DIVERSITY-PRESERVING RLVR VIA INFORMATION PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR), exemplified by Group Relative Policy Optimization (GRPO), is widely used for post-training of modern generative models. Although GRPO improves reasoning accuracy, it can also over-sharpen model output distributions, suppressing alternative correct responses and reducing solution diversity. We mitigate this issue through information projection, seeking the policy that best preserves the response diversity of the base model at a target expected reward. Divergence from the base policy serves as a proxy for diversity loss. To this end, we propose β-Bregman Policy Optimization (βBPO), a family of representation-based policy updates derived from information-projection dynamics under β-Bregman divergences. In a tabular model, we characterize the resulting projection flow and prove that it minimizes the corresponding divergence from the base policy at every attained reward level. Experiments on the Hyper-Grid task illustrate that βBPO attains a reward-maximizing policy while preserving diversity among high-reward actions. Extensive experiments on mathematical reasoning further show that βBPO maintains competitive Pass@1 while improving Pass@k at large sampling budgets, and achieves the KL-divergence-accuracy Pareto frontier, supporting information projection as a principled framework for mitigating over-sharpening during RLVR.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) becomes a standard post-training method for generative models following the success of DeepSeek-R1 (Guo et al., 2025), using verifiable rewards from ground-truth answers. Growing evidence suggests that its benefits primarily arise from sharpening behaviors already present in the base model (Yue et al., 2025; Huang et al., 2025; Wu et al., 2026). Such sharpening is beneficial to a degree, since improving Pass@1 requires increasing the total probability assigned to correct responses (Huang et al., 2025; He et al., 2025a).

However, this sharpening effect is not always desired. Specifically, the verifier adopted by policy optimization algorithms like GRPO assigns the reward to distinct correct responses and provides *no preference* among the correct alternatives, and the sharpening effect can amplify the existing probability differences among the correct ones. In particular, Yue et al. (2025) observe that the models trained via Group Relative Policy Optimization (GRPO) show a compromised Pass@k compared to the base model when k is large, indicating a decreased diversity or coverage over the correct responses. This over-sharpening effect is harmful to the output diversity (Fan et al., 2026), and raises a question: *Can RLVR improve accuracy without unnecessarily concentrating probability among equally correct responses?*

We approach this question through information projection (Bregman, 1967; Censor & Lent, 1981), and investigate the algorithm design that yields policy updates that can lead to the minimum-divergence policies at every targeted reward. Specifically, our contributions are three-fold: we **firstly** show that the canonical GRPO algorithm fails to achieve the information projection of KL-divergence by analyzing its gradient flow. Specifically, under the binary reward and single-context tabular model, we show that the KL information projection will maintain the conditional distribution over the actions within each reward set. However, the softmax policy trained via GRPO will amplify the probability differences among equally correct responses. Thus, it fails to preserve the conditional distributions within the correct action set, yielding a less diverse policy.

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.

To overcome this flaw of GRPO, we **further** propose a family of representation-based policy optimization algorithms, β-Bregman Policy Optimization (βBPO) inspired by the information-projection dynamics under β-Bregman divergences (Bregman, 1967; Censor & Lent, 1981). Under the single-context tabular model, we show that the exact projection flow minimizes the chosen divergence at every attained reward level, preserving the diversity of correct responses. Moreover, we show the updates of βBPO follow an empirical projection flow, which is proven to converge to the exact projection flow as the group size G increases with diminishing step size η.

Finally, extensive experiments on the Hyper-grid model and Large Language Model (LLM) post-training are conducted to corroborate the efficacy of βBPO. The Hyper-grid model (Sinha et al., 2026) offers exact characterizations of the trained policy, where βBPO manages to achieve nearly optimal expected reward with diverse high-reward actions and minimum divergence during training. For LLM post-training, we implement βBPO with Qwen2.5-Math-1.5B (Yang et al., 2024), Qwen3-4B and Qwen3-8B models (Yang et al., 2025) on MATH (Hendrycks et al., 2021) and DAPO (Yu et al., 2025) datasets. The results show that βBPO yields improved diversity measured by Pass@k, with up to a 6.5% increment on Pass@512 with Qwen3-8B. It also achieves the KL-divergence–accuracy Pareto frontier, which corroborates our theoretical findings. Lastly, we illustrate that βBPO can be integrated with the existing cross-query optimization methods like MaxRL (Tajwar et al., 2026), yielding further diversity improvement.

2 PRELIMINARIES

Reinforcement Learning with Verifiable Rewards RLVR is widely adopted to enhance the reasoning ability of LLMs (Shao et al., 2024; Guo et al., 2025). It optimizes the LLM parameters to maximize the expected reward of the generated responses. Let $\mathcal{X}$ be the alphabet and $\mathcal{X}^*$ be the set of all sequences whose elements are in $\mathcal{X}$. Given an input query $x \in \mathcal{X}^*$, an LLM $\pi_\theta : \mathcal{X}^* \to \Delta(\mathcal{X})$, parameterized by $\theta \in \mathbb{R}^{d_\theta}$, generates a response $y \in \mathcal{X}^*$ auto-regressively. Specifically, y_s is sampled from $\pi_\theta(\cdot \mid x, y_{<s})$ where $y_{<s} = (y_{s'})_{s'<s}$ denotes the previous generated tokens. For a whole response y, we also say y is sampled from $P_\theta(\cdot \mid x)$ with the induced full-response distribution $P_\theta(y \mid x) = \prod_{s=1}^{|y|} \pi_\theta(y_s \mid x, y_{<s})$. The reward of the query-response pair (x, y) is denoted as $r(x, y) = \phi(x, y) \in \mathbb{R}$, which is given by a verifier $\phi : \mathcal{X}^* \times \mathcal{X}^* \to \mathbb{R}$. The goal of RLVR is to learn a policy that maximizes the expected reward: $J(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\big[\mathbb{E}_{y\sim P_\theta(\cdot|x)}[r(x, y)]\big]$. where $\mathcal{D}$ is the curated query prompt set, and $x \sim \mathcal{D}$ denotes the uniform sampling on $\mathcal{D}$. To effectively optimize the expected reward $J(\theta)$, GRPO (Shao et al., 2024; Guo et al., 2025) iteratively optimizes a surrogate expected reward, starting from the base model π_{θ_0}. Concretely, for the t-th iteration of GRPO, the parameter θ_{t+1} is optimized according to the gradient of the following surrogate expected reward

$$J_t(\theta) = \mathbb{E}_{x,\{y_{t,g}\}_{g=1}^G}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|y_{t,g}|}\sum_{s=1}^{|y_{t,g}|}\min\Big(w_{t,g,s}(\theta)\widehat{A}_{t,g}, \text{clip}\left(w_{t,g,s}(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{A}_{t,g}\Big)\right],$$

where the expectation is taken with $x \sim \mathcal{D}$ and $\{y_{t,g}\}_{g=1}^G \overset{i.i.d.}{\sim} P_{\theta_t}(\cdot \mid x)$. In the expression, $|y|$ denotes the length of the response y, $\text{clip}(w, 1-\varepsilon, 1+\varepsilon) = \min\{\max\{w, 1-\varepsilon\}, 1+\varepsilon\}$ is the clip function used to control the parameter update. The importance ratio of $y_{t,g,s}$ is defined as $w_{t,g,s}(\theta) = \pi_\theta(y_{t,g,s} \mid x, y_{t,g,<s})/\pi_{\theta_t}(y_{t,g,s} \mid x, y_{t,g,<s})$ and $\widehat{A}_{t,g} = (r_{t,g} - \bar{r}_t)/\text{std}(\{r_{t,g'}\}_{g'=1}^G)$ is the normalized advantage of response $y_{t,g}$ where $r_{t,g} = r(x, y_{t,g})$,

$$\bar{r}_t = \text{mean}(\{r_{t,g}\}_{g=1}^G) = \frac{\sum_g r_{t,g}}{G} \quad \text{and} \quad \text{std}(\{r_{t,g}\}_{g=1}^G) = \sqrt{\frac{\sum_g (r_{t,g} - \bar{r}_t)^2}{G}}.$$

We set $\widehat{A}_{t,g} = 0$ when the standard deviation is zero. A close variant of GRPO, Dr. GRPO (Liu et al., 2025), removes the length normalization $|y_{t,g}|^{-1}$ and the standard deviation $\text{std}(\{r_{t,g}\}_{g=1}^G)$ in the advantage $\widehat{A}_{t,g}$, to keep the unbiasedness and to improve the token efficiency. More related works are in Appendix A.

The Sharpening Effect RLVR, for instance through GRPO (Guo et al., 2025), biases the output distribution toward high-reward responses and typically reduces the policy entropy. This tendency is often termed the *sharpening effect* (Huang et al., 2025; Fan et al., 2026). A moderate sharpening

effect is desirable, as it increases the probability of sampling correct responses and consequently improves the accuracy. However, the *over-sharpening* effect has been observed, i.e., the sampling probabilities of some correct responses get decreased after update (Fan et al., 2026). Specifically, Yue et al. (2025) find that Pass@$k(\mathcal{D})$ decreases after RLVR training compared to the base model when K is large (e.g., $k = 1024$ for math tasks), where

$$\text{Pass@}k(\mathcal{D}_{\text{test}}) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathbb{E}_{\{y^{(m)}\}_{m=1}^{k}\overset{\text{i.i.d.}}{\sim}P_\theta(\cdot|x)}\left[\max_{m\in[k]} r(x, y^{(m)})\right]\right],$$

for a given $k \in \mathbb{N}$, a test prompts $\mathcal{D}_{\text{test}}$ and binary rewards $r(x, y) \in \{0, 1\}$. To interpret, the policy becomes concentrated and homogeneous responses are sampled during evaluation, which reduces exploration over the candidate responses. As a result, some problems that can be solved by the base model become unsolved even with a large number of independent attempts by the trained model.

Information Projection We introduce the notions of D-*projection* and *information projection*, which define the policy that has the minimum divergence within an admissible set. We will adopt this notion to characterize the trained policy in the following sections.

Definition 2.1 (D-projection/Information projection (Bregman, 1967; Censor & Lent, 1981))**.** *Let $\mathcal{P}$ denote a space of probability distributions, let $\mathcal{M} \subseteq \mathcal{P}$ be an admissible set, and let $\mathrm{D} : \mathcal{P}\times\mathcal{P} \to \mathbb{R}_+ \cup \{+\infty\}$ be a divergence measure. Given a reference distribution $Q \in \mathcal{P}$, the D-projection of Q onto $\mathcal{M}$ with respect to D is defined as $\Pi^{\mathrm{D}}_{\mathcal{M}}(Q) \in \operatorname{argmin}_{P\in\mathcal{M}} \mathrm{D}(P\|Q)$. In particular, when $\mathrm{D}(\cdot\|\cdot) = \mathrm{KL}(\cdot\|\cdot)$, it is called the information projection.*

When $\mathrm{D}(\cdot\|\cdot)$ is clear from the context, we use "information projection" to refer to either projection. We further introduce β-Bregman divergence, which generalizes the KL-divergence.

Definition 2.2 (β-Bregman divergence (Bregman, 1967))**.** *Define the convex potential function $\Phi_\beta(\pi) = \sum_{i=1}^{K} \pi_i^\beta/[\beta(\beta-1)]$. where $\beta \in \mathbb{R}\setminus\{0, 1\}$. The associated β-Bregman divergence is*

$$\mathrm{D}_\beta(\pi\|\pi') = \Phi_\beta(\pi) - \Phi_\beta(\pi') - \nabla\Phi_\beta(\pi')^\top(\pi - \pi') \tag{1}$$

The β-Bregman divergence can be extended to $\beta \in \{0, 1\}$. In the limit $\beta \to 1$, when π and π' are probability distributions, $\mathrm{D}_\beta(\pi\|\pi') \to \mathrm{KL}(\pi\|\pi') = \mathrm{D}_1(\pi\|\pi')$. Furthermore, when $\beta \to 0$, $\mathrm{D}_\beta(\pi\|\pi')$ converges to the Itakura-Saito divergence $\mathrm{D}_\beta(\pi\|\pi') \to \mathrm{D}_0(\pi\|\pi') = \sum_i(\pi_i/\pi'_i) - \log(\pi_i/\pi'_i) - 1$. Lastly, when $\beta = 2$, it is the squared Euclidean distance $\mathrm{D}_2(\pi\|\pi') = \|\pi - \pi'\|^2/2$.

3 RLVR WITH INFORMATION PROJECTION PRINCIPLE

In this section, we aim to identify a policy that can achieve a targeted expected reward while demonstrating good response diversity. We compare the diversity of the policies at a given expected reward level. Specifically, given a target expected reward $u \in \mathbb{R}$, we define the *admissible policy set* $\mathcal{M}_u = \{\pi \in \Delta_K : J(\pi) = u\}$. We also write $\mathcal{M}_{\geq u} = \{\pi \in \Delta_K : J(\pi) \geq u\}$. Among all policies attaining the same expected reward u, we ask *which policy can achieve a good "diversity"?*

Given that the sharpening effect concentrates the probability mass on a subset of responses, the base model π^0 can serve as a reference policy. We adopt the divergence between a policy π and the base model, $\mathrm{D}(\pi\|\pi^0)$, as a proxy measure of the diversity of π. We avoid the forward divergence $\mathrm{D}(\pi^0\|\pi)$ because its mode-covering behavior (Chen et al., 2026) encourages retaining mass on low-reward responses to which π^0 assigns non-negligible probability. The problem becomes

Which policy $\pi \in \mathcal{M}_u$ achieves the minimum divergence $\mathrm{D}(\pi\|\pi^0)$ from the initial base policy π^0?

This is precisely the *information projection* problem on the admissible policy set $\mathcal{M}_u$ (Bregman, 1967; Censor & Lent, 1981). In the following, we show that GRPO fails to achieve the information projection under the *single-context tabular model* with the softmax policy. This model aligns with the commonly used outcome-reward setup and is widely used in the literature (Mei et al., 2020; Lazić et al., 2026; Sinha et al., 2026).

Single-Context Tabular Model Let $\mathcal{A} = [K]$ be a finite action set, where $\mathbf{e}_i \in \mathbb{R}^K$ denotes the one-hot vector with the i-th coordinate being 1 and others being 0. For each action $i \in \mathcal{A}$, $\mathbf{e}_i$ denotes its standard-basis feature vector. A policy π is a vector on the probability simplex $\Delta_K = \{\pi \in$

Algorithm 1 GRPO for Single-Context Tabular Model

Require: Logits $z_t \in \mathbb{R}^K$, rewards $r = (r_i)_{i=1}^K$, group size G, learning rate $\eta > 0$.

1: Set $\pi_t = \mathrm{softmax}(z_t)$ and sample $a_{t,1}, \ldots, a_{t,G} \overset{\text{i.i.d.}}{\sim} \pi_t$.

2: Observe $r_{t,g} = r_{a_{t,g}}$, compute $\widehat{b}_t = \mathrm{mean}(\{r_{t,g}\}_{g=1}^G)$ and $A_{t,g} = r_{t,g} - \widehat{b}_t$.

3: Update $z_{t+1} = z_t + \eta \nabla_z \widehat{J}_t(z)|_{z=z_t}$ and set $\pi_{t+1} = \mathrm{softmax}(z_{t+1})$, where

$$\widehat{J}_t(z) = \frac{1}{G}\sum_{g=1}^{G} \min\left\{w_{t,g}(z)A_{t,g}, \mathrm{clip}(w_{t,g}(z), 1-\varepsilon, 1+\varepsilon)A_{t,g}\right\}, \quad w_{t,g}(z) = \frac{\pi_{a_{t,g}}(z)}{\pi_{t,a_{t,g}}}. \tag{2}$$

$\mathbb{R}_+^K : \sum_{i=1}^K \pi_i = 1\}$. Each action i is also associated with a reward $r_i \in \mathbb{R}$ which is assumed to be deterministic. The expected reward objective for a given policy π is $J(\pi) = \sum_{i\in[K]} \pi_i r_i$. We start training from a pretrained or base policy π^0, where $\pi_i^0 > 0$ for every action $i \in [K]$. We follow the usual tabular softmax parameterization of the policy, i.e., each action i is associated with a logit $z_i \in \mathbb{R}$ and the softmax policy is a distribution parameterized by the logits $z \in \mathbb{R}^K$: $\pi_i(z) = \mathrm{softmax}(z)_i = e^{z_i}/\sum_{j=1}^K e^{z_j}$. For binary rewards, define the nonempty correct and incorrect action sets $\mathcal{C} = \{i : r_i = 1\}$ and $\mathcal{B} = \{i : r_i = 0\}$, respectively. Assume that $\mathcal{C}$ and $\mathcal{B}$ are both nonempty. Denote $q = \sum_{i\in\mathcal{C}} \pi_i$ as the cumulative probability of correct actions.

We also define the *conditional success and failure probabilities* as ρ and σ with $\rho_i = \pi_i/q$for $i \in \mathcal{C}$, and $\sigma_j = \pi_j/(1-q)$ for $j \in \mathcal{B}$, respectively. We present the training dynamics for one iteration of GRPO on the single-context tabular model in Algorithm 1. For analytical clarity, we ignore the standard deviation, and we show in Lemma F.1 that it induces the same policy trajectory, differing only by a reparameterization of time.

The KL-divergence between two discrete distributions is defined as $\mathrm{KL}(\pi\|\pi^0) = \sum_{i:\pi_i>0} \pi_i \log(\pi_i/\pi_i^0)$, and $\mathrm{KL}(p\|q) = \mathrm{KL}(\mathrm{Bern}(p)\|\mathrm{Bern}(q))$ for $p, q \in (0,1)$. Because the KL-divergence has been widely used in the Reinforcement Learning (RL) literature (Schulman et al., 2015; Ouyang et al., 2022; Shenfeld et al., 2026), we put our emphasis on the KL-divergence in the following discussions.

3.1 GRPO Fails to Achieve the KL Information Projection

We focus on GRPO, a memory-efficient and widely adopted RLVR method. We show that its policy trajectory generally differs from the KL information-projection path. The analysis is carried out on the *gradient flow* of GRPO: $\dot{z} = \mathrm{d}z(\tau)/\mathrm{d}\tau = \mathbb{E}_{a\sim\pi(z)}\left[(r_a - J(\pi(z)))\nabla_z \log \pi_a(z)\right] = \xi(z)$, where we use τ to index the time of the flow and consider the binary reward case. We provide the detailed derivation of $\xi(z)$ in Appendix F.4. Denote initial accuracy $q_0 = \pi^0(\mathcal{C})$, and let ρ^0, σ^0 denote the conditional success and failure distributions of the base policy π^0.

In order to show GRPO is not a KL information projection, we firstly characterize the property that a KL information projection needs to satisfy.

Proposition 3.1 (Information projection at matched accuracy)**.** *For every binary-reward policy π satisfying $q \in (0,1)$, $\pi_i = q\rho_i$ for $i \in \mathcal{C}$, and $\pi_i = (1-q)\sigma_i$ for $i \in \mathcal{B}$,*

$$\mathrm{KL}(\pi\|\pi^0) = \mathrm{KL}(q\|q_0) + q\,\mathrm{KL}(\rho\|\rho^0) + (1-q)\,\mathrm{KL}(\sigma\|\sigma^0). \tag{3}$$

Consequently, for each fixed $q \in (0,1)$, $\pi^{\mathrm{KL},q} = \mathrm{argmin}_{\pi\in\Delta_K:\,\pi(\mathcal{C})=q} \mathrm{KL}(\pi\|\pi^0)$ is unique and

$$\pi_i^{\mathrm{KL},q} = \frac{q}{q_0}\pi_i^0 \quad \textit{for } i \in \mathcal{C}, \qquad \pi_j^{\mathrm{KL},q} = \frac{1-q}{1-q_0}\pi_j^0 \quad \textit{for } j \in \mathcal{B}. \tag{4}$$

The decomposition in (3) indicates that for any policy $\pi \in \mathcal{M}_q$, we have $\mathrm{KL}(\pi\|\pi^0)g \geq \mathrm{KL}(q\|q_0)$, and equality is obtained when the latter two terms are 0, i.e., (4) is satisfied. In other words, as the expected reward increases, the projection *only changes* the mass assigned to the two reward sets but *preserves* the conditional probabilities within each set. This property is desired as it maintains the correct response diversity of the base model. We now turn to the updates of GRPO.

Algorithm 2 β-BREGMAN POLICY OPTIMIZATION (βBPO)

Require: Prompt $x \in \mathcal{X}^*$, current parameters θ_t, group size G, $\beta \in \mathbb{R}$, $\lambda > 0$, verifier ϕ, feature map $h : \mathcal{X}^* \to \mathbb{R}^d$, and learning rate $\eta > 0$.

1: Sample $\{y_{t,g}\}_{g=1}^{G} \overset{\text{i.i.d.}}{\sim} P_{\theta_t}(\cdot \mid x)$ and set $T_{t,g} = |y_{t,g}|$.

2: Compute reward $r_{t,g} = \phi(x, y_{t,g})$ and representation $h_{t,g} = \text{stopgrad}(h(y_{t,g}))$.

3: Set $\Sigma_t = \lambda I_d + \sum_{g'=1}^{G} h_{t,g'} h_{t,g'}^\top$ and $\ell_{t,g} = h_{t,g}^\top \Sigma_t^{-1} h_{t,g}$.

4: Set $\zeta_{t,g} = G\ell_{t,g}$ and $\bar{\zeta}_{t,g} = \text{stopgrad}\Big(\frac{\zeta_{t,g}}{(G^{-1}\sum_{g'} \zeta_{t,g'})}\Big)$.

5: Compute $\widehat{b}_{\beta,t} = \frac{\sum_g \bar{\zeta}_{t,g}^{\beta-1} r_{t,g}}{\sum_g \bar{\zeta}_{t,g}^{\beta-1}}$ and $A_{t,g}^{\beta} = r_{t,g} - \widehat{b}_{\beta,t}$.

6: Compute $\widehat{\mu}_{\beta,t} = \text{stopgrad}\Big(\frac{\sum_{g'} \bar{\zeta}_{t,g'}^{\beta} A_{t,g'}^{\beta}}{\sum_{g'} \bar{\zeta}_{t,g'}}\Big)$ and $\tilde{A}_{t,g} = \text{stopgrad}\Big(\bar{\zeta}_{t,g}\big(\bar{\zeta}_{t,g}^{\beta-1} A_{t,g}^{\beta} - \widehat{\mu}_{\beta,t}\big)\Big)$.

7: Update $\theta_{t+1} = \theta_t + \eta \nabla_\theta \widehat{J}_{\beta,t}(\theta)\big|_{\theta=\theta_t}$ where

$$\widehat{J}_{\beta,t}(\theta) = \frac{1}{G} \sum_{g=1}^{G} \sum_{s=1}^{T_{t,g}} \min\Big\{ w_{t,g,s}(\theta) \tilde{A}_{t,g},\ \text{clip}\Big(w_{t,g,s}(\theta), 1-\varepsilon, 1+\varepsilon\Big) \tilde{A}_{t,g} \Big\}.$$

where $w_{t,g,s}(\theta) = \dfrac{\pi_\theta(y_{t,g,s} \mid x, y_{t,g,<s})}{\pi_{\theta_t}(y_{t,g,s} \mid x, y_{t,g,<s})}$.

8: **return** θ_{t+1}

Lemma 3.2 (GRPO over-sharpens)**.** *Under the gradient flow of GRPO $\dot{z} = \xi(z)$ with $\pi(0) = \pi^0$, write $q(\tau) = \sum_{n \in \mathcal{C}} \pi_n(\tau)$, and suppose nondegenerate binary rewards are used, if there exists a correct action pair (i, j) with $\pi_i(0) > \pi_j(0)$, then for every $\tau > 0$,*

$$\frac{d}{d\tau} \log \frac{\pi_i}{\pi_j} = \dot{z}_i - \dot{z}_j = \big(1 - q\big)\big(\pi_i - \pi_j\big) > 0 \quad \textit{and} \quad \frac{\pi_i(\tau)}{\pi_j(\tau)} > \frac{\pi_i(0)}{\pi_j(0)}. \tag{5}$$

Furthermore, $\pi(\tau) \notin \operatorname{argmin}_{\pi \in \mathcal{M}_{J(\pi(\tau))}} \mathrm{KL}(\pi \| \pi^0)$.

The first inequality in (5) shows that the larger correct-action probability receives the larger logit velocity, even though both actions are correct. This results in increased probability ratios. We describe this phenomenon as "rich-get-richer". Consequently, it yields a more concentrated policy even within the correct action set, and the conditional probabilities σ^0 and ρ^0 in (4) are not retained. Hence, the policy trained by GRPO is not an information projection.

3.2 OUR METHOD: β-BREGMAN POLICY OPTIMIZATION

To mitigate the over-sharpening effect of GRPO characterized in Lemma 3.2, we seek to balance the updates assigned to different responses. In this section, we consider β-*Bregman divergence*, a generalization of KL-divergence, as the geometry underlying the information projection.

We present our algorithm, β-BREGMAN POLICY OPTIMIZATION (βBPO), in Algorithm 2. The algorithm first samples G i.i.d. responses from the current policy $P_{\theta_t}(\cdot \mid x)$ for a given query x (Line 1). The verifier then assigns each response a verified reward $r_{t,g}$ and extracts its representation $h_{t,g} \in \mathbb{R}^d$, and the operator stopgrad$(\cdot)$ is the identity function in the forward pass and has zero derivative during optimization. (Line 2). Next, the design matrix Σ_t and leverage scores $\ell_{t,g}$ are computed (Line 3), which quantify the level of uncertainty and are widely adopted in the RL literature (Soare et al., 2014; Tuyls et al., 2026). For stability, leverage scores are then normalized (Line 4). The weighted baseline $\widehat{b}_{\beta,t}$ is then calculated, where the rare responses have larger weights in the baseline when $\beta \geq 1$ (Line 5). The trajectory coefficients are adjusted so that the under-represented responses obtain larger advantages through $\bar{\zeta}_{t,g}^{\beta-1}$ in $\tilde{A}_{g,t}$ and the outer $\bar{\zeta}_{t,g}$ compensates for the sampling frequency of these rare responses (Line 6). This yields more balanced updates across actions or responses, thus mitigating the over-sharpening effect of GRPO as in Lemma 3.2. Finally, the algorithm optimizes the GRPO-style objective using the carefully constructed advan-

tage $\tilde{A}_{t,g}$ (Line 7). During implementation, the only modification to the GRPO-style objective is the replacement of its original advantage with $\tilde{A}_{t,g}$.

For the simplicity of the analysis, we consider the single-context tabular model. We present the simplified version of Algorithm 2 in Algorithm 3 in Appendix C. In the following, we show that the update follows the empirical projection flow induced by the β-Bregman divergence in Definition 2.2.

Regularized Projection Path To characterize the policy under the training dynamics induced by Algorithm 3, we introduce the regularized projection path (Csiszár, 1975; Beck & Teboulle, 2003; Zhan et al., 2023). Given a pretrained policy π^0, a β-Bregman divergence $\mathrm{D}_\beta(\cdot\|\cdot)$ and $\tau \geq 0$, we define the *Regularized Projection Path* as

$$\pi_\beta(\tau) = \operatorname*{argmax}_{\pi \in \Delta_K} \tau J(\pi) - \mathrm{D}_\beta(\pi\|\pi^0), \tag{6}$$

where the parameter τ is a reward-constraint Lagrange multiplier for the information projection problem. When $\tau = 0$, the solution $\pi_\beta(0) = \pi^0$; as τ increases, $\pi_\beta(\tau)$ tends to maintain a high expected reward. We denote $\mathcal{S}_\beta(\tau) = \{i : \pi_{\beta,i}(\tau) > 0\}$ as the set of active coordinates.

While the closed-form projection path can be obtained (see Lemma F.2), the policy $\pi_\beta(\tau)$ is usually a parameterized neural network in modern generative models. Therefore, a gradient-based update that leads to $\pi_\beta(\tau)$ is required. We present the following *projection flow* of the policy, which characterizes the gradient. We suppress the notation τ when there is no risk of confusion.

Theorem 3.3 (Projection flow)**.** *Fix $\beta \in \mathbb{R}$, on every interval on which the active support $\mathcal{S}_\beta$ is fixed, the path $\pi_\beta(\tau)$* (6) *and the logit of the softmax policy z_β satisfy*

$$\dot{\pi}_{\beta,i} = \pi_{\beta,i}^{2-\beta}(r_i - b_\beta(\pi_\beta)),\ \forall i \in \mathcal{S}_\beta \quad \textit{and} \quad \dot{z}_{\beta,i} = \pi_{\beta,i}^{1-\beta}\big(r_i - b_\beta(\pi_\beta)\big) \cdot \mathbb{1}\{i \in \mathcal{S}_\beta\} \tag{7}$$

where $b_\beta(\pi_\beta) = \sum_{i\in\mathcal{S}_\beta} \pi_{\beta,i}^{2-\beta} r_i / \sum_{i\in\mathcal{S}_\beta} \pi_{\beta,i}^{2-\beta}$. Equivalently, the policy velocity is the natural gradient induced by the β-Bregman divergence on the active coordinates

$$\dot{\pi_\beta} = \operatorname*{argmax}_{v \in \{\mathbf{1}_K^\top v = 0, v_i = 0, i \notin \mathcal{S}_\beta\}} \left\{ r^\top v - \frac{1}{2} v^\top H_\beta(\pi_\beta) v \right\} \quad \textit{where } H_\beta(\pi) = \nabla^2 \Phi_\beta(\pi). \tag{8}$$

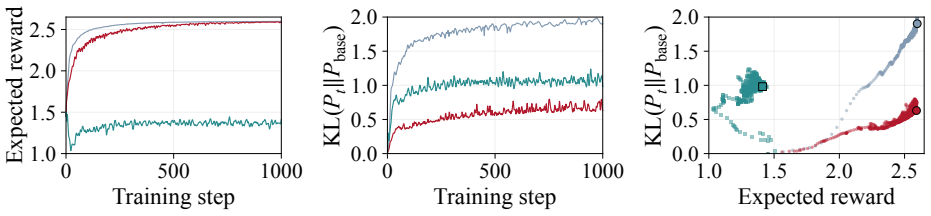
Given $\beta \geq 1$, Theorem 3.3 indicates that the rare and correct actions get more logit increments. In particular, for two actions i, j with $r_i = r_j$ and $\pi_i > \pi_j$, both actions get positive update as $\pi_{\beta,k} > 0$ for $k = i, j$, but action j get *more* logit updates as $\pi_{\beta,j}^{1-\beta} > \pi_{\beta,i}^{1-\beta}$. This mitigates the over-sharpening effect of GRPO in Lemma 3.2. We further show that the projection flow in Theorem 3.3 leads to the policy with minimum divergence at matched expected reward.

Proposition 3.4 (Minimum divergence at matched expected reward)**.** *Fix $\beta \in \mathbb{R}$ and let $q(\tau) = J(\pi_\beta(\tau))$. For every $\tau \geq 0$, $\pi_\beta(\tau) = \operatorname{argmin}_{\pi \in \mathcal{M}_{q(\tau)}} \mathrm{D}_\beta(\pi\|\pi^0)$. More strongly, $\pi_\beta(\tau) = \operatorname{argmin}_{\pi \in \mathcal{M}_{\geq q(\tau)}} \mathrm{D}_\beta(\pi\|\pi^0)$, and the minimizer is unique.*

This proposition suggests that the projection path defined in (6) is an information projection at an expected reward level. By following the projection flow (7), we can arrive at this desired policy. We show that the updates given by Algorithm 3 satisfy an empirical projection flow, given by the empirical observations. Denote $N_{t,i} = \sum_{g=1}^G \mathbb{1}\{a_{t,g} = i\}$, $\widehat{\pi}_{t,i} = N_{t,i}/G$, and $\widehat{\mathcal{S}}_t = \{i : \widehat{\pi}_{t,i} > 0\}$.

Proposition 3.5 (Empirical Projection Flow)**.** *For deterministic action rewards, Algorithm 3 can be written as $z_{t+1} = z_t + \eta \widehat{g}_{\beta,t}$, and $\widehat{g}_{\beta,t,i} = \widehat{\pi}_{t,i}^{1-\beta}(r_i - \widehat{b}_{\beta,t}) \cdot \mathbb{1}\{i \in \widehat{\mathcal{S}}_t\}$, where $\widehat{b}_{\beta,t} = \sum_{i\in\widehat{\mathcal{S}}_t} \widehat{\pi}_{t,i}^{2-\beta} r_i / \sum_{i\in\widehat{\mathcal{S}}_t} \widehat{\pi}_{t,i}^{2-\beta}$ Thus $\widehat{g}_{\beta,t}$ is exactly the β-projection logit flow for the empirical policy on its active coordinates.*

Proposition 3.5 is a pathwise statement about the empirical active simplex, where the update shares the same form as (7) with π_β replaced by the empirical $\hat{\pi}$. In Theorem F.5, we show that the expected update direction converges to the exact projection flow of π. In conclusion, Algorithm 3 serves as an empirical flow of the exact projection flow in (7). We further compare βBPO with the existing methods in Appendix B.



(a) Expected reward (b) KL($P_t \| P_{\text{base}}$) (c) Reward vs. KL

Figure 1: Training dynamics of the reward and KL-divergence on the Hyper-grid task. βBPO achieves near-optimal expected reward, and maintains the lowest KL-divergence during training and at matched expected reward level.

4 Experiments

In this section, we conduct extensive experiments to evaluate the design of βBPO in terms of accuracy and maintenance of diversity. We also compare βBPO with the other baselines. Specifically, we consider two sets of experiments. First, to enable clear interpretation and exact characterization of the response distribution, we conduct experiments on the Hyper-grid task, a controlled discrete grid-based environment adopted in prior work (Sinha et al., 2026). We then extend our method to the post-training of LLMs on popular mathematical tasks to showcase its empirical gains.

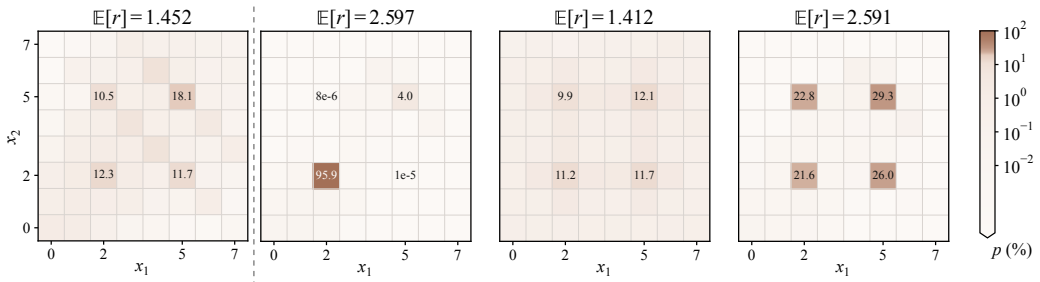
4.1 Experiment 1: Hyper-grid

Setup For the Hyper-grid task, we consider an n-dim discrete space $\mathcal{G} = \{0, 1, \ldots, H-1\}^n$ where n is the dimensionality of the hyper-grid and H denotes the horizon along each coordinate. The model will always start at $\mathbf{0} = (0, \ldots, 0)$ and at each step it can either choose and increase one coordinate by 1 without exceeding $H-1$ or terminate the rollout to receive the terminal reward associated with the grid. The reward landscape is inspired by IPS-GRPO, with explicitly defined peak and adjacent regions. Let $\mathcal{P}_{\texttt{peak}} \subset \mathcal{G}$ denote a set of P peak states, and define their adjacent states as $\mathcal{P}_{\texttt{adj}} = \{\mathbf{x} \in \mathcal{G} \setminus \mathcal{P}_{\texttt{peak}} : \exists\, \mathbf{p} \in \mathcal{P}_{\texttt{peak}},\ \|\mathbf{x} - \mathbf{p}\|_1 = 1\}$. The terminal reward associated with $\mathbf{x}_{\texttt{ter}}$ is $R = R_0 + R_1 \mathbf{1}\{\mathbf{x}_{\texttt{ter}} \in \mathcal{P}_{\texttt{peak}} \cup \mathcal{P}_{\texttt{adj}}\} + R_2 \mathbf{1}\{\mathbf{x}_{\texttt{ter}} \in \mathcal{P}_{\texttt{peak}}\}$ with $0 < R_0 < R_1 < R_2$. In our experiments, we set $n = 2$, $H = 8$ and $P = 4$, with $\mathcal{P}_{\texttt{peak}} = \{(2,2), (2,5), (5,2), (5,5)\}$. We also set $R_0 = 0.1, R_1 = 0.5, R_2 = 2$ as IPS-GRPO (Sinha et al., 2026). Therefore, the reward landscape has four global optima. We use a two-layer decoder-only Transformer as the policy model. Similar to the standard generative model training pipeline, we first perform Supervised Fine-Tuning (SFT) to obtain a base policy π_{θ_0} that induces a terminal distribution P_{base} with good coverage over the high-reward regions (Figure 2a). We then conduct training from this shared base using βBPO, IPS-GRPO, and GRPO. Further model and training details are provided in Appendix D.1.

Experimental Results At each training step t, we compute the exact terminal distribution P_t by enumerating action sequences and multiplying the conditional probabilities assigned to the actions: $P_t(\mathbf{x}) = \sum_{a \in \mathcal{T}(\mathbf{x})} \prod_{s=1}^{|a|} \pi_{\theta_t}(a_s \mid a_{<s})$, where θ_t denotes the policy parameters and $\mathcal{T}(a)$ contains all sequences starting at $\mathbf{0}$ and terminating at $\mathbf{x}$, including the final termination action. We evaluate $\mathrm{KL}(P_t \| P_{\text{base}})$ with these exact terminal distributions P_t.

Figure 1 presents the training dynamics of the expected reward and KL-divergence. From Figure 1(a) and (b), GRPO exhibits a sharp increase in both expected reward and divergence, indicating rapid concentration of probability mass on a subset of high-reward states, while IPS-GRPO maintains a good divergence but has degraded expected reward. In comparison, βBPO steadily improves the expected reward with a more controlled growth in divergence, indicating better coverage of high-reward states. Moreover, in Figure 1(c), by comparing checkpoints at matched expected reward, βBPO shows substantially lower divergence than GRPO. These results suggest βBPO can attain the target reward while remaining closer to the base distribution, consistent with Proposition 3.4.

Figure 2 presents the final distribution over the states. It shows that βBPO achieves near-optimal reward comparable to GRPO while preserving all four peak probabilities, as the training dynamics suggest. In comparison, GRPO greatly concentrates mass on a **single** terminal state, supporting our



(a) Base (b) GRPO (c) IPS-GRPO (d) βBPO

Figure 2: Terminal probability distributions of the states on the Hyper-grid task ($n = 2$, $H = 8$) with cell annotations indicating peak probabilities. βBPO achieves near-optimal expected reward while preserving all four peak probabilities that existed in the base distribution.

theoretical findings in Lemma 3.2. IPS-GRPO instead approaches the reward-proportional target $P_t(\mathbf{x}) \propto R(\mathbf{x})$ (Fan et al., 2026), yielding a diverse but **suboptimal** policy. These results strongly support our information projection approach, which can preserve the diversity in the correct actions while maximizing the expected reward.

4.2 Experiment 2: LLM Post-training

Setup For LLM post-training, we consider the RLVR setup on mathematics tasks. Specifically, we adopt Qwen2.5-Math-1.5B (Yang et al., 2024), Qwen3-4B and Qwen3-8B models (Yang et al., 2025) as the base models and they are trained on the MATH (Hendrycks et al., 2021) and DAPO (Yu et al., 2025) datasets. Our algorithm and the baselines are implemented via slime (Zhu et al., 2025b) framework with the standard parameter setup: during training, 8 rollouts will be generated for each query, the generation length is 3072 for Qwen2.5-Math-1.5B and Qwen3-4B, and 8192 for Qwen3-8B, the learning rate is 10^{-6}, the KL and entropy coefficients are 0 by default and the rollout temperature is 1.0. For the construction of the hidden states, βBPO adopts the same approach as RepExp (Tuyls et al., 2026), which has been proved to be effective under their setup. We consider the baseline methods, including GRPO (Guo et al., 2025), MaxRL (Tajwar et al., 2026), RexExp (Tuyls et al., 2026) and IAC-GRPO (Fan et al., 2026). The trained models are evaluated by the Pass@k metric on AIME 2025 (Zhang & Math-AI, 2025), AIME 2026 (Zhang & Math-AI, 2026) and BeyondAIME (ByteDance-Seed, 2025).

Furthermore, as Proposition 3.4 indicates, the β-Bregman projection flow can lead to the minimum divergence at the matched expected reward level, we evaluate the token-level KL-divergences[1],

$$\widehat{D}_{\mathrm{KL}}(P_\theta \| P_0) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}} \frac{1}{|y|} \sum_{t=1}^{|y|} \mathrm{KL}\left(\pi_\theta(v \mid x, y_{<t}) \middle\| \pi_0(v \mid x, y_{<t})\right).$$

which is computed by randomly choosing 5 responses for each query in the test set.

Lastly, βBPO involves weighting the responses **within each group**, and there are **cross-query** weighting methods, where all responses of a query share the same weight in the objective function and different query groups carry different weights (Tajwar et al., 2026). We conduct βBPO with MaxRL (Tajwar et al., 2026), where each group is weighted by the reciprocal of the number of correct responses, to demonstrate that βBPO is readily integrated with these methods.

Experimental Results Figure 3 showcases the Pass@k performance of our method βBPO and the baseline methods over the test benchmarks. Following previous works, the Pass@k metrics usually do not require an error bar, as each data point is already averaged over large samples (Tajwar et al., 2026; Tuyls et al., 2026). Across the models, βBPO and the baselines consistently improve mean accuracy (Pass@1) compared to the base model, indicating better inference efficiency yield by the RL post-training stage. As k increases, the performance of models diverges: the standard GRPO algorithm shows diminishing gains with more responses, whereas βBPO keeps the upward trend up

[1]Although the token-level KL-divergence differs from the trajectory-level KL-divergence, it is widely used in the literature as a tractable proxy, since direct computation over the full trajectory space $\mathcal{X}^*$ is impractical.

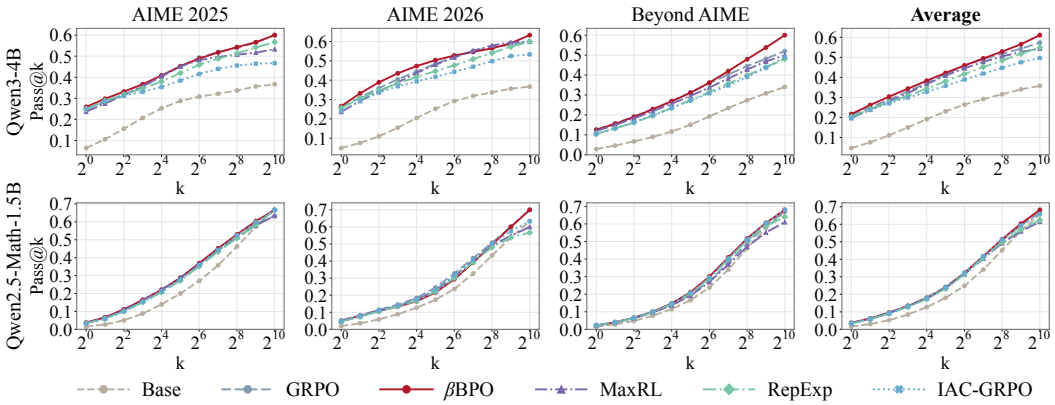
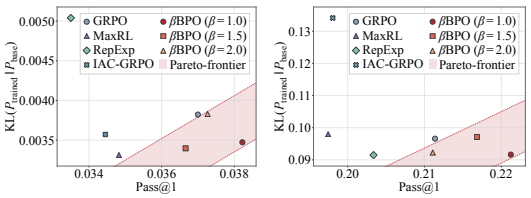


Figure 3: Pass@k performance on the test benchmarks. βBPO demonstrates better performance, especially when k is large. This indicates better diversity thanks to the principled algorithmic design, inspired by information projection.



(a) Qwen2.5-Math-1.5B (b) Qwen3-4B

Figure 4: Average token-level KL-divergence versus average accuracy. βBPO achieves the Pareto frontier of these two objectives.

Pass@k	MaxRL	βBPO	βBPO + MaxRL
Qwen2.5-Math-1.5B			
$k = 1$	0.035	0.038	0.037
$k = 1024$	0.633	0.667	**0.733**
Qwen3-4B			
$k = 1$	0.236	0.260	0.260
$k = 1024$	0.533	**0.600**	**0.600**

Table 1: Integration of βBPO with MaxRL. Both methods benefit from the integration.

to $k = 1024$. Specifically, on Qwen3-4B, βBPO achieves a Pass@1024 of 61.1%, surpassing 57.3% of the runner-up model by 3.8 percentage points. Results for Qwen3-8B, reported in Appendix D, show a gain of 6.5 percentage points in Pass@512.

Figure 4 reports the average divergence with respect to accuracy. It shows that βBPO has managed to achieve the Pareto frontier of KL-divergence and accuracy, i.e., with matched accuracy, βBPO achieves smaller divergence compared to other baselines. This corroborates the intuition given by our theoretical derivation in Proposition 3.4.

Finally, Table 1 presents the evaluation results of the integrated methods on AIME 2025. The results show that both MaxRL and βBPO benefit from the integration, suggesting that both within-group diversity, which is the focus of βBPO, and cross-query weighting, which is targeted by MaxRL, are both important in preserving the diversity of the responses of the base model. This supports that βBPO is **orthogonal** and can also be integrated with the existing cross-query methods. We present more discussion regarding the choice of β, the use of entropy regularization, and the construction of the representation in Appendix D.

5 Conclusions

We introduced βBPO to mitigate the over-sharpening effect in RLVR through information projection. In a single-context tabular model, we showed that GRPO amplifies probability differences among equally correct actions, whereas the β-Bregman projection flow minimizes divergence from the base policy at every attained reward level. These findings are corroborated by the empirical studies on Hyper-grid and mathematical reasoning, which demonstrate improved diversity and a better divergence–accuracy tradeoff, while maintaining competitive Pass@1 and improving Pass@k at large sampling budgets. These findings support information projection as a principled framework for diversity-preserving RLVR. The limitation of our study is its focus on LLMs, and the applicability of our method to multi-modal models remains to be investigated.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis. We have not used generative AI tools for implementing methods, assisting with translation, cleaning and reformatting datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, interpreting results. Generating synthetic data sets is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 3 authors. 3 authors checked the raw and processed datasets and the experimental results. For manuscript preparation, AI was used only to suggest wording improvements; the manuscript was written by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7.

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Yair Censor and Arnold Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353, 1981.

Howard Chen, Noam Razin, Karthik R Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *Forty-third International Conference on Machine Learning*, 2026.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models, 2025.

I. Csiszár. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. ISSN 00911798, 2168894X.

Mingyuan Fan, Weiguang Han, Daixin Wang, Cen Chen, Zhiqiang Zhang, and Jun Zhou. When sharpening becomes collapse: Sampling bias and semantic coupling in rl with verifiable rewards, 2026.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 09–15 Jun 2019.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 9 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z.

Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25548–25560. Association for Computational Linguistics, November 2025a. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1298.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, et al. Justrl: Scaling a 1.5 b llm with a simple rl recipe. *arXiv preprint arXiv:2512.16649*, 2025b.

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.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhiyuan Hu, Yucheng Wang, Yufei He, Jiaying Wu, Yilun Zhao, See-Kiong Ng, Cynthia Breazeal, Anh Tuan Luu, Hae Won Park, and Bryan Hooi. Rewarding the rare: Uniqueness-aware RL for creative problem solving in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 39765–39790, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1982.

Audrey Huang, Adam Block, Dylan Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 76687–76739, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui

Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025.

Nevena Lazić, Liam Fowl, András György, and Csaba Szepesvári. To see the unseen: on the generalization ability of transformers in symbolic reasoning, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020.

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

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg,

Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning, 2026.

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

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026.

Abhijeet Sinha, Sundari Elango, and Dianbo Liu. Expected return causes outcome-level mode collapse in reinforcement learning and how to fix it with inverse probability scaling. In *Forty-third International Conference on Machine Learning*, 2026.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. Maximum likelihood reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Jens Tuyls, Dylan Foster, Akshay Krishnamurthy, and Jordan Ash. Representation-based exploration for language models: From test-time to post-training. In *International Conference on Learning Representations*, volume 2026, pp. 120621–120645, 2026.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023. doi: 10.1137/21M1456789.

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

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

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

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. `https://github.com/THUDM/slime`, 2025b. GitHub repository. Corresponding author: Xin Lv.

A Related Works

A.1 Reinforcement Learning Yields Strong Reasoning Ability

Reinforcement Learning, especially RLVR, has emerged as a powerful paradigm for post-training LLMs with strong reasoning capabilities (OpenAI et al., 2026; Guo et al., 2025; Kimi Team et al., 2025; Yang et al., 2025; Yu et al., 2025) since the release of the open recipe of GRPO, a critic-free policy optimization algorithm (Guo et al., 2025). Zeng et al. (2025) and Hu et al. (2025) provide

reproducible RL for several family of base models. Many variants of GRPO have appeared to improve the vanilla GRPO algorithm: DAPO (Yu et al., 2025) improves clipping and sampling, whereas Dr. GRPO (Liu et al., 2025) removes length biases and standard deviation reweighting from policy optimization. Followup works (Zheng et al., 2025; MiniMax et al., 2025; Qi et al., 2026) introduce sequence-level importance ratios, importance-weighting clipping, and divergence-based constraints, respectively. These algorithms are primarily designed to stabilize RL training, for example by modifying trust-region approximations, importance weighting, clipping mechanisms, or strategies for mitigating off-policy staleness.

A.2 RL TRAINING SHARPENS DISTRIBUTION

Although RLVR significantly improves the reasoning performance of LLMs, recent studies suggest that these gains primarily arise from reshaping the output distribution of the model rather than learning fundamentally new behaviors (Huang et al., 2025; Yue et al., 2025; Wu et al., 2026). Huang et al. (2025) formalizes sharpening as concentrating probability on high-quality responses, hence reducing inference-time computation through self-improvement. Yu et al. (2025) finds that RLVR improves Pass@1 but can compromise base models at large k, indicating improved sampling efficiency without consistently broader responses coverage. Wu et al. (2026) further observes that answer-level diversity can decline even when token-level entropy increases. As the sharpening effect generally reduces the diversity or coverage of the base model, our work adopts the base model as the reference model when measuring the diversity via the divergences.

A.3 MAINTAINING THE DIVERSITY DURING RL TRAINING

In order to mitigate the sharpening effect, recent studies propose many approaches, for instance Pass@k-aware advantage estimation (Chen et al., 2025), trajectory reweighting (Zhu et al., 2025a; Tajwar et al., 2026), exploration-based objectives (Tuyls et al., 2026), and inverse-probability correction for outcome-level mode collapse (Sinha et al., 2026). These approaches encourage the model carry out more exploration during training or put more emphasis on the hard queries. Our work instead proposes a principled representation-based information-projection inspired advantage computation and trajectory reweighting scheme that aim to preserve the response diversity existed in the base model, while improving the expected reward.

B DISCUSSION AND COMPARISON WITH EXISTING WORKS

We compare our information-projection guided approach with the existing literature.

Firstly, our approach is related to KL-regularized policy optimization, which typically maximizes $J(\pi)-\lambda_{\mathrm{KL}}\,\mathrm{KL}(\pi\|\pi^0)$ for a fixed reference policy π^0 (Geist et al., 2019; Ouyang et al., 2022). When this objective is optimized *exactly* over the probability simplex, its solution coincides with the $\beta = 1$ in our projection path, with projection time $\tau = \lambda_{\mathrm{KL}}^{-1}$. However, standard Euclidean-gradient PPO or GRPO updates **need not trace the corresponding minimum-KL path** at intermediate reward levels. Our update is derived directly from the projection flow and extends it to the β-Bregman family. The parameter β controls the geometry independently of the reward-regularization tradeoff: KL ($\beta = 1$) preserves the reference ratios among equally rewarded actions, whereas $\beta > 1$ can redistribute mass toward underrepresented rewarded actions.

Secondly, both Tuyls et al. (2026) and βBPO adopt a representation-based approach. Specifically, the aggregated last-layer hidden states are used as the features of the generated responses. However, they are used for **different purposes**. Tuyls et al. (2026) makes use of the hidden states to get the *exploration bonus*, which is **added** to the advantage in the standard GRPO pipeline and encourages exploration over the more uncertain directions. In contrast, βBPO uses these representations to construct a proxy for *inverse effective sample counts* in feature space, which **multiplicatively** reweights the adjusted advantage (see Line 6). In the single-context tabular setting, we show that this proxy is exactly proportional to the inverse empirical action probabilities (see Lemma C.1).

Thirdly, Shenfeld et al. (2026) shows that conditioning a policy on *binary-reward* success is a KL I-projection and analyze an alternating procedure with exact or summably inexact M-projections that selects a KL-minimal optimal policy. During experiment, they conduct GRPO for binary reward

Algorithm 3 β-BREGMAN POLICY OPTIMIZATION

Require: Logits $z_t \in \mathbb{R}^K$, rewards $\{r_i\}_{i=1}^K$, $\beta \in \mathbb{R}$, group size $G \in \mathbb{N}$, learning rate $\eta > 0$

1: Set $\pi_t = \mathrm{softmax}(z_t)$ and sample $a_{t,1}, \ldots, a_{t,G} \overset{\text{i.i.d.}}{\sim} \pi_t$.
2: Observe $r_{t,g} = r_{a_{t,g}}$ and compute $\widehat{\pi}_{t,i} = G^{-1} \sum_{g=1}^{G} \mathbb{1}\{a_{t,g} = i\}$.
3: Compute $\widehat{b}_{\beta,t} = \mathrm{stopgrad}\left(\frac{\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta} r_{t,g}}{\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta}} \right)$ and $z_{t+1} \leftarrow z_t$.
4: **for** $g = 1, \ldots, G$ **do**
5: $\quad z_{t+1,a_{t,g}} \leftarrow z_{t+1,a_{t,g}} + \frac{\eta}{G} \widehat{\pi}_{t,a_{t,g}}^{-\beta} \big(r_{t,g} - \widehat{b}_{\beta,t}\big)$.
6: **end for**
7: **return** z_{t+1}

without explicit KL regularization. The Euclidean parameter-gradient updates used by GRPO generally **do not realize this operator**: a gradient step toward an M-projection is not its minimizer, and ordinary logit-gradient dynamics amplify probability differences among responses with identical rewards. We characterize this projection gap and **derive dynamics that achieve the minimum-divergence** at *every attained reward level* for general reward *beyond the binary case*. The construction is further extended to a β-Bregman family whose geometry controls whether reward-equivalent reasoning responses or actions are sharpened, preserved, or equalized.

Lastly, a related line of work maintains policy diversity through **reweighting**. He et al. (2025a) and Hu et al. (2026) down-weight the advantages of high-probability actions, scaling them by the rank of a response's unlikeliness and by an inverse power of its cluster size, respectively. Sinha et al. (2026) instead rescales each reward by the inverse of its probability, which induces a stationary distribution proportional to the reward vector; the resulting policy is therefore optimal only for binary rewards. Fan et al. (2026) down-weights the advantages of positive samples alone. In comparison, our algorithm is derived from a **principled information-projection view**, in which the coefficients are determined by the requirement that the update follow the projection flow.

C SIMPLIFICATION OF βBPO ON SINGLE-CONTEXT TABULAR MODEL

In this appendix, we provide a simplification of Algorithm 2 on the single-context tabular model, as presented Algorithm 3. It firstly initialize the current policy π_t by the logits z_t, and also samples a group of G i.i.d. actions $\{a_{t,g}\}_{g=1}^G$ (Line 1). Then the rewards of the sampled actions $\{r_{a_{t,g}}\}_{g=1}^G$ are observed and the empirical distribution $\widehat{\pi}_t$ can be computed (Line 2). Afterwards, the baseline $\widehat{b}_{\beta,t}$ is computed (Line 3). Then, the coordinate increment for each sample is accumulated, where the power is $-\beta$ as the aggregation over the G occurrences contributes one additional factor of $\widehat{\pi}_{t,i}$ (Line 5). Lastly, it returns the updated logits (Line 7).

To show that it is a simplification of Algorithm 2, we present the following lemma.

Lemma C.1 (Tabular Reduction of Algorithm 2)**.** *Consider the single-context tabular model with $\theta_t = z_t$, $T_{t,g} = 1$, $h(i) = \mathbf{e}_i$, and $d = K$. Let $N_{t,i} = \sum_{g=1}^{G} \mathbb{1}\{a_{t,g} = i\}$, $\widehat{\pi}_{t,i} = \frac{N_{t,i}}{G}$, and suppose $N_{t,i} > 0$ for every $i \in [K]$. In the ridgeless limit $\lambda \downarrow 0$, Algorithm 2 has*

$$\bar{\zeta}_{t,g} = \frac{1}{K} \widehat{\pi}_{t,a_{t,g}}^{-1} \quad \text{and} \quad \widehat{b}_{\beta,t} = \frac{\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta} r_{t,g}}{\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta}}.$$

Define the empirical logit direction generated by Algorithm 3 as

$$\widehat{g}_{\beta,t,i} = \begin{cases} \widehat{\pi}_{t,i}^{1-\beta}(r_i - \widehat{b}_{\beta,t}), & \widehat{\pi}_{t,i} > 0, \\ 0, & \widehat{\pi}_{t,i} = 0. \end{cases}$$

The gradient of Algorithm 2 then satisfies $\nabla_z \hat{J}_{\beta,t}(z)\Big|_{z=z_t} = K^{-\beta} \widehat{g}_{\beta,t} - K^{-1} \widehat{\mu}_{\beta,t} \mathbf{1}_K$ on the event that every action is observed. Consequently, Algorithm 2 and Algorithm 3 produce the same updated softmax policy after the learning-rate identification $\eta = K^{-\beta}\eta$.

Proof. Since $h(i) = \mathbf{e}_i$, $\Sigma_t = \sum_{g=1}^{G} \mathbf{e}_{a_{t,g}} \mathbf{e}_{a_{t,g}}^\top = \mathrm{diag}(N_{t,1}, \ldots, N_{t,K})$ in the limit $\lambda \downarrow 0$. Hence, whenever $a_{t,g}$ is action i, we have

$$\ell_{t,g} = N_{t,i}^{-1}, \quad \text{and} \quad \zeta_{t,g} = G\ell_{t,g} = \widehat{\pi}_{t,i}^{-1}.$$

Because every action is observed,

$$\frac{1}{G}\sum_{g=1}^{G} \zeta_{t,g} = \frac{1}{G}\sum_{i=1}^{K} N_{t,i}\frac{G}{N_{t,i}} = K,$$

which proves

$$\bar{\zeta}_{t,g} = K^{-1}\widehat{\pi}_{t,a_{t,g}}^{-1}.$$

Substitution into the definition of $\widehat{b}_{\beta,t}$ cancels the common factor $K^{1-\beta}$ and gives the stated baseline. Recall that

$$\widehat{\mu}_{\beta,t} = \frac{\sum_{g=1}^{G} \bar{\zeta}_{t,g}^{\beta} A_{t,g}^{\beta}}{\sum_{g=1}^{G} \bar{\zeta}_{t,g}}, \qquad \tilde{A}_{t,g} = \bar{\zeta}_{t,g}^{\beta} A_{t,g}^{\beta} - \widehat{\mu}_{\beta,t}\bar{\zeta}_{t,g}.$$

It follows that

$$\sum_{g=1}^{G} \tilde{A}_{t,g} = \sum_{g=1}^{G} \bar{\zeta}_{t,g}^{\beta} A_{t,g}^{\beta} - \widehat{\mu}_{\beta,t}\sum_{g=1}^{G} \bar{\zeta}_{t,g} = 0.$$

At $z = z_t$, every importance ratio equals one, so clipping is locally inactive and

$$\nabla_z w_{t,g,1}(z_t) = \mathbf{e}_{a_{t,g}} - \pi_t.$$

Therefore,

$$\nabla_z \hat{J}_{\beta,t}(z)\Big|_{z=z_t} = \frac{1}{G}\sum_{g=1}^{G} \tilde{A}_{t,g}(\mathbf{e}_{a_{t,g}} - \pi_t) = \frac{1}{G}\sum_{g=1}^{G} \tilde{A}_{t,g}\mathbf{e}_{a_{t,g}},$$

where the last equality follows from $\sum_{g=1}^{G} \tilde{A}_{t,g} = 0$.

Since $A_{t,g}^{\beta} = r_{t,g} - \widehat{b}_{\beta,t}$ and $\bar{\zeta}_{t,g} = K^{-1}\widehat{\pi}_{t,a_{t,g}}^{-1}$, we have

$$\tilde{A}_{t,g} = K^{-\beta}\widehat{\pi}_{t,a_{t,g}}^{-\beta}(r_{t,g} - \widehat{b}_{\beta,t}) - K^{-1}\widehat{\mu}_{\beta,t}\widehat{\pi}_{t,a_{t,g}}^{-1}.$$

Grouping the $N_{t,i}$ occurrences of action i yields

$$\begin{aligned}\left[\nabla_z \hat{J}_{\beta,t}(z)\Big|_{z=z_t}\right]_i &= \frac{N_{t,i}}{G}\left[K^{-\beta}\widehat{\pi}_{t,i}^{-\beta}(r_i - \widehat{b}_{\beta,t}) - K^{-1}\widehat{\mu}_{\beta,t}\widehat{\pi}_{t,i}^{-1}\right] \\ &= K^{-\beta}\widehat{\pi}_{t,i}^{1-\beta}(r_i - \widehat{b}_{\beta,t}) - K^{-1}\widehat{\mu}_{\beta,t} \\ &= K^{-\beta}\widehat{g}_{\beta,t,i} - K^{-1}\widehat{\mu}_{\beta,t}.\end{aligned}$$

Thus,

$$\nabla_z \hat{J}_{\beta,t}(z)\Big|_{z=z_t} = K^{-\beta}\widehat{g}_{\beta,t} - K^{-1}\widehat{\mu}_{\beta,t}\mathbf{1}_K.$$

Algorithm 3 accumulates the direction $\widehat{g}_{\beta,t}$. The remaining term is common to every logit and therefore does not affect softmax:

$$\mathrm{softmax}\left(z_t + \eta K^{-\beta}\widehat{g}_{\beta,t} - \eta K^{-1}\widehat{\mu}_{\beta,t}\mathbf{1}_K\right) = \mathrm{softmax}\left(z_t + \eta K^{-\beta}\widehat{g}_{\beta,t}\right).$$

Taking $\eta = K^{-\beta}\eta$ proves the result. □

D More Experimental Results

D.1 Further Details about the Hyper-grid experiment

Model architecture We use a two-layer decoder-only Transformer with hidden dimension 128, four attention heads per layer, and feed-forward networks with intermediate dimension 256 and GELU activations.

SFT initialization We construct a fixed demonstration dataset from valid trajectories starting at $\mathbf{0}$ and terminating at $\mathbf{x} \in \mathcal{P}_{\texttt{peak}}$. SFT minimizes next-action cross-entropy over these trajectories, including the final termination action. Each update samples 128 trajectories with replacement from the fixed demonstration dataset and applies Adam with learning rate 10^{-4}. The checkpoint after 75 updates serves as the shared base for βBPO, IPS-GRPO, and GRPO.

RL training We train βBPO, IPS-GRPO, and GRPO for 1,000 updates using Adam with learning rate 3×10^{-5}, sampling 1,024 fresh on-policy trajectories per update.

D.2 Experiments with Qwen3-8B

We further examine βBPO on a larger model: Qwen3-8B. The training configurations are the same as the one in Section 4, but with rollout length 8192. We plot the results in Figure 5. βBPO consistently achieves better performance compared to the baseline methods. In particular, it improves the Pass@512 performance of the runner-up method from 74.9% to 81.4%, a +6.5 absolute percentage point increment. This further strengthens the efficacy of the principled diversity-preserving algorithmic design, inspired by the information projection of the β-Bregman divergences.

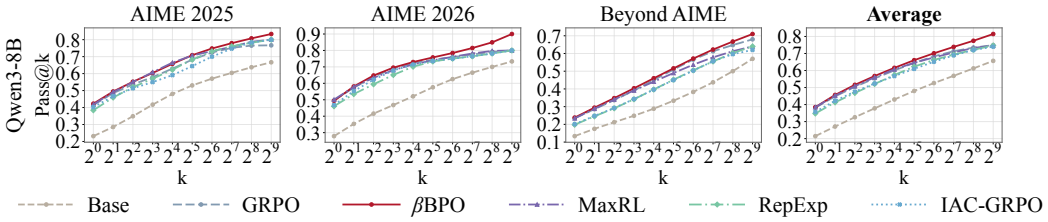
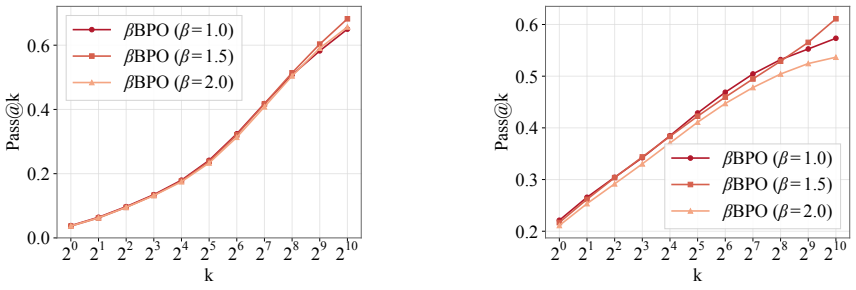


Figure 5: Pass@k performance of Qwen3-8B on the test benchmarks. βBPO achieves better diversity performance compared to the baseline methods.

D.3 Ablation on β

We carry out an ablation study on β for the LLM post-training experiments. We choose $\beta \in \{1.0, 1.5, 2.0\}$ and follow the same experimental setup as in Section 4. Figure 6 presents the evaluation results, which indicates $\beta \in \{1.0, 1.5\}$ show great practical gains across the tested models.

We remark that while the theoretical derivation in Lemma F.3 indicates larger β put more emphasis on the diversity when reaching the limit, but these policies converge to the limit at different speed and the training stops after finite time during practical implementations. Therefore, an overly large β may not yield a diversity-preserving policy. Our ablation study suggest $\beta \in \{1.0, 1.5\}$ are good choices in practice.



(a) Qwen2.5-Math-1.5B (b) Qwen3-4B

Figure 6: Ablation study on β of βBPO. Our algorithm demonstrates good performance with $\beta \in \{1.0, 1.5\}$.

D.4 Influence of the Entropy Regularizer

In Section 4, we follow the widely used RL configurations where the entropy regularizer is not enabled (He et al., 2025b). As the entropy regularizer can encourage exploration and helps to maintain the diversity in the output, we examine βBPO with it. Specifically, we add an entropy regularizer $-\lambda_{\text{ent}}\bar{H}_\theta(x, y)$, to the expected reward objective of each method, where the entropy coefficient λ_{ent} controls the intensity of the regularizer and is set to be 0.001 following the common practice, and the average entropy of a response y is defined as

$$\bar{H}_\theta(x, y) = \frac{1}{|y|}\sum_{s=1}^{|y|} H_\theta^{\text{tok}}(x, y_{<s}),$$
$$\text{where} \quad H_\theta^{\text{tok}}(x, y_{<s}) = -\sum_{v\in\mathcal{X}} \pi_\theta(v \mid x, y_{<s}) \log \pi_\theta(v \mid x, y_{<s}).$$

Figure 7 presents our result. Most methods show some improvement compared to the results without the entropy regularizer. In particular, βBPO demonstrate good performance compared to the baselines in Pass@256, and also better performance compared to the setup without the entropy regularizer.

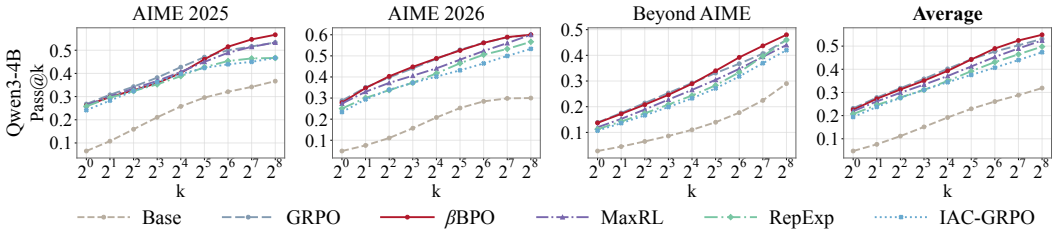


Figure 7: Pass@k performance of Qwen3-4B on the test benchmarks with entropy regularizer enabled. βBPO still demonstrate better improvement compared to the baselines.

D.5 Details of the Hidden State

For the LLM post-training, we construct the hidden state by the approach of RexExp (Tuyls et al., 2026).

We elaborate their approach here for completeness: given a query x, G responses $\{y_g\}_{g=1}^G$ are generated. We save $\tilde{h}_{s,g}$, the hidden state of the s-th token of y_g. As $\tilde{h}_{s,g}$ has large dimension, Tuyls et al. (2026) adopts a random matrix projection which turn $\tilde{h}_{s,g}$ into a smaller vector $\hat{h}_{s,g}$. The representation of the response y_g is the centered aggregated hidden state:

$$h_g = \frac{1}{|y_g|}\sum_{s=1}^{|y_g|} \hat{h}_{s,g} - \bar{h}, \quad \text{where} \quad \bar{h} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{|y_g|}\sum_{s=1}^{|y_g|}\hat{h}_{s,g}.$$

We remark that the quality of the representation can influence the final performance. There are other construction approaches, but we choose the one from Tuyls et al. (2026) as it shares similar goal and has proven to be effective. In addition, our experiments in Figure 3 and 5 suggest larger models produce more reliable representations with the same construction method, as the models demonstrate more evident improvement as the size grows.

E PROOFS OF MAIN RESULTS

E.1 PROOF OF PROPOSITION 3.1

Proof of Proposition 3.1. As we are considering the softmax policy, all actions have positive probability at finite time. By the definition of KL-divergence,

$$\mathrm{KL}(\pi\|\pi^0) = \sum_{i\in\mathcal{A}} \pi_i \log \frac{\pi_i}{\pi_i^0} = \sum_{i\in\mathcal{C}} \pi_i \log \frac{\pi_i}{\pi_i^0} + \sum_{i\in\mathcal{B}} \pi_i \log \frac{\pi_i}{\pi_i^0}.$$

For the correct set $\mathcal{C}$,

$$\sum_{i\in\mathcal{C}} \pi_i \log \frac{\pi_i}{\pi_i^0} = \sum_{i\in\mathcal{C}} q\rho_i \log \frac{q\rho_i}{q_0\rho_i^0} = q \log \frac{q}{q_0} + q\,\mathrm{KL}(\rho\|\rho^0).$$

Similarly on $\mathcal{B}$, we have

$$\sum_{i\in\mathcal{B}} \pi_i \log \frac{\pi_i}{\pi_i^0} = (1-q) \log \frac{1-q}{1-q_0} + (1-q)\,\mathrm{KL}(\sigma\|\sigma^0).$$

Together, it yields

$$\mathrm{KL}(\pi\|\pi^0) = \mathrm{kl}(q\|q_0) + q\,\mathrm{KL}(\rho\|\rho^0) + (1-q)\,\mathrm{KL}(\sigma\|\sigma^0).$$

This indicates that for any policy $\pi \in \{\pi \in \Delta_K : \pi(\mathcal{C}) = q\}$, the one that attains the minimum divergence $\mathrm{KL}(\pi\|\pi^0)$ needs to satisfy $\rho_i = \rho_i^0$ and $\sigma_j = \sigma_j^0$ for $i \in \mathcal{C}, j \in \mathcal{B}$. Consequently, for a fixed $q \in (0,1)$, $\pi^{\mathrm{KL},q} = \mathrm{argmin}_{\pi\in\Delta_K:\,\pi(\mathcal{C})=q} \mathrm{KL}(\pi\|\pi^0)$ is unique and satisfies

$$\pi_i^{\mathrm{KL},q} = \frac{q}{q_0}\pi_i^0 \quad \text{for } i \in \mathcal{C}, \qquad \pi_j^{\mathrm{KL},q} = \frac{1-q}{1-q_0}\pi_j^0 \quad \text{for } j \in \mathcal{B}.$$

The same identity for zero coordinates follows by continuity under the convention above. □

E.2 PROOF OF LEMMA 3.2

Proof of Lemma 3.2. In the binary reward model, according to Lemma F.4, we obtain

$$\dot{z}_i(\tau) = \pi_i(\tau)(1-q(\tau)) \quad \text{for } i \in \mathcal{C}, \qquad \dot{z}_j(\tau) = -\pi_j(\tau)q(\tau) \quad \text{for } j \in \mathcal{B}.$$

Under the gradient flow $\dot{z}(\tau) = \xi(z(\tau))$, for any two $i, j \in \mathcal{C}$,

$$\frac{d}{d\tau} \log \frac{\pi_i(\tau)}{\pi_j(\tau)} = \dot{z}_i(\tau) - \dot{z}_j(\tau) = (1-q(\tau))(\pi_i(\tau) - \pi_j(\tau)).$$

Similarly, for two $i, j \in \mathcal{B}$,

$$\frac{d}{d\tau} \log \frac{\pi_i(\tau)}{\pi_j(\tau)} = \dot{z}_i(\tau) - \dot{z}_j(\tau) = -q(\tau)(\pi_i(\tau) - \pi_j(\tau)).$$

Because (19) is ordinary gradient ascent on J,

$$\frac{d}{d\tau} J(\pi(\tau)) = \nabla_z J(\pi(\tau))^\top \dot{z}(\tau) = \|\nabla_z J(\pi(\tau))\|_2^2 > 0$$

for a nondegenerate binary problem at an interior policy. Hence the GRPO trajectory can be parameterized by its strictly increasing matched accuracy $q(\tau)$.

Therefore, if there exist a correct action pair (i, j) with unequal initial probabilities, i.e., $\pi_i(0) > \pi_j(0)$, we have

$$\frac{d}{d\tau} \log \frac{\pi_i(\tau)}{\pi_j(\tau)} = (1-q(\tau))(\pi_i(\tau) - \pi_j(\tau)) > 0.$$

We claim that $\frac{d}{d\tau} \log \frac{\pi_i(\tau)}{\pi_j(\tau)} > 0$ for all τ, as long as $q(\tau) \in (0,1)$ and the policy is still an interior policy. If not, denote $\delta_{ij}(\tau) = \log \frac{\pi_i(\tau)}{\pi_j(\tau)}$ and $\tau_\star = \inf\{\tau' \geq 0 : \delta_{ij}(\tau') = 0\}$. Because $\delta_{ij}(0) > 0$

and by the continuity of the softmax policy, $\tau_\star > 0$. If $\tau_\star < \infty$, then $\delta_{ij}(\tau) > 0$ for all $\tau \in (0, \tau_\star)$ and

$$\delta_{ij}(\tau_\star) - \delta_{ij}(0) = \int_0^{\tau_\star} \frac{d}{d\tau'} \log \frac{\pi_i(\tau')}{\pi_j(\tau')} \mathrm{d}\tau' = \int_0^{\tau_\star} (1 - q(\tau'))\big(\pi_i(\tau') - \pi_j(\tau')\big)\mathrm{d}\tau' > 0.$$

This contradicts with $\delta_{ij}(\tau_\star) = 0 < \delta_{ij}(0)$. Therefore, $\delta_{ij}(\tau) > 0$ for all τ and therefore, $\delta_{ij}(\tau) > \delta_{ij}(0)$ for $\tau > 0$, i.e.,

$$\frac{\pi_i(\tau)}{\pi_j(\tau)} > \frac{\pi_i(0)}{\pi_j(0)}.$$

As the ratio between two correct actions are not maintained during GRPO training, thus, the policy trained via GRPO is not the information projection induced by the KL-divergence by Proposition 3.1: $\pi(\tau) \notin \operatorname{argmin}_{\pi \in \mathcal{M}_{J(\pi(\tau))}} \mathrm{KL}(\pi \| \pi^0)$.

□

E.3 Proof of Theorem 3.3

Proof of Theorem 3.3. Fix an open interval on which $\mathcal{S}_\beta(\tau) = \mathcal{S}$ is constant. Recall that $\nu_\beta(\tau)$ denote the multiplier of the simplex-normalization constraint in the active-coordinate KKT system. Thus, by (12)

$$\chi_\beta(\pi_{\beta,i}(\tau)) - \chi_\beta(\pi_i^0) = \tau r_i - \nu_\beta(\tau), \qquad i \in \mathcal{S}.$$

Differentiating gives

$$\dot{\pi}_{\beta,i}(\tau) = \pi_{\beta,i}(\tau)^{2-\beta}\big(r_i - \dot{\nu}_\beta(\tau)\big).$$

Since $\sum_{i\in\mathcal{S}} \dot{\pi}_{\beta,i}(\tau) = 0$,

$$\dot{\nu}_\beta(\tau) = \frac{\sum_{i\in\mathcal{S}} \pi_{\beta,i}(\tau)^{2-\beta} r_i}{\sum_{i\in\mathcal{S}} \pi_{\beta,i}(\tau)^{2-\beta}} = b_\beta(\pi_\beta(\tau)). \tag{9}$$

For $z_{\beta,i}$, note that the face-restricted softmax Jacobian on the active coordinates satisfies

$$\frac{\partial \pi_i}{\partial z_{\beta,j}} = \pi_i(\mathbb{1}\{i = j\} - \pi_j), \quad \text{for } i, j \in \mathcal{S}_\beta(\tau).$$

Therefore, for the logit of the active coordinate, we have

$$\dot{\pi}_{\beta,i}(\tau) = \pi_{\beta,i}(\tau)\big(\dot{z}_{\beta,i}(\tau) - \sum_{j\in\mathcal{S}_\beta(\tau)} \pi_{\beta,j}(\tau)\dot{z}_{\beta,j}(\tau)\big) = (\pi_{\beta,i}(\tau))^{2-\beta}\big(r_i - b_\beta(\pi_\beta(\tau))\big)$$

which further yields

$$\dot{z}_{\beta,i}(\tau) - \sum_{j\in\mathcal{S}_\beta(\tau)} \pi_{\beta,j}(\tau)\dot{z}_{\beta,j}(\tau) = (\pi_{\beta,i}(\tau))^{1-\beta}\big(r_i - b_\beta(\pi_\beta(\tau))\big)$$

As adding the same scalar $\gamma(\tau)$ to every active $\dot{z}_{\beta,i}(\tau)$ does not change the softmax policy velocity, the general logit representative is

$$\dot{z}_{\beta,i}(\tau) = (\pi_{\beta,i}(\tau))^{1-\beta}\big(r_i - b_\beta(\pi_\beta(\tau))\big) + \gamma(\tau).$$

We choose the representative $\gamma(\tau) = 0$, which gives the final $z_{\beta,i}$ in (7).

For the equivalent tangent-space characterization, the first-order condition on each active coordinate is

$$r_i - (\pi_{\beta,i}(\tau))^{\beta-2} v_i - \lambda_{\tan} = 0, \qquad i \in \mathcal{S}_\beta(\tau).$$

Together with $\sum_{i\in\mathcal{S}_\beta(\tau)} v_i = 0$, this gives

$$\lambda_{\tan} = b_\beta(\pi_\beta(\tau)), \qquad v_i = (\pi_{\beta,i}(\tau))^{2-\beta}\big(r_i - b_\beta(\pi_\beta(\tau))\big) = \dot{\pi}_{\beta,i}(\tau),$$

which proves (8).

□

E.4 Proof of Proposition 3.4

Proof of Proposition 3.4. The optimality in (6) implies that for every $\pi \in \Delta_K$,

$$\mathrm{D}_\beta(\pi\|\pi^0) - \mathrm{D}_\beta(\pi_\beta(\tau)\|\pi^0) \geq \tau\big[J(\pi) - J(\pi_\beta(\tau))\big].$$

On the equality, the right-hand side is zero. On the superlevel set, it is nonnegative because $\tau \geq 0$. Therefore, the projection flow gives the information projection on the set $\{\pi : J(\pi) \geq J(\pi_\beta(\tau))\}$, which achieves the minimum β-Bregman divergence with respect to the base policy. Uniqueness follows from the strict convexity of $\mathrm{D}_\beta(\cdot\|\pi^0)$ on the convex reward slice and superlevel set. □

E.5 Proof of Proposition 3.5

Proof of Proposition 3.5. Fix an iteration t. Grouping all sample occurrences over $i \in \widehat{\mathcal{S}}_t$ in the baseline gives

$$\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta} r_{t,g} = G \sum_{i \in \widehat{\mathcal{S}}_t} \widehat{\pi}_{t,i}^{2-\beta} r_i,$$

and similarly,

$$\sum_{g=1}^{G} \widehat{\pi}_{t,a_{t,g}}^{1-\beta} = G \sum_{i \in \widehat{\mathcal{S}}_t} \widehat{\pi}_{t,i}^{2-\beta}.$$

This yields the baseline $\widehat{b}_{\beta,t}$. Grouping the logit increments instead gives

$$\frac{1}{G} \sum_{g:a_{t,g}=i} \widehat{\pi}_{t,i}^{-\beta}(r_i - \widehat{b}_{\beta,t}) = \widehat{\pi}_{t,i}^{1-\beta}(r_i - \widehat{b}_{\beta,t}),$$

which proves the statement of $\widehat{g}_{\beta,t,i}$. These updates precisely follow the update in Theorem 3.3. □

F Supporting Lemmas

F.1 The effect of the standard deviation

In the single-context tabular model and binary reward case, each action represents one complete response. We show that the standard-deviation normalization used by vanilla GRPO changes only the speed of the expected gradient flow, rather than its trajectory. As the objective is equivalent to Dr. GRPO when the standard deviation is absent in the single-context tabular model, this comparison is also a comparison between GRPO and Dr. GRPO. We formalize this relationship below.

Lemma F.1. *Consider the single-context tabular model with nonempty correct and incorrect action sets $\mathcal{C} = \{i : r_i = 1\}, \mathcal{B} = \{i : r_i = 0\}$. At iteration t, let*

$$q_t = \sum_{i \in \mathcal{C}} \pi_{t,i} \in (0,1), \qquad N_t = \sum_{g=1}^{G} r_{t,g}.$$

Then $N_t \mid z_t \sim \mathrm{Binomial}(G, q_t)$. Denote the fresh-gradient estimator of Dr. GRPO as

$$\widehat{\xi}_t^{\mathrm{Dr}} = \frac{1}{G} \sum_{g=1}^{G} (r_{t,g} - \widehat{b}_t) \nabla_z \log \pi_{a_{t,g}}(z)\big|_{z=z_t}, \qquad \widehat{b}_t = \frac{1}{G} \sum_{g=1}^{G} r_{t,g}.$$

Define the corresponding vanilla-GRPO estimator by

$$\widehat{\xi}_t^{\mathrm{GRPO}} = \begin{cases} \dfrac{1}{G} \displaystyle\sum_{g=1}^{G} \frac{r_{t,g} - \widehat{b}_t}{\widehat{s}_t} \nabla_z \log \pi_{a_{t,g}}(z)\big|_{z=z_t}, & \widehat{s}_t > 0, \\ 0, & \widehat{s}_t = 0, \end{cases}$$

where

$$\widehat{s}_t = \left(\frac{1}{G} \sum_{g=1}^{G} (r_{t,g} - \widehat{b}_t)^2 \right)^{1/2}.$$

Denote the population mean-centered score gradient as

$$\xi(z) = \mathbb{E}_{a\sim\pi(z)}\left[\big(r_a - J(\pi(z))\big)\nabla_z \log \pi_a(z)\right]$$

Then

$$\mathbb{E}\Big[\widehat{\xi}_t^{\mathrm{Dr}} \mid z_t\Big] = \frac{G-1}{G}\,\xi(z_t), \tag{10}$$

$$\mathbb{E}\Big[\widehat{\xi}_t^{\mathrm{GRPO}} \mid z_t\Big] = \kappa_G(q_t)\,\xi(z_t), \tag{11}$$

where

$$\kappa_G(q) = \frac{\mathbb{E}_{N\sim\mathrm{Binomial}(G,q)}\left[\sqrt{N(G-N)}\,\mathbf{1}\{0<N<G\}\right]}{Gq(1-q)} > 0.$$

Consequently, the expected gradient flows of vanilla GRPO and mean-centered Dr. GRPO trace the same trajectories in policy space, up to a strictly increasing, state-dependent reparameterization of time.

Proof. For Dr. GRPO, exchangeability of the i.i.d. samples gives

$$\mathbb{E}[\widehat{\xi}_t^{\mathrm{Dr}} \mid z_t] = \mathbb{E}\left[(r_{t,1} - \widehat{b}_t)\nabla_z \log \pi_{a_{t,1}}(z_t)\,\middle|\, z_t\right].$$

The terms in $\widehat{b}_t$ involving $r_{t,2}, \ldots, r_{t,G}$ vanish because they are independent of $a_{t,1}$ and

$$\mathbb{E}_{a\sim\pi_t}[\nabla_z \log \pi_a(z_t)] = 0.$$

It follows that

$$\mathbb{E}[\widehat{\xi}_t^{\mathrm{Dr}} \mid z_t] = \frac{G-1}{G}\mathbb{E}_{a\sim\pi_t}\left[r_a \nabla_z \log \pi_a(z_t)\right] = \frac{G-1}{G}\xi(z_t),$$

which proves (10).

We next consider vanilla GRPO. Conditional on $N_t = n$ with $0 < n < G$, binary rewards give

$$\widehat{b}_t = \frac{n}{G}, \qquad \widehat{s}_t = \frac{\sqrt{n(G-n)}}{G}.$$

Hence the normalized advantage equals

$$\sqrt{\frac{G-n}{n}} \quad \text{for a correct action,} \qquad -\sqrt{\frac{n}{G-n}} \quad \text{for an incorrect action.}$$

The normalized advantages sum to zero. Therefore, using

$$\nabla_{z_i} \log \pi_a(z_t) = \mathbf{1}\{a = i\} - \pi_{t,i},$$

the common term involving $-\pi_{t,i}$ cancels.

For $i \in \mathcal{C}$, let

$$N_{t,i} = \sum_{g=1}^{G} \mathbf{1}\{a_{t,g} = i\}.$$

Conditional on $N_t = n$,

$$\mathbb{E}[N_{t,i} \mid N_t = n, z_t] = n\frac{\pi_{t,i}}{q_t}.$$

It follows that

$$\mathbb{E}[\widehat{\xi}_{t,i}^{\mathrm{GRPO}} \mid N_t = n, z_t] = \frac{\pi_{t,i}}{Gq_t}\sqrt{n(G-n)} = \frac{\sqrt{n(G-n)}}{Gq_t(1-q_t)}\,\xi_i(z_t),$$

where $\xi_i(z_t) = \pi_{t,i}(1-q_t)$ for $i \in \mathcal{C}$. Similarly, for $j \in \mathcal{B}$,

$$\mathbb{E}[\widehat{\xi}_{t,j}^{\mathrm{GRPO}} \mid N_t = n, z_t] = -\frac{\pi_{t,j}}{G(1-q_t)}\sqrt{n(G-n)} = \frac{\sqrt{n(G-n)}}{Gq_t(1-q_t)}\,\xi_j(z_t),$$

where $\xi_j(z_t) = -\pi_{t,j}q_t$ for $j \in \mathcal{B}$. When $N_t \in \{0, G\}$, one has $\widehat{s}_t = 0$, and the estimator is zero by convention. Averaging over $N_t \mid z_t \sim \mathrm{Binomial}(G, q_t)$ proves (11). Since $q_t \in (0,1)$ and $G \geq 2$, the event $0 < N_t < G$ has positive probability, so $\kappa_G(q_t) > 0$.

Thus, the two expected vector fields are positive scalar multiples of the same population field ξ. Multiplication by a positive, state-dependent scalar changes only the speed along a trajectory, not the trajectory itself. □

F.2 Closed-form β-Projection Path

Lemma F.2 (Closed-form β-Projection Path)**.** *For every $\beta \in \mathbb{R}$ and $\tau \geq 0$, the maximizer in (6) exists and is unique. In the unified dual notation it is given for each $i \in [K]$ by*

$$\pi_{\beta,i}(\tau) = \mathcal{E}_\beta\big(\chi_\beta(\pi_i^0) + \tau r_i - \nu_\beta(\tau)\big)\,,$$

where $\nu_\beta(\tau)$ is the unique scalar that normalizes the distribution and

$$\chi_\beta(x) = \begin{cases} \dfrac{x^{\beta-1}-1}{\beta-1}, & \beta \neq 1, \\ \log x, & \beta = 1, \end{cases} \qquad \mathcal{E}_\beta(y) = \begin{cases} [1+(\beta-1)y]_+^{1/(\beta-1)}, & \beta > 1, \\ e^y, & \beta = 1, \\ [1+(\beta-1)y]^{1/(\beta-1)}, & \beta < 1, \end{cases}$$

The domain of $\chi_\beta(x)$ is $x > 0$. Moreover, $[x]_+ = \max\{x, 0\}$. For $\beta < 1$, $\mathcal{E}_\beta(y)$ is real-valued on $\{y : 1 + (\beta-1)y > 0\}$. Specifically, for $\beta \neq 1$,

$$\pi_{\beta,i}(\tau) = \begin{cases} \big[(\pi_i^0)^{\beta-1} + (\beta-1)(\tau r_i - \nu_\beta(\tau))\big]_+^{1/(\beta-1)}, & \beta > 1, \\ \big[(\pi_i^0)^{\beta-1} + (\beta-1)(\tau r_i - \nu_\beta(\tau))\big]^{1/(\beta-1)}, & \beta < 1, \end{cases} \tag{12}$$

where every bracket in the $\beta < 1$ branch is strictly positive. At $\beta = 1$,

$$\pi_{1,i}(\tau) = \frac{\pi_i^0 e^{\tau r_i}}{\sum_j \pi_j^0 e^{\tau r_j}}. \tag{13}$$

Proof of Lemma F.2. Recall the projection path:

$$\pi_\beta(\tau) = \operatorname*{argmax}_{\pi \in \Delta_K} \big\{\tau J(\pi) - \mathrm{D}_\beta(\pi \| \pi^0)\big\}.$$

For $\beta > 0$, the divergence is finite on the compact probability simplex Δ_K. For $\beta \leq 0$, it tends to $+\infty$ at the boundary, so the objective tends to $-\infty$ there and its upper level sets are compact subsets of the interior. By the continuity of the objective function and the set is compact, the existence of the maximum point can be obtained.

By the definition of β-Bregman divergence in (1), we have $\nabla_\pi^2 \mathrm{D}_\beta(\pi \| \pi') \succ 0$ on the interior of its effective domain (or on the active coordinates), so the objective function is strictly concave on its effective domain, hence the maximizer is unique.

- On every positive coordinate, the KKT condition gives

$$\chi_\beta(\pi_{\beta,i}(\tau)) - \chi_\beta(\pi_i^0) = \tau r_i - \nu_\beta(\tau). \tag{14}$$

- For $\beta > 1$, complementary slackness gives the positive part in (12). Its normalizing sum is continuous and nonincreasing in $\nu_\beta(\tau)$, strictly decreasing while positive, and ranges from $+\infty$ to zero.
- For $\beta < 1$, the divergence regularizer becomes steep when $0 < \beta < 1$ and becomes a barrier that forces every coordinate to be positive when $\beta \leq 0$. Denote

$$\nu_{\beta,\min}(\tau) = \max_i \left\{\tau r_i + \frac{(\pi_i^0)^{\beta-1}}{\beta-1}\right\}.$$

  For $\nu_\beta(\tau) > \nu_{\beta,\min}(\tau)$ all brackets in (12) are positive, and it decreases continuously from $+\infty$ as $\nu_\beta(\tau) \downarrow \nu_{\beta,\min}(\tau)$ to zero as $\nu_\beta(\tau) \to \infty$. This proves the existence and uniqueness of the normalizer in every non-KL case.
- At $\beta = 1$, direct normalization of (14) gives (13).

□

F.3 Properties of β-Bregman Information Projection

Lemma F.3 (Reward-optimal limit for all real orders)**.** *Let $r_\star = \max_{i\in[K]} r_i$ and $\mathcal{A}_\star = \{i \in [K] : r_i = r_\star\}$. For $\beta \neq 1$, let $a^\star_\beta \geq 0$ be the unique solution of*

$$\sum_{i\in\mathcal{A}_\star} \left[(\pi^0_i)^{\beta-1} + (\beta-1)a^\star_\beta\right]^{1/(\beta-1)} = 1, \tag{15}$$

where the brackets are required to be strictly positive when $\beta < 1$. Then, for every $\beta \in \mathbb{R}$, $\pi_\beta(\tau) \to \pi^\infty_\beta$ with

$$\pi^\infty_{\beta,i} = \begin{cases} \left[(\pi^0_i)^{\beta-1} + (\beta-1)a^\star_\beta\right]^{1/(\beta-1)}, & i \in \mathcal{A}_\star,\ \beta \neq 1, \\ \pi^0_i/\pi^0(\mathcal{A}_\star), & i \in \mathcal{A}_\star,\ \beta = 1, \\ 0, & i \notin \mathcal{A}_\star. \end{cases} \tag{16}$$

- *If $\mathcal{A}_\star = \mathcal{A}$, then $a^\star_\beta = 0$ for $\beta \neq 1$ and $\pi_\beta(\tau) = \pi^0$ for all β and τ.*
- *If $\mathcal{A}_\star \neq \mathcal{A}$ and $\beta > 1$, the limit is reached at*

$$\tau_\star(\beta) = \max_{j\notin\mathcal{A}_\star} \frac{(\pi^0_j)^{\beta-1} + (\beta-1)a^\star_\beta}{(\beta-1)(r_\star - r_j)},$$

and $\pi_\beta(\tau) = \pi^\infty_\beta$ for $\tau \geq \tau_\star(\beta)$.

- *If $\mathcal{A}_\star \neq \mathcal{A}$ and $\beta \leq 1$, then $\mathcal{S}_\beta(\tau) = \mathcal{A}$ at every finite τ, and the limit is attained only as $\tau \to \infty$. For $\beta > 0$, the limit is the unique minimum-divergence reward-optimal policy:*

$$\pi^\infty_\beta = \operatorname*{argmin}_{\pi\in\mathcal{M}_{r_\star}} \mathrm{D}_\beta(\pi\|\pi^0). \tag{17}$$

For $\beta \leq 0$ and $\mathcal{A}_\star \neq \mathcal{A}$, $\mathrm{D}_\beta(\pi\|\pi^0) = +\infty$ for every $\pi \in \mathcal{M}_{r_\star}$.

Proof. **We firstly prove** (15)**.** Because $r_i \leq r_\star$ for every i, so $\mathcal{M}_{r_\star} = \{\pi \in \Delta_K : \mathrm{supp}(\pi) \subseteq \mathcal{A}_\star\}$. For $\beta \neq 1$, denote $\vartheta = \beta - 1$ and define

$$G_\beta(a) = \sum_{i\in\mathcal{A}_\star} \left[(\pi^0_i)^\vartheta + \vartheta a\right]^{1/\vartheta}.$$

Its maximal domain with $a \geq 0$ is $[0,\infty)$ when $\vartheta > 0$, and $[0, \bar{a}_\beta)$ when $\vartheta < 0$, where $\bar{a}_\beta = \min_{i\in\mathcal{A}_\star} \frac{(\pi^0_i)^\vartheta}{-\vartheta}$. On either domain,

$$G'_\beta(a) = \sum_{i\in\mathcal{A}_\star} \left[(\pi^0_i)^\vartheta + \vartheta a\right]^{1/\vartheta - 1} > 0.$$

Moreover, $G_\beta(0) = \pi^0(\mathcal{A}_\star) \leq 1$, and $G_\beta(a)$ tends to $+\infty$ at the right endpoint of its domain. Hence, by the continuity of $G_\beta(a)$, (15) has a unique solution.

If $\mathcal{A}_\star = \mathcal{A}$, then $J(\pi) = r_\star$ on Δ_K, so (6) uniquely selects π^0 for every τ. In the following dicusions, we assume $\mathcal{A}_\star \neq \mathcal{A}$.

We then prove (16) • First suppose $\beta > 1$ and set $\vartheta = \beta - 1 > 0$. Consider the candidate normalizer

$$\bar{\nu}_\beta(\tau) = \tau r_\star - a^\star_\beta.$$

With this choice, the close-form solution in (12) is

$$(\pi^0_i)^\vartheta + \vartheta a^\star_\beta \quad \text{for } i \in \mathcal{A}_\star \quad \text{and} \quad B_j(\tau) = (\pi^0_j)^\vartheta + \vartheta a^\star_\beta - \vartheta\tau(r_\star - r_j) \quad \text{for } j \notin \mathcal{A}_\star.$$

where $B_j(\tau) \leq 0$ when $\tau \geq \tau_\star(\beta)$. For such τ, the positive-part branch and (15) show that $\bar{\nu}_\beta(\tau)$ normalizes the policy, and uniqueness in Lemma F.2 gives $\pi_\beta(\tau) = \pi^\infty_\beta$. Conversely, if $\tau < \tau_\star(\beta)$, equality with π^∞_β would force the same normalizer on $\mathcal{A}_\star$, while some $B_j(\tau) > 0$ would give positive suboptimal mass. Thus $\tau_\star(\beta)$ is the first attainment time and (16) holds in this regime.

• At $\beta = 1$, rewrite (13) as

$$\pi_{1,i}(\tau) = \frac{\pi_i^0 e^{-\tau(r_\star - r_i)}}{\sum_k \pi_k^0 e^{-\tau(r_\star - r_k)}}.$$

Letting $\tau \to \infty$ gives the KL branch of (16). Every coordinate is positive at finite τ, so the limit cannot be attained at finite time.

• It remains to consider $\beta < 1$. Set $\vartheta = \beta - 1 < 0$. The closed-form solution in (12) is strictly positive, so $\mathcal{S}_\beta(\tau) = \mathcal{A}$ at every finite τ. Define

$$\Delta_\beta(\tau) = \vartheta\big(\tau r_\star - \nu_\beta(\tau)\big).$$

The KKT equations of the regularized projection path in (6) give

$$\begin{cases} \pi_{\beta,i}(\tau)^\vartheta = (\pi_i^0)^\vartheta + \Delta_\beta(\tau), \ i \in \mathcal{A}_\star, \\ \pi_{\beta,j}(\tau)^\vartheta = (\pi_j^0)^\vartheta + \vartheta\tau(r_j - r_\star) + \Delta_\beta(\tau), \ j \notin \mathcal{A}_\star. \end{cases}$$

Because $\beta < 1$ gives full support at every finite τ, (9) applies. Therefore,

$$\dot{\Delta}_\beta(\tau) = \vartheta\big(r_\star - \dot{\nu}_\beta(\tau)\big) = \vartheta\big(r_\star - b_\beta(\pi_\beta(\tau))\big) < 0.$$

The strict inequality follows because the finite-time policy has positive suboptimal mass. Moreover,

$$\sum_{i \in \mathcal{A}_\star} \left[(\pi_i^0)^\vartheta + \Delta_\beta(\tau)\right]^{1/\vartheta} < 1 = \sum_{i \in \mathcal{A}_\star} \left[(\pi_i^0)^\vartheta + \vartheta a_\beta^\star\right]^{1/\vartheta}.$$

The displayed sum is strictly decreasing in its common additive argument, so $\Delta_\beta(\tau) > \vartheta a_\beta^\star$. Thus, $\Delta_\beta(\tau)$ is decreasing and bounded below, and has a finite limit Δ_∞. For $j \notin \mathcal{A}_\star$, $\vartheta\tau(r_j - r_\star) \to +\infty$, so $\pi_{\beta,j}(\tau) \to 0$. Because the suboptimal mass vanishes, passing to the limit in $\sum_{k \in \mathcal{A}} \pi_{\beta,k}(\tau) = 1$ gives

$$1 = \sum_{i \in \mathcal{A}_\star} \left[(\pi_i^0)^\vartheta + \Delta_\infty\right]^{1/\vartheta}.$$

Comparing this identity with (15) and using uniqueness of its level-one solution yields $\Delta_\infty = \vartheta a_\beta^\star$. This proves (16). The path has full support at every finite time so the limit is attained only as $\tau \to \infty$.

Lastly, we prove (17) The preceding arguments establish the limit and when it is attained. We now further characterize its variational meaning.

For $\beta > 0$, consider the restriction of $\mathrm{D}_\beta(\cdot\|\pi^0)$ to the relative interior of $\mathcal{M}_{r_\star}$. Differentiating the divergence with respect to its first argument gives, for every $i \in \mathcal{A}_\star$,

$$\frac{\partial}{\partial \pi_i} \mathrm{D}_\beta(\pi\|\pi^0)\bigg|_{\pi = \pi_\beta^\infty} = \begin{cases} \dfrac{(\pi_{\beta,i}^\infty)^{\beta-1} - (\pi_i^0)^{\beta-1}}{\beta - 1} = a_\beta^\star, & \beta \neq 1, \\ \log \dfrac{\pi_{1,i}^\infty}{\pi_i^0} = -\log \pi^0(\mathcal{A}_\star), & \beta = 1. \end{cases}$$

Thus there exists a scalar c_β, independent of i, such that

$$\frac{\partial}{\partial \pi_i} \mathrm{D}_\beta(\pi\|\pi^0)\bigg|_{\pi = \pi_\beta^\infty} = c_\beta, \qquad i \in \mathcal{A}_\star.$$

The tangent space of the optimal face is

$$\mathcal{M}_{r_\star}^\perp = \left\{ v \in \mathbb{R}^K : v_j = 0 \text{ for } j \notin \mathcal{A}_\star, \quad \sum_{i \in \mathcal{A}_\star} v_i = 0 \right\}.$$

Therefore, for every feasible tangent direction $v \in \mathcal{M}_{r_\star}^\perp$,

$$\begin{aligned} \mathrm{D}'_\beta(\pi_\beta^\infty\|\pi^0; v) &= \sum_{i \in \mathcal{A}_\star} \frac{\partial \mathrm{D}_\beta(\pi\|\pi^0)}{\partial \pi_i}\bigg|_{\pi = \pi_\beta^\infty} v_i \\ &= c_\beta \sum_{i \in \mathcal{A}_\star} v_i = 0. \end{aligned}$$

Hence π_β^∞ satisfies the first-order optimality condition on $\mathcal{M}_{r_\star}$. Since $\mathrm{D}_\beta(\cdot\|\pi^0)$ is strictly convex on this face for $\beta > 0$, π_β^∞ is its unique minimizer.

For $\beta \leq 0$, every policy supported on $\mathcal{M}_{r_\star}$ has a zero coordinate against the full-support reference π^0, and therefore has infinite divergence. □

F.4 GRADIENT OF GRPO

In the gradient flow analysis, the update is small at each step, i.e., the learning $\eta \to 0$. Therefore, the importance ratio $w_{t,g} \to 1$ and the objective function in (2) can be simplified as $\widehat{J}_t(z) = \frac{1}{G}\sum_{g=1}^{G} w_{t,g}(z)A_{t,g}$. Hence, the gradient estimator $\hat{\xi}_t$ and the population mean-centered score gradient $\xi(z)$ are given respectively by

$$\hat{\xi}_t = \frac{1}{G}\sum_{g=1}^{G}(r_{a_{t,g}} - \widehat{b}_t)\nabla_z \log \pi_{a_{t,g}}(z)\big|_{z=z_t} \text{ and } \xi(z) = \mathbb{E}_{a\sim\pi(z)}\big[(r_a - J(\pi(z)))\nabla_z \log \pi_a(z)\big]. \tag{18}$$

We show in Lemma F.4 that the expected finite-group gradient estimator $\mathbb{E}[\hat{\xi}_t]$ has exactly the population direction $\xi(z)$, up to a positive common factor.

Lemma F.4 (Gradient of GRPO)**.** *For every action i,*

$$\xi_i(z) = \pi_i(z)(r_i - J(\pi(z))). \tag{19}$$

Moreover, for deterministic rewards and $G \geq 2$, the expected finite-group GRPO direction satisfies

$$\mathbb{E}[\hat{\xi}_{t,i}] = \frac{G-1}{G}\pi_{t,i}(r_i - J(\pi_t)) = \frac{G-1}{G}\xi_i(z_t).$$

Proof of Lemma F.4. We suppress the notation z in the policy parameterization to have a cleaner presentation in the proof. We first note that the score function for the softmax policy satisfies

$$\frac{\partial}{\partial z_i}\log \pi_a = \mathbb{1}\{a = i\} - \pi_i.$$

Therefore, the gradient of the expected reward becomes

$$\begin{aligned}\xi_i &= \sum_a \pi_a(r_a - J(\pi))(\mathbb{1}\{a=i\} - \pi_i)\\ &= \pi_i(r_i - J(\pi)) - \pi_i\sum_a \pi_a(r_a - J(\pi))\\ &= \pi_i(r_i - J(\pi)),\end{aligned}$$

where $\pi_i\sum_a \pi_a(r_a - J(\pi)) = 0$ by the definition of $J(\pi)$.

For $i \in [K]$, denote $N_{t,i} = \sum_g \mathbb{1}\{a_{t,g} = i\}$. From (18), we have

$$\begin{aligned}\hat{\xi}_{t,i} &= \frac{1}{G}\sum_{g=1}^{G}(r_{a_{t,g}} - \widehat{b}_t)(\nabla_z \log \pi_{a_{t,g}})_i\\ &= \frac{1}{G}\sum_{g=1}^{G}(r_{a_{t,g}} - \widehat{b}_t)(\mathbb{1}\{a_{t,g} = i\} - \pi_i)\\ &= \frac{N_{t,i}}{G}(r_i - \widehat{b}_t)\end{aligned}$$

Because $\mathbb{E}\big[\frac{N_{t,i}}{G}\big] = \pi_i$ and

$$\begin{aligned}\mathbb{E}\big[\frac{N_{t,i}}{G}\widehat{b}_t\big] &= \frac{1}{G^2}\sum_{g=1}^{G}\sum_{g'=1}^{G}\mathbb{E}[\mathbb{1}\{a_{t,g} = i\}r_{a_{t,g'}}]\\ &= \frac{1}{G^2}\big(G\pi_i r_i + G(G-1)\pi_i J(\pi)\big)\\ &= \frac{1}{G}\pi_i r_i + \frac{G-1}{G}\pi_i J(\pi),\end{aligned}$$

therefore, we have

$$\mathbb{E}[\hat{\xi}_{t,i}] = \pi_i r_i - \frac{1}{G}\pi_i r_i - \frac{G-1}{G}\pi_i J(\pi) = \frac{G-1}{G}\xi_i.$$

□

F.5 Convergence to the Expected-flow

Theorem F.5 (Convergence to the expected-flow). *Fix $\beta \in \mathbb{R}$ and an interior policy π, and suppose $\pi_{\min} = \min_i \pi_i > 0$ and $\max_i |r_i| \le R$. For each $G \in \mathbb{N}$, draw $a_1, \dots, a_G \overset{\text{i.i.d.}}{\sim} \pi$ and define $N_i^{(G)} = \sum_{g=1}^{G} \mathbb{1}\{a_g = i\}$, $\widehat{\pi}_i^{(G)} = N_i^{(G)}/G$, and $\widehat{\mathcal{S}}^{(G)} = \{i : \widehat{\pi}_i^{(G)} > 0\}$. Set*

$$\widehat{b}_\beta^{(G)} = \frac{\sum_{i \in \widehat{\mathcal{S}}^{(G)}} (\widehat{\pi}_i^{(G)})^{2-\beta} r_i}{\sum_{i \in \widehat{\mathcal{S}}^{(G)}} (\widehat{\pi}_i^{(G)})^{2-\beta}}, \qquad \widehat{g}_{\beta,i}^{(G)} = \begin{cases} (\widehat{\pi}_i^{(G)})^{1-\beta}(r_i - \widehat{b}_\beta^{(G)}), & i \in \widehat{\mathcal{S}}^{(G)}, \\ 0, & i \notin \widehat{\mathcal{S}}^{(G)}. \end{cases}$$

Write $x_+ = \max\{x, 0\}$. Define

$$g_{\beta,i}(\pi) = \pi_i^{1-\beta}(r_i - b_\beta(\pi)), \qquad f_\beta(\pi) = F(\pi) g_\beta(\pi),$$

where $F(\pi) = \operatorname{diag}(\pi) - \pi\pi^\top$. Then $\widehat{g}_\beta^{(G)} \to g_\beta(\pi)$ almost surely and $\mathbb{E}\|\widehat{g}_\beta^{(G)} - g_\beta(\pi)\|_\infty \to 0$. In particular, there are constants $C_{\mathrm{err}}, c_{\mathrm{tail}} > 0$, depending only on $(K, \beta, R, \pi_{\min})$, such that

$$\left\| \mathbb{E}[\widehat{g}_\beta^{(G)}] - g_\beta(\pi) \right\|_\infty \le C_{\mathrm{err}} G^{-1/2} + C_{\mathrm{err}} G^{(\beta-1)_+} e^{-c_{\mathrm{tail}} G}, \tag{20}$$

$$\left\| F(\pi)\mathbb{E}[\widehat{g}_\beta^{(G)}] - f_\beta(\pi) \right\|_\infty \le C_{\mathrm{err}} G^{-1/2} + C_{\mathrm{err}} G^{(\beta-1)_+} e^{-c_{\mathrm{tail}} G}. \tag{21}$$

Consequently, for a sufficiently small learning rate η

$$\mathbb{E}\Big[\pi(z + \eta \widehat{g}_\beta^{(G)}) - \pi(z)\Big] = \eta f_\beta(\pi) + O\Big(\eta G^{-1/2} + \eta G^{(\beta-1)_+} e^{-c_{\mathrm{tail}} G} + \eta^2\Big). \tag{22}$$

Hence the limits $G \to \infty$ and then $\eta \downarrow 0$ recover the projection flow (7) on every compact interior portion of the path.

Proof of Theorem F.5. For the almost-sure assertion, note that the samples are i.i.d., i.e., $a_1, a_2, \dots \overset{\text{i.i.d.}}{\sim} \pi$. For a given $G \in \mathbb{N}$, we denote

$$N_i^{(G)} = \sum_{g=1}^{G} \mathbb{1}\{a_g = i\}, \qquad \widehat{\pi}_i^{(G)} = \frac{N_i^{(G)}}{G}.$$

The strong law of large numbers gives $\widehat{\pi}_i^{(G)} \to \pi_i$ almost surely for all $i \in [K]$ simultaneously. In particular, because $\pi_i > 0$, almost every sample sequence has a random G_0 such that $\widehat{\pi}_i^{(G)} \ge \pi_i/2 > 0$ for every i and every $G \ge G_0$. We further set

$$\alpha = 2 - \beta, \qquad Z_\alpha(\psi) = \sum_{j=1}^{K} {\psi_j}^\alpha, \qquad b_\beta(\psi) = \frac{\sum_{j=1}^{K} {\psi_j}^\alpha r_j}{Z_\alpha(\psi)}, \qquad g_{\beta,i}(\psi) = {\psi_i}^{1-\beta}\big(r_i - b_\beta(\psi)\big)$$

in the interior of the probability simplex. When all actions are observed, $\widehat{b}_\beta^{(G)} = b_\beta(\widehat{\pi}^{(G)})$ and $\widehat{g}_\beta^{(G)} = g_\beta(\widehat{\pi}^{(G)})$, whereas the population quantities are $b_\beta(\pi)$ and $g_\beta(\pi)$. These maps are continuous in the interior of the probability simplex. Therefore, $\widehat{g}_\beta^{(G)} \to g_\beta(\pi)$ almost surely.

We then prove the bias bound and L^1 convergence. Denote the good event

$$\mathcal{E}_G = \Big\{ \widehat{\pi}_i^{(G)} \ge \frac{\pi_i}{2} \text{ for all } i \Big\}, \qquad \underline{\pi} = \frac{1}{2}\pi_{\min}.$$

On $\mathcal{E}_G$, both π and $\widehat{\pi}^{(G)}$ belong to the convex compact set

$$\Delta_K(\underline{\pi}) = \{\psi \in \Delta_K : \psi_i \ge \underline{\pi} \text{ for every } i\}.$$

For $\psi \in \Delta_K(\underline{\pi})$, the baseline is a convex combination of the rewards, so $|b_\beta(\psi)| \le R$, and

$$Z_\alpha(\psi) \ge K\underline{\pi}^{\alpha_+} > 0.$$

Direct differentiation gives, for $i, n \in [K]$,

$$\frac{\partial b_\beta}{\partial \psi_n}(\psi) = \frac{\alpha {\psi_n}^{\alpha-1}\big(r_n - b_\beta(\psi)\big)}{Z_\alpha(\psi)}, \tag{23}$$

$$\frac{\partial g_{\beta,i}}{\partial \psi_n}(\psi) = (1-\beta)\mathbb{1}\{i=n\}{\psi_i}^{-\beta}\big(r_i - b_\beta(\psi)\big) - {\psi_i}^{1-\beta}\frac{\partial b_\beta}{\partial \psi_n}(\psi).$$

For $\kappa \in \mathbb{R}$, $\sup_{\underline{\pi} \le \varrho \le 1} \varrho^\kappa = \underline{\pi}^{\min\{\kappa,0\}}$. Consequently, combining (23) and $|r_i - b_\beta(\psi)| \le 2R$, we have

$$\left|\frac{\partial b_\beta}{\partial \psi_n}(\psi)\right| \le \frac{2R|\alpha|}{K\underline{\pi}^{\alpha_+}}\underline{\pi}^{\min\{\alpha-1,0\}},$$
$$\left|\frac{\partial g_{\beta,i}}{\partial \psi_n}(\psi)\right| \le 2R|1-\beta|\underline{\pi}^{\min\{-\beta,0\}} + \underline{\pi}^{\min\{1-\beta,0\}}\frac{2R|\alpha|}{K\underline{\pi}^{\alpha_+}}\underline{\pi}^{\min\{\alpha-1,0\}}.$$

Thus there is a finite constant $L_{\mathrm{loc}} = L_{\mathrm{loc}}(K,\beta,R,\pi_{\min})$ such that the mean-value theorem on $\Delta_K(\underline{\pi})$ yields

$$\|g_\beta(\psi) - g_\beta(\pi)\|_\infty \le L_{\mathrm{loc}}\|\psi - \pi\|_1, \qquad \psi \in \Delta_K(\underline{\pi}). \tag{24}$$

The multinomial variances and Cauchy–Schwarz give

$$\mathbb{E}\|\widehat{\pi}^{(G)} - \pi\|_1 \le \sum_{i=1}^K \sqrt{\mathbb{E}(\widehat{\pi}_i^{(G)} - \pi_i)^2} = \frac{1}{\sqrt{G}}\sum_{i=1}^K \sqrt{\pi_i(1-\pi_i)} \le \sqrt{\frac{K}{G}}.$$

Therefore (24) implies

$$\mathbb{E}\big[\|\widehat{g}_\beta^{(G)} - g_\beta(\pi)\|_\infty \mathbb{1}\{\mathcal{E}_G\}\big] \le L_{\mathrm{loc}}\sqrt{\frac{K}{G}}. \tag{25}$$

It remains to control the event on which the empirical policy approaches a face of the simplex. Since $N_i^{(G)}$ is binomial with mean $G\pi_i$, the multiplicative Chernoff inequality gives

$$\mathbb{P}\Big(N_i^{(G)} < \frac{1}{2}G\pi_i\Big) \le \exp\Big(-\frac{1}{8}G\pi_i\Big).$$

A union bound hence yields

$$\mathbb{P}\big(\mathcal{E}_G^c\big) \le K\exp\Big(-\frac{1}{8}G\pi_{\min}\Big). \tag{26}$$

On $\mathcal{E}_G^c$, the empirical baseline is a convex combination of the observed rewards, because its weights $(\widehat{\pi}_i^{(G)})2-\beta$ are strictly positive on the empirical support. Hence $|\widehat{b}_\beta^{(G)}| \le R$. Moreover, every observed empirical mass is at least $1/G$, and every unobserved coordinate of $\widehat{g}_\beta^{(G)}$ is zero. This indicates that

$$\|\widehat{g}_\beta^{(G)}\|_\infty \le 2RG^{(\beta-1)_+}. \tag{27}$$

The population direction satisfies the fixed bound

$$\|g_\beta(\pi)\|_\infty \le 2R\pi_{\min}^{-(\beta-1)_+}.$$

Combining these two bounds with (26), and using $G \ge 1$, gives

$$\mathbb{E}\big[\|\widehat{g}_\beta^{(G)} - g_\beta(\pi)\|_\infty \mathbb{1}\{\mathcal{E}_G^c\}\big] \le C_{\mathrm{err}}G^{(\beta-1)_+}e^{-c_{\mathrm{tail}}G}, \qquad c_{\mathrm{tail}} = \frac{1}{8}\pi_{\min}, \tag{28}$$

where $C_{\mathrm{err}} = C_{\mathrm{err}}(K,\beta,R,\pi_{\min})$. Combining (25) and (28), we get

$$\mathbb{E}\|\widehat{g}_\beta^{(G)} - g_\beta(\pi)\|_\infty \le C_{\mathrm{err}}G^{-1/2} + C_{\mathrm{err}}G^{(\beta-1)_+}e^{-c_{\mathrm{tail}}G}.$$

The right-hand side tends to zero for every fixed $\beta \in \mathbb{R}$, proving the L^1 convergence. $\|\mathbb{E}X\|_\infty \le \mathbb{E}\|X\|_\infty$ gives (20).

Because $F(\pi)$ is deterministic,

$$\big\|F(\pi)\Delta z\big\|_\infty = \max_i \pi_i\Big|\Delta z_i - \sum_j \pi_j \Delta z_j\Big| \le 2\|\Delta z\|_\infty.$$

Furthermore, the definition of $b_\beta(\pi)$ gives $\pi^\top g_\beta(\pi) = 0$, and therefore

$$F(\pi)g_\beta(\pi) = \mathrm{diag}(\pi)g_\beta(\pi) = \big(\pi_i^{2-\beta}(r_i - b_\beta(\pi))\big)_{i=1}^K = f_\beta(\pi).$$

Applying (20) proves (21).

Finally, we convert the logit-direction estimate into a one-step policy estimate. Recall $\pi(z) = \mathrm{softmax}(z)$ and the first derivative of its i-th coordinate is

$$\frac{\partial \pi_i}{\partial z_j}(z) = \pi_i(z)\big(\mathbb{1}\{i=j\} - \pi_j(z)\big), \qquad \nabla_z \pi(z) = \mathrm{diag}(\pi) - \pi\pi^\top = F(\pi).$$

Differentiating once more gives the explicit Hessian entries

$$\frac{\partial^2 \pi_i}{\partial z_j \partial z_n}(z) = \pi_i(z)\big(\mathbb{1}\{i=n\} - \pi_n(z)\big)\big(\mathbb{1}\{i=j\} - \pi_j(z)\big) - \pi_i(z)\pi_j(z)\big(\mathbb{1}\{j=n\} - \pi_n(z)\big).$$

As the softmax probability lies in $[0,1]$, and $|\mathbb{1}\{i=j\} - \pi_j(z)| \le 1$, hence,

$$\left|\frac{\partial^2 \pi_i}{\partial z_j \partial z_n}(z)\right| \le 2 \qquad \text{for all } i,j,n \text{ and all } z \in \mathbb{R}^K. \tag{29}$$

We apply Taylor's theorem with integral remainder to the function $\vartheta_i(\omega) = \pi_i(z + \omega\Delta z)$. Since

$$\vartheta_i'(0) = \nabla \pi_i(z)^\top \Delta z, \qquad \vartheta_i''(\omega) = \Delta z^\top \nabla^2 \pi_i(z+\omega\Delta z)\Delta z,$$

we have the exact identity

$$\pi_i(z+\Delta z) - \pi_i(z) - \nabla \pi_i(z)^\top \Delta z = \int_0^1 (1-\omega)\Delta z^\top \nabla^2 \pi_i(z+\omega\Delta z)\Delta z \, d\omega.$$

By (29),

$$\big|\Delta z^\top \nabla^2 \pi_i(z+\omega\Delta z)\Delta z\big| \le 2\sum_{j=1}^K \sum_{n=1}^K |\Delta z_j||\Delta z_n| = 2\|\Delta z\|_1^2 \le 2K\|\Delta z\|_2^2.$$

Because $\int_0^1 (1-\omega)\, d\omega = 1/2$, taking the maximum over i proves the uniform bound

$$\|\pi(z+\Delta z) - \pi(z) - F(\pi(z))\Delta z\|_\infty \le K\|\Delta z\|_2^2. \tag{30}$$

The same good/bad-event split also supplies the moment needed to average this remainder. On $\mathcal{E}_G$, $\widehat{\pi}_i^{(G)} \ge \underline{\pi}$ for all $i \in [K]$, so

$$\|\widehat{g}_\beta^{(G)}\|_\infty \le 2R\underline{\pi}^{\min\{1-\beta,0\}},$$

whereas (27) holds on $\mathcal{E}_G^c$. Thus

$$\mathbb{E}\|\widehat{g}_\beta^{(G)}\|_2^2 \le 4KR^2\underline{\pi}^{2\min\{1-\beta,0\}} + 4K^2R^2G^{2(\beta-1)_+}e^{-c_{\mathrm{tail}}G} \le C_{\mathrm{err}},$$

because a polynomial times $e^{-c_{\mathrm{tail}}G}$ is bounded. Taking the logit update $\Delta z = \eta\widehat{g}_\beta^{(G)}$ in (30), taking expectations, and then using (21) gives

$$\begin{aligned}\mathbb{E}[\mathrm{softmax}(z+\eta\widehat{g}_\beta^{(G)}) - \mathrm{softmax}(z)] &= \eta F(\pi)\mathbb{E}[\widehat{g}_\beta^{(G)}] + O(\eta^2) \\ &= \eta f_\beta(\pi) + O\big(\eta G^{-1/2} + \eta G^{(\beta-1)_+}e^{-c_{\mathrm{tail}}G} + \eta^2\big),\end{aligned}$$

which proves (22). If π ranges over a compact subset of the simplex interior, then $\pi_{\min}$ has a common positive lower bound, so every constant above can be chosen uniformly. Dividing the last display by η, then taking $G \to \infty$ and $\eta \downarrow 0$, proves the final claim. □