Conditional Equivalence of DPO and RLHF: Assumptions, Failure Modes, and Provable Alignment

Zhiqin Yang[♯], Yonggang Zhang[♯], Wei Xue[♯], Dong Fang[⊥], Bo Han[♮], Yike Guo[♯]

[♯]The Hong Kong University of Science and Technology [♮]Hong Kong Baptist University [⊥]LIGHTSPEED

ICML 2026

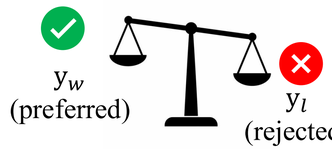
Question: Under what conditions can DPO be derived through

Why this paper?

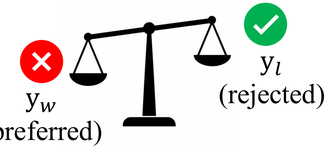
DPO is widely used because it is simpler than RLHF and often claimed to be theoretically equivalent.

This paper shows the equivalence is conditional, not universal.

Aligned reference



Misaligned reference



Key failure condition

Assumption violation when $\boldsymbol{\delta_{ref} \leq -\frac{\Delta r^*}{\beta}}$

DPO still minimize loss while preferring the rejected response.

Experimental results

Method	AlpacaEval 2			Arena-Hard	
	WR(%)	LC(%)	Avg Length	WR (%)	90% CI
SFT-Base	14.22	13.47	1972	19.2	(−1.4 / +1.5)
SLiC-HF	16.33	15.06	1998	22.9	(−1.4 / +1.6)
Contrastive-PO	18.94	15.95	2169	26.6	(−1.6 / +1.8)
DPO	24.60	25.09	1896	28.9	(−1.7 / +1.5)
RRHF	15.53	16.40	1830	20.5	(−1.5 / +1.6)
RDPO	24.25	25.01	1895	28.5	(−1.9 / +1.5)
ORPO	17.00	17.29	1876	21.7	(−1.1 / +1.3)
KTO	17.72	17.40	1925	23.1	(−1.0 / +1.1)
IPO	21.48	20.41	1993	26.9	(−1.6 / +1.6)
SimPO	23.48	25.91	1810	30.0	(−2.7 / +2.3)
CPO (Ours)	**25.15**	**26.57**	1879	**32.6**	(−1.9 / +2.4)

- CPO achieves the best overall performance across benchmarks.
- CPO remains strong while preserving competitive response length.

Methodology: Conditional equivalence analysis

➢ 1. Core relation from RLHF optimum

At the RLHF-optimal policy $\boldsymbol{\pi^*}$, the margin satisfies:

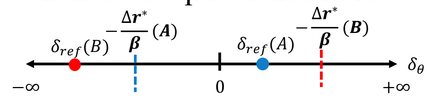
$$\boldsymbol{\delta_{\pi^*} = \delta_{ref} + \frac{\Delta r^*}{\beta}}$$

- $\boldsymbol{\delta_{ref} = log\ \frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}}$: reference log-margin
- $\boldsymbol{\delta_{\pi^*} = log\ \frac{\pi^*(y_w|x)}{\pi^*(y_l|x)}}$: optimal policy log-margin
- $\Delta \boldsymbol{r^* = r(y_w, x) - r(y_l, x)}$: reward gap (> 0)
- $\boldsymbol{\beta > 0}$: RLHF/DPO temperature

For RLHF-optimal policy $\boldsymbol{\pi^*}$ to still prefer the human-preferred response, we must have:

$$\boldsymbol{\delta_{\pi^*} > 0}$$

➢ 2. When does equivalence holds?



Case A
Aligned reference

$\boldsymbol{\delta_{ref} > -\frac{\Delta r^*}{\beta}}$

↓

$\boldsymbol{\delta_{\pi^*} > 0}$

Equivalence holds:
RLHF-optimal still prefers human preference

Case B
Misaligned reference

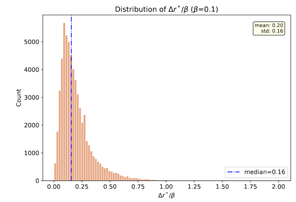
$\boldsymbol{\delta_{ref} \leq -\frac{\Delta r^*}{\beta}}$

↓

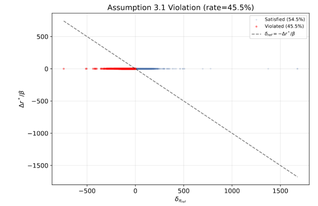
$\boldsymbol{\delta_{\pi^*} \leq 0}$

Equivalence fails:
RLHF-optimal no loger prefers human preference

➢ 3. Failure evidence



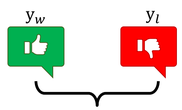
The distribution of $\frac{\Delta r^*}{\beta}$



Violation fraction: 45.5%

How CPO fix the problem

1. Reference pair (y_w preferred over y_l)



2. Compute margins $\boldsymbol{\delta_{ref}}$ and $\boldsymbol{\delta_\theta}$

$\boldsymbol{\delta_{ref}}$ from reference
$\boldsymbol{\delta_\theta}$ from current policy

3. Compute corrected margin

$$\tilde{\gamma}_{\text{ref}}(x, y_w, y_l) = \gamma\left(\frac{1}{\pi_{\text{ref}}(y_w|x)} + \frac{1}{\pi_{\text{ref}}(y_l|x)}\right)$$

Use reference-based correction term

4. Add adaptive correction or conservative margin

Apply $\tilde{\gamma}_{ref}$ or $\boldsymbol{\beta \Phi_{cons}(\delta_{ref})}$

5. Obtain non-negative effective margin

$\boldsymbol{\delta_\theta - \delta_{ref} - corr \geq 0}$
Optimize to increase it

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(\beta(\delta_{\pi_\theta} - \delta_{\pi_{\text{ref}}}))\right]$$

$$\mathcal{L}_{\text{CPO}}(\pi_\theta) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta(\delta_{\pi_\theta} - \delta_{\pi_{\text{ref}}}) - \tilde{\gamma}_{\text{ref}}(x, y_w, y_l)\right)\right]$$

$$\mathcal{L}_{\text{E-CPOC}}(\pi_\theta) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\left(\delta_{\pi_\theta} - \delta_{\pi_{\text{ref}}}\right) - \beta\Phi_{\text{cons}}(\delta_{\pi_{\text{ref}}})\right)\right]$$

Theoretical guarantees

- Conditional equivalence criterion: DPO = RLHF iff $\boldsymbol{\delta_{ref} > -\frac{\Delta r^*}{\beta}}$
- CPO enforces a non-negative effective margin.
- CPO avoids convergence to undesirable space $\mathcal{U}$.

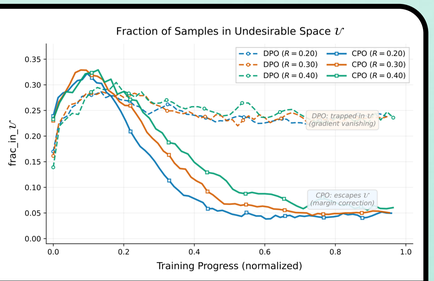
Analysis

Varying reference policy quality

Corruption Ratio	$\delta_{\pi_{ref}} < 0$ (%)	Assumption 3.1 Violated (%)
$R = 0.2$	53.2	52.9
$R = 0.3$	56.9	56.8
$R = 0.4$	60.1	60.0

AlpacaEval 2 performance under misaligned reference policies.

Corruption Ratio	Method	WR(%)	LC(%)	Avg Length
$R = 0.2$	DPO	16.82	17.23	1958
	CPO	**22.47**	**27.60**	1699
$R = 0.3$	DPO	14.99	15.48	1894
	CPO	**22.91**	**27.35**	1686
$R = 0.4$	DPO	15.43	15.98	1907
	CPO	**20.54**	**24.34**	1714



Contact Me

Linkedin

Wechat

Paper Detail

ArXiv

Code