Robust Direct Preference Optimization for Offline Learning

Extended Abstract

Shihab Ahmed
University of Central Florida
Orlando, FL, United States
Shihab.Ahmed@ucf.edu

Zhenyi Wang
University of Central Florida
Orlando, FL, United States
Zhenyi.Wang@ucf.edu

Yue Wang
University of Central Florida
Orlando, FL, United States
Yue.Wang@ucf.edu

ABSTRACT

Recent preference-based alignment methods, such as Direct Preference Optimization (DPO), perform well under comprehensive offline datasets. However, practical datasets often contain sparse, noisy, and unevenly distributed comparisons, which can degrade model performance. To address this, we first adopt the principle of pessimism and propose a Robust DPO framework that optimizes for the worst-case reward within some data-dependent uncertainty set, and show its effectiveness in offline problems. Moreover, we show that the resulting robust optimal policy can be obtained by directly fine-tuning a baseline DPO model, avoiding the need for retraining. We further construct an uncertainty set to tackle the data uncertainty, based on the graph Laplacian, and show the set contains the true underlying reward with high probability. We then further evaluate the effectiveness of our method in controlled tabular and LLM setting, which validate our theoretical finds.

KEYWORDS

preference alignment, offline learning, robust optimization

ACM Reference Format:
Shihab Ahmed, Zhenyi Wang, and Yue Wang. 2026. Robust Direct Preference Optimization for Offline Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/UTXZ7394

1 INTRODUCTION

Aligning language models with human preferences is critical for producing helpful and safe outputs [2, 9]. Reinforcement Learning from Human Feedback (RLHF) [6, 17, 21] has become the standard approach, training a reward model on preference data and optimizing policies via reinforcement learning. Direct Preference Optimization (DPO) [19] and its variants [1, 7] simplify this pipeline by directly optimizing policies on preference data without explicit reward modeling.

However, DPO and its variants generally assume preference dataset is accurate and comprehensive. In practice, datasets are often *sparse*, *noisy*, and *unevenly distributed*—certain actions may have few comparisons, labels may be inconsistent, and coverage may vary across the action space [4, 5]. Under such conditions, DPO can become overconfident in poorly supported directions, amplifying noise and degrading robustness.

This challenge mirrors the distributional shift problem in *offline reinforcement learning*, where policies must be learned from fixed datasets without further interaction [8, 15, 22]. The principle of *pessimism*—optimizing under conservative value estimates—has proven effective in offline RL [12–14, 23, 24]. Identifying the success of it in offline RL, we naturally ask: *Can pessimism be adapted to preference alignment to tackle imperfect offline datasets?*

In this paper, we answer it affirmatively by developing **Robust DPO**, a framework that pessimistically optimizes for the worst-case reward within a data-dependent uncertainty set.

2 PRELIMINARIES: RLHF AND DPO

For an underlying implicit reward $r(s, a) \in \mathbb{R}$ for all (s, a)-pairs, and the probability of preference with the Bradley-Terry-Luce (BTL) model [3, 6, 17]:

$$\mathbb{P}(a_1 \succ a_2 \mid s) = \sigma(r(s, a_1) - r(s, a_2)) .$$

The goal is to find the optimal policy $\pi^*(s)$:

$$\arg\max_{\pi} \left\{ \mathbb{E}_{a\sim\pi(\cdot|s)} \left[r(s, a) \right] - \beta \, \mathrm{KL}(\pi(\cdot \mid s) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid s)) \right\} \quad (1)$$

Standard RLHF first learns the reward model $\hat{r}(s, \cdot)$ from the dataset through maximum likelihood estimator (MLE) and then applies standard RL methods. Direct Preference Optimization (DPO) [19] bypasses explicit reward modeling and directly optimizes the policy using the preference data. Given a reference policy π_{ref}, DPO directly solves (1) for its closed-form solution:

$$\pi^*(a|s) = \frac{1}{Z(s)} \pi_{\mathrm{ref}}(a|s) \exp\left(\frac{r(s, a)}{\beta} \right), \quad (2)$$

where $Z(s) = \sum_{a'} \pi_{\mathrm{ref}}(a'|s) \exp(r(s, a')/\beta)$ is the normalization function. Hence, DPO performs a soft reward-weighted improvement step from the pre-trained policy π_{ref}.

3 ROBUST DPO

Consider a state s and suppose we have obtained an estimate $\hat{r}(s, \cdot)$ of the latent reward function. Let $\kappa(s, a) \geq 0$ represent a per-action uncertainty radius which quantifies how accurate the estimation is. We define the uncertainty set as

$$\mathcal{R}_{s,a} = \left\{ r : |r(s, a) - \hat{r}(s, a)| \leq \kappa(s, a) \right\}. \quad (3)$$

This interval-based set captures all reward vectors whose deviation from $\hat{r}$ is controlled action-wise by the radius R_0. Rather than optimizing expected reward under the nominal estimate $\hat{r}$, we can

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UTXZ7394

maximize the *worst-case* reward over $\mathcal{R}_{s,a}$:

$$\pi_r^*(s) = \arg\max_{\pi\in\Delta(\mathcal{A})} \min_{r\in\mathcal{R}_{s,a}} \left\{ \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)] - \beta\,\mathrm{KL}(\pi \,\|\, \pi_{\mathrm{ref}}) \right\}. \tag{4}$$

This max – min formulation (also known as robust optimization [11, 16]) optimize for the worst-case performance in the uncertainty set. When the uncertainty set is carefully designed to include the true reward (with high probability), the learned policy π_r^* provides an optimized lower bound on the true performance and ensures a guarnteed baseline performance.

Since the uncertainty set (3) is interval-based, the inner minimization over $r \in \mathcal{R}_{s,a}$ admits a simple solution: for any distribution π, the worst-case reward is attained at $\underline{r}(s,a) = \hat{r}(s,a) - \kappa(s,a)$. Substituting this into (4) and solving the KL-regularized maximization yields the optimal policy:

$$\pi_r^*(a|s) = \frac{\pi_{\mathrm{ref}}(a|s)\exp\left(\frac{\hat{r}(s,a)-\kappa(s,a)}{\beta}\right)}{Z'(s)}, \tag{5}$$

where $Z'(s)$ is the normalization constant. Equation (5) shows that robustness enters as an *action-wise penalty*: each action's estimated reward $\hat{r}(s,a)$ is reduced by its uncertainty radius $\kappa(s,a)$ before exponentiation. Actions with larger uncertainty receive stronger down-weighting, implementing pessimism at the per-action level.

Comparing with (5), we obtain the key relationship:

$$\frac{\pi_r^*(a|s)}{\pi^*(a|s)} = \frac{Z(s)}{Z'(s)}\exp\left(-\tfrac{\kappa(s,a)}{\beta}\right). \tag{6}$$

This indicates that we can directly obtain π_r^* by *reweighting* π^* with the corresponding uncertainty radii. The multiplicative factor $\exp(-\kappa(s,a)/\beta)$ down-weights actions proportionally to their uncertainty, ensuring that the robust policy remains conservative in data-scarce or noisy regions of the action space.

3.1 Design of Uncertainty Sets

To apply our robust DPO to offline alignment problems, we need to design such uncertainty sets to include the true rewards. We note that pairwise comparisons induce a natural graph structure [20, 25]. For each state s, define the comparison graph G_s with $\mathcal{A}$ actions as vertices and comparison counts $N_{ij}(s)$ as edge weights. The weighted graph Laplacian $L_s \in \mathbb{R}^{\mathcal{A}\times\mathcal{A}}$ encodes data connectivity via quadratic form: for any $v \in \mathbb{R}^{\mathcal{A}}$,

$$v^\top L_s v = \sum_{i<j} N_{ij}(s)(v_i - v_j)^2. \tag{7}$$

Applied to the reward vector $r(s) \in \mathbb{R}^{\mathcal{A}}$, this measures total variation of rewards across observed comparisons. Rarely compared actions yield high effective resistance $\Omega_s(i,j) = (e_i - e_j)^\top L_s^\dagger (e_i - e_j)$, indicating higher uncertainty.

Theorem 3.1 (Ellipsoidal Concentration). *For a state $s \in \mathcal{S}$ with connected comparison graph, let $\hat{r}(s)$ be the maximum-likelihood estimate under identifiability $\mathbf{1}^\top r(s) = 0$, with bounded rewards $|r(s,a)| \le B$. Then with probability $1-\delta$:*

$$\|r(s) - \hat{r}(s)\|_{L_s} \le \sqrt{\rho_s(\delta)}. \tag{8}$$

where the ellipsoid radius is $\rho_s(\delta) = \frac{C}{\gamma(B)^2}\left(\mathcal{A} + \log\frac{1}{\delta}\right)$, with $\gamma(B) = \min_{|z|\le 2B}\sigma(z)[1-\sigma(z)] > 0$, $\sigma(z) = 1/(1+e^{-z})$.

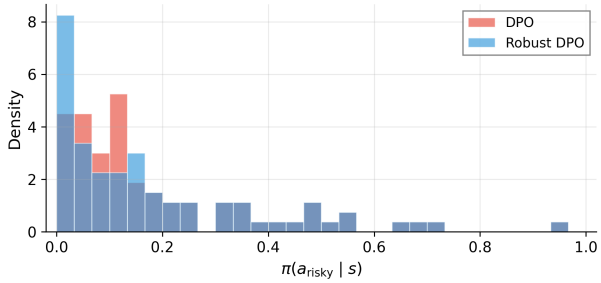

Figure 1: Robust DPO suppresses risky-action probabilities in uncertain states (left) and shifts margin distributions leftward (right), reflecting pessimistic contraction.

The proof combines M-estimation theory with the Hanson–Wright inequality for score concentration. From Theorem 3.1, the per-action penalty follows via the support function of the ellipsoid:

$$\kappa(s,a) = \sqrt{\rho_s(\delta)} \cdot \sqrt{e_a^\top L_s^\dagger e_a}, \tag{9}$$

which ensures $|\hat{r}(s,a) - r(s,a)| \le \kappa(s,a)$ with probability $\ge 1-\delta$. We thus utilize it to construct the uncertainty set, which ensures that the true reward $r(s,a) \in \mathcal{R}_{s,a}$ as required by (3).

4 EXPERIMENTS

We validate robust DPO in both controlled tabular settings and large-scale LLM alignment.

Tabular Domain. In a gridworld with risky/safe actions and Zipf-distributed comparison coverage, we test under 2× environment stochasticity. Robust DPO achieves 6.9% higher return than DPO (0.609 vs. 0.570) with 73% lower constraint violations. Figure 1 shows that π_r^* suppresses risky-action probabilities in uncertain states.

LLM Alignment. We validate inference-time adaptation without retraining. Given a base DPO model π_0, we compute an ambiguity score u_A based on the margin $m_i = \log(p_{\pi_0}(y^+|x)/p_{\pi_0}(y^-|x))$ and reweight outputs as $\tilde{\pi}(y|x) \propto \pi(y|x)\exp(-\rho\, u_A)$, downweighting responses on ambiguous prompts. Table 1 shows that on Qwen2.5-0.5B [18] with LoRA [10], reweighting preserves accuracy while increasing the log-probability gap, validating that robust DPO generalizes to deployment-time updates.

Table 1: Post-hoc reweighting on HH-RLHF (Qwen2.5-0.5B) increases preference margin while preserving accuracy.

Method	Pref. Acc. (95% CI)	logp gap
Vanilla DPO	0.58 [0.48, 0.67]	21.97
Weighted DPO (ρ=1)	0.58 [0.48, 0.67]	22.20
Weighted DPO (ρ=3)	0.58 [0.48, 0.68]	23.05

5 ACKNOWLEDGEMENT

The work is partially supported by DARPA under Agreement No. HR0011-24-9-0427.

REFERENCES

[1] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4447–4455.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[3] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[4] Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. 2024. Robust reinforcement learning from corrupted human feedback. *Advances in Neural Information Processing Systems* 37 (2024), 124093–124113.

[5] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably Robust DPO: Aligning Language Models with Noisy Feedback. arXiv:2403.00409 [cs.LG] https://arxiv.org/abs/2403.00409

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[7] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. arXiv:2402.01306 [cs.LG] https://arxiv.org/abs/2402.01306

[8] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. arXiv:1812.02900 [cs.LG] https://arxiv.org/abs/1812.02900

[9] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462 [cs.CL] https://arxiv.org/abs/2009.11462

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[11] Garud N Iyengar. 2005. Robust dynamic programming. *Mathematics of Operations Research* 30, 2 (2005), 257–280.

[12] Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2020. Is Pessimism Provably Efficient for Offline RL? *arXiv preprint arXiv:2012.15085* (2020).

[13] Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021. Is Pessimism Provably Efficient for Offline RL?. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5084–5096. https://proceedings.mlr.press/v139/jin21e.html

[14] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems* 33 (2020), 1179–1191.

[15] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643 [cs.LG] https://arxiv.org/abs/2005.01643

[16] Arnab Nilim and Laurent El Ghaoui. 2004. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*. 839–846.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744.

[18] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[20] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. 2016. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research* 17, 58 (2016), 1–47.

[21] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems* 33 (2020), 3008–3021.

[22] Yue Wang, Jinjun Xiong, and Shaofeng Zou. 2024. Achieving the Asymptotically Minimax Optimal Sample Complexity of Offline Reinforcement Learning: A DRO-Based Approach. *Transactions on Machine Learning Research* (2024).

[23] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. 2021. Bellman-consistent Pessimism for Offline Reinforcement Learning. *arXiv preprint arXiv:2106.06926* (2021).

[24] Chi Zhang, Ziying Jia, George K Atia, Sihong He, and Yue Wang. 2025. Pessimism Principle Can Be Effective: Towards a Framework for Zero-Shot Transfer Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*.

[25] Banghua Zhu, Michael Jordan, and Jiantao Jiao. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*. PMLR, 43037–43067.