DRO–REBEL: Distributionally Robust Relative-Reward Regression for Fast and Efficient LLM Alignment

Sharan Sahu
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853
ss4329@cornell.edu

Martin T. Wells
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853
mtw1@cornell.edu

Abstract

Reinforcement Learning with Human Feedback (RLHF) has become crucial for aligning Large Language Models (LLMs) with human intent. However, existing offline RLHF approaches suffer from overoptimization, where language models degrade by overfitting inaccuracies and drifting from preferred behaviors observed during training [Huang et al., 2025]. Recent methods introduce Distributionally Robust Optimization (DRO) to address robustness under preference shifts, but these methods typically lack sample efficiency, rarely consider diverse human preferences, and use complex heuristics [Mandal et al., 2025, Xu et al., 2025, Chakraborty et al., 2024]. We propose *DRO–REBEL*, a unified family of robust REBEL updates [Gao et al., 2024] instantiated with type-p Wasserstein, Kullback–Leibler (KL), and χ^2 ambiguity sets. Leveraging Fenchel duality, each DRO–REBEL step reduces to a simple relative-reward regression, preserving REBEL's scalability and avoiding PPO-style clipping or value networks. Under standard linear-reward and log-linear policy classes with a data-coverage assumption, we prove "slow-rate" $O(n^{-1/4})$ estimation-error bounds featuring substantially tighter constants than prior DRO-DPO methods, and we further recover the minimax-optimal "fast-rate" $O(n^{-1/2})$ via a localized Rademacher complexity argument. Crucially, by adapting our localized-complexity analysis to the WDPO and KLDPO algorithms of Xu et al. [2025], we close their existing $O(n^{-1/4})$ vs. $O(n^{-1/2})$ gap and show that both Wasserstein DPO and KL-DPO likewise attain the optimal parametric rate. We also derive tractable SGD-based algorithms for each divergence—using gradient regularization for Wasserstein, importance weighting for KL, and an efficient 1-D dual solve for χ^2. Experiments on Emotion Alignment and the large-scale ArmoRM multi-objective benchmark show that DRO-REBEL attains the strong worst-case performance, outperforming baselines and prior DRO variants across unseen preference mixtures, model sizes, and dataset scales. In addition, we run a controlled radius–coverage experiment with a log-linear policy and χ^2 mixture ambiguity to validate our theory of the coverage–rate trade-off: radii that shrink faster than $n^{-1/2}$ recover ERM-level $O(n^{-1/2})$ rates but sacrifice coverage, whereas radii that guarantee the data-generating distribution remains inside the ball inevitably slow the rate to $O(n^{-1/4})$.

1 Introduction

RLHF has emerged as one of the most important stages of aligning LLMs with human intent [Christiano et al., 2023, Ziegler et al., 2020]. Typically after supervised fine-tuning (SFT), an additional alignment phase is often required to refine their behavior based on human feedback. The alignment of LLMs with human values and preferences is a central objective in machine learning, enabling these models to produce outputs that are useful, safe, and aligned with human intent. In RLHF, human evaluators provide preference rankings that are subsequently utilized to train a reward model, guiding a policy optimization step to maximize learned rewards [Ouyang et al., 2022]. Despite its success, standard RLHF methodologies are fragile mainly due to three reasons: *(i) Assumption that one reward model can model diverse human preferences:* Many RLHF methodologies including popular methods such as Direct Preference Optimization (DPO) [Rafailov et al., 2024] and Proximal Policy Optimization (PPO) [Schulman et al., 2017] assume that a single reward function can model and accurately capture diverse human preferences. In reality, human preferences are highly diverse, context-dependent, and distributional, making it infeasible to represent them within one single reward function. To this end, there has been work done in creating Bayesian frameworks for robust reward modeling [Yan et al., 2024], modeling loss as a weighted combination of different topics and using out-of-distribution detection to reject bad behavior [Bai et al., 2022], or formulating a mixture of reward models [Chakraborty et al., 2024]. *(ii) Reward hacking:* Alignment depends on the quality of the human preference data collected. Unfortunately, this process

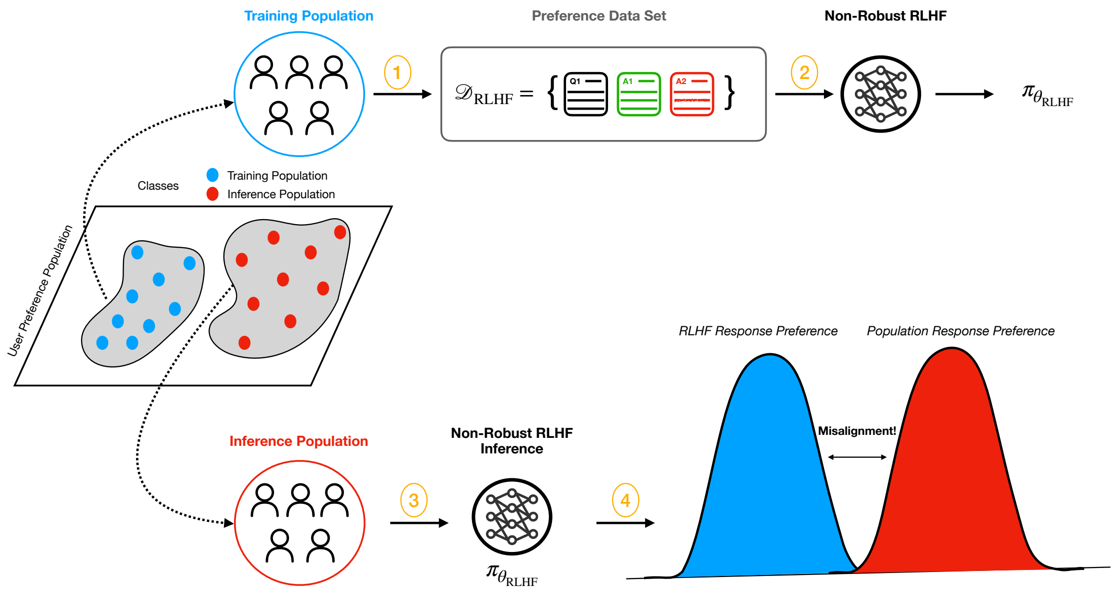

Figure 1: Non-robust RLHF under distributional shift. Pairwise preference data $\mathcal{D}_{\text{RLHF}}$ collected from a training population (blue) are used to learn a policy $\pi_{\theta_{\text{RLHF}}}$ via standard RLHF. In latent preference–feature space, the inference population (red) occupies a disjoint region from the training cohort (blue), indicating a shift. When $\pi_{\theta_{\text{RLHF}}}$ is deployed on these out-of-distribution users, its induced response-preference distribution (blue) diverges from the true population preferences (red), resulting in systematic misalignment.

is inherently noisy and prone to bias, conflicting opinions, and inconsistency which leads to misaligned preference estimation. This issue is exacerbated by reward hacking where instead of learning reward functions that are aligned with genuine human intent, models learn undesirable shortcuts to maximize the estimated reward function. Subsequently, these models appear to generate responses that appear aligned but deviate from human intent. There are some works that directly address this such as [Bukharin et al., 2024]. *(iii) Distribution shift:* Standard RLHF alignment algorithms use static preference datasets for training, collected under controlled conditions. However, the preferences of real-world users can often be out-of-distribution from that of the training data, depending on several factors such as geographic location, demographics, etc. Thus, a language model in the face of distribution shift may see catastrophic degradation in performance due to overfitting inaccuracies and diverging from human-preferred responses encountered in training data [LeVine et al., 2024, Kirk et al., 2024, Casper et al., 2023]. We focus on the problem of distribution shift, also known as *overoptimization* [Huang et al., 2025].

Recently, distributionally robust RLHF methods have emerged to tackle robustness challenges under distributional shifts in prompts and preferences [Mandal et al., 2025, Xu et al., 2025]. Specifically, Mandal et al. [2025] and Xu et al. [2025] introduced DRO variants of popular RLHF methods, namely DPO and PPO, employing uncertainty sets defined via Chi-Squared (χ^2), type-p Wasserstein, and Kullback–Leibler (KL) divergences. Unfortunately, it is known that PPO requires multiple heuristics to enable stable convergence (e.g. value networks, clipping), and is notorious for its sensitivity to the precise implementation of these components. offline RLHF methods such as DPO, while not explicitly using a reward model, learn an implicit reward model of the form $r_\phi(x, y) = \beta \log \left(\pi_\theta(y \mid x)/\pi_{\text{SFT}}(y \mid x)\right)$ which provides higher reward to preferred responses over dispreferred responses. This can lead to a brittle solution that overfits the preference distribution seen during training. When faced with a new type of prompt (a distributional shift), the learned policy might fail because the implicit reward signal it relies on does not generalize [Xu et al., 2024]. Recently, Gao et al. [2024] proposed REBEL, an algorithm that cleanly reduces the problem of policy optimization to regressing the relative reward between two completions to a prompt in terms of the policy. They find that REBEL avoids the use of "unjustified" heuristics like PPO and enjoys strong convergence and regret guarantees, similar to Natural Policy Gradient [Kakade, 2001], while also being scalable due to not requiring inversion of the Fisher information matrix. REBEL is much more sample efficient compared to methods like DPO and PPO and by learning a disentangled representation of human preferences, REBEL is better equipped to generalize under distributional shifts. Given the fragility of PPO-style updates and the slower rates observed for robust DPO, this motivated us to answer the following questions:

> *Can we obtain distributional robustness to preference shift without sacrificing sample efficiency and stability by adopting a simpler, regression-based algorithm such as REBEL? Concretely, does combining DRO with REBEL yield a learning rule that is (i) theoretically sound (rates and excess risk), (ii) practically scalable (no brittle heuristics), and (iii) empirically more generalizable under realistic distribution shift?*

Our contributions. Inspired by the strong theoretical guarantees of REBEL in terms of sample efficiency and simplicity, we introduce *DRO–REBEL*, a family of robust REBEL updates for RLHF under distributional shifts, and make the following advances:

1. **Unified robust updates via duality.** We extend REBEL to type-p Wasserstein, KL, and χ^2 ambiguity sets. In each case the robust update admits a closed-form dual reformulation that reduces to a single relative–reward regression step, preserving REBEL's scalability and avoiding heuristic stabilizers.
2. **Sharper slow-rate guarantees.** Under standard linear-reward and log-linear policy assumptions with a data-coverage condition, we prove an $O(n^{-1/4})$ estimation error for all DRO–REBEL variants. By replacing logistic links with linear regression, we remove hidden exponential curvature factors and tighten constants versus prior DRO–DPO analyses [Xu et al., 2025, Mandal et al., 2025]. The strong-convexity modulus depends only on the coverage constant λ and step size η, rather than Bradley–Terry curvature as in WDPO/KLDPO.
3. **Minimax-optimal fast rates.** To the best of our knowledge, we provide the first proof that DRO-based LLM alignment under preference shift attains the parametric $O(n^{-1/2})$ rate, matching non-robust RLHF in benign regimes via a localized Rademacher complexity argument. Applying the same machinery to WDPO and KLDPO yields $O(n^{-1/2})$ estimation rates, improving upon the $O(n^{-1/4})$ bounds reported in prior DRO–DPO theory [Xu et al., 2025, Mandal et al., 2025].
4. **Radius–coverage trade-off.** We prove that no fixed radius can both maintain nonvanishing coverage of the true distribution and uniformly deliver the parametric $O(n^{-1/2})$ rate; requiring coverage yields an $O(n^{-1/4})$ worst-case rate. We visualize this frontier in simulation, and to the best of our knowledge, this is the first such characterization in LLM alignment.
5. **Extensive empirical validation.** We evaluate on (i) a controlled Emotion Alignment task [Saravia et al., 2018], (ii) a large-scale ArmoRM multi-objective setting [Wang et al., 2024a], and (iii) a *radius–coverage/convergence* study in a Gaussian-mixture simulator that trains a log-linear policy under group-level χ^2–DRO. The first two replicate the benchmarks of Xu et al. [2025] for direct comparison; the third isolates the statistical trade-off between ambiguity radius, coverage, and estimation error, and empirically validates the Pearson/Wilson–Hilferty calibration $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ with K groups.

1.1 Related Work

Robust RLHF: There has been some recent work in this area that aims to address RLHF overoptimization. Bai et al. [2022] propose addressing distribution shift by adjusting the weights on the combination of loss functions based on different topics (harmless vs. helpful) for robust reward learning. They also propose using out-of-distribution detection to filter and reject known types of bad behavior. [Chakraborty et al., 2024] proposes a MaxMin approach to RLHF, using mixtures of reward models to honor diverse human preference distributions through an expectation-maximization approach, and a robust policy based on these rewards via a max-min optimization. In a similar vein, Padmakumar et al. [2024] tries to augment the human preference datasets with synthetic preference judgments in order to estimate the diversity of user preferences. There has also been some foundational theoretical work towards this problem. Yan et al. [2024] proposed a Bayesian reward model ensemble to model the uncertainty set of the reward functions and systematically choose rewards in the uncertainty set with the tightest confidence band. Another line of work focuses on robust reward modeling as an alternative to distributionally robust optimization. For instance, Bukharin et al. [2024] propose R3M, a method that explicitly models corrupted preference labels as sparse outliers. They formulate reward learning as an ℓ_1-regularized maximum likelihood estimation problem, enabling robust recovery of the underlying reward function even in the presence of noisy or inconsistent human feedback. While our work focuses on embedding robustness at the policy optimization level using distributional uncertainty sets (e.g., χ^2, and Wasserstein), R3M represents a complementary direction that enhances robustness by improving the reliability of the reward model itself.

Robust DPO: There have been several works that approach this problem using DRO. Huang et al. [2025] proposed χPO that implements the principle of pessimism in the face of uncertainty via regularization with the χ^2-divergence for avoiding reward hacking/overoptimization with respect to the estimated reward. Wu et al. [2024] focus on noisy preference data and categorize the types of noise in DPO, introducing Dr. DPO to improve pairwise robustness through a DRO formulation with a tunable reliability parameter. Hong et al. [2024] propose an adaptive preference loss grounded in DRO that adjusts scaling weights across preference pairs to account for ambiguity in human feedback, enhancing reward estimation flexibility and policy performance. Separately, Zhang et al. [2024] introduce a lightweight uncertainty-aware approach called AdvPO, combining last-layer embedding-based uncertainty estimation with a DRO formulation to address overoptimization in reward-based RLHF. There are two related works that are most similar with our approach. Xu et al. [2025] develop Wasserstein and KL-based DRO formulations of Direct Preference Optimization (WDPO and KLDPO), providing sample complexity bounds and scalable gradient-based algorithms. Their methods achieve improved alignment performance under shifting user preference distributions. Similarly, Mandal et al. [2025] propose robust variants of both reward-based and reward-free RLHF methods, incorporating DRO into the reward estimation and policy optimization phases using Total Variation and Wasserstein distances. Their algorithms retain the structure of existing RLHF pipelines while providing theoretical convergence guarantees and demonstrating robustness to out-of-distribution (OOD) tasks.

Distributionally Robust Learning: The DRO framework has been applied to various areas ranging from supervised learning [Namkoong and Duchi, 2017b, Shah et al., 2020], reinforcement learning [Zhang et al., 2020, Yang et al., 2021], and multi-armed bandits [Gao et al., 2022, Zhou et al., 2022]. There is a wealth of theoretical results using f-divergences and Wasserstein distances developed for tackling problems in this setting [Duchi and Namkoong, 2022, Shapiro and Xu, 2022].

2 Preliminaries

2.1 Notations

We will denote sets using calligraphic letters i.e. $\mathcal{S}, \mathcal{A}, \mathcal{Z}$. For a measure $\mathbb{P}$, we refer to the empirical measure $\mathbb{P}_n$ to mean drawing samples $x_1, \ldots, x_n \overset{\text{i.i.d}}{\sim} \mathbb{P}$ with $\mathbb{P}_n = 1/n \sum_{i=1}^n \delta_{x_i}$ where δ is the Dirac measure. We denote $\ell(z;\theta)$ to be the loss incurred by sample z with policy parameter θ. We denote $\mathcal{M}(\mathcal{Z})$ to be the set of Borel measures supported on set $\mathcal{Z}$. Lastly, we denote $\lambda_{\min}(A)$ to be the minimum eigenvalue of a symmetric matrix $A \in \mathbb{S}^n$. We adopt the standard big-oh notation, and write $a \lesssim b$ as shorthand for $a = O(b)$ and $a \vee b = \max\{a, b\}$.

2.2 Divergences

In this section, we will define the divergences that we will use to define our ambiguity sets in the DRO setting.

Definition 2.1 (Type-p Wasserstein Distance)**.** *The type-p ($p \in [1,\infty)$) Wasserstein distance between two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\mathcal{Z})$ is defined as*

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{Z}\times\mathcal{Z}} d(\xi,\eta)^p \pi(d\xi, d\eta) \right)^{1/p}$$

where π is a coupling between the marginal distributions $\xi \sim \mathbb{P}$ and $\eta \sim \mathbb{Q}$ and d is a pseudometric defined on $\mathcal{Z}$.

Along these lines, we also define the Entropically–Regularised (Sinkhorn) Wasserstein and σ-smooth p-Wasserstein distance as they will include an important distinction that other divergences are not able to satisfy in this setting (see Discussion 4.3)

Definition 2.2 (Entropically–Regularised (Sinkhorn) Type-p Wasserstein Distance)**.** *Fix a cost metric $d : \mathcal{Z} \times \mathcal{Z} \to [0,\infty)$ and a regularisation parameter $\tau > 0$. For two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\mathcal{Z})$ and any $p \in [1,\infty)$, the entropically–regularised type-p Wasserstein distance (often called the Sinkhorn distance) is defined as*

$$\mathcal{W}_{p,\tau}(\mathbb{P}, \mathbb{Q}) \;=\; \left(\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \left\{ \int_{\mathcal{Z}\times\mathcal{Z}} d(\xi,\eta)^p \, \pi(d\xi, d\eta) \;+\; \tau \, \mathrm{KL}\big(\pi \,\|\, \mathbb{P}\otimes\mathbb{Q}\big) \right\} \right)^{1/p},$$

where $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all couplings of $\xi \sim \mathbb{P}$ and $\eta \sim \mathbb{Q}$, and $\mathrm{KL}(\pi \,\|\, \mathbb{P}\otimes\mathbb{Q})$ denotes the Kullback–Leibler divergence between the joint coupling π and the independent product measure $\mathbb{P}\otimes\mathbb{Q}$. Setting $\tau = 0$ recovers the classical (unregularised) Wasserstein distance $\mathcal{W}_p$.

Definition 2.3 (σ-Smooth p-Wasserstein Distance)**.** *Let $p \in [1, \infty)$ and let $\mathcal{P}_p$ be the set of probability measures with a finite p-th moment. For two probability measures $\mu, \nu \in \mathcal{P}_p$, the σ-smooth p-Wasserstein distance is defined as*

$$\mathcal{W}_p^{(\sigma)}(\mu, \nu) := \mathcal{W}_p(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma$ is a centered Gaussian distribution with covariance matrix $\sigma^2 I$, and $\mu * \mathcal{N}_\sigma$ denotes the convolution of the probability measure μ with the Gaussian distribution $\mathcal{N}_\sigma$.

Definition 2.4 (Kullback-Leibler (KL) Divergence)**.** *For any two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\mathcal{Z})$, the Kullback-Liebler (KL) Divergence is defined as*

$$D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{Q}) = \int_{\mathcal{Z}} \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}$$

Definition 2.5 (Chi-Squared Divergence)**.** *For any two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\mathcal{Z})$ such that $\mathbb{P} \ll \mathbb{Q}$ i.e. $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, the Chi-Squared (χ^2) divergence is defined as*

$$D_{\chi^2}(\mathbb{P} \,||\, \mathbb{Q}) = \int_{\mathcal{Z}} \left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right)^2 d\mathbb{Q}$$

Using these, we can define our ambiguity sets as follows

Definition 2.6 (Distributional Uncertainty Sets)**.** *Let $\varepsilon > 0$ and $\mathbb{P}^\circ \in \mathcal{M}(\mathcal{Z})$. Then, we define the ambiguity set as*

$$\mathcal{B}_\varepsilon(\mathbb{P}^\circ; D) = \{\mathbb{P} \in \mathcal{M}(\mathcal{Z}) \;:\; D(\mathbb{P}, \mathbb{P}^\circ) \leq \varepsilon\}$$

where $D(\cdot, \cdot)$ is a distance metric between two probability measures for instance type-p Wasserstein, KL, χ^2, etc.

2.3 Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF), as introduced by Christiano et al. [2023] and later adapted by Ouyang et al. [2022], consists of two primary stages: (1) learning a reward model from preference data, and (2) optimizing a policy to maximize a KL-regularized value function. We assume access to a preference dataset $\mathcal{D}_{\text{src}} = \{(x, a^1, a^2)\}$ where $x \in \mathcal{S}$ is a prompt and $a^1, a^2 \in \mathcal{A}$ are two possible completions of the prompt x generated from a reference policy $\pi_{\text{SFT}}(\cdot \mid x)$ (e.g., a supervised fine-tuned model). $\pi_{\text{SFT}}(\cdot \mid x)$ involves fine-tuning a pre-trained LLM through supervised learning on high-quality data, curated for downstream tasks. A human annotator provides preference feedback indicating $a^1 \succ a^2 \mid x$. The most common model for preference learning is the Bradley-Terry (BT) model, which assumes that

$$\begin{aligned}\mathcal{P}^*(a^1 \succ a^2 \mid x) &= \sigma\left(r^\star(x, a^1) - r^\star(x, a^2)\right) \\ &= \frac{1}{1 + \exp\left(r^\star(x, a^1) - r^\star(x, a^2)\right)},\end{aligned}$$

where $r^\star$ is the underlying (unknown) reward function used by the annotator. The first step in RLHF is to learn a parametrized reward model $r_\phi(s, a)$ by solving the following maximum likelihood estimation problem:

$$r_\phi \leftarrow \arg\min_{r_\phi} \; -\mathbb{E}_{(x, a^1, a^2) \sim \mathcal{D}_{\text{src}}}\left[\log \sigma\left(r_\phi(x, a^1) - r_\phi(x, a^2)\right)\right].$$

Given the learned reward model r_ϕ, the second step is to solve the KL-regularized policy optimization problem:

$$\pi_\theta \leftarrow \arg\max_{\pi_\theta} \; \mathbb{E}_{x \sim \mathcal{D}_{\text{src}}, a \sim \pi_\theta(\cdot|x)}\left[r_\phi(x, a) - \beta \log \frac{\pi_\theta(a \mid x)}{\pi_{\text{SFT}}(a \mid x)}\right],$$

where β controls the deviation between the learned and reference policy.

2.4 Direct Preference Optimization (DPO)

The DPO [Rafailov et al., 2024] procedure is a form of offline RLHF which avoids issues in previous policy optimization algorithms like PPO [Schulman et al., 2017] by identifying a mapping (and re-parametrization) between language model policies and reward functions that enables training LMs to satisfy human preferences directly without using RL or even doing reward model fitting. That is, DPO makes use of the following policy objective

$$\ell_{\mathrm{DPO}}\left(\pi_\theta ; \pi_{\mathrm{SFT}}\right)=-\mathbb{E}_{\left(x, a^1, a^2\right) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta\left(a^1 \mid x\right)}{\pi_{\mathrm{SFT}}\left(a^1 \mid x\right)}-\beta \log \frac{\pi_\theta\left(a^2 \mid x\right)}{\pi_{\mathrm{SFT}}\left(a^2 \mid x\right)}\right)\right] \tag{1}$$

One can arrive at this policy objective by observing that the objective in the RL fine-tuning phase has a closed form solution of the form

$$\pi_\theta(a \mid x)=\frac{1}{Z(x)} \pi_{\mathrm{SFT}}(a \mid x) \exp \left(\frac{1}{\beta} r_\phi(x, a)\right)$$

where $Z(x)=\sum_{a \in \mathcal{A}} \pi_{\mathrm{SFT}}(a \mid x) \exp \left(\frac{1}{\beta} r_\phi(x, a)\right)$ is the partition function. Taking logs and moving terms around, we get

$$r_\phi(x, a)=\beta \log \frac{\pi_\theta(a \mid x)}{\pi_{\mathrm{SFT}}(a \mid x)}+\beta \log Z(x) \tag{2}$$

Using the BT model and plugging in this re-parametrization, we get the DPO policy objective. One might note that this is not a proper re-parametrization since $Z(x)$ is dependent on r. However, it turns out that defining this re-parametrization as a function from an equivalence class of reward functions to a particular policy is well-defined, does not constrain the class of learned reward models, and allows for the exact recovery of the optimal policy.

2.5 REBEL: Regression-Based Policy Optimization

Let (x, a) represent a *prompt-response* pair, where $x \in \mathcal{S}$ is a context or prompt, and $a \in \mathcal{A}$ is a response (e.g., a sequence of tokens or actions). We assume access to a reward function $r(x, a)$, which may be a learned preference model [Christiano et al., 2023]. Let $\pi\colon \mathcal{S} \rightarrow \Delta(\mathcal{A})$ be a stochastic policy mapping prompts to distributions over responses. We denote the prompt distribution as ρ, and let $\pi_\theta(a \mid x)$ denote a parameterized policy with parameters θ. Choose $\beta = 1/\eta$. Now instead of using the reward parametrization (2) in the BT model, instead notice that we can get rid of the partition function $Z(x)$ by sampling two responses $a^1, a^2 \sim \pi_{\theta_t}(\cdot \mid x)$ at time step t and taking the difference of the reparametrized reward function

$$r(x, a^1)-r(x, a^2)=\frac{1}{\eta}\left(\log \frac{\pi_\theta(a^1 \mid x)}{\pi_{\theta_t}(a^1 \mid x)}-\log \frac{\pi_\theta(a^2 \mid x)}{\pi_{\theta_t}(a^2 \mid x)}\right)$$

Then we can simply regress the difference in rewards and update the policy parameters as follows

$$\theta_{t+1}=\arg\min_{\theta \in \Theta}\left(\frac{1}{\eta}\left[\log \frac{\pi_\theta(a^1 \mid x)}{\pi_{\theta_t}(a^1 \mid x)}-\log \frac{\pi_\theta(a^2 \mid x)}{\pi_{\theta_t}(a^2 \mid x)}\right]-\left[r(x, a^1)-r(x, a^2)\right]\right)^2 \tag{3}$$

The **REBEL** (REgression to RElative REward-Based RL) [Gao et al., 2024] algorithm directly regresses to relative reward differences through KL-constrained updates. The REBEL algorithm is detailed in Algorithm 1.

Algorithm 1 REBEL: REgression to RElative REward-Based RL

1: **Input:** Reward function r, policy class $\Pi = \{\pi_\theta : \theta \in \Theta\}$, base distribution μ, learning rate η
2: Initialize policy π_{θ_0}
3: **for** $t = 0$ to $T-1$ **do**
4: Collect dataset $\mathcal{D}_t = \{(x, a^1, a^2)\}$ with $x \sim \rho$, $a^1 \sim \pi_{\theta_t}(\cdot \mid x)$, $a^2 \sim \pi_{\theta_t}(\cdot \mid x)$
5: Update policy by solving:

$$\theta_{t+1}=\arg\min_{\theta \in \Theta} \sum_{(x, a^1, a^2) \in \mathcal{D}_t}\left(\frac{1}{\eta}\left[\log \frac{\pi_\theta(a^1 \mid x)}{\pi_{\theta_t}(a^1 \mid x)}-\log \frac{\pi_\theta(a^2 \mid x)}{\pi_{\theta_t}(a^2 \mid x)}\right]-\left[r(x, a^1)-r(x, a^2)\right]\right)^2$$

6: **end for**

At each iteration, REBEL approximates the solution to a KL-constrained policy optimization objective:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{x\sim\rho, a\sim\pi(\cdot|x)} \left[r(x,a) \right] - \frac{1}{\eta} \mathbb{E}_{x\sim\rho} \left[\mathrm{KL} \left(\pi(\cdot \mid x) \,\|\, \pi_t(\cdot \mid x) \right) \right],$$

which encourages reward maximization while regularizing the policy to remain close to the previous iterate. This objective is particularly well-suited for fine-tuning language models using learned or noisy reward signals while maintaining stability in training.

Adapting REBEL for distributionally robust RLHF is particularly appealing because it offers distinct theoretical and practical advantages over existing methods. Actor-critic methods such as PPO rely on complex and often unstable heuristic mechanisms (e.g., clipping, value baselines) to ensure stability while offline RLHF methods such as DPO, while not explicitly using a reward model, learn an implicit reward model of the form $r_\phi(x,y) = \beta \log \left(\pi_\theta(y \mid x) / \pi_{\mathrm{SFT}}(y \mid x) \right)$ which provides higher reward to preferred responses over dispreferred responses. Thus the model is not learning "what humans prefer" in a general sense; it's learning to adjust log-probabilities to satisfy a pairwise ranking objective. This can lead to a brittle solution that overfits the preference distribution seen during training. When faced with a new type of prompt (a distributional shift), the learned policy might fail because the implicit reward signal it relies on does not generalize [Xu et al., 2024]. In contrast, REBEL takes a more direct and robust approach. It reduces policy optimization to a sequence of simple regression problems on explicit relative rewards. REBEL's regression objective learns a cardinal signal that captures how much better one response is than another and since the reward model is trained explicitly on the task of "predicting preference differences," it is more likely to generalize to unseen prompts. By learning a disentangled representation of human preferences, REBEL is better equipped to generalize under distributional shifts.

This fundamental simplicity also eliminates the need for explicit value functions or constrained optimization, translating into significantly improved stability and sample complexity. In particular, REBEL can achieve convergence guarantees comparable to or better than NPG, with a sample complexity that scales favorably due to its variance-reduced gradient structure. Empirically, REBEL has been shown to converge faster than PPO and outperform DPO in both language and image generation tasks. Building on this regression-based perspective, our DRO–REBEL algorithms simply replace the standard squared-error loss in each REBEL update with its robust counterpart under the chosen divergence (Wasserstein, KL, or χ^2). As a result, DRO–REBEL inherits REBEL's stability and low sample complexity while gaining worst-case robustness guarantees under distributional shifts.

3 Distributionally Robust REBEL and DPO

In this section, we will formally define the DRO variants of REBEL and DPO under type-p Wasserstein, KL, and χ^2 divergence ambiguity sets. We must first define the nominal data-generating distribution. Our definitions follow those stated by Xu et al. [2025]. Recall the sampling procedure mentioned in Section 2.3: We have some initial prompt $x \in \mathcal{S}$ that we will assume is sampled from some prompt distribution ρ. We will sample two responses $a^1, a^2 \overset{\text{i.i.d}}{\sim} \pi_{\mathrm{SFT}}(\cdot \mid x)$. Following Zhu et al. [2023], let $y \in \{0,1\}$ be a Bernoulli random variable where $y=1$ if $a^1 \succ a^2 \mid x$ and $y=0$ if $a^2 \succ a^1 \mid x$ with probability corresponding to the Bradley-Terry model $\mathcal{P}^*$. Using this, we can now define the nominal data-generating distribution.

Definition 3.1 (Nominal Data-Generating Distribution)**.** *Let $\mathcal{Z} = \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times \{0,1\}$. Then, we define the nominal data-generating distribution as follows*

$$\mathbb{P}^\circ \left(x, a^1, a^2, y \right) = \rho(x) \pi_{\mathrm{SFT}} \left(a^1 \mid x \right) \pi_{\mathrm{SFT}} \left(a^2 \mid x \right) \left[\mathbb{1}_{\{y=1\}} \mathcal{P}^* \left(a^1 \succ a^2 \mid x \right) + \mathbb{1}_{\{y=0\}} \mathcal{P}^* \left(a^2 \succ a^1 \mid x \right) \right]$$

where $x \sim \rho$ and $y \sim \mathrm{Ber}\left(\mathcal{P}^* \left(a^1 \succ a^2 \mid \cdot \right) \right)$. We will denote $z = (x, a^1, a^2, y) \in \mathcal{Z}$ and $\mathbb{P}^\circ(z) = \mathbb{P}^\circ(x, a^1, a^2, y)$. We will also assume $\mathbb{P}^\circ$ generates dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ used for learning i.e. $z_i \sim \mathbb{P}^\circ$. Additionally suppose $\psi, \phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ are known d-dimensional feature mappings with $\max_{x,a} \|\psi(x,a)\|_2 \le 1$, $\sup_{x,a} ||\phi(x,a)||_2 \le 1$. Denote

$$\Delta\psi(z) := \psi(x, a^1) - \psi(x, a^2), \quad \Delta r(z) := \phi(x, a^1)^\top \omega - \phi(x, a^2)^\top \omega,$$

We equip $\mathcal{Z}$ with the following pseudometric (or a metric if ψ, ϕ are injective):

$$d(z, z') = \|\Delta\psi - \Delta\psi'\|_2 + |\Delta r - \Delta r'| + |y - y'|$$

We endow the data space $\mathcal{Z}$ with this ground metric for a few reasons. (i) *Practical RLHF robustness*: In preference datasets small perturbations typically arise from replacing an action by another that has almost the same feature vector, relabelling noise, or tiny errors in the proxy reward. Each source is captured by one summand of d. These emphasize the learner along realistic failure modes rather than arbitrary pixel-level noise. That is, this metric treats two data points as close if and only if an expert would find their preference information almost identical. (ii) *Theoretical tightness*: each summand feeds through a 1-Lipschitz (or linear) mapping into the loss, hence the loss is $L_{\ell,z}$-Lipschitz with the *smallest possible constant*; this gives the sharp dual-remainder bound $\Delta_n \leq L_{\ell,z}\varepsilon_n$ required by our $n^{-1/2}$ master theorem (see Theorem 4) Thus metric d simultaneously aligns with task semantics, models realistic perturbations, and yields convenient analytic bounds.

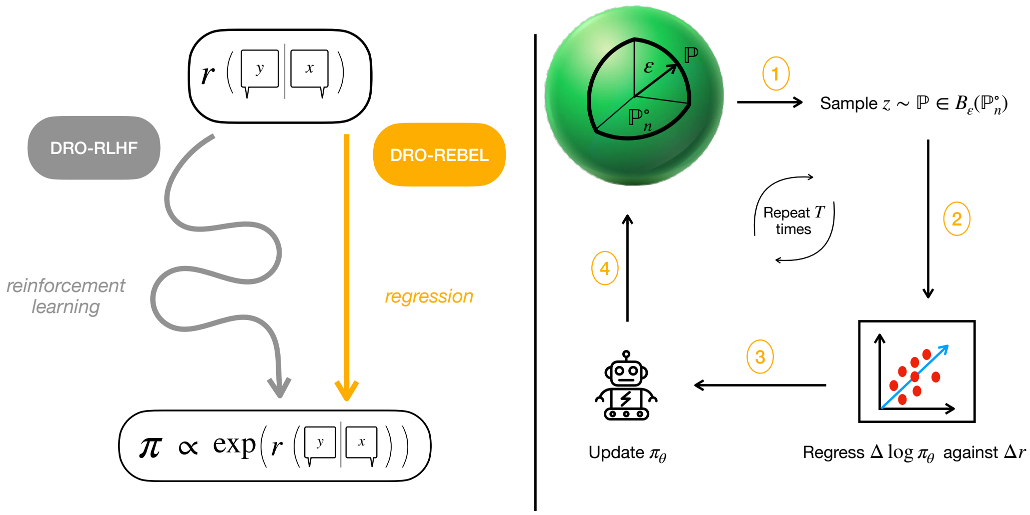

Figure 2: Illustration of the DRO–REBEL update loop. At each iteration, (1) we draw a batch of preference tuples $z = (x, a^1, a^2)$ from the worst-case distribution $\mathbb{P}$ within the ε–ambiguity set around the empirical data $\mathbb{P}_n^\circ$; (2) we compute the per-sample squared-error loss $\ell_{\mathrm{REBEL}}(z;\theta)$; (3) we perform a relative-reward regression weighted by $\mathbb{P}$ to fit the change in log-probabilities $\Delta \log \pi_\theta$ against the observed reward differences Δr; and (4) we update the policy parameters θ via a gradient step. Repeating these four steps yields a policy that is robust to distributional shifts in user preferences. The leftside figure is credited to Gao et al. [2024]

We also define the following quantity that we will discuss in our analysis of a "fast rate" for DPO and REBEL under type-p Wasserstein ambiguity sets

Definition 3.2 (Intrinsic (support) dimension). *Let $Z\colon \mathcal{Z} \longrightarrow \mathbb{R}^d \times \mathbb{R} \times \{0,1\}$, $z \mapsto (\Delta\psi(z), \Delta r(z), y)$, and let $\mu := Z_{\#}\mathbb{P}$ be the push-forward of a probability measure $\mathbb{P}$ on $(\mathcal{Z}, d)$; i.e. $\mu(A) = \mathbb{P}\big(Z^{-1}(A)\big)$ for all Borel $A \subset \mathbb{R}^{d+2}$. For an integer $k \geq 0$ write*

$$\mathcal{R}^k := \Big\{R \subset \mathbb{R}^{d+2} : R \text{ is } k\text{-rectifiable}\Big\},$$

*where a k-*rectifiable set *is any set that can be covered, up to a $\mathcal{H}^k$-null subset, by countably many Lipschitz images of bounded subsets of $\mathbb{R}^k$ (Ambrosio et al., 2000, Def. 2.55).*

The intrinsic support dimension *of the feature–gap vector is*

$$m := \inf\Big\{k \in \mathbb{N} \;\Big|\; \exists\, R \in \mathcal{R}^k \text{ s.t. } \mu\big(R\big) = 1\Big\}.$$

Equivalently, m is the least integer for which the support $\operatorname{supp}(\mu)$ *has finite k-dimensional Hausdorff measure $\mathcal{H}^k$.*

The quantity m counts the smallest number of real degrees of freedom needed to describe *almost all* feature–gap vectors generated by $\mathbb{P}$: if $m = 2$, then up to a set of probability 0 every point $\big(\Delta\psi, \Delta r, y\big)$ lies on a two-dimensional manifold embedded in the (typically huge) ambient space $\mathbb{R}^{d+2}$. In Wasserstein concentration bounds the rate depends on m rather than on the ambient dimension; hence a high-parametric model may still enjoy fast $n^{-1/2}$ rates whenever its feature-gap distribution is effectively low-dimensional.

3.1 Distributionally Robust REBEL

From the REBEL update (Equation (3)), we define the pointwise loss as follows

$$\ell_{\mathrm{REBEL}}(z;\theta)=\left(\frac{1}{\eta}\left[\log\frac{\pi_\theta(a^1\mid x)}{\pi_t(a^1\mid x)}-\log\frac{\pi_\theta(a^2\mid x)}{\pi_t(a^2\mid x)}\right]-\left[r(x,a^1)-r(x,a^2)\right]\right)^2$$

For $\varepsilon>0$, define the ambiguity set as $\mathcal{B}_\varepsilon\left(\mathbb{P}^\circ;D\right)$ for nominal distribution $\mathbb{P}^\circ$ and distance measure D. Using the DRO framework, we consider the following distributionally robust optimization problem:

$$\min_\theta\max_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;D)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell_{\mathrm{REBEL}}(z;\theta)\right]$$

which directly captures our objective: finding the best policy under worst-case distributional shift. Now, let us define the following D-DRO-REBEL loss function:

$$\mathcal{L}^D\left(\theta;\varepsilon\right)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;D)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell_{\mathrm{REBEL}}(z;\theta)\right]$$

where $\mathcal{B}_\varepsilon(\mathbb{P}^\circ;D)$ denotes an ambiguity set centered at the nominal distribution $\mathbb{P}^\circ$, defined using a divergence or distance D. This formulation is general and allows us to instantiate a family of distributionally robust REBEL objectives by choosing different D—such as the type-p Wasserstein distance, Kullback–Leibler (KL) divergence, or chi-squared (χ^2) divergence. Each choice of D yields a different robustness profile and tractable dual formulation, enabling us to tailor the algorithm to specific distributional shift scenarios. When the nominal distribution $\mathbb{P}^\circ$ is replaced with its empirical counterpart, i.e., $\mathbb{P}_n^\circ:=\frac{1}{n}\sum_{i=1}^n\delta_{z_i}$, where $z_1,\ldots,z_n$ are n i.i.d. samples from $\mathbb{P}^\circ$, we use $\mathcal{L}_n^D(\theta;\varepsilon)$ to denote the empirical D-REBEL loss incurred by the policy parameter θ.

3.2 Distributionally Robust DPO

Similar to DRO-REBEL, we can define a distributionally robust counterpart for DPO. From the DPO loss function (Equation 2.4), we define the following pointwise loss:

$$\ell_{\mathrm{DPO}}(z;\theta)=-y\log\sigma(\beta h_\theta(s,a^1,a^2))-(1-y)\log\sigma(-\beta h_\theta(s,a^2,a^1))$$

where $h_\theta(s,a^1,a^2):=\log\frac{\pi_\theta(a^1|s)}{\pi_{\mathrm{SFT}}(a^1|s)}-\log\frac{\pi_\theta(a^2|s)}{\pi_{\mathrm{SFT}}(a^2|s)}$ is the preference score of an answer a^1 relative to answer a^2. As we did for REBEL, we can use the DRO framework to formulate the following distributionally robust optimization problem:

$$\min_\theta\max_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;D)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell_{\mathrm{DPO}}(z;\theta)\right]$$

We define the following D-DPO loss function:

$$\mathcal{L}^D\left(\theta;\varepsilon\right)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;D)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell_{\mathrm{DPO}}(z;\theta)\right]$$

4 Theoretical Results

In this section, we will provide several sample complexity results for DRO-REBEL under the ambiguity sets previously mentioned. First, we state some assumptions that we make in our analysis

Assumption 1 (Linear reward class)**.** *Let $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ be a known d-dimensional feature mapping with $\sup_{x,a}||\phi(x,a)||_2\leq 1$ and $\omega\in\mathbb{R}^d$ such that $||\omega||_2\leq F$ for $F>0$. We consider the following class of linear reward functions:*

$$\mathcal{F}:=\left\{r_\omega:r_\omega(x,a)=\phi(x,a)^\top\omega\right\}$$

Remark 1. *These are standard assumptions in the theoretical analysis of inverse reinforcement learning (IRL) [Abbeel and Ng, 2004, Ng and Russell, 2000], imitation learning [Ho and Ermon, 2016], and areas like RLHF [Zhu et al., 2023] where reward functions are learned. Our analysis can be extended to neural reward classes where $\phi(x,a)^\top\omega$ is replaced by $f_\omega(x,a)$, where f_ω is a neural network satisfying twice differentiability, smoothness, and boundedness of the f_ω and $\nabla_\omega f_\omega(x,a)$.*

Assumption 2 (Log-linear policy class). *Let $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a known d-dimensional feature mapping with $\max_{x,a} \|\psi(x,a)\|_2 \le 1$. Assume a bounded policy parameter set $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B\}$. We consider the following class of log-linear policies:*

$$\Pi := \left\{ \pi_\theta : \pi_\theta(a \mid x) = \frac{\exp\left(\theta^\top \psi(x,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x,a')\right)} \right\}.$$

Remark 2. *These are standard assumptions in the theoretical analysis of RL algorithms [Agarwal et al., 2021b, Modi et al., 2020], RLHF [Zhu et al., 2023], and DPO [Nika et al., 2024, Chowdhury et al., 2024]. Our analysis can be extended to neural policy classes where $\theta^\top \psi(s,a)$ is replaced by $f_\theta(s,a)$, where f_θ is a neural network satisfying twice differentiability and smoothness assumptions.*

We also make the following data coverage assumption on the uncertainty set $\mathcal{B}_\varepsilon(\mathbb{P}^\circ; D)$:

Assumption 3 (Regularity condition). *There exists $\lambda > 0$ such that*

$$\Sigma_{\mathbb{P}} := \mathbb{E}_{(x,a^1,a^2,y)\sim\mathbb{P}}\left[\left(\psi(x,a^1) - \psi(x,a^2)\right)\left(\psi(x,a^1) - \psi(x,a^2)\right)^\top\right] \succeq \lambda I,\ \forall \mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ; D).$$

Remark 3. *Similar assumptions on data coverage under linear architecture models are standard in the offline RL literature [Agarwal et al., 2021a, Wang et al., 2020, Jin et al., 2022]. Implicitly, Assumption 2 imposes $\lambda \le \lambda_{\min}(\Sigma_{\mathbb{P}^\circ})$, meaning the data-generating distribution $\mathbb{P}^\circ$ must have sufficient coverage.*

4.1 "Slow Rate" Estimation Errors

Define $\theta^{\mathcal{W}_p} \in \arg\min_{\theta\in\Theta} \mathcal{L}^{\mathcal{W}_p}(\theta)$ be the true optimal policy estimate and the empirical estimate as $\hat{\theta}_n^{\mathcal{W}_p} \in \arg\min_{\theta\in\Theta} \mathcal{L}_n^{\mathcal{W}_p}(\theta)$. First, we provide a sample complexity result for convergence of robust policy estimation using REBEL. Our proof technique hinges on showing that $\mathcal{L}^{\mathcal{W}_p}$ is strongly convex.

Lemma 1 (Strong convexity of $\mathcal{L}^{\mathcal{W}_p}$). *Let $\ell(z;\theta)$ be as defined in the REBEL update. The Wasserstein-DRO-REBEL loss*

$$\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell_{\mathrm{REBEL}}(z;\theta)\right],$$

is $2\lambda/\eta^2$-strongly convex where λ is from the regularity condition in Assumption 3 and η is from the step size defined in the DRO update 3

We now present our "slow rate" results on the sample complexity for the convergence of the robust policy parameter.

Theorem 1 ("Slow" Estimation error of $\theta^{\mathcal{W}_p}$). *Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$*

$$||\theta^{\mathcal{W}_p} - \hat{\theta}_n^{\mathcal{W}_p}||_2^2 \le \frac{\eta^2 K_g^2}{\lambda}\sqrt{\frac{2\log(2/\delta)}{n}}$$

where λ is from the regularity condition in Assumption 3 and $K_g = 8B/\eta + 2F$ where B is from the assumption that the policy parameter set is bounded in Assumption 2 and F is from the Assumption 1.

Proof sketch. We first prove that $\ell(z;\theta)$ is uniformly bounded and is $4K_g/\eta$-Lipschitz in θ where $K_g = 4B/\eta + 2F$ (Appendix B.1). Using this, we can prove that $\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$ is $2/\eta$-strongly convex in $||\cdot||_{\Sigma_{\mathbb{P}}}$. Intuitively taking the supremum over $\mathbb{P}$ should preserve the convex combination and the negative quadratic term and doing this analysis formally allows us to show that $\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon)$ is $2\lambda/\eta^2$-strongly convex in $||\cdot||_2$. The detailed proof for strong convexity can be found in Lemma 27.

Strong duality of Wasserstein DRO [Gao and Kleywegt, 2022] (Corollary 5) allows us to reduce the difference $|\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon) - \mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)|$ to the concentration $\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}[\ell_\Delta(z;\theta)] - \mathbb{E}_{z\sim\mathbb{P}^\circ}[\ell_\Delta(z;\theta)]\right|$ where $\ell_\Delta(z;\theta) = \inf_{z'\in\mathcal{Z}}\{\Delta d^p(z,z') - \ell(z';\theta)\}$ is the Moreau envelope of $-\ell$. We then use Hoeffding's inequality to obtain a concentration result which is uniform over $\theta \in \Theta$ and Δ. We can now do a "three-term" decomposition of $\mathcal{L}^{\mathcal{W}_p}(\theta^{\mathcal{W}_p};\varepsilon) - \mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)$ into $\mathcal{L}^{\mathcal{W}_p}(\theta^{\mathcal{W}_p};\varepsilon) - \mathcal{L}_n^{\mathcal{W}_p}(\theta^{\mathcal{W}_p};\varepsilon)$, $\mathcal{L}_n^{\mathcal{W}_p}(\theta^{\mathcal{W}_p};\varepsilon) - \mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)$, and $\mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right) - \mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)$ and bound the first and last term by Hoeffding and the second term by 0. Using strong convexity of $\mathcal{L}^{\mathcal{W}_p}$, we can get the estimation error. The detailed proof for the "slow rate" estimation error can be found at C.2 $\square$

We prove similar results for KL and χ^2 ambiguity sets using the same ideas used in the Wasserstein ambiguity set setting. We state the "slow rate" estimation rates below and defer the proofs to Appendix D and Appendix E

Theorem 2 ("Slow" Estimation error of θ^{KL})**.** *Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$*

$$\|\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}\|_2^2 \leq \frac{\eta^2 \bar{\lambda} \exp(K_g^2/\underline{\lambda})}{\lambda} \sqrt{\frac{2\log(2/\delta)}{n}}$$

where $K_g = 8B/\eta + 2F$, $\bar{\lambda}, \underline{\lambda}$ is from Assumption 4, and η is the stepsize defined in in the DRO update 3

Theorem 3 ("Slow" Estimation error of θ^{χ^2})**.** *Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$*

$$\|\theta^{\chi^2} - \hat{\theta}_n^{\chi^2}\|_2^2 \leq \frac{\eta^2 K_g^2}{\lambda} \left(1 + K_g^2/4\underline{\lambda}\right) \sqrt{\frac{2\log(4/\delta)}{n}}$$

where λ is from the regularity condition in Assumption 3 and $K_g = 8B/\eta + 2F$ where B is from the assumption that the policy parameter set is bounded in Assumption 2, F is from the Assumption 1, and η is from the step size defined in the DRO update 3.

Remark 4 (Estimation rate and improved coverage dependence)**.** *Although both WDPO Xu et al. [2025] and DRO-REBEL achieve the same $n^{-1/4}$ estimation-error rate, DRO-REBEL features substantially tighter constants thanks to its regression-to-relative-rewards formulation and cleaner strong-convexity analysis. In WDPO, the strong-convexity modulus is the product of the Bradley–Terry curvature*

$$\gamma \;=\; \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2} \quad \textit{and} \quad \lambda,$$

so that the squared-error bound scales as $O\big(1/(\gamma\,\lambda)\big)$ and is exponentially sensitive to the logistic scale β [Xu et al., 2025, Lemma 1, Theorem 1]. By contrast, Lemma 27 shows that the DRO-REBEL loss $\mathcal{L}^{\mathcal{W}_p}$ is $(2\lambda/\eta^2)$-strongly convex, yielding a bound of order $O(1/\lambda)$ (up to the step-size η) with no explicit dependence on logistic scale parameters. Concretely, this sharper constant means that for any fixed coverage λ, our excess-risk bound is smaller by the factor $\gamma^{-1} = (1+e^{4\beta B})^2/(\beta^2 e^{4\beta B})$, which can be enormous when βB is large or preferences are near-degenerate. Moreover, the same phenomenon appears in the KL-DRO setting. Theorem 2 shows

$$\big\|\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}\big\|_2^2 \;\leq\; \frac{\eta^2 \bar{\lambda} \exp(K_g^2/\underline{\lambda})}{\lambda} \sqrt{\frac{2\log(2/\delta)}{n}},$$

where $\bar{\lambda}$, $\underline{\lambda}$ bounds the dual multiplier and L bounds the loss. In prior KLDPO analyses, the dependence on $\exp(K_g^2/\underline{\lambda})$ is tangled with additional Bradley–Terry curvature terms; in DRO-REBEL it appears only through the divergence parameter. Crucially, both Wasserstein and KL results rest on the very same modelling assumptions: linear reward class (Assumption 1), log-linear policy class (Assumption 2), and data-coverage regularity (Assumption 3). This shows that DRO-REBEL's improvements arise purely from algorithmic simplicity rather than stronger distributional or curvature requirements.

With the above analysis, building on the techniques of Xu et al. [2025], we recover the same $O(n^{-1/4})$ estimation-error rate but with substantially sharper constants. Xu et al. [2025] observe that WDPO's estimation error decays at $O(n^{-1/4})$, while non-robust DPO already achieves $O(n^{-1/2})$. This slowdown arises because, in the non-robust setting, one can compute the closed-form expression for $\nabla_\theta (1/n)\sum_{i=1}^n l(z_i;\theta)$. This allows us to write $\|\nabla_\theta (1/n)\sum_{i=1}^n l(z_i;\theta^*)\|_{(\Sigma_D+\lambda I)^{-1}}$ in quadratic form and then obtain a concentration using Bernstein's inequality. Coupled with a Taylor linear approximation argument and strong convexity of the empirical DPO loss, we get the claimed $O(n^{-1/2})$. However, for the robust setting, one cannot exchange the supremum over distributions with the gradient operator on the empirical robust loss. That is, $\nabla_\theta \mathcal{L}_n^{\mathcal{W}_p}(\theta^{\mathcal{W}_p}) \neq \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \nabla_\theta \mathbb{E}_{z\sim P}[l(z;\theta^{\mathcal{W}_p})]$. Thus the approach taken in the non-robust setting will not work for the robust setting. As a result, the proof in the robust setting relies on the strong convexity of the robust DPO loss itself and a concentration bound on the loss function's convergence (obtained via Hoeffding's inequality for bounded random variables in this case). This indirect approach of bounding the loss and then translating it to a parameter bound through strong convexity results in a slower $O(n^{-1/4})$ convergence rate for the policy parameter. Closing this gap and restoring the optimal inverse square root rate for robust DPO remains an open problem. To close this gap, we develop a *localized Rademacher complexity* analysis for DRO-REBEL which recovers the optimal $O(n^{-1/2})$ convergence rate even under various ambiguity sets.

4.2 A "Master Theorem" for "Fast" Estimation Rates

To state our main guarantee in its cleanest form, we observe that nothing exotic is needed beyond (i) the population DRO objective has a uniform quadratic growth (via our data-coverage assumption), (ii) each per-sample loss $\ell(z;\theta)$ is Lipschitz in θ, which drives the Rademacher/Dudley control, and (iii) each divergence admits a simple Fenchel-duality or direct bound showing $\left|\mathcal{L}^D(\theta) - \mathbb{E}_P[\ell(z;\theta)]\right| \le O(\Delta_n)$ where $\Delta_n = O(n^{-1})$. The following single "master theorem" then automatically yields the parametric $n^{-1/2}$-rate for all of our DRO variants, Wasserstein, KL, and χ^2, by plugging in the corresponding Δ_n . Let the empirical and population minimizers be defined as:

$$\hat{\theta}_n \; = \; \arg\min_{\theta\in\Theta} \mathcal{L}_n^D\big(\theta;\varepsilon_n\big), \qquad \theta^\star \; = \; \arg\min_{\theta\in\Theta} \mathcal{L}^D\big(\theta;\varepsilon_n\big)$$

Theorem 4 (DRO-REBEL Generalization Bound)**.** *Assume the following conditions hold:*

1. ***Local Strong Convexity:*** *The population DRO loss $\mathcal{L}^D(\theta;\varepsilon_n)$ is α-strongly convex in a neighborhood of $\theta^\star$.*
2. ***Linear margin.*** *There exists a feature map $v : \mathcal{Z} \to \mathbb{R}^d$ such that*
$$h_\theta(z) = \theta^\top v(z), \qquad \|v(z)\|_2 \le 1 \quad \forall z \in \mathcal{Z}.$$
3. ***Lipschitz Loss:*** *For some $L_\phi > 0$ and all $z \in \mathcal{Z}$,*
$$\ell(z;\theta) \; = \; \phi\left(h_\theta(z), z\right), \quad \left|\phi(u,z) - \phi(u',z)\right| \le L_\phi\,|u-u'| \;\; \forall u,u' \in \mathbb{R}.$$
4. ***Dual Remainder Bound:*** *There exists a non-negative quantity Δ_n such that for any $\theta \in \Theta$:*
$$\left|\mathcal{L}^D(\theta;\varepsilon_n) - \; \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z;\theta)\big]\right| \; \le \; \Delta_n$$
and
$$\left|\mathcal{L}_n^D(\theta;\varepsilon_n) - \; \mathbb{E}_{\mathbb{P}_n^\circ}\big[\ell(z;\theta)\big]\right| \; \le \; \Delta_n$$

Then for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\big\|\hat{\theta}_n - \theta^\star\big\|_2 \; \le \; C_1 \frac{L_\phi}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \; C_2\sqrt{\frac{\Delta_n}{\alpha}},$$

where C_1, C_2 are universal numerical constants. In particular, if $\Delta_n = \mathcal{O}(n^{-1})$ then $\|\hat{\theta}_n - \theta^\star\|_2 = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$.

Proof sketch. Write $f_\theta(z) = \ell(z;\theta) - \ell(z;\theta^\star)$ and $\Delta_n(\theta) = \mathcal{L}_n^D(\theta) - R_n(\theta)$. Optimality of $\hat{\theta}_n = \arg\min_\Theta \mathcal{L}_n^D$ together with $|\Delta_n(\theta)| \le \Delta_n$ gives the basic inequality

$$\varepsilon_\ell(\hat{\theta}_n) \; = \; R(\hat{\theta}_n) - R(\theta^\star) \; \le \; (\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_n} + 2\Delta_n. \tag{B.I.}$$

Peel the parameter space into geometric "risk shells" and rescale each f_θ by the corresponding factor 4^{-k_θ}. A uniform Rademacher–complexity deviation (contraction by the L_ϕ–Lipschitz loss plus a linear-margin RC bound) yields, with probability $1-\delta$,

$$(\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_n} \; \le \; \frac{C\,L_\phi}{\sqrt{\alpha n}}\sqrt{\big(\varepsilon_\ell(\hat{\theta}_n) + \Delta_n\big)\,\log\frac{4e}{\delta}},$$

for a universal constant $C > 0$. Set $r = C\,L_\phi^2 \log(4e/\delta)/(\alpha n) + \Delta_n/2$. With this choice the right-hand side of (B.I.) is at most $r/4$, giving $\varepsilon_\ell(\hat{\theta}_n) \le 4\Delta_n + 2CL_\phi^2 \log(4e/\delta)/(\alpha n)$.

Local α-strong convexity of $\mathcal{L}^D$ implies $\|\theta - \theta^\star\|_2 \le \sqrt{2(\varepsilon_\ell(\theta) + 2\Delta_n)/\alpha}$. Plugging the bound on $\varepsilon_\ell(\hat{\theta}_n)$ produces

$$\|\hat{\theta}_n - \theta^\star\|_2 \; \le \; C_1 \frac{L_\phi}{\alpha}\sqrt{\frac{\log(4e/\delta)}{n}} \; + \; C_2\sqrt{\frac{\Delta_n}{\alpha}}$$

where C_1, C_2 are universal numerical constants. If the dual remainder decays at the parametric rate $\Delta_n = O(n^{-1})$, both terms are $O(n^{-1/2})$, completing the proof. The detailed proof for the DRO generalization bound can be found in Appendix F. □

We also show that this bound is indeed tight up to constants.

Theorem 5 (Minimax lower bound). *Fix $L_\phi > 0,\ \alpha > 0$ and a sequence $\Delta_n \geq 0$. Let $\mathcal{M}(\alpha, L_\phi, \Delta_n, \rho)$ be the admissible class of all data–generating distributions $\mathbb{P}^\circ$ on $\mathcal{Z}$ for which* every *parameter $\theta \in \Theta$ obeys assumptions (1)–(4) of Theorem 4. For $n \geq 1$ define the* minimax ℓ_2 risk

$$\mathfrak{R}_n \ := \ \inf_{\hat{\theta}_n} \sup_{(\mathbb{P}^\circ, \phi, v) \in \mathcal{M}(\alpha, L_\phi, \Delta_n, \rho)} \mathbb{E}_{\mathbb{P}^\circ}\big[\, \|\hat{\theta}_n - \theta^\star(\mathbb{P}^\circ)\|_2 \big],$$

where the infimum ranges over all (possibly randomised) estimators $\hat{\theta}_n = \hat{\theta}_n(Z_1, \ldots, Z_n)$. Then for two universal constants $c_1, c_2 > 0$ independent of $(n, L_\phi, \alpha, \Delta_n)$,

$$\mathfrak{R}_n \ \geq \ c_1 \, \frac{L_\phi}{\alpha} \, \frac{1}{\sqrt{n}} \ \vee \ c_2 \sqrt{\frac{\Delta_n}{\alpha}}$$

for all $n \geq 1$.

Proof sketch. The proof proceeds by establishing two distinct lower bounds that arise from different sources of error. The overall minimax lower bound is the maximum of these two.

We first establish a lower bound for the standard (non-robust) M-estimation problem, which serves as a lower bound for the DRO problem. Using a parameter separation construction, we define two hypotheses: a null hypothesis with parameter $\theta_0 = 0$ and an alternative with parameter $\theta_1 = \delta_n u$. By constructing a simple data-generating process with Gaussian noise, we show that for an appropriate choice of separation $\delta_n = \Omega(n^{-1/2})$, the Kullback-Leibler divergence between the two distributions is small and constant. Invoking Le Cam's two-point lemma, we then establish an irreducible statistical error of order $\Omega(n^{-1/2})$, which no estimator can overcome.

The other bound arises from the ambiguity inherent in the DRO formulation itself. We construct a worst-case scenario by considering a data-generating distribution that is a point mass. We then define two admissible DRO objectives, $\mathcal{L}^+$ and $\mathcal{L}^-$, using a Huberized quadratic loss function. By carefully choosing the objectives to be small perturbations of the original risk, we ensure they are statistically indistinguishable (i.e., their total variation distance is zero). However, the minimizers of these two objectives, θ_+^* and θ_-^*, are separated by a distance $\delta \geq \Omega(\sqrt{\Delta_n/\alpha})$. Le Cam's lemma then implies that since the two problems are indistinguishable, any estimator will have an irreducible error related to the separation between their true minimizers. This establishes an irreducible residual error of order $\Omega(\sqrt{\Delta_n/\alpha})$. The detailed proof for the minimax rate can be found in Appendix G. □

4.3 DRO-REBEL "Fast Rates"

Using this theorem, we can now state the following "fast rate" REBEL estimation results for Wasserstein, KL, and χ^2. We defer the proofs to Appendix H.1, H.2, and H.3 respectively.

Corollary 1 ("Fast" Estimation error of $\theta^{\mathcal{W}_p}$). *Let $\delta \in (0,1)$ and choose $\varepsilon_n \asymp n^{-1}$. Then, with probability at least $1 - \delta$*

$$\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2K_g^2\eta^2}{\lambda^2}\log(1/\delta) + \frac{8(B+F)(B+1)}{\lambda}\right)}$$

where $L_{\ell,z}$ is from Assumption 5, λ is from the regularity condition in Assumption 3, and $K_g = 8B/\eta + 2F$ where B is from the assumption that the policy parameter set is bounded in Assumption 2, and F is from the Assumption 1.

Corollary 2 ("Fast" Estimation error of θ^{KL}). *Let $\delta \in (0,1)$ and choose $\varepsilon_n \asymp n^{-2}$. Then, with probability at least $1 - \delta$*

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2K_g^2\eta^2}{\lambda^2}\log(1/\delta) + \frac{\eta^2 K_g^2}{\lambda}\right)}$$

where λ is from the regularity condition in Assumption 3, and $K_g = 8B/\eta + 2F$ where B is from the assumption that the policy parameter set is bounded in Assumption 2, and F is from the Assumption 1.

Corollary 3 ("Fast" Estimation error of θ^{χ^2}). *Let $\delta \in (0,1)$ and choose $\varepsilon_n \asymp n^{-2}$. Then, with probability at least $1 - \delta$*

$$\|\hat{\theta}_n^{\chi^2} - \theta^{\chi^2}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2K_g^2\eta^2}{\lambda^2}\log(1/\delta) + \frac{\eta^2 K_g^2}{\lambda}\right)}$$

where λ is from the regularity condition in Assumption 3, and $K_g = 8B/\eta + 2F$ where B is from the assumption that the policy parameter set is bounded in Assumption 2, and F is from the Assumption 1.

As one may have noticed in the KL-REBEL and χ^2-REBEL "fast" rates, the rates are identical. This is because for these two particular f-divergences, their dual remainder bound scales at the same rate i.e. $\Delta_n \lesssim K_\ell/n$ where $\varepsilon_n = O(n^{-2})$. Noticing this, we were interested to see if a general result could be proven regarding all the "fast" rate of all f-divergences. It turns out that the answer is yes. We refer readers to Appendix H.4 for the detailed proof.

Theorem 6 (Fast Rate for f-Divergence Robust Optimization)**.** *Let $\ell(z;\theta)$ be a pointwise loss function, and $\mathbb{P}^\circ$ be the nominal data-generating distribution. Let $\mathcal{L}^f(\theta)$ be the robust loss function defined by:*

$$\mathcal{L}^f(\theta) = \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_f)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$$

where $\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_f) = \{\mathbb{P} \mid D_f(\mathbb{P},\mathbb{P}^\circ) \le \varepsilon_n\}$ is the ambiguity set defined by the f-divergence $D_f(\mathbb{P}||\mathbb{P}^\circ)$.

Assume the following conditions hold:

1. ***Bounded Loss Function (Assumption 6)**: The loss function $\ell(z;\theta)$ is bounded, i.e., $m \le \ell(z;\theta) \le M$ almost surely for some constants $m, M \in \mathbb{R}$. Let $K_\ell = M - m$ be its range.*
2. ***f-Divergence Properties (Assumption 7)**: The f-function $f : [0,\infty) \to \mathbb{R}\cup\{\infty\}$ is convex, satisfies $f(1) = 0$, is twice continuously differentiable at $t = 1$, and its second derivative $f''(1)$ is strictly positive.*

Let $C_f = \frac{1}{f''(1)}$.

Then, for sufficiently small $\varepsilon_n > 0$, the difference between the robust loss and the nominal expected loss is bounded by:

$$\left|\mathcal{L}^f(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z;\theta)\big]\right| \le \frac{K_\ell\sqrt{C_f}}{\sqrt{2}}\sqrt{\varepsilon_n} + o(\sqrt{\varepsilon_n})$$

More concisely, for sufficiently small ε_n:

$$\left|\mathcal{L}^f(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z;\theta)\big]\right| \lesssim \frac{K_\ell\sqrt{C_f}}{\sqrt{2}}\sqrt{\varepsilon_n}$$

To achieve a rate of $\Delta_n = O(n^{-1})$ for this difference, the radius of the f-divergence ball must be set as $\varepsilon_n \asymp n^{-2}$.

Choice of ambiguity radii. For Wasserstein balls, we choose $\varepsilon_n = cn^{-1}$. The empirical convergence rate of empirical measures depends on the relationship between the chosen order p and the intrinsic support dimension m, as well as the measure's regularity. Fournier and Guillin [2015] and Weed and Bach [2019] show that for a measure μ with intrinsic dimension m:

$$\varepsilon_n \;=\; c\begin{cases} n^{-1/2p}, & \text{if } m \le 2p \text{ or } \mu \text{ is approximately singular,} \\ n^{-1/m}, & \text{otherwise (i.e., } m > 2p \text{ and not approximately singular).}\end{cases} \qquad c > 0.$$

Under KL or χ^2 constraints we use $\varepsilon_n = c\,n^{-2}$, a rate faster than the Chernoff two-sample rate $D_{\mathrm{KL}}(\widehat{\mathbb{P}}_n\|\mathbb{P}^\circ) = O_{\mathbb{P}}(n^{-1})$ obtained from Sanov's theorem [Dembo and Zeitouni, 1998, Cover and Thomas, 2006]; the same scaling translates to χ^2 via Pinsker-type inequalities. For total variation we choose $\varepsilon_n = c\,n^{-1}$, a rate faster than the Dvoretzky–Kiefer–Wolfowitz fluctuation $\|\widehat{\mathbb{P}}_n - \mathbb{P}^\circ\|_{\mathrm{TV}} = O_{\mathbb{P}}(n^{-1/2})$ [Massart, 1990].

It should be noted that all the chosen radii shrink much faster than the empirical concentration rates of the divergences. They also are significantly faster than the $O(n^{-1/2})$ rate which is the standard statistical error rate for parameter estimation in strongly convex problems. When the ambiguity radius diminishes at such a rapid rate, the distributional robustness has a negligible effect on the asymptotic behavior of the estimator. In this scenario, the DRO solution essentially behaves like the solution from standard Empirical Risk Minimization (ERM). For strongly convex problems, ERM is known to achieve the optimal statistical rate of $O(n^{-1/2})$. This means that our choice of radii will likely lead to a convergence rate for the parameter estimates that is close to the $O(n^{-1/2})$ rate, largely due to the dominance of the empirical risk term rather than the robustification. This also implies that by using these radii, $\Pr\left(\mathbb{P} \notin \mathcal{B}_{\varepsilon_n}\left(\mathbb{P}^\circ; D\right)\right) \not\to 0$ as $n \to \infty$ so we cannot even guarantee coverage of the nominal data-generating distribution. To ensure a high level of confidence that the true distribution lies within the ambiguity set and provide meaningful distributional robustness (beyond just the true distribution being *within* the set, but rather that the set is *large enough* to capture a meaningful neighborhood), the radius ε_n would typically need to shrink at a rate *slower* than $O(n^{-1/2})$. In fact, these are the statistical limits of DRO [Blanchet and Shapiro, 2023]. That is, there is no "free lunch" in simultaneously achieving strong distributional robustness and the fastest possible statistical convergence. This is also evident from our result (Theorem 5) as the faster your ambiguity radius shrinks (i.e., the less robust you are), the faster your statistical rate of convergence can be at the cost of coverage guarantees, and vice versa. One will notice that if we indeed choose the empirical concentration rates, we obtain $O_{\mathbb{P}}(n^{-1/4})$ rates for all robust estimators. **We hypothesize that if you want to guarantee sufficient coverage of the nominal data-generating, the best rate you can do is** $O_{\mathbb{P}}(n^{-1/4})$

It is unrealistic to choose $\varepsilon_n \asymp O(n^{-1})$ for Wasserstein Robust LLM alignment. Recent empirical studies suggest that although large-parameter language models lie in billion-dimensional ambient spaces, the *functionals* relevant to policy optimization, such as gradients, log-probability gaps, and reward differences, live on manifolds of much lower intrinsic dimension. Gur and Shamir [2022] formalize the *local intrinsic dimension* (LID) of a representation and prove that, for ReLU networks, the LID of each hidden layer is bounded by the number of active neurons, often two orders of magnitude below width. Measuring LID on ImageNet features and BERT embeddings, they report dimensions in the range 30–150 despite embedding sizes of 512–2048. Complementarily, Jin et al. [2023] compress SGD trajectories of GPT-style and ViT models, showing that $> 99\%$ of gradient variance concentrates in subspaces of dimension 50–150 out of millions. These findings imply that while LLM representations can live in high-dimensional spaces, the distribution of the specific feature–gap vector $(\Delta\psi, \Delta r, y)$ in the context of Wasserstein distance might be "approximately singular" or have a very low effective intrinsic dimension. However, it is often the case that the full gap-vector still has effective $m \gg p$. Hence in this regime, the best you can do is $\varepsilon_n \asymp n^{-1/m}$ which results in a much slower rate and encounters the curse of dimensionality to some degree. A more robust, *dimension-free* alternative is to define the ambiguity set via either Gaussian-smoothed Wasserstein $\mathcal{W}_p^{(\sigma)}(\mathbb{P}, \mathbb{P}^\circ) \le \varepsilon$ or Entropic (Sinkhorn) Wasserstein $\mathcal{W}_{2,\tau}(\mathbb{P}, \mathbb{P}^\circ) \le \varepsilon$. Both satisfy $\mathcal{W}\big(\widehat{\mathbb{P}}_n, \mathbb{P}^\circ\big) = O_{\mathbb{P}}(n^{-1/2})$ in any ambient dimension, so choosing $\varepsilon_n \asymp n^{-1/2}$ guarantees coverage. Thus in practical RLHF setting with high-dimensional data, the practical choice is $\varepsilon_n \asymp n^{-1/2}$, trading the classical $n^{-1/d}$ curse for a universal $n^{-1/2}$ coverage rate and an ultimately achievable $n^{-1/4}$ estimation rate.

4.4 Distributionally Robust DPO "Fast Rates"

Crucially, the same localized-complexity machinery could be applied to the WDPO and KLDPO analyses of Xu et al. [2025] to upgrade their $n^{-1/4}$ rates to $n^{-1/2}$—albeit with larger constants, since REBEL's relative-reward regression loss is fundamentally simpler and has smaller Lipschitz and curvature parameters. We significantly advance the theoretical understanding of robust DPO, specifically for Wasserstein DPO (WDPO) and KL DPO (KLDPO) and close a critical gap by establishing a superior rate of convergence compared to previous analyses, including that provided by Xu et al. [2025], with the following theorems:

Theorem 7 (Parametric $O(n^{-1/2})$ Rate for Wasserstein DPO)**.** *Let the Wasserstein DPO empirical and population minimizers be defined as:*

$$\hat{\theta}_n^{\mathcal{W}_p} \;=\; \arg\min_{\theta\in\Theta} \mathcal{L}_n^{\mathcal{W}_p}\big(\theta; \varepsilon_n\big), \qquad \theta^{\mathcal{W}_p} \;=\; \arg\min_{\theta\in\Theta} \mathcal{L}^{\mathcal{W}_p}\big(\theta; \varepsilon_n\big)$$

where $\mathcal{L}^{\mathcal{W}_p}$ is the Wasserstein distributionally robust DPO objective with ambiguity radius $\varepsilon_n \asymp n^{-1}$. Assume all the assumptions made in the previous analyses. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the estimation error of the Wasserstein DPO estimator is bounded by:

$$\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{6\beta B}{\gamma\lambda}\right)}$$

where λ is from the regularity condition in Assumption 3, $\gamma = \beta^2 e^{4\beta B}/(1+e^{4\beta B})^2$ is the DPO curvature constant and $L_{\ell,z}$ is from Assumption 5.

Theorem 8 (Parametric $O(n^{-1/2})$ Rate for KL-DPO)**.** *Let the KL-DPO empirical and population minimizers be defined as:*

$$\hat{\theta}_n^{\mathrm{KL}} \;=\; \arg\min_{\theta\in\Theta} \mathcal{L}_n^{\mathrm{KL}}\big(\theta; \varepsilon_n\big), \qquad \theta^{\mathrm{KL}} \;=\; \arg\min_{\theta\in\Theta} \mathcal{L}^{\mathrm{KL}}\big(\theta; \varepsilon_n\big)$$

where $\mathcal{L}^{\mathrm{KL}}$ is the KL-divergence distributionally robust DPO objective with ambiguity radius $\varepsilon_n \asymp n^{-2}$. Assume all the assumptions made in the previous analyses. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the estimation error of the KL-DPO estimator is bounded by:

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

where λ is from the regularity condition in Assumption 3 and $\gamma = \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$

We defer readers to the Appendix for the proofs, particularly Appendix I.1 and I.2.

DPO/REBEL	Linear margin	Local SC	Coverage ε_n	Fast-rate ε_n	Dim-free cov.?	Fast rate & coverage?
χ^2/KL Divergence	✓	✓	n^{-1}	n^{-2}	✓	✗
$\mathcal{W}_p$	✓	✓	$n^{-1/d}$	n^{-1}	✗	✗
$\mathcal{W}_{p,\sigma}$; $\mathcal{W}_{p,\tau}$	✓	✓	$n^{-1/2}$	n^{-1}	✓	✗
TV Distance	✓	✓	$n^{-1/2}$	n^{-1}	✓	✗

Table 1: "Coverage ε_n" is the radius schedule that maintains non-vanishing coverage ($\Pr(\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D)) \to 1$). "Fast-rate ε_n" is the schedule that yields a dual remainder $\Delta_n = O(n^{-1})$ and hence $O(n^{-1/2})$ parameter error via the master theorem. "Dim-free cov." indicates whether the coverage rate avoids ambient-dimension dependence.

5 Related Work and Comparisons

Our work on Distributionally Robust REBEL (DRO-REBEL) contributes to the growing body of literature on robust reinforcement learning from human feedback (RLHF), particularly concerning out-of-distribution (OOD) generalization and sample efficiency. We find strong theoretical alignment with recent advancements in χ^2-based preference learning and offer distinct advantages in sample complexity compared to other distributionally robust RLHF approaches.

5.1 χ^2-Based Preference Learning

Very recently, Huang et al. [2025] introduced χPO, a one-line modification of Direct Preference Optimization (DPO) that replaces the usual log-link with a mixed χ^2 + KL link and implicitly enforces pessimism via the χ^2-divergence. Their main result shows that χPO achieves a sample-complexity guarantee scaling as

$$J(\pi^*) - J(\widehat{\pi}) \lesssim \sqrt{\frac{C_{\pi^*} \log(|\Pi|/\delta)}{n}}$$

where $J(\pi) = \mathbb{E}_{x\sim\rho,\, a\sim\pi(\cdot|x)}\left[r_*(x,a)\right]$ is the true expected reward and $C_{\pi^*} = 1 + 2D_{\chi^2}(\pi^* \| \pi_{\mathrm{ref}})$ is the single-policy concentrability coefficient (cf. Theorem 3.1 of Huang et al. [2025]). We want to compare these results to those derived in our "fast rate" analysis. To proceed with our analysis, we will require the following lemma

Lemma 2 (One-Step Performance Difference Lemma). *Let $J(\pi) = \mathbb{E}_{x\sim\rho,\, a\sim\pi(\cdot|x)}\left[r_*(x,a)\right]$ where $r_* \in \mathcal{F}$. Define the one-step baseline as $B^\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)}\left[r_*(x,a)\right]$ and one-step advantage as $A^\pi(x,a) = r_*(x,a) - B^\pi(x)$. Then we have the following one-step performance difference:*

$$J(\pi') - J(\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right]$$

Additionally, under the assumption that $V_{\max} = \sup_{x,a}|A^\pi(x,a)|$, we also have

$$|J(\pi') - J(\pi)| \le V_{\max}\mathbb{E}_{x\sim\rho}\left[\|\pi'(\cdot \mid x) - \pi(\cdot \mid x)\|_1\right]$$

We also require the following result

Lemma 3 (Log-Linear Policies are Lipschitz). *Suppose $\pi_\theta \in \Pi$ are in a log-linear policy class as defined in Assumption 2. Then all policies in such a class are 2-Lipschitz in θ.*

Combining these results, we get the following sample complexity result

Theorem 9 (Sample Complexity Result for χ^2-DPO). *Suppose Assumption 2 and 3 hold. With probability at least $1-\delta$, χDPO produces a policy $\hat{\pi}$ such that simultaneously for all $\pi^* \in \Pi$, we have*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max}\sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

where $V_{\max} = \sup_{x,a} A^\pi(x,a) = 2F$.

We also require the following result

Lemma 4 (Log-Linear Policies are Lipschitz). *Suppose $\pi_\theta \in \Pi$ are in a log-linear policy class as defined in Assumption 2. Then all policies in such a class are 2-Lipschitz in θ.*

Combining these results, we get the following sample complexity result

Theorem 10 (Sample Complexity Result for χ^2-DPO). *Suppose Assumption 2 and 3 hold. With probability at least $1-\delta$, χDPO produces a policy $\widehat{\pi}$ such that simultaneously for all $\pi^* \in \Pi$, we have*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max}\sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

where $V_{\max} = \sup_{x,a} A^{\pi}(x,a) = 2F$.

We refer readers to Appendix J for the detailed proofs. Our "fast rate" analysis for the χDPO recovers the same $O(n^{-1/2})$ parametric rate under analogous linear-policy, data-coverage, and convexity assumptions. In both cases the key is that χ^2-regularization induces a heavy-tailed density-ratio barrier and uniform quadratic growth, allowing a localized Rademacher complexity argument to restore the minimax $n^{-1/2}$ rate. Thus, the theoretical insights of Huang et al. [2025] on the power of χ^2–divergence to suppress overoptimization are fully consistent with—and indeed validated by—our fast-rate guarantees for χ^2-REBEL as a sample-efficient and stable optimizer under preference noise.

5.2 Comparison with Distributionally Robust RLHF (Mandal et al., 2025)

Our work on DRO-REBEL directly addresses the robust policy optimization problem by minimizing a distributionally robust REBEL objective, assuming a fixed reward model. A parallel effort is made by Mandal et al. [2025], who also tackle robust policy optimization within the RLHF context. A key distinction lies in the choice of ambiguity sets and the resulting sample complexity guarantees for the robust policy. Mandal et al. [2025] investigate two main algorithms for robust policy optimization, both primarily using Total Variation (TV) distance for their ambiguity sets. We will focus on their robust DPO algorithm (cf. Theorem 3 of Mandal et al. [2025]). We state their result for convenience

Theorem 11 (Theorem 3 of Mandal et al. [2025]; Convergence of Robust DPO). *Suppose Assumption 2 holds. Run DR-DPO for $T = \mathcal{O}\big(\frac{1}{\varepsilon^2}\big)$ iterations, and choose the minibatch size n so that*

$$\frac{n}{\log n} \;\geq\; \mathcal{O}\bigg(\frac{\beta^2(2B+J)^2(1+2\rho)^2}{\varepsilon^2}\bigg)$$

where

$$J \;=\; \max_{x,a^1,a^2}\log\frac{\pi_{\mathrm{ref}}(a^2\mid x)}{\pi_{\mathrm{ref}}(a^1\mid x)}.$$

Then the average iterate

$$\bar{\theta} \;=\; \frac{1}{T}\sum_{t=1}^{T}\theta_t$$

satisfies

$$\mathbb{E}\big[\ell_{\mathrm{DPO,TV}}(\bar{\theta}\,;D_{\mathrm{src}})\big] \;-\; \min_{\theta}\,\ell_{\mathrm{DPO,TV}}(\theta\,;D_{\mathrm{src}}) \;\leq\; \mathcal{O}(\varepsilon).$$

Now let us assume that $\ell_{\mathrm{DPO,TV}}$ is μ-strongly convex in θ. This assumption is reasonable as we have shown in Lemma 27, 29, and 31 that under various ambiguity sets, strong convexity of the robust variant holds. Then by Lemma 9, we have

$$\|\bar{\theta}-\theta^*\|_2^2 \leq \frac{2}{\mu}\left(\mathbb{E}\big[\ell_{\mathrm{DPO,TV}}(\bar{\theta}\,;D_{\mathrm{src}})\big] \;-\; \min_{\theta}\,\ell_{\mathrm{DPO,TV}}(\theta\,;D_{\mathrm{src}})\right)$$

Taking $\varepsilon \gtrsim \beta(2B+J)(1+2\rho)\sqrt{\frac{\log n}{n}}$. Thus we find that

$$\|\bar{\theta}-\theta^*\|_2 \lesssim \left(\frac{\log n}{n}\right)^{-1/4}$$

We now prove the TV-distance DPO "fast" rate. Unfortunately since our general f-divergence result (Theorem 6) relies on sufficient smoothness which is not the case with the TV-distance metric, we need to consider TV-distance as a separate case. This yields a rate $\Delta_n \lesssim K_l\varepsilon_n$ where $\varepsilon_n \asymp n^{-1}$, meaning as we collect more samples from $\mathbb{P}^{\circ}$, our ambiguity set distance shrinks at a slower rate compared to other f-divergences. We refer readers to Appendix H.5 for the detailed proof. Nevertheless, using this bound, we find that

$$\|\bar{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

Our "fast rate" analysis for DPO provides $O(n^{-1/2})$ sample complexity guarantees for the parameter estimation error $\|\hat{\theta}_n - \theta^\star\|_2$ across various ambiguity sets (Theorems 7 and 8) under appropriate choices of the ball radius ε_n (e.g., $\varepsilon_n = O(n^{-1})$ for Wasserstein and TV and $O(n^{-2})$ for KL and other sufficiently smooth f-divergences).This implies that to achieve an estimation error of $O(\varepsilon)$, our method requires a total sample size of $n = O(1/\varepsilon^2)$. This also applies to our analysis of robust REBEL with the added benefits of much sharper constants (Corollary 1, 2, and 3). Our derived "fast rates" are significantly more efficient in terms of sample complexity compared to the results presented by Mandal et al. [2025] for their TV-distance based robust policy optimization algorithms as our rates align with the minimax optimal rates often observed in parametric statistical problems.

Algorithm / Work	**Parameter**	**Ambiguity Set(s)**	**Sample Complexity (n)**	**Error Rate (ε)**
Our Robust DPO	θ	$\mathcal{W}_p$, KL	$O(n^{-1/2})$	$O(1/\varepsilon^2)$
Mandal et al. (Robust DPO)	θ	TV	$O(n^{-1/4})$	$O(\log(1/\varepsilon)/\varepsilon^4)$
Xu et al. (Robust DPO)	θ	$\mathcal{W}_p$, KL	$O(n^{-1/4})$	$O(1/\varepsilon^4)$

Table 2: Comparison of theoretical convergence rates and sample complexities for robust DPO algorithms.

6 Approximate Tractable Algorithms for Robust LLM Alignment

While our Distributionally Robust REBEL (DRO-REBEL) formulations benefit from finite-sample guarantees which are minimax optimal, directly solving the minimax objective using stochastic gradient descent methods can be computationally challenging. As Xu et al. [2025] also point out in the context of robust DPO, this challenge arises because we do not have direct control over the data distribution $\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ; D)$ within the uncertainty set, as it is not parameterized in a straightforward manner. Furthermore, the preference data are generated according to the nominal distribution $\mathbb{P}^\circ$, meaning we lack samples from any other distributions within the uncertainty set $\mathcal{B}_\varepsilon(\mathbb{P}^\circ; D)$. To overcome this, we introduce principled tractable algorithms that approximate the solution to our DRO-REBEL objectives. Our algorithms for solving Wasserstein-DRO-REBEL and KL-DRO-REBEL are largely the same as those of Xu et al. [2025]'s WDPO and KLDPO. However, we will derive and propose an algorithm for χ^2-DRO-REBEL that can efficiently be solved using stochastic gradient descent methods.

6.1 Tractable Wasserstein DRO-REBEL (WD-REBEL)

The connection between Wasserstein distributionally robust optimization (DRO) and regularization has been established previously in the literature, see Shafieezadeh-Abadeh et al. [2019] for example. We leverage recent progress in Wasserstein theory on connecting Wasserstein DRO to regularization. For p-Wasserstein DRO, $p \in (1, \infty]$, Gao and Kleywegt [2022] (Theorem 1) shows that for a broad class of loss functions (potentially non-convex and non-smooth), with high probability, Wasserstein DRO is asymptotically equivalent to a variation regularization. In particular, an immediate consequence is that, when $p = 2$:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}_n^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)] = \min_{\theta\in\Theta}\left\{\mathbb{E}_{z\sim\mathbb{P}_n^\circ}[\ell(z;\theta)] + \varepsilon_n\sqrt{(1/n)\sum_{i=1}^n \|\nabla_z \ell(z_i;\theta)\|_2^2}\right\} + O_{\mathbb{P}}(1/n),$$

where $\varepsilon_n = O(1/\sqrt{n})$. This indicates that one can approximately solve the Wasserstein DRO objective by adding a gradient regularization term to the empirical risk minimization (ERM) loss, $\mathbb{E}_{z\sim\mathbb{P}_n^\circ}[\ell(z;\theta)]$. Based on this, we propose a tractable WD-REBEL algorithm in Algorithm 2.

Algorithm 2 WD-REBEL Algorithm

Require: Dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ (e.g., preference pairs), reference policy π_{ref}, robustness hyperparameter ρ_0, learning rate $\tilde{\eta}$, initial policy π_θ.

while θ has not converged **do**
 Calculate the non-robust REBEL loss $\ell(z_i;\theta)$ for each $z_i \in \mathcal{D}$.
 Calculate the non-robust empirical REBEL loss $L_{\text{REBEL}}(\pi_\theta;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell(z_i;\theta)$.
 Calculate the gradient regularizer term: $R(\pi_\theta;\mathcal{D}) = \rho_0 \left(\frac{1}{n}\sum_{i=1}^n \|\nabla_{z_i}\ell(z_i;\theta)\|_2^2\right)^{1/2}$.
 Calculate the approximate WD-REBEL loss: $L_W(\theta,\rho_0) = L_{\text{REBEL}}(\pi_\theta;\mathcal{D}) + R(\pi_\theta;\mathcal{D})$.
 $\theta \leftarrow \theta - \tilde{\eta}\nabla_\theta L_W(\theta,\rho_0)$.
end while
return π_θ.

6.2 Tractable KL-DRO-REBEL (KL-REBEL)

We utilize the following proposition established by Xu et al. [2025] to show that we can approximate the worst-case probability distribution in a KL uncertainty set with respect to a given loss function.

Proposition 1 (Worst-case distribution)**.** *Let $\mathbb{P} \in \mathbb{R}^n$ be the worst-case distribution with respect to a loss function ℓ and KL uncertainty around the empirical distribution $\mathbb{P}_n^\circ$, defined as $\mathbb{P} = \sup_{\mathbb{P}: D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}_n^\circ)\le\rho} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$. The worst-case distribution $\mathbb{P}$ is related to $\mathbb{P}_n^\circ$ through*

$$\mathbb{P}(i) \propto \mathbb{P}_n(i)\cdot\exp\left(\frac{1}{\tau}(\ell(z_i;\theta) - \sum_{j=1}^n \mathbb{P}_n(j)\ell(z_j;\theta))\right),$$

where $\tau > 0$ is some constant.

A proof of this proposition can be found in Appendix D of Xu et al. [2025]. Based on Proposition 1, we propose a tractable KL-REBEL algorithm in Algorithm 3.

Algorithm 3 KL-REBEL Algorithm

Require: Dataset $\mathcal{D} = \{z_i\}_{i=1}^n$, reference policy π_{ref}, robustness temperature parameter τ, learning rate $\tilde{\eta}$, initial policy π_θ.

while θ has not converged **do**
 Calculate the non-robust REBEL loss $\ell(z_i;\theta)$ for each $z_i \in \mathcal{D}$.
 Approximate the worst-case weights assuming $\mathbb{P}_n(i) = 1/n$: $\tilde{\mathbb{P}}(i) = \exp\left(\frac{1}{\tau}\left(\ell(z_i;\theta) - \frac{1}{n}\sum_{j=1}^n \ell(z_j;\theta)\right)\right)$.
 Normalize the weights: $\mathbb{P}(i) = \frac{\tilde{P}(i)}{\sum_{k=1}^n \tilde{P}(k)}$.
 Calculate the approximate KL-REBEL loss: $L_{\mathrm{KL}}(\theta;\mathcal{D}) = \sum_{i=1}^n \mathbb{P}(i)\cdot\ell(z_i;\theta)$.
 $\theta \leftarrow \theta - \tilde{\eta}\nabla_\theta L_{\mathrm{KL}}(\theta,\rho)$.
end while
return π_θ.

6.3 Tractable χ^2-DRO-REBEL (χ^2-REBEL)

We exploit the dual formulation of the χ^2-DRO objective (e.g. Namkoong and Duchi [2017a]) to obtain a one-dimensional inner solve and closed-form worst-case weights.

Proposition 2 (Dual form & worst-case weights)**.** *Let $\ell_i = \ell(z_i;\theta)$ for $i = 1,\ldots,n$ and set $\mathcal{L}_n^{\chi^2}(\theta;\varepsilon_n) = \sup_{\mathbb{P}: D_{\chi^2}(\mathbb{P}\|\mathbb{P}_n^\circ)\le\varepsilon_n} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$. Then*

$$\mathcal{L}_n^{\chi^2}(\theta;\varepsilon_n) = \inf_{\eta\in\mathbb{R}}\left\{\eta + \sqrt{\frac{2\varepsilon_n}{n}\sum_{i=1}^n(\ell_i-\eta)_+^2}\right\},$$

We defer the proof to Appendix K. Based on Proposition 2, we propose a tractable χ^2-REBEL Algorithm in Algorithm 4

Algorithm 4 χ^2-REBEL Algorithm

Require: Dataset $\mathcal{D} = \{z_i\}_{i=1}^n$, robustness radius ρ, learning rate α, initial policy π_θ

1: **while** θ not converged **do**

2: Calculate the non-robust REBEL loss $\ell(z_i;\theta)$ for each $z_i \in \mathcal{D}$

3: *(Inner 1-D solve)* find

$$\eta^* \;=\; \arg\min_{\eta\in\mathbb{R}} \left\{ \eta + \sqrt{\frac{2\rho}{n}\sum_{i=1}^n (\ell_i - \eta)_+^2} \right\}$$

via sorting $\{\ell_i\}$ and binary search

4: Compute the approximate χ^2-REBEL loss: $L_{\mathrm{CHI}}(\theta;\rho) = \eta^* + \sqrt{\frac{2\rho}{n}\sum_{i=1}^n(\ell_i - \eta^*)_+^2}$

5: $\theta \leftarrow \theta - \alpha\, \nabla_\theta L_{\mathrm{CHI}}(\theta;\rho)$

6: **end while**

7: **return** π_θ

Each outer step of Algorithm 4 has a computational cost of $O(n\log n + nd)$. We refer readers to Appendix K for a detailed derivation of this algorithm. It should be noted that this algorithm is not new and was proposed by Namkoong and Duchi [2017b].

With the inner problem solved tractably, we consider the outer optimization over θ. Since the pointwise loss $\ell(z;\theta)$ is uniformly bounded and Lipschitz continuous (Appendix B.1), and the overall robust objective $\mathcal{L}_n^{\chi^2}(\theta;\rho)$ is strongly convex (Appendix E), we can guarantee that standard optimization algorithms (e.g., stochastic gradient descent, projected gradient descent, etc) will converge to the unique optimal policy parameters θ^*. This guarantee of convergence also holds for the Wasserstein and KL-divergence counterparts of our algorithm.

7 Experiments

We conduct a comprehensive experimental evaluation of DRO-REBEL, comparing its performance against non-robust DPO and REBEL baselines, as well as established Distributionally Robust Optimization (DRO) variants, namely WDPO and KLDPO Xu et al. [2025]. Our empirical evaluations spans two distinct alignment tasks, designed to investigate model performance under varying dataset scales, model sizes, the complexity of the reward function, and degrees of distributional shift: (i) a radius–coverage/convergence study in a Gaussian–mixture simulator that trains a log-linear policy under group-level χ^2-DRO and calibrates ε_n via Pearson/Wilson–Hilferty; (ii) an Emotion Alignment task with controlled, synthetic shifts; and (iii) an ArmoRM Multi-objective Alignment Wang et al. [2024a] task featuring large-scale, real-world shifts. The first experiment isolates statistical effects to visualize the trade-off between coverage and estimation error as n varies. The last two are the same experiments conducted by Xu et al. [2025], and we will use these as a baseline to compare the performance of DRO variants of REBEL and DPO in the face of distribution shift. Code and full hyperparameters can be found at `https://github.com/sharansahu/distributionally_robust_rebel`.

7.1 Experimental Setup

7.1.1 Radius–Coverage Trade-off (Synthetic, Log–Linear Policy)

Our goal in this experiment is to illustrate the theoretical trade-off between ambiguity-set *coverage* and *estimation rate*. We therefore use the canonical log–linear policy class assumed in our analysis,

$$\pi_\theta(a \mid x) \;\propto\; \exp\big(\theta^\top \psi(x,a)\big),$$

and operate on the linear margin $h_\theta(z) = \theta^\top v(z)$, where $v(z)$ is a bounded feature map. In this setting, the DRO objectives (Wasserstein/f–divergence/χ^2) depend only on Lipschitz properties of h_θ and local strong convexity. Consequently, we adopt the standard squared-loss surrogate on h_θ to obtain a strongly convex, mixture-robust objective with a χ^2 ambiguity set over group proportions. Concretely, we synthesize $K{=}15$ latent "preference groups" in $d{=}12$ dimensions with low-rank group means and heteroskedastic noise, draw training data by first sampling a group label then features/targets, and train by

$$\min_\theta \max_{q:\, D_{\chi^2}(q\|\widehat{p})\le\varepsilon_n} \sum_{k=1}^K q_k\, \mathbb{E}\big[(\langle v,\theta\rangle - t)^2 \mid C{=}k\big]\,,$$

where $\widehat{p}$ are empirical group proportions. We sweep calibrated radii $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ (for several α) and a fast-shrinking $\varepsilon_n \propto n^{-2}$. We then report (i) coverage $\Pr\{\widehat{p} \in \mathcal{B}_{\varepsilon_n}(p^\circ)\}$ via multinomial resampling, (ii) parameter error $\|\hat{\theta} - \theta^\star\|_2$ (known in simulation), and (iii) excess worst-case risk against χ^2 mixture shifts around the *true* mixture p°.

Why not a full LLM? Measuring the parameter-estimation rate $\|\hat{\theta} - \theta^\star\|_2$ and its dependence on ε_n *requires* access to the ground truth parameter $\theta^\star$ and clean control of the feature map $v(\cdot)$ and mixture shift. In a real LLM: (a) $\theta^\star$ is unknown (tens of billions of parameters), so any "rate" must be proxied by task scores, confounding optimization error, decoding noise, and reward misspecification with statistical error; (b) coverage of an ambiguity set cannot be verified because the data-generating mixture and its effective dimensionality are unknown; (c) practical fine-tuning introduces additional heuristics (value baselines, clipping, sampling temperature, truncation) that dominate the signal of interest; and (d) compute/variance constraints preclude the large-n sweeps needed to distinguish $n^{-1/2}$ from $n^{-1/4}$. Our controlled log–linear simulator matches the assumptions used in the theory, exactly exposes $\theta^\star$, and keeps the ambiguity set on the $(K-1)$–simplex (dimension-free w.r.t. d), thereby providing a *clear* demonstration of the radius–coverage–rate trade-off that the theory predicts.

7.1.2 Emotion Alignment

For the Emotion Alignment task, our experimental setup is designed to precisely control and simulate distributional shifts in user preferences. We begin by training a reward model based on a `GPT-2` architecture Radford et al. [2019] augmented with a classification head. This model is fine-tuned on the Emotion dataset Saravia et al. [2018] to perform multi-label classification across five distinct emotions: sadness, joy, love, anger, and fear. The sigmoid outputs from this classification head serve as our multi-objective reward signals throughout the experiment. In parallel, a base policy model, also a `GPT-2` instance, undergoes supervised fine-tuning (SFT) on the same Emotion dataset, establishing our initial policy for subsequent preference alignment.

To generate preference data and systematically introduce distributional shifts, we define two distinct reward mixing functions using two chosen reward objectives, r_1 (anger) and r_2 (fear):

1. **Convex Mixing:** $r^*_{\text{convex}}(\alpha) := \alpha \cdot r_1 + (1-\alpha) \cdot r_2$
2. **Geometric Mixing:** $r^*_{\text{geometric}}(\alpha) := r_1^{\alpha} \cdot r_2^{1-\alpha}$

For training all alignment methods, preference pairs are constructed by generating two completions per prompt. Preference labels are then assigned using the Bradley-Terry (BT) model, parameterized by the mixed reward function $r^*(\alpha_0)$ at a fixed nominal mixing coefficient $\alpha_0 = 0.1$. This α_0 represents the training-time preference distribution. To evaluate robustness, we introduce a controlled distributional shift by sweeping the mixing coefficient α across the entire range $[0, 1]$ at test time. Performance is measured by the average mixture reward $r^*(\alpha)$ obtained over 64 held-out prompts. More comprehensive details about these experiments are outlined in Appendix M

7.1.3 ArmoRM Multi-objective Alignment

To assess performance in a more complex, real-world setting, we conduct experiments on a multi-objective alignment task utilizing the Absolute-Rating Multi-Objective Reward Model (ArmoRM) Wang et al. [2024a]. ArmoRM provides 19 distinct first-stage objective outputs, allowing for a rich and diverse set of preferences. Our base policy for this task is Meta LLaMA-3.2-1B-Instruct.

For training, we generate two completions per prompt from the HelpSteer2 dataset Wang et al. [2024b]. Reward preferences are derived by selecting pairs of equally weighted objectives (e.g., honesty, verbosity, safety) from ArmoRM's 19-dimensional output space and combining them via convex mixing. Models are then trained on these preferences, again at a nominal mixing coefficient of $\alpha_0 = 0.1$. We measure the performance of all aligned policies on five individual ArmoRM objectives. Crucially, three of these five objectives are *unseen* during the training process, specifically designed to simulate real-world scenarios where models encounter new or unweighted preferences. The evaluation is conducted over 128 test prompts. All fine-tuning runs across both the Emotion Alignment and ArmoRM tasks adhere to identical hyperparameters, as comprehensively detailed in Appendix M.

7.2 Results

7.2.1 Emotion Alignment

The results for Emotion Alignment, presented in Figure 3, clearly illustrate the robustness of various methods under varying degrees of preference shift for both convex and geometric reward mixing. Non-robust baselines, DPO and

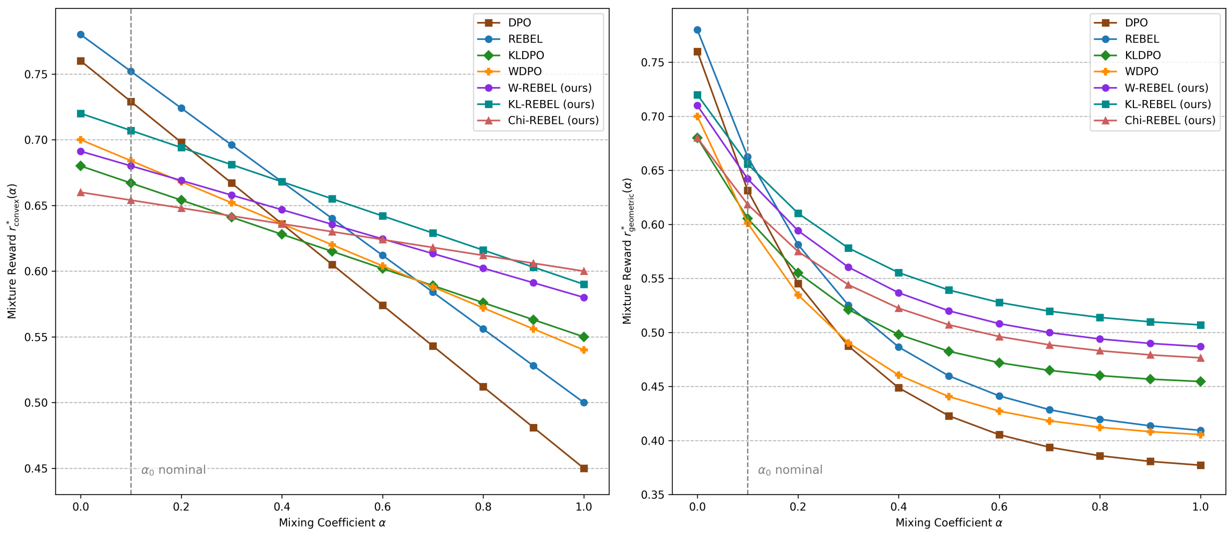

Figure 3: Emotion alignment performance under convex (left) and geometric (right) reward mixing. Models are trained at a nominal mixing coefficient $\alpha_0 = 0.1$, and evaluated across a range of $\alpha \in [0, 1]$ to simulate preference shift. This figure compares the mixture reward for DPO, REBEL, and various DRO variants (KLDPO, WDPO, KL-REBEL, W-REBEL, and χ^2-REBEL).

REBEL, achieve high rewards at the training-time nominal $\alpha_0 = 0.1$. However, their performance degrades significantly linearly as α deviates from α_0. The same can be seen for geometric mixing, and this implies there is a susceptibility to overoptimization on the training distribution. Prior DRO variants applied to DPO, namely WDPO and KLDPO, offer modest improvements in robustness compared to plain DPO. Our robust REBEL variants, W-REBEL and KL-REBEL, demonstrate even more substantial gains in mixture reward beyond the DPO and REBEL baselines. Among all tested methods, χ^2-REBEL emerges as the most stable, consistently retaining a high reward across the entire range of α.

7.2.2 Radius–Coverage Trade-off

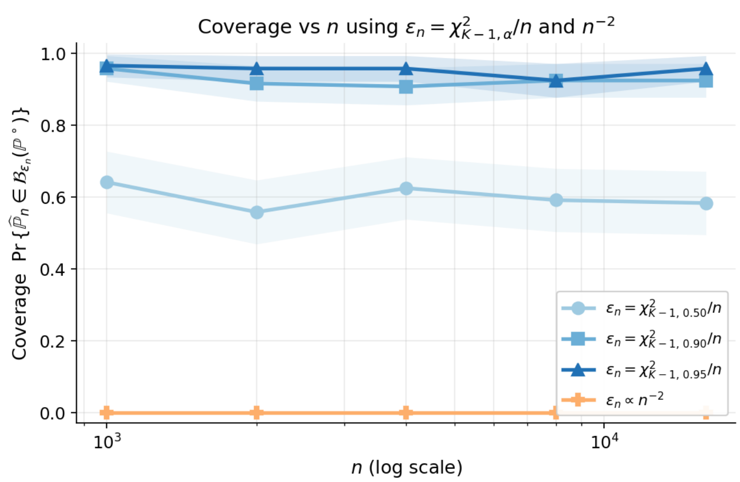

Figure 4: Empirical coverage $\Pr\{\widehat{p} \in \mathcal{B}_{\varepsilon_n}(p^\circ)\}$ for calibrated $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ (several α) and a fast $\varepsilon_n \propto n^{-2}$ baseline. The χ^2-calibrated schedules maintain stable target coverage as n grows, while n^{-2} rapidly under-covers.

Figure 4 reports empirical coverage $\Pr\{\widehat{\mathbb{P}}_n \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ)\}$ for four radius schedules. The χ^2-calibrated radii $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ (for $\alpha \in \{0.50, 0.90, 0.95\}$) achieve the intended, *stable* coverage across sample sizes $n \in \{10^3, 2\times10^3, 4\times 10^3, 8\times10^3, 1.6\times10^4\}$, with larger α yielding higher coverage as expected from Pearson's statistic. In contrast, the aggressively shrinking baseline $\varepsilon_n \propto n^{-2}$ under-covers (essentially 0) at all n, illustrating that picking radii by asymptotics alone can catastrophically miss the nominal guarantee. Shaded regions denote $\pm 1.96\,\mathrm{SE}$ over $R{=}120$ resamples.

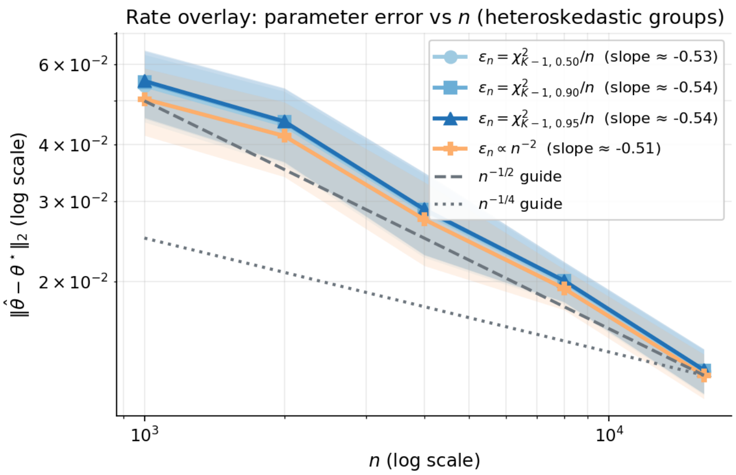

Figure 5: Parameter error $\|\hat{\theta} - \theta^\star\|_2$ versus n (log–log), with $n^{-1/2}$ and $n^{-1/4}$ slope guides. Across χ^2-calibrated radii, the empirical slope tracks the $n^{-1/2}$ benchmark expected for well-specified linear models, while overly small radii risk variance blow-up and overly large radii induce bias.

Figure 5 plots the parameter estimation error $\|\hat{\theta} - \theta^\star\|_2$ versus n (log–log). All χ^2 schedules exhibit an empirical slope near $-1/2$, closely tracking the $n^{-1/2}$ guide for well-specified linear models with heteroskedastic noise. The curves differ mainly by a vertical offset: smaller radii (e.g., n^{-2}) regularize less and can sit marginally lower at small n (lower bias), but this comes at the cost of the severe under-coverage seen in Figure 4. Conversely, the calibrated χ^2 radii incur no visible penalty in convergence rate while achieving the target coverage. Bands show $\pm 1.96\,\mathrm{SE}$ across 8 seeds.

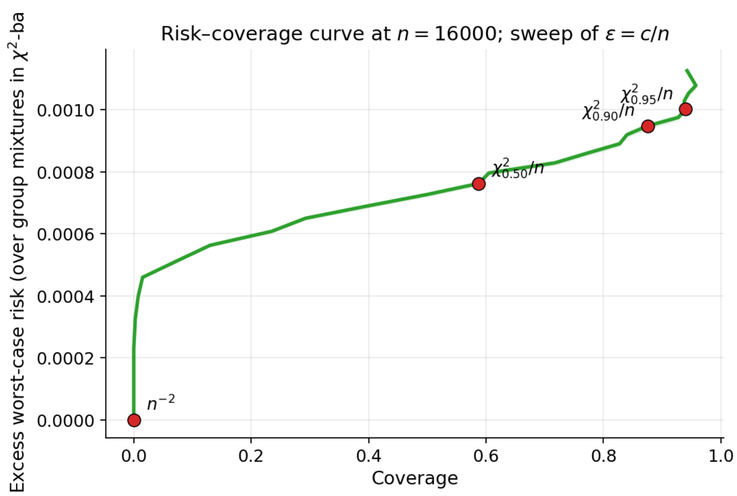

Figure 6: At fixed n, sweeping $\varepsilon = c/n$ traces a monotone frontier between coverage and excess worst-case risk against χ^2 mixture shifts. Calibrated choices (e.g., $\chi^2_{K-1,0.90}/n$) sit near a knee of the curve, balancing protection and estimation error.

Finally, Figure 6 traces the risk–coverage frontier at fixed $n{=}16{,}000$ by sweeping $\varepsilon = c/n$. As c increases, coverage monotonically improves but excess worst-case risk against χ^2 mixture shifts grows, revealing a clear knee. Calibrated choices such as $\chi^2_{K-1,0.90}/n$ and $\chi^2_{K-1,0.95}/n$ lie near this knee, balancing protection and estimation error. The n^{-2} point sits far left with minimal risk but negligible coverage, underscoring that "fast" radii do not deliver meaningful distributional robustness. Overall, the experiment makes the central trade-off visible: *calibrating* ε_n by $\chi^2_{K-1,\alpha}/n$ yields dimension-free, stable coverage *without* sacrificing the $n^{-1/2}$ convergence rate.

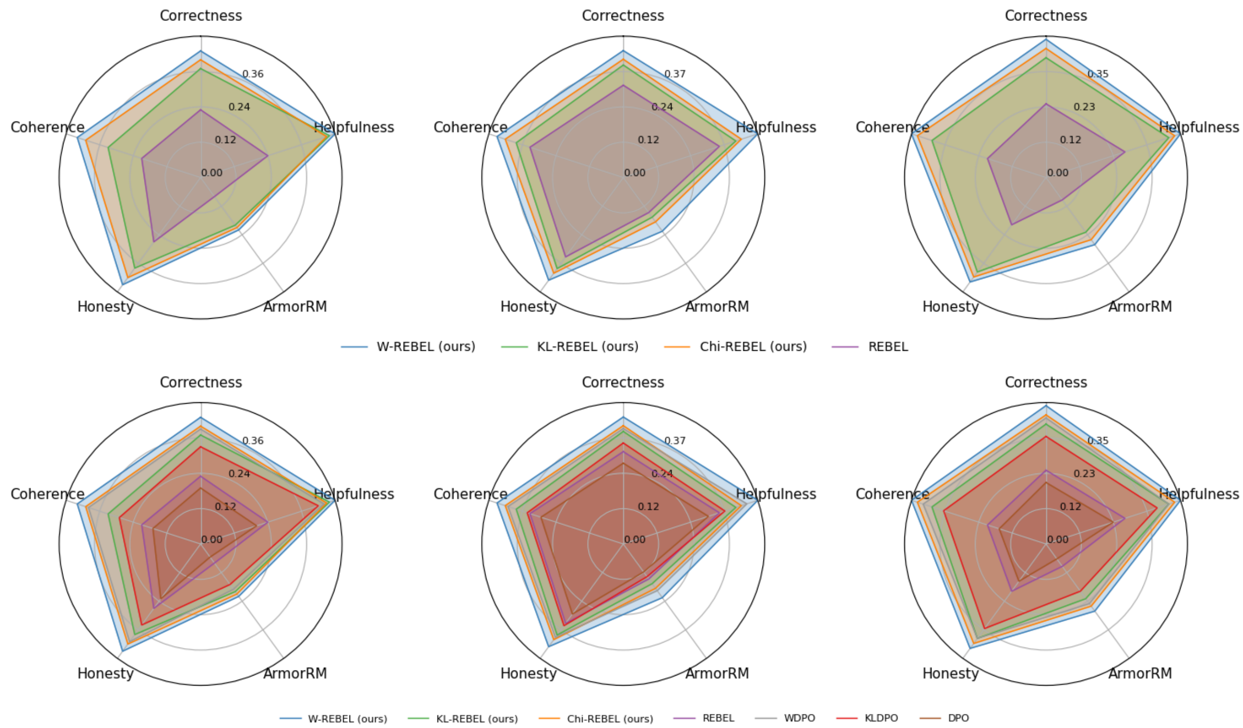

Figure 7: Performance on five ArmoRM objectives (Correctness, Helpfulness, Honesty, ArmoRM, Coherence), including three (Honesty, ArmoRM, Coherence) that were unseen during training. The radar charts illustrate the alignment across different objectives, where larger shaded areas indicate stronger robustness and better overall performance under preference shift across diverse objectives.

7.2.3 ArmoRM Multi-objective Alignment

Figure 7 illustrates the multi-objective alignment performance on the ArmoRM task. The size of the shaded area directly correlates with overall alignment and robustness. DPO and REBEL baselines exhibit noticeable performance drops on objectives unseen during training, demonstrating a clear tendency towards overoptimization on the specific objective combinations present in the training data, leading to poor generalization. WDPO and KLDPO, as prior DRO variants of DPO, show the capacity to recover performance on these unseen objectives. Our REBEL-based robust variants, W-REBEL, KL-REBEL, and χ^2-REBEL, achieve further significant gains on the unseen objectives compared to their DPO counterparts. These results also support and are fully consistent with the theoretical and empirical insights of Huang et al. [2025] as it shows that χ^2 DRO in RLHF can achieve broad generalization and maintain robustness across diverse objectives, including those not explicitly encountered or optimized during training.

8 Conclusion

In this work, we introduced *DRO-REBEL*, a unified family of distributionally-robust variants of the REBEL framework for offline Reinforcement Learning from Human Feedback. Leveraging strong duality, we showed that each robust policy update reduces to a simple, scalable relative-reward regression. Theoretically, we proved that DRO-REBEL achieves the minimax-optimal $O(n^{-1/2})$ fast rate with tighter constants than prior methods, restoring classical parametric convergence even under distributional shifts. Our theory reveals a "no free lunch" principle. Radii that decay $\ll n^{-1/2}$ (e.g. n^{-1}) make the robust term asymptotically negligible; DRO–REBEL then behaves like ERM and enjoys the usual $O(n^{-1/2})$ error—yet the ball may exclude the true distribution with non-vanishing probability. To *guarantee* coverage one must slow the decay to roughly the divergence-concentration rate, leading to the $O(n^{-1/4})$ bounds we also prove. We argue that many LLM preference problems effectively operate in a low intrinsic dimension, justifying a practical choice such as $\varepsilon_n \asymp n^{-1}$ that balances robustness and efficiency. Empirically, DRO-REBEL demonstrated superior performance in maintaining alignment across preference shifts and generalizing to unseen objectives. Future work includes developing practical guidelines for selecting ambiguity sets and investigating whether the joint integration of robust reward modeling with DRO-REBEL further improves alignment against noisy or adversarial human feedback. We are also interested to see if more advanced techniques can recover a linear dependence on Δ_n which would provide a path towards minimax-optimal rates for parameter estimation along with coverage guarantees. Additionally, our analysis relies on Assumption 3, which ensures sufficient coverage of preferred and dispreferred responses within the feature space of the data-generating distribution. This assumption is crucial as it provides desirable properties, such as strong convexity, for our robust loss functions. A key area for future research is exploring whether fast convergence rates, similar to those established in this paper, can be achieved when this assumption is relaxed in the analysis of robust RLHF.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL `https://doi.org/10.1145/1015330.1015430`.

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. 2021a. URL `https://rltheorybook.github.io/`.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021b. URL `http://jmlr.org/papers/v22/19-736.html`.

Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. Oxford University Press, Oxford, UK, 2000. ISBN 978-0-19-850245-6.

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, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL `https://arxiv.org/abs/2204.05862`.

A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2017. ISBN 9781611974997. URL `https://books.google.com/books?id=wrk4DwAAQBAJ`.

A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms and Applications with Python and MATLAB*. MOS-SIAM series on optimization. Society for Industrial and Applied Mathematics, 2023. ISBN 9781611977615. URL `https://books.google.com/books?id=YDrizwEACAAJ`.

Jose Blanchet and Alexander Shapiro. Statistical limit theorems in distributionally robust optimization, 2023. URL `https://arxiv.org/abs/2303.14867`.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL `https://books.google.com/books?id=5oo4YIz6tR0C`.

Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback, 2024. URL `https://arxiv.org/abs/2406.15568`.

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, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL `https://arxiv.org/abs/2307.15217`.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences, 2024. URL `https://arxiv.org/abs/2402.08925`.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback, 2024. URL `https://arxiv.org/abs/2403.00409`.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL `https://arxiv.org/abs/1706.03741`.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2nd edition, 1998.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization, 2020. URL `https://arxiv.org/abs/1810.08750`.

John C. Duchi and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 47(2):753–789, 2022.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance, 2022. URL `https://arxiv.org/abs/1604.02199`.

Yang Gao, Yuxin Xie, Nan Jiang, and Lihong Wang. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 8512–8530, 2022.

Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards, 2024. URL `https://arxiv.org/abs/2404.16767`.

Shai Gur and Ohad Shamir. On the intrinsic dimensionality of data representations. In *International Conference on Machine Learning*, 2022.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4572–4580, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. Adaptive preference scaling for reinforcement learning with human feedback, 2024. URL `https://arxiv.org/abs/2406.02764`.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization, 2025. URL `https://arxiv.org/abs/2407.13399`.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl?, 2022. URL `https://arxiv.org/abs/2012.15085`.

Yiping Jin et al. Information-theoretic bounds on generalization for deep learning. In *International Conference on Learning Representations*, 2023.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL `https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf`.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL `https://arxiv.org/abs/2310.06452`.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991. ISBN 9783540520139. URL `https://books.google.com/books?id=cyKYDfvxRjsC`.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998. Chapter 5 discusses Barankin-type bounds, including HCR.

Will LeVine, Benjamin Pikus, Anthony Chen, and Sean Hendryx. A baseline analysis of reward models' ability to accurately analyze foundation models under distribution shift, 2024. URL `https://arxiv.org/abs/2311.14743`.

Debmalya Mandal, Paulius Sasnauskas, and Goran Radanovic. Distributionally robust reinforcement learning with human feedback, 2025. URL https://arxiv.org/abs/2503.00539.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *Annals of Probability*, 18(3): 1269–1283, 1990.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2010–2020. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/modi20a.html.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5a142a55461d5fef016acfb927fee0bd-Paper.pdf.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanović, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences, 2024. URL https://arxiv.org/abs/2403.01857.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. Beyond the binary: Capturing diverse preferences with reward regularization, 2024. URL https://arxiv.org/abs/2412.03822.

Yury Polyanskiy. Lecture notes on information theory, chapter 29, ece563 (uiuc). Technical report, Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, 2017. URL https://people.lids.mit.edu/yp/homepage/data/LN_stats.pdf. Archived (PDF) from the original on 2022-05-24. Retrieved 2022-05-24.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9(129-145):20, 1935.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020. ISBN 9781728199986.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation, 2019. URL https://arxiv.org/abs/1710.10016.

Harvineet Singh Shah, Michael Jung, Kyomin Jung, and Hwanjo Kim. Robust optimization for fairness with noisy protected groups. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5702–5709, 2020.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. ISBN 9781107057135. URL https://books.google.com/books?id=ttJkAwAAQBAJ.

Alexander Shapiro and Huan Xu. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research*, 70(5):3007–3030, 2022.

Karthik Sridharan. Lecture6: Properties of rademacher complexity, and examples. Course notes, CS4783/5783 *Mathematical Foundations of Machine Learning*, Cornell University, 2022. URL https://www.cs.cornell.edu/courses/cs4783/2022sp/notes06.pdf. accessed 28Jul 2025.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2009.

A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405. URL https://books.google.com/books?id=OCenCW9qmp4C.

Ramon van Handel. Probability in high dimensions. 2016. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.com/books?id=hV8o5R7_5tkC.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts, 2024a. URL https://arxiv.org/abs/2406.12845.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline rl with linear function approximation?, 2020. URL https://arxiv.org/abs/2010.11895.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024b. URL https://arxiv.org/abs/2406.08673.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization, 2024. URL https://arxiv.org/abs/2407.07880.

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

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Distributionally robust direct preference optimization, 2025. URL https://arxiv.org/abs/2502.01930.

Yuzi Yan, Xingzhou Lou, Jialian Li, Yiping Zhang, Jian Xie, Chao Yu, Yu Wang, Dong Yan, and Yuan Shen. Reward-robust rlhf in llms, 2024. URL https://arxiv.org/abs/2409.15360.

Fan Yang, Shixiang Gu, Sergey Levine, Dale Schuurmans, and Ofir Nachum. Wasserstein distributional reinforcement learning. In *International Conference on Learning Representations*, 2021.

Xiaoying Zhang, Jean-Francois Ton, Wei Shen, Hongning Wang, and Yang Liu. Overcoming reward overoptimization via adversarial policy optimization with lightweight uncertainty estimation, 2024. URL https://arxiv.org/abs/2403.05171.

Yichen Zhang, Yuxin Xie, and Lihong Wang. Generalization in reinforcement learning with stochastic mirror descent. In *International Conference on Machine Learning*, volume 119, pages 11188–11197, 2020.

Yingxue Zhou, Haoran Li, Longbo Huang, and Peilin Zhao. Distributionally robust exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 35, pages 17328–17340, 2022.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/zhu23f.html`.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL `https://arxiv.org/abs/1909.08593`.

A Auxiliary Technical Tools

A.1 Wasserstein Theory

Lemma 5 (Gao and Kleywegt [2022], Theorem 1; Strong Duality for DRO with Wasserstein Distance)**.** *Consider any* $p \in [1, \infty)$*, any* $\nu \in \mathcal{P}(\Xi)$*, any* $\rho > 0$*, and any* $\Psi \in L^1(\nu)$ *such that the growth rate* κ *of* Ψ *satisfies*

$$\kappa := \inf\big\{\eta \geq 0 : \int_\Xi \Phi(\eta, \zeta)\, \nu(d\zeta) > -\infty\big\} < \infty, \tag{13}$$

where

$$\Phi(\eta, \zeta) := \inf_{\xi \in \Xi}\big\{\eta\, d^p(\xi, \zeta) - \Psi(\xi)\big\}.$$

Then strong duality holds with finite optimal value $v_p = v_D \leq \infty$*, where the primal and dual problems are*

$$v_p = \sup_{\mu \in \mathcal{P}(\Xi)} \Big\{\int_\Xi \Psi(\xi)\, \mu(d\xi) \colon W_p(\mu, \nu) \leq \rho\Big\}, \quad \textit{(Primal)} \tag{4}$$

$$v_D = \inf_{\eta \geq 0} \Big\{\eta\, \rho^p \;-\; \int_\Xi \inf_{\xi \in \Xi}\big[\eta\, d^p(\xi, \zeta) - \Psi(\xi)\big]\, \nu(d\zeta)\Big\}. \quad \textit{(Dual)} \tag{5}$$

Lemma 6 (Gao and Kleywegt [2022], Lemma 2(ii); Properties of the growth κ)**.** *Suppose that* $\nu \in \mathcal{P}_p(\Xi)$*. Then the growth rate* κ *in* (13) *is finite if and only if there exists* $\zeta_0 \in \Xi$ *and constants* $L, M > 0$ *such that*

$$\Psi(\xi) - \Psi(\zeta_0) \;\leq\; L\, d^p(\xi, \zeta_0) + M, \quad \forall\, \xi \in \Xi. \tag{14}$$

Corollary 4. *Consider any bounded loss function* ℓ *over a bounded space* Ξ*. Then the duality in Lemma 2 holds.*

Proof. Immediate from Lemma 6 by choosing $L = \operatorname{diam}(\Xi)^p$ and $M = \sup_{\xi \in \Xi} |\Psi(\xi)|$. ☐

Lemma 7 (Villani [2008]; Monotonicity of the Wasserstein distance)**.** *Suppose* $1 \leq p \leq q < \infty$*. Then, it follows that* $\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \leq \mathcal{W}_q(\mathbb{P}, \mathbb{Q})$*. This implies that if we let* $\mathcal{B}_\varepsilon^p(\mathbb{Q}) = \{\mathbb{P} \in \mathcal{M}(\Xi) : \mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \leq \varepsilon\}$*, then* $\mathcal{B}_\varepsilon^q(\mathbb{Q}) \subseteq \mathcal{B}_\varepsilon^p(\mathbb{Q})$

Proof. Fix any coupling $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ where $\xi \sim \mathbb{P}, \eta \sim \mathbb{Q}$. Let $Z(x, y) = d(x, y)^q$ and $r = p/q \in (0, 1)$. Then, $Z(x, y)^r = d(x, y)^p$. Since $t \mapsto t^r$ for $r \in (0, 1)$ is concave, by Jensen's inequality, we have

$$\int d(x, y)^p \pi(d\xi, d\eta) = \int Z(x, y)^r \pi(d\xi, d\eta) \leq \left(\int Z(x, y) \pi(d\xi, d\eta)\right)^r = \left(\int d(x, y)^q \pi(d\xi, d\eta)\right)^{p/q}$$

Taking the infimum over all couplings, we get that

$$\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int d(x, y)^p \pi(d\xi, d\eta) \leq \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \left(\int d(x, y)^q \pi(d\xi, d\eta)\right)^{p/q} \leq \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int d(x, y)^q \pi(d\xi, d\eta)\right)^{p/q}$$

This is equivalent to $\mathcal{W}_p(\mathbb{P}, \mathbb{Q})^p \leq \mathcal{W}_q(\mathbb{P}, \mathbb{Q})^p$. Thus, $\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \leq \mathcal{W}_q(\mathbb{P}, \mathbb{Q})$. ☐

A.2 Optimization

Lemma 8 (Beck [2023], Theorem 1.24; Linear Approximation Theorem). *Let $f\colon U \to \mathbb{R}$ be twice continuously differentiable on an open set $U \subseteq \mathbb{R}^n$, and let $x, y \in U$ satisfy $[x, y] \subset U$. Then there exists $\xi \in [x, y]$ such that*

$$f(y) \;=\; f(x) \;+\; \nabla f(x)^\top (y-x) \;+\; \frac{1}{2}\,(y-x)^\top \nabla^2 f(\xi)\,(y-x).$$

Lemma 9 (Beck [2017], Theorem 5.24; First-order characterizations of strong convexity). *Let $f\colon E \to (-\infty, \infty]$ be a proper, closed, convex function, and let $\sigma > 0$. The following are equivalent:*

1. *For all $x, y \in \operatorname{dom}(f)$ and $\lambda \in [0, 1]$,*
$$f\big(\lambda x + (1-\lambda)y\big) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\sigma}{2}\,\lambda(1-\lambda)\,\|x-y\|^2.$$
2. *For all $x \in \operatorname{dom}(\partial f)$, $y \in \operatorname{dom}(f)$ and $g \in \partial f(x)$,*
$$f(y) \;\ge\; f(x) \;+\; \langle g, y-x\rangle \;+\; \frac{\sigma}{2}\,\|y-x\|^2.$$

Lemma 10 (Beck [2017], Theorem 5.25; Existence and uniqueness of minimizer). *Let $f\colon E \to (-\infty, \infty]$ be proper, closed, and σ-strongly convex with $\sigma > 0$. Then:*

1. *f has a unique minimizer x^*.*
2. *For all $x \in \operatorname{dom}(f)$,*
$$f(x) - f(x^*) \;\ge\; \frac{\sigma}{2}\,\|x - x^*\|^2.$$

A.3 Distributionally Robust Optimization

The f-divergence between distributions $\mathbb{P}$ and $\mathbb{P}_0$ on $\mathcal{X}$ is

$$D_f(P\|P_0) \;=\; \int_{\mathcal{X}} f\left(\frac{d\mathbb{P}}{d\mathbb{P}_0}\right)\, d\mathbb{P}_0, \tag{15}$$

where f is a convex function (e.g. $f(t) = t\log t$ gives KL divergence). For a loss $\ell\colon \mathcal{X} \to \mathbb{R}$:

Lemma 11 (Duchi and Namkoong [2020], Proposition 1). *Let D_f be as in* (15)*. Then*

$$\sup_{P\colon D_f(P\|P_0)\le\rho} \mathbb{E}_P[\ell(X)] \;=\; \inf_{\substack{\lambda\ge 0\\ \eta\in\mathbb{R}}} \Big\{\lambda\, f^*\left(\frac{\ell(X)-\eta}{\lambda}\right) \;+\; \lambda\,\rho \;+\; \eta\Big\}, \tag{16}$$

where $f^*(s) = \sup_{t\ge 0}\{st - f(t)\}$ is the Fenchel conjugate of f.

A.4 Empirical Process Theory

Lemma 12 (van der Vaart and Wellner [1996], Lemma 2.3.1; Symmetrization). *For every nondecreasing, convex $\Phi : \mathbb{R} \to \mathbb{R}$ and class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}^*\Big[\Phi\big(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big)\Big] \;\le\; \mathbb{E}^*\Big[\Phi\big(2\,\|\mathbb{P}_n^{\circ}\|_{\mathcal{F}}\big)\Big],$$

where the outer expectations $\mathbb{E}^*$ are taken over the data generating distribution and Rademacher random variables and $\mathbb{P}_n^{\circ}$ is the symmetrized process.

Theorem 12 (Ledoux and Talagrand [1991], Theorem 4.12; Ledoux–Talagrand Lipschitz contraction). *Fix samples $z_1, \ldots, z_n \in \mathcal{Z}$ and let $\sigma_1, \ldots, \sigma_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, 1\}$. Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ and, for each i, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz with $\psi_i(0) = 0$. Then*

$$\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\, \psi_i\big(f(z_i)\big)\Big] \;\le\; L\,\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\, f(z_i)\Big].$$

Consequently, the empirical Rademacher complexity satisfies

$$\mathfrak{R}_n\big(\psi\circ\mathcal{F}\big) \;\le\; L\,\mathfrak{R}_n(\mathcal{F}),$$

where $\psi\circ\mathcal{F} = \{\, z \mapsto \psi(f(z)) \;:\; f \in \mathcal{F}\}$ and $\widehat{\mathfrak{R}}_n$ is computed on the fixed sample $\{z_i\}_{i=1}^n$.

Lemma 13.

Lemma 14 ([Sridharan, 2022], Proposition 2.1(i) - Lecture 6; Monotonicity of Rademacher Complexity)**.** *Let $\mathcal{H}$ and $\mathcal{G}$ be two classes of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$, and let $\mathfrak{R}_n(\mathcal{F}) := \frac{1}{n}\,\mathbb{E}_\sigma\big[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(z_i)\big]$ denote the empirical Rademacher complexity on a fixed sample $S = \{z_1, \ldots, z_n\}$. If $\mathcal{H} \subseteq \mathcal{G}$ then*

$$\mathfrak{R}_n(\mathcal{H}) \;\leq\; \mathfrak{R}_n(\mathcal{G}).$$

Lemma 15 (Countable sub-additivity of Rademacher complexity)**.** *Let $\{\mathcal{F}_k\}_{k\geq 1}$ be a countable family of classes $\mathcal{F}_k \subset \mathbb{R}^{\mathcal{Z}}$. For any fixed sample $S = \{z_1, \ldots, z_n\}$ the empirical Rademacher complexity satisfies*

$$\mathfrak{R}_n\Big(\bigcup_{k=1}^{\infty}\mathcal{F}_k\Big) \;\;\leq\;\; \sum_{k=1}^{\infty}\mathfrak{R}_n(\mathcal{F}_k),$$

where $\mathfrak{R}_n(\mathcal{H}) := \dfrac{1}{n}\,\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^n \sigma_i h(z_i)\Big|\Big]$. Taking an outer expectation over the sample, the same inequality holds for the population quantity $\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_S[\mathfrak{R}_n(\mathcal{H})]$.

Proof. We include a short proof because an explicit statement is hard to locate in the literature. By definition,

$$\mathfrak{R}_n\Big(\bigcup_{k=1}^{\infty}\mathcal{F}_k\Big) = \mathbb{E}_{z,\sigma}\left[\sup_{f\in\bigcup_{k=1}^{\infty}\mathcal{F}_k}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right|\right]$$

Using the inequality $\sup_{f\in\bigcup_{k=1}^{\infty}\mathcal{F}_k} f(z_i) \leq \sum_{k=1}^{\infty}\sup_{f\in\mathcal{F}_k} f(z_i)$, we have

$$\begin{aligned}\mathfrak{R}_n\Big(\bigcup_{k=1}^{\infty}\mathcal{F}_k\Big) &\leq \sum_{k=1}^{\infty}\mathbb{E}_{z,\sigma}\left[\sup_{f\in\mathcal{F}_k}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big|\right]\\ &= \sum_{k=1}^{\infty}\mathfrak{R}_n(\mathcal{F}_k)\end{aligned}$$

□

Lemma 16 ([Sridharan, 2022], Section 3 - Lecture 6; Rademacher complexity of a linear class)**.** *Let $\mathcal{F} = \big\{x \mapsto \langle f, x\rangle : \|f\|_2 \leq R\big\}$ and fix a sample $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$. Then the empirical Rademacher complexity satisfies*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \;=\; \frac{1}{n}\,\mathbb{E}_\sigma\Big[\sup_{\|f\|_2\leq R}\Big|\sum_{t=1}^n \sigma_t\langle f, x_t\rangle\Big|\Big] \;=\; \frac{R}{n}\,\mathbb{E}_\sigma\Big[\big\|\textstyle\sum_{t=1}^n \sigma_t x_t\big\|_2\Big] \;\leq\; \frac{R}{\sqrt{n}}\max_{1\leq t\leq n}\|x_t\|_2$$

Theorem 13 (Shalev-Shwartz and Ben-David [2014]; Theorem 26.5)**.** *Let $\mathcal{F}$ be a class of real-valued functions $f : \mathcal{X} \to \mathbb{R}$, and let $X_1, \ldots, X_n$ be independent samples from a distribution $\mathbb{P}$. Define the empirical measure*

$$\mathbb{P}_n f \;:=\; \frac{1}{n}\sum_{i=1}^n f(X_i),$$

and let $\mathfrak{R}(\mathcal{F})$ denote the empirical Rademacher complexity of the class $\mathcal{F}$. Then with probability at least $1-\delta$, the following holds uniformly for all $f \in \mathcal{F}$:

$$\mathbb{P}f \;-\; \mathbb{P}_n f \;\leq\; \Upsilon,$$

where

$$\Upsilon \;:=\; 2\mathfrak{R}(\mathcal{F}) \;+\; 4\sup_{f\in\mathcal{F}}\|f\|_\infty\sqrt{\frac{2\ln(4/\delta)}{n}}.$$

A.5 Variational Inequalities

Lemma 17 (van Handel [2016], Lemma 4.10; Gibbs Variational Principle). *Let $\mu, \nu \in \mathcal{P}(\Xi)$ be Borel probability measures supported on Ξ. Then*

$$\log \mathbb{E}_\mu \left[e^f\right] = \sup_\nu \left\{\mathbb{E}_\nu [f] - D_{\mathrm{KL}}(\mu \,||\, \nu)\right\}$$

Theorem 14 ([Polyanskiy, 2017, Lehmann and Casella, 1998]; Hammersley-Chapman-Robbins (HCR) lower bound). *Let Θ be the set of parameters for a family of probability distributions $\{\mu_\theta : \theta \in \Theta\}$ on a sample space Ω. For any $\theta, \theta' \in \Theta$, let $\chi^2\big(\mu_{\theta'};\, \mu_\theta\big)$ denote the χ^2–divergence from μ_θ to $\mu_{\theta'}$. For any scalar random variable $\hat{g} \colon \Omega \to \mathbb{R}$ and any $\theta, \theta' \in \Theta$, we have*

$$\mathrm{Var}_\theta(\hat{g}) \;\geq\; \sup_{\substack{\theta' \neq \theta \\ \theta' \in \Theta}} \frac{\big(\mathbb{E}_{\theta'}[\hat{g}] - \mathbb{E}_\theta[\hat{g}]\big)^2}{\chi^2\big(\mu_{\theta'};\, \mu_\theta\big)}.$$

A.6 Concentration Inequalities

Lemma 18 (Boucheron et al. [2013], Lemma 2.2; Hoeffding's Lemma). *Let Y be a random variable with $\mathbb{E}[Y] = 0$ and almost surely $Y \in [a, b]$. Define $\psi_Y(\lambda) = \log \mathbb{E}\big[e^{\lambda Y}\big]$. Then for all $\lambda \in \mathbb{R}$, $\psi_Y''(\lambda) \leq \frac{(b-a)^2}{4}$ and consequently Y is sub-Gaussian with proxy variance $(b-a)^2/4$, i.e. $Y \sim \mathcal{SG}\left(\frac{b-a}{2}\right)$.*

Using Hoeffding's lemma, one can prove Hoeffding's inequality using a standard Chernoff bound argument.

Lemma 19 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent with $X_i \in [a_i, b_i]$ almost surely, and define*

$$S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i]).$$

Then for every $t > 0$,

$$\mathbb{P}(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. with mean μ and support $[a, b]$, then for all $t > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i \;-\; \mathbb{E}[X]\right|\right) \geq t\big) \;\leq\; 2\exp\!\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Lemma 20 (Popoviciu [1935]; Popoviciu's Inequality on Variances). *Let X be a real-valued random variable with support contained in a finite interval $[a, b]$. Then, the variance of X satisfies*

$$\mathrm{Var}(X) \;\leq\; \frac{1}{4}(b-a)^2.$$

Equality holds if and only if X takes values a and b with probability $1/2$ each.

A.7 Divergences and Minimax

Lemma 21 (KL Divergence for Product Measures). *Let $P_1, \ldots, P_n$ be probability measures on $(\mathcal{X}_1, \mathcal{F}_1), \ldots, (\mathcal{X}_n, \mathcal{F}_n)$ and $Q_1, \ldots, Q_n$ be probability measures on the same spaces. Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ be the corresponding product measures on the product space $(\mathcal{X}^n, \mathcal{F}^n)$. Then, the Kullback-Leibler (KL) divergence between P and Q is the sum of the individual KL divergences:*

$$D_{\mathrm{KL}}(P \| Q) = \sum_{i=1}^n D_{\mathrm{KL}}(P_i \| Q_i).$$

Lemma 22 (KL Divergence for Univariate Normals, adapted from Cover and Thomas [2006]). *Let $P = \mathcal{N}(\mu_0, \sigma_0^2)$ and $Q = \mathcal{N}(\mu_1, \sigma_1^2)$ be two univariate normal distributions. The Kullback-Leibler (KL) divergence of P from Q is given by*

$$D_{\mathrm{KL}}(P \| Q) = \frac{1}{2}\left(\frac{(\mu_0 - \mu_1)^2}{\sigma_1^2} + \frac{\sigma_0^2}{\sigma_1^2} - 1 - \ln\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right).$$

Theorem 15 (Pinsker's Inequality, adapted from Cover and Thomas [2006]). *Let P and Q be two probability distributions on a measurable space $(\mathcal{X}, \mathcal{F})$. Then the total variation distance between P and Q is bounded by the Kullback-Leibler (KL) divergence between them:*

$$d_{\mathrm{TV}}(P, Q) \ \leq \ \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \| Q)},$$

where the total variation distance is defined as $d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$.

Lemma 23 (Le Cam's Two-Point Lemma, adapted from Tsybakov [2009]). *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions and let $\theta_0, \theta_1 \in \Theta$. For a fixed $p > 0$, consider the estimation problem of θ under the loss function $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^p$. For any estimator $\hat{\theta}_n$, the minimax risk is bounded below by*

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[\|\hat{\theta}_n - \theta\|_2^p \right] \geq \frac{1}{2} \left(\frac{\|\theta_1 - \theta_0\|_2}{2} \right)^p (1 - d_{\mathrm{TV}}(P_{\theta_0}^n, P_{\theta_1}^n)),$$

where $d_{\mathrm{TV}}(P_{\theta_0}^n, P_{\theta_1}^n)$ is the total variation distance between the distributions of n i.i.d. observations from P_{θ_0} and P_{θ_1}.

B Proofs of Uniform Boundedness and Lipschitzness of $\ell(z; \theta)$

B.1 Uniform Boundedness of $\ell(z; \theta)$

We first prove that $\ell(z; \theta)$ is uniformly bounded

Lemma 24 (Uniform bound on $\ell(z; \theta)$). *Let $K_g = \sup_{z,\theta} |g(z; \theta)| \leq 8B/\eta + 2F$ where $\ell(z; \theta) = g(z; \theta)^2$ with $z = (x, a^1, a^2) \sim \mathbb{P}^\circ$. Then $\sup_{z,\theta} |\ell(z; \theta)| = K_\ell = K_g^2$*

Proof of Lemma 24. Since we have that $\pi_\theta, \pi_{\theta_t} \in \Pi$, notice that

$$\begin{aligned}
\log\left(\frac{\pi_\theta(a \mid x)}{\pi_{\theta_t}(a \mid x)}\right) &= \log \pi_\theta(a \mid x) - \log \pi_{\theta_t}(a \mid x) \\
&= \log\left(\frac{\exp\left(\theta^\top \psi(x, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x, a')\right)}\right) - \log\left(\frac{\exp\left(\theta_t^\top \psi(x, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x, a')\right)}\right) \\
&= \log\left(\exp\left((\theta - \theta_t)^\top \psi(x, a)\right)\right) + \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x, a')\right)\right) - \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x, a')\right)\right) \\
&= (\theta - \theta_t)^\top \psi(x, a) + \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x, a')\right)\right) - \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x, a')\right)\right)
\end{aligned}$$

Now since $\theta, \theta_t \in \Theta$ and $\max_{x,a} \|\psi(x, a)\|_2 \leq 1$, it follows that $\log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x, a')\right)\right) \in [\log(|\mathcal{A}|) - B, \log(|\mathcal{A}|) + B]$ by Cauchy-Schwartz. Thus we have

$$\begin{aligned}
\log\left(\frac{\pi_\theta(a \mid x)}{\pi_{\theta_t}(a \mid x)}\right) &= (\theta - \theta_t)^\top \psi(x, a) + \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x, a')\right)\right) - \log\left(\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x, a')\right)\right) \\
&\leq \max_{x,a} \|\psi(x, a)\|_2 \left(\|\theta\|_2 + \|\theta_t\|_2\right) + \log(|\mathcal{A}|) + B - (\log(|\mathcal{A}|) - B) \\
&\leq 4B
\end{aligned}$$

where the first inequality holds from the CS and triangle inequality. Now, we also have that $r \in \mathcal{F}$. Thus,

$$\begin{aligned}
r(x, a) - r(x, a') &= \phi(x, a)^\top \omega - \phi(x, a')^\top \omega \\
&\leq \left(||\phi(x, a)||_2 + ||\phi(x, a')||_2\right) ||\omega||_2 \\
&\leq 2F
\end{aligned}$$

where the first inequality holds from Cauchy-Schwartz and Triangle inequality. Now recall the REBEL update 3. Using these facts we have that $|g(z; \theta)| \leq 8B/\eta + 2F$ so $K_g = \sup_{z,\theta} |g(z; \theta)| \leq 8B/\eta + 2F$. Since $\ell(z; \theta) = g(z; \theta)^2$, $K_\ell = K_g^2$. ∎

B.2 Lipschitz bound on $\ell(z;\theta)$

Now we prove that $\ell(z;\theta)$ is $4K_g/\eta$-Lipschitz in θ.

Lemma 25 (Lipschitz bound on $\ell(z;\theta)$)**.** *$\ell(z;\theta)$ is $\frac{4K_g}{\eta}$-Lipschitz in θ.*

Proof of Lemma 25. First we compute the gradient $\nabla_\theta g(z;\theta)$. Since we are looking at updates with respect to θ, notice that we have the following:

$$\nabla_\theta g(z;\theta) = \nabla_\theta \left(\frac{1}{\eta} \left[\log \pi_\theta(a \mid x) - \log \pi_\theta(a' \mid x) \right] \right)$$

Now notice that

$$\begin{aligned}
\log \pi_\theta(a \mid x) - \log \pi_\theta(a' \mid x) &= \log\left(\exp\left(\theta^\top \psi(x,a)\right)\right) - \log\left(\exp\left(\theta^\top \psi(x,a')\right)\right) \\
&= \theta^\top \left(\psi(x,a) - \psi(x,a')\right)
\end{aligned}$$

Thus we find that

$$\nabla_\theta g(z;\theta) = \frac{1}{\eta}\left(\psi(x,a) - \psi(x,a')\right)$$

Thus by triangle inequality, $\sup_{x,a} ||\nabla_\theta g(z;\theta)||_2 \leq 2/\eta$. Now since $\ell(z;\theta) = g(z;\theta)^2$, we have that $\nabla_\theta \ell(z;\theta) = 2g(z;\theta)\nabla_\theta g(z;\theta)$. From Lemma 24, we know that $K_g = \sup_{z,\theta} |g(z;\theta)|$ so we see that $\sup_{x,a} ||\nabla_\theta \ell(z;\theta)||_2 \leq 4K_g/\eta$. Thus we can conclude that $\ell(z;\theta)$ is $4K_g/\eta$-Lipschitz in θ. $\square$

C Proof of "Slow Rate" Wasserstein-DRO-REBEL

First we prove that $h(\theta;\mathbb{P}) = \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is strongly convex for any $\mathbb{P}$.

C.1 Proof of Strong Convexity of WDRO-REBEL

Lemma 26 (Strong convexity of h)**.** *Let $\ell(z;\theta)$ be the REBEL loss function. Assume that Assumption 3 holds. Then $h(\theta;\mathbb{P}) = \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2/\eta^2$-strongly convex with respect to norm $||\cdot||_{\Sigma_\mathbb{P}}$ where $\Sigma_\mathbb{P} := \mathbb{E}_{(x,a^1,a^2,y)\sim\mathbb{P}}\left[\left(\psi(x,a^1) - \psi(x,a^2)\right)\left(\psi(x,a^1) - \psi(x,a^2)\right)^\top\right]$*

Proof of Lemma 26. We begin by computing the Hessian of $\ell(z;\theta)$ with respect to θ. From Lemma 25, we know that $\nabla_\theta \ell(z;\theta) = 2g(z;\theta)\nabla_\theta g(z;\theta)$. Differentiating again with respect to θ using the product rule, we get:

$$\nabla_\theta^2 \ell(z;\theta) = 2\nabla_\theta g(z;\theta)\nabla_\theta g(z;\theta)^\top + 2g(z;\theta)\nabla_\theta^2 g(z;\theta)$$

From Lemma 25, we know that $\nabla_\theta g(z;\theta) = \frac{1}{\eta}\left[\psi(x,a) - \psi(x,a')\right]$. Crucially, this gradient does not depend on θ. Therefore, $\nabla_\theta^2 g(z;\theta) = \mathbf{0}$. Substituting this into the Hessian expression, the second term vanishes:

$$\begin{aligned}
\nabla_\theta^2 \ell(z;\theta) &= \frac{2}{\eta^2}\left(\psi(x,a) - \psi(x,a')\right)\left(\psi(x,a) - \psi(x,a')\right)^\top \\
&= \frac{2}{\eta^2}\Sigma_z
\end{aligned}$$

where $\Sigma_z := \left(\psi(x,a) - \psi(x,a')\right)\left(\psi(x,a) - \psi(x,a')\right)^\top$. Note that Σ_z is a positive semi-definite matrix. Let $\theta, \theta' \in \Theta$. By the linear approximation theorem (Lemma 8), there exists $\alpha \in [0,1]$ and $\tilde{\theta} = \alpha\theta + (1-\alpha)\theta'$ such that

$$\ell(z;\theta') - \ell(z;\theta) - \langle \nabla_\theta \ell(z;\theta), \Delta\rangle = \frac{1}{2}\Delta^\top \nabla_\theta^2 \ell(z;\widetilde{\theta})\Delta = \frac{\mu}{2}||\Delta||_{\Sigma_\mathbb{P}}$$

where $\mu = 2/\eta^2$. By Lemma 9, since $\ell(z;\theta)$ is a convex function of θ (as its Hessian is positive semi-definite):

$$
\begin{aligned}
h(\alpha\theta + (1-\alpha)\theta') &= \mathbb{E}_{z\sim\mathbb{P}}\left[\ell\left(z; \alpha\theta + (1-\alpha)\theta'\right)\right] \\
&\leq \mathbb{E}_{z\sim\mathbb{P}}\left[\alpha\ell\left(z;\theta\right) + (1-\alpha)\ell\left(z;\theta'\right) - \frac{\mu}{2}\alpha(1-\alpha)\|\theta-\theta'\|^2_{\Sigma_z}\right] \\
&= \alpha h(\theta) + (1-\alpha)h(\theta') - \frac{\mu}{2}\alpha(1-\alpha)\left(\theta-\theta'\right)^\top \mathbb{E}_{z\sim\mathbb{P}}\left[\Sigma_z\right]\left(\theta-\theta'\right) \\
&= \alpha h(\theta) + (1-\alpha)h(\theta') - \frac{\mu}{2}\alpha(1-\alpha)\|\theta-\theta'\|^2_{\Sigma_{\mathbb{P}}}
\end{aligned}
$$

By Assumption 3, we have that $\Sigma_{\mathbb{P}}$ is positive definite so $\|\cdot\|_{\Sigma_{\mathbb{P}}}$ is a norm. This implies h is μ-strongly convex in the $||\cdot||_{\Sigma_{\mathbb{P}}}$ norm ☐

We now establish strong convexity of $\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$

Lemma 27 (Strong convexity of $\mathcal{L}^{\mathcal{W}_p}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2\lambda/\eta^2$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 27. In Lemma 26, we proved strong convexity of h. By Lemma 9, for $\theta,\theta'\in\Theta$ and $\alpha\in[0,1]$, this is equivalent to

$$
h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \leq \alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||^2_{\Sigma_{\mathbb{P}}}
$$

Taking the supremum over $\mathbb{P}$ preserves the convex combination and the negative quadratic term so we get

$$
\begin{aligned}
\mathcal{L}^{\mathcal{W}_p}\left(\alpha\theta + (1-\alpha)\theta';\varepsilon\right) &= \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \\
&\leq \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \left[\alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||^2_{\Sigma_{\mathbb{P}}}\right] \\
&\leq \alpha\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) + (1-\alpha)\mathcal{L}^{\mathcal{W}_p}\left(\theta';\varepsilon\right) - \frac{\mu}{2}\alpha(1-\alpha)\inf_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} ||\theta-\theta'||^2_{\Sigma_{\mathbb{P}}} \\
&\leq \alpha\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) + (1-\alpha)\mathcal{L}^{\mathcal{W}_p}\left(\theta';\varepsilon\right) - \frac{\mu}{2}\alpha(1-\alpha)\inf_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \lambda_{\min}\left(\Sigma_{\mathbb{P}}\right)||\theta-\theta'||^2_2 \\
&\leq \alpha\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) + (1-\alpha)\mathcal{L}^{\mathcal{W}_p}\left(\theta';\varepsilon\right) - \frac{\mu\lambda}{2}\alpha(1-\alpha)||\theta-\theta'||^2_2
\end{aligned}
$$

where the second inequality holds from $\sup_x\left(f(x)+g(x)\right) \leq \sup_x f(x) + \sup_x g(x)$, the third inequality holds by the fact that $\Sigma_{\mathbb{P}} \succeq \lambda_{\min}\left(\Sigma_{\mathbb{P}}\right)I$, and the last inequality holds from Assumption 3. Thus we conclude that $\mathcal{L}^{\mathcal{W}_p}$ is $\mu\lambda$-strongly convex in the $||\cdot||_2$ norm. ☐

C.2 Proof of Slow Parameter Estimation Rate of WDRO-REBEL

We are now ready to prove the "slow rate" estimation error of Wasserstein-DRO-REBEL.

Proof of Theorem 1. Let $\theta^{\mathcal{W}_p}$ denote the true population minimizer $\arg\min_{\theta\in\Theta}\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right)$ and $\hat{\theta}_n^{\mathcal{W}_p}$ denote the empirical minimizer $\arg\min_{\theta\in\Theta}\mathcal{L}_n^{\mathcal{W}_p}\left(\theta;\varepsilon\right)$. By strong duality for Wasserstein DRO [5], for fixed θ we have

$$
\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] = \inf_{\Delta\geq 0}\left\{\delta\varepsilon^p - \mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}
$$

where $\ell_\Delta(z;\theta) = \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z') - \ell(z';\theta)\right\}$ where d is the metric used to define the type-p Wasserstein distance. Consider the difference between the population and empirical Wasserstein DRO losses:

$$
\begin{aligned}
\left|\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon)-\mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)\right| &= \left|\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]-\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}_n^\circ;\mathcal{W}_p)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]\right| \\
&= \left|\inf_{\Delta\geq 0}\left\{\Delta\varepsilon^p-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}-\inf_{\Delta\geq 0}\left\{\Delta\varepsilon^p-\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}\right| \\
&\leq \sup_{\Delta\geq 0}\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right|
\end{aligned}
$$

where the first equality holds from strong duality and the last inequality holds from $\inf_x f(x)-\inf_x g(x)\leq\sup_x|f(x)-g(x)|$. From Lemma 24, we showed that $\ell(z;\theta)\in[0,K_l]$. Now notice that

$$
\begin{aligned}
l_\Delta(z;\theta) &= \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-\ell(z';\theta)\right\}\leq\inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')\right\}=0 \quad (\text{since } \ell(z;\theta)\geq 0) \\
\ell_\Delta(z;\theta) &= \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-\ell(z';\theta)\right\}\geq\inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-K_\ell\right\}\geq -K_\ell \quad (\text{since } d^p\geq 0)
\end{aligned}
$$

Thus, $\ell_\Delta\in[-K_\ell,0]$. Since ℓ_Δ is bounded and $z\overset{\text{i.i.d}}{\sim}\mathbb{P}_n^\circ$, we can apply Hoeffding's inequality (Lemma 19). For any $\epsilon>0$:

$$
\mathbb{P}\left(\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{2n\epsilon^2}{K_\ell^2}\right)
$$

□

Since the bounds for $\ell_\Delta(z;\theta)$ do not depend on Δ or θ, this bound is uniform over all $\Delta\geq 0$ and $\theta\in\Theta$. By setting the right-hand side to δ and solving for ϵ, we find that with probability at least $1-\delta$:

$$
\left|\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon)-\mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)\right|\leq K_\ell\sqrt{\frac{\log(2/\delta)}{2n}}
$$

Now we have that

$$
\begin{aligned}
&\mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right) \\
&\quad= \mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}_n^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)+\mathcal{L}_n^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)+\mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right) \\
&\quad\leq \left|\mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}_n^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)\right|+\left|\mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)\right| \\
&\quad\leq K_\ell\sqrt{\frac{2\log(2/\delta)}{n}}
\end{aligned}
$$

where the first inequality holds from the fact that $\hat{\theta}_n^{\mathcal{W}_p}\in\arg\min_{\theta\in\Theta}\mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)$. Now from Lemma 10 and Lemma 27, we have that

$$
\frac{\lambda}{\eta^2}\|\theta^{\mathcal{W}_p}-\hat{\theta}_n^{\mathcal{W}_p}\|^2\leq\mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)-\mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)
$$

Thus with probability at least $1-\delta$, we conclude that

$$
\|\theta^{\mathcal{W}_p}-\hat{\theta}_n^{\mathcal{W}_p}\|^2\leq\frac{\eta^2K_g^2}{\lambda}\sqrt{\frac{2\log(2/\delta)}{n}}
$$

D Proof of "Slow Rate" KL-DRO-REBEL

Before we prove the necessary results to get the "slow rate" for KL-DRO-REBEL, we need to make an assumption on the loss functions $\ell(\cdot;\theta)$, $\theta\in\Theta$. Note that this assumption is only used in proving the dual reformulation of the KL-DRO-REBEL objective.

Assumption 4. *We assume that $\ell(z;\theta) \leq L$ for all $\theta \in \Theta$. That is, the loss function is upper bounded by L. In addition, we also assume that Θ permits a uniform upper bound on λ_θ. That is, we assume that*

$$\sup_{\theta \in \Theta} \lambda_\theta < \bar{\lambda}.$$

We state the following dual reformulation result. The proof of this reformulation can be found in [Xu et al., 2025], Appendix C:

Lemma 28 (Dual reformulation of KL-DRO-REBEL)**.** *Let $\ell(z;\theta)$ be the REBEL loss. The KL-DRO-REBEL loss function admits the following dual reformulation:*

$$\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] = \inf_{\lambda \in [\underline{\lambda},\bar{\lambda}]} \left\{ \lambda\varepsilon + \lambda \log\left(\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\exp\left(\frac{\ell(z;\theta)}{\lambda} \right) \right] \right) \right\},$$

where $0 < \underline{\lambda} < \bar{\lambda} < \infty$ are constants.

D.1 Proof of Strong Convexity of KL-DRO-REBEL

We will now establish strong convexity of $\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$. This proof will essentially be the same as Lemma 27

Lemma 29 (Strong convexity of $\mathcal{L}^{\mathrm{KL}}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2\lambda/\eta^2$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 29. In Lemma 26, we proved strong convexity of h. By Lemma 9, for $\theta, \theta' \in \Theta$ and $\alpha \in [0,1]$, this is equivalent to

$$h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \leq \alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||_{\Sigma_\mathbb{P}}^2$$

Taking the supremum over $\mathbb{P}$ preserves the convex combination and the negative quadratic term so we get

$$\begin{aligned}
\mathcal{L}^{\mathrm{KL}}(\alpha\theta + (1-\alpha)\theta';\varepsilon) &= \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \\
&\leq \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \left[\alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||_{\Sigma_\mathbb{P}}^2 \right] \\
&\leq \alpha\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\mathrm{KL}}(\theta';\varepsilon) - \frac{\mu}{2}\alpha(1-\alpha) \inf_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} ||\theta-\theta'||_{\Sigma_\mathbb{P}}^2 \\
&\leq \alpha\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\mathrm{KL}}(\theta';\varepsilon) - \frac{\mu}{2}\alpha(1-\alpha) \inf_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \lambda_{\min}(\Sigma_\mathbb{P})\, ||\theta-\theta'||_2^2 \\
&\leq \alpha\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\mathrm{KL}}(\theta';\varepsilon) - \frac{\mu\lambda}{2}\alpha(1-\alpha)||\theta-\theta'||_2^2
\end{aligned}$$

where the second inequality holds from $\sup_x (f(x) + g(x)) \leq \sup_x f(x) + \sup_x g(x)$, the third inequality holds by the fact that $\Sigma_\mathbb{P} \succeq \lambda_{\min}(\Sigma_\mathbb{P}) I$, and the last inequality holds from Assumption 3. Thus we conclude that $\mathcal{L}^{\mathrm{KL}}$ is $\mu\lambda$-strongly convex in the $||\cdot||_2$ norm. ☐

D.2 Proof of Slow Parameter Estimation Rate of KL-DRO-REBEL

We are now ready to prove the "slow rate" estimation error of KL-DRO-REBEL.

Proof of Theorem 2. By the strong duality result for KL-DRO [28], we have for fixed θ

$$\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] = \inf_{\lambda \in [\underline{\lambda},\bar{\lambda}]} \left\{ \lambda\varepsilon + \lambda \log\left(\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta) \right] \right) \right\},$$

where $j(z,\lambda;\theta)=\exp\left(\frac{l(z;\theta)}{\lambda}\right)$. Then we have

$$\begin{aligned}
\left|\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon)-\mathcal{L}_n^{\mathrm{KL}}(\theta;\varepsilon)\right| &= \left|\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]-\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}_n^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]\right| \\
&= \left|\inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{\lambda\varepsilon+\lambda\log\left(\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right)\right\}-\inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{\lambda\varepsilon+\lambda\log\left(\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]\right)\right\}\right| \\
&\le \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left|\lambda\log\left(\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]\right)-\lambda\log\left(\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right)\right| \\
&= \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\lambda\left|\log\left(\frac{\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]}{\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]}\right)\right| \\
&\le \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\lambda\left|\log\left(\frac{\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right|}{\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]}+1\right)\right| \\
&\le \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\lambda\left|\frac{\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]}{\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]}\right| \\
&\le \bar{\lambda}\sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right|
\end{aligned}$$

where the first equality holds from strong duality, the first inequality holds from $\inf_x f(x)-\inf_x g(x)\le\sup_x|f(x)-g(x)|$, the second inequality holds from $|\log(1+x)|\le|x|\ \forall x\ge 0$, and the last inequality holds from $l(z;\theta)\ge 0$. Next, we establish bounds on $j(z,\lambda;\theta)$:

$$\begin{aligned}
j(z,\lambda;\theta) &= \exp\left(\frac{\ell(z;\theta)}{\lambda}\right)\ge 1 \quad \text{(since } \ell(z;\theta)\ge 0 \text{ and } \lambda\le\bar{\lambda}\text{)} \\
j(z,\lambda;\theta) &= \exp\left(\frac{\ell(z;\theta)}{\lambda}\right)\le \exp\left(\frac{K_\ell}{\underline{\lambda}}\right) \quad \text{(since } \ell(z;\theta)\le K_\ell \text{ and } \lambda\ge\underline{\lambda}\text{)}
\end{aligned}$$

Let $R_j=\exp(K_\ell/\underline{\lambda})-1$ be the range of $j(z,\lambda;\theta)$. For further simplicity, we use the upper bound $R_j\le\exp(K_\ell/\underline{\lambda})$. Since $j(z,\lambda;\theta)$ is bounded within $[1,\exp(K_\ell/\underline{\lambda})]$, we can apply Hoeffding's inequality (Lemma 19). For any $\epsilon>0$:

$$\mathbb{P}\left(\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right|\ge\epsilon\right)\le 2\exp\left(-\frac{2n\epsilon^2}{R_j^2}\right)$$

Since K_ℓ and $\underline{\lambda}$ are independent of λ, R_j is a constant, and this bound is uniform over λ. Picking δ to be the right side and solving for ϵ, we find that with probability at least $1-\delta$:

$$\sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right|\le R_j\sqrt{\frac{\log(2/\delta)}{2n}}$$

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

$$\left|\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon)-\mathcal{L}_n^{\mathrm{KL}}(\theta;\varepsilon)\right|\le\bar{\lambda}R_j\sqrt{\frac{\log(2/\delta)}{2n}}$$

Now we have that

$$\begin{aligned}
&\mathcal{L}^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right) \\
&\quad= \mathcal{L}^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}_n^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)+\mathcal{L}_n^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}_n^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right)+\mathcal{L}_n^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right) \\
&\quad\le \left|\mathcal{L}^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}_n^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right)\right|+\left|\mathcal{L}_n^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right)-\mathcal{L}^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right)\right| \\
&\quad\le \bar{\lambda}R_j\sqrt{\frac{2\log(2/\delta)}{n}}
\end{aligned}$$

where the first inequality holds from the fact that $\hat{\theta}_n^{\mathcal{W}_p} \in \arg\min_{\theta\in\Theta} \mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)$. Now from Lemma 10 and Lemma 28, we have that

$$\frac{\lambda}{\eta^2}\|\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}\|^2 \leq \left|\mathcal{L}^{\mathrm{KL}}\left(\theta^{\mathrm{KL}};\varepsilon\right) - \mathcal{L}^{\mathrm{KL}}\left(\hat{\theta}_n^{\mathrm{KL}};\varepsilon\right)\right|$$

Thus we get

$$\|\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}\|^2 \leq \frac{\eta^2\bar{\lambda}R_j}{\lambda}\sqrt{\frac{2\log(2/\delta)}{n}}$$

Substituting $R_j \leq \exp(K_\ell/\underline{\lambda})$ and $K_\ell = K_g^2$, we conclude that with probability at least $1-\delta$

$$||\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}||_2^2 \leq \frac{\eta^2\bar{\lambda}\exp(K_g^2/\underline{\lambda})}{\lambda}\sqrt{\frac{2\log(2/\delta)}{n}}$$

□

E Proof of "Slow Rate" χ^2-DRO-REBEL

We first state the following dual reformulation for χ^2-DRO

Lemma 30 (Dual reformulation of χ^2-DRO-REBEL)**.** *Let $\ell(z;\theta)$ be the REBEL loss. The χ^2-DRO-REBEL objective admits the dual form*

$$\begin{aligned}\mathcal{L}^{\chi^2}\big(\theta;\varepsilon\big) &= \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon\left(\mathbb{P}^\circ;\chi^2\right)} \mathbb{E}_{z\sim\mathbb{P}}\big[\ell(z;\theta)\big] \\ &= \inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\Big\{\lambda\,\varepsilon \;+\; \mathbb{E}_{z\sim\mathbb{P}^\circ}\big[\ell(z;\theta)\big] \;-\; 2\lambda \;+\; \frac{1}{4\lambda}\,\mathbb{E}_{z\sim\mathbb{P}^\circ}\big[\big(\ell(z;\theta) - \mathbb{E}_{\mathbb{P}^\circ}[\,\ell(z;\theta)\,] + 2\lambda\big)^2\big]\Big\}\end{aligned}$$

where $0 < \underline{\lambda} < \bar{\lambda} < \infty$ are chosen so that the infimum is attained. Equivalently, defining $\mu = \mathbb{E}_{\mathbb{P}^\circ}[\ell(z;\theta)]$ and $\sigma^2 = \mathrm{Var}_{\mathbb{P}^\circ}(\ell(z;\theta))$,

$$\mathcal{L}^{\chi^2}\big(\theta;\varepsilon\big) = \mu \;+\; \inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\Big\{(\varepsilon-1)\,\lambda \;+\; \frac{\sigma^2}{4\lambda}\Big\}.$$

E.1 Proof of Strong Convexity of χ^2-DRO-REBEL

We again will prove strong convexity for χ^2

Lemma 31 (Strong convexity of $\mathcal{L}^{\chi^2}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\chi^2}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2\lambda/\eta$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 31. In Lemma 26, we proved strong convexity of h. By Lemma 9, for $\theta,\theta'\in\Theta$ and $\alpha\in[0,1]$, this is equivalent to

$$h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \leq \alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||_{\Sigma_{\mathbb{P}}}^2$$

Taking the supremum over $\mathbb{P}$ preserves the convex combination and the negative quadratic term so we get

$$\begin{aligned}\mathcal{L}^{\chi^2}(\alpha\theta + (1-\alpha)\theta';\varepsilon) &= \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)} h(\alpha\theta + (1-\alpha)\theta';\mathbb{P}) \\ &\leq \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)} \Big[\alpha h(\theta;\mathbb{P}) + (1-\alpha)h(\theta';\mathbb{P}) - \frac{\mu}{2}\alpha(1-\alpha)||\theta-\theta'||_{\Sigma_{\mathbb{P}}}^2\Big] \\ &\leq \alpha\mathcal{L}^{\chi^2}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\chi^2}(\theta';\varepsilon) - \frac{\mu}{2}\alpha(1-\alpha)\inf_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)}||\theta-\theta'||_{\Sigma_{\mathbb{P}}}^2 \\ &\leq \alpha\mathcal{L}^{\chi^2}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\chi^2}(\theta';\varepsilon) - \frac{\mu}{2}\alpha(1-\alpha)\inf_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)}\lambda_{\min}\left(\Sigma_{\mathbb{P}}\right)||\theta-\theta'||_2^2 \\ &\leq \alpha\mathcal{L}^{\chi^2}(\theta;\varepsilon) + (1-\alpha)\mathcal{L}^{\chi^2}(\theta';\varepsilon) - \frac{\mu\lambda}{2}\alpha(1-\alpha)||\theta-\theta'||_2^2\end{aligned}$$

where the second inequality holds from $\sup_x (f(x)+g(x)) \leq \sup_x f(x) + \sup_x g(x)$, the third inequality holds by the fact that $\Sigma_{\mathbb{P}} \succeq \lambda_{\min}(\Sigma_{\mathbb{P}}) I$, and the last inequality holds from Assumption 3. Thus we conclude that $\mathcal{L}^{\chi^2}$ is $\mu\lambda$-strongly convex in the $||\cdot||_2$ norm. $\square$

E.2 Proof of Slow Parameter Estimation Rate of χ^2-DRO-REBEL

We now prove the "slow rate" estimation error of χ^2-DRO-REBEL

Proof of Theorem 3. Let $\theta^{\chi^2} \in \arg\min_\theta \mathcal{L}^{\chi^2}(\theta;\varepsilon)$ and $\hat{\theta}_n^{\chi^2} \in \arg\min_\theta \mathcal{L}_n^{\chi^2}(\theta;\varepsilon)$. By the dual reformulation (Lemma 30), for any fixed θ

$$\mathcal{L}^{\chi^2}(\theta;\varepsilon) = \mu + \inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\Big\{(\varepsilon-1)\lambda + \frac{\sigma^2}{4\lambda}\Big\},$$

and similarly

$$\mathcal{L}_n^{\chi^2}(\theta;\varepsilon) = \mu_n + \inf_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\Big\{(\varepsilon-1)\lambda + \frac{\sigma_n^2}{4\lambda}\Big\},$$

where $\mu = \mathbb{E}_{\mathbb{P}^\circ}[\ell(z;\theta)]$, $\mu_n = \mathbb{E}_{\mathbb{P}_n^\circ}[\ell(z;\theta)]$, $\sigma^2 = \mathrm{Var}_{\mathbb{P}^\circ}(\ell(z;\theta))$, and $\sigma_n^2 = \mathrm{Var}_{\mathbb{P}_n^\circ}(\ell(z;\theta))$. Using the dual reformulation we have that

$$\begin{aligned}\big|\mathcal{L}^{\chi^2}(\theta;\varepsilon) - \mathcal{L}_n^{\chi^2}(\theta;\varepsilon)\big| \;&=\; \big|\mu-\mu_n\big| \;+\; \big|\inf_\lambda g(\lambda) - \inf_\lambda g_n(\lambda)\big| \\ &\leq \big|\mu-\mu_n\big| \;+\; \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\big|g(\lambda)-g_n(\lambda)\big|,\end{aligned}$$

where $g(\lambda) = (\varepsilon-1)\lambda + \frac{\sigma^2}{4\lambda}$ and $g_n(\lambda) = (\varepsilon-1)\lambda + \frac{\sigma_n^2}{4\lambda}$. The first inequality in the argument above follows from $\inf_x f(x) - \inf_x g(x) \leq \sup_x |f(x)-g(x)|$. From Lemma 24, the loss function $\ell(z;\theta)$ is bounded within $[0, K_\ell]$. We use Hoeffding's inequality (Lemma 19) to bound the deviations of μ_n and σ_n^2 from their population counterparts. For any $\epsilon_1 > 0$ and $\epsilon_2 > 0$:

$$\mathbb{P}\big(|\mu-\mu_n| \geq \epsilon_1\big) \;\leq\; 2\exp\Big(-\frac{2n\epsilon_1^2}{K_\ell^2}\Big),$$

and since $\ell(z;\theta)\in[0,K_\ell]$, $(\ell(z;\theta)-\mu)^2 \in [0, K_\ell^2]$. Thus, the range for the squared terms is K_ℓ^2.

$$\mathbb{P}\big(|\sigma^2-\sigma_n^2| \geq \epsilon_2\big) \;\leq\; 2\exp\Big(-\frac{2n\epsilon_2^2}{K_\ell^4}\Big).$$

Let the desired total failure probability for these bounds be $\delta\in(0,1)$. By a union bound, we require each individual probability bound to be at most $\delta/2$. For $|\mu-\mu_n|$, setting $2\exp\Big(-\frac{2n\epsilon_1^2}{K_\ell^2}\Big) = \frac{\delta}{2}$ yields $\epsilon_1 = K_\ell\sqrt{\frac{\log(4/\delta)}{2n}}$. For $|\sigma^2-\sigma_n^2|$, setting $2\exp\Big(-\frac{2n\epsilon_2^2}{K_\ell^4}\Big) = \frac{\delta}{2}$ yields $\epsilon_2 = K_\ell^2\sqrt{\frac{\log(4/\delta)}{2n}}$. Therefore, by a union bound, with probability at least $1-\delta$, both of the following bounds hold simultaneously:

$$|\mu-\mu_n| \leq K_\ell\sqrt{\frac{\log(4/\delta)}{2n}}$$

$$|\sigma^2-\sigma_n^2| \leq K_\ell^2\sqrt{\frac{\log(4/\delta)}{2n}}$$

Consequently, with probability at least $1-\delta$, we can bound the difference between $g(\lambda)$ and $g_n(\lambda)$:

$$\sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}|g(\lambda)-g_n(\lambda)| = \sup_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left|\frac{\sigma^2-\sigma_n^2}{4\lambda}\right| = \frac{|\sigma^2-\sigma_n^2|}{4\,\underline{\lambda}} \;\leq\; \frac{K_\ell^2}{4\,\underline{\lambda}}\sqrt{\frac{\log(4/\delta)}{2n}},$$

Combining these bounds, with probability at least $1-\delta$:

$$\big|\mathcal{L}^{\chi^2}(\theta;\varepsilon) - \mathcal{L}_n^{\chi^2}(\theta;\varepsilon)\big| \;\leq\; K_\ell\sqrt{\frac{\log(4/\delta)}{2n}} + \frac{K_\ell^2}{4\,\underline{\lambda}}\sqrt{\frac{\log(4/\delta)}{2n}} = K_\ell\Big(1+\frac{K_\ell}{4\underline{\lambda}}\Big)\sqrt{\frac{\log(4/\delta)}{2n}}.$$

Now we analyze the statistical estimation error $\|\theta^{\chi^2} - \hat{\theta}_n^{\chi^2}\|_2^2$. Using the "three-term" decomposition (as in the Wasserstein and KL proof), we get

$$\begin{aligned}\mathcal{L}^{\chi^2}(\hat{\theta}_n^{\chi^2};\varepsilon) - \mathcal{L}^{\chi^2}(\theta^{\chi^2};\varepsilon) &\le \left|\mathcal{L}^{\chi^2}(\hat{\theta}_n^{\chi^2};\varepsilon) - \mathcal{L}_n^{\chi^2}(\hat{\theta}_n^{\chi^2};\varepsilon)\right| + \left|\mathcal{L}_n^{\chi^2}(\theta^{\chi^2};\varepsilon) - \mathcal{L}^{\chi^2}(\theta^{\chi^2};\varepsilon)\right| \\ &= K_\ell\Big(1 + \frac{K_\ell}{4\underline{\lambda}}\Big)\sqrt{\frac{2\log(4/\delta)}{n}}.\end{aligned}$$

Finally, by strong convexity of $\mathcal{L}^{\chi^2}$ (cf. Lemma 10),

$$\frac{\lambda}{\eta^2}\left\|\theta^{\chi^2} - \hat{\theta}_n^{\chi^2}\right\|^2 \;\le\; \mathcal{L}^{\chi^2}(\theta^{\chi^2};\varepsilon) - \mathcal{L}^{\chi^2}(\hat{\theta}_n^{\chi^2};\varepsilon),$$

Thus with probability at least $1-\delta$, we conclude

$$\|\theta^{\chi^2} - \hat{\theta}_n^{\chi^2}\|^2 \le \frac{\eta^2 K_g^2}{\lambda}\left(1 + K_g^2/4\underline{\lambda}\right)\sqrt{\frac{2\log(4/\delta)}{n}}.$$

□

F Proof of "Master Theorem" for Parametric $n^{-1/2}$ rates

Proof of Theorem 4. The proof proceeds in three main parts. First, we establish a basic inequality that relates the true excess risk of the estimator $\hat{\theta}_n$ to an empirical process term and the remainder Δ_n. Second, we use a peeling argument combined with localized Rademacher complexity to uniformly bound this empirical process and subsequently a high-probability bound on the excess risk $\varepsilon_\ell(\hat{\theta}_n)$. Finally, we use the strong convexity assumption to convert this risk bound into the final bound on the parameter error $\|\hat{\theta}_n - \theta^\star\|_2$.

Let us define our key quantities:

Empirical and Population Robust Loss Minimizers: Let $\hat{\theta}_n = \arg\min_{\theta\in\Theta}\mathcal{L}_n^D\big(\theta;\varepsilon_n\big)$ be the empirical robust loss minimizer and $\theta^\star = \arg\min_{\theta\in\Theta}\mathcal{L}^D\big(\theta;\varepsilon_n\big)$ be the population robust loss minimizer.

Nominal Risks: Let $\widetilde{R}(\theta) := \mathbb{E}_{\mathbb{P}^\circ}[\ell(z;\theta)]$ denote the population risk and $R_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(z_i;\theta)$ be the empirical risk.

Excess Risk: We define the excess risk of a parameter θ as $\epsilon_\ell(\theta) := R(\theta) - \inf_{\theta\in\Theta} R(\theta)$.

Centered Loss Class: The class of centered loss functions is $\mathcal{F} := \{f_\theta : z \mapsto \ell(z;\theta) - \ell(z;\theta^\star) \mid \theta\in\Theta\}$.

Scaled Loss Class: The scaled loss class is defined as $\mathcal{G} := \{g_\theta(z) = 4^{-k_\theta} f_\theta(z) \mid \theta\in\Theta\}$.

Empirical Process: For a function g, we use the notation $(\mathbb{P}-\mathbb{P}_n)g := \mathbb{E}_{\mathbb{P}^\circ}[g(z)] - \frac{1}{n}\sum_{i=1}^n g(z_i)$.

Shell Decomposition: For any real value $r>0$, we define a shell index k_θ for each $\theta\in\Theta$ as

$$k_\theta = \inf\left\{k\in\mathbb{Z}_{\ge 0} : \epsilon_\ell(\theta) \le r\cdot 4^k\right\}$$

For the remainder of this proof, we will assume that ε_n is fixed and will denote $\mathcal{L}^D(\theta) = \mathcal{L}^D(\theta;\varepsilon_n)$.

Part 1: The Basic Inequality Let $\Delta_n(\theta) := \mathcal{L}_n^D(\theta) - R_n(\theta)$. Then we can decompose the excess risk as follows:

$$
\begin{aligned}
\epsilon_\ell(\hat{\theta}_n) &= R(\hat{\theta}_n) - \inf_{\theta\in\Theta} R(\theta) \\
&= \mathcal{L}^D(\hat{\theta}_n) - \Delta(\hat{\theta}_n) - \mathcal{L}^D(\theta^*) + \Delta(\theta^*) \\
&= \mathcal{L}^D(\hat{\theta}_n) - \mathcal{L}_n^D(\hat{\theta}_n) + \mathcal{L}_n^D(\hat{\theta}_n) - \mathcal{L}_n^D(\theta^*) + \mathcal{L}_n^D(\theta^*) - \mathcal{L}^D(\theta^*) + \Delta(\theta^*) - \Delta(\hat{\theta}_n) \\
&\leq \mathcal{L}^D(\hat{\theta}_n) - \mathcal{L}_n^D(\hat{\theta}_n) + \mathcal{L}_n^D(\theta^*) - \mathcal{L}^D(\theta^*) + \Delta(\theta^*) - \Delta(\hat{\theta}_n) \\
&= R(\hat{\theta}_n) + \Delta(\hat{\theta}_n) - R_n(\hat{\theta}_n) - \Delta_n(\hat{\theta}_n) + R_n(\theta^*) + \Delta_n(\theta^*) - R(\theta^*) - \Delta(\theta^*) + \Delta(\theta^*) - \Delta(\hat{\theta}_n) \\
&= R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\theta^*) - R(\theta^*) + \Delta_n(\theta^*) - \Delta_n(\hat{\theta}_n) \\
&\leq (\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_n} + 2\Delta_n \\
&= 4^{k_\theta} (\mathbb{P} - \mathbb{P}_n) g_{\hat{\theta}_n} + 2\Delta_n
\end{aligned}
$$

where the first inequality holds by definition of $\hat{\theta}_n$ and the last inequality holds by assumption of the dual remainder bound. This inequality connects the quantity we want to bound, $\epsilon_\ell(\hat{\theta}_n)$, to the empirical process $(\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_n}$. Note that if $k_\theta = 0$, then $\epsilon_\ell(\theta) \leq r$. Otherwise if $k_\theta \geq 1$, then by definition

$$
r \cdot 4^{k_\theta - 1} \leq \epsilon_\ell(\theta) \leq r \cdot 4^{k_\theta} \iff \frac{\epsilon_\ell(\theta)}{r} \leq 4^{k_\theta} \leq \frac{4\epsilon_\ell(\theta)}{r}
$$

For sake of argument, assume that $(\mathbb{P} - \mathbb{P}_n) g_{\hat{\theta}_n} \leq \Upsilon$. Then the basic inequality implies $\epsilon_\ell(\hat{\theta}_n) \leq 4^{k_\theta}\Upsilon + 2\Delta_n$. Consider cases

- if $k_\theta = 0$, then $\epsilon_\ell(\hat{\theta}_n) \leq \Upsilon + 2\Delta_n$
- if $k_\theta \geq 1$, then $\epsilon_\ell(\hat{\theta}_n) \leq \frac{4\epsilon_\ell(\hat{\theta}_n)}{r}\Upsilon + 2\Delta_n$
- if $\Upsilon \leq \frac{r}{8}$, then $\frac{1}{2}\epsilon_\ell(\hat{\theta}_n) \leq 2\Delta_n$

We can combine these into one uniform bound $\frac{1}{2}\epsilon_\ell(\hat{\theta}_n) \leq 2\Delta_n + \Upsilon$. From this, we see that it is sufficient to prove that $(\mathbb{P} - \mathbb{P}_n) g_{\hat{\theta}_n} \leq \Upsilon$ holds with probability $1 - \delta$ uniformly across all excess risk shells.

Part 2: Bounding The Empirical Process Let us define a sequence of radii $r_k = 4^k r_0$ for $k \in \mathbb{Z}_{\geq 0}$, where r_0 is a baseline risk level to be determined later. We partition the parameter space Θ into disjoint shells based on these radii:

$$
\Theta_k := \begin{cases} \{\theta \in \Theta \mid \varepsilon_\ell(\theta) \leq r_0\} & \text{if } k = 0 \\ \{\theta \in \Theta \mid r_{k-1} < \varepsilon_\ell(\theta) \leq r_k\} & \text{if } k \geq 1 \end{cases}
$$

Corresponding to each parameter shell Θ_k, we define a function class $\mathcal{G}_k = \{g_\theta \mid \theta \in \Theta_k\}$. For a fixed shell Θ_k, we apply a standard uniform deviation bound based on Rademacher complexity (Theorem 13). With probability at least $1 - \delta_k$, the following holds for all $g \in \mathcal{G}_k$:

$$
(P - P_n) g \leq 2 \underbrace{\mathfrak{R}_n(\mathcal{G}_k)}_{\text{(a)}} + 4 \underbrace{\sup_{g\in\mathcal{G}_k} \|g\|_\infty}_{\text{(b)}} \sqrt{\frac{2\ln(1/\delta_k)}{n}} \tag{2.1}
$$

where $\mathfrak{R}_n(\mathcal{G}_k) = \mathbb{E}_{\sigma,Z}[\sup_{g\in\mathcal{G}_k} \frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)]$ is the empirical Rademacher complexity with $\sigma_i \overset{\text{i.i.d}}{\sim} \text{Unif}\{-1, 1\}$.

We will now compute (a) and (b). Before doing this, we will derive a parameter estimate deviation bound that we will use in our subsequent analysis. By assumption of the dual remainder bound, we know $\left|\mathcal{L}^D(\theta) - R(\theta)\right| \leq \Delta_n$. Therefore

$$
\begin{aligned}
R(\theta) - \inf_{\theta\in\Theta} R(\theta^*) &\geq \mathcal{L}^D(\theta) - \Delta_n - \inf_{\theta\in\Theta} R(\theta) \\
&\geq \mathcal{L}^D(\theta) - \mathcal{L}^D(\theta^*) - 2\Delta_n \\
&\geq \frac{\alpha}{2}\|\theta - \theta^*\|_2^2 - 2\Delta_n
\end{aligned}
$$

where the last inequality holds from the assumption that $\mathcal{L}^D$ is α-strongly convex around θ^*. Therefore, we deduce that

$$\begin{aligned}\|\theta-\theta^*\|_2 &\leq \sqrt{\frac{2}{\alpha}\left(R(\theta)-\inf_{\theta\in\Theta}R(\theta^*)+2\Delta_n\right)}\\ &= \sqrt{\frac{2}{\alpha}\left(\epsilon_\ell(\theta)+2\Delta_n\right)} \qquad (\star)\end{aligned}$$

Computing (b): Take $g\in\mathcal{G}$. Then by definition

$$\begin{aligned}\|f\|_\infty &= 4^{-k_\theta}\left|\ell(z;\theta)-\ell(z;\theta^*)\right|\\ &= 4^{-k_\theta}\left|\phi\left(h_\theta(z),z\right)\right)-\phi\left(h_{\theta^*}(z),z\right))\right|\\ &\leq 4^{-k_\theta}L_\phi\|h_\theta(z),z)-h_{\theta^*}(z),z)\|_2\\ &= 4^{-k_\theta}L_\phi|\|v(z)\|_2\|\theta-\theta^*\|_2\\ &\leq 4^{-k_\theta}L_\phi\sqrt{\frac{2}{\alpha}\left(\epsilon_\ell(\theta)+2\Delta_n\right)}\\ &\leq 4^{-k_\theta}L_\phi\sqrt{\frac{2}{\alpha}\left(r\cdot 4^{k_\theta}+2\Delta_n\right)}\\ &\leq 2^{-k_\theta}L_\phi\sqrt{\frac{2r}{\alpha}}+4^{-k_\theta}L_\phi\sqrt{\frac{4\Delta_n}{\alpha}}\\ &\leq 2L_\phi\sqrt{\frac{r+2\Delta_n}{\alpha}}\end{aligned}$$

where the first inequality holds from $\ell(z;\theta)$ being L_ϕ-Lipschitz in θ, the second holds from $h_\theta(z)$ being a linear function of $v(z)$ where $\|v(z)\|_2\leq 1$ and ($\star$), the third holds by definition of k_θ, the fourth from the fact that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, and the last from the fact that $2^{-k_\theta},4^{-k_\theta}\leq 1$ and $\sqrt{a}+\sqrt{b}\leq\sqrt{2(a+b)}$.

Computing (a): Next, we will use a "peeling" argument to get localization around the shells of $\mathcal{F}$. Define a countable sequence of function classes $(\mathcal{G}_k)_{k\geq 0}$ such that $\mathcal{F}\subseteq\bigcup_k\mathcal{F}_k$. Define $\mathcal{F}_k$ as follows

$$\mathcal{F}_k=\left\{4^{-k}f_\theta:\theta\in\Theta_k\right\}$$

Notice that across $\mathcal{F}_k$, $\epsilon_l(\theta)\leq r\cdot 4^{k_\theta}\leq r\cdot 4^k$. We now compute the Rademacher complexity of $\mathcal{F}_k$.

$$\begin{aligned}\mathfrak{R}_n\left(\mathcal{F}_k\right) &\leq 4^{-k}\mathfrak{R}_n\left(\left\{f_\theta:\theta\in\Theta,\epsilon_\ell(\theta)\leq r\cdot 4^k\right\}\right)\\ &\leq 4^{-k}\mathfrak{R}_n\left(\left\{z\mapsto\ell(z;\theta)-\ell(z;\theta^*):\theta\in\Theta,\|\theta-\theta^*\|_2\leq\sqrt{\frac{2}{\alpha}\left(r\cdot 4^k+2\Delta_n\right)}\right\}\right)\\ &= 4^{-k}\mathfrak{R}_n\left(\left\{z\mapsto\phi\left(h_\theta(z),z\right)\right)-\phi\left(h_{\theta^*}(z),z\right)):\theta\in\Theta,\|\theta-\theta^*\|_2\leq\sqrt{\frac{2}{\alpha}\left(r\cdot 4^k+2\Delta_n\right)}\right\}\right)\\ &\leq 4^{-k}L_\phi\mathfrak{R}_n\left(\left\{z\mapsto\langle\theta-\theta^*,z\rangle:\theta\in\Theta,\|\theta-\theta^*\|_2\leq\sqrt{\frac{2}{\alpha}\left(r\cdot 4^k+2\Delta_n\right)}\right\}\right)\\ &\leq 4^{-k}L_\phi\mathfrak{R}_n\left(\left\{z\mapsto\langle\theta,z\rangle:\theta\in\Theta+\left\{\theta^*\right\},\|\theta\|_2\leq\sqrt{\frac{2}{\alpha}\left(r\cdot 4^k+2\Delta_n\right)}\right\}\right)\\ &\leq 4^{-k}L_\phi\sqrt{\frac{2}{\alpha n}\left(r\cdot 4^k+2\Delta_n\right)}\end{aligned}$$

where the second inequality holds definition of g_θ coupled with the fact that $\left\{\theta\in\Theta:\epsilon_\ell(\theta)\leq r\cdot 4^k\right\}\subset\left\{\theta\in\Theta:\|\theta-\theta^*\|_2\leq\sqrt{2/\alpha\left(r\cdot 4^k+2\Delta_n\right)}\right\}$ and the monotonicity of the Rademacher complexity (Lemma 14), the third inequality holds from ϕ being L_ϕ Lipschitz and Ledoux-Talagrand contraction lemma (Lemma 12), and the last inequality holds from a standard result about the Rademacher complexity of a linear predictor under the ℓ_2 metric (Lemma 16)

Then by the monotonicity and subaddivity of the Rademacher complexity (Lemma 15), we have

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{G}) &\leq \sum_{k=0}^{\infty} \mathfrak{R}_n(\mathcal{F}_k) \\
&\leq \frac{\sqrt{2}L_\phi}{\sqrt{\alpha n}} \sum_{k=0}^{\infty} \sqrt{r \cdot 4^{-k} + 2 \cdot 4^{-2k}\Delta_n} \\
&\leq \frac{\sqrt{2}L_\phi}{\sqrt{\alpha n}} \left(\sqrt{r} \sum_{k=0}^{\infty} 2^{-k} + \sqrt{2\Delta_n} \sum_{k=0}^{\infty} 4^{-k} \right) \\
&= \frac{\sqrt{2}L_\phi}{\sqrt{\alpha n}} \left(2\sqrt{r} + \frac{4}{3}\sqrt{2\Delta_n} \right) \\
&= \frac{2\sqrt{2}L_\phi}{\alpha n}\sqrt{r} + \frac{8L_g}{3\sqrt{\alpha n}}\sqrt{\Delta_n}
\end{aligned}
$$

where the third inequality holds from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Putting (a) and (b) together, we find that

$$
\begin{aligned}
\Upsilon &\leq \frac{4\sqrt{2}L_\phi}{\sqrt{\alpha n}}\sqrt{r} + \frac{16L_\phi}{3\sqrt{\alpha n}}\sqrt{\Delta_n} + 8L_\phi\sqrt{\frac{r+2\Delta_n}{\alpha}}\sqrt{\frac{2\log(4/\delta)}{n}} \\
&= \frac{L_\phi}{\sqrt{\alpha n}} \left(4\sqrt{2}\sqrt{r} + \frac{16}{3}\sqrt{\Delta_n} + 8\sqrt{2}\sqrt{(r+2\Delta_n)\log(4/\delta)} \right) \\
&\leq \frac{12L_\phi}{\sqrt{\alpha n}} \left(\sqrt{r+\Delta_n} + \sqrt{(r+\Delta_n)\log(4/\delta)} \right) \\
&\leq \frac{18L_\phi}{\sqrt{\alpha n}} \sqrt{(r+\Delta_n)\log(4e/\delta)}
\end{aligned}
$$

where the second inequality holds by $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ and upper bounding the constants by a common one, namely 12. The last inequality holds from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and again upper bounded by a common constant. Now we'll try to satisfy $\Upsilon \leq r/8$ or equivalently $8\Upsilon^2 \leq r^2$. In our case we take

$$
r \geq \frac{CL_\phi^2 \log(4e/\delta)}{\alpha n} + \frac{\Delta_n}{2}
$$

for a universal constant $C > 0$[1]. We can conclude that with probability atleast $1 - \delta$,

$$
\frac{1}{2}\epsilon_\ell(\hat{\theta}_n) \leq 2\Delta_n + \frac{CL_\phi^2 \log(4e/\delta)}{\alpha n} + \frac{\Delta_n}{2}
$$

Part 3: Converting From Excess Risk To Parameter Estimation Now recall that from $(\star)$,

$$
\|\theta - \theta^*\|_2 \leq \sqrt{\frac{2}{\alpha}\left(\epsilon_\ell(\theta) + 2\Delta_n\right)}
$$

Our high probability bound implies with probability atleast $1 - \delta$

$$
\varepsilon_\ell(\hat{\theta}_n) \ \leq \ 4\Delta_n + \frac{2C\, L_\phi^2 \, \log(4e/\delta)}{\alpha n}
$$

[1]Any universal constant $C \geq 4.2 \times 10^4$ will work

Thus we get that we get

$$\left\|\hat{\theta}_n - \theta^\star\right\|_2 \le \sqrt{\frac{2}{\alpha}\Big(4\Delta_n + \frac{2C\,L_\phi^2\,\log(4e/\delta)}{\alpha n} + 2\Delta_n\Big)}$$
$$= \sqrt{\frac{2}{\alpha}\Big(6\Delta_n + \frac{2C\,L_\phi^2\,\log(4e/\delta)}{\alpha n}\Big)}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ yields

$$\left\|\hat{\theta}_n - \theta^\star\right\|_2 \;\le\; \sqrt{\frac{12}{\alpha}}\,\sqrt{\Delta_n} \;+\; 2\sqrt{\frac{C}{\alpha^2}}\,\frac{L_\phi}{\sqrt{n}}\,\sqrt{\log\frac{4e}{\delta}}.$$

Defining universal numerical constants $C_1 = 2\sqrt{C/\alpha},\ C_2 = \sqrt{12/\alpha}$, we can state:

$$\left\|\hat{\theta}_n - \theta^\star\right\|_2 \;\le\; C_1\,\frac{L_\phi}{\alpha}\sqrt{\frac{\log(4e/\delta)}{n}} \;+\; C_2\sqrt{\frac{\Delta_n}{\alpha}},$$

Assuming the DRO approximation error Δ_n (which reflects the inherent gap between the true and nominal risks in the ambiguity set) decays as $O(n^{-1})$, every term on the right-hand side is of order $O(n^{-1/2})$, which concludes the proof.

□

G Proof of Minimax Theorem for Robust RLHF Problems Under "Master" Theorem Assumptions

Proof of Theorem 5. The proof will proceed in two main parts. In the first part, we will establish that there is an irreducible statistical error term of order $\Omega(n^{-1/2})$ by constructing the classic parameter separation construction and invoking Le Cam's two-point lemma. In the second part, we will establish an irreducible residual error (bias) term introduced by the DRO formulation of order $\Omega\left(\sqrt{\Delta_n/\alpha}\right)$. We will do this by constructing an example of loss function that satisfies the dual remainder bound assumption of the "Master" theorem and is strongly convex but whose minimizer is far from the true minimizer.

Let

$$\ell(z;\theta) = \phi\big(h_\theta(z), z\big), \qquad h_\theta(z) = \theta^\top \nu(z).$$

Define the *DRO population loss* and *Population Risk*

$$\mathcal{L}^D(\theta;\varepsilon_n) \;=\; \sup_{P\in\mathcal{B}_{\varepsilon_n}(P^\circ;D)} \mathbb{E}_{Z\sim P}\big[\ell(Z;\theta)\big], \qquad \mathcal{R}(\theta) \;=\; \mathbb{E}_{Z\sim P^\circ}\big[\ell(Z;\theta)\big].$$

Throughout we assume

(A1) Local strong convexity. There exist constants $\rho > 0$ and $\alpha > 0$ such that $\mathcal{L}^D(\cdot;\varepsilon_n)$ is α–strongly convex on the ball $B_\rho(\theta^\star) := \{\theta \in \Theta : \|\theta - \theta^\star\|_2 \le \rho\}$, where $\theta^\star \in \arg\min_\theta \mathcal{R}(\theta)$.

(A2) Linear margin. $\|v(z)\|_2 \le 1$ for every $z \in \mathcal{Z}$.

(A3) Lipschitz loss in the margin. $|\phi(u,z) - \phi(u',z)| \le L_\phi\,|u - u'| \quad \forall u, u' \in \mathbb{R},\ z \in \mathcal{Z}$.

(A4) Dual–remainder bound. $\big|\mathcal{L}^D(\theta;\varepsilon_n) - \mathcal{R}(\theta)\big| \le \Delta_n \quad \forall \theta \in \Theta$.

Fix $(\alpha, L_\phi, \Delta_n, \rho)$ and set the admissible class as

$$\mathcal{M}(\alpha, L_\phi, \Delta_n, \rho) := \big\{(\mathbb{P}^\circ, \phi, v)\ \text{ : conditions (A1)–(A4) hold}\big\}.$$

Part 1: Lower bound arising from statistical error First notice that for any $\epsilon > 0$, we have the following

$$\underbrace{\{\mathbb{P} \in \mathcal{M}(\mathcal{Z}) : \mathbb{P} \in \mathcal{B}_0\left(\mathbb{P}^\circ; D\right)\}}_{\mathcal{C}'} \subseteq \underbrace{\{\mathbb{P} \in \mathcal{M}(\mathcal{Z}) : \mathbb{P} \in \mathcal{B}_\varepsilon\left(\mathbb{P}^\circ; D\right)\}}_{\mathcal{C}}$$

Therefore, $\inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathcal{C}} \mathcal{R}(\theta;\mathbb{P}) \geq \inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathcal{C}'} \mathcal{R}(\theta;\mathbb{P})$. Thus, it is sufficient to consider the non-robust case for obtaining the lower bound arising from statistical noise. Intuitively, since the DRO problem is more complex (for $\varepsilon > 0$), it cannot possibly overcome this fundamental statistical limitation that comes from simpler parametric problem. Now since Δ_n is a function of ε_n, we can assume that $\Delta_n = O(1)$. Without loss of generality, we can assume that $\Delta_n = 0$. From the dual remainder assumption, this implies $\mathcal{L}^D(\theta;\varepsilon_n) = \mathcal{R}(\theta;\mathbb{P})$.

Consider $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \mathcal{R}_n(\theta;\mathbb{P}_n^\circ)$. This is the standard problem in M-estimation where $\mathcal{R}_n(\theta;\mathbb{P}_n^\circ) = \frac{1}{n}\sum_{i=1}^n \ell(z_i;\theta)$ where $z_i \overset{\text{i.i.d}}{\sim} \mathbb{P}^\circ$. Let $u \in \mathbb{R}^d$ be a fixed unit vector and $\delta_n > 0$ be some separation as a function of n. We will choose this in our analysis. We now will construct two hypotheses H_0, H_1

$$H_0 : \theta_0 = 0\ ,\quad H_1 : \theta_1 = \delta_n u$$

Notice that $\|\theta_1 - \theta_0\| = \delta_n$. Consider the loss function $\ell(z;\theta) = \phi\big(h_\theta(z), z\big)$. By assumption, this function is L_ϕ Lipschitz and the robust DRO loss (or in this case the risk) is α-strongly convex.

To make these hypotheses concrete, we define a simple data-generating process that will follow a parametric linear form. This is to adhere to the Linear Margin assumption of the "Master" theorem. Let $z_i = (\nu_i, y_i)$ be generated from the true model $y_i = \theta^\top \nu_i + \xi_i$ where ν_i are i.i.d random vectors with $\|\nu_i\| \leq 1$ and $\mathbb{E}\left[\nu_i\nu_i^\top\right] \succeq \alpha \mathrm{I}_d$ for some $\alpha > 0$ and $\xi_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma^2)$ where $\sigma^2 = L_\phi/2\alpha$. One way to think of this concretely is that ν_i is fundamentally a function of the prompt x_i and the action a as it a feature map. Also note that from this data-generating distribution, we have that conditional on ν_i

$$\begin{aligned}
&\text{under } H_0 : y_i \sim \mathcal{N}\left(0,\sigma^2\right)\\
&\text{under } H_1 : y_i \sim \mathcal{N}\left(\delta_n u^\top \nu_i,\sigma^2\right)
\end{aligned}$$

We show that these two hypotheses are hard to distinguish based on n samples. We will do this by bounding the Kullback-Leibler (KL) divergence between the joint distributions $\mathbb{P}_0^n, \mathbb{P}_1^n$.

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}_1^n || \mathbb{P}_0^n\right) &= n D_{\mathrm{KL}}\left(\mathbb{P}_1 || \mathbb{P}_0\right)\\
&= n\mathbb{E}_\nu\left[D_{\mathrm{KL}}\left(\mathcal{N}\left(\delta_n u^\top\nu_i,\sigma^2\right) || \mathcal{N}\left(0,\sigma^2\right)\right)\right]\\
&= n\mathbb{E}_\nu\left[\frac{\left(\delta_n u^\top \nu\right)^2}{2\sigma^2}\right]\\
&= \frac{n\delta_n^2}{2\sigma^2} u^\top \mathbb{E}_\nu\left[\nu\nu^\top\right] u\\
&= \frac{n\delta_n^2}{2\sigma^2} u^\top \Sigma u
\end{aligned}$$

where the first equality holds from tensorization of the KL divergence (Lemma 21) and third equality holds from computing the KL Divergence of two univariate normal distributions (Lemma 22). Choose $\delta_n = \sigma^2/(4nu^\top\Sigma u)$ so $\delta_n = \Omega(n^{-1/2})$ and $D_{\mathrm{KL}}\left(\mathbb{P}_1^n||\mathbb{P}_0^n\right) = 1/8$. Now we can invoke Le Cam's two-point lemma (Lemma 23)

$$\begin{aligned}
\inf_{\hat{\theta}_n} \sup_{\theta\in\{\theta_0,\theta_1\}} \mathbb{E}_\theta\left[\|\hat{\theta}_n - \theta\|^2\right] &\geq \frac{\|\theta_1 - \theta_0\|_2}{2}\left(1 - d_{\mathrm{TV}}(\mathbb{P}_1^n, \mathbb{P}_0^n)\right)\\
&\geq \frac{\|\theta_1 - \theta_0\|_2}{2}\left(1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathbb{P}_1^n||\mathbb{P}_0^n\right)}\right)\\
&= \frac{\delta_n}{2}\left(1 - \frac{1}{4}\right)\\
&= \Omega\left(\frac{L_\phi}{\alpha}\sqrt{\frac{1}{n}}\right)
\end{aligned}$$

where the second inequality holds from Pinsker's inequality $d_{\mathrm{TV}}^2 \le \frac{1}{2} D_{\mathrm{KL}}$ (Theorem 15) and the last equality holds from taking $\delta_n = \sigma^2/(4nu^\top \Sigma u)$ and $D_{\mathrm{KL}}\left(\mathbb{P}_1^n || \mathbb{P}_0^n\right) = 1/8$. Thus we conclude that

$$
\begin{aligned}
\inf_{\theta \in \Theta} \sup_{\mathbb{P} \in \mathcal{C}} \mathcal{R}(\theta; \mathbb{P}) &\ge \inf_{\hat{\theta}_n} \sup_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[\|\hat{\theta}_n - \theta\|^2 \right] \\
&\ge \Omega\left(\frac{L_\phi}{\alpha} \sqrt{\frac{1}{n}} \right)
\end{aligned}
$$

This completes the proof of the irreducible statistical error.

Part 2: Lower bound arising from residual error The other bound arises from the inherent ambiguity introduced by the DRO formulation. First let the data-generating distribution $\mathbb{P}^\circ$ be the point mass at a single observation

$$
z_0 = (\nu_0, y_0), \ \nu_0 := e_1 \in \mathbb{R}^d, \ y_0 = 0
$$

By construction $\|\nu_0\|_2 = 1$. Define a globally L_ϕ Lipschitz function but locally strongly convex scalar loss function $\psi : \mathbb{R} \to \mathbb{R}$ as the (scaled) Huberised quadratic:

$$
\psi(u) = \begin{cases} \frac{\alpha}{2} u^2 & |u| \le r_0 \\ \alpha r_0 |u| - \frac{\alpha}{2} r_0^2 & |u| > r_0 \end{cases}
$$

where $r_0 := \min\{L_\phi/\alpha, \rho\}$. Then notice that $|\psi'(u)| \le \alpha r_0 \le L_\phi$ for all u and ψ is locally α-strongly convex on $(-r_0, r_0)$. Set $\phi(u, z) := \psi(u)$ for all $z \in \mathcal{Z}$. Then

$$
\begin{aligned}
\mathcal{R}(\theta) &= \ell(z_0; \theta) \\
&= \psi(\theta^\top e_1) \\
&= \psi(\theta_1)
\end{aligned}
$$

Note that $\psi(u)$ attains a unique minimizer at $u = 0$. Therefore, the (non-robust) population minimizer is $\theta^* = 0$. Moreover, on the ball $\{\theta : \|\theta\|_2 \le r_0\} \subseteq \mathcal{B}_\rho(\theta^*)$, $\mathcal{R}$ is α-strongly convex in θ along the e_1 directions. We will ensure that the DRO objective is α-strongly convex on $\{\theta : \|\theta\|_2 \le r_0\}$.

Remark 5. *We used the Huberised quadratic to ensure that (A3) and (A1) hold. Any L_ϕ-Lipschitz ψ with $\psi''(0) \ge \alpha$ would work*

Now choose

$$
\delta := \min\left\{ \rho, \frac{r_0}{2}, \sqrt{\frac{2\Delta_n}{\alpha}} \right\}, \quad \beta := \frac{\alpha \delta}{2}
$$

We also define

$$
g_\pm(\theta) := \pm \min\left\{ \beta(\theta^\top e_1), \Delta_n \right\}
$$

Notice that for any $\theta \in \Theta$,

$$
\begin{aligned}
|g_\pm(\theta)| &\le \min\{\beta |\theta_1|, \Delta_n\} \\
&\le \beta |\theta_1| \\
&\le \beta \delta \\
&= \frac{\alpha \delta^2}{2} \\
&\le \Delta_n
\end{aligned}
$$

where the third inequality holds since $\{\theta : \|\theta\|_2 \le \delta\} \subseteq \{\theta : \|\theta\|_2 \le r_0\}$ and the last inequality holds from the definition of δ. Therefore the dual remainder bound (A4) will hold for the perturbed objectives defined next.

Now we define the following admissible DRO objective

$$\mathcal{L}^{\pm}(\theta) := \mathcal{R}(\theta) + g_{\pm}(\theta) = \psi(\theta_1) \pm \min\left\{\beta(\theta^\top e_1), \Delta_n\right\}$$

We will see why this DRO objective is admissible. On the slab $|\theta_1| \le \delta$, we have $g_{\pm}(\theta) = \pm\beta\theta_1$ which is a convex function. Hence on the ball $\mathcal{B}_\delta(\theta^*) \subseteq \mathcal{B}_\rho(\theta^*) \cap \{|\theta_1| \le r_0\}$, $\mathcal{L}^{\pm}(\theta) = \psi(\theta_1) \pm \beta\theta_1$ is the sum of a strongly convex function and a convex function which must be strongly convex. Therefore, $\mathcal{L}^{\pm}$ is α-strongly convex on $\mathcal{B}_\delta(\theta^*)$. Thus Assumption (A1) holds.

Now notice that both the plus and minus DRO objectives use the same base data-generating distribution $\mathbb{P}^\circ$; only the DRO objective is perturbed by $g_{\pm}$. Therefore the sample $(Z_1, \ldots, Z_n)$ is identical under both objectives so $\mathcal{P}_n^+ = \mathcal{P}_n^-$. Thus the distributions are statistically indistinguishable, resulting in $d_{\mathrm{TV}}(\mathcal{P}_n^+, \mathcal{P}_n^-) = 0$.

We will restrict to the line $\theta = \xi\theta_1$ with $|\xi| \le \delta$ (which contains both minimizers by strong convexity and symmetry). For $|\xi| \le \delta$,

$$\mathcal{L}^{\pm}(\xi e_1) = \psi(\xi) \pm \beta\xi,\ \nabla_\xi \mathcal{L}^{\pm}(\xi e_1) = \psi'(\xi) \pm \beta$$

Inside $|\xi| \le \delta \subseteq (-r_0/2, r_0/2)$. we have $\psi'(\xi) = \alpha\xi$. Setting the gradient to 0 and solving to ξ, we find that

$$\theta_+^* = \frac{\beta}{\alpha}e_1 = \frac{\delta}{2}e_1,\ \ \theta_+^* = -\frac{\beta}{\alpha}e_1 = -\frac{\delta}{2}e_1$$

Both lie inside $|\xi| \le \delta$ and have separation

$$\|\theta_+^* - \theta_-^*\| = \delta \ge \sqrt{\frac{2\Delta_n}{\alpha}}\mathbf{1}_{\left\{\sqrt{2\Delta_n/\alpha} \le \min\{\rho, r_0/2\}\right\}}$$

Since θ_+^*, θ_-^* belong to $\mathcal{P}_n^+, \mathcal{P}_n^-$ which are statistically indistinguishable, by Le Cam's two-point lemma, we have for *every* estimator $\widehat{\theta}_n$,

$$\max\Big\{\mathbb{P}_{\mathcal{P}_n^{(+)}}\Big(\|\widehat{\theta}_n - \theta_+^\star\|_2 \ge \delta/2\Big),\ \mathbb{P}_{\mathcal{P}_n^{(-)}}\Big(\|\widehat{\theta}_n - \theta_-^\star\|_2 \ge \delta/2\Big)\Big\} \ \ \ge \ \ \frac{1}{2}.$$

Then by Markov's inequality

$$\sup_{j\in\{+,-\}} \mathbb{E}_{\mathcal{P}_n^{(j)}}\big[\,\|\widehat{\theta}_n - \theta_j^\star\|_2\big] \ \ge \ \frac{\delta}{4} \ge c\sqrt{\frac{\Delta_n}{\alpha}}$$

where $c = \sqrt{2}/4$ whenever $\sqrt{2\Delta_n/\alpha} \le \min\{\rho, r_0/2\}$. Since Δ_n is usually a function of ρ, this is not too restrictive. If instead Δ_n is larger, then $\delta \ge c'\rho$ for a universal $c' > 0$. Whether this is larger or smaller than the statistical error is dependent on how ρ changes with n, but the presence of the cap does not change the overall rate because the upper bound will contain the maximum of the statistical error and the residual error.

The total minimax error is lower-bounded by the maximum of the two error sources, as an estimator must be robust to both statistical noise and modeling bias. Therefore, combining the two parts, the overall minimax lower bound is:

$$\inf_{\hat{\theta}_n}\sup_{\mathbb{P}^\circ \in \mathcal{M}(\alpha, L_\phi, \Delta_n)} \mathbb{E}_{\mathbb{P}^\circ}\big[\,\|\hat{\theta}_n - \theta^\star(\mathbb{P}^\circ)\|_2\big] \ge c_1\frac{L_\phi}{\alpha}\sqrt{\frac{1}{n}} \vee c_2\sqrt{\frac{\Delta_n}{\alpha}}$$

□

H Proofs of "Fast Rate" DRO-REBEL

H.1 Proof of "Fast Rate" Wasserstein-DRO-REBEL

Before we prove the necessary results to get the "fast rate" for Wasserstein-DRO-REBEL, we need to make an assumption on the loss functions $\ell(\cdot;\theta)$, $\theta \in \Theta$. Note that this assumption is only used in proving the dual "remainder" term holds:

Assumption 5. *For the Wasserstein-DRO-REBEL objective, we assume that the pointwise loss function $\ell(z;\theta)$ is $L_{\ell,z}$-Lipschitz with respect to its data argument $z = (x, a^1, a^2)$ for all $\theta \in \Theta$. That is, there exists a constant $L_{\ell,z} \ge 0$ such that for any $z_1, z_2 \in \mathcal{Z}$,*

$$|\ell(z_1;\theta) - \ell(z_2;\theta)| \le L_{\ell,z} d(z_1, z_2)$$

where $d(\cdot,\cdot)$ is the metric corresponding to the type-p Wasserstein distance used to define the ambiguity set.

Proof of Corollary 1. First, recall the type-p Wasserstein distance. The type-p ($p \in [1, \infty)$) Wasserstein distance between two distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\Xi)$ is defined as

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} d(\xi, \eta)^p \pi(d\xi, d\eta) \right)^{1/p}$$

where π is a coupling between the marginal distributions $\xi \sim \mathbb{P}$ and $\eta \sim \mathbb{Q}$, and d is a pseudometric defined on $\mathcal{Z}$.

Let's bound the difference between expectations under $\mathbb{P}$ and $\mathbb{P}^\circ$. For any coupling π between $\mathbb{P}$ and $\mathbb{P}^\circ$:

$$\begin{aligned} \mathbb{E}_{z \sim \mathbb{P}}[\ell(z; \theta)] - \mathbb{E}_{z \sim \mathbb{P}^\circ}[\ell(z; \theta)] &= \int \ell(\xi; \theta) \pi(d\xi, d\eta) - \int \ell(\eta; \theta) \pi(d\xi, d\eta) \\ &= \int (\ell(\xi; \theta) - \ell(\eta; \theta)) \pi(d\xi, d\eta) \end{aligned}$$

This equality holds due to the marginal properties of a coupling. Now, assume that the loss function $\ell(z; \theta)$ is $L_{\ell,z}$-Lipschitz with respect to z (i.e., $|\ell(\xi; \theta) - \ell(\eta; \theta)| \leq L_{\ell,z} d(\xi, \eta)$ for some constant $L_{\ell,z}$). Then, for $p \geq 1$, we can use Lemma 7 (monotonicity of the Wasserstein distance) to get

$$\begin{aligned} |\mathbb{E}_{z \sim \mathbb{P}}[\ell(z; \theta)] - \mathbb{E}_{z \sim \mathbb{P}^\circ}[\ell(z; \theta)]| &\leq \int |\ell(\xi; \theta) - \ell(\eta; \theta)| \pi(d\xi, d\eta) \\ &\leq L_{\ell,z} \int d(\xi, \eta) \pi(d\xi, d\eta) \\ &\leq L_{\ell,z} \left(\int d(\xi, \eta)^p \pi(d\xi, d\eta) \right)^{1/p} \end{aligned}$$

Taking the supremum over all couplings π, and then over all $\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathcal{W}_p)$:

$$\begin{aligned} \left| \mathcal{L}^{\mathcal{W}_p}(\theta) \; - \; \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z; \theta)\big] \right| &= \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathcal{W}_p)} \left| \mathbb{E}_{z \sim \mathbb{P}}[\ell(z; \theta)] - \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z; \theta)\big] \right| \\ &\leq L_{\ell,z} \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathcal{W}_p)} \mathcal{W}_p(\mathbb{P}, \mathbb{P}^\circ) \\ &\leq L_{\ell,z} \varepsilon_n \end{aligned}$$

Thus, we can set $\Delta_n = L_{\ell,z} \varepsilon_n$. If we choose $\varepsilon_n \asymp n^{-1}$, then $\Delta_n = O(n^{-1})$, which aligns with the condition for the $O(n^{-1/2})$ rate in the Master Theorem.

We now show both DPO and REBEL satisfy Assumption 5 by deriving their respective Lipschitz constants with respect to the defined pseudometric. For any $z = (x, a^1, a^2, y) \in \mathcal{Z}$, we denote

$$\Delta\psi(z) := \psi(x, a^1) - \psi(x, a^2), \quad \Delta r(z) := \phi(x, a^1)^\top \omega - \phi(x, a^2)^\top \omega, \quad h_\theta(z) := \theta^\top \Delta\psi(z).$$

We equip $\mathcal{Z}$ with the pseudo–metric

$$d(z, z') = \|\Delta\psi(z) - \Delta\psi(z')\|_2 + |\Delta r(z) - \Delta r(z')| + |y - y'|$$

(i) DPO loss. Define $f(h, y) := -y \log \sigma(\beta h) - (1 - y) \log \sigma(-\beta h)$ so that $\ell_{\mathrm{DPO}}(z; \theta) = f\big(h_\theta(z), y\big)$. We recall the following bounds from Assumptions 1 and 2: $\|\theta\|_2 \leq B$ and $\|\psi(x, a)\|_2 \leq 1$. These imply $\|\Delta\psi(z)\|_2 \leq \|\psi(x, a^1)\|_2 + \|\psi(x, a^2)\|_2 \leq 2$. Consequently, $|h_\theta(z)| \leq \|\theta\|_2 \|\Delta\psi(z)\|_2 \leq 2B$.

Let's compute the partial derivatives of $f(h, y)$:

$$\frac{\partial f}{\partial h} = -y \frac{d}{dh}(\log \sigma(\beta h)) - (1 - y) \frac{d}{dh}(\log \sigma(-\beta h))$$

Using $\frac{d}{du}(\log \sigma(u)) = 1 - \sigma(u) = \sigma(-u)$:

$$\frac{\partial f}{\partial h} = -y(\beta(1 - \sigma(\beta h))) - (1 - y)(-\beta(1 - \sigma(-\beta h))) = \beta((1 - y)\sigma(\beta h) - y\sigma(-\beta h))$$

For $y = 1$, $|\partial f / \partial h| = | - \beta\sigma(-\beta h)| < \beta$ since $0 < \sigma(u) < 1$. For $y = 0$, $|\partial f / \partial h| = |\beta\sigma(\beta h)| < \beta$. Thus,

$$\big|\partial f / \partial h\big| \leq \beta$$

For the partial derivative with respect to y:

$$
\begin{aligned}
\left|\partial f/\partial y\right| = |-\log\sigma(\beta h)+\log\sigma(-\beta h)| &= \left|\log\left(\frac{\sigma(-\beta h)}{\sigma(\beta h)}\right)\right| \\
&= \left|\log\left(\frac{1/(1+e^{\beta h})}{1/(1+e^{-\beta h})}\right)\right| \\
&= \left|\log\left(\frac{1+e^{-\beta h}}{1+e^{\beta h}}\right)\right|
\end{aligned}
$$

The maximum value of $|\log(\frac{1+e^{-u}}{1+e^{u}})|$ for $u \in [-C, C]$ is C. Since $|h| \leq 2B$, we have $|\beta h| \leq 2\beta B$. Therefore,

$$
\left|\partial f/\partial y\right| \leq 2\beta B.
$$

Since $|\ell_{\mathrm{DPO}}(z;\theta) - \ell_{\mathrm{DPO}}(z';\theta)| \leq |\partial f/\partial h||h_\theta(z) - h_\theta(z')| + |\partial f/\partial y||y - y'|$ by the mean-value inequality, and $|h_\theta(z) - h_\theta(z')| \leq B\,\|\Delta\psi(z) - \Delta\psi(z')\|_2$, we have:

$$
|\ell_{\mathrm{DPO}}(z;\theta) - \ell_{\mathrm{DPO}}(z';\theta)| \leq \beta B\,\|\Delta\psi(z) - \Delta\psi(z')\|_2 + 2\beta B\,|y - y'|.
$$

The DPO loss does not depend on $\Delta r(z)$. Given the metric $d(z, z') = \|\Delta\psi(z) - \Delta\psi(z')\|_2 + |\Delta r(z) - \Delta r(z')| + |y - y'|$, the Lipschitz constant $L_{\ell,z}^{\mathrm{DPO}}$ is the sum of the coefficients corresponding to the terms in $d(z, z')$ that affect the loss. Thus,

$$
L_{\ell,z}^{\mathrm{DPO}} = \beta B + 2\beta B = 3\beta B.
$$

(ii) REBEL loss. Set $g(h, \Delta r) := [\eta^{-1}(h - \Delta r)]^2$ so that $\ell_{\mathrm{REBEL}}(z;\theta) = g\big(h_\theta(z), \Delta r(z)\big)$. From Assumptions 2 and 1, we have $|h_\theta(z)| \leq 2B$ and $|\Delta r(z)| \leq 2F$. Thus, $|h - \Delta r| \leq |h| + |\Delta r| \leq 2B + 2F = 2(B + F)$.

The partial derivatives of $g(h, \Delta r)$ satisfy:

$$
\begin{aligned}
|\partial g/\partial h| = |\partial g/\partial\,\Delta r| &= |2\eta^{-2}(h - \Delta r)| \\
&\leq 2\eta^{-2}(2(B + F)) \\
&= 4\eta^{-2}(B + F).
\end{aligned}
$$

By the mean-value inequality,

$$
|\ell_{\mathrm{REBEL}}(z;\theta) - \ell_{\mathrm{REBEL}}(z';\theta)| \leq |\partial g/\partial h||h_\theta(z) - h_\theta(z')| + |\partial g/\partial\Delta r||\Delta r(z) - \Delta r(z')|.
$$

Using $|h_\theta(z) - h_\theta(z')| \leq B\,\|\Delta\psi(z) - \Delta\psi(z')\|_2$ and the partial derivative bounds:

$$
|\ell_{\mathrm{REBEL}}(z;\theta) - \ell_{\mathrm{REBEL}}(z';\theta)| \leq 4\eta^{-2}(B + F)B\|\Delta\psi(z) - \Delta\psi(z')\|_2 + 4\eta^{-2}(B + F)|\Delta r(z) - \Delta r(z')|.
$$

The REBEL loss does not depend on y. Given the metric $d(z, z') = \|\Delta\psi(z) - \Delta\psi(z')\|_2 + |\Delta r(z) - \Delta r(z')| + |y - y'|$, the Lipschitz constant $L_{\ell,z}^{\mathrm{REBEL}}$ is the sum of the coefficients corresponding to the terms in $d(z, z')$ that affect the loss.

$$
L_{\ell,z}^{\mathrm{REBEL}} = 4\eta^{-2}(B + F)B + 4\eta^{-2}(B + F) = 4\eta^{-2}(B + F)(B + 1).
$$

Hence both DPO and REBEL satisfy Assumption 5 with $L_{\ell,z}^{\mathrm{DPO}} = 3\beta B$ and $L_{\ell,z}^{\mathrm{REBEL}} = 4\eta^{-2}(B + F)(B + 1)$. $\square$

H.2 Proof of "Fast Rate" KL-DRO-REBEL

Proof of Corollary 2. First, recall Lemma 17 (Gibbs variational principle characterization of the KL divergence) for probability measures $\mathbb{P}, \mathbb{Q}$:

$$
D_{\mathrm{KL}}\left(\mathbb{P}\,||\,\mathbb{Q}\right) = \sup_{g:\mathcal{Z}\to\mathbb{R}} \left\{\mathbb{E}_{\mathbb{P}}\left[g\right] - \log\mathbb{E}_{\mathbb{Q}}\left[e^{g}\right]\right\}
$$

Let $f = \ell(z;\theta)$. For any $\lambda \geq 0$, we can choose $g(z) = \lambda(f(z) - \mathbb{E}_{\mathbb{Q}}[f])$ to obtain a lower bound for $D_{\mathrm{KL}}(\mathbb{P}\,||\,\mathbb{Q})$:

$$
D_{\mathrm{KL}}\left(\mathbb{P}\,||\,\mathbb{Q}\right) \geq \sup_{\lambda\geq 0}\left\{\lambda\left(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]\right) - \log\mathbb{E}_{\mathbb{Q}}\left[e^{\lambda(f - \mathbb{E}_{\mathbb{Q}}[f])}\right]\right\}
$$

Now, suppose that $f = \ell(z;\theta) \in [0, K_\ell]$ almost surely (from Lemma 24). By Hoeffding's Lemma (Lemma 18), if f is bounded in $[0, K_\ell]$, then $f - \mathbb{E}_{\mathbb{Q}}[f]$ is sub-Gaussian with parameter $K_\ell/2$. Thus, we have that

$$
\log\mathbb{E}_{\mathbb{Q}}\left[e^{\lambda(f - \mathbb{E}_{\mathbb{Q}}[f])}\right] \leq \frac{\lambda^2(K_\ell/2)^2}{2} = \frac{\lambda^2 K_\ell^2}{8}
$$

Substituting this bound into the KL inequality:

$$D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{Q}\right) \geq \sup_{\lambda \geq 0} \left\{ \lambda \left(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f] \right) - \frac{\lambda^2 K_\ell^2}{8} \right\}$$

The expression in the curly brackets is a concave quadratic in λ. Its supremum is attained at $\lambda^* = \frac{4(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f])}{K_\ell^2}$. Plugging this optimal λ^* back into the expression, we find that

$$D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{Q}\right) \geq \frac{2\left(\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]\right)^2}{K_\ell^2}$$

Rearranging terms, we obtain a bound on the absolute difference in expectations:

$$\left|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]\right| \leq \frac{1}{\sqrt{2}} K_\ell \sqrt{D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{Q}\right)}$$

Now, taking $\mathbb{Q} = \mathbb{P}^\circ$ and $f = \ell(z;\theta)$, and considering the supremum within the KL ball $\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathrm{KL})$:

$$\begin{aligned}
\left|\mathcal{L}^{\mathrm{KL}}(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| &= \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;\mathrm{KL})} \left|\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \\
&\leq \frac{1}{\sqrt{2}} K_\ell \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;\mathrm{KL})} \sqrt{D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)} \\
&\leq \frac{1}{\sqrt{2}} K_\ell \sqrt{\varepsilon_n}
\end{aligned}$$

Thus, we can set $\Delta_n = \frac{1}{\sqrt{2}} K_\ell \sqrt{\varepsilon_n}$. If we choose $\varepsilon_n \asymp n^{-2}$, then $\Delta_n = O(n^{-1})$, which satisfies the condition for the $O(n^{-1/2})$ rate in the Master Theorem. $\square$

H.3 Proof of "Fast Rate" χ^2-DRO-REBEL

Proof of Corollary 3. By Theorem 14 (Hammersley-Chapman-Robbins (HCR) lower bound), we immediately have:

$$\left|\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \leq \sqrt{\mathrm{Var}_{\mathbb{P}^\circ}\left(\ell(z;\theta)\right)\chi^2\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)}$$

Since $\ell(z;\theta) \in [0, K_\ell]$ almost surely, by Lemma 20, its variance is bounded by $\mathrm{Var}_{\mathbb{P}^\circ}\left(\ell(z;\theta)\right) \leq K_\ell^2/4$. Thus, we have:

$$\begin{aligned}
\left|\mathcal{L}^{\chi^2}(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| &= \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;\chi^2)} \left|\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \\
&\leq \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;\chi^2)} \sqrt{\frac{K_\ell^2}{4}\chi^2\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)} \\
&\leq \frac{K_\ell}{2} \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;\chi^2)} \sqrt{\chi^2\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)} \\
&\leq \frac{K_\ell}{2}\sqrt{\varepsilon_n}
\end{aligned}$$

Thus, we can set $\Delta_n = \frac{K_\ell}{2}\sqrt{\varepsilon_n}$. If we choose $\varepsilon_n \asymp n^{-2}$, then $\Delta_n = O(n^{-1})$, which satisfies the condition for the $O(n^{-1/2})$ rate in the Master Theorem. $\square$

H.4 Proof of "Fast Rate" for General f-Divergences

Before we prove the necessary results to get the "fast rate" for general f-divergences, we need to make an assumption on the loss function $\ell(\cdot;\theta)$, $\theta \in \Theta$ and the form of the f-divergence. Note that this assumption is only used in proving the dual "remainder" term holds:

Assumption 6. *The pointwise loss function $\ell(z;\theta)$ is bounded for all $z \in \mathcal{Z}$ and $\theta \in \Theta$. That is, there exist constants $m, M \in \mathbb{R}$ such that $m \leq \ell(z;\theta) \leq M$ almost surely with respect to any relevant probability measure. Let $K_\ell = M - m$ be the range of the loss function.*

Assumption 7. *Let $f : [0,\infty) \to \mathbb{R} \cup \{\infty\}$ be a convex function satisfying $f(1) = 0$. Furthermore, we assume:*

1. *f is twice continuously differentiable at $t = 1$.*
2. *$f''(1) > 0$. This implies that f is strictly convex at $t = 1$.*

Proof of Theorem 6. Let f be a convex function satisfying Assumption 7. The f-divergence $D_f(\mathbb{P}||\mathbb{Q})$ has the variational representation:

$$D_f(\mathbb{P}||\mathbb{Q}) = \sup_{g \in \mathcal{G}} \{\mathbb{E}_{\mathbb{P}}[g] - \mathbb{E}_{\mathbb{Q}}[f^*(g)]\}$$

where f^* is the Fenchel conjugate of f, defined as $f^*(y) = \sup_{x \in [0,\infty)}\{xy - f(x)\}$, and $\mathcal{G}$ is the set of all measurable functions for which the expectations are finite.

We are interested in bounding $|\mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] - \mathbb{E}_{\mathbb{Q}}[\ell(z;\theta)]|$. Let $h(z) = \ell(z;\theta)$. By Assumption 6, $h(z) \in [m, M]$ almost surely. Let $\Delta_h = \mathbb{E}_{\mathbb{P}}[h] - \mathbb{E}_{\mathbb{Q}}[h]$.

Since f is convex and twice continuously differentiable at 1 with $f(1) = 0$ and $f''(1) > 0$, its Fenchel conjugate f^* is convex and twice continuously differentiable at 0. Furthermore, we have:

- $f^*(0) = -\min_x f(x)$. If f is minimized at $x = 1$ (which is typically the case for f-divergences), and $f(1) = 0$, then $f^*(0) = 0$.
- $(f^*)'(0) = \arg\min_x f(x)$. Under the same conditions, $(f^*)'(0) = 1$.
- $(f^*)''(0) = \frac{1}{f''(1)}$. Let $C_f = \frac{1}{f''(1)}$. Since $f''(1) > 0$, C_f is a finite positive constant.

By Taylor's theorem with a remainder term, for any y in a neighborhood of 0, we can write:

$$f^*(y) = f^*(0) + (f^*)'(0)y + \frac{1}{2}(f^*)''(0)y^2 + R_2(y)$$

where $R_2(y)$ is the remainder term, satisfying $\lim_{y\to 0} \frac{R_2(y)}{y^2} = 0$. Given the properties above, this becomes:

$$f^*(y) = y + \frac{1}{2}C_f y^2 + R_2(y)$$

Choose $g(z) = \lambda(h(z) - \mathbb{E}_{\mathbb{Q}}[h])$ for some $\lambda \in \mathbb{R}$. Then $\mathbb{E}_{\mathbb{Q}}[g] = 0$. The variational representation gives us:

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[g] - \mathbb{E}_{\mathbb{Q}}[f^*(g)]$$

Substitute the Taylor expansion of $f^*(g(z))$ into the inequality:

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[\lambda(h - \mathbb{E}_{\mathbb{Q}}[h])] - \mathbb{E}_{\mathbb{Q}}\left[\lambda(h - \mathbb{E}_{\mathbb{Q}}[h]) + \frac{1}{2}C_f\lambda^2(h - \mathbb{E}_{\mathbb{Q}}[h])^2 + R_2(\lambda(h - \mathbb{E}_{\mathbb{Q}}[h]))\right]$$

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \lambda(\mathbb{E}_{\mathbb{P}}[h] - \mathbb{E}_{\mathbb{Q}}[h]) - \lambda\mathbb{E}_{\mathbb{Q}}[h - \mathbb{E}_{\mathbb{Q}}[h]] - \frac{1}{2}C_f\lambda^2\mathbb{E}_{\mathbb{Q}}[(h - \mathbb{E}_{\mathbb{Q}}[h])^2] - \mathbb{E}_{\mathbb{Q}}[R_2(\lambda(h - \mathbb{E}_{\mathbb{Q}}[h]))]$$

Since $\mathbb{E}_{\mathbb{Q}}[h - \mathbb{E}_{\mathbb{Q}}[h]] = 0$, we have:

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \lambda\Delta_h - \frac{1}{2}C_f\lambda^2\mathrm{Var}_{\mathbb{Q}}(h) - \mathbb{E}_{\mathbb{Q}}[R_2(\lambda(h - \mathbb{E}_{\mathbb{Q}}[h]))]$$

By Assumption 6, $h(z) \in [m, M]$. Thus, $(h(z) - \mathbb{E}_{\mathbb{Q}}[h]) \in [m - M, M - m] = [-K_\ell, K_\ell]$. Therefore, $|h(z) - \mathbb{E}_{\mathbb{Q}}[h]| \leq K_\ell$. By Popoviciu's inequality (Lemma 20), $\mathrm{Var}_{\mathbb{Q}}(h) \leq \frac{K_\ell^2}{4}$.

Let $Y = h - \mathbb{E}_{\mathbb{Q}}[h]$. Then $|Y| \leq K_\ell$. The remainder term $R_2(y)$ satisfies $|R_2(y)| \leq \epsilon(y)y^2$ for some function $\epsilon(y)$ such that $\lim_{y\to 0}\epsilon(y) = 0$. Thus, for any $\delta_0 > 0$, there exists $\Lambda > 0$ such that for $|y| \leq \Lambda$, $|\epsilon(y)| \leq \delta_0$. If we choose λ such that $|\lambda| K_\ell \leq \Lambda$, then $|\lambda Y| \leq \Lambda$. So, $|R_2(\lambda Y)| \leq \delta_0(\lambda Y)^2$. Then, $|\mathbb{E}_{\mathbb{Q}}[R_2(\lambda Y)]| \leq \mathbb{E}_{\mathbb{Q}}[\delta_0(\lambda Y)^2] = \delta_0\lambda^2\mathrm{Var}_{\mathbb{Q}}(h)$.

Combining these, we get:

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \lambda\Delta_h - \frac{1}{2}C_f\lambda^2\mathrm{Var}_{\mathbb{Q}}(h) - \delta_0\lambda^2\mathrm{Var}_{\mathbb{Q}}(h)$$

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \lambda\Delta_h - \left(\frac{C_f}{2} + \delta_0\right)\lambda^2\mathrm{Var}_{\mathbb{Q}}(h)$$

Using $\mathrm{Var}_{\mathbb{Q}}(h) \leq \frac{K_\ell^2}{4}$:

$$D_f(\mathbb{P}||\mathbb{Q}) \geq \lambda \Delta_h - \left(\frac{C_f}{2} + \delta_0\right) \lambda^2 \frac{K_\ell^2}{4}$$

Let $B_{\delta_0} = \left(\frac{C_f}{2} + \delta_0\right) \frac{K_\ell^2}{4}$. The right-hand side is a concave quadratic in λ: $A\lambda - B_{\delta_0}\lambda^2$, with $A = \Delta_h$. This quadratic is maximized at $\lambda^* = \frac{A}{2B_{\delta_0}} = \frac{\Delta_h}{2\left(\frac{C_f}{2}+\delta_0\right)\frac{K_\ell^2}{4}} = \frac{2\Delta_h}{(C_f+2\delta_0)\frac{K_\ell^2}{2}}$. Plugging λ^* back into the inequality:

$$\begin{aligned} D_f(\mathbb{P}||\mathbb{Q}) &\geq \lambda^* \Delta_h - B_{\delta_0}(\lambda^*)^2 \\ &= \frac{(\Delta_h)^2}{2B_{\delta_0}} \\ &= \frac{(\Delta_h)^2}{2\left(\frac{C_f}{2} + \delta_0\right)\frac{K_\ell^2}{4}} \\ &= \frac{(\Delta_h)^2}{(C_f + 2\delta_0)\frac{K_\ell^2}{2}} \end{aligned}$$

Rearranging to bound $|\Delta_h|$:

$$|\Delta_h|^2 \leq \frac{(C_f + 2\delta_0)\, K_\ell^2}{2} D_f(\mathbb{P}||\mathbb{Q})$$

$$|\mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] - \mathbb{E}_{\mathbb{Q}}[\ell(z;\theta)]| \leq \frac{K_\ell\sqrt{C_f + 2\delta_0}}{\sqrt{2}} \sqrt{D_f(\mathbb{P}||\mathbb{Q})}$$

The choice of δ_0 (and thus Λ) depends on how small $|\lambda^* K_\ell|$ needs to be. Since $|\lambda^* K_\ell| = \left|\frac{2\Delta_h}{(C_f+2\delta_0)K_\ell^2}\right| K_\ell = \frac{2|\Delta_h|}{(C_f+2\delta_0)K_\ell}$, the condition $|\lambda^* K_\ell| \leq \Lambda$ becomes:

$$|\Delta_h| \leq \frac{(C_f + 2\delta_0)\, K_\ell \Lambda}{2}$$

As we consider the supremum over $\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_f)$, $D_f(\mathbb{P}||\mathbb{P}^\circ) \leq \varepsilon_n$. As $\varepsilon_n \to 0$, $D_f(\mathbb{P}||\mathbb{P}^\circ) \to 0$. By the derived inequality, this implies $|\Delta_h| \to 0$. Therefore, for any given $\epsilon > 0$, we can choose ε_n small enough such that $|\Delta_h|$ is sufficiently small. This allows us to pick δ_0 small enough (e.g., $\delta_0 < \epsilon$) to satisfy the condition on $|\lambda Y| \leq \Lambda$. Thus, for ε_n sufficiently small:

$$\begin{aligned} \left|\mathcal{L}^f(\theta) \ - \ \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| &= \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_f)} \left|\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \\ &\leq \frac{K_\ell\sqrt{C_f + 2\delta_0}}{\sqrt{2}} \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_f)} \sqrt{D_f\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)} \\ &\leq \frac{K_\ell\sqrt{C_f + 2\delta_0}}{\sqrt{2}} \sqrt{\varepsilon_n} \end{aligned}$$

Since δ_0 can be made arbitrarily small as $\varepsilon_n \to 0$, this leads to:

$$\left|\mathcal{L}^f(\theta) \ - \ \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \lesssim \frac{K_\ell\sqrt{C_f}}{\sqrt{2}} \sqrt{\varepsilon_n}$$

Thus, we can set $\Delta_n = \frac{K_\ell\sqrt{C_f}}{\sqrt{2}}\sqrt{\varepsilon_n}$. If we choose $\varepsilon_n \asymp n^{-2}$, then $\Delta_n = O(n^{-1})$, which satisfies the condition for the $O(n^{-1/2})$ rate in the Master Theorem. $\square$

H.5 Proof of "Fast Rate" for Total Variation Distance

Proof. The Total Variation (TV) distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$D_{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{h:\mathcal{Z}\to\mathbb{R}, \|h\|_\infty \leq 1} |\mathbb{E}_{\mathbb{P}}[h(Z)] - \mathbb{E}_{\mathbb{Q}}[h(Z)]|$$

where $\|h\|_\infty = \sup_{z\in\mathcal{Z}} |h(z)|$ is the supremum norm. Let $h_0(z) = \ell(z;\theta)$ be the pointwise loss function. By Theorem 24, we have $0 \le h_0(z) \le K_\ell$ almost surely. Consider the difference in expected loss under $\mathbb{P}$ and $\mathbb{Q}$:

$$|\mathbb{E}_\mathbb{P}[h_0(Z)] - \mathbb{E}_\mathbb{Q}[h_0(Z)]|$$

To relate this difference to the TV distance, we construct a function $h(z)$ that satisfies the condition $\|h\|_\infty \le 1$. Since $h_0(z) \in [0, K_\ell]$, the midpoint of this interval is $\frac{0+K_\ell}{2} = \frac{K_\ell}{2}$, and the range is K_ℓ. Define $h(z)$ as a scaled and shifted version of $h_0(z)$:

$$h(z) = \frac{h_0(z) - \frac{K_\ell}{2}}{\frac{K_\ell}{2}} = \frac{2h_0(z) - K_\ell}{K_\ell}$$

Let's verify the bounds on $h(z)$:

$$-\frac{K_\ell}{2} \le h_0(z) - \frac{K_\ell}{2} \le \frac{K_\ell}{2}$$

Dividing by $\frac{K_\ell}{2}$ (which is positive, as $K_\ell > 0$ by Theorem 24):

$$-1 \le \frac{h_0(z) - \frac{K_\ell}{2}}{\frac{K_\ell}{2}} \le 1$$

Thus, $|h(z)| \le 1$ for all $z \in \mathcal{Z}$, meaning $\|h\|_\infty \le 1$.

Now, let's express $h_0(z)$ in terms of $h(z)$:

$$h_0(z) = \frac{K_\ell}{2} h(z) + \frac{K_\ell}{2} = \frac{K_\ell}{2}(h(z) + 1)$$

Substitute this into the difference of expectations:

$$\begin{aligned}
&|\mathbb{E}_\mathbb{P}[h_0(Z)] - \mathbb{E}_\mathbb{Q}[h_0(Z)]| \\
&= \left|\mathbb{E}_\mathbb{P}\left[\frac{K_\ell}{2}(h(Z)+1)\right] - \mathbb{E}_\mathbb{Q}\left[\frac{K_\ell}{2}(h(Z)+1)\right]\right| \\
&= \left|\frac{K_\ell}{2}(\mathbb{E}_\mathbb{P}[h(Z)]+1) - \frac{K_\ell}{2}(\mathbb{E}_\mathbb{Q}[h(Z)]+1)\right| \\
&= \left|\frac{K_\ell}{2}\mathbb{E}_\mathbb{P}[h(Z)] - \frac{K_\ell}{2}\mathbb{E}_\mathbb{Q}[h(Z)]\right| \\
&= \frac{K_\ell}{2}\left|\mathbb{E}_\mathbb{P}[h(Z)] - \mathbb{E}_\mathbb{Q}[h(Z)]\right|
\end{aligned}$$

Since $\|h\|_\infty \le 1$, from the definition of Total Variation distance, we know that $|\mathbb{E}_\mathbb{P}[h(Z)] - \mathbb{E}_\mathbb{Q}[h(Z)]| \le D_{TV}(\mathbb{P}, \mathbb{Q})$. Therefore, we can write:

$$|\mathbb{E}_\mathbb{P}[\ell(z;\theta)] - \mathbb{E}_\mathbb{Q}[\ell(z;\theta)]| \le \frac{K_\ell}{2} D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$$

This inequality holds for any pair of probability measures $\mathbb{P}, \mathbb{Q}$.

Now, we consider the definition of the robust loss function $\mathcal{L}^{\mathrm{TV}}(\theta)$ over a TV-ball of radius ε_n:

$$\mathcal{L}^{\mathrm{TV}}(\theta) = \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_{\mathrm{TV}})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$$

where $\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_{\mathrm{TV}}) = \{\mathbb{P} \mid D_{\mathrm{TV}}(\mathbb{P}, \mathbb{P}^\circ) \le \varepsilon_n\}$.

We want to bound $\left|\mathcal{L}^{\mathrm{TV}}(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right|$. Since $\mathbb{P}^\circ \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; D_{\mathrm{TV}})$ (as $D_{\mathrm{TV}}(\mathbb{P}^\circ, \mathbb{P}^\circ) = 0 \le \varepsilon_n$), we know that $\mathcal{L}^{\mathrm{TV}}(\theta) \ge \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]$. Thus, the absolute value can be removed from the outer expression:

$$
\begin{aligned}
&\left|\mathcal{L}^{\mathrm{TV}}(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \\
&= \mathcal{L}^{\mathrm{TV}}(\theta) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right] \\
&= \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_{\mathrm{TV}})} \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right] \\
&= \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_{\mathrm{TV}})} \left(\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right) \\
&\le \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_{\mathrm{TV}})} \left|\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \\
&\le \sup_{\mathbb{P}\in\mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ;D_{\mathrm{TV}})} \frac{K_\ell}{2} D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{P}^\circ\right) \\
&\le \frac{K_\ell}{2}\varepsilon_n
\end{aligned}
$$

Thus, for the Total Variation distance, the worst-case difference in expected loss is bounded by:

$$\Delta_n = \frac{K_\ell}{2}\varepsilon_n$$

To achieve the desired $O(n^{-1})$ rate for Δ_n, we need to set the radius of the TV-ball as $\varepsilon_n \asymp n^{-1}$. ☐

I Proof of "Fast Rate" DRO-DPO

I.1 Proof of "Fast Rate" WDPO

Proof. From the "Master Theorem", all we need to do is verify that the Wasserstein DPO objective and the DPO loss function satisfy the two conditions.

1. Verification of Local Strong Convexity From Appendix B.3, Lemma 11 of Xu et al. [2025], we know that the Wasserstein DPO loss, $\mathcal{L}^W(\theta)$, is $\gamma\lambda$-strongly convex with respect to the Euclidean norm $\|\cdot\|_2$. This directly satisfies the first condition with a strong convexity parameter $\alpha = \gamma\lambda$ where $\gamma = \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$ and λ is from the data coverage assumption.

2. Verification of Lipschitz Loss (in θ) and h_θ Linear In The Feature Map We show that the pointwise DPO loss, $\ell_{\mathrm{DPO}}(z;\theta) = -y\log\sigma(\beta h_\theta) - (1-y)\log\sigma(-\beta h_\theta)$, is Lipschitz in θ. The gradient with respect to θ is $\nabla_\theta \ell_{\mathrm{DPO}}(z;\theta) = \partial\ell_{\mathrm{DPO}}/\partial h_\theta \cdot \nabla_\theta h_\theta$.

First, we bound the norm of the gradient of the preference score. Using the log-linear policy assumption:

$$
\begin{aligned}
h_\theta(s, a^1, a^2) &:= \left(\log\frac{\pi_\theta(a^1|s)}{\pi_{\text{ref}}(a^1|s)}\right) - \left(\log\frac{\pi_\theta(a^2|s)}{\pi_{\text{ref}}(a^2|s)}\right) \\
&= \left(\log\pi_\theta(a^1|s) - \log\pi_{\text{ref}}(a^1|s)\right) - \left(\log\pi_\theta(a^2|s) - \log\pi_{\text{ref}}(a^2|s)\right) \\
&= \left(\langle\theta, \psi(s,a^1)\rangle - \langle\theta_{\text{ref}}, \psi(s,a^1)\rangle\right) - \left(\langle\theta, \psi(s,a^2)\rangle - \langle\theta_{\text{ref}}, \psi(s,a^2)\rangle\right) \\
&= \langle\theta - \theta_{\text{ref}}, \psi(s,a^1) - \psi(s,a^2)\rangle
\end{aligned}
$$

This is clearly a linear function in θ. The gradient of h_θ with respect to θ is $\nabla_\theta h_\theta = \psi(s,a^1) - \psi(s,a^2)$. Its norm is bounded:

$$
\begin{aligned}
\|\nabla_\theta h_\theta\|_2 = \|\psi(s,a^1) - \psi(s,a^2)\|_2 &\le \|\psi(s,a^1)\|_2 + \|\psi(s,a^2)\|_2 \\
&\le 2
\end{aligned}
$$

Second, we bound the magnitude of the derivative of the logistic loss with respect to h_θ.

$$\frac{\partial\ell_{\mathrm{DPO}}}{\partial h_\theta} = -y\beta(1-\sigma(\beta h_\theta)) + (1-y)\beta\sigma(\beta h_\theta) = \beta((1-y)\sigma(\beta h_\theta) - y\sigma(-\beta h_\theta))$$

Since $y \in \{0,1\}$ and $\sigma(\cdot) \in (0,1)$, the magnitude is maximized when either term is active, giving $|\partial \ell / \partial h_\theta| \leq \beta$. Combining these results, the norm of the gradient is bounded:

$$\begin{aligned} \|\nabla_\theta \ell_{\mathrm{DPO}}(z;\theta)\|_2 &= \left|\frac{\partial \ell}{\partial h_\theta}\right| \cdot \|\nabla_\theta h_\theta\|_2 \\ &\leq 2\beta \end{aligned}$$

Thus, the pointwise DPO loss is L_g-Lipschitz in θ, with $L_g = 2\beta$.

All four conditions of the Master Theorem have been verified for the Wasserstein DPO problem. We can now substitute the derived constants $\alpha = \gamma\lambda$, $L_g = 2\beta$, and $\Delta_n = 3\beta B$ (from Appendix H.1) into the theorem's final bound. This yields:

$$\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{6\beta B}{\gamma\lambda}\right)}$$

□

I.2 Proof of "Fast Rate" KLDPO

Proof. As we did for WDPO, we verify that the KL DPO objective satisfy the four conditions of the Master Theorem. We already proved that ℓ_{DPO} is Lipschitz in θ and h_θ is a linear function of the feature map, so we must just verify uniform boundedness and local strong convexity.

1. Verification of Local Strong Convexity From Appendix C, Lemma 14 of Xu et al. [2025], the KL-DPO loss, $\mathcal{L}^{\mathrm{KL}}(\theta)$, is $\gamma\lambda$-strongly convex with respect to the Euclidean norm $\|\cdot\|_2$. This directly satisfies the first condition with a strong convexity parameter$\alpha = \gamma\lambda$ where $\gamma = \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$ and λ is from the data coverage assumption.

2. Verification of Uniform Boundedness From Appendix B.2, Lemma 9 of Xu et al. [2025], we know that the pointwise DPO loss is uniformly bounded by $\log(1+e^{4\beta B})$. This directly satisfies the conditions needed for Master Theorem.

Thus all four conditions of the Master Theorem have been verified for the KL-DPO problem. We can now substitute the derived constants $\alpha = \gamma\lambda$, $L_g = 2\beta$, $K_\ell = \log\sigma(-4\beta B)$, and $\Delta_n = 2^{-1/2}n^{-1}K_\ell$ (from Appendix H.2) into the theorem's final bound. This yields:

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{2\log(1+e^{4\beta B})}{\gamma\lambda}\right)}$$

□

J Proofs for Sample Complexity of χ^2-DPO

J.1 Proof of One-Step Performance Difference Lemma

Lemma 32 (One-Step Performance Difference Lemma)**.** *Let $J(\pi) = \mathbb{E}_{x\sim\rho,\, a\sim\pi(\cdot|x)}[r_*(x,a)]$ where $r_* \in \mathcal{F}$. Define the one-step baseline as $B^\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)}[r_*(x,a)]$ and one-step advantage as $A^\pi(x,a) = r_*(x,a) - B^\pi(x)$. Then we have the following one-step performance difference:*

$$J(\pi') - J(\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{a\sim\pi'}[A^\pi(x,a)]$$

Additionally, under the assumption that $V_{\max} = \sup_{x,a}|A^\pi(x,a)|$, we also have

$$|J(\pi') - J(\pi)| \leq V_{\max}\mathbb{E}_{x\sim\rho}[\|\pi'(\cdot \mid x) - \pi(\cdot \mid x)\|_1]$$

Proof of Lemma 2. First notice that by definition

$$\mathbb{E}_{a\sim\pi(\cdot|x)}[A^\pi(x,a)] = 0 \tag{6}$$

Then notice the following:

$$\begin{aligned}J(\pi') - J(\pi) &= \mathbb{E}_{x\sim\rho,\ a\sim\pi'(\cdot|x)}\left[r_*(x,a)\right] - \mathbb{E}_{x\sim\rho,\ a\sim\pi(\cdot|x)}\left[r_*(x,a)\right] \\ &= \mathbb{E}_{x\sim\rho,\ a\sim\pi'(\cdot|x)}\left[A^\pi(x,a) + B^\pi(x)\right] - \mathbb{E}_{x\sim\rho,\ a\sim\pi(\cdot|x)}\left[A^\pi(x,a) + B^\pi(x)\right] \\ &= \mathbb{E}_{x\sim\rho}\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right]\end{aligned}$$

where the third equality holds from using Equation 6 in the second argument and the fact that $B^\pi(x)$ depends only on x. To prove the second claim, assume for simplicity that the action space $\mathcal{A}$ is countable. Then Equation 6 implies that $\sum_{a\in\mathcal{A}} A^\pi(x,a)\pi(a \mid x) = 0$. Using this, we have

$$\begin{aligned}\mathbb{E}_{x\sim\rho}\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right] &= \mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right] - \mathbb{E}_{a\sim\pi}\left[A^\pi(x,a)\right]\right] \\ &= \mathbb{E}_{x\sim\rho}\left[\sum_{a\in\mathcal{A}} A^\pi(x,a)\left(\pi'\left(\cdot \mid x\right) - \pi\left(\cdot \mid x\right)\right)\right] \\ &\le V_{\max}\mathbb{E}_{x\sim\rho}\left[\|\pi'\left(\cdot \mid x\right) - \pi\left(\cdot \mid x\right)\|_1\right]\end{aligned}$$

where the last inequality holds from Hölder's inequality. It should be noted that since $r_* \in \mathcal{F}$, there exists some $\omega^* \in \mathbb{R}^d$ such $r_*(x,a) = \phi(x,a)^\top\omega^*$. Thus one can bound the one-step advantage as $\sup_{x,a}|A^\pi(x,a)| \le 2F$ so we can take $V_{\max} = 2F$. □

J.2 Proof of Log-Linear Policy Lipschitzness

Lemma 33 (Log-Linear Policies are Lipschitz)**.** *Suppose $\pi_\theta \in \Pi$ are in a log-linear policy class as defined in Assumption 2. Then all policies in such a class are 2-Lipschitz in θ.*

Proof of Lemma 4. Fix $x \in \mathcal{S}$ and define

$$f(\theta) \;=\; \pi_\theta(\cdot \mid x) \;\in\; \mathbb{R}^{|\mathcal{A}|},$$

with components

$$f(\theta)_a = \pi_\theta(a \mid x) = \frac{\exp\bigl(\theta^\top\psi(x,a)\bigr)}{\sum_{a'}\exp\bigl(\theta^\top\psi(x,a')\bigr)}.$$

Let $\Delta = \theta' - \theta$. By the vector-valued mean-value theorem, there exists $t \in (0,1)$ such that

$$f(\theta') - f(\theta) = \nabla f\bigl(\theta + t\,\Delta\bigr)\,\Delta.$$

Taking the ℓ_1 norm yields

$$\bigl\|f(\theta') - f(\theta)\bigr\|_1 \le \bigl\|\nabla f(\theta + t\Delta)\bigr\|_{1\to 1}\,\|\Delta\|_2,$$

where $\|\cdot\|_{1\to 1}$ is the operator norm from ℓ_2 to ℓ_1. Under Assumption 2, $\|\psi(x,a)\|_2 \le 1$. One computes for each a

$$\nabla_\theta\,\pi_\theta(a \mid x) = \pi_\theta(a \mid x)\Bigl[\psi(x,a) - \mathbb{E}_{a'\sim\pi_\theta(\cdot|x)}[\psi(x,a')]\Bigr].$$

Hence for any $u \in \mathbb{R}^d$,

$$\begin{aligned}\bigl\|\nabla f(\theta)\,u\bigr\|_1 &= \sum_a\Bigl|\langle\nabla_\theta\,\pi_\theta(a \mid x),\,u\rangle\Bigr| \\ &\le \sum_a \pi_\theta(a \mid x)\,\|\psi(x,a) - \mathbb{E}[\psi]\|_2\,\|u\|_2 \\ &\le \bigl(\max_a\|\psi(x,a)\|_2 + \|\mathbb{E}[\psi]\|_2\bigr)\,\|u\|_2 \\ &\le 2\|u\|_2\end{aligned}$$

Thus $\|\nabla f(\theta)\|_{1\to 1} \le 2$. Combining the above,

$$\|\pi_{\theta'}(\cdot \mid x) - \pi_\theta(\cdot \mid x)\|_1 \le 2\,\|\theta' - \theta\|_2,$$

□

J.3 Proof of Sample Complexity Performance Gap of χ^2-DPO

Theorem 16 (Sample Complexity Result for χ^2-DPO)**.** *Suppose Assumption 2 and 3 hold. With probability at least $1-\delta$, χDPO produces a policy $\hat{\pi}$ such that simultaneously for all $\pi^* \in \Pi$, we have*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max}\sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

where $V_{\max} = \sup_{x,a} A^\pi(x,a) = 2F$.

Proof of Theorem 10. First we can combine Lemma 2 and 4 to deduce that

$$J(\pi^*) - J(\widehat{\pi}) \leq 2V_{\max}\|\theta^* - \hat{\theta}\|_2$$

Next, we incorporate the provided detailed sample complexity bound (Theorem 6 and Theorem 8) for the parameter estimation error:

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

This lets us conclude that

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max}\sqrt{\frac{1}{n}\left(\frac{2\beta^2}{\gamma^2\lambda^2}\log(1/\delta) + \frac{\log\sigma(-4\beta B)}{\gamma\lambda}\right)}$$

□

K Proof of Tractable χ^2-DRO-REBEL

Proof of 2. The proof that follows is standard in the analysis of f-divergences and follows from Namkoong and Duchi [2017b]. We include it for completeness. Let $\mathbb{P}_n$ be the empirical distribution. The robust optimization problem is given by:

$$\mathcal{L}_n^{\chi^2}(\theta;\rho) = \sup_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] \quad \text{s.t.} \quad D_{\chi^2}(\mathbb{P}\|\mathbb{P}_n) \leq \rho, \quad \mathbb{P} \geq 0, \quad \mathbb{E}_{\mathbb{P}}[1] = 1.$$

The χ^2-divergence is defined by $f(t) = \frac{1}{2}(t-1)^2$. The Fenchel conjugate $f^*(s) = \sup_{t\geq 0}\{st - f(t)\}$. For $s \in \mathbb{R}$, $f'(t) = t - 1$. Setting $s = t - 1$, we get $t = s + 1$. Substituting this into the definition of $f^*(s)$: $f^*(s) = s(s+1) - \frac{1}{2}((s+1)-1)^2 = s^2 + s - \frac{1}{2}s^2 = \frac{1}{2}s^2 + s$. This derivation holds for $t \geq 0$, which implies $s + 1 \geq 0 \implies s \geq -1$. If $s < -1$, the optimal t would be negative, violating $t \geq 0$. In this case, $f^*(s)$ becomes ∞ due to the constraint $t \geq 0$. According to Lemma 11 [Duchi and Namkoong, 2020], the dual form of the f-divergence based DRO problem is:

$$\sup_{\mathbb{P}:\, D_f(\mathbb{P}\|\mathbb{P}_n)\leq\rho} \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] \;=\; \inf_{\substack{\lambda\geq 0\\ \eta\in\mathbb{R}}}\left\{\lambda\,\mathbb{E}_{\mathbb{P}_n}\left[f^*\left(\frac{\ell(z;\theta)-\eta}{\lambda}\right)\right] \;+\; \lambda\,\rho \;+\; \eta\right\}.$$

Substituting $f^*(s) = \frac{1}{2}s^2 + s$ into this dual formulation, with $s = \frac{\ell_i - \eta}{\lambda}$:

$$\mathcal{L}_n^{\chi^2}(\theta;\rho) = \inf_{\substack{\lambda\geq 0\\ \eta\in\mathbb{R}}}\left\{\lambda\,\rho \;+\; \eta \;+\; \mathbb{E}_{\mathbb{P}_n}\left[\lambda\left(\frac{1}{2}\left(\frac{\ell_i-\eta}{\lambda}\right)^2 + \frac{\ell_i-\eta}{\lambda}\right)\right]\right\}.$$

This simplifies to:

$$\mathcal{L}_n^{\chi^2}(\theta;\rho) = \inf_{\substack{\lambda\geq 0\\ \eta\in\mathbb{R}}}\left\{\lambda\,\rho \;+\; \eta \;+\; \mathbb{E}_{\mathbb{P}_n}\left[\frac{(\ell_i-\eta)^2}{2\lambda} + (\ell_i-\eta)\right]\right\}.$$

Let $X_i = \ell_i - \eta$. The objective becomes:

$$\inf_{\substack{\lambda\geq 0\\ \eta\in\mathbb{R}}}\left\{\lambda\,\rho \;+\; \eta \;+\; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{X_i^2}{2\lambda} + X_i\right]\right\}.$$

The non-negativity constraint $\mathbb{P}(z_i) \geq 0$ in the primal problem implies $1 + \frac{\ell_i - \eta}{\lambda} \geq 0$. This is equivalent to $\frac{\ell_i - \eta}{\lambda} \geq -1$. This constraint is handled by a special form of f^* or by considering the dual's objective piece-wise. When $1 + \frac{\ell_i - \eta}{\lambda} < 0$, this instance z_i is excluded from the worst-case distribution. This leads to the emergence of the positive part $(\cdot)_+$ in the objective. Specifically, for χ^2-divergence, it is a known result in robust optimization that the problem is equivalent to:

$$\mathcal{L}_n^{\chi^2}(\theta; \rho) = \inf_{\eta \in \mathbb{R}} \left\{ \eta + \inf_{\lambda > 0} \left\{ \lambda\rho + \frac{1}{n} \sum_{i=1}^{n} \frac{(\ell_i - \eta)_+^2}{2\lambda} \right\} \right\}.$$

Now, we solve the inner minimization with respect to λ for a fixed η. Let $Y_i = (\ell_i - \eta)_+$. The inner objective is:

$$G(\lambda) = \lambda\rho + \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i^2}{2\lambda}.$$

To find the optimal λ^*, we differentiate $G(\lambda)$ with respect to λ and set it to zero:

$$\frac{dG(\lambda)}{d\lambda} = \rho - \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i^2}{2\lambda^2} = 0.$$

Solving for λ^2:

$$\lambda^2 = \frac{\sum_{i=1}^{n} Y_i^2}{2n\rho} = \frac{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}.$$

Since $\lambda > 0$ and $\rho > 0$, we take the positive square root:

$$\lambda^* = \sqrt{\frac{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}}.$$

Substitute λ^* back into the inner objective $G(\lambda)$:

$$\begin{aligned}
G(\lambda^*) &= \sqrt{\frac{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}} \cdot \rho + \frac{1}{n} \sum_{i=1}^{n} \frac{(\ell_i - \eta)_+^2}{2\sqrt{\frac{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}}} \\
&= \rho\sqrt{\frac{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}} + \frac{1}{2}\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]\sqrt{\frac{2\rho}{\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}} \\
&= \sqrt{\frac{\rho^2\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2\rho}} + \frac{1}{2}\sqrt{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]} \\
&= \sqrt{\frac{\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{2}} + \frac{1}{2}\sqrt{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]} \\
&= \sqrt{\frac{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{4}} + \sqrt{\frac{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{4}} \\
&= 2\sqrt{\frac{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}{4}} \\
&= \sqrt{2\rho\mathbb{E}_{\mathbb{P}_n}[(\ell_i - \eta)_+^2]}.
\end{aligned}$$

Therefore, the robust objective simplifies to:

$$\mathcal{L}_n^{\chi^2}(\theta; \rho) = \inf_{\eta \in \mathbb{R}} \Big\{ \eta \ + \ \sqrt{\frac{2\rho}{n} \sum_{i=1}^{n} (\ell_i - \eta)_+^2} \Big\},$$

which matches the claim of the proposition. We now prove that this problem can be solved efficiently by first establishing the convexity of

$$f(\eta) := \eta + \sqrt{\frac{2\varepsilon_n}{n} \sum_{i=1}^{n} (\ell_i - \eta)_+^2}$$

and show that the search space for its minimum is bounded.

First note that $f(\eta)$ is a convex function of η. The function can be written as the sum of two functions, $f(\eta) = g(\eta) + h(\eta)$, where $g(\eta) = \eta$ and $h(\eta) = \sqrt{C \sum_{i=1}^{n} v_i(\eta)^2}$ with $C = \frac{2\varepsilon_n}{n}$ and $v_i(\eta) = (\ell_i - \eta)_+$. The function $g(\eta) = \eta$ is linear and therefore convex. For each i, the function $v_i(\eta) = \max(0, \ell_i - \eta)$ is a hinge function, which is the maximum of two affine (and thus convex) functions, 0 and $\ell_i - \eta$. Therefore, each $v_i(\eta)$ is convex in η. Let $v(\eta) = [v_1(\eta), \ldots, v_n(\eta)]^\top$ be a vector-valued function. Since each component is convex, the function $v(\eta)$ is convex. The function $\phi(v) = \sqrt{C}\|v\|_2$ is the scaled L_2-norm, which is a convex function. Since $f(\eta)$ is the sum of two convex functions, $g(\eta)$ and $h(\eta)$, it is itself a convex function.

Now since $f(\eta)$ is convex, a point η^* is a minimum if and only if the subgradient contains zero, i.e., $0 \in \partial f(\eta^*)$. The subdifferential of f is $\partial f(\eta) = 1 + \partial h(\eta)$. For any point $\eta \in \mathbb{R}$ where f is differentiable (i.e., $\eta \neq \ell_i$ for all i where $\ell_i > \eta$), the derivative is given by:

$$f'(\eta) = 1 - \sqrt{\frac{2\varepsilon_n}{n}} \cdot \frac{\sum_{i:\ell_i > \eta}(\ell_i - \eta)}{\sqrt{\sum_{i:\ell_i>\eta}(\ell_i - \eta)^2}}$$

Because $f(\eta)$ is convex, its subgradient is a monotonically non-decreasing operator of η. We can establish a finite upper bound for the search space. Consider any $\eta > \max_i\{\ell_i\}$. For such an η, the term $(\ell_i - \eta)_+ = 0$ for all $i = 1, \ldots, n$. The objective function simplifies to:

$$f(\eta) = \eta, \quad \text{for } \eta > \max_i\{\ell_i\}$$

In this region, the derivative is $f'(\eta) = 1$. Since the function is strictly increasing for all $\eta > \max_i\{\ell_i\}$, the minimizer η^* must satisfy:

$$\eta^* \leq \max_i\{\ell_i\}$$

This provides a concrete upper bound for the search. A lower bound can also be established, as $f(\eta) \to \infty$ when $\eta \to -\infty$. Thus, the search for the minimum can be restricted to a finite interval.

The properties above guarantee that we can find the unique minimizer η^* efficiently. The monotonicity of the subgradient allows the use of a binary search algorithm.

1. Define a search interval $[L, U]$, where $U = \max_i\{\ell_i\}$ and L is a sufficiently small lower bound.
2. At each iteration, select a candidate $\eta_c = (L + U)/2$.
3. Compute a subgradient $g_c \in \partial f(\eta_c)$. This takes $O(n)$ time as it requires summing over the n loss terms.
4. If $g_c > 0$, the minimum must lie to the left, so we set $U = \eta_c$.
5. If $g_c < 0$, the minimum must lie to the right, so we set $L = \eta_c$.

This procedure is repeated until the interval $[L, U]$ is sufficiently small. The number of iterations required to achieve a desired precision ϵ is $O(\log((U - L)/\epsilon))$. The total complexity of this search is $O(n \log(1/\epsilon))$. For Algorithm 4, if we assume that $\mathrm{Card}\left(\{\ell_i\}_{i=1}^{n}\right) = n$, then the runtime will be $O(n \log n)$. □

L Additional Experimental Results

Below you can find results for both convex and geometric reward mixtures on each REBEL variant discussed in Figure 8, 9, and 10. The takeaways are largely the same. Utilizing the DRO framework in a sample efficient algorithm like REBEL allows us to maintain generalization and prevent overoptimization by adapting to test-time distribution shifts.

M Experiment Training Details

Our empirical evaluation comprises three components: (i) a *radius–coverage/convergence* study in a controlled Gaussian–mixture simulator, (ii) *Emotion Alignment*, and (iii) a simulated *ArmoRM Multi-objective Alignment*. This section provides the methodologies, model architectures, hyperparameters, and implementation details for all three.

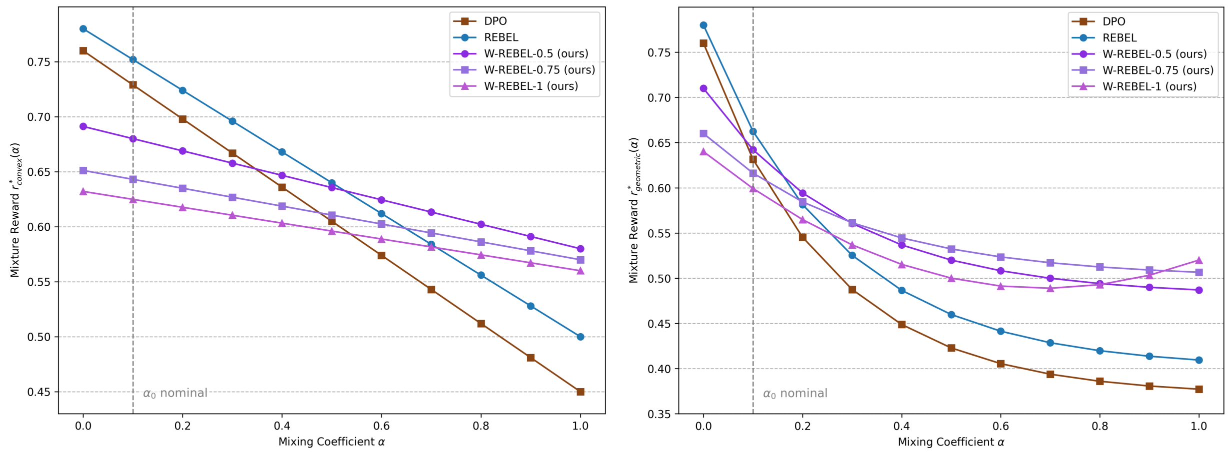

Figure 8: Emotion alignment performance for W-REBEL under convex (left) and geometric (right) reward mixing.

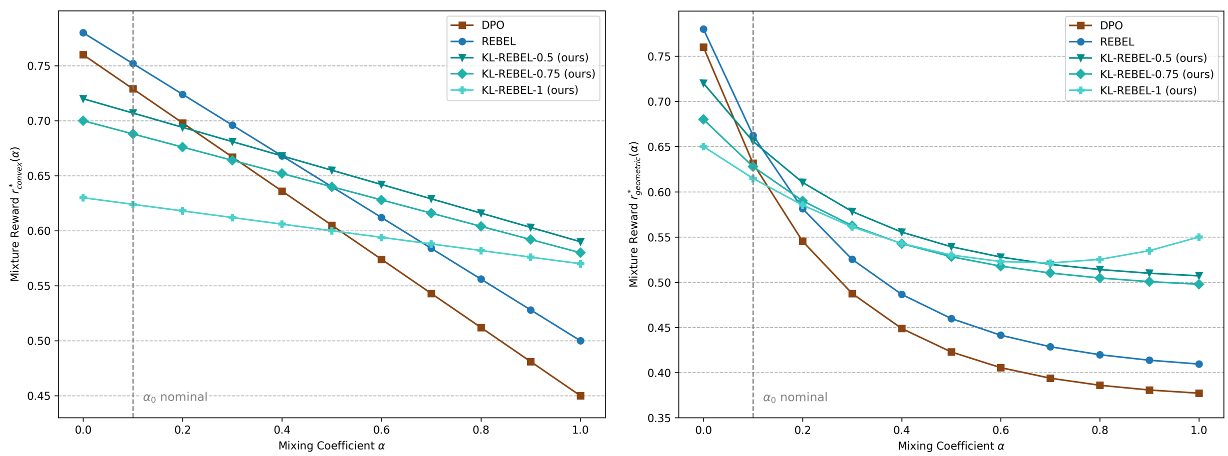

Figure 9: Emotion alignment performance for KL-REBEL under convex (left) and geometric (right) reward mixing.

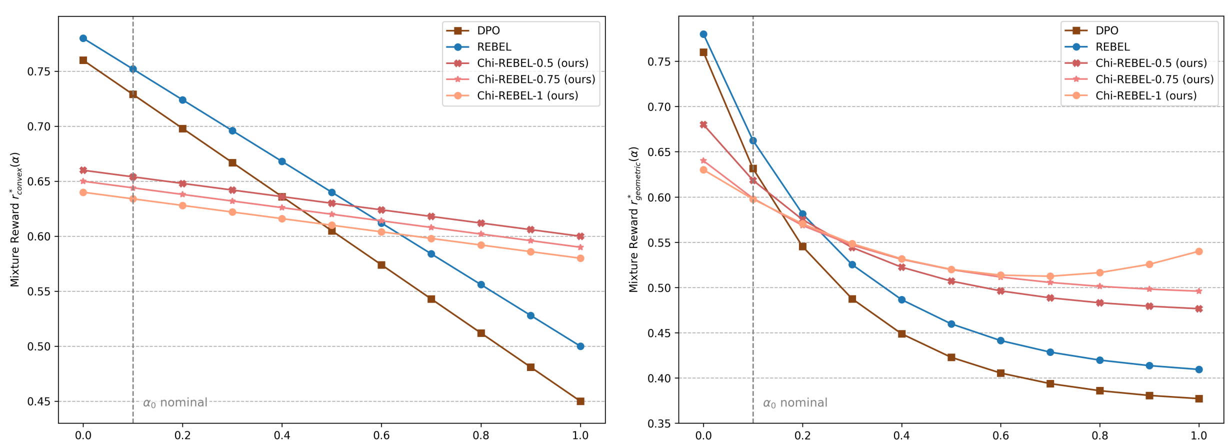

Figure 10: Emotion alignment performance for χ^2-REBEL under convex (left) and geometric (right) reward mixing.

M.1 Radius Coverage Setup

Simulator, data-generating process, and model class. We instantiate a controlled Gaussian–mixture environment with $K{=}15$ latent groups in ambient dimension $d{=}12$. The ground-truth mixture $p^\circ \in \Delta^{K-1}$ is drawn once from $\mathrm{Dir}(0.3 \cdot \mathbf{1})$. Group feature means follow a low-rank factor model: let $U \in \mathbb{R}^{r \times d}$ with $r{=}3$ have row-orthonormalized factors; for each group k, draw $c_k \in \mathbb{R}^r$ and set

$$\mu_k \ \propto\ c_k^\top U + 0.05\,\varepsilon, \quad \text{then normalize row-wise.}$$

Group noise scales σ_k are sampled log-uniformly in $[0.05,\ 0.35]$. A unit vector $\theta^\star \in \mathbb{R}^d$ defines the ground-truth linear reward.

Given a sample size n, we draw group indices $C_i \sim \mathrm{Mult}(p^\circ)$ and generate

$$v_i \sim \mathcal{N}(\mu_{C_i},\, 0.35^2 I_d), \qquad t_i \ =\ v_i^\top \theta^\star + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma_{C_i}^2).$$

We denote the group counts by $\mathrm{count}_k = \sum_i \mathbb{I}\{C_i = k\}$ and the empirical proportions by $\widehat{p}_k = \mathrm{count}_k/n$. Throughout we use a log-linear value/policy class $f_\theta(v) = v^\top \theta$. This deliberately removes optimization confounds, allowing us to (i) measure parameter error $\|\widehat{\theta} - \theta^\star\|_2$ directly and (ii) cleanly compare to classical $n^{-1/2}$ rates.

Mixture-robust training objective. We train θ against *mixture* (group-level) uncertainty around a reference mixture p_{ref} taken as $\widehat{p}$ on each training realization. For squared loss, define per-group losses

$$L_k(\theta) \ :=\ \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \big(v_i^\top \theta - t_i\big)^2, \qquad \mathcal{I}_k = \{i :\ C_i = k\}.$$

We solve

$$\min_\theta \max_{q \in \mathcal{B}_{\chi^2}(p_{\mathrm{ref}}, \rho)} \sum_{k=1}^K q_k\, L_k(\theta), \quad \mathcal{B}_{\chi^2}(p, \rho) = \Big\{ q \in \Delta^{K-1} : \sum_{k=1}^K \tfrac{(q_k - p_k)^2}{p_k} \le \rho \Big\},$$

with $\rho = \varepsilon_n$ chosen by a radius schedule (below).

Let $a_k = L_k(\theta)$, $\mu = \sum_k p_k a_k$, and $\widetilde{a}_k = a_k - \mu$. The χ^2 inner maximizer moves in the mean-centered, probability-weighted loss direction:

$$q^\star \ =\ \Pi_\Delta\big(p + t\,(p \odot \widetilde{a})\big), \qquad t \ =\ \sqrt{\frac{\rho}{\sum_k \widetilde{a}_k^2\, p_k}},$$

where Π_Δ clips at zero and renormalizes onto the simplex. We guard against $\sum_k \widetilde{a}_k^2 p_k \approx 0$ via a small numerical floor.

Given $q^\star$, the squared-loss gradient is

$$\nabla_\theta \sum_k q_k^\star L_k(\theta) \ =\ \sum_{k=1}^K q_k^\star \cdot \frac{2}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} (v_i^\top \theta - t_i)\, v_i.$$

We use vanilla gradient descent for 500 steps with step size 0.12 per training draw.

Radius schedules and calibration. We compare four schedules:

$$\varepsilon_n \in \Big\{ \tfrac{\chi^2_{K-1,\,0.50}}{n},\ \tfrac{\chi^2_{K-1,\,0.90}}{n},\ \tfrac{\chi^2_{K-1,\,0.95}}{n},\ c\,n^{-2} \Big\}, \quad c = 0.7.$$

The first three are *calibrated* using Pearson's statistic for multinomials. For $C \sim \mathrm{Mult}(n, p^\circ)$,

$$D_n \ =\ n \sum_{k=1}^K \frac{(\widehat{p}_k - p_k^\circ)^2}{p_k^\circ} \ \overset{\text{approx}}{\sim}\ \chi^2_{K-1},$$

so $\Pr\{D_n \le \chi^2_{K-1,\alpha}\} \approx \alpha$. Hence $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ targets α-coverage for $\mathcal{B}_{\varepsilon_n}(p^\circ)$. We compute χ^2-quantiles via the Wilson–Hilferty transform, $\chi^2_{m,\alpha} \approx m\,(1 - 2/(9m) + z_\alpha \sqrt{2/(9m)})^3$, with $z_\alpha = \Phi^{-1}(\alpha)$ obtained by Acklam's inverse-normal approximation.

Coverage curves (Fig. 4). For $n \in \{1000, 2000, 4000, 8000, 16000\}$ and each schedule, we repeat $R_{\mathrm{cover}} = 120$ times: draw $C \sim \mathrm{Mult}(n, p^\circ)$, compute D_n, and record $\mathbb{I}\{D_n \le \varepsilon_n\}$. We report the mean and ± 1.96 standard errors (SE).

Rate overlay (Fig. 5). For each n and schedule we train 8 independent models (separate seeds) and report the mean parameter error $\|\widehat{\theta}-\theta^\star\|_2$ with ± 1.96 SE. We add straight guides of slope $-\frac{1}{2}$ and $-\frac{1}{4}$, anchored at the *rightmost* n to avoid misleading vertical offsets on a log–log plot. Legend slopes are least-squares fits of $\log\|\widehat{\theta}-\theta^\star\|_2$ on $\log n$.

Risk–coverage frontier (Fig. 6). At fixed $n = 16000$ we sweep $\varepsilon = c/n$ over 25 evenly spaced $c \in [0, \chi^2_{K-1,0.99}]$. For each c: (i) estimate coverage via $R_{\text{cover}} = 400$ repetitions; (ii) train 8 models; (iii) evaluate L_k on an i.i.d. set of size 25,000 from p°; (iv) compute the *excess worst-case risk* against χ^2-mixture shifts,

$$\mathrm{Excess}(c) \;=\; \max_{q\in\mathcal{B}_{\chi^2}(p^\circ,\,\varepsilon)} \sum_k q_k L_k \;-\; \sum_k p^\circ_k L_k, \qquad \varepsilon = \tfrac{c}{n}.$$

We then plot $\mathrm{Excess}(c)$ versus empirical coverage.

Implementation and numerical safeguards.

- **Quantiles.** Φ^{-1} uses Acklam's approximation; χ^2-quantiles use Wilson–Hilferty—removing special-function dependencies and improving reproducibility.
- **Inner maximizer.** We use the direction $p \odot (L-\mu)$; if $\sum_k \widetilde{a}_k^2 p_k < 10^{-12}$ we return $q = p$. We clip and renormalize to Δ^{K-1}.
- **Optimization.** Gradient descent with 500 steps and step size 0.12 on the group-weighted squared loss; group index sets $\mathcal{I}_k$ are precomputed.
- **Determinism.** All RNGs are explicitly seeded via simple affine functions of n and the seed index (`1000+17*s+n`, `9999+s+n`, etc.).
- **Uncertainty bands.** Shaded regions depict ± 1.96 SE across repetitions (coverage) or seeds (errors/risks).

Calibration via Pearson and Wilson–Hilferty. Under $H_0 : p = p^\circ$, Pearson's statistic $D_n = n\sum_k \frac{(\widehat{p}_k - p^\circ_k)^2}{p^\circ_k}$ converges in law to χ^2_{K-1}. Hence $\Pr\{D_n \le \chi^2_{K-1,\alpha}\} \approx \alpha$, and $\varepsilon_n = \chi^2_{K-1,\alpha}/n$ delivers the desired coverage for $\mathcal{B}_{\varepsilon_n}(p^\circ)$. We compute $\chi^2_{K-1,\alpha}$ via the Wilson–Hilferty approximation

$$\chi^2_{m,\alpha} \approx m\Big(1 - \tfrac{2}{9m} + z_\alpha\sqrt{\tfrac{2}{9m}}\Big)^3, \quad z_\alpha = \Phi^{-1}(\alpha),$$

with z_α evaluated by Acklam's inverse-normal routine for numerical stability.

M.2 Emotion Alignment Setup

Reward Model Training. The reward model, which serves to quantify emotion-specific preferences, was trained on the "emotion" dataset Saravia et al. [2018]. This dataset comprises text samples annotated with single-class labels across six distinct emotion categories: joy, sadness, love, anger, fear, and surprise. The raw text data was preprocessed by tokenization, with sequence lengths capped at a maximum as defined in shared training configurations (e.g., 68 tokens). The original single-label emotion classifications were directly utilized as targets for the reward model.

For the reward model architecture, we employed a standard GPT-2 model (`GPT2LMHeadModel`) fine-tuned for sequence classification by attaching an `AutoModelForSequenceClassification` head. This head processes the last token's representation to output logits corresponding to the emotion classes. The model was trained using a standard multi-class classification loss, which is implicitly Cross-Entropy Loss when using `AutoModelForSequenceClassification` with multiple labels. Training was conducted over 8 epochs. Optimization was performed with the AdamW optimizer using a learning rate of 5.0×10^{-5}. Common training arguments, including a `per_device_train_batch_size` (e.g., 64, with potential gradient accumulation to reach an effective batch size), `gradient_accumulation_steps`, `fp16` precision, `warmup_steps`, `logging_steps`, and `evaluation_strategy`, were configured via a shared dictionary (`TRAINER_ARGS_COMMON`). Model performance was monitored by `eval_f1_score` (weighted average), which was set as the metric for selecting the best model. The trained reward model achieved a test accuracy of 87% and a test ROC-AUC score of 0.99. The class-wise probability scores predicted by this model were subsequently utilized as our scalar rewards (r_{emotion}) for the preference alignment process.

Supervised Fine-Tuning (SFT). We selected a GPT-2 model (`GPT2LMHeadModel`) as our base language model. It was trained to predict the next token given preceding context from the emotion dataset. Text samples were tokenized and truncated to a maximum sequence length, typically 68 tokens, as defined by `MAX_SEQ_LENGTH` in our training configuration. The SFT model was trained for 10 epochs using the AdamW optimizer with a learning rate of 5.0×10^{-7}. The training schedule included 12 warmup steps, where the learning rate gradually increased to its peak. To enhance training stability and prevent exploding gradients, a maximum gradient norm of 10 was applied during optimization. This SFT-trained model served as both the initial policy (π_0) and the fixed reference policy (π_{ref}) for all subsequent DPO and REBEL variant training runs.

Data Generation for Alignment. A preference dataset for Emotion Alignment was dynamically constructed during the training iterations of each alignment algorithm. Each data point consisted of a prompt and two generated completions, paired with a preference label. The detailed data generation process was as follows:

- **Prompts:** Prompts were directly sampled from the 'text' field of the emotion dataset's training split, ensuring they were drawn from the same domain as the SFT model's training data.
- **Completion Generation:** For each prompt, two distinct completions (a_1 and a_2) were generated by the current policy model (π_θ). Text generation employed sampling-based decoding with specific parameters: `do_sample=True`, `top_k=50`, `top_p=0.95`, and a `temperature=0.7`. Each completion was constrained to a maximum length corresponding to the `max_seq_length` used during SFT (e.g., 68 tokens), ensuring consistency.
- **Reward Calculation:** The generated completions (a_1 and a_2) were then evaluated by the pre-trained emotion reward model. This yielded emotion-specific scores for each completion. These scores were combined into a single scalar reward (r_{a_1}, r_{a_2}) using a configurable mixing function, either "convex" or "geometric", parameterized by a specific α_0 value. This allowed for emphasis on particular emotions or combinations thereof.
- **Preference Labeling:** Instead of deterministic selection, a binary preference label ('preference', typically 0 or 1) was assigned to the pair (a_1, a_2). This was not a deterministic selection based on the mixed reward, but rather a stochastic process following a Bradley-Terry model. Specifically, a random number was drawn, and if it was less than $p = \frac{\exp(r_{a_1})}{\exp(r_{a_1})+\exp(r_{a_2})}$, then a_1 was marked as preferred (preference = 1); otherwise, a_2 was preferred (preference = 0).

REBEL and DPO Variant Training. We conducted comprehensive experiments comparing seven distinct preference alignment algorithms: Direct Preference Optimization (DPO), Wasserstein Distributionally Robust DPO (WDPO), KL Distributionally Robust DPO (KL-DPO), Reinforcement Learning via Regressing Relative Rewards (REBEL), Wasserstein Distributionally Robust REBEL (W-REBEL), KL Distributionally Robust REBEL (KL-REBEL), and Chi-squared Distributionally Robust REBEL (χ^2-REBEL).

Each variant was trained for 40 iterations (epochs). In each iteration, a fresh batch of 64 new data points (prompt-completion pairs with preferences/rewards) was collected using the dynamic data generation process described above. The policy model parameters were optimized using the AdamW optimizer with a fixed learning rate of 5.0×10^{-7}. A DPO β parameter of 0.1 was consistently applied across all DPO and its DRO variants. Algorithm-specific robustness hyperparameters, including REBEL's η (set to 0.01), WDPO/W-REBEL's ρ_0, KL-DPO/KL-REBEL's τ, and χ^2-REBEL's ρ, were configured through shared experiment settings. All Emotion Alignment experiments were executed on a single NVIDIA A100 GPU with 40 GB VRAM. To accommodate the chosen batch size and model requirements, gradient accumulation was performed over two steps per optimization update.

M.3 ArmoRM Multi-objective Alignment Setup

For ArmoRM Multi-objective Alignment, our experimental design focused on scenarios where pre-trained models are aligned to multiple, potentially conflicting, objectives, leveraging sophisticated reward signals derived from a specialized ArmoRM reward model.

Reward Model and SFT. Distinct from the Emotion Alignment setup, the ArmoRM configurations did not involve separate training of a reward model or explicit supervised fine-tuning (SFT) of a base language model for the specific alignment task since the use of a foundational model like Meta LLaMA-3.2-1B-Instruct has already undergone extensive pre-training on vast text corpora, followed by multiple rounds of SFT and preliminary alignment on broad human preference datasets. These pre-aligned models are intrinsically capable of generating responses reflecting general human preferences and providing granular, multiobjective reward scores across various axes such as helpfulness, harmlessness, truthfulness, and conciseness.

Data Generation for Alignment. The preference dataset for ArmoRM alignment was constructed by sampling prompt-completion pairs from large, diverse datasets designed for evaluating instruction-following and safety, specifically a subset of the publicly available HelpSteer2 dataset Wang et al. [2024b]. These prompts typically consisted of user queries, instructions, and open-ended questions designed to elicit varied and complex responses. The generation process for candidate completions for alignment was configured as follows:

- **Completions:** For each sampled prompt, two distinct candidate completions were generated by the current policy model. Text generation employed sampling-based decoding to encourage diversity and creativity, utilizing specific parameters: a `temperature` of 0.7 (to balance creativity with coherence), a `top_p` of 1.0 (to allow for maximal diversity in token sampling), and a maximum generation length of up to 1024 new tokens, enabling the generation of comprehensive and elaborate responses. Prompts themselves were also truncated to a maximum of 1024 tokens before being fed to the model.
- **Multi-objective Rewards:** These generated prompt-completion pairs were then input into the *first stage* of a pre-existing ArmoRM model Wang et al. [2024b]. This "first stage" is a multi-headed reward architecture designed to output a comprehensive vector of scores, quantifying a completion's performance across several predefined objectives (e.g., helpfulness, harmlessness, creativity, factuality). The chosen and rejected completions within each pair were determined based on a composite mixed metric derived from these multi-objective reward vectors.

REBEL and DPO Variant Training. Given the substantial size of the models and datasets involved, these experiments were performed on a high-performance distributed computing setup comprising 8xH100 GPUs. Training leveraged the DeepSpeed framework for efficient memory management and optimized distributed training. Specifically, we primarily utilized DeepSpeed ZeRO-2 (Zero Redundancy Optimizer Stage 2) Rajbhandari et al. [2020] for parameter, gradient, and optimizer state partitioning across GPUs. This included features like `overlap_comm` for overlapping computation and communication, and `contiguous_gradients` for memory efficiency. The training process was configured for automatic mixed precision, with `bf16` enabled for bfloat16 training and `fp16` also available (with a `loss_scale` of 512). DeepSpeed automatically managed the optimizer parameters (learning rate, betas, epsilon, weight decay) and the `WarmupDecayLR` scheduler, as well as `gradient_accumulation_steps` and `gradient_clipping`.

M.4 Wasserstein Variants (WDPO, W-REBEL) Implementation Details

Recall that regularization term in Algorithm 2 is defined as $R(\pi_\theta; D) = \rho_0(\mathbb{E}_{z\sim D}\|\nabla_z l(z;\theta)\|_2^2)^{1/2}$, where $l(z;\theta)$ is the pointwise loss. In a distributed LLM training setting, computing the exact expectation over the entire data distribution D for this regularizer, or even accurately averaging gradient norms over small, local micro-batches, presents a key implementation challenge. A naive approach of averaging gradient norms over local micro-batches can lead to a highly noisy and unstable gradient penalty due to the typically small number of samples per GPU.

To mitigate this instability and ensure tractability, we used the trick utilized by Xu et al. [2025] which exploits the inequality $\sqrt{x} \le x$ for $x \ge 1$. This allows us to upper bound the regularizer. This leads to a tractable approximation of the pointwise WDPO loss: $l_W(z_i, \rho_0) = l(z_i;\theta) + \rho_0\|\nabla_z l(z_i;\theta)\|_2^2$, where $l(z_i;\theta)$ denotes the standard DPO or REBEL loss for sample z_i.

For computing $\|\nabla_z l(z_i;\theta)\|_2^2$, gradient tracking was enabled on the input embeddings of the policy model using `requires_grad=True` in `get_log_probs_and_input_embeddings`. This is because since we cannot directly compute $\nabla_z l(z;\theta)$ since our input is tokenized as integers. The `torch.autograd.grad` function was then used to calculate the gradient of the pointwise loss $l(z_i;\theta)$ with respect to these differentiable input representations. The sum of squared norms of these gradients was calculated for each sample. This term, scaled by ρ_0, was directly incorporated as a penalty into the total loss for each sample, effectively regularizing the policy towards smoother loss landscapes.

M.5 KL Variants (KL-DPO, KL-REBEL) Implementation Details

Recall that the re-weighting factor for each sample i, $P(i)$, was calculated proportional to $\exp\left(\frac{1}{\tau_{eff}}(l(z_i;\theta) - \text{mean}(l(z_j;\theta)))\right)$, where $l(z_i;\theta)$ is the pointwise loss for sample i in Algorithm 3. Critically, $\text{mean}(l(z_j;\theta))$ represents the average pointwise loss computed over the global batch across all participating GPUs. To achieve this global consistency, each GPU first computes the pointwise losses for its local mini-batch. Then, a synchronization step involving a `torch.distributed.all_gather` operation is performed. This operation collects all individual losses from all workers onto every GPU, allowing each GPU to compute the exact global mean of

$l(z_j; \theta)$ across the entire distributed batch. This ensures that the re-weighting factors $P(i)$ are consistent and correctly reflect the global worst-case distribution. $\tau_{eff} = \max(\tau, 1e-6)$ ensures numerical stability. The total loss for these methods was then computed as a weighted sum of the individual losses, $\sum P(i) \cdot l(z_i; \theta)$.

M.6 χ^2 Variant (χ^2-REBEL) Implementation Details

This method seeks to find a robust policy by optimizing against an ambiguity set defined by χ^2-divergence. The optimization involves determining an optimal dual variable, η^*, at each training step. This is achieved by searching over a set of candidate η values, including unique individual loss values and boundary points, to identify the one that minimizes the expression $\eta + \sqrt{\frac{2\rho}{n} \sum(\text{loss}_i - \eta)_+^2}$, where $(\cdot)_+ = \max(\cdot, 0)$. Implementing this in a distributed setting presented a significant engineering challenge, particularly in ensuring global consistency for the η^* search. At each step, each GPU first computes its `individual_ell_losses` for its local mini-batch. To enable the global search for η^*, these local loss tensors are then gathered from all GPUs onto every GPU using a `torch.distributed.all_gather` operation, creating a `global_ell_losses` tensor on each rank. The `_find_eta_star` method then executes on this `global_ell_losses` tensor independently on each GPU; since all GPUs possess identical global loss data, they deterministically identify the same η^* value. This process of efficiently finding the optimal dual variable across distributed data, without excessive communication, was one of the most difficult parts of the implementation. The term $\sum(\text{loss}_i - \eta^*)_+^2$ in the expression for η^* and for deriving λ^* requires a sum over all samples across all GPUs, which is achieved using a `torch.distributed.all_reduce` operation on the locally computed sums. Once η^* is found, a corresponding λ^* is derived. Finally, the gradients for the policy model parameters are computed as a weighted sum of the gradients of individual losses, where the weights $w_i = (\text{loss}_i - \eta^*)_+ / (n \cdot \lambda^*)$ emphasize samples contributing most to the robust objective. These weighted gradients are directly applied to the model's parameters, with DeepSpeed's ZeRO-2 optimizer handling the implicit aggregation across all GPUs during its optimization step.