L_{G,\alpha}\eta\Gamma^2_{G,\alpha}}{B}.$$

Summing over $t = 0, \dots, T-1$ gives

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2\right] &\le \frac{2}{\eta}\mathbb{E}\left[J_{G,\alpha}(\theta_T) - J_{G,\alpha}(\theta_0)\right] + \frac{TL_{G,\alpha}\eta\Gamma^2_{G,\alpha}}{B} \\ &\le \frac{2\big(J^*_{G,\alpha} - J_{G,\alpha}(\theta_0)\big)}{\eta} + \frac{TL_{G,\alpha}\eta\Gamma^2_{G,\alpha}}{B}.\end{aligned}$$

Since $R \sim \mathrm{Unif}\{0, \dots, T-1\}$ is independent of the algorithmic randomness,

$$\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_R)\right\|^2\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_t)\right\|^2\right] \le \frac{2\big(J^*_{G,\alpha} - J_{G,\alpha}(\theta_0)\big)}{\eta T} + \frac{L_{G,\alpha}\eta\Gamma^2_{G,\alpha}}{B}.$$

This proves the theorem. □

The result is the standard nonconvex stochastic-gradient guarantee for the finite-group objective. The optimization term decays as $1/(\eta T)$, while the stochastic term scales as η/B. The group size and shaping parameter enter through the objective $J_{G,\alpha}$, its smoothness constant $L_{G,\alpha}$, and the refined variance bound $\Gamma_{G,\alpha}$.

For completeness, define

$$\Delta_{G,\alpha,0} := J^*_{G,\alpha} - J_{G,\alpha}(\theta_0).$$

Then Theorem 17 gives

$$\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_R)\right\|^2\right] \le \frac{2\Delta_{G,\alpha,0}}{\eta T} + \frac{L_{G,\alpha}\eta\Gamma^2_{G,\alpha}}{B}.$$

Balancing the two terms suggests

$$\eta^* = \min\left\{\frac{1}{L_{G,\alpha}}, \sqrt{\frac{2\Delta_{G,\alpha,0}B}{L_{G,\alpha}\Gamma^2_{G,\alpha}T}}\right\}.$$

Whenever the square-root stepsize is admissible,

$$\mathbb{E}\left[\left\|\nabla J_{G,\alpha}(\theta_R)\right\|^2\right] \le 2\sqrt{\frac{2\Delta_{G,\alpha,0}L_{G,\alpha}\Gamma^2_{G,\alpha}}{BT}}.$$

Each iteration samples BG completions, so the total rollout budget is

$$N_{\mathrm{roll}} = TBG.$$

Consequently,

$$\mathbb{E}\left[\|\nabla J_{G,\alpha}(\theta_R)\|^2\right] \le 2\sqrt{\frac{2G\Delta_{G,\alpha,0}L_{G,\alpha}\Gamma_{G,\alpha}^2}{N_{\mathrm{roll}}}}.$$

For fixed G and α, the idealized on-policy update therefore has the standard nonconvex stochastic-gradient rollout complexity

$$N_{\mathrm{roll}} = O(\varepsilon^{-4})$$

for obtaining

$$\mathbb{E}\left[\|\nabla J_{G,\alpha}(\theta_R)\|^2\right] \le \varepsilon^2.$$

The hidden constant depends on G and α through $\Delta_{G,\alpha,0}$, $L_{G,\alpha}$, and $\Gamma_{G,\alpha}^2$.

The proof of our main result for the normalized objective is a direct consequence. Since $J_{G,\alpha} = C_{G,\alpha}\widetilde{J}_{G,\alpha}$, the effective stepsize for the normalized objective is $\gamma = \eta C_{G,\alpha}$. Hence, the condition $\eta \le 1/(C_{G,\alpha}L)$ becomes $\gamma \le 1/L$, while $0 \le \widetilde{J}_{G,\alpha} \le 1$, $C_{G,\alpha} \ge 1/2$, and $\Gamma_{G,\alpha} \le 2\Pi_*$. Thus, the corresponding normalized stationarity bound constants are uniformly bounded over G and α and we obtain the following stationarity guarantee.

Proof of Theorem 8. Define $C_{G,\alpha} := F_{G,\alpha}(1) - F_{G,\alpha}(0)$. By the definition of the normalized objective,

$$\nabla J_{G,\alpha} = C_{G,\alpha}\nabla\widetilde{J}_{G,\alpha}, \qquad J_{G,\alpha}^\star - J_{G,\alpha}(\theta_0) = C_{G,\alpha}\big(\widetilde{J}_{G,\alpha}^\star - \widetilde{J}_{G,\alpha}(\theta_0)\big).$$

Applying the finite-group stationarity bound from Theorem 17 and dividing by $C_{G,\alpha}^2$ gives, with $\gamma := \eta C_{G,\alpha}$,

$$\mathbb{E}\left[\|\nabla\widetilde{J}_{G,\alpha}(\theta_R)\|^2\right] \le \frac{2(\widetilde{J}_{G,\alpha}^\star - \widetilde{J}_{G,\alpha}(\theta_0))}{\gamma T} + \frac{L\gamma\Gamma_{G,\alpha}^2}{C_{G,\alpha}^2 B}.$$

Since $0 \le \widetilde{J}_{G,\alpha} \le 1$, $C_{G,\alpha} \ge 1/2$, and $\Gamma_{G,\alpha} \le 2\Pi_*$,

$$\mathbb{E}\left[\|\nabla\widetilde{J}_{G,\alpha}(\theta_R)\|^2\right] \le \frac{2}{\gamma T} + \frac{16L\Pi_*^2\gamma}{B}.$$

Finally, $\eta \le (C_{G,\alpha}L)^{-1}$ implies $\gamma \le 1/L$, while $C_{G,\alpha} \ge 1/2$ gives $\gamma \ge \eta/2$. □

F.4 Biased Stochastic Gradient Ascent on J

We now analyze annealed α-GRPO relative to the original reward objective

$$J(\theta) = \mathbb{E}_Q\left[\rho_Q(\theta)\right].$$

Throughout this subsection, the group size $G \ge 2$ is fixed, and we write

$$c_G := \frac{G-1}{G}.$$

At the endpoint $\alpha = 1$, the finite-group transformation satisfies

$$f_{G,1}(\rho) = c_G,$$

and hence

$$\nabla J_{G,1}(\theta) = c_G\nabla J(\theta).$$

The factor c_G is not an optimization bias: it preserves the stationary points and ascent directions of J and can be absorbed into the stepsize. The relevant bias for annealing is therefore the deviation of $\nabla J_{G,\alpha}$ from $c_G\nabla J$.

F.4.1 Bias Relative to the Original Reward Objective

For fixed $G \geq 2$ and $\alpha \in [0, 1]$, define the scaled bias

$$e_{G,\alpha}(\theta) := \nabla J_{G,\alpha}(\theta) - c_G \nabla J(\theta).$$

Using

$$\nabla J_{G,\alpha}(\theta) = \mathbb{E}_Q \left[f_{G,\alpha}(\rho_Q(\theta)) \nabla \rho_Q(\theta) \right]$$

and

$$\nabla J(\theta) = \mathbb{E}_Q \left[\nabla \rho_Q(\theta) \right],$$

we obtain

$$e_{G,\alpha}(\theta) = \mathbb{E}_Q \left[\left(f_{G,\alpha}(\rho_Q(\theta)) - c_G \right) \nabla \rho_Q(\theta) \right]. \tag{26}$$

The following proposition shows that this scaled bias vanishes uniformly as $\alpha \to 1$.

Proposition 18 (Vanishing scaled bias for α-GRPO)**.** *Suppose Assumption 7 holds. For every fixed $G \geq 2$, every $\alpha \in [0, 1]$, and every $\theta \in \mathbb{R}^d$,*

$$\|e_{G,\alpha}(\theta)\| \leq \Pi_* c_G \left[\left(\frac{G-1}{G^2} \right)^{\alpha-1} - 1 \right].$$

Consequently,

$$\sup_{\theta \in \mathbb{R}^d} \|\nabla J_{G,\alpha}(\theta) - c_G \nabla J(\theta)\| \longrightarrow 0 \qquad \textit{as} \quad \alpha \to 1.$$

Proof. Define

$$a_G := \min_{1 \leq k \leq G-1} \frac{k}{G} \left(1 - \frac{k}{G} \right) = \frac{G-1}{G^2}.$$

Since $\alpha - 1 \leq 0$, for every $k = 1, \ldots, G-1$,

$$1 \leq \left[\frac{k}{G} \left(1 - \frac{k}{G} \right) \right]^{\alpha-1} \leq a_G^{\alpha-1}.$$

Recall that

$$f_{G,\alpha}(\rho) = c_G \sum_{s=0}^{G-2} \left[\frac{s+1}{G} \left(1 - \frac{s+1}{G} \right) \right]^{\alpha-1} \binom{G-2}{s} \rho^s (1-\rho)^{G-2-s}.$$

The Bernstein basis polynomials are nonnegative and sum to one. Therefore, for every $\rho \in [0, 1]$,

$$c_G \leq f_{G,\alpha}(\rho) \leq c_G a_G^{\alpha-1}.$$

It follows that

$$\sup_{\rho \in [0,1]} |f_{G,\alpha}(\rho) - c_G| \leq c_G \left(a_G^{\alpha-1} - 1 \right). \tag{27}$$

Using Equation 26, Lemma 14, and Equation 27, we obtain

$$\begin{aligned} \|e_{G,\alpha}(\theta)\| \leq \mathbb{E}_Q \left[|f_{G,\alpha}(\rho_Q(\theta)) - c_G| \, \|\nabla \rho_Q(\theta)\| \right] &\leq \Pi_* \sup_{\rho \in [0,1]} |f_{G,\alpha}(\rho) - c_G| \\ &\leq \Pi_* c_G \left(a_G^{\alpha-1} - 1 \right) \\ &= \Pi_* c_G \left[\left(\frac{G-1}{G^2} \right)^{\alpha-1} - 1 \right]. \end{aligned}$$

Since $a_G \in (0, 1)$ is fixed,

$$a_G^{\alpha-1} \to 1 \qquad \text{as } \alpha \to 1.$$

The upper bound is independent of θ, proving uniform convergence. □

The proposition isolates the bias induced by choosing $\alpha < 1$. At $\alpha = 1$, it vanishes exactly:

$$e_{G,1}(\theta) = 0.$$

For fixed $\alpha < 1$, the finite-group expected update generally remains a nonconstant reweighting of the prompt gradients and therefore need not have the same stationary points as J.

F.4.2 Biased Stochastic Gradient Ascent for Annealed α-GRPO

Let $(\alpha_t)_{t\geq 0} \subset [0,1]$ be an annealing schedule and define

$$\beta_t := \Pi_* c_G \Big[\Big(\frac{G-1}{G^2} \Big)^{\alpha_t - 1} - 1 \Big]. \tag{28}$$

Set

$$e_t := e_{G,\alpha_t}(\theta_t).$$

Under fresh on-policy sampling, Proposition 2 and Proposition 18 imply

$$\mathbb{E}\big[\widehat{g}^{\alpha_t}_{B,G}(\theta_t) \mid \mathcal{F}_t\big] = c_G \nabla J(\theta_t) + e_t, \qquad \|e_t\| \leq \beta_t. \tag{29}$$

For fixed G,

$$\beta_t \to 0 \qquad \text{whenever} \qquad \alpha_t \to 1.$$

Moreover, Lemma 16 gives the conditional variance bound

$$\mathbb{E}\left[\|\widehat{g}^{\alpha_t}_{B,G}(\theta_t) - \mathbb{E}[\widehat{g}^{\alpha_t}_{B,G}(\theta_t) \,\Big|\, \mathcal{F}_t]\|^2 \,\Big|\, \mathcal{F}_t \right] \leq \frac{\Gamma^2}{B}, \tag{30}$$

where the uniform choice $\Gamma := 2\Pi_*$ is valid simultaneously for all $G \geq 2$ and $\alpha_t \in [0,1]$. Finally, as established in Section F.2, the original reward objective is L_J-smooth with $L_J := L_s + \Pi_*^2$.

Proof of Theorem 9. By L_J-smoothness of J and the update

$$\theta_{t+1} = \theta_t + \eta \widehat{g}^{\alpha_t}_{B,G}(\theta_t),$$

we have

$$J(\theta_{t+1}) \geq J(\theta_t) + \eta \langle \nabla J(\theta_t), \widehat{g}^{\alpha_t}_{B,G}(\theta_t) \rangle - \frac{L_J \eta^2}{2} \|\widehat{g}^{\alpha_t}_{B,G}(\theta_t)\|^2.$$

Taking conditional expectation given $\mathcal{F}_t$ and using Equation 29 and Equation 30 yields

$$\begin{aligned} \mathbb{E}\left[J(\theta_{t+1}) \,|\, \mathcal{F}_t\right] \geq\ & J(\theta_t) + \eta \left\langle \nabla J(\theta_t), c_G \nabla J(\theta_t) + e_t \right\rangle \\ & - \frac{L_J \eta^2}{2} \left\| c_G \nabla J(\theta_t) + e_t \right\|^2 - \frac{L_J \eta^2 \Gamma^2}{2B}. \end{aligned}$$

Indeed, the conditional second-moment decomposition gives

$$\mathbb{E}[\|\widehat{g}^{\alpha_t}_{B,G}(\theta_t)\|^2 \,|\, \mathcal{F}_t] = \|c_G \nabla J(\theta_t) + e_t\|^2 + \mathbb{E}\big[\|\widehat{g}^{\alpha_t}_{B,G}(\theta_t) - \mathbb{E}[\widehat{g}^{\alpha_t}_{B,G}(\theta_t) \,|\, \mathcal{F}_t]\|^2 \,\big|\, \mathcal{F}_t\big].$$

By Young's inequality,

$$\langle \nabla J(\theta_t), e_t \rangle \geq -\frac{c_G}{4} \|\nabla J(\theta_t)\|^2 - \frac{1}{c_G} \|e_t\|^2,$$

and

$$\|c_G \nabla J(\theta_t) + e_t\|^2 \leq 2c_G^2 \|\nabla J(\theta_t)\|^2 + 2\|e_t\|^2.$$

Consequently,

$$\mathbb{E}[J(\theta_{t+1}) \mid \mathcal{F}_t] \geq J(\theta_t) + \big(\frac{3c_G\eta}{4} - L_J c_G^2 \eta^2\big) \|\nabla J(\theta_t)\|^2 - \big(\frac{\eta}{c_G} + L_J \eta^2\big) \|e_t\|^2 - \frac{L_J \eta^2 \Gamma^2}{2B}.$$

Since $0 < \eta \leq \frac{1}{2L_J c_G}$, we have

$$L_J c_G^2 \eta^2 \leq \frac{c_G \eta}{2} \qquad \text{and} \qquad \frac{3c_G\eta}{4} - L_J c_G^2 \eta^2 \geq \frac{c_G \eta}{4}.$$

Moreover,

$$L_J \eta^2 \leq \frac{\eta}{2c_G} \leq \frac{\eta}{c_G}.$$

Using $\|e_t\| \leq \beta_t$, we therefore obtain

$$\mathbb{E}\left[J(\theta_{t+1}) \,|\, \mathcal{F}_t\right] \geq J(\theta_t) + \frac{c_G \eta}{4} \|\nabla J(\theta_t)\|^2 - \frac{2\eta}{c_G} \beta_t^2 - \frac{L_J \eta^2 \Gamma^2}{2B}. \tag{31}$$

Taking total expectation, summing over $t = 0, \dots, T-1$, and using $J(\theta_T) \le J^*$ for $J^* := \sup_\theta J(\theta)$ gives

$$\frac{c_G \eta}{4} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big] \le J^* - J(\theta_0) + \frac{2\eta}{c_G} \sum_{t=0}^{T-1} \beta_t^2 + \frac{L_J \eta^2 \Gamma^2 T}{2B}.$$

Since R is uniformly distributed on $\{0, \dots, T-1\}$, independently of the algorithmic randomness,

$$\mathbb{E}\big[\|\nabla J(\theta_R)\|^2\big] = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big].$$

Dividing by T and multiplying by $4/(c_G \eta)$ yields

$$\mathbb{E}\big[\|\nabla J(\theta_R)\|^2\big] \le \frac{4(J^* - J(\theta_0))}{c_G \eta T} + \frac{8}{c_G^2 T} \sum_{t=0}^{T-1} \beta_t^2 + \frac{2 L_J \eta \Gamma^2}{c_G B}.$$

This proves the theorem. □

The bound separates three effects. The first term is the usual optimization error and decays as $1/(\eta T)$. The third is the stochastic variance term and scales as η/B. The middle term is specific to annealed α-GRPO and measures the average squared bias induced by using $\alpha_t < 1$. A fixed value $\alpha < 1$ therefore leaves a nonvanishing bias term relative to J in this upper bound, whereas an annealing schedule reduces this term in the bound in proportion to

$$\frac{1}{T} \sum_{t=0}^{T-1} \beta_t^2.$$

Proposition 19 (Variable-stepsize biased stochastic ascent)**.** *Suppose Assumption 7 holds, and let* $J^* := \sup_\theta J(\theta)$. *Consider the update*

$$\theta_{t+1} = \theta_t + \eta_t \widehat{g}^{\alpha_t}_{B,G}(\theta_t),$$

where

$$0 < \eta_t \le \frac{1}{2 L_J c_G}.$$

Let $T \ge 2$ *and* $R \in \{0, \dots, T-1\}$ *be sampled independently of* $\mathcal{F}_{T-1}$ *according to*

$$\mathbb{P}(R = t) = \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s}.$$

Then

$$\mathbb{E}[\|\nabla J(\theta_R)\|^2] \le \frac{4\big(J^* - J(\theta_0)\big)}{c_G \sum_{t=0}^{T-1} \eta_t} + \frac{8}{c_G^2} \frac{\sum_{t=0}^{T-1} \eta_t \beta_t^2}{\sum_{t=0}^{T-1} \eta_t} + \frac{2 L_J \Gamma^2}{c_G B} \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}.$$

Proof. Repeating the one-step argument from the proof of Theorem 9 with stepsize η_t gives

$$\mathbb{E}\left[J(\theta_{t+1}) \,|\, \mathcal{F}_t\right] \ge J(\theta_t) + \frac{c_G \eta_t}{4} \|\nabla J(\theta_t)\|^2 - \frac{2\eta_t}{c_G} \beta_t^2 - \frac{L_J \eta_t^2 \Gamma^2}{2B}.$$

Taking total expectation, summing over $t = 0, \dots, T-1$, and using $J(\theta_T) \le J^*$, we obtain

$$\frac{c_G}{4} \sum_{t=0}^{T-1} \eta_t \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \le J^* - J(\theta_0) + \frac{2}{c_G} \sum_{t=0}^{T-1} \eta_t \beta_t^2 + \frac{L_J \Gamma^2}{2B} \sum_{t=0}^{T-1} \eta_t^2.$$

By the definition of R,

$$\mathbb{E}\big[\|\nabla J(\theta_R)\|^2\big] = \frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big]}{\sum_{t=0}^{T-1} \eta_t}.$$

Dividing the preceding inequality by $\sum_{t=0}^{T-1} \eta_t$ and multiplying by $4/c_G$ proves the result. □

The variable-stepsize bound shows that the relevant bias quantity is the stepsize-weighted average

$$\frac{\sum_{t=0}^{T-1} \eta_t \beta_t^2}{\sum_{t=0}^{T-1} \eta_t}.$$

Thus, the bias must vanish along the portion of the trajectory receiving non-negligible optimization weight.

Corollary 20 (Vanishing stationarity under Robbins–Monro stepsizes)**.** *Suppose the assumptions of Proposition 19 hold and*

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \qquad 0 < \eta_t \le \frac{1}{2L_J c_G}.$$

If

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} \eta_t \beta_t^2}{\sum_{t=0}^{T-1} \eta_t} = 0,$$

then

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big]}{\sum_{t=0}^{T-1} \eta_t} = 0.$$

For fixed G, the weighted-bias condition is satisfied whenever $\alpha_t \to 1$.

Proof. Proposition 19 gives

$$\frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big]}{\sum_{t=0}^{T-1} \eta_t} \le \frac{4\big(J^* - J(\theta_0)\big)}{c_G \sum_{t=0}^{T-1} \eta_t} + \frac{8}{c_G^2} \frac{\sum_{t=0}^{T-1} \eta_t \beta_t^2}{\sum_{t=0}^{T-1} \eta_t} + \frac{2L_J \Gamma^2}{c_G B} \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}.$$

The first term converges to zero because $\sum_{t=0}^{\infty} \eta_t = \infty$. The second term converges to zero by assumption. For the third term,

$$\sum_{t=0}^{T-1} \eta_t^2 \le \sum_{t=0}^{\infty} \eta_t^2 < \infty,$$

whereas

$$\sum_{t=0}^{T-1} \eta_t \to \infty.$$

Hence the third term also converges to zero, proving the first claim.

For the final claim, recall that

$$\beta_t = \Pi_* c_G \Big[\Big(\frac{G-1}{G^2} \Big)^{\alpha_t - 1} - 1 \Big].$$

For fixed G, the map

$$\alpha \longmapsto \Big(\frac{G-1}{G^2} \Big)^{\alpha - 1} - 1$$

is continuous on $[0, 1]$ and vanishes at $\alpha = 1$. Therefore, $\alpha_t \to 1$ implies $\beta_t \to 0$.

Moreover, $(\beta_t)_{t\ge 0}$ is bounded because $\alpha_t \in [0, 1]$. The weighted Cesàro argument therefore gives

$$\frac{\sum_{t=0}^{T-1} \eta_t \beta_t^2}{\sum_{t=0}^{T-1} \eta_t} \longrightarrow 0.$$

This proves the result. □

The corollary establishes asymptotic stationarity of J in the stepsize-weighted average, or equivalently in the corresponding random-iterate sense. A fixed value $\alpha < 1$ does not satisfy the weighted-bias condition, since then

$$\beta_t \equiv \beta_\alpha > 0.$$

The bound consequently retains a nonzero bias floor relative to J. By contrast, for fixed G, every schedule satisfying $\alpha_t \to 1$ eliminates the scaled bias asymptotically.

Remark 21 (Finite-horizon annealing schedules). The preceding corollary concerns a single infinite-time schedule. For each horizon $T \geq 2$, consider for example the finite-horizon linear schedule

$$\alpha_t^{(T)} = \frac{t}{T-1}, \qquad t = 0, \ldots, T-1.$$

The bias quantity appearing in the constant-stepsize finite-time bound is

$$\frac{1}{T} \sum_{t=0}^{T-1} \beta_t^2.$$

Although $\alpha_{T-1} = 1$, increasing the horizon stretches the full annealing path over more iterations. The average squared bias therefore does not generally converge to zero as $T \to \infty$. For example, for the linear schedule it converges to the corresponding integral of the squared bias over $[0, 1]$, which is strictly positive.

The finite-time theorem should therefore be interpreted literally: finite-horizon linear or cosine annealing improves the guarantee only insofar as it reduces the average squared bias over the realized trajectory. Asymptotic stationarity of J follows instead from a single infinite-time schedule satisfying the weighted-bias condition in Corollary 20.