Robust Direct Preference Optimization is Effective in Offline Learning

Anonymous ACL submission

Abstract

Direct Preference Optimization (DPO) is widely studied and used in preference alignment problems. However, when the offline datasets are sparse, limited, imbalanced, or noisy, due to the constrained collection processes, DPO may suffer from performance degradation. Inspired by the pessimism principle in offline learning, we propose Robust DPO (rDPO), a pessimistic preference-optimization framework that accounts for dataset uncertainties by optimizing against the worst-case latent reward within a data-dependent uncertainty set. We show that our rDPO enjoys a simple structure and can be directly fine-tuned from the vanilla DPO policies. Moreover, We theoretically prove the effectiveness of our rDPO, showing it learns a policy robust to the dataset uncertainties. We further empirically verify that rDPO improves robustness in both controlled synthetic environments under sparse/noisy comparisons, and language-model preference tuning under targeted corruption.

1 Introduction

One of the central challenges of Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; Achiam et al., 2023) is the alignment with human values or preferences, whereas non-aligned models may generate outputs that are unhelpful, toxic, or factually incorrect (Gehman et al., 2020; Bai et al., 2022b; Ganguli et al., 2023). Such alignment models are generally trained on offline datasets of preferences collected from human experts. Besides supervised fine tuning (SFT) (Wei et al., 2022; Ouyang et al., 2022; Zhou et al., 2023), Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020) is widely studied and used to align the model with the preference data (Bradley and Terry, 1952; Luce et al., 1959). Despite successes, RLHF learns the reward model first, which can be complex and prone to reward hacking (Casper et al., 2023; Perez et al., 2022; Amodei et al., 2016). Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants (Rafailov et al., 2023; Xu et al., 2024; Azar et al., 2024; Ethayarajh et al., 2024) elegantly sidestep these issues via a reparameterization enabling direct policy optimization on preference data, and receive great success.

However, the performance of these alignment methods heavily relies on the quality and quantity of the preference dataset. In practice, the datasets are generally limited, sparse, noisy, or unevenly distributed due to the costly data collection process, which introduces uncertainty to the dataset. Yet vanilla RLHF do not account for these uncertainties, resulting in an inaccurate reward, and can suffer from severe performance degradation (Casper et al., 2023; Banerjee and Gopalan, 2024; Zhang et al., 2025).

Our approach. To address these issues and tackle dataset uncertainties, we develop a robust preference alignment framework that adopts the principle of pessimism in tackling uncertainty. This principle has been widely studied in robust and offline reinforcement learning (Buckman et al., 2020; Wang et al., 2024; Xie et al., 2021a; Jin et al., 2020; Iyengar, 2005; Nilim and El Ghaoui, 2004; Bagnell et al., 2001; Satia and Lave Jr, 1973; Wiesemann et al., 2013; Tamar et al., 2014; Xu and Mannor, 2010), and robust optimization (Rahimian and Mehrotra, 2019; Kuhn et al., 2025) which penalizes overconfident extrapolation, mitigating distributional shift (Fujimoto et al., 2019; Kumar et al., 2020; Uehara and Sun, 2023). Inspired by these studies, we first characterize the uncertainties within the preference data by constructing an uncertainty set on the estimated reward. The size of such an uncertainty set is determined by the statistical errors of reward learning from the dataset, based on the comparison graph geometry. We then optimize for the worst case within it, to conservatively

tackle the aforementioned uncertainty. We further show that, despite the construction of the uncertainty set being based on the estimated reward, we can bypass the reward learning phase, and directly learn the robust policy through our robust variant of DPO. We hence propose our robust DPO (rDPO), which is implementable via two complementary views: (i) a *margin-shift* approach that subtracts uncertainty-dependent corrections from pairwise preference margins (a drop-in replacement for standard DPO loss), and (ii) a *reweighting* approach that robustifies learning by downweighting statistically unreliable comparisons with no additional retraining at all. We also theoretically characterize the sub-optimality of the learned policy, and empirically validate our claims in both controlled synthetic settings and LLM preference tuning under realistic dataset imperfections.

2 Related Work

Reinforcement Learning from Human Feedback. RLHF has emerged as the dominant paradigm for aligning language models with human values (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022b). The standard pipeline involves supervised fine-tuning, training a reward model on preference data using Bradley-Terry or Thurstone-Mosteller formulations (Bradley and Terry, 1952; Thurstone, 1927), and policy optimization via PPO (Schulman et al., 2017). RLHF demonstrated success in instruction-following (Ouyang et al., 2022), summarization (Stiennon et al., 2020), and safety alignment (Bai et al., 2022a), but suffers from computational complexity and training instability (Amodei et al., 2016; Casper et al., 2023; Perez et al., 2022).

Direct Preference Optimization (DPO) (Rafailov et al., 2023) massively simplifies RLHF by bypassing explicit reward modeling through a closed-form objective built on the Bradley-Terry model. This sparked development of contrastive methods including IPO (Azar et al., 2024), which addresses DPO overfitting via regularization; SLiC (Zhao et al., 2023), using hinge loss; and KTO (Ethayarajh et al., 2024). Cal-DPO (Xiao et al., 2024) addresses DPO's contrastive nature ignoring absolute reward values by proposing calibration to ground truth rewards.

Offline RLHF and Pessimism. A central challenge in offline preference optimization is distributional shift between behavior and learned policies, extensively studied in offline RL (Levine et al., 2020; Fujimoto et al., 2019; Kumar et al., 2020; Kidambi et al., 2020; Uehara and Sun, 2021; Wang et al., 2016). The pessimism principle—systematically underestimating values for uncertain state-action pairs—yields provably robust algorithms (Buckman et al., 2020; Jin et al., 2021a). CQL (Kumar et al., 2020) penalizes Q-values for out-of-distribution actions; MOPO (Yu et al., 2020) and MOReL (Kidambi et al., 2020) use uncertainty penalties in model-based RL. These works establish that avoiding overoptimization in poorly supported directions is crucial in data-limited regimes. Preference optimization can thus be viewed as offline policy optimization in comparison space. Recent work (Chowdhury et al., 2024; Xu et al., 2024) studies DPO's susceptibility to noise and proposes variance-based weighting, but lacks systematic uncertainty quantification characterizing pessimism.

Robust Uncertainty-Aware Preference Optimization. Recent work aims to robustify preference optimization via uncertainty signals. Some approaches calibrate or rescale logits/rewards to mitigate temperature mismatch (Xiao et al., 2024; Mao et al., 2024), treating uncertainty as global scaling rather than local, pair-specific confidence. Another thread introduces distributional robustness (Wu et al., 2024; Bukharin et al., 2024; Mandal et al., 2025), formulating distributionally robust optimization over rewards or preference distributions to guard against worst-case perturbations and OOD drift. These provide strong distribution-level guarantees but often require auxiliary critics, adversarial inner loops, or additional networks, increasing computational overhead.

Our work contributes local, per-comparison pessimism with finite-sample guarantees in the tabular setting. We introduce spectral uncertainty penalties derived from comparison graph structure and design a robust DPO framework adopting the pessimism principle from offline RL. Unlike prior heuristic treatments, we provide finite-sample guarantees connecting policy suboptimality directly to spectral properties of the offline comparison graph, offering a principled foundation for robust preference optimization.

3 Preliminaries

3.1 Setup and Notations

We consider the problem of offline learning from preferences. Let $\mathcal{S}$ be the prompt space and $\mathcal{A} =$

$\{1, \ldots, A\}$ be the action space containing all possible responses in a finite setting. For each prompt $s \in \mathcal{S}$, a pairwise comparison dataset is collected:

$$\mathcal{D}_s = \{(a_i, a_j, y) \mid a_i, a_j \in \mathcal{A}, y \in \{0, 1\}\},$$

where $y = 1$ indicates preference for response a_i over a_j, and $y = 0$ indicates a_j is preferred over a_i. The learner's goal is to infer a policy $\pi(a|s)$ that produces preferred responses.

Notations. Let $N_{ij}(s)$ denote the count of comparisons between actions i and j, and $W_{ij}(s)$ the number of times i is preferred over j. We denote by $\delta_a \in \mathbb{R}^A$ the indicator vector for action a (i.e., $\delta_a(a') = \mathbb{I}[a' = a]$). The space of probability distributions over $\mathcal{A}$ is denoted by $\Delta(\mathcal{A})$. We let $\mathcal{D} = \cup_s \mathcal{D}_s$ denote the global dataset and express expectations $\mathbb{E}_{(s,a^+,a^-)\sim\mathcal{D}}$ over the empirical distribution of comparisons.

3.2 RLHF and DPO

The Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022) provides a probabilistic framework for pairwise preference generation. An underlying implicit reward $r(s, a) \in \mathbb{R}$ is assumed for all (s, a)-pairs, and the probability of preference over action is given by the logistic distribution:

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

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

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

where π_{ref} is some reference policy from pretraining, and the temperature β controls the sharpness of preference alignment. Without loss of generality, we assume $\sum_{a\in\mathcal{A}} r(s, a) = 0, \forall s \in \mathcal{S}$.

RLHF. Standard RLHF first learns the reward model $\hat{r}(s, \cdot)$ from the dataset through maximum likelihood estimator (MLE) as:

$$\arg\max_{r} \Big\{ \sum_{(a_i,a_j,y)\in\mathcal{D}_s} \Big[y \log \sigma\big(r(a_i) - r(a_j)\big) + (1 - y) \log \sigma\big(r(a_j) - r(a_i)\big) \Big] \Big\},$$

and then applies standard RL methods.

DPO. Direct Preference Optimization (DPO) (Rafailov et al., 2023) 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:

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

where $Z(s) = \sum_{a'} \pi_{\mathrm{ref}}(a'|s) \exp(r(s, a')/\beta)$ is the normalization function. It then directly plugs in the MLE for r, resulting in an optimization problem

$$\arg\max_{\pi} \Big\{ \mathbb{E}_{(s,a^+,a^-)\sim\mathcal{D}} \Big[\log \sigma\Big(\beta \log \frac{\pi(a^+ \mid s)}{\pi_{\mathrm{ref}}(a^+ \mid s)} - \beta \log \frac{\pi(a^- \mid s)}{\pi_{\mathrm{ref}}(a^- \mid s)}\Big)\Big] \Big\},$$

where a^+, a^- denote the preferred and unpreferred actions, and the optimization can be directly solved through the dataset, without estimating the reward function. Notably, DPO performs a soft reward-weighted improvement step from the pretrained policy π_{ref}.

4 Robust DPO for Offline Learning

We then present our robust DPO method.

4.1 Robust Alignment and Robust DPO

As mentioned, DPO may struggle when preference data is limited or noisy. First, the learned MLE reward estimate $\hat{r}(s, a)$ (RLHF) may be inaccurate when the dataset is limited or imbalanced (Jin et al., 2021b; Uehara and Sun, 2021; Xie et al., 2021a,b; Rashidinejad et al., 2021; Zanette et al., 2021; Yin and Wang, 2021; Shi et al., 2022; Zhan et al., 2022; Wang et al., 2023). Second, preferences themselves may be non-stationary and time-varying due to new information, social influences, and cultural trends (Zafari et al., 2019; Johnson and Mayorga, 2020; Caldwell, 1981), or the dataset generation may contain noise and uncertainties (Yang et al., 2024a; Chowdhury et al., 2024; Liang et al., 2024; Bukharin et al., 2024; Huang et al., 2025; Nishimori et al., 2025; Sahu and Wells, 2025). These mismatches motivate a robust approach grounded in data-driven uncertainty quantification.

To address estimation uncertainty, we adopt the principle of *pessimism*, widely used in offline reinforcement learning (Jin et al., 2021b; Uehara and Sun, 2021; Xie et al., 2021a,b; Rashidinejad et al.,

2021; Zanette et al., 2021; Yin and Wang, 2021; Shi et al., 2022; Zhan et al., 2022; Wang et al., 2023). Our approach frames robust preference optimization as a minimax problem over an uncertainty set, leading naturally to action-wise reweighting.

Uncertainty sets around nominal rewards. 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

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

This interval-based set captures all reward vectors whose deviation from $\hat{r}$ is controlled action-wise by the radius R_0.

Robust minimax objective. Rather than optimizing expected reward under the nominal estimate $\hat{r}$, we maximize the *worst-case* reward over $\mathcal{R}_{s,a}$:

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

The inner minimization selects the reward vector that is least favorable to the current policy, while the outer maximization finds the policy that performs best even under this pessimistic scenario.

The intuition of our method is that, if the uncertainty set is constructed so that the true reward $r(s,a) \in \mathcal{R}_{s,a}$, then the robust policy π_{r}^* provides an optimized lower bound on the true performance, and is hence more robust to the dataset uncertainties (Iyengar, 2005).

Closed-form robust policy and re-wrighting formulation of rDPO. We then derive the closed-form solution to (4). 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

$$r^{\mathrm{worst}}(s,a) \;=\; \hat{r}(s,a) - \kappa(s,a).$$

Substituting this into (4) and solving the KL-regularized maximization yields the optimal policy:

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

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.

Recall that the nominal DPO policy (without robustness) takes the form

$$\pi^*(a|s) \;=\; \frac{1}{Z(s)}\, \pi_{\mathrm{ref}}(a|s)\, \exp\big(\tfrac{\hat{r}(s,a)}{\beta}\big).$$

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

$$\frac{\pi_{\mathrm{r}}^*(a|s)}{\pi^*(a|s)} \;=\; \frac{Z(s)}{Z'(s)}\, \exp\big(-\tfrac{\kappa(s,a)}{\beta}\big). \quad (6)$$

This indicates that we can directly obtain π_{r}^* by *reweighting* π^*. 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. So, in summary, the minimax formulation (4) over interval uncertainty sets (3) leads directly to the robust policy (5), which incorporates pessimism through an intuitive action-wise reweighting scheme. The next subsections will instantiate $\kappa(s,a)$ via graph-theoretic and spectral analysis of the preference data, grounding these abstract radii in finite-sample guarantees.

Margin-shift formulation of rDPO. A key advantage of the robust DPO policy (5) is that it can be directly derived as the optimizer of a modified DPO loss objective. Define the DPO logit margin relative to π_{ref} as

$$\Delta\ell_\pi(s; a^+, a^-) \;=\; \log \frac{\pi(a^+|s)}{\pi_{\mathrm{ref}}(a^+|s)} - \log \frac{\pi(a^-|s)}{\pi_{\mathrm{ref}}(a^-|s)}.$$

Then we set the **Robust DPO loss** as

$$\mathcal{L}_{\mathrm{r}}(\pi) = \mathbb{E}_{(s,a^+,a^-)\sim\mathcal{D}} \Big[\log \sigma\big(\beta\, \Delta\ell_\pi(s; a^+, a^-) - \big(\kappa(s,a^+) - \kappa(s,a^-)\big)\big) \Big], \quad (7)$$

where the penalty difference $\kappa(s,a^+) - \kappa(s,a^-)$ acts as a deterministic offset to each pairwise margin, directly controlling the degree of pessimism.

We then show the following result.

Proposition 1. *The optimizer of the robust loss* (7) *coincides with the policy* (5).

This establishes an equivalence between the robust-optimization view (4) and the margin-shift view (7), both yielding the same robust policy. The margin-shift formulation is particularly convenient for practical optimization, as it reduces to a standard logistic regression with modified targets.

We then present our rDPO algorithm as in Algorithm 1.

Algorithm 1 Robust DPO Algorithm

Inputs: preference data $\mathcal{D}$, reference policy π_{ref}, temperature β, penalty map $\kappa(x, y)$, step size η
Initialize $\theta \leftarrow \theta_{\text{ref}}$ (or a copy of SFT weights)
repeat
  Sample minibatch $\mathcal{B} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^B \subset \mathcal{D}$
  for $(x, y^+, y^-) \in \mathcal{B}$ **do**
    Compute margin $m_\theta \leftarrow \beta\Delta\ell_\theta(x; y^+, y^-) - (\kappa(x, y^+) - \kappa(x, y^-))$
    Accumulate loss $\mathcal{L} \mathrel{+}= -\log\sigma(m_\theta)$
  end for
  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$
until convergence
Return: π_θ

4.2 Graph Laplacian and Ellipsoidal Confidence Sets

Our rDPO relies on the construction of the state-action wise error quantification $\kappa(s, a)$, i.e., to ensure $|r(s, a) - \hat{r}(s, a)| \le \kappa(s, a)$. However, such a quantification is generally difficult in RLHF: the MLE for BTL models estimates reward differences $r(s, a_1) - r(s, a_2)$ rather than absolute rewards, and concentration bounds apply to these differences rather than individual reward values.

The next subsection addresses this challenge by developing an ellipsoidal uncertainty set that quantifies reward estimation error through the spectral structure of the comparison graph.

To address this issue and construct data-dependent penalties $\kappa(s, a)$, in this section, we leverage the graph structure induced by pairwise comparisons and the spectral properties of the associated Laplacian matrix (Grone et al., 1990; Grone and Merris, 1994).

Comparison graph and Laplacian. The pairwise comparison data induces a natural graph structure for each state s. Define the *comparison graph* G_s with vertices $\mathcal{A}$ and edge weights $N_{ij}(s)$ (the count of comparisons between actions i and j). The weighted graph Laplacian $L_s \in \mathbb{R}^{A\times A}$ is

$$L_s = \sum_{i<j} N_{ij}(s)(\delta_i - \delta_j)(\delta_i - \delta_j)^\top.$$

The Laplacian is positive semidefinite with nullspace spanned by the all-ones vector $\mathbf{1}$. Its second smallest eigenvalue $\lambda_2(L_s) > 0$ (algebraic connectivity) when the graph is connected, and it provides a natural way to quantify uncertainty, encapsulating both connectivity and data density of the comparison structure. If certain actions are rarely compared, the corresponding nodes are weakly connected, yielding a small $\lambda_2(L_s)$. Larger $\lambda_2(L_s)$ indicates richer coverage and well-balanced comparisons. For any vector $v \in \mathbb{R}^A$ satisfying $\mathbf{1}^\top v = 0$, the quadratic form $v^\top L_s v$ measures how rapidly v varies across connected actions.

The pseudoinverse $L_s^\dagger$ defines a covariance-like matrix, and the *effective resistance*

$$\mathcal{R}_s(i, j) = (\delta_i - \delta_j)^\top L_s^\dagger(\delta_i - \delta_j)$$

quantifies how uncertain the relative ranking between actions i and j remains given the observed comparisons. Regions of the comparison graph with high effective resistance correspond to poorly supported preferences. Thus the Laplacian encodes the data geometry while $L_s^\dagger$ encodes uncertainty geometry.

Ellipsoidal concentration for BTL MLE. The key technical result enabling our construction is an ellipsoidal concentration bound for the BTL maximum-likelihood estimator.

Lemma 2 (Ellipsoidal Concentration)**.** *Fix a state s with connected comparison graph and let $\hat{r}(s)$ be the MLE under the gauge $\mathbf{1}^\top r(s) = 0$, with bounded dynamic range $|r(s, a)| \le B$. Then for any $\delta \in (0, 1)$, there exists a constant $\rho_s(\delta) > 0$ such that*

$$(r(s) - \hat{r}(s))^\top L_s(r(s) - \hat{r}(s)) \;\le\; \rho_s(\delta)$$

with probability at least $1 - \delta$.

This result follows from the structure of the BTL likelihood under the L_s-norm and the concentration of its score function (see Appendix A.1 and discussions in (Shah et al., 2016; Zhu et al., 2023)).

Per-action penalty construction. Lemma 2 implies that for any vector $u \in \mathbb{R}^A$,

$$|\langle u, r(s) - \hat{r}(s)\rangle| \;\le\; \sqrt{\rho_s(\delta)}\,\sqrt{u^\top L_s^\dagger u}.$$

Setting $u = \delta_a - \frac{1}{A}\mathbf{1}$ (which respects the gauge constraint) yields a per-action penalty. Formally, let $\hat{r}(s, a)$ be the maximum-likelihood estimate of

the latent reward, and define the pessimistic reward estimate:

$$r_{\text{pess}}(s,a) = \hat{r}(s,a) - \kappa(s,a),$$
$$\kappa(s,a) = \sqrt{\rho_s(\delta)}\,\sqrt{\delta_a^\top L_s^\dagger \delta_a}. \tag{8}$$

Then $r_{\text{pess}}(s,a) \le \hat{r}(s,a)$ holds with probability at least $1-\delta$. The quadratic form $\delta_a^\top L_s^\dagger \delta_a$ captures the effective resistance of action a in the comparison graph, quantifying how weakly connected that action is to others in the dataset. Weakly connected (rarely compared) actions receive larger penalties, making the resulting policy more conservative in uncertain regions.

This construction ties $\kappa(s,a)$ directly to observable graph-theoretic quantities, providing a bridge between statistical estimation and policy regularization. Crucially, it enables us to construct an uncertainty set for the reward function itself, rather than only for reward differences.

Hence we can plug this quantification term κ to our rDPO method and learn the robust policy. In the next section, we derive theoretical guarantees of the policy learned.

4.3 Finite-Sample and Spectral Guarantees

We now establish finite-sample guarantees for the robust DPO policy by combining the ellipsoidal confidence set from Section 4.2. We first state our assumptions.

Assumption 3. The reference policy assigns nonzero probability to every action: $\pi_{\text{ref}}(a|s) > 0$ for all (s,a).

Assumption 4. Rewards are mean-zero within each state, i.e., $\sum_a r(s,a) = 0$, ensuring identifiability.

Assumption 5. $|r(s,a)| \le B$ for all (s,a) and some $B > 0$.

Note that Assumptions 3 and 5 are standard in RLHF (), while Assumption 4 is without loss of generality as the BTL model depends only on reward differences.

Denote the robust policy learned through our rDPO as $\pi_{\text{r}}^*(a|s) \propto \pi_{\text{ref}}(a|s)\exp((\hat{r}(s,a) - \kappa(s,a))/\beta)$ and the optimal policy (with true rewards) as $\pi^*(a|s) \propto \pi_{\text{ref}}(a|s)\exp(r(s,a)/\beta)$. Let $\kappa_{\max}(s) = \max_a \kappa(s,a)$.

Theorem 6. *With probability at least $1-\delta$, it holds that*

$$0 \le J_s(\pi^*) - J_s(\pi_{\text{r}}^*) \le \mathcal{O}\left(\frac{\kappa_{\max}(s)}{\beta}\right), \tag{9}$$

where $J_s(\pi) = \sum_a \pi(a|s)\, r(s,a)$.

These results indicate that the robust policy remains close to the optimal policy. The deviation scales linearly with $\kappa_{\max}(s)/\beta$: increasing pessimism or lowering temperature contracting logits more strongly, while in well-supported regions (small κ), the two policies coincide.

Spectral characterization. We further derive the guarantees with the Laplacian pseudoinverse.

Theorem 7 (Spectral dependence)**.** *Let $\lambda_2(L_s)$ denote the algebraic connectivity (Fiedler value) of the comparison graph G_s. Then*

$$\max_a \kappa(s,a) \;=\; \mathcal{O}\Big(\sqrt{\tfrac{\rho_s(\delta)}{\lambda_2(L_s)}}\Big).$$

This follows from the Rayleigh quotient inequality $u^\top L_s^\dagger u \le \|u\|_2^2/\lambda_2(L_s)$ applied to $u = \delta_a - \frac{1}{A}\mathbf{1}$.

Corollary 8 (Sample-complexity scaling)**.** *If $L_s = N_s \bar{L}_s$ with $\bar{L}_s$ the normalized Laplacian and N_s the total number of comparisons at state s, then*

$$\kappa_{\max}(s) = \mathcal{O}\big(1/\sqrt{N_s \lambda_2(\bar{L}_s)}\big),$$

$$\|\pi_{\text{r}}^* - \pi^*\|_{\text{TV}} = \tilde{\mathcal{O}}\big(1/(\beta\sqrt{N_s \lambda_2(\bar{L}_s)})\big).$$

This establishes that uncertainty decays as $\mathcal{O}(1/\sqrt{N_s})$ with more preference samples, mirroring classical offline RL confidence bounds. Dense, well-connected comparison graphs (large $\lambda_2(L_s)$ and N_s) lead to smaller penalties and weaker pessimism, whereas sparse or disconnected graphs produce stronger regularization. This parallels confidence-set shrinkage in linear bandits, now expressed through the spectral geometry of preference data.

Remark 9. Our algorithm design and theoretical results are derived for tabular settings. In practical LLM setting where the problem scale is large, our tabular approach becomes inefficient. To address this issue, we further extend our rDPO to function approximation settings, which we deferred to Appendix A.5.

5 Experiments

In this section, we empirically verify that the effectiveness and robustness of our rDPO under both tabular environments and practical LLM alignment.

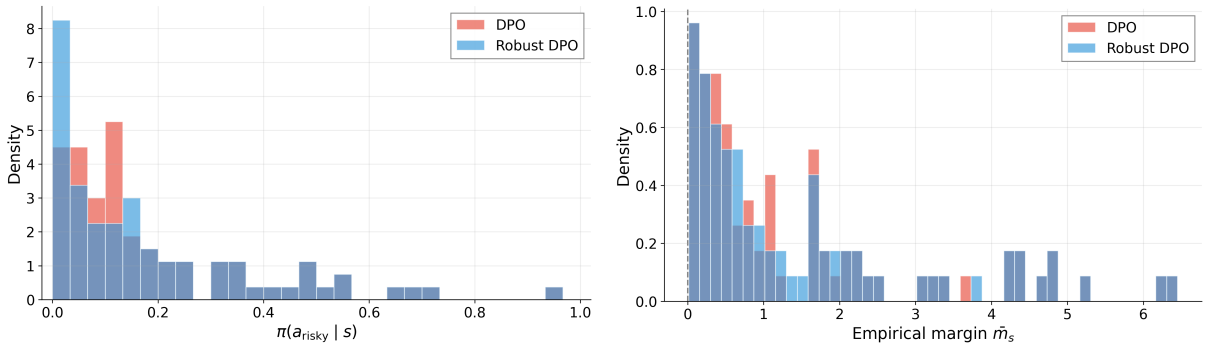

Figure 1: **Pessimistic correction induces conservative policy shifts. (Left)** Robust DPO (blue) concentrates mass away from risky actions compared to DPO (red). **(Right)** Margin distributions show that the pessimistic update $m \leftarrow m - (\kappa^+ - \kappa^-)$ produces tighter, left-shifted empirical margins, reducing overconfidence on uncertain comparisons.

5.1 Tabular Domain

We construct a gridworld environment with $S = 8$ states and $A = 5$ actions. The reward function includes a risky action (high reward, violates constraints) and safe alternatives. The preference dataset follows a Zipf distribution over states ($\alpha = 1.25$), creating heavy-tailed imbalance: $\approx 70\%$ of states have dense comparison graphs ($N_s > 250$, $\lambda_2(L_s) > 0.6$), while a long tail has sparse, noisy data ($N_s < 80$, $\lambda_2 < 0.3$, 10% label flips) to model realistic preference data where some prompts receive abundant feedback while others are undersampled (Ethayarajh et al., 2022).

We instantiate penalties via ellipsoidal confidence sets and train for 5000 steps with $\beta = 0.5$, batch size 64. Policies are tested on an environment with $2\times$ stochasticity (out-of-distribution evaluation), taken average over 10 seeds.

For each state s, we compute $\kappa_{\max}(s)$ and $(N_s\lambda_2(L_s))^{-1/2}$. Long-tail states ($N_s = 50$, $\lambda_2 = 0.2$) have $\kappa_{\max} \approx 0.75$, while well-covered states ($N_s = 300$, $\lambda_2 = 0.8$) have $\kappa_{\max} \approx 0.12$. This validates that uncertainty penalties adapt to local data geometry.

Robustness under distribution shift On the perturbed test environment, Robust DPO achieves 6.9% higher return than DPO (0.609 ± 0.027 vs. 0.570 ± 0.026) with 73% lower violation rate (0.8% vs. 3.1%). Figure 1 (left) shows that π_r^* suppresses risky-action probabilities in uncertain states, realizing the contraction mechanism. The margin distribution (Fig. 1, right) shifts leftward for Robust DPO, reflecting the margin correction $m \leftarrow m - (\kappa^+ - \kappa^-)$.

In a sample-complexity sweep, we vary $N_s \in [50, 500]$ and observe that $\|\pi_r^* - \pi^*\|_{\text{TV}}$ decays as $\mathcal{O}(N_s^{-0.48})$, close to the theoretical $\mathcal{O}(N_s^{-0.5})$ (Corollary 8).

5.2 LLM Fine-tuning

From theory to practice Theorem 7 predicts that uncertainty scales with comparison-graph sparsity and disconnectedness. In LLM domains, we cannot compute graph Laplacians explicitly, but we can proxy uncertainty via gradient variance. For each comparison (x, y^+, y^-), we estimate:

$$\hat{\kappa}(x, y^+, y^-) = \alpha \cdot \text{Var}_{\theta' \sim \mathcal{B}}\Big[\nabla_\theta \mathcal{L}_{\text{DPO}}\big(x, y^+, y^-; \theta'\big)\Big].$$

where $\mathcal{B}$ is a buffer of recent checkpoints and α is a scaling hyperparameter. High variance indicates inconsistent gradients across training, signaling ambiguous or low-quality comparisons.

Setup **Datasets:** HH (Helpful-Harmless) (Bai et al., 2022a): 161k pairs with dual objectives (helpfulness vs. safety); SHP (Stanford Human Preferences) (Ethayarajh et al., 2022): 348k pairs from Reddit, diverse domains. **Models:** Pythia-2.8B (Biderman et al., 2023), LLaMA-7B (Touvron et al., 2023), initialized from SFT checkpoints. **Training:** Batch size 128, learning rate 10^{-6}, $\beta = 0.5$, 2 epochs, H100 GPUs.

Baselines: DPO (no penalty), IPO (linearized reward mapping), VPO (auxiliary value model), Cal-DPO (temperature calibration). rDPO uses margin-driven, per-comparison penalties without additional models.

Metrics: Reward Margin $(r^+ - r^-)$ on held-out pairs (frozen reward model) quantifies preference

Table 1: LLM alignment results. rDPO achieves $2-3\times$ higher reward margins while maintaining accuracy, consistent with improved preference separation under the value-gap bound (Theorem 6).

Method	HH (Pythia-2.8B)		SHP (LLaMA-7B)	
	Margin ↑	Acc. (%)	Margin ↑	Acc. (%)
DPO	0.217	57	0.037	48
IPO	0.011	61	0.075	62
Cal-DPO	0.035	54	0.080	60
VPO	0.302	53	0.074	58
rDPO	**0.602**	**56**	**0.109**	**57**

Table 2: Reweighting evaluation on 100 HH-RLHF test pairs (Qwen2.5-0.5B with LoRA). Weighted DPO maintains accuracy while increasing preference margin strength, consistent with conservative updates on ambiguous data.

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

separation (Razin et al., 2025); Pairwise Accuracy measures ordering correctness.

Results Table 1 shows rDPO improves margins by $2.7\times$ (HH, Pythia) and $2.9\times$ (SHP, LLaMA) over DPO while preserving accuracy within 1%. Other baselines here as in VPO requires a separate value network; Cal-DPO applies global rescaling. rDPO's margin-driven, local penalties achieve superior margins without architectural changes, suggesting that the the developed framework is solid.

We note that in our results, our rDPO achieves a much higher reward margin, showing enhanced robustness against dataset uncertainty. Moreover, our rDPO maintains a comparable (slightly worst) accuracy with other baselines. We highlight that such results are expected, known as the robustness-accuracy trade-off.

5.3 Reweighting: Deployment Without Retraining

In this section, we validate our re-weighting formulation of rDPO. We define **Ambiguity score:**

$$u_{A,i} = 1 - \min\left(1, \frac{|m_i|}{c_{0.8}}\right),$$
$$m_i = \log \frac{p_{\pi_0}(y^+ \mid x)}{p_{\pi_0}(y^- \mid x)}.$$

where π_0 is the base model and $c_{0.8}$ is the 80th percentile of margins on training data ($c_{0.8} = 1.27$). High u_A means the base model finds the comparison ambiguous, suggesting high label noise or underspecification. We then reweight outputs:

$$\tilde{\pi}_\theta(y \mid x) \propto \pi_\theta(y \mid x) \exp\big(-\rho\, u_A\big).$$

where, $\rho \in \{1.0, 3.0\}$ Intuitively, we *downweight* responses on ambiguous prompts, shifting probability mass toward safer, less controversial outputs.

Results Table 2 shows, with Qwen2.5-0.5B (Yang et al., 2024b; Team, 2024) with LoRA adapter (Hu et al., 2021) on hh-rlhf subset, reweighting with $\rho = 3.0$ preserves accuracy (0.58, 95% CI [0.48, 0.68]) while *increasing* the gap to 23.05—a 5% improvement in preference discrimination without changing model weights. This validates that robust DPO's mechanism generalizes to inference-time adaptation, enabling rapid deployment updates.

6 Conclusion

We presented Robust Direct Preference Optimization (rDPO), a principled framework that integrates pessimism into preference learning via uncertainty quantification. By constructing ellipsoidal confidence sets from comparison-graph Laplacians, we derived data-dependent error quantifications $\kappa(s, a)$, naturally adapting to local data quality, and further derive the sub-optimality gap of our robust policy. Empirical validation across three paradigms further confirms tight theory-practice alignment, where our rDPO achieves $2-3\times$ higher reward margins than DPO while maintaining similar accuracy. Data-dependent reweighting experiments on Qwen2.5-0.5B further demonstrate that base-model margins provide a practical instantiation of uncertainty-driven down-weighting.

Rather than globally rescaling rewards or adding auxiliary models, rDPO adaptively contracts uncertain preference margins at the per-comparison level. This mechanism improves preference separation in well-supported regions while preventing overconfident updates on ambiguous data. The consistency across tabular, gradient-variance, and margin-based uncertainty proxies suggests that the core insight (margin-driven pessimism proportional to uncertainty) is fundamental to robust preference learning, independent of specific implementation details. By unifying statistical estimation theory with practical alignment objectives, rDPO provides a scalable, interpretable approach to uncertainty-aware preference optimization.

Limitations

The theoretical guarantees and results are developed in finite settings with Bradley–Terry preference models and assumes identifiable rewards, boundedness, and well-connected comparison graphs. For large-scale LLMs, prompts vary widely, comparison graphs are implicit, and per-action confidence sets are not directly available.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

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*, pages 4447–4455. PMLR.

J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. 2001. Solving uncertain markov decision processes.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

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, and 32 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Debangshu Banerjee and Aditya Gopalan. 2024. Towards reliable alignment: Uncertainty-aware rlhf. *Preprint*, arXiv:2410.23726.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

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

Jacob Buckman, Carles Gelada, and Marc G. Bellemare. 2020. The importance of pessimism in fixed-dataset policy optimization. *Preprint*, arXiv:2009.06799.

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*, 37:124093–124113.

John C Caldwell. 1981. The mechanisms of demographic change in historical perspective. *Population studies*, 35(1):5–27.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably robust dpo: Aligning language models with noisy feedback. *Preprint*, arXiv:2403.00409.

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*, volume 30. Curran Associates, Inc.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Miroslav Fiedler. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. *Preprint*, arXiv:1812.02900.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion,

Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, and 30 others. 2023. The capacity for moral self-correction in large language models. *Preprint*, arXiv:2302.07459.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Preprint*, arXiv:2009.11462.

Robert Grone and Russell Merris. 1994. The laplacian spectrum of a graph ii. *SIAM Journal on discrete mathematics*, 7(2):221–229.

Robert Grone, Russell Merris, and Viakalathur Shankar Sunder. 1990. The laplacian spectrum of a graph. *SIAM Journal on matrix analysis and applications*, 11(2):218–238.

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. *Preprint*, arXiv:2106.09685.

Shuaiyi Huang, Mara Levy, Anubhav Gupta, Daniel Ekpo, Ruijie Zheng, and Abhinav Shrivastava. 2025. Trend: Tri-teaching for robust preference-based reinforcement learning with demonstrations. *arXiv preprint arXiv:2505.06079*.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021a. Is pessimism provably efficient for offline rl? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021b. Is pessimism provably efficient for offline RL? In *Proc. International Conference on Machine Learning (ICML)*, pages 5084–5096.

Branden B Johnson and Marcus Mayorga. 2020. Temporal shifts in americans' risk perceptions of the zika outbreak. *Human and Ecological Risk Assessment: An International Journal*, 27(5):1242–1257.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. 2025. Distributionally robust optimization. *Acta Numerica*, 34:579–804.

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

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *Preprint*, arXiv:2005.01643.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Zhihao Shi, Feng Wu, and Jieping Ye. 2024. Ropo: Robust preference optimization for large language models. *arXiv preprint arXiv:2404.04102*.

R Duncan Luce and 1 others. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Debmalya Mandal, Paulius Sasnauskas, and Goran Radanovic. 2025. Distributionally robust reinforcement learning with human feedback. *arXiv preprint arXiv:2503.00539*.

Xin Mao, Feng-Lin Li, Huimin Xu, Wei Zhang, and Anh Tuan Luu. 2024. Don't forget your reward values: Language model alignment via value-based calibration. *arXiv preprint arXiv:2402.16030*.

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)*, pages 839–846.

Soichiro Nishimori, Yu-Jie Zhang, Thanawat Lodkaew, and Masashi Sugiyama. 2025. On symmetric losses for robust policy optimization with noisy preferences. *arXiv preprint arXiv:2505.24709*.

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*, volume 35, pages 27730–27744. Curran Associates, Inc.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

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*, 36:53728–53741.

Hamed Rahimian and Sanjay Mehrotra. 2019. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. 2021. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D Lee, and Sanjeev Arora. 2025. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv:2503.15477*.

Mark Rudelson and Roman Vershynin. 2013. Hanson-wright inequality and sub-gaussian concentration. *Preprint*, arXiv:1306.2872.

Sharan Sahu and Martin T Wells. 2025. Dro-rebel: Distributionally robust relative-reward regression for fast and efficient llm alignment. *arXiv preprint arXiv:2509.19104*.

Jay K Satia and Roy E Lave Jr. 1973. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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):1–47.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. 2022. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *Proc. International Conference on Machine Learning (ICML)*, pages 19967–20025.

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*, 33:3008–3021.

Aviv Tamar, Shie Mannor, and Huan Xu. 2014. Scaling up robust MDPs using function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 181–189. PMLR.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Louis L Thurstone. 1927. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Masatoshi Uehara and Wen Sun. 2021. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Masatoshi Uehara and Wen Sun. 2023. Pessimistic model-based offline reinforcement learning under partial coverage. *Preprint*, arXiv:2107.06226.

Lingxiao Wang, Xiao Zhang, and Quanquan Gu. 2016. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. 2024. A unified principle of pessimism for offline reinforcement learning under model mismatch. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Yue Wang, Jinjun Xiong, and Shaofeng Zou. 2023. Achieving the minimax optimal sample complexity of offline reinforcement learning: A dro-based approach. *Preprint*, arXiv:2305.13289.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. 2013. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024. Cal-dpo: Calibrated direct preference optimization for language model alignment. *Preprint*, arXiv:2412.14516.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. 2021a. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. 2021b. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 27395–27407.

Huan Xu and Shie Mannor. 2010. Distributionally robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 2505–2513.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.

Adam X Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. 2024a. Bayesian reward models for llm alignment. *arXiv preprint arXiv:2402.13210*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ming Yin and Yu-Xiang Wang. 2021. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. Mopo: Model-based offline policy optimization. *Preprint*, arXiv:2005.13239.

Farhad Zafari, Irene Moser, and Tim Baarslag. 2019. Modelling and analysis of temporal preference drifts using a component-based factorised latent approach. *Expert systems with applications*, 116:186–208.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. 2021. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. 2022. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*.

Chuheng Zhang, Wei Shen, Li Zhao, Xuyun Zhang, Xiaolong Xu, Wanchun Dou, and Jiang Bian. 2025. Policy filtration for rlhf to mitigate noise in reward models. *Preprint*, arXiv:2409.06957.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *Preprint*, arXiv:2305.10425.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

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*, pages 43037–43067. PMLR.

A Theoretical Results

A.1 Finite-Sample Concentration for BTL

This appendix provides the technical ingredients and a concise proof of the finite-sample confidence ellipsoid used to construct pessimistic penalties $\kappa(s, a)$.

Fix a state $s \in \mathcal{S}$ with actions $\mathcal{A} = \{1, \dots, A\}$. Let the (count) Laplacian of the comparison graph be

$$L_s \;=\; \sum_{i<j} N_{ij}(s)\,(\delta_i - \delta_j)(\delta_i - \delta_j)^\top, \tag{10}$$

and write $\|v\|_{L_s}^2 := v^\top L_s v$ and $\|g\|_{L_s^\dagger}^2 := g^\top L_s^\dagger g$ on $1^\perp := \{v : \mathbf{1}^\top v = 0\}$.

Assumption 10 (Connectivity)**.** The comparison graph is connected. Equivalently, $\lambda_2(L_s) > 0$ and $L_s \succ 0$ on $1^\perp$.

Let $\ell_s(r)$ denote the *negative* BTL log-likelihood at state s.

Curvature in Laplacian geometry.

Lemma 11 (BTL curvature in L_s-norm)**.** *Under Assumptions 3–5 and 10, define* $\gamma(B) := \min_{|z|\le 2B} \sigma(z)\,[1-\sigma(z)] > 0$*, where* $\sigma(z) = 1/(1+e^{-z})$*. Then for all* $\Delta \in 1^\perp$*,*

$$\ell_s(r+\Delta) - \ell_s(r) - \langle \nabla \ell_s(r), \Delta\rangle \;\ge\; \tfrac{\gamma(B)}{2}\,\Delta^\top L_s \Delta.$$

Equivalently, on $1^\perp$*,* $\nabla^2 \ell_s(\tilde r) \succeq \gamma(B)\, L_s$ *for all* $\tilde r$ *on the segment* $[r,\, r+\Delta]$*.*

Proof. For one comparison (i, j) with logit $z = r_i - r_j$, $\phi''_{ij}(z) = \sigma(z)[1-\sigma(z)] \in (0, 1/4]$ and $\phi''_{ij}(z) \ge \gamma(B)$ for $|z| \le 2B$. Summing over samples/pairs yields a weighted Laplacian, $\nabla^2 \ell_s(\tilde r) \succeq \gamma(B)\, L_s$ on $1^\perp$. A Taylor remainder identity gives the claim (Shah et al., 2016). □

M-estimation inequality. Let $\hat r(s)$ be the MLE (constrained to $1^\perp$), $\Delta := \hat r(s) - r(s) \in 1^\perp$. By optimality, $\ell_s(r+\Delta) \le \ell_s(r)$; Lemma 11 implies

$$\tfrac{\gamma(B)}{2}\,\|\Delta\|_{L_s}^2 \;\le\; \langle -\nabla \ell_s(r), \Delta\rangle \;\le\; \|\nabla \ell_s(r)\|_{L_s^\dagger}\,\|\Delta\|_{L_s},$$

hence

$$\|\Delta\|_{L_s} \;\le\; \frac{2}{\gamma(B)}\,\|\nabla \ell_s(r)\|_{L_s^\dagger}. \tag{11}$$

Concentration of the score. For one comparison (i, j) with label $y \in \{0, 1\}$ and $z = r_i - r_j$, $\phi_{ij}(z) = -y \log \sigma(z) - (1-y)\log(1-\sigma(z))$ has $\phi'_{ij}(z) = \sigma(z) - y$ and, by the chain rule,

$$\nabla_r \phi_{ij}(r) \;=\; (\sigma(r_i - r_j) - y)\,(\delta_i - \delta_j).$$

Summing over all pairs/samples at s,

$$\nabla \ell_s(r) \;=\; \sum_{i<j} \sum_{t=1}^{N_{ij}(s)} \big(\sigma(r_i - r_j) - y_{ij}^{(t)}\big)\,(\delta_i - \delta_j).$$

At the true r, the residuals are mean-zero, bounded in $[-1, 1]$, and independent across samples. Stack them as $X \in \mathbb{R}^m$ and the pairwise directions as columns of U so that $\nabla \ell_s(r) = UX$ and $\|\nabla \ell_s(r)\|_{L_s^\dagger}^2 = X^\top M X$ with $M := U^\top L_s^\dagger U$. Since $UU^\top = L_s$, one checks $\|M\|_{\mathrm{op}} = 1$ and $\|M\|_{\mathrm{fro}}^2 = A - 1$ (on $1^\perp$).

By the Hanson–Wright inequality (Rudelson and Vershynin, 2013; Shah et al., 2016), there is a universal $C > 0$ such that, with probability at least $1 - \delta$,

$$\|\nabla \ell_s(r)\|_{L_s^\dagger}^2 \;\le\; C\big(A + \log \tfrac{1}{\delta}\big). \tag{12}$$

Proof of Lemma 2: Ellipsoidal concentration.

Proof. Combine (11) and (12), then square:

$$\begin{aligned}\|\hat{r}(s) - r(s)\|_{L_s}^2 &\le \frac{4}{\gamma(B)^2} \|\nabla \ell_s(r)\|_{L_s^\dagger}^2 \\ &\le \frac{C}{\gamma(B)^2}\Big(A + \log\tfrac{1}{\delta}\Big) \\ &=: \rho_s(\delta).\end{aligned}$$

This is $(\hat{r} - r)^\top L_s(\hat{r} - r) \le \rho_s(\delta)$, completing the proof. □

Support function and per-action radii. Let $\mathcal{C}_s(\rho) := \{v \in 1^\perp : \ v^\top L_s v \le \rho\}$. For any $u \in 1^\perp$,

$$\sup_{v \in \mathcal{C}_s(\rho)} \langle u, v\rangle \ = \ \sqrt{\rho}\,\sqrt{u^\top L_s^\dagger u}. \tag{13}$$

Applying (13) to $e := \hat{r}(s) - r(s) \in \mathcal{C}_s(\rho_s(\delta))$ yields

$$|\langle u,\, \hat{r}(s) - r(s)\rangle| \ \le \ \sqrt{\rho_s(\delta)}\,\sqrt{u^\top L_s^\dagger u}.$$

For the per-action direction $u_a := \delta_a - \frac{1}{A}\mathbf{1} \in 1^\perp$, define the *per-action radius*

$$\kappa(s,a) \ := \ \sqrt{\rho_s(\delta)}\,\sqrt{u_a^\top L_s^\dagger u_a}. \tag{14}$$

This is the penalty construction used in (8) of the main paper.

Corollary 12 (One-sided LCB)**.** *With probability at least* $1 - \delta$*, for all* $a \in \mathcal{A}$*,*

$$\hat{r}(s,a) - \kappa(s,a) \ \le \ r(s,a).$$

Scaling with sample size. If $L_s = N_s\,\bar{L}_s$ ($N_s = \sum_{i<j} N_{ij}(s)$, $\bar{L}_s$ fixed shape), then $L_s^\dagger = N_s^{-1}\bar{L}_s^\dagger$ and

$$\kappa(s,a) \ = \ \frac{1}{\sqrt{N_s}}\sqrt{\rho_s(\delta)\,u_a^\top \bar{L}_s^\dagger u_a} \ = \ \mathcal{O}\big(N_s^{-1/2}\big).$$

This $1/\sqrt{N_s}$ decay matches classical statistical rates.

A.2 Policy Deviation Bounds

We now analyze how far the pessimistic DPO policy can deviate from the oracle policy with true rewards. All results hold with probability at least $1 - \delta$, under the concentration guarantees of Appendix A.1.

Proof of Theorem 6: Oracle value gap.

Proof. Let π^* be the oracle DPO policy with logits from the true rewards $r(s,\cdot)$, and suppose $|r(s,a)| \le B$ for all a. Define

$$J_s(\pi) := \sum_{a \in \mathcal{A}} \pi(a|s)\, r(s,a).$$

For any two distributions μ, ν over $\mathcal{A}$,

$$\begin{aligned}\big|J_s(\mu) - J_s(\nu)\big| &= \Big|\sum_a \big(\mu(a \mid s) - \nu(a \mid s)\big)\, r(s,a)\Big| \\ &\le \|r(s,\cdot)\|_\infty \, \|\mu(\cdot \mid s) - \nu(\cdot \mid s)\|_1 \\ &\le 2B\, \|\mu - \nu\|_{\mathrm{TV}}.\end{aligned}$$

Apply this with $\mu = \pi_{\mathrm{r}}^*(\cdot|s)$ and $\nu = \pi^*(\cdot|s)$. With probability at least $1-\delta$,

$$\|\pi_{\mathrm{r}}^*(\cdot|s) - \pi^*(\cdot|s)\|_{\mathrm{TV}} \le e^{\varepsilon} - 1, \qquad \varepsilon := \frac{\kappa_{\max}(s)}{\beta},$$

where $\kappa_{\max}(s) := \max_a \kappa(s,a)$.

Therefore,

$$0 \le J_s(\pi^*) - J_s(\pi_{\mathrm{r}}^*) \le 2B\,(e^{\varepsilon} - 1) = \mathcal{O}\Big(\tfrac{B\,\kappa_{\max}(s)}{\beta}\Big).$$

Equivalently, if one prefers an explicit constant for small $\varepsilon \le 1$, use $e^{\varepsilon} - 1 \le 2\varepsilon$ to obtain

$$J_s(\pi^*) - J_s(\pi_{\mathrm{r}}^*) \le \frac{4B}{\beta}\,\kappa_{\max}(s).$$

□

Discussion. These bounds show that pessimism reduces logits, contracts policy ratios within $\exp(\pm\kappa/\beta)$, keeps the policies close in total variation, and incurs only a small value gap relative to the oracle (Theorem 6). As $\kappa_{\max}(s) \to 0$ (e.g., $\mathcal{O}(1/\sqrt{N_s})$ with more comparisons), pessimistic and oracle policies coincide.

A.3 Policy Deviation Bounds

A.4 Spectral Dependence and Effective Resistance

The pessimism penalties $\kappa(s,a)$ defined in (14) involve quadratic forms of the Laplacian pseudoinverse. This section develops their interpretation in spectral graph theory and proves the spectral scaling results from the main paper.

Effective resistance and pairwise bounds. For a comparison graph $G_s = (\mathcal{A}, E)$ with Laplacian L_s, the effective resistance between nodes i and j is

$$R_{\mathrm{eff}}(i,j) := (\delta_i - \delta_j)^\top L_s^\dagger\,(\delta_i - \delta_j).$$

Proposition 13 (Pairwise uncertainty equivalent with effective resistance)**.** *For any $(a,b) \in \mathcal{A}\times\mathcal{A}$, the ellipsoidal confidence set (Appendix A.1) implies*

$$\begin{aligned}\big|\big(r(s,a) - r(s,b)\big) \;-\; \big(\hat r(s,a) - \hat r(s,b)\big)\big| \\ \le \sqrt{\rho_s(\delta)}\,\sqrt{R_{\mathrm{eff}}(a,b)}.\end{aligned}$$

Thus the confidence radius on a pairwise margin scales with the square root of effective resistance. Pairs connected by many redundant comparisons have small resistance (small uncertainty), while sparsely connected pairs have large resistance (large uncertainty).

Per-action penalties as average resistance. For $u_a = \delta_a - \frac{1}{A}\mathbf{1} \in 1^\perp$, one can show

$$u_a^\top L_s^\dagger u_a = \frac{1}{A^2}\sum_{b\in\mathcal{A}} R_{\mathrm{eff}}(a,b).$$

Proposition 14 (Penalty as average resistance)**.** *For each action a,*

$$\kappa(s,a) = \sqrt{\rho_s(\delta)}\,\sqrt{\tfrac{1}{A^2}\sum_b R_{\mathrm{eff}}(a,b)}.$$

Hence $\kappa(s,a)$ reflects the average effective resistance from action a to all others.

Proof of Theorem 7: Spectral scaling.

Proof. Let $\lambda_2(L_s)$ denote the algebraic connectivity (Fiedler value (Fiedler, 1973)) of G_s. By the Rayleigh quotient inequality,

$$u^\top L_s^\dagger u \le \frac{\|u\|_2^2}{\lambda_2(L_s)}, \quad u \in 1^\perp.$$

For the per-action direction $u_a = \delta_a - \frac{1}{A}\mathbf{1}$, we have $\|u_a\|_2^2 = 1 + 1/A^2 \le 2$ (for $A \ge 1$). This gives

$$\begin{aligned}\kappa(s,a) &= \sqrt{\rho_s(\delta)}\,\sqrt{u_a^\top L_s^\dagger u_a} \\ &\le \sqrt{\rho_s(\delta)}\,\frac{\|u_a\|_2}{\sqrt{\lambda_2(L_s)}} \\ &\le \sqrt{\frac{2\rho_s(\delta)}{\lambda_2(L_s)}}.\end{aligned}$$

In particular,

$$\max_a \kappa(s,a) \le \sqrt{\tfrac{2\rho_s(\delta)}{\lambda_2(L_s)}} = \mathcal{O}\Big(\sqrt{\tfrac{\rho_s(\delta)}{\lambda_2(L_s)}}\Big).$$

Thus graphs with better connectivity (larger λ_2) yield smaller pessimism penalties, completing the proof. □

Proof of Corollary 8: Sample size scaling.

Proof. If $L_s = N_s\,\bar{L}_s$ with N_s total comparisons at state s and normalized Laplacian $\bar{L}_s$, then

$$L_s^\dagger = \tfrac{1}{N_s}\,\bar{L}_s^\dagger.$$

Substituting into the penalty formula,

$$\begin{aligned}\kappa(s,a) &= \sqrt{\rho_s(\delta)}\,\sqrt{u_a^\top L_s^\dagger u_a} \\ &= \sqrt{\rho_s(\delta)}\,\sqrt{\tfrac{1}{N_s}\,u_a^\top \bar{L}_s^\dagger u_a} \\ &= \frac{1}{\sqrt{N_s}}\,\sqrt{\rho_s(\delta)\,u_a^\top \bar{L}_s^\dagger u_a}.\end{aligned}$$

By Theorem 7,

$$u_a^\top \bar{L}_s^\dagger u_a \le \frac{\|u_a\|_2^2}{\lambda_2(\bar{L}_s)} \le \frac{2}{\lambda_2(\bar{L}_s)}.$$

Hence,

$$\kappa_{\max}(s) = \max_a \kappa(s,a) = \mathcal{O}\big(1/\sqrt{N_s \lambda_2(\bar{L}_s)}\big).$$

$$\begin{aligned}\|\pi_{\mathrm{r}} - \pi^*\|_{\mathrm{TV}} &= \mathcal{O}\left(\frac{\kappa_{\max}(s)}{\beta}\right) \\ &= \mathcal{O}\left(\frac{1}{\beta\,\sqrt{N_s\,\lambda_2(\bar{L}_s)}}\right).\end{aligned}$$

Since $\rho_s(\delta) = \mathcal{O}(A + \log(1/\delta))$, the $\tilde{\mathcal{O}}$ notation absorbs logarithmic factors. □

Asymptotic shrinkage. As $N_s \to \infty$, $\kappa(s,a) = \mathcal{O}(1/\sqrt{N_s})$, so pessimistic and oracle policies coincide in the large-sample limit. This matches classical statistical rates and demonstrates that robustness is strongest precisely where data is weakest.

A.5 Linear Function Approximation

A.6 Linear Setting

In LLM alignment settings, rewards often exhibit shared structure across contexts and responses. We extend our uncertainty-aware pessimism framework to the linear function approximation regime and show that the tabular Laplacian penalty emerges as a special case.

Linear reward model. Let $\phi(s,a) \in \mathbb{R}^d$ be a known feature map and assume a linear latent reward

$$r_{\theta^\star}(s,a) \;=\; \langle \theta^\star, \phi(s,a)\rangle, \qquad \theta^\star \in \mathbb{R}^d. \tag{15}$$

Under the Bradley–Terry–Luce (BTL) logistic comparison model,

$$\begin{aligned} \mathbb{P}(a_i^+ \succ a_i^- \mid s_i) &= \sigma\big(\langle \theta^\star, \Delta\phi_i\rangle\big), \\ \Delta\phi_i &:= \phi(s_i, a_i^+) - \phi(s_i, a_i^-). \end{aligned}$$

Ellipsoidal confidence set. Let $\hat{\theta}$ be the regularized MLE obtained from the preference dataset. Define the pairwise design matrix

$$\Sigma_D \;:=\; \sum_{i=1}^{n} \Delta\phi_i \Delta\phi_i^\top. \tag{16}$$

Existing analyses of preference-based reward learning in linear models (e.g., (Zhu et al., 2023)) imply ellipsoidal concentration of the MLE in the Σ_D-geometry. Under standard regularity conditions, there exists a radius $\alpha_n(\delta) > 0$ such that with probability at least $1-\delta$,

$$\|\hat{\theta} - \theta^\star\|_{\Sigma_D} \;:=\; \sqrt{(\hat{\theta} - \theta^\star)^\top \Sigma_D (\hat{\theta} - \theta^\star)} \;\le\; \alpha_n(\delta). \tag{17}$$

Per-action pessimism. Equation (17) yields a uniform reward lower bound via Cauchy–Schwarz in the $(\Sigma_D, \Sigma_D^\dagger)$ dual norms: for any (s,a),

$$\begin{aligned} |r_{\hat{\theta}}(s,a) - r_{\theta^\star}(s,a)| &= \big|\langle \hat{\theta} - \theta^\star, \phi(s,a)\rangle\big| \\ &\le \|\hat{\theta} - \theta^\star\|_{\Sigma_D}\, \|\phi(s,a)\|_{\Sigma_D^\dagger}. \end{aligned} \tag{18}$$

Therefore, with probability at least $1-\delta$,

$$\begin{aligned} r_{\theta^\star}(s,a) &\ge r_{\hat{\theta}}(s,a) - \kappa(s,a), \\ \kappa(s,a) &:= \alpha_n(\delta)\, \sqrt{\phi(s,a)^\top \Sigma_D^\dagger \phi(s,a)}. \end{aligned} \tag{19}$$

This converts the coupled ellipsoidal uncertainty in θ into a rectangular (per-action) pessimistic envelope in reward space, enabling the same robust policy construction (5) as in the tabular case.

Connection to tabular case. When features are one-hot encodings of state-action pairs (i.e., $\phi(s,a) = e_{(s,a)}$ for the tabular setting), the design matrix Σ_D becomes the Laplacian matrix L_s from Section 4.2, and the penalty (19) recovers the graph-Laplacian-based penalty (8).

The regularized MLE $\hat{\theta}$ satisfies ellipsoidal concentration in the Σ_D-geometry. Under standard regularity conditions (bounded logit range, Lipschitz gradients), existing analyses (Zhu et al., 2023) yield:

Lemma 15 (Ellipsoidal concentration for linear BTL)**.** *Under Assumption 5 (with bounded $|\langle \theta^\star, \Delta\phi_i\rangle| \le B$ for all i) and connectivity of the design (i.e., $\Sigma_D \succ \lambda I_d$), there exists a radius $\alpha_n(\delta) > 0$ such that with probability at least $1-\delta$,*

$$\|\hat{\theta} - \theta^\star\|_{\Sigma_D} := \sqrt{(\hat{\theta} - \theta^\star)^\top \Sigma_D (\hat{\theta} - \theta^\star)} \le \alpha_n(\delta),$$

where $\alpha_n(\delta) = \mathcal{O}(\sqrt{d + \log(1/\delta)})$.

The proof follows the same M-estimation + concentration template as the tabular case (Appendix A.1), but now operating in feature space. The key ingredients are:

- Strong convexity of the logistic loss in the Σ_D-norm (analogous to Lemma 11).
- Self-concordance of the logistic link, which bounds Hessian variation.
- Concentration of the score $\nabla \ell(\theta^\star)$ via martingale or Hanson–Wright inequalities.

See Zhu et al. (2023) (Theorem 1) for the complete proof.

Per-action pessimism via Cauchy–Schwarz. The ellipsoidal set (17) implies a uniform reward lower bound. For any (s, a), by Cauchy–Schwarz in the $(\Sigma_D, \Sigma_D^\dagger)$ dual norms:

$$\begin{aligned} |r_{\hat{\theta}}(s,a) - r_{\theta^\star}(s,a)| &= \left|\langle \hat{\theta} - \theta^\star, \phi(s,a)\rangle\right| \\ &\le \|\hat{\theta} - \theta^\star\|_{\Sigma_D} \, \|\phi(s,a)\|_{\Sigma_D^\dagger} \\ &\le \alpha_n(\delta)\sqrt{\phi(s,a)^\top \Sigma_D^\dagger \phi(s,a)}, \end{aligned} \tag{20}$$

where $\|v\|^2_{\Sigma_D^\dagger} := v^\top \Sigma_D^\dagger v$.

Therefore, with probability at least $1 - \delta$,

$$r_{\theta^\star}(s,a) \ge r_{\hat{\theta}}(s,a) - \kappa(s,a),$$

where

$$\kappa(s,a) := \alpha_n(\delta)\sqrt{\phi(s,a)^\top \Sigma_D^\dagger \phi(s,a)}. \tag{21}$$

This converts the coupled ellipsoidal uncertainty in θ into a rectangular (per-action) pessimistic envelope in reward space, enabling the same robust policy construction (5) as in the tabular case.

Connection to tabular case: One-hot features. When features are one-hot encodings of state-action pairs, i.e.,

$$\phi(s,a) = e_{(s,a)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|},$$

the design matrix Σ_D becomes (up to regularization) the state-wise comparison-graph Laplacian L_s.

B Experimental Details

This appendix provides implementation details and hyperparameters for the experiments in Section 5.

B.1 Tabular Domain

Environment. We construct an abstract 8-state MDP with $S = 8$ states and $A = 5$ actions to enable exact computation of all theoretical quantities as in penalties $\kappa(s,a)$, comparison graph properties $N_s, \lambda_2(L_s)$, total-variation distance, etc. The environment structure is:

- **States:** $\mathcal{S} = \{s_1, s_2, \ldots, s_8\}$, representing abstract contexts.
- **Actions:** $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$.
- **Rewards:** Designed to include a risky action (higher reward but uncertain) and safe alternatives. Rewards are gauged to zero mean per state: $\sum_a r(s,a) = 0$.
- **Dynamics:** Deterministic transitions during training; out-of-distribution evaluation uses $2\times$ increased stochasticity to test robustness under distribution shift.

Preference data generation. The key design choice is to create a long-tailed, imbalanced preference dataset that mirrors real RLHF data (Ethayarajh et al., 2022), where some prompts receive abundant feedback while others are under-sampled.

State distribution: We sample states according to a Zipf distribution with parameter $\alpha = 1.25$:

$$\Pr(s_i) \propto \frac{1}{i^{1.25}}, \quad i \in \{1, \dots, 8\}.$$

This creates heavy-tailed imbalance: approximately 70% of comparisons concentrate on the first 3–4 states, while the remaining states form a long tail.

Comparison graphs: For each state s, we construct a comparison graph over actions $\mathcal{A}$ and sample N_s pairwise comparisons. The resulting dataset exhibits:

- **Dense states** (70% of states): $N_s > 250$ comparisons, dense comparison graphs with $\lambda_2(L_s) > 0.6$ (high algebraic connectivity), low uncertainty.
- **Sparse states** (30% of states, long tail): $N_s < 80$ comparisons, sparse or path-like comparison graphs with $\lambda_2(L_s) < 0.3$ (low algebraic connectivity), high uncertainty.

Label noise: To model annotation errors, we inject label flips with 10% probability in sparse states (where data is weakest), flipping the preference outcome $y_{ij} \mapsto 1 - y_{ij}$ for individual comparisons.

BTL sampling: Preference labels are drawn from the Bradley–Terry–Luce model:

$$\Pr(a_i \succ a_j \mid s) = \sigma\big(r(s, a_i) - r(s, a_j)\big), \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Training configuration.

- **Penalty construction:** Ellipsoidal confidence sets (Eq. 8) with $\kappa(s, a) = \sqrt{\rho_s(\delta)}\sqrt{u_a^\top L_s^\dagger u_a}$, where $\rho_s(\delta)$ is computed from Lemma 2 with $\delta = 0.05$.
- **Reference policy:** Uniform $\pi_{\text{ref}}(a|s) = 1/A = 0.2$ for all (s, a).
- **Temperature:** $\beta = 0.5$.
- **Optimization:** Softmax policy parameterization with exact gradient computation (no sampling). Training for 5000 steps with batch size 64, using Adam optimizer with learning rate 10^{-3}.
- **Evaluation:** Policies tested on perturbed environment with $2\times$ transition stochasticity (OOD evaluation). All metrics averaged over 10 independent seeds.

B.2 LLM Preference Learning

Datasets. **HH (Helpful-Harmless)** (Bai et al., 2022a): ∼161k preference pairs collected from human annotators, balancing helpfulness and harmlessness objectives. Each sample consists of a prompt and two assistant responses, with one labeled as preferred.

SHP (Stanford Human Preferences) (Ethayarajh et al., 2022): ∼348k preference pairs scraped from Reddit, covering diverse open-domain topics. The dataset exhibits natural long-tailed distribution over prompts and domains.

Preprocessing: Standard train/validation/test splits. Sequences truncated to 384 tokens (Pythia) or 2048 tokens (LLaMA). Prompts and responses formatted with instruction template: `[INST] prompt [RESP] response`. Right-padding with tokenizer pad token.

Models and reference policies. **Pythia-2.8B** (Biderman et al., 2023): Decoder-only transformer with 2.8 billion parameters.

LLaMA-7B (Touvron et al., 2023): Decoder-only transformer with 7 billion parameters.

Reference policies π_{ref}: Both models initialized from supervised fine-tuned (SFT) checkpoints trained on the same instruction corpus, then frozen during preference optimization.

Training configuration.

- **Optimizer:** AdamW with learning rate 1×10^{-6}, linear warmup over 2% of steps, gradient clipping at 1.0.
- **Batch size:** 128 (Pythia-HH), 64 (LLaMA-SHP), with gradient accumulation of 8 steps.
- **Temperature:** $\beta = 0.5$ (main experiments); $\beta \in \{0.1, 0.3, 0.5\}$ (ablation).
- **Training duration:** 2 epochs over training data.
- **Precision:** Mixed-precision (FP16) training with Fully Sharded Data Parallel (FSDP).
- **Hardware:** NVIDIA H100 GPUs (80GB).
- **Evaluation:** Every 500 steps on held-out validation set. Results averaged over last 5 checkpoints and 3 random seeds.

Gradient variance proxy for uncertainty. Since we cannot compute comparison-graph Laplacians explicitly in LLM domains, we proxy uncertainty via gradient variance. For each comparison (x, y^+, y^-), we estimate:

$$\hat{\kappa}(x, y^+, y^-) = \alpha \cdot \mathrm{Var}_{\theta' \sim \mathcal{B}}[\nabla_\theta \mathcal{L}_{\mathrm{DPO}}(x, y^+, y^-; \theta')],$$

where $\mathcal{B}$ is a buffer of recent checkpoints (last 10 iterations) and $\alpha = 0.2$ is a scaling hyperparameter. High variance indicates inconsistent gradients across training, signaling ambiguous or low-quality comparisons—analogous to low λ_2 or small N_s in the tabular case.

Evaluation metrics. **Reward margin:** The primary metric quantifying preference separation strength:

$$\mathrm{Margin} = \mathbb{E}_{(x,y^+,y^-)}[r_\phi(x, y^+) - r_\phi(x, y^-)],$$

where r_ϕ is a frozen reward model trained on the same dataset using Bradley–Terry likelihood. Larger margins indicate greater separation between preferred and rejected responses, reflecting stronger consistency with human judgments (Razin et al., 2025).

Pairwise accuracy: Binary preference prediction accuracy:

$$\mathrm{Acc} = \mathbb{P}_{(x,y^+,y^-)}[\log \pi_\theta(y^+ \mid x) > \log \pi_\theta(y^- \mid x)].$$

Measures how often the model assigns higher log-probability to the preferred continuation.

Temperature ablation. To validate the κ/β scaling prediction, we train with $\beta \in \{0.1, 0.3, 0.5\}$ on HH+Pythia. As β decreases, margin curves become smoother and oscillations diminish, directly confirming that effective contraction strength is κ/β. For $\beta = 0.1$, chosen/rejected reward curves are nearly monotonic; for $\beta = 0.5$, oscillations increase. This empirically validates the theory-predicted trade-off: smaller β induces stronger pessimism and more conservative learning, while larger β allows faster but noisier updates.

B.3 Data-Dependent Reweighting Experiments

This section provides details for the reweighting experiment (Section 5).

Ambiguity metric and weighting. Our reweighting approach computes a deterministic ambiguity score from the base model π_0 (frozen Qwen2.5-0.5B-Instruct). For each training pair (x_i, y_i^+, y_i^-), we compute:

$$m_i = \mathrm{meanlogp}_{\pi_0}(y_i^+ | x_i) - \mathrm{meanlogp}_{\pi_0}(y_i^- | x_i),$$

where $\mathrm{meanlogp}$ is the mean token log-probability over completion tokens.

Table 3: Additional results with heuristic margin-shift DPO.

Task / Model	Method	Margin	Accuracy
Toxicity (GPT2–L)	DPO	0.022	0.052
	rDPO	**0.025**	**0.055**
Toxicity (LLaMA–7B)	DPO	0.052	0.58
	rDPO	**0.060**	**0.58**
IMDB (GPT2–L)	DPO	3.10	0.80
	rDPO	**3.80**	**0.82**

Let $c_{0.8}$ be the 80th percentile of $|m|$ over valid training examples. For our clean HH-RLHF training subset, $c_{0.8} = 1.2666$. We define an ambiguity score $u_{A,i} \in [0, 1]$:

$$u_{A,i} = 1 - \text{clip}\left(\frac{|m_i|}{c_{0.8}},\, 0,\, 1\right).$$

Pairs with small $|m_i|$ (weak base-model preference signal) receive larger $u_{A,i}$ and are treated as more ambiguous.

Given $u_{A,i}$, we define penalty-like weights:

$$w_i(\rho) = \exp(-\rho\, u_{A,i}),$$

and optimize a weighted DPO objective:

$$\mathcal{L}_{\text{wDPO}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i \cdot \ell_{\text{DPO}}(x_i, y_i^+, y_i^-), \qquad \tilde{w}_i = \frac{w_i}{\bar{w}},$$

where $\bar{w} = \mathbb{E}_{\text{train}}[w]$ is a precomputed global mean weight (used to keep effective loss scale stable). We report $\rho \in \{1.0, 3.0\}$.

Model and training configuration. **Base model:** Qwen/Qwen2.5-0.5B-Instruct.

Parameter-efficient tuning: LoRA on attention projections (`q_proj, k_proj, v_proj, o_proj`) with rank $r = 8$, $\alpha = 16$, dropout 0.05.

DPO hyperparameters: $\beta = 0.1$ (sigmoid loss), learning rate 10^{-4}, 1 epoch, per-device batch size 1 with gradient accumulation 16, warmup ratio 0.03, max prompt length 512, max sequence length 1024.

Dataset: HH-RLHF (Anthropic Helpful-Harmless) with standard train/test split. Training on clean subset; evaluation on 100-example test set.

Evaluation metrics. We evaluate preference-following by computing the completion log-probability gap:

$$\text{gap}_i = \log p_\theta(y_i^+|x_i) - \log p_\theta(y_i^-|x_i),$$

where $\log p_\theta$ is the sum of token log-probabilities over completion tokens. We report:

- **Preference accuracy:** $\frac{1}{N}\sum_i \mathbb{I}[\text{gap}_i > 0]$ with bootstrap 95% CI (2,000 resamples).
- **Mean log-probability gap:** $\frac{1}{N}\sum_i \text{gap}_i$.

B.4 Exploratory additional tasks

We ran pilot studies on Toxicity and IMDB using GPT2-L and LLaMA-7B to test the margin gains, like the proxy uncertainty experiment addressed in the Experiments. Here we used a simplified margin-shift variant of DPO: a per-example scalar shift m_i (constant/length/rarity) is subtracted from the DPO logit before the sigmoid (Eq. (7) with $\kappa(x, y^+) - \kappa(x, y^-) = m_i$).