A Common Pitfall of Margin-based Language Model Alignment: Gradient Entanglement

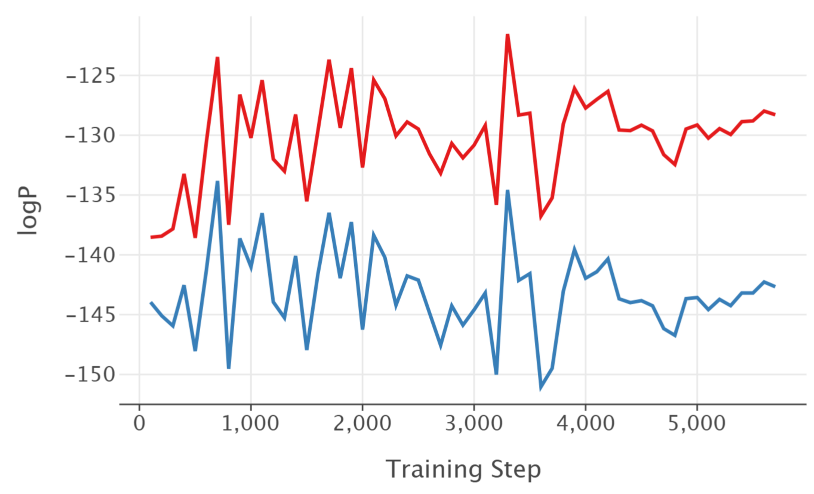
Hui Yuan*, Yifan Zeng*, Yue Wu*, Huazheng Wang, Mengdi Wang and Liu Leqi*

PRINCETON UNIVERSITY

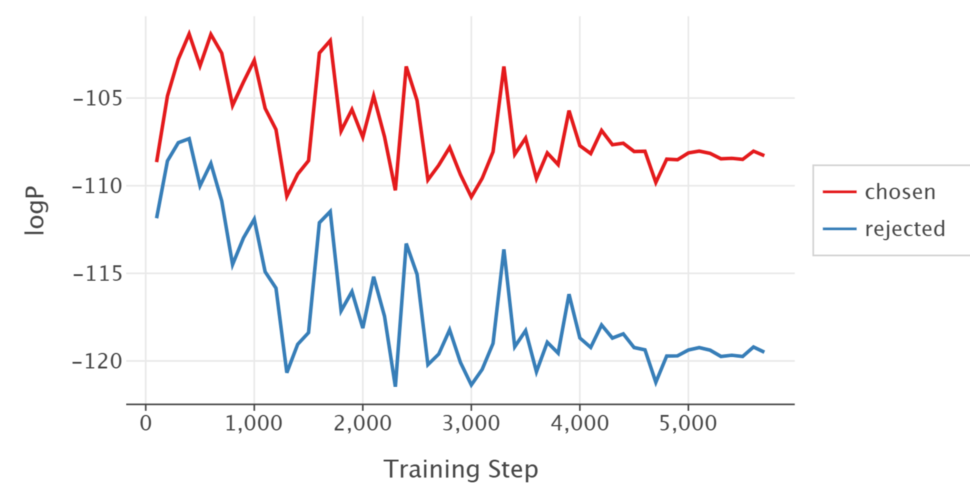
Oregon State University

TEXAS
The University of Texas at Austin

Problematic Behavior



Model log-likelihood on rejected response may increase



Model log-likelihood on chosen response may decrease

Key Takeaways

A Pitfall of RLHF: Underspecify the ideal behavior of model log-probabilities

The Cause: Gradient Entanglement effect passed through the gradient inner product

Why Large Gradient Inner Product: Non-contrastive tokens are involved

Case Study: DPO

DPO Objective: $\ell_{DPO} = -\log\sigma(a-b)$ with $a := \beta \log\left(\frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{ref}(\mathbf{y}_w|\mathbf{x})}\right)$ and $b := \beta \log\left(\frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{ref}(\mathbf{y}_l|\mathbf{x})}\right)$

After one step optimizing ℓ_{DPO}:

$$\Delta \log \pi_w \approx C \cdot \left(\|\nabla \log \pi_w\|^2 - \langle \nabla \log \pi_w, \nabla \log \pi_l \rangle\right),$$
$$\Delta \log \pi_l \approx C \cdot \left(\langle \nabla \log \pi_w, \nabla \log \pi_l \rangle - \|\nabla \log \pi_l\|^2\right).$$

Case	$\log \pi_w, \log \pi_l$	Condition
1	$\log \pi_w \uparrow \log \pi_l \downarrow$	$\langle \nabla \log \pi_w, \nabla \log \pi_l \rangle \le \min(\|\nabla \log \pi_w\|^2, \|\nabla \log \pi_l\|^2)$
2	$\log \pi_w \downarrow \log \pi_l \downarrow$	$\|\nabla \log \pi_w\|^2 \le \langle \nabla \log \pi_w, \nabla \log \pi_l \rangle \le \|\nabla \log \pi_l\|^2$
3	$\log \pi_w \uparrow \log \pi_l \uparrow$	$\|\nabla \log \pi_l\|^2 \le \langle \nabla \log \pi_w, \nabla \log \pi_l \rangle \le \|\nabla \log \pi_w\|^2$

The Cause: Gradient Entanglement

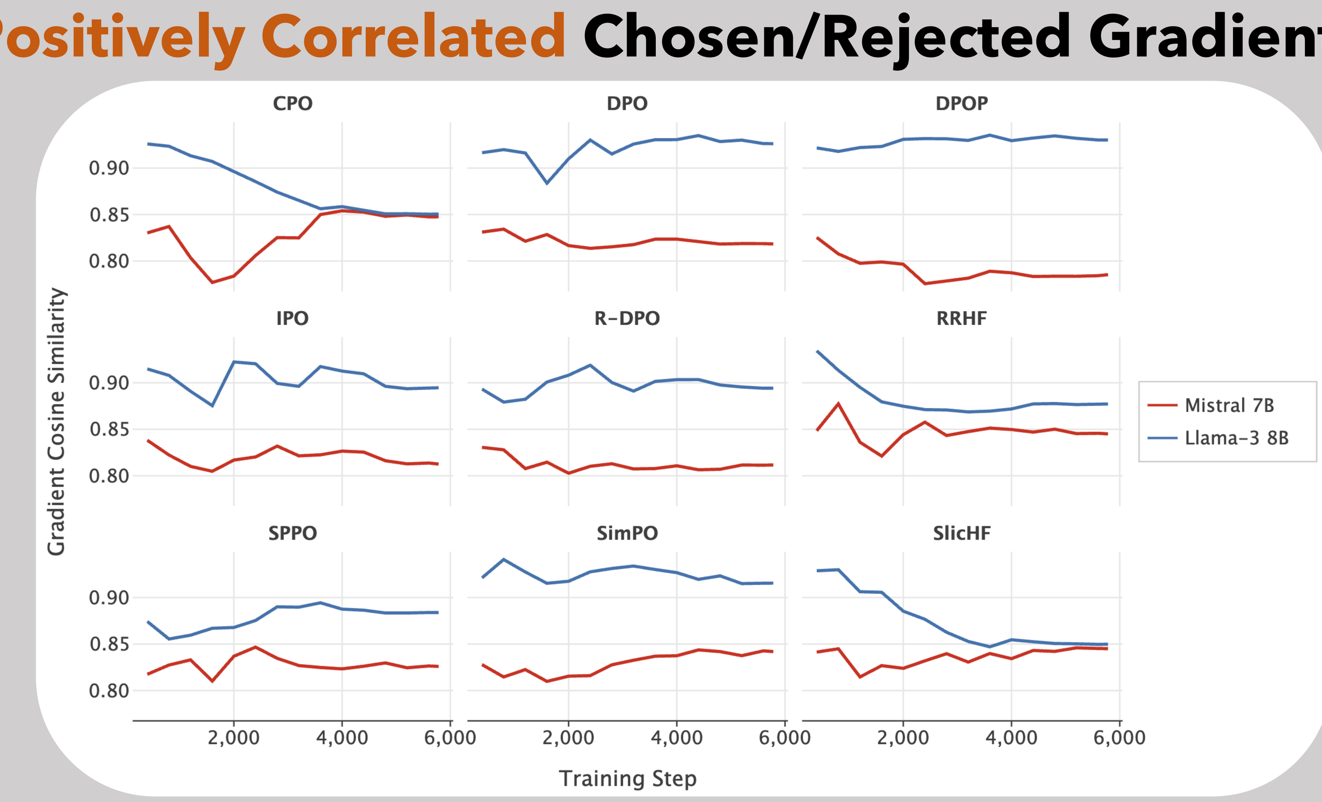
General Margin-Based RLHF Objective $\ell(\theta) = -\Big(m(h_w(\log \pi_w) - h_l(\log \pi_l)) + \Lambda(\log \pi_w)\Big)$

	$m(a)$	$h_w(a)$
DPO (Rafailov et al.)	$\log \sigma(a - c_{\text{ref}})$	βa
R-DPO (Park et al.)	$\log \sigma(a - (c_{\text{ref}} + \alpha(\lvert y_w\rvert - \lvert y_l\rvert)))$	[illegible]
SimPO (Meng et al.)	$\log \sigma(a - \gamma)$	[illegible]
IPO (Azar et al.)	$(a - (c_{\text{ref}} + \frac{1}{2\beta}))^2$	[illegible]
RRHF (Yuan et al.)	$\min(0, a)$	[illegible]
SlicHF (Zhao et al.)	$\min(0, a - \delta)$	[illegible]
CPO (Xu et al.)	$\log \sigma(a)$	[illegible]
DPOP (Pal et al.)	$\log \sigma(a - c_{\text{ref}})$	[illegible]
KTO (Ethayarajh et al.)	a	[illegible]
SPPO (Wu et al.)	a	$(a - \beta^{-1})^2$

Table 2: Instantiation of margin-based preference optimization losses.

Scan for the full Table 2 & paper !

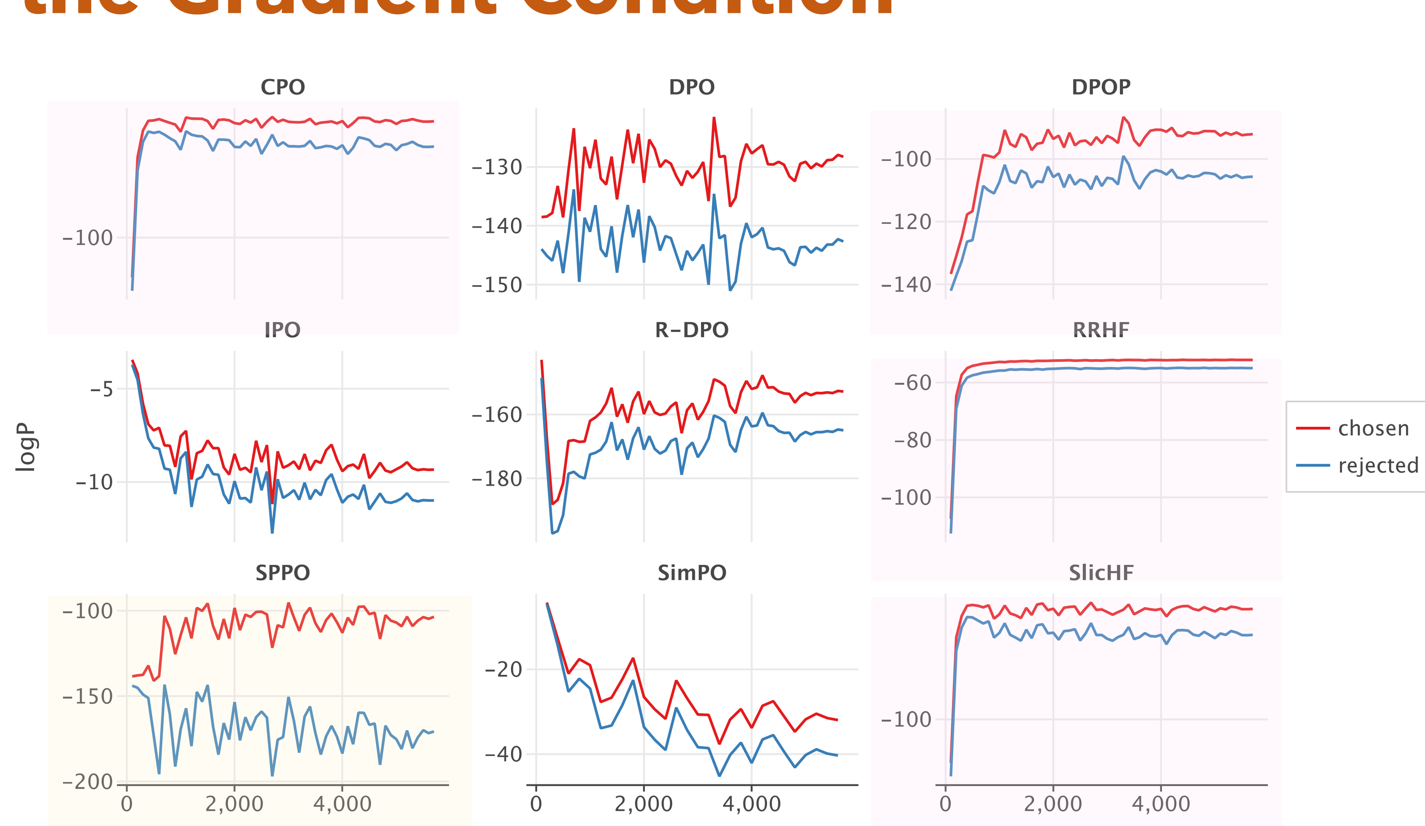
Positively Correlated Chosen/Rejected Gradients



The chosen log-probability change depends on the rejected gradient, and vice versa. The mutual dependency is characterized by: (**d_w and d_l are objective-dependent scalars**)

$$\Delta \log \pi_w \approx \eta \left(d_w \|\nabla_\theta \log \pi_w\|^2 - d_l \langle \nabla_\theta \log \pi_w, \nabla_\theta \log \pi_l \rangle\right),$$
$$\Delta \log \pi_l \approx \eta \left(d_w \langle \nabla_\theta \log \pi_w, \nabla_\theta \log \pi_l \rangle - d_l \|\nabla_\theta \log \pi_l\|^2\right).$$

Explainable Training Dynamics with the Gradient Condition



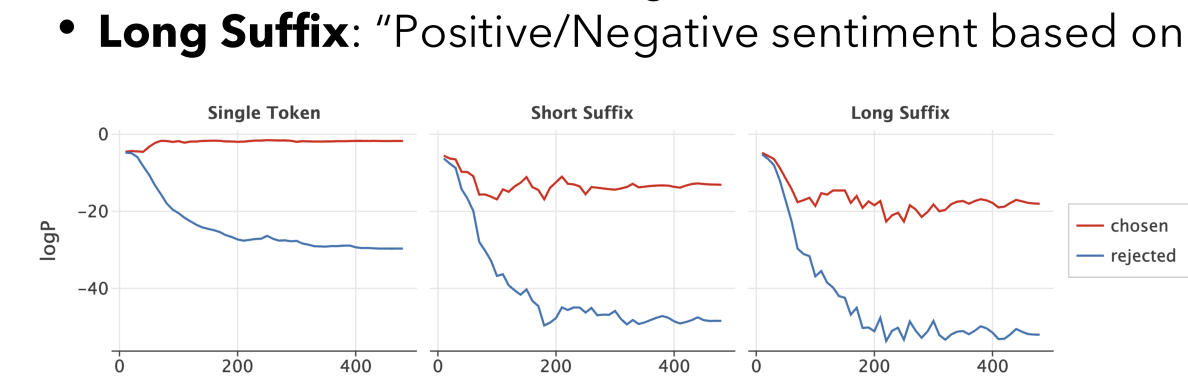
Explicit regularization on $\log \pi_w$: d_w is greater so that $\log \pi_w$ is more likely to increase

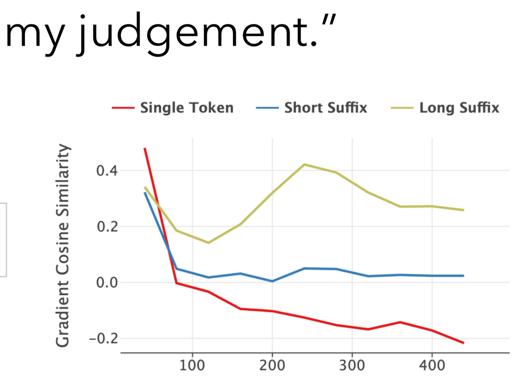
SPPO: $\frac{d_w}{d_l} > 1$ and $\|\nabla \log \pi_l\|^2 > \|\nabla \log \pi_w\|^2$ is observed, thus the gradient condition are more lenient to be satisfied.

Investigation: Why the gradient inner product is large?

Consider a synthetic RLHF dataset (x: *statement*, $\mathbf{y}_w$: **true sentiment**, $\mathbf{y}_l$: **false sentiment**) with three configurations of $\mathbf{y}_w$ and $\mathbf{y}_l$:

- **Single Token**: "Positive/Negative."
- **Short Suffix**: "Positive/Negative sentiment."
- **Long Suffix**: "Positive/Negative sentiment based on my judgement."





Theoretical Results

- **(Theorem 1) Single Token**: $\langle \nabla \log \pi_w, \nabla \log \pi_l \rangle < 0$, thus $\log \pi_w \uparrow$ and $\log \pi_l \downarrow$.
- **(Theorem 3) Short/Long Suffix**: $\langle \nabla \log \pi_w, \nabla \log \pi_l \rangle > 0$ as the suffix length goes up, both $\log \pi_w$ and $\log \pi_l$ decrease.
  - **The token-wise inner product can be negative**: $\langle \nabla \log \pi_w^i, \nabla \log \pi_l^i \rangle < 0$, i is the index of "Positive/Negative".