Geometric-Averaged Preference Optimization for Soft Preference Labels

Hiroki Furuta[1,2], Kuang-Huei Lee[1], Shixiang Shane Gu[1], Yutaka Matsuo[2], Aleksandra Faust[1], Heiga Zen[1], Izzeddin Gur[1]

[1]Google DeepMind, [2]The University of Tokyo

RLHF & DPO only consider binary preference labels

- Most prior works to align LLMs (RLHF & DPO) only assume binary preference labels.
  - y_1 is better than y_2 (with probability/confidence 1)
  - E.g. reward modeling objective only considers the positive term of binary cross entropy:

$$\min_\psi -\mathbb{E}\left[\log \sigma(r_\psi(x, y_1) - r_\psi(x, y_2))\right]$$

- However, human preference can vary across individuals, and should be represented distributionally → proportional **soft preference labels**

Soft Preference Labels

- Soft preference labels are proportional
  - E.g. y_1 is better than y_2 in 70% (y_2 is better than y_1 in 30%)
- We define soft labels as an approximation of true preference probability p*, and estimate it with an average of sampled binary preference labels $l_i \in \{0, 1\}$
  - Monte-Carlo sampling, Majority Voting, etc

$$\hat{p}_{x,y_1,y_2} := \hat{p}(y_1 \succ y_2|x) \approx p^*(y_1 \succ y_2|x) \qquad \hat{p} = \frac{1}{M}\sum_{i=1}^{M} l_i$$

- (to estimate soft preference labels, we may leverage AI feedback with token logits and Bradley-Terry models)

Proposal: Weighted Geometric-Averaging of Output Likelihoods

$$y_w \sim \bar{\pi}(y_w \mid x) := \frac{1}{Z_{\pi,w}(x)} \pi(y_1 \mid x)^{\hat{p}} \pi(y_2 \mid x)^{1-\hat{p}}$$
$$y_l \sim \bar{\pi}(y_l \mid x) := \frac{1}{Z_{\pi,l}(x)} \pi(y_1 \mid x)^{1-\hat{p}} \pi(y_2 \mid x)^{\hat{p}},$$

- Replace the original likelihoods in DPO objective with their weighted geometric average (while ignoring normalization term)

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}} \left[\log \sigma \left(h_\theta(x, y_1, y_2)\right)\right]$$
$$= -\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)\pi_\theta(y_2 \mid x)}\right)\right]$$

$$\pi(y_1 \mid x) \rightarrow \pi(y_1 \mid x)^{\hat{p}} \pi(y_2 \mid x)^{1-\hat{p}} \qquad \pi(y_2 \mid x) \rightarrow \pi(y_1 \mid x)^{1-\hat{p}} \pi(y_2 \mid x)^{\hat{p}}$$

Geometric Direct Preference Optimization (GDPO)

$$\mathcal{L}_{\text{GDPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_1 \mid x)^{\hat{p}} \pi_\theta(y_2 \mid x)^{1-\hat{p}} \pi_{\text{ref}}(y_1 \mid x)^{1-\hat{p}} \pi_{\text{ref}}(y_2 \mid x)^{\hat{p}}}{\pi_{\text{ref}}(y_1 \mid x)^{\hat{p}} \pi_{\text{ref}}(y_2 \mid x)^{1-\hat{p}} \pi_\theta(y_1 \mid x)^{1-\hat{p}} \pi_\theta(y_2 \mid x)^{\hat{p}}}\right)\right]$$
$$= -\mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}} \left[\log \sigma \left(\beta(2\hat{p} - 1) \log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)\pi_\theta(y_2 \mid x)}\right)\right],$$

Proposal: Geometric-DPO and its variant (GIPO)

- Such an geometric-averaging can be applicable to any method based on DPO

$$\mathcal{L}_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}\left[\left(h_\theta(x, y_1, y_2) - \frac{1}{2\beta}\right)^2\right]$$

$$\mathcal{L}_{\text{cIPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}}\left[\left(h_\theta(x, y_1, y_2) - \frac{2\hat{p}-1}{2\beta}\right)^2\right]$$

Geometric Identity Preference Optimization (GIPO)

$$\mathcal{L}_{\text{GIPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}}\left[(2\hat{p}-1)^2\left(h_\theta(x, y_1, y_2) - \frac{1}{2\beta}\right)^2\right]$$

Proposal: Geometric-DPO and its variant (GROPO)

- Such an geometric-averaging can be applicable to any method based on DPO

$$\mathcal{L}_{\text{ROPO}}(\pi_\theta, \pi_{\text{ref}}) = \alpha \mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}} \left[\sigma\left(h_\theta(x, \mathbf{y_2}, \mathbf{y_1})\right)\right] - \gamma \mathbb{E}_{(x,y_2,y_1)\sim\mathcal{D}} \left[\log \sigma\left(h_\theta(x, y_1, y_2)\right)\right]$$
$$= \alpha \left(1 - \mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}} \left[\sigma\left(h_\theta(x, y_1, y_2)\right]\right)\right) + \gamma \mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}),$$

Geometric Robust Preference Optimization (GROPO)

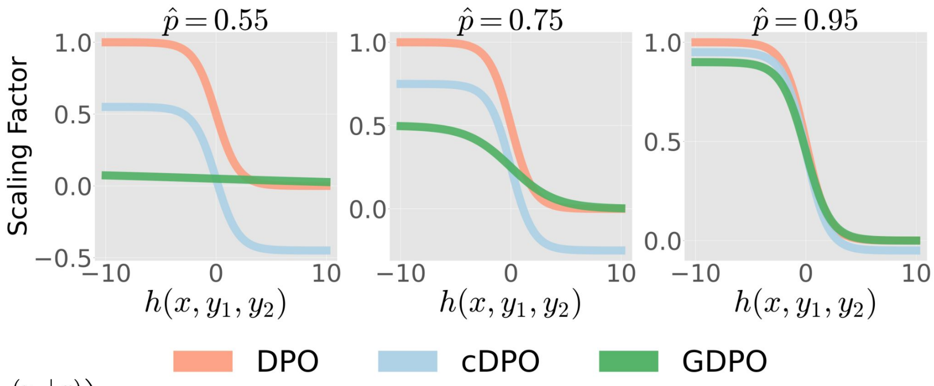
$$\mathcal{L}_{\text{GROPO}}(\pi_\theta, \pi_{\text{ref}}) = \alpha \left(1 - \mathbb{E}_{\mathcal{D}} \left[\sigma\left(\beta(2\hat{p}-1) \log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)\pi_\theta(y_2 \mid x)}\right)\right]\right) + \gamma \mathcal{L}_{\text{GDPO}}(\pi_\theta, \pi_{\text{ref}})$$

Adjust the Scale of Gradients

- Geometric-Averaging can adjust the norm of gradient based on soft preference
  - Make the scale of gradients from the equally-good samples close to zero (i.e. ignoring gradients around p=0.5)

$$\nabla_\theta \mathcal{L} = -\beta \mathbb{E}_{(x,y_1,y_2,\hat{p})\sim\mathcal{D}}\left[\underbrace{w_\theta(x,y_1,y_2,\hat{p})}_{\text{scaling factor}}\underbrace{[\nabla_\theta \log \pi_\theta(y_1 \mid x) - \nabla_\theta \log \pi_\theta(y_2 \mid x)]}_{\text{positive and negative policy gradients}}\right]$$

Method	Scaling Factor w_θ
DPO [38]	$1-\rho_\theta$
cDPO [28]	$\hat{p}-\rho_\theta$
GDPO (ours)	$(2\hat{p}-1)(1-\rho'_\theta)$
IPO [38]	$\frac{1}{\beta^2}-\frac{2}{\beta}\log\frac{\rho_\theta}{1-\rho_\theta}$
cIPO [26]	$\frac{2\hat{p}-1}{\beta^2}-\frac{2}{\beta}\log\frac{\rho_\theta}{1-\rho_\theta}$
GIPO (ours)	$(2\hat{p}-1)^2\left(\frac{1}{\beta^2}-\frac{2}{\beta}\log\frac{\rho'_\theta}{1-\rho'_\theta}\right)$
ROPO [26]	$(\gamma-\alpha\rho_\theta)(1-\rho_\theta)$
GROPO (ours)	$(2\hat{p}-1)(\gamma-\alpha\rho'_\theta)(1-\rho'_\theta)$



$\rho_\theta := \sigma\left(\beta \log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)\pi_\theta(y_2 \mid x)}\right)$ $\rho'_\theta := \sigma\left(\beta(2\hat{p}-1)\log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)\pi_\theta(y_2 \mid x)}\right)$

Soft Preference Labels from AI Feedback

- Ask LLM which output (1) or (2) is preferable, compute the logit of (1) and (2) tokens, and then transform them into AI preference probability through Bradley-Terry model

$$\hat{p}_{\mathrm{AI}}(y_1 \succ y_2 \mid x) = \frac{\exp(\texttt{score}(\texttt{(1)}))}{\exp(\texttt{score}(\texttt{(1)})) + \exp(\texttt{score}(\texttt{(2)}))}$$

Prompt for AI Feedback (Train/Eval) on Plasma Plan

Task: Judge the quality of two plans, choose the option among (1) or (2). A good plan should be well-ordered, complete, informative and contains no repetitive steps.

Goal: {goal}
Plan (1): {plan_1}
Plan (2): {plan_2}
Choose among (1) or (2):

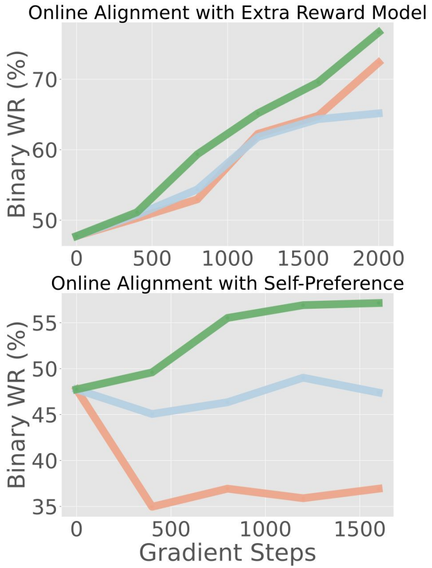
Results with Common RLHF benchmarks

- In standard RLHF benchmarks (Reddit TL;DR, Helpfulness & Harmlessness), Geometric-Averaging consistently outperforms original methods

	Reddit TL;DR				Anthropic Helpful				Anthropic Harmless			
	v.s. PaLM 2-L		v.s. GPT-4		v.s. PaLM 2-L		v.s. GPT-4		v.s. PaLM 2-L		v.s. GPT-4	
Methods	Binary	%	Binary	%	Binary	%	Binary	%	Binary	%	Binary	%
SFT	16.20%	41.08%	3.80%	33.38%	62.60%	56.69%	5.74%	20.67%	62.76%	57.83%	31.54%	36.42%
DPO [41]	16.90%	40.91%	4.00%	33.51%	86.21%	75.40%	16.23%	33.98%	75.40%	65.95%	41.02%	42.79%
cDPO [30]	17.20%	41.61%	3.80%	33.38%	83.28%	74.04%	16.11%	33.28%	74.97%	65.91%	39.53%	40.52%
GDPO (ours)	**19.30%**	**41.69%**	**4.70%**	**33.56%**	**88.90%**	**76.59%**	**19.83%**	**36.07%**	**77.70%**	**67.43%**	**43.31%**	**44.33%**
IPO [2]	20.40%	42.79%	5.00%	34.22%	91.09%	78.91%	21.66%	38.84%	80.36%	68.85%	43.37%	44.72%
cIPO [27]	19.70%	42.04%	4.40%	33.52%	90.24%	77.84%	18.18%	36.88%	81.85%	69.92%	44.80%	45.03%
GIPO (ours)	**21.90%**	**43.03%**	**5.30%**	**34.84%**	**92.56%**	**79.48%**	**21.90%**	**39.04%**	**87.24%**	**71.75%**	**51.92%**	**47.86%**
ROPO [27]	16.20%	40.20%	4.20%	33.40%	86.33%	74.96%	17.45%	34.83%	74.10%	65.74%	43.37%	44.72%
GROPO (ours)	**18.50%**	**41.56%**	**5.30%**	**34.84%**	**88.71%**	**77.10%**	**20.13%**	**36.42%**	**77.26%**	**67.38%**	**44.80%**	**45.03%**
Ave.Δ(+Geom.)	**+2.10%**	**+0.69%**	**+0.78%**	**+0.72%**	**+2.90%**	**+1.62%**	**+2.79%**	**+1.77%**	**+4.09%**	**+1.87%**	**+4.63%**	**+2.33%**

Results with Online Feedback

- By preparing extra reward models, or calculating the reward with the likelihood of LLM itself (self-preference), we can extend offline DPO into online settings
- With self-preference (inaccurate in many cases), GDPO significantly outperforms others.

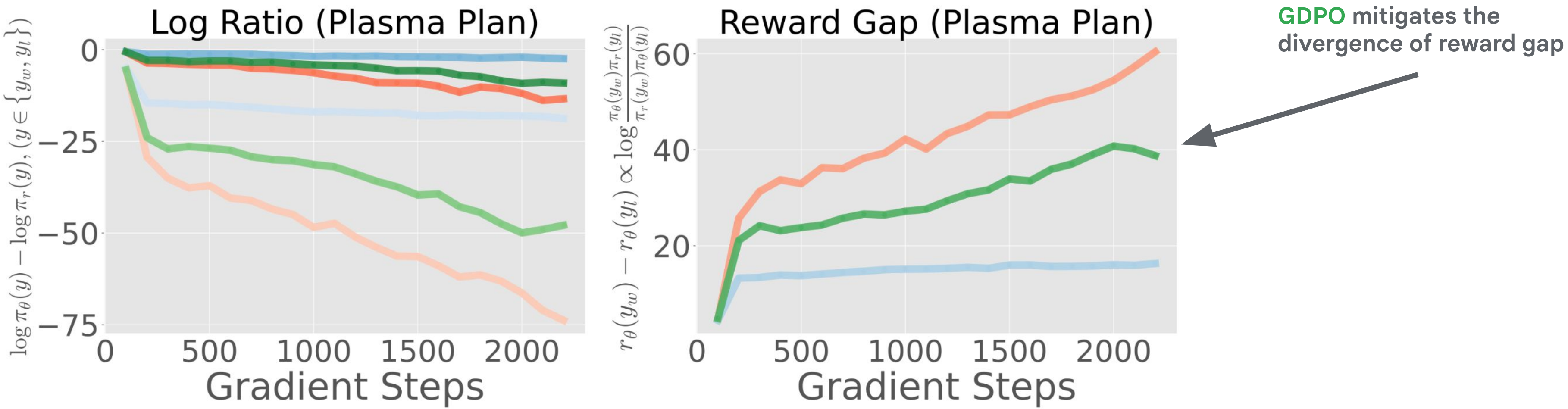


Self-Preference

$$\rho_\theta = \sigma\left(\beta \log \frac{\pi_\theta(x,y_w)\pi_{\text{ref}}(x,y_l)}{\pi_{\text{ref}}(x,y_w)\pi_\theta(x,y_l)}\right)$$

Issue 1: Over-Optimization (in DPO)

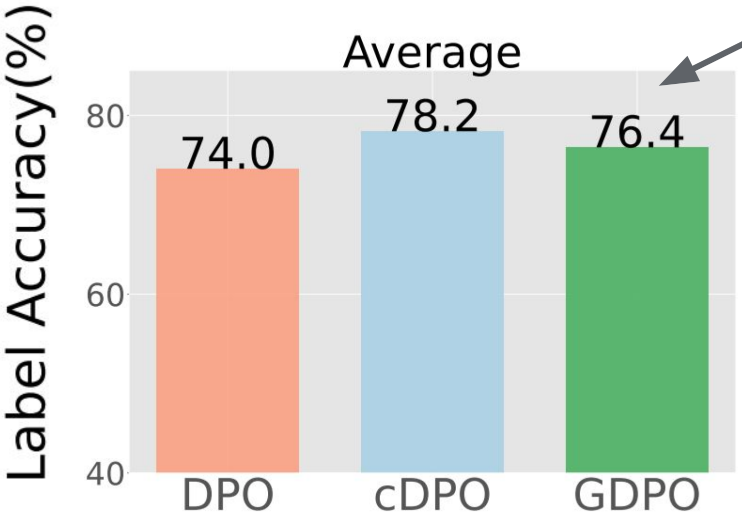
- It is pointed out that DPO objective forces reward gap increase to infinity
- This causes unnecessary update of positive/negative likelihoods (i.e. over-optimization) $r_\theta(x, y_w) - r_\theta(x, y_l) \rightarrow \infty$



Issue 2: Objective Mismatch (in cDPO)

- Conservative DPO (cDPO) have binary-cross entropy objective by leveraging soft preference labels, which is good at preference modeling, but not always lead to better greedy decoding for text generation (objective mismatch)

GDPO balances the preference modeling and text generation (but cDPO fails)



Methods	Plasma Plan v.s. PaLM 2-L Binary	Plasma Plan v.s. PaLM 2-L %	Plasma Plan v.s. GPT-4 Binary	Plasma Plan v.s. GPT-4 %
SFT	47.74%	48.87%	24.51%	39.88%
DPO [41]	83.16%	63.88%	66.20%	54.34%
cDPO [30]	75.96%	60.25%	35.66%	45.49%
GDPO (ours)	**85.48%**	**64.83%**	**72.36%**	**55.90%**

Conclusion

- Introduce soft preference labels, in contrast to binary labels
  - Majority Voting, AI feedback, etc
- Propose weighted geometric averaging of output likelihood
  - Applicable to any method based on DPO
  - Make the scale of gradients from the equally-good samples close to zero
- Geometric-DPO/IPO/ROPO consistently outperforms original methods
- GDPO can mitigate over-optimization and objective mismatch issues