Importance Weighting for Aligning Language Models under Deployment Distribution Shift

Thanawat Lodkaew[1] Tongtong Fang[2] Takashi Ishida[3,1] Masashi Sugiyama[3,1]

[1]The University of Tokyo [2]The Institute of Statistical Mathematics [3]RIKEN AIP

TRANSACTIONS on ML RESEARCH (tmlr)

I. Summary

i. **Motivation.** Training and deployment objectives often differ. For example, models are trained for helpfulness but deployed for harmlessness, creating a *deployment distribution shift*.

ii. **Key assumption.** Within the training dataset, some instances are *useful* (relevant), such as those containing helpful and harmless responses, for optimizing performance under the deployment distribution. In contrast, others are *not useful* (irrelevant), such as those that are helpful but harmful responses.

iii. **Method.** Inspired by [1], we propose an *importance weighting (IW)* method tailored for *direct preference optimization (DPO)* [2], IW-DPO, to mitigate this distribution shift by estimating *importance weights* through density ratio estimation between training and validation data, upweighting relevant instances and downweighting irrelevant ones to better align with the deployment distribution.

iv. **Results.** Experimental results under various distribution shift scenarios using multiple datasets demonstrate the effectiveness of our approach, with approximately *4% overall win rate improvement* over the standard DPO.

II. Deployment Distribution Shift

i. Definition

The deployment environment (deployment dist.) changes in ways not reflected in the training dataset (training dist.) due to *changes in end-user behavior, preferences, etc.*

$$p_{\mathrm{tr}}(x, y_1, y_2, b) \neq p_{\mathrm{te}}(x, y_1, y_2, b)$$

ii. Factors of distribution shift

$$p(x, y_1, y_2, b) = \underbrace{p(x)}_{1}\underbrace{p(y_1, y_2 \mid x)}_{2}\underbrace{p(b \mid x, y_1, y_2)}_{3}$$

$p_{\mathrm{tr}}(x) \neq p_{\mathrm{te}}(x)$ — Prompt

$p_{\mathrm{tr}}(y_1, y_2 \mid x) \neq p_{\mathrm{te}}(y_1, y_2 \mid x)$ — Response

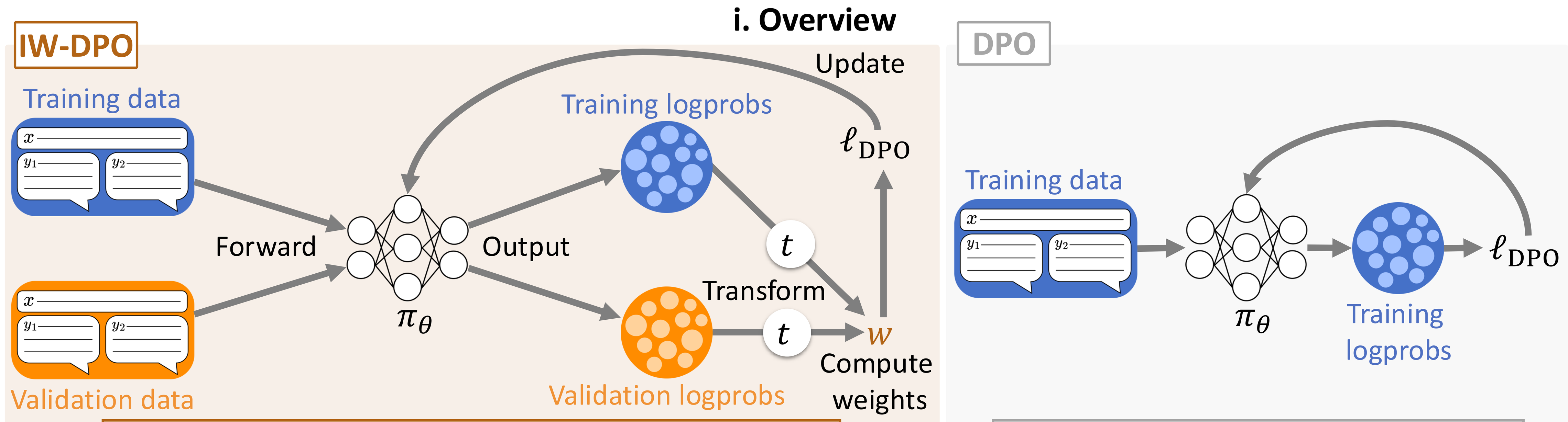
$p_{\mathrm{tr}}(b \mid x, y_1, y_2) \neq p_{\mathrm{te}}(b \mid x, y_1, y_2)$ — Preference label

iii. Distribution shift types

	Type of shift	Factor 1	Factor 2	Factor 3
a	No shift			
b	Full shift	✓	✓	✓
c	Prompt shift	✓		
d	Response shift		✓	
e	Preference label shift			✓
f	Prompt + response shift	✓	✓	
g	Prompt + preference label shift	✓		✓
h	Response + preference label shift		✓	✓

III. Importance Weighted Direct Preference Optimization

i. Overview



IW-DPO: $\min_{\pi_\theta} \mathbb{E}_{p_{\mathrm{tr}}(x,y_1,y_2,b)}[w(x, y_1, y_2, b)\ell_{\mathrm{DPO}}(x, y_1, y_2, b)]$

DPO: $\min_{\pi_\theta} \mathbb{E}_{p_{\mathrm{tr}}(x,y_1,y_2,b)}[\ell_{\mathrm{DPO}}(x, y_1, y_2, b)]$

ii. Problem setting

Training and validation datasets are available, with the constraint that $N_{\mathrm{v}} \ll N_{\mathrm{tr}}$

$$\mathcal{D}_{\mathrm{tr}} = \{(x^{\mathrm{tr},i}, y_1^{\mathrm{tr},i}, y_2^{\mathrm{tr},i}, b^{\mathrm{tr},i})\}_{i=1}^{N_{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{tr}}(x, y_1, y_2, b)$$

$$\mathcal{D}_{\mathrm{v}} = \{(x^{\mathrm{v},i}, y_1^{\mathrm{v},i}, y_2^{\mathrm{v},i}, b^{\mathrm{v},i})\}_{i=1}^{N_{\mathrm{v}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{te}}(x, y_1, y_2, b)$$

Goal is to *optimize for the test distribution*

$$J(\pi_\theta) = \mathbb{E}_{p_{\mathrm{te}}(x,y_1,y_2,b)}[\ell_{\mathrm{DPO}}(x, y_1, y_2, b)]$$

iii. Definition of importance weight and training objective

Assume support of the training distribution covers that of the test distribution

$$\mathrm{supp}(p_{\mathrm{te}}) \subseteq \mathrm{supp}(p_{\mathrm{tr}})$$

$$w^\star(x, y_1, y_2, b) = p_{\mathrm{te}}(x, y_1, y_2, b)/p_{\mathrm{tr}}(x, y_1, y_2, b)$$

$$J(\pi_\theta) = \mathbb{E}_{p_{\mathrm{tr}}(x,y_1,y_2,b)}[w^\star(x, y_1, y_2, b)\ell_{\mathrm{DPO}}(x, y_1, y_2, b)] = J_{\mathrm{tr}}(\pi_\theta, w^\star)$$

Empirical training objective

$$\hat{J}(\pi_\theta) = \frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} w^{\mathrm{tr},i}\ell_{\mathrm{DPO}}(x^{\mathrm{tr},i}, y_1^{\mathrm{tr},i}, y_2^{\mathrm{tr},i}, b^{\mathrm{tr},i})$$

iv. Importance weight estimation

Transformation function $t: (x, y_1, y_2, b) \mapsto z$

↓

Transformed data

$$Z_{\mathrm{tr}} = \{t(x^{\mathrm{tr},i}, y_1^{\mathrm{tr},i}, y_2^{\mathrm{tr},i}, b^{\mathrm{tr},i})\}_{i=1}^{N_{\mathrm{tr}}}$$

$$Z_{\mathrm{v}} = \{t(x^{\mathrm{v},i}, y_1^{\mathrm{v},i}, y_2^{\mathrm{v},i}, b^{\mathrm{v},i})\}_{i=1}^{N_{\mathrm{v}}}$$

↓

Density ratio estimator $\boldsymbol{w} = \omega(Z_{\mathrm{tr}}, Z_{\mathrm{v}})$

v. Choices of transformation function

Loss (IW-DPO-L)

$$t: (x, y_1, y_2, b) \mapsto \ell_{\mathrm{DPO}}(x, y_1, y_2, b)$$

$$\ell_{\mathrm{DPO}}(x, y_1, y_2, b) = -\log\sigma(b \cdot (r(x, y_1) - r(x, y_2)))$$

Reward (IW-DPO-R)

$$t: (x, y_1, y_2, b) \mapsto \hat{r}(x, y_1, y_2, b)$$

$$\hat{r}(x, y_1, y_2, b) = (r(y_1), r(y_2))$$

$$r(x, y) = \beta\log\frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}$$

IV. Experimental Scenarios

We simulated three deployment distribution shift scenarios

Helpful-Harmless LM

Training data: Helpful-Harmful responses + *Helpful-Harmless* responses

Test data: *Helpful-Harmless* responses

Shift type: d or h

Dataset: SafeRLHF [3]

Science LM

Training data: Science fiction-domain prompts + *Science-domain* prompts

Test data: *Science-domain* responses

Shift type: b or f

Dataset: SHP [4]

Culture-Aware LM

Training data: American-culture preference labels + *Indian-culture* preference labels

Test data: *Indian-culture* preference labels

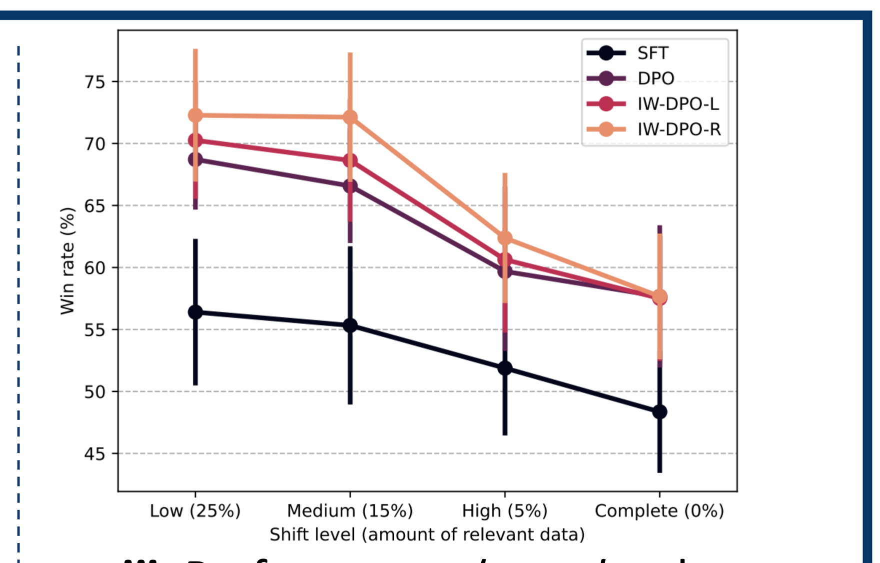
Shift type: e

Dataset: CALI [5]

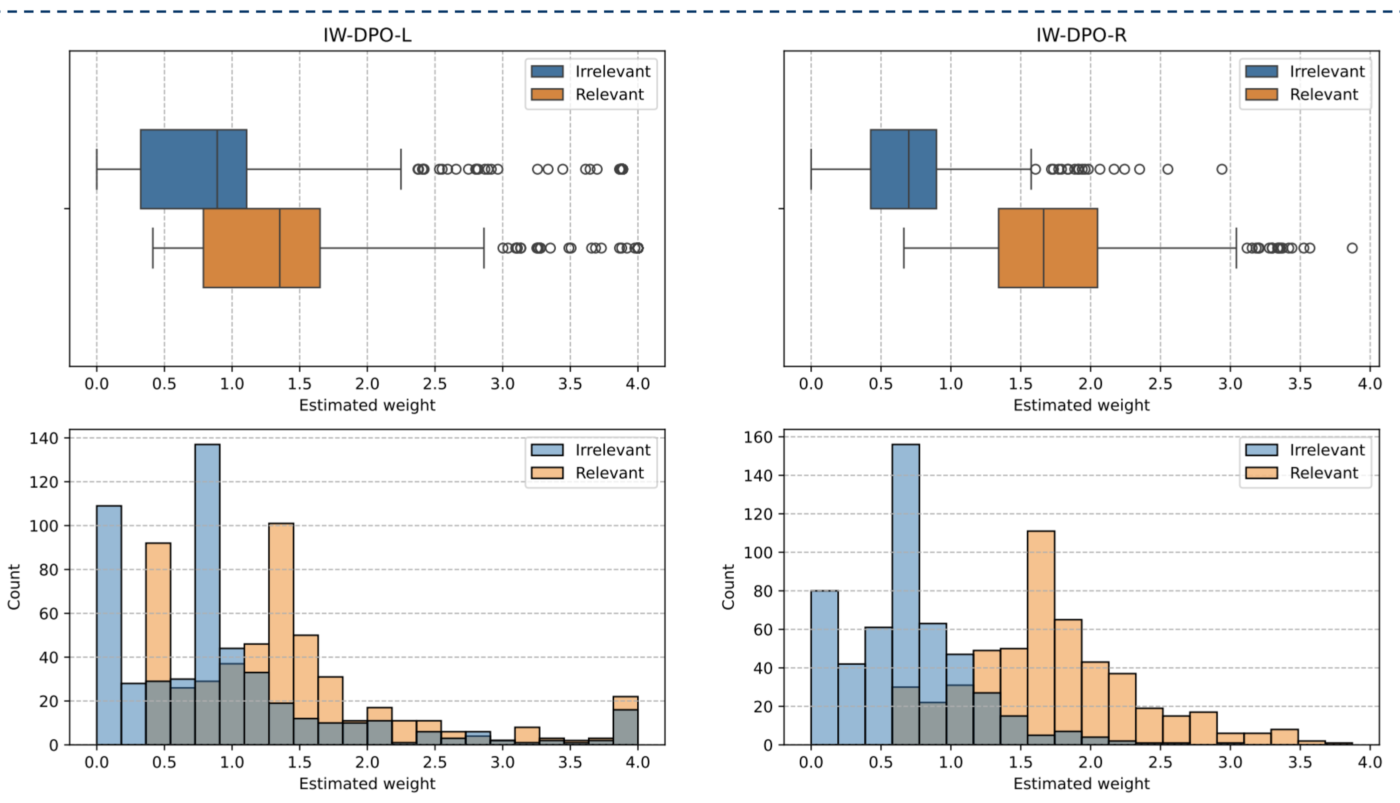
V. Results

Method	Helpful-Harmless LM	Science LM	Culture-Aware LM
SFT w/ $\mathcal{D}_{\mathrm{tr}}+\mathcal{D}_{\mathrm{v}}$	56.40 ± 5.12	47.06 ± 5.59	31.72 ± 3.13
DPO w/ $\mathcal{D}_{\mathrm{v}}$	60.48 ± 4.25	53.20 ± 5.14	32.15 ± 3.56
DPO w/ $\mathcal{D}_{\mathrm{tr}}+\mathcal{D}_{\mathrm{v}}$	68.71 ± 3.45	63.79 ± 3.45	35.62 ± 0.97
WPO (Zhou et al., 2024) w/ $\mathcal{D}_{\mathrm{tr}}+\mathcal{D}_{\mathrm{v}}$	70.26 ± 4.05	64.84 ± 5.22	$36.41 \pm 1.25^*$
IW-DPO-L	70.50 ± 3.46	$65.88 \pm 6.96^*$	$36.49 \pm 1.39^*$
IW-DPO-R	$\mathbf{72.28 \pm 4.62}$	$\mathbf{70.59 \pm 3.01}$	$\mathbf{36.92 \pm 1.77}$

i. *More* improvement in win rate in the *Helpful-Harmless LM* and *Science LM* scenarios, but *less* in the *Culture-Aware LM* scenario



iii. Performance *degrade* when distribution shift becomes *more severe*



ii. Importance weight differences become *clearer* with *IW-DPO-R*!

Scenario	Method	Density ratio estimator	Win/Match rate (%)
Helpful-Harmless LM	IW-DPO-L	KMM	$70.50 \pm 3.46^*$
		KLIEP	70.10 ± 4.39
		RuLSIF	$\mathbf{72.28 \pm 4.94}$
	IW-DPO-R	KMM	$72.28 \pm 4.62^*$
		KLIEP	$71.88 \pm 4.20^*$
		RuLSIF	$\mathbf{73.19 \pm 3.39}$
Science LM	IW-DPO-L	KMM	$65.88 \pm 6.96^*$
		KLIEP	$\mathbf{68.10 \pm 2.66}$
		RuLSIF	$67.58 \pm 3.32^*$
	IW-DPO-R	KMM	$\mathbf{70.59 \pm 3.01}$
		KLIEP	$69.28 \pm 4.45^*$
		RuLSIF	$70.59 \pm 4.68^*$
Culture-Aware LM	IW-DPO-L	KMM	$36.49 \pm 1.39^*$
		KLIEP	$\mathbf{37.83 \pm 2.68}$
		RuLSIF	$36.45 \pm 0.70^*$
	IW-DPO-R	KMM	$36.92 \pm 1.77^*$
		KLIEP	$36.25 \pm 1.36^*$
		RuLSIF	$\mathbf{38.38 \pm 1.46}$

iv. Choice of density ratio estimator is *not significant*

References

[1] T. Fang et al. Rethinking Importance Weighting for Deep Learning under Distribution Shift. In NeurIPS, 2020.
[2] R. Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In NeurIPS, 2024.
[3] J. Ji et al. BeaverTails: Towards Improved Safety Alignment of LLM via a Human Preference Dataset. In NeurIPS, 2023.
[4] K. Ethayarajh et al. Understanding Dataset Difficulty with V-Usable Information. In ICML, 2022.
[5] J. Huang et al. Culturally Aware Natural Language Inference. In EMNLP, 2023.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所 The Institute of Statistical Mathematics

RIKEN

東京大学 THE UNIVERSITY OF TOKYO