On the Effect of Negative Gradient in Group Relative Deep Reinforcement Optimization

THE UNIVERSITY OF BRITISH COLUMBIA [1] · VECTOR INSTITUTE [2] · amii [3]

Wenlong Deng[1,2], Yi Ren[1], Muchen Li[1], Danica J. Sutherland[1,3], Xiaoxiao Li[1,2*] and Christos Thrampoulidis[1*]

NTHR

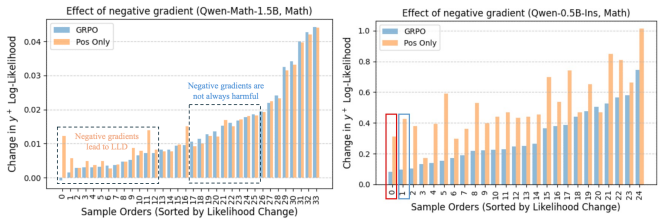
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Background: We identify Lazy Likelihood Displacement (LLD), wherein the likelihood of correct responses marginally increases or even decreases during training.

Related Work: Several approaches have been proposed to mitigate the reduced probabilities of preferred responses in DPO, which is undesirable as it often degrades model performance by diverting probability mass away from optimal response.

Observations



Definition 4.1 *Let $\pi_{\theta_{init}}$ and $\pi_{\theta_{fin}}$ denote the initial and final language models, before and after optimizing a preference learning objective $\mathcal{J}$ (e.g., Eq. (1)) over a dataset $\mathcal{D}$, such that $\mathcal{J}(\theta_{fin}) < \mathcal{J}(\theta_{init})$. We say that **LLD** occurs for a tuple $(\mathbf{x}, \boldsymbol{y}^+) \in \mathcal{D}$ if, for small nonnegative constant $\epsilon \geq 0$,*

$$\ln \pi_{\theta_{fin}}(\boldsymbol{y}^+|\mathbf{x}) < \ln \pi_{\theta_{init}}(\boldsymbol{y}^+|\mathbf{x}) + \epsilon\,. \tag{3}$$

Focus & Contribution

Questions:

(1) Why does GRPO training sometimes reduce the probability of correct responses?

(2) Can we identify the sources of LLD?

(3) Can we mitigate this misalignment without losing data efficiency?

Our Contribution:

Q.1 GRPO Learning Dynamics → **Q.2** LLD Analysis → **Q.3** Selective NTHR Penalization

Q.1 Theory Undersntanding

GRPO training's influence of the likelihood change of correct responses:

$$\frac{d}{dt}\ln \pi_{\theta(t)}(\boldsymbol{y}_i^+|\boldsymbol{x}) = \left\langle \nabla \ln \pi_{\theta(t)}(\boldsymbol{y}_i^+|\boldsymbol{x}),\ \frac{d\theta(t)}{dt} \right\rangle$$

Theorem (GRPO learning dynamics, informal): The likelihood change of **correct responses** will be ***lazy*** if the following quantity ***increases***

$$p^- \underbrace{\sum_{k=1}^{|\boldsymbol{y}_i^+|}\sum_{j=1}^{N^-}\sum_{k'=1}^{|\boldsymbol{y}_j^-|} \alpha_{k,k'}^- \cdot \langle \mathbf{h}_{\mathbf{x},\boldsymbol{y}_{i,<k}^+}, \mathbf{h}_{\mathbf{x},\boldsymbol{y}_{j,<k'}^-} \rangle}_{\text{impact of negative gradient}} - p^+ \sum_{k=1}^{|\boldsymbol{y}_i^+|}\sum_{i'=1}^{N^+}\sum_{k''=1}^{|\boldsymbol{y}_{i'}^+|} \alpha_{k,k''}^+ \cdot \langle \mathbf{h}_{\mathbf{x},\boldsymbol{y}_{i,<k}^+}, \mathbf{h}_{\mathbf{x},\boldsymbol{y}_{i',<k''}^+} \rangle. \tag{5}$$

Illustration of the sample highlighted by the red block in first figure. Mostly, LLD samples are:

(1) Almost Correct

(2) Format Issue

Math Question: Find the greatest integer value of b for which the expression has a domain of all real numbers.

Ground-truth Answer: 5 *Sample: [Math dataset], Model: [Qwen-0.5B-Ins]*

Answer: To solve this problem, we need to **...**

Now, we find the approximate value of $\sqrt{32}$: $\sqrt{32} \approx 5.656$ So, we have: $b > 5.656$

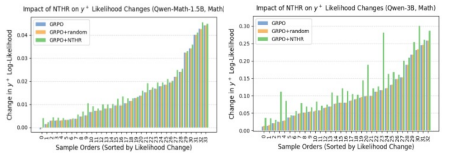
5) Identify the greatest integer value of b : The greatest integer greater than 5.656 is 6. Therefore, the greatest integer value of b for which the expression $\frac{9x^3+4x^2+11x+7}{x^2+bx+8}$ has a domain of all real numbers is 6.

Q.2 &Q.3 Analysis and NTHR

Top-K	Qwen-1.5B-deepseek GWHES	Qwen-1.5B-deepseek Random	Qwen-1.5B-math GWHES	Qwen-1.5B-math Random
10	50%	17.5%	60%	21.3%
15	75%	26.3%	75%	31.9%

Using GWHES (Eq. (5)) score can effectively identify LLD samples compared with random selection.

NTHR Selective Penalization: We use GWHES score and reduce the penalty for negative tokens whose GWHES scores exceed a given threshold.



Effectively mitigates the LLD issue: green lines show a larger increased likelihood

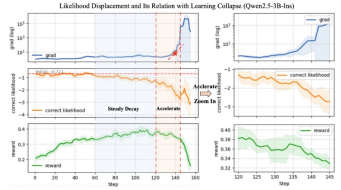
Base model + Method	AIME24	AMC	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-Math-1.5B						
Base	3.3	20.0	39.6	7.7	24.9	19.10
GRPO	13.3	57.5	**71.8**	29.0	34.1	41.14
Pos Only	10.0	57.5	70.6	30.1	31.0	39.84
NTHR	**16.7**	57.5	70.8	**30.5**	**34.2**	41.94
Qwen2.5-0.5B-Ins						
Base	0.0	2.5	33.4	4.4	7.0	9.46
GRPO	0.0	7.5	33.8	**9.2**	8.1	11.72
NTHR	0.0	**10.0**	**36.6**	8.1	8.6	**12.66**
Qwen2.5-1.5B-Ins						
Base	0.0	22.5	53.0	19.1	20.7	23.06
GRPO	3.3	32.5	57.2	18.8	**23.0**	26.96
NTHR	**6.7**	**35.0**	58.8	21.0	20.9	**28.48**
Qwen2.5-Math-1.5B (deepscaler)						
Base	3.3	20.0	39.6	7.7	24.9	19.10
GRPO	10.0	42.5	72.4	**32.4**	31.9	37.80
NTHR	**16.7**	47.5	73.2	29.4	31.4	39.60
Qwen2.5-3B						
Base	10.0	37.5	58.6	26.1	24.6	31.36
GRPO	6.7	35.0	66.6	31.2	**29.9**	33.88
NTHR	**10.0**	**47.5**	65.6	31.6	26.8	**36.30**

Improved greedy decoding performance.

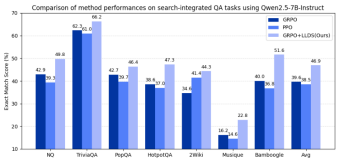
Borader Impact

We take this opportunity to **highlight two of our recent related works**.

[1] LLD is the **source** of **GRPO collapse** in Tool-integrated **mult-turn RL.**



Case study using Search-R1, demonstrating strong LLD.



We provide a simple regularization to make training stable.

[2] We unify direct optimization and advantage shaping of recent GRPO-variants.

Table 1: Unified view of main algorithms studied in this work. Rows grouped by target evaluation metric. ρ, ρ_K are per-example expected 0/1 and Pass@K rewards; $\hat{\rho}$, $\hat{\rho}_K$ are their empirical estimators

Algorithm	Advantage Scores (A^+, A^-)	Weighted Empirical Gradient	Population Surrogate Reward
Targets 0/1 evaluation (K=1)			
RLOO [KHW19]	$\left(\frac{N^- - \hat{\rho}N^-}{\hat{\rho}N^+}, -\frac{\hat{\rho}N^+}{\hat{\rho}N^+}\right)$	$\hat{\rho}(1-\hat{\rho})[\hat{\nabla}_+ - \hat{\nabla}_-]$	ρ
GRPO [Sha+24]	$\left(\sqrt{\frac{1-\hat{\rho}}{\hat{\rho}}}, -\sqrt{\frac{\hat{\rho}}{1-\hat{\rho}}}\right)$	$\sqrt{\hat{\rho}(1-\hat{\rho})}[\hat{\nabla}_+ - \hat{\nabla}_-]$	$2\arcsin(\sqrt{\rho})$
Skew-It [This work]	$\left((1-\hat{\rho})\sqrt{\frac{\hat{\rho}}{1-\hat{\rho}}}, -(1-\hat{\rho})\sqrt{\frac{1-\hat{\rho}}{\hat{\rho}}}\right)$	$(1-\hat{\rho})\sqrt{\hat{\rho}(1-\hat{\rho})}[\hat{\nabla}_+ - \hat{\nabla}_-]$	$\arcsin(\sqrt{\rho}) + \sqrt{\rho(1-\rho)}$
Targets Pass@K evaluation ($K \geq 2$)			
RLOO$_K$ [This work]	$\left(\hat{P}_{K-1}\frac{N^+(N^+-1)}{N^+N^-}, -\hat{P}_{K-1}\frac{N^+}{N^-}\right)^\dagger$	$\hat{P}_{K-1}\hat{\rho}(1-\hat{\rho})\hat{\nabla}_+ - \hat{P}_{K-1}\hat{\rho}(1-\hat{\rho})\hat{\nabla}_-$	ρ_K
GRPO$_K$ [Cha+25a]	$\left(\alpha_K\sqrt{\frac{N^-}{N^+}}, -\alpha_K\sqrt{\frac{N^+}{N^-}}\right)^\dagger$	$\sqrt{\frac{1-\hat{\rho}_K}{\hat{\rho}_K}}\hat{\rho}[\hat{\nabla}_+ - \hat{\nabla}_-]$	$\frac{\pi}{2}\arcsin(\sqrt{\rho_K})$
GRPO$_K$ [This work]	$\left(\hat{P}_{K-1}\sqrt{\frac{1-\hat{\rho}}{\hat{\rho}}}, -\hat{P}_{K-1}\sqrt{\frac{\hat{\rho}}{1-\hat{\rho}}}\right)$	$\hat{P}_{K-1}\sqrt{\hat{\rho}(1-\hat{\rho})}\hat{\nabla}_+ - \hat{P}_{K-1}\sqrt{\hat{\rho}(1-\hat{\rho})}\hat{\nabla}_-$	$\mathcal{B}(1-(1-\rho_K)^{1/K}; \frac{1}{2}, K-\frac{1}{2})$ *

[1] On GRPO Collapse in Search-R1: The Lazy Likelihood-Displacement Death Spiral.

[2] Advantage Shaping as Surrogate Reward Maximization: Unifying Pass@K Policy Gradients

[1] Tool-RL-LLD

[2] ASSR