NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF CAMBRIDGE

On Extending Direct Preference Optimization to Accommodate Ties

Jinghong Chen (Eric)
jc2124@cam.ac.uk

Motivation

- Direct Preference Optimization (DPO) trains LLMs on pairs of preferred and dispreferred responses:

$$\{(y_i, y_j): y_i \succ y_j\} \rightarrow \text{DPO}$$

- But not all pairs have a clear winner. Current practice (Llama 3, Qwen 2) discards them.

$$\{(y_i, y_j): y_i \sim y_j\} \rightarrow \text{Discarded}$$

"Any model which does not allow for the possibility of ties is not making full use of the information contained in the no-preference class." *- A Generalization of the Bradley-Terry Model, Rao and Kupper, 1967*

DPO does not Accommodate Ties

- The Bradley-Terry (BT) comparison model assign probability of item y_i beating item y_j

$$P^{BT}(y_i \succ y_j) = \frac{e^{r_i}}{e^{r_i} + e^{r_j}}$$

$$= \sigma(r_i - r_j) = \sigma(d_{ij})$$

- DPO uses policy-reference log-likelihood ratio as the reward.

$$r_i = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{ref}(y_i|x)}$$

- and maximizes the following log-probability over all pairs in the training set

$$\log P^{BT}(y_i \succ y_j)$$

But Ties Emerge Naturally

Machine Translation & Summarization

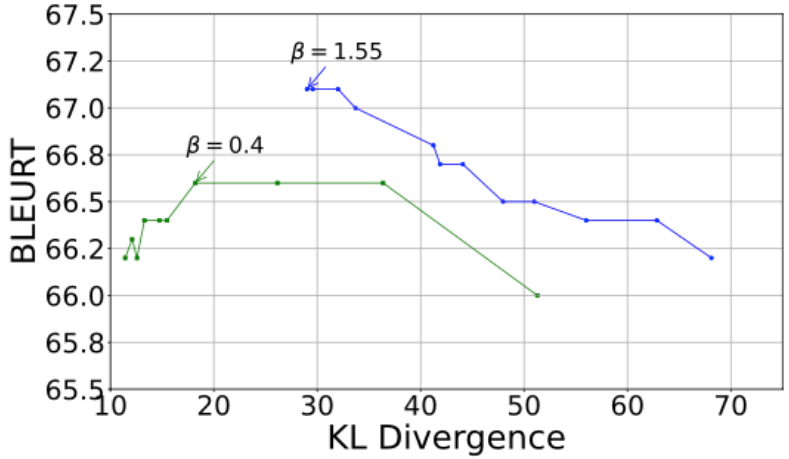
Many possible responses with similar quality (WMT, IWSLT, TL;DR)

Mathematical Reasoning

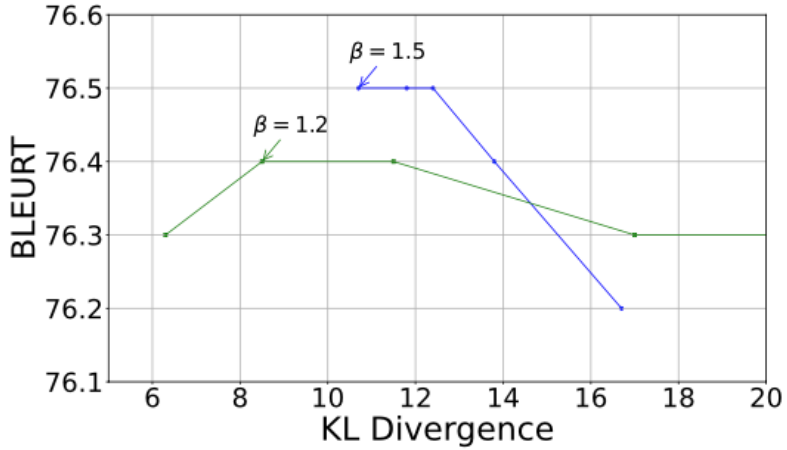
Different reasoning steps leading to the correct answer (GSM8K)

Running DPO on Ties Hurts Performance

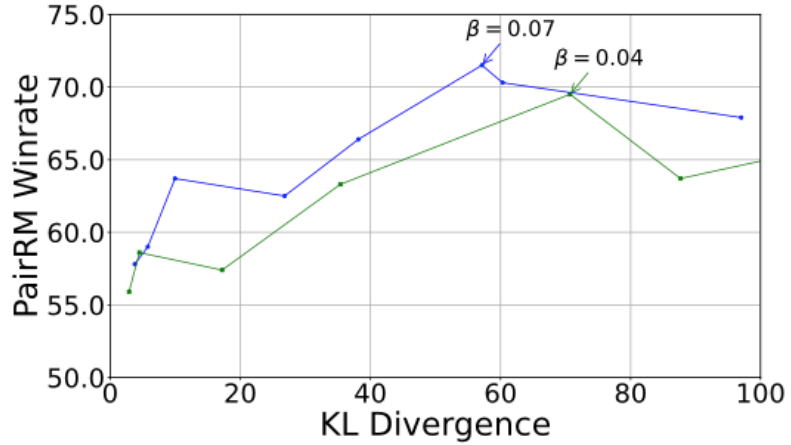
Tied Pairs = responses with the closest scores among 8 samples.



(a) WMT21 ZH-EN.
Baseline = 66.0



(b) IWSLT17 FR-EN.
Baseline = 75.7



(c) TL;DR PairRM Win-Rate;
Baseline = 37.5%

Task Performance v.s. KL Divergence. We train systems with different beta values to produce operating curves. Systems on Blue curves are trained with Clear Preference only. Systems on green curves are trained with Clear Preference and Tied Pairs.

Tie-Compatible Bradley-Terry Variants

Rao-Kupper Model (RK)

- BT with a margin ν_{RK}: item with similar strengths are likely to tie

$$P^{RK}(y_i \succ y_j) = \frac{\lambda_i}{\lambda_i + \nu_{RK}\lambda_j} = \sigma(d_{ij} - \log \nu_{RK})$$

$$P^{RK}(y_i \sim y_j) = \frac{(\nu_{RK}^2 - 1)\lambda_i\lambda_j}{(\lambda_i + \nu_{RK}\lambda_j)(\lambda_j + \nu_{RK}\lambda_j)} = (\nu_{RK}^2 - 1)\sigma(-d_{ij} - \log \nu_{RK})P^{RK}(y_i \succ y_j)$$

Davidson Model (D)

- Respect Luce's Choice Theorem $\frac{P(y_i \succ y_j)}{P(y_j \succ y_i)} = \frac{\lambda_i}{\lambda_j}$

$$P^{D}(y_i \succ y_j) = \frac{\lambda_i}{\lambda_i + \lambda_j + \nu_D\sqrt{\lambda_i\lambda_j}} = \frac{1}{1 + \exp(-d_{ij}) + 2\nu_D \exp(-d_{ij}/2)}$$

$$P^{D}(y_i \sim y_j) = 2\nu_D \exp(-d_{ij}/2)P^{D}(y_w \succ y_l)$$

DPO-RK and DPO-D Objectives

- Using the same reward parameterization as DPO $r_i = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{ref}(y_i|x)}$
- Optimize both on Clear Preference and Tied Pairs

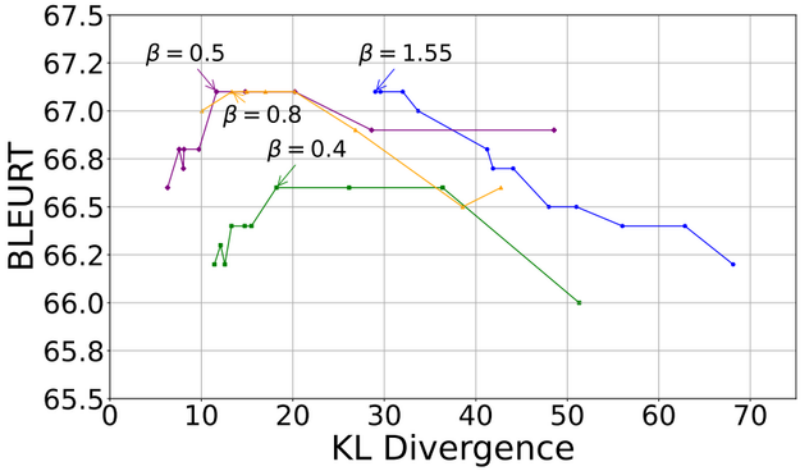
$\log P(y_i \succ y_j)$ on Clear Preferences; $\log P(y_i \sim y_j)$ on Tied Pairs

- Now the pairwise dataset can contain ties!

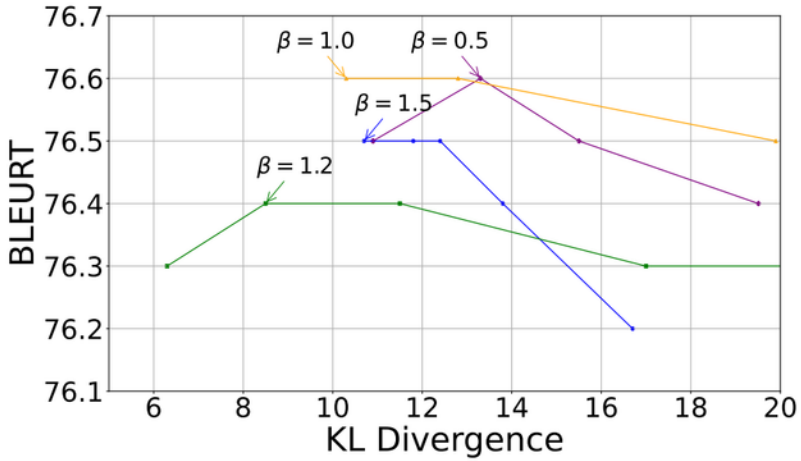
Benefits of Including Ties

Tied Pairs = responses with the closest scores among 8 samples.

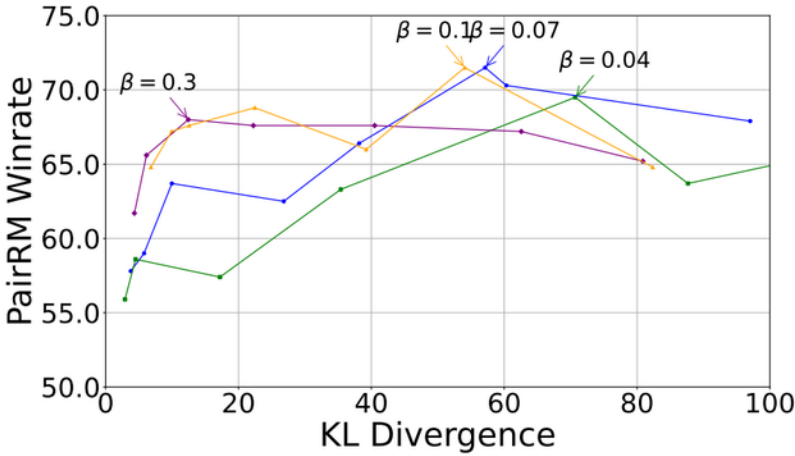
Similar Performance as DPO(CP), but at much smaller KL divergence



(a) WMT21 ZH-EN frontiers.
Baseline BLEURT = 66.0



(b) IWSLT FR-EN frontiers.
Baseline BLEURT = 75.7



(c) TL;DR frontiers.
Baseline win-rate = 37.5%

Performance v.s. KL operating curves. Purple: DPO-RK(CP+TP) systems; Yellow: DPO-D(CP+TP) systems, Blue: DPO(CP) systems; Green: DPO-D(CP+TP) systems. We find that DPO-RK(CP+TP) and DPO-D(CP+TP) perform similary as DPO-D(CP), but does this at smaller KL divergence.

Benefits of Including Ties

Tied Pairs = responses ranked differently by different metrics

Better performance across all metrics

Model	COMET	KIWI-22	XCOMET	KIWI-XXL
ALMA-7B-LoRA Xu et al. [2024]	79.78	76.81	83.94	73.65
+ DPO(CP)	79.66	77.73	88.87	74.12
+ DPO-RK(CP+TP)	**80.63**	**78.91**	**90.40**	**75.77**
+ DPO-D(CP+TP)	80.38	78.27	90.09	75.54

Table 1. ZH-EN translation performance on ALMA-R-Preference test set. The best result is reported for DPO(CP), DPO-RK(CP+TP) and DPO-D(CP+TP) over a beta sweep in [0.1, 0.3, 0.5, 0.7, 0.9].

Benefits of Including Ties

Tied Pairs = a random pair when all 8 samples are correct

Better performance via Higher Preservation of Correctness

β	SimPO (CP)	CPO (CP)	DPO (CP)	DPO-RK (CP+TP)	DPO-D (CP+TP)
0.1	82.5%	82.6%	76.4%	**83.5%**	81.7%
0.3	81.8%	83.1%	83.7%	**84.4%**	83.2%
0.5	81.8%	83.1%	83.6%	83.8%	**84.5%**
0.7	81.6%	82.6%	83.3%	83.7%	**84.5%**
1.0	82.2%	83.6%	83.5%	**84.1%**	83.7%

Table 2. GSM8K test set performance with greedy decoding after one-epoch of preference optimization for a range of beta values, evaluated by exact match after "####". The base Qwen2.5-3B-Instruct model scores 70.9%.

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF CAMBRIDGE

“DPO-RK and DPO-D explicitly model and consume ties compared to DPO. Using the otherwise discarded no-preference pairs can lead to stronger regularization and better task performance."

Paper: https://arxiv.org/abs/2409.17431

Code: https://github.com/EriChen0615/DPO-RKD

Correspondence: jc2124@cam.ac.uk