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

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—and demonstrate DRO–REBEL's superior robustness and sample efficiency on both synthetic benchmarks and language model alignment tasks.

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 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. 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. It should be noted that REBEL is much more sample efficient compared to methods like DPO and PPO.

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:

- **Unified robust updates via duality.** We extend REBEL to type-p Wasserstein, KL and χ^2 ambiguity sets, showing that each robust policy update admits a simple relative-reward regression via Fenchel (or strong) duality, preserving REBEL's scalability and eliminating heuristic tweaks.
- **Sharper "slow-rate" guarantees.** Under standard linear-reward and log-linear-policy assumptions plus a data-coverage condition, we prove an $O(n^{-1/4})$ estimation-error bound for every DRO–REBEL variant. By replacing logistic links with linear regression, we eliminate hidden exponential factors and tighten constants over prior DRO-DPO analyses [Xu et al., 2025, Mandal et al., 2025].
- **Minimax-optimal "fast" rates.** Through a localized Rademacher-complexity argument, we recover the optimal $O(n^{-1/2})$ parametric convergence rate for all DRO–REBEL variants, closing the gap between robust and non-robust RLHF methods.
- **Improved curvature and stability.** We show DRO–REBEL's strong-convexity modulus scales only with the data-coverage constant λ and step-size η, avoiding the exponential sensitivity to the Bradley–Terry curvature in WDPO/KLDPO [Xu et al., 2025]. This yields uniformly smaller excess-risk bounds and more stable updates.
- **Accelerated robust DPO analysis.** We extend our fast-rate machinery to distributionally robust Direct Preference Optimization (WDPO and KLDPO), proving an $O(n^{-1/2})$ estimation-error rate, improving on the $O(n^{-1/4})$ rates in prior DRO-DPO theory [Xu et al., 2025, Mandal et al., 2025].

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. $\mathcal{S}, \mathcal{A}, \mathcal{Z}$. When we refer to the norm $||x||$, we are referring to the Euclidean norm. 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 $a = O(b)$.

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 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 $\mathcal{Z}$.

Definition 2.2 (Kullback-Leibler (KL) Divergence)**.**

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

Definition 2.3 (Chi-Squared Divergence)**.** *If $\mathbb{P} \ll \mathbb{Q}$,*

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

Using these, we can define our ambiguity sets as follows

Definition 2.4 (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 distributions i.e. type-p Wasserstein, KL, χ^2.

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:

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

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

2.4 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 $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ 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. 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:

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

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. Adapting REBEL for distributionally robust RLHF is particularly appealing because it offers both theoretical and practical advantages over existing methods like PPO and DPO. Whereas PPO relies on heuristic mechanisms (e.g., clipping, value baselines) and DPO requires strong assumptions about preference modeling, REBEL reduces policy optimization to a sequence of regression problems on relative rewards—eliminating the need for explicit value functions or constrained optimization. This simplicity translates into significantly lower 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.

2.5 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

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

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

$$\pi^*(y \mid x)=\frac{1}{Z(x)} \pi_{\mathrm{SFT}}(y \mid x) \exp \left(\frac{1}{\beta} r(x, y)\right)$$

where $Z(x)=\sum_y \pi_{\mathrm{SFT}}(y \mid x) \exp \left(\frac{1}{\beta} r(x, y)\right)$. Taking logs and moving terms around, we get

$$r(x, y)=\beta \frac{\pi(y \mid x)}{\pi_{\mathrm{SFT}}(y \mid x)}+\beta \log Z(x)$$

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.

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 $a^1, a^2 \stackrel{\text{i.i.d}}{\sim} \pi_{\mathrm{SFT}}(\cdot \mid x)$. 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 $y \sim \operatorname{Ber}\left(\mathcal{P}^*\left(a^1 \succ a^2 \mid \cdot\right)\right)$. We will denote $z=\left(x, a^1, a^2, y\right) \in \mathcal{Z}$ 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. $z_i \sim \mathbb{P}^{\circ}$.

3.1 Distributionally Robust REBEL

From the REBEL update (Equation (1)), we define the pointwise loss as follows

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

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:

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

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:

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

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 2.5), we define the following pointwise loss:

$$\ell_{\text{DPO}}(z;\theta) = -y\log\sigma(\beta h_\theta) - (1-y)\log\sigma(-\beta h_\theta)$$

where $h_\theta(s, a^1, a^2) := \log\frac{\pi_\theta(a^1|s)}{\pi_{\text{SFT}}(a^1|s)} - \log\frac{\pi_\theta(a^2|s)}{\pi_{\text{SFT}}(a^2|s)}$ 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:

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

We define the following D-DPO loss function:

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

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 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 settings involving neural network-parameterized policy or reward classes, commonly found in modern RLHF for LLM alignment. In such cases, the key properties of the loss function $\ell(z;\theta)$ (e.g., uniform boundedness,* *Lipschitz continuity, and strong convexity with respect to θ) would require corresponding assumptions on the neural network architectures, such as twice differentiability, smoothness (bounded Hessian), and specific architectural choices (e.g., linear output layers for strong convexity) that ensure these properties hold for the composite function.*

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*

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

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 1

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. For the full proof, we refer readers to Appendix C. At a high level, we first prove that $\ell(z; \theta)$ is uniformly bounded and is $4K_g/\eta$-Lipschitz in θ where $K_g = 4B/\eta + 2F$. 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 17.

Strong duality of Wasserstein DRO [Gao and Kleywegt, 2022] (Corollary 4) 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 □

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 1

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

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

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, F is from the Assumption 1, and η is from the step size defined in the DRO update 1.

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 17 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 hidden logistic factors. 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

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

where $\bar{\lambda}$ bounds the dual multiplier and L bounds the loss. In prior KLDPO analyses, the dependence on $\exp(K_g^2/\bar{\lambda})$ 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 and refining 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.

Theorem 4 (Parametric $n^{-1/2}$-rate for DRO-REBEL)**.** *Let the empirical and population minimizers be defined as:*

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

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. ***Lipschitz Loss:*** *The pointwise loss function $\ell(z;\theta)$ is L_g-Lipschitz with respect to θ for all z.*
3. ***Bounded Loss:*** *The pointwise loss $\ell(z;\theta)$ is uniformly bounded by K_ℓ for all $z \in \mathcal{Z}, \theta \in \Theta$.*
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 any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, the estimation error is bounded as:

$$\|\hat{\theta}_n - \theta^*\|_2 \le \frac{16 c_0 L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{8 c_1 L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8\Delta_n}{\alpha}} + \sqrt{\frac{4 c_2 K_\ell \log(1/\delta)}{\alpha n}}$$

where the universal constants c_0, c_1, c_2 arise from concentration inequalities. In particular, if the dual remainder bound has rate $\Delta_n = O(n^{-1})$, then the estimator achieves a parametric rate of:

$$\|\hat{\theta}_n - \theta^\star\|_2 = O\left(n^{-1/2}\right)$$

Proof Sketch of Theorem 4. For the full proof, we refer readers to Appendix F. At a high level, for fixed $\theta \in \Theta$, by using the triangle-inequality and the dual-remainder bound, we get $\left|\mathcal{L}^D(\theta) - \mathcal{L}_n^D(\theta)\right| \;\le\; 2\,\Delta_n + \left|M(\theta) - M_n(\theta)\right|$. Next, let $r = \|\theta-\theta^*\|_2$ and define $\mathcal{F}_r = \big\{\ell(\cdot,\theta) - \ell(\cdot,\theta^*) : \|\theta-\theta^*\| \le r\big\}$. By a standard symmetrization (Lemma 10), Dudley entropy-integral (Corollary 5), and Bousquet's inequality (Theorem 9) argument, one shows with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}_r}\left|P_n f - P f\right| = O\big(\Delta_n + L_g\, r\sqrt{(d+\ln(1/\delta))/n} + K_\ell\, \ln(1/\delta)/n\big)$$

Let $\mathcal{E}_k = \left\{\sup_{\|\theta-\theta^*\|_2\le r_k}\left|\mathcal{L}_n^D\left(\theta;\varepsilon_n\right) - \mathcal{L}^D\left(\theta;\varepsilon_n\right)\right| \le \psi_n(r_k,\delta_k)\right\}$ where

$$\psi_n(r,\delta) = O\big(\Delta_n + L_g\, r\sqrt{(d+\ln(1/\delta))/n} + K_\ell\, \ln(1/\delta)/n\big)$$

We can relate this uniform bound to the estimation error $\widetilde{r} = \|\hat{\theta}_n - \theta^*\|_2$ using a **localization argument**. We define a dyadic sequence of radii $r_k = 2^k r_0$ for $k = 0, 1, \ldots, \lceil \log_2\left(R_{\max}/r_0\right)\rceil$. Using a union bound over these radii, we establish that the uniform convergence event holds simultaneously for all balls $B(\theta^*, r_k)$ with high probability (e.g., $1-\delta$). Let k^* be the smallest integer such that $\widetilde{r} \le r_{k^*}$. By this very definition of k^*, $\hat{\theta}_n \in B(\theta^*, r_{k^*})$. Furthermore, if $k^* > 0$, then $r_{k^*-1} < \widetilde{r} \le r_{k^*}$. Since $r_{k^*} = 2\cdot r_{k^*-1}$ (for a dyadic sequence), this implies $r_{k^*} < 2\widetilde{r}$. For $k^* = 0$, $\widetilde{r} \le r_0$. In general, we can use $r_{k^*} \le 2\max(\widetilde{r}, r_0)$ for the linear term of ψ_n, which simplifies to $2\widetilde{r}$ when $\widetilde{r}$ is large enough (as expected asymptotically). Combining this with the uniform convergence bound for r_{k^*}, α-strong convexity of $\mathcal{L}^D(\theta)$ around θ^*, and the definition of the empirical minimizer, we arrive at a quadratic inequality in $\widetilde{r}$:

$$\frac{\alpha}{2}\widetilde{r}^2 - \left(4c_0 L_g\sqrt{\frac{d}{n}} + 2c_1 L_g\sqrt{\frac{\log(1/\delta)}{n}}\right)\widetilde{r} - \left(4\Delta_n + 2c_2\frac{K_\ell}{n}\log(1/\delta)\right) \le 0$$

Solving this quadratic inequality for $\widetilde{r}$ yields

$$\widetilde{r} = O\left(\frac{L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\Delta_n}{\alpha}} + \sqrt{\frac{K_\ell}{\alpha}\frac{\log(1/\delta)}{n}}\right)$$

Taking $\Delta_n = O(n^{-1})$, all terms in the bound become $O(n^{-1/2})$, which establishes the claimed convergence rate. $\square$

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 G, H, and I respectively.

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

$$\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \le \frac{16c_0K_g\eta}{\lambda}\sqrt{\frac{d}{n}} + \left(\frac{8c_1K_g\eta}{\lambda} + \frac{2K_g\eta\sqrt{c_2}}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{n}} + \frac{2\eta\sqrt{L_{\ell,z}}}{\sqrt{\lambda n}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, $L_{\ell,z}$ 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 $\delta \in (0,1)$. Then, with probability at least $1-\delta$*

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \le \frac{16c_0K_g\eta}{\lambda}\sqrt{\frac{d}{n}} + \left(\frac{8c_1K_g\eta}{\lambda} + \frac{2K_g\eta\sqrt{c_2}}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{n}} + \frac{2^{3/4}\eta K_g}{\sqrt{\lambda n}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, 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 $\delta \in (0,1)$. Then, with probability at least $1-\delta$*

$$\|\hat{\theta}_n^{\chi^2} - \theta^{\chi^2}\|_2 \le \frac{16c_0K_g\eta}{\lambda}\sqrt{\frac{d}{n}} + \left(\frac{8c_1K_g\eta}{\lambda} + \frac{2K_g\eta\sqrt{c_2}}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\sqrt{2}\eta K_g}{\sqrt{\lambda n}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, 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.

Our fast-rate analysis shows that DRO-REBEL achieves the minimax-optimal $O(n^{-1/2})$ estimation error, even under a Wasserstein, KL, or χ^2 ambiguity set, simply by combining strong convexity with a localized Rademacher complexity argument. This matches the classical parametric rate in M-estimation theory and parallels the finite-sample guarantees for generic Wasserstein DRO obtained by Gao [2022]. Crucially, the same localized-complexity machinery could be applied to other DRO settings under similar assumptions. In this way, DRO-REBEL both tightens constants in the slow regime and offers a clear recipe for restoring the optimal $n^{-1/2}$ convergence rate for common robust preference-learning algorithms.

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 (no logistic link) 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 5 (Parametric $O(n^{-1/2})$ Rate for Wasserstein DPO)**.** *Let the Wasserstein DPO empirical and population minimizers be defined as:*

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

where $\mathcal{L}^{\mathcal{W}_p}$ is the Wasserstein distributionally robust DPO objective with ambiguity radius ε_n. 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:

$$\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \leq \frac{32c_0\beta}{\gamma\lambda}\sqrt{\frac{d}{n}} + \frac{16c_1\beta}{\gamma\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8L_{\ell,z}}{\gamma\lambda n}} + \sqrt{\frac{4c_2\log(1+e^{4\beta B})\log(1/\delta)}{\gamma\lambda n}}$$

where $\gamma = \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$ is the DPO curvature constant, c_0, c_1, c_2 are universal constants from concentration inequalities, and $L_{\ell,z}$ is from Assumption 5.

Theorem 6 (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 ε_n. 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:

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \leq \frac{32c_0\beta}{\gamma\lambda}\sqrt{\frac{d}{n}} + \frac{16c_1\beta}{\gamma\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8\log(1+e^{4\beta B})}{\gamma\lambda n}} + \sqrt{\frac{4c_2\log(1+e^{4\beta B})\log(1/\delta)}{\gamma\lambda n}}$$

where $\gamma = \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$ and c_0, c_1, c_2 are universal constants.

We defer readers to the Appendix for the proofs, particularly Appendix J and K

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. Although we have not explicitly stated the exact parametric rate for χDPO, using a similar analysis to that of Wasserstein DPO and KL-DPO, we can recover a parametric $O(n^{-1/2})$ rate. 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

$$|J(\pi') - J(\pi)| \leq V_{\max}\mathbb{E}_{x\sim\rho}\|\pi'(\cdot \mid x) - \pi(\cdot \mid x)\|_1$$

Proof. First notice that by definition $\mathbb{E}_{a\sim\pi(\cdot|x)}\left[A^\pi(x,a)\right] = 0$ ($\star$). 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 inequality holds from using $\star$ 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 discrete and countable. Then $\star$ implies that $\sum_{a\in\mathcal{A}} A^\pi(x,a)\pi(a \mid x) = 0$. Using this, we have

$$\begin{aligned}
\mathbb{E}_{x\sim\rho}\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right] &= \mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{a\sim\pi'}\left[A^\pi(x,a)\right] - \mathbb{E}_{a\sim\pi}\left[A^\pi(x,a)\right]\right] \\
&= \mathbb{E}_{x\sim\rho}\left[\sum_{a\in\mathcal{A}} A^\pi(x,a)\left(\pi'(\cdot \mid x) - \pi(\cdot \mid x)\right)\right] \\
&\leq V_{\max}\mathbb{E}_{x\sim\rho}\|\pi'(\cdot \mid x) - \pi(\cdot \mid x)\|_1
\end{aligned}$$

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 $A^\pi(x,a) \leq 2F$ so we can take $V_{\max} = 2F$. $\square$

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.*

Proof. 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

$$\bigl\|f(\theta') - f(\theta)\bigr\|_1 \leq \bigl\|\nabla f(\theta + t\Delta)\bigr\|_{1\to1}\,\|\Delta\|_2,$$

where $\|\cdot\|_{1\to1}$ is the operator norm from ℓ_2 to ℓ_1. Under Assumption 2, $\|\psi(x,a)\|_2 \leq 1$. 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 $u \in \mathbb{R}^d$,

$$\begin{aligned}
\bigl\|\nabla f(\theta)\,u\bigr\|_1 &= \sum_a\Bigl|\langle\nabla_\theta\,\pi_\theta(a \mid x),\,u\rangle\Bigr| \\
&\leq \sum_a \pi_\theta(a \mid x)\,\|\psi(x,a) - \mathbb{E}[\psi]\|_2\,\|u\|_2 \\
&\leq \bigl(\max_a\|\psi(x,a)\|_2 + \|\mathbb{E}[\psi]\|_2\bigr)\,\|u\|_2 \\
&\leq 2\|u\|_2
\end{aligned}$$

Thus $\|\nabla f(\theta)\|_{1\to1} \leq 2$. Combining the above,

$$\|\pi_{\theta'}(\cdot \mid x) - \pi_\theta(\cdot \mid x)\|_1 \leq 2\,\|\theta' - \theta\|_2,$$

$\square$

Combining these results, we get the following sample complexity result

Theorem 7 (Sample Complexity Result For χDPO). *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*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max} C_{\beta,B,c_0,c_1,c_2,\gamma,\lambda} \sqrt{\frac{d + \log(1/\delta)}{n}}$$

where $V_{\max} = \sup_{x,a} A^{\pi}(x,a) = 2F$ and $C_{\beta,B,c_0,c_1,c_2,\gamma,\lambda}$ is a universal constant.

Proof. First we can combine Lemma 2 and 3 to deduce that

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max} \|\theta^* - \hat{\theta}\|_2$$

From Corollary 3 and Theorem 6, we know that

$$\begin{aligned}
\|\hat{\theta}_n^{\chi^2} - \theta^{\chi^2}\|_2 &\leq C_{c_0,\beta,\gamma,\lambda,B} \sqrt{\frac{d}{n}} + C_{c_1,c_2,\beta,\gamma,\lambda,B} \sqrt{\frac{\log(1/\delta)}{n}} \\
&\leq \max\{C_{c_0,\beta,\gamma,\lambda,B}, C_{c_1,c_2,\beta,\gamma,\lambda,B}\} \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \\
&\leq \sqrt{2} \max\{C_{c_0,\beta,\gamma,\lambda,B}, C_{c_1,c_2,\beta,\gamma,\lambda,B}\} \sqrt{\frac{d + \log(1/\delta)}{n}}
\end{aligned}$$

where the last inequality holds from the fact $(\sqrt{a} + \sqrt{b})^2 = a + b + 2\sqrt{ab}$ and the AM-GM inequality. If we let $C_{\beta,B,c_0,c_1,c_2,\gamma,\lambda} = \sqrt{2}\max\{C_{c_0,\beta,\gamma,\lambda,B}, C_{c_1,c_2,\beta,\gamma,\lambda,B}\}$, we conclude

$$J(\pi^*) - J(\widehat{\pi}) \lesssim V_{\max} C_{\beta,B,c_0,c_1,c_2,\gamma,\lambda} \sqrt{\frac{d + \log(1/\delta)}{n}}$$

$\square$

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 8 (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*

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

where

$$J = \max_{x,y^+,y^-} \log \frac{\pi_{\text{ref}}(y^+ \mid x)}{\pi_{\text{ref}}(y^- \mid x)}.$$

Then the average iterate

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

satisfies

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

Now let us assume that $\ell_{\mathrm{DPO,TV}}$ is μ-strongly convex in θ. This assumption is completely reasonable as we have shown in Lemma 17, 19, and 21 that under various ambiguity sets, strong convexity of the robust variant holds. Then by Lemma 7, we have

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

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

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

In contrast, 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 5 and 6) under appropriate choices of the ball radius ε_n (e.g., $\varepsilon_n = O(n^{-1})$ for Wasserstein and $O(n^{-2})$ for KL).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 1: 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 $p = 2$:

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

where $\rho_n = O(1/\sqrt{n})$. 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 η, 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: $L_W(\theta,\rho_0) = L_{\text{REBEL}}(\pi_\theta;\mathcal{D}) + R(\pi_\theta;\mathcal{D})$.
    $\theta \leftarrow \theta - \eta\nabla_\theta L_W(\theta,\rho_0)$.
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 $\mathbb{P} = \sup_{\mathbb{P}:D_{\text{KL}}(\mathbb{P}\|\mathbb{P}_n^\circ)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$. 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 η, 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: $L_{\text{KL}}(\theta;\mathcal{D}) = \sum_{i=1}^n \mathbb{P}(i)\cdot\ell(z_i;\theta)$.
    $\theta \leftarrow \theta - \eta\nabla_\theta L_{\text{KL}}(\theta,\rho)$.
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 $i = 1,\dots,n$ 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*

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

and the supremum is attained by the "re-weighted" empirical distribution

$$w_i \;=\; \frac{(\ell_i - \eta^*)_+}{n\,\lambda^*}, \quad \lambda^* \;=\; \sqrt{\frac{2\varepsilon_n}{n}\sum_{j=1}^{n}(\ell_j - \eta^*)_+^2},$$

where η^* minimizes the inner objective.

We defer the proof to Appendix L. 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

while θ not converged **do**

  Compute pointwise REBEL loss $\ell_i \leftarrow \ell(z_i;\theta)$ and gradients $g_i \leftarrow \nabla_\theta \ell(z_i;\theta)$

  (Inner 1-D solve) find

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

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

  Compute $\lambda^* = \sqrt{\tfrac{2\rho}{n}\sum_{i=1}^{n}(\ell_i - \eta^*)_+^2}$

  Form weights $w_i = \dfrac{(\ell_i - \eta^*)_+}{n\,\lambda^*}$

  Robust gradient: $G \;=\; \sum_{i=1}^{n} w_i\, g_i$

  Gradient step: $\theta \leftarrow \theta - \alpha\, G$

end while

return π_θ

Each outer step of Algorithm 4 has a computational cost of $O(n \log n + nd)$. This complexity arises from the two primary components of the update: the inner one-dimensional solve for the optimal dual variable η^* and the subsequent robust gradient computation and parameter update. While the inner solve is an optimization over the unbounded domain $\eta \in \mathbb{R}$, its specific structure allows for a highly efficient solution. Let the inner objective function be denoted by $f(\eta)$:

$$f(\eta) = \eta + \sqrt{\frac{2\rho}{n}\sum_{i=1}^{n}(\ell_i - \eta)_+^2}$$

where $(\cdot)_+ := \max(0,\cdot)$. This function $f(\eta)$ can be shown to be convex in η. It is the sum of a linear function, η, and a second term which is a composition of the convex L2-norm with a vector of convex functions, $v_i(\eta) = (\ell_i - \eta)_+$. The convexity of $f(\eta)$ ensures the existence of a unique global minimum and implies that its subgradient, $\partial f(\eta)$, is a monotonically non-decreasing operator.

This monotonicity allows for the use of an efficient binary search to find the unique η^* where $0 \in \partial f(\eta^*)$. One can show that the optimal solution is bounded, i.e., $\eta^* \in (-\infty, \max_i \ell_i]$, because for any $\eta > \max_i \ell_i$, the objective simplifies to $f(\eta) = \eta$, which is strictly increasing. The cost of this inner step is therefore dominated by the initial sorting of the loss values $\{\ell_i\}_{i=1}^n$ to identify the function's non-differentiable "breakpoints," which takes $O(n \log n)$ time.

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), 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., projected gradient descent) 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 Conclusion

In this work, we introduced *DRO–REBEL*, a unified family of distributionally-robust variants of the REBEL framework for offline RLHF. By instantiating ambiguity sets via KL, χ^2, and type-p Wasserstein divergences and exploiting strong/Fenchel duality, each robust policy update reduces to a simple relative-reward regression. Under standard linear-policy and data-coverage assumptions, we proved that all DRO–REBEL variants achieve an $O(n^{-1/4})$ "slow" estimation-error rate with substantially tighter constants than prior DRO-DPO methods and, via a localized Rademacher-complexity argument, recover the minimax-optimal $O(n^{-1/2})$ "fast" parametric rate. Our analysis shows that DRO–REBEL not only tightens slow-rate bounds but also restores the classical parametric rate under distributional shifts.

Along the way, we learned that strong/Fenchel duality can collapse complex DRO updates into tractable regressions, that improving constant factors in slow-rate bounds can have outsized practical impact and, perhaps most surprisingly, that worst-case robustness and optimal $O(n^{-1/2})$ convergence need not be at odds. This work also raises several open questions: How should practitioners choose or adapt between KL, χ^2, and Wasserstein ambiguity sets in real RLHF pipelines? Can the same dual-regression approach and fast-rate analysis be extended to neural (nonlinear) policy classes? And how might one jointly integrate robust reward modeling with DRO–REBEL to further bolster alignment under noisy or adversarial feedback?

Looking ahead, we plan to validate these theoretical findings empirically. We will train our robust RLHF algorithms on the Unified-Feedback dataset [Jiang et al., 2023] and evaluate out-of-distribution robustness on established reward-evaluation benchmarks such as Reward-Bench [Lambert et al., 2024], MT-Bench [Zheng et al., 2023], and HHH-Alignment [Askell et al., 2021] With more time, we would also conduct systematic hyperparameter sweeps (e.g. on ε, η, dual bounds), prototype DRO–REBEL with transformer-parameterized policies, and explore adaptive schemes for selecting or mixing ambiguity sets online. We anticipate that these experiments will corroborate our theoretical guarantees, demonstrating superior generalization under preference shifts and resistance to overoptimization, thereby providing a unified, principled methodology for reliably aligning LLMs with diverse and uncertain human preferences.

References

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`.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL `https://arxiv.org/abs/2112.00861`.

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`.

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`.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002. ISSN 1631-073X. doi: https://doi.org/10.1016/S1631-073X(02)02292-6. URL `https://www.sciencedirect.com/science/article/pii/S1631073X02022926`.

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.

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.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality, 2022. URL https://arxiv.org/abs/2009.04382.

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.

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.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023. URL https://arxiv.org/abs/2306.02561.

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

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.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024. URL https://arxiv.org/abs/2403.13787.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

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 4 (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*

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

where

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

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

$$v_p = \sup_{\mu \in \mathcal{P}(\Xi)} \left\{\int_\Xi \Psi(\xi)\, \mu(d\xi) \colon W_p(\mu, \nu) \leq \rho\right\}, \quad \textit{(Primal)} \tag{3}$$

$$v_D = \inf_{\eta \geq 0}\left\{\eta\, \rho^p \; - \; \int_\Xi \inf_{\xi \in \Xi}\left[\eta\, d^p(\xi, \zeta) - \Psi(\xi)\right] \nu(d\zeta)\right\}. \quad \textit{(Dual)} \tag{4}$$

Lemma 5 (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 5 by choosing $L = \operatorname{diam}(\Xi)^p$ and $M = \sup_{\xi \in \Xi} |\Psi(\xi)|$. $\square$

A.2 Optimization

Lemma 6 (Beck [2023], Theorem 1.24; Linear Approximation Theorem)**.** *Let $f : U \to \mathbb{R}$ 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*

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

Lemma 7 (Beck [2017], Theorem 5.24; First-order characterizations of strong convexity)**.** *Let $f : E \to (-\infty, \infty]$ be a proper, closed, convex function, and let $\sigma > 0$. The following are equivalent:*

1. *For all $x, y \in \operatorname{dom}(f)$ and $\lambda \in [0, 1]$,*
$$f\big(\lambda x + (1-\lambda) y\big) \le \lambda f(x) + (1-\lambda) f(y) - \tfrac{\sigma}{2}\,\lambda(1-\lambda)\,\|x-y\|^2.$$

2. *For all $x \in \operatorname{dom}(\partial f)$, $y \in \operatorname{dom}(f)$ and $g \in \partial f(x)$,*
$$f(y) \;\ge\; f(x) \;+\; \langle g, y-x\rangle \;+\; \tfrac{\sigma}{2}\,\|y-x\|^2.$$

Lemma 8 (Beck [2017], Theorem 5.25; Existence and uniqueness of minimizer)**.** *Let $f : E \to (-\infty, \infty]$ be proper, closed, and σ-strongly convex with $\sigma > 0$. Then:*

1. *f has a unique minimizer x^*.*

2. *For all $x \in \operatorname{dom}(f)$,*
$$f(x) - f(x^*) \;\ge\; \tfrac{\sigma}{2}\,\|x - x^*\|^2.$$

A.3 Distributionally Robust Optimization

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

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

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

Lemma 9 (Duchi and Namkoong [2020], Proposition 1)**.** *Let D_f be as in* (15)*. Then*

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

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

A.4 Empirical Process Theory

Lemma 10 (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 $\mathcal{F}$,*

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

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

Corollary 5 (Boucheron et al. [2013], Corollary 13.2; Dudley's Entropy Integral)**.** *Let $\mathcal{T}$ be a finite pseudometric space and let $(X_t)_{t \in \mathcal{T}}$ be a collection of random variables such that, for all $t, t' \in \mathcal{T}$ and all $\lambda > 0$,*

$$\log \mathbb{E}\big[e^{\lambda\,(X_t - X_{t'})}\big] \;\le\; \frac{\lambda^2\, d^2(t, t')}{2}\,.$$

Then for any fixed $t_0 \in \mathcal{T}$, if we set

$$\delta \;=\; \sup_{t \in \mathcal{T}} d(t, t_0) \quad \textit{and} \quad H(u, \mathcal{T}) \;=\; \log N\big(u, \mathcal{T}, d\big)$$

denoting the covering-number entropy at scale u, it holds that

$$\mathbb{E}\Big[\sup_{t \in \mathcal{T}} \big(X_t - X_{t_0}\big)\Big] \;\le\; 12 \int_0^{\delta/2} \sqrt{H\big(u, \mathcal{T}\big)}\;\mathrm{d}u\,.$$

Theorem 9 (Bousquet [2002], Theorem 2.1)**.** *Let $c > 0$, let X_i be independent random variables with distribution P, and let $\mathcal{F}$ be a class of functions $f : X \to \mathbb{R}$. Assume that for all $f \in \mathcal{F}$,*

$$\mathbb{E}[f(X_i)] = 0, \qquad \|f\|_\infty \le c.$$

Let $\sigma > 0$ satisfy

$$\sigma^2 \;\ge\; \sup_{f \in \mathcal{F}} \mathrm{Var}\big(f(X_i)\big).$$

Define

$$Z \;=\; \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i), \qquad v = n\sigma^2 + 2c\,\mathbb{E}[Z], \qquad h(u) = (1+u)\ln(1+u) - u.$$

Then for any $x \ge 0$,

$$\Pr\big(Z \ge \mathbb{E}[Z] + x\big) \;\le\; \exp\Big(-v\,h\big(\tfrac{x}{c\,v}\big)\Big).$$

Moreover, with probability at least $1 - e^{-x}$,

$$Z \;\le\; \mathbb{E}[Z] \;+\; \sqrt{2\,x\,v} \;+\; \frac{c\,x}{3}.$$

A.5 Variational Inequalities

Lemma 11 (van Handel [2016], Lemma 4.10; Gibbs Variational Principle)**.** *Let $\mu, \nu \in \mathcal{P}(\Xi)$ be Borel probability measures supported on Ξ. Then*

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

Theorem 10 ([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*

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

A.6 Concentration Inequalities

Lemma 12 (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 $\lambda \in \mathbb{R}$, $\psi''_Y(\lambda) \le \frac{(b-a)^2}{4}$ 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 13 (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 $t > 0$,

$$\mathbb{P}\left(S \ge t\right) \le \exp\left(-2t^2 / \sum_{i=1}^{n} (b_i - a_i)^2\right)$$

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

$$\mathbb{P}\big(\big|\tfrac{1}{n}\sum_{i=1}^{n} X_i \;-\; \mathbb{E}\left[X\right]\big| \ge t\big) \;\le\; 2\exp\Big(-2nt^2/(b-a)^2\Big).$$

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

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

Lemma 14 (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 14. Since we have that $\pi_\theta, \pi_{\theta_t} \in \Pi$, notice that

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

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

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

where the first inequality holds from Hölder's 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 1. 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 $\ell(z;\theta) = g(z;\theta)^2$, $K_\ell = K_g^2$. ☐

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

Lemma 15 (Lipschitz bound on $\ell(z;\theta)$)**.** *$\ell(z;\theta)$ is $\frac{4K_g}{\eta}$-Lispchitz in θ.*

Proof of Lemma 15. 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 14, 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. ☐

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

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

Lemma 16 (Strong convexity of h)**.** *Let $\ell(z;\theta)$ be the REBEL loss function. Assume that Assumption 3 holds. Then $h(\theta;\mathbb{P}) = \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$ is $2/\eta$-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 16. We begin by computing the Hessian of $\ell(z;\theta)$ with respect to θ. From Lemma 15, 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:

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

From Lemma 15, we know that $\nabla_\theta g(z;\theta) = \frac{1}{\eta}[\psi(x,a)-\psi(x,a')]$. Crucially, this gradient does not depend on θ. Therefore, $\nabla^2_\theta g(z;\theta) = \mathbf{0}$. Substituting this into the Hessian expression, the second term vanishes:

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

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 6), there exists $\alpha \in [0,1]$ and $\widetilde{\theta} = \alpha\theta + (1-\alpha)\theta'$ such that

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

where $\mu = 2/\eta^2$. By Lemma 7, 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 $\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$

Lemma 17 (Strong convexity of $\mathcal{L}^{\mathcal{W}_p}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$ 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 17. In Lemma 16, we proved strong convexity of h. By Lemma 7, 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. □

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 [4], for fixed θ we have

$$
\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]=\inf_{\Delta\geq0}\left\{\delta\varepsilon^p-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}
$$

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:

$$
\begin{aligned}
\left|\mathcal{L}^{\mathcal{W}_p}\left(\theta;\varepsilon\right)-\mathcal{L}_n^{\mathcal{W}_p}\left(\theta;\varepsilon\right)\right| &= \left|\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathcal{W}_p)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]-\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}_n^\circ;\mathcal{W}_p)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]\right| \\
&= \left|\inf_{\Delta\geq0}\left\{\Delta\varepsilon^p-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}-\inf_{\Delta\geq0}\left\{\Delta\varepsilon^p-\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]\right\}\right| \\
&\leq \sup_{\Delta\geq0}\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right|
\end{aligned}
$$

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 14, we showed that $\ell(z;\theta)\in[0,K_l]$. Now notice that

$$
\begin{aligned}
l_\Delta(z;\theta) &= \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-\ell(z';\theta)\right\}\leq\inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')\right\}=0 \quad (\text{since } \ell(z;\theta)\geq0) \\
\ell_\Delta(z;\theta) &= \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-\ell(z';\theta)\right\}\geq\inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z')-K_\ell\right\}\geq-K_\ell \quad (\text{since } d^p\geq0)
\end{aligned}
$$

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 13). For any $\epsilon>0$:

$$
\mathbb{P}\left(\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right|\geq\epsilon\right)\leq2\exp\left(-\frac{2n\epsilon^2}{K_\ell^2}\right)
$$

□

Since the bounds for $\ell_\Delta(z;\theta)$ do not depend on Δ or θ, this bound is uniform over all $\Delta\geq0$ and $\theta\in\Theta$. By setting the right-hand side to δ and solving for ϵ, we find that with probability at least $1-\delta$:

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

Now we have that

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

where the first inequality holds from the fact that $\hat{\theta}_n^{\mathcal{W}_p}\in\arg\min_{\theta\in\Theta}\mathcal{L}_n^{\mathcal{W}_p}\left(\theta;\varepsilon\right)$. Now from Lemma 8 and Lemma 17, 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 18 (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:*

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

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

We will now establish strong convexity of $\mathcal{L}^{\mathrm{KL}}\left(\theta;\varepsilon\right)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$. This proof will essentially be the same as Lemma 17

Lemma 19 (Strong convexity of $\mathcal{L}^{\mathrm{KL}}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\mathrm{KL}}\left(\theta;\varepsilon\right)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ 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 19. In Lemma 16, we proved strong convexity of h. By Lemma 7, 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

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

where the second inequality holds from $\sup_x\left(f(x)+g(x)\right)\le\sup_x f(x)+\sup_x g(x)$, 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 $\mathcal{L}^{\mathrm{KL}}$ is $\mu\lambda$-strongly convex in the $||\cdot||_2$ norm. □

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 [18], we have for fixed θ

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

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

$$
\begin{aligned}
\left|\mathcal{L}^{\mathrm{KL}}\left(\theta;\varepsilon\right)-\mathcal{L}^{\mathrm{KL}}_n\left(\theta;\varepsilon\right)\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}^\circ_n;\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}^\circ_n}\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}^\circ_n}\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}^\circ_n}\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}^\circ_n}\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}^\circ_n}\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}^\circ_n}\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 13). 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

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

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 8 and Lemma 18, we have that

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

Thus we get

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

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

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

□

E Proof of "Slow Rate" χ^2-DRO-REBEL

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

Lemma 20 (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*

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

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 $\sigma^2 = \mathrm{Var}_{\mathbb{P}^\circ}(\ell(z;\theta))$,

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

We again will prove strong convexity for χ^2

Lemma 21 (Strong convexity of $\mathcal{L}^{\chi^2}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\chi^2}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)} \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]$ 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 21. In Lemma 16, we proved strong convexity of h. By Lemma 7, 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

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

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. ☐

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 $\hat{\theta}^{\chi^2}_n \in \arg\min_\theta \mathcal{L}^{\chi^2}_n(\theta;\varepsilon)$. By the dual reformulation (Lemma 20), 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

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

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

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

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 14, the loss function $\ell(z;\theta)$ is bounded within $[0, K_\ell]$. We use Hoeffding's inequality (Lemma 13) to bound the deviations of μ_n and σ_n^2 from their population counterparts. For any $\epsilon_1 > 0$ and $\epsilon_2 > 0$:

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

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.

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

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:

$$|\mu - \mu_n| \le K_\ell\sqrt{\tfrac{\log(4/\delta)}{2n}}$$
$$|\sigma^2 - \sigma_n^2| \le K_\ell^2\sqrt{\tfrac{\log(4/\delta)}{2n}}$$

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

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

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

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

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 8),

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

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

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

□

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

Proof of Theorem 4. Let $M(\theta) = \mathbb{E}_{z\sim\mathbb{P}^\circ}[\ell(z;\theta)]$ and $M_n(\theta) = \mathbb{E}_{z\sim\mathbb{P}_n^\circ}[\ell(z;\theta)]$. First notice that we can do a loss decomposition as follows: for fixed $\theta \in \Theta$

$$\begin{aligned}\left|\mathcal{L}^D(\theta;\varepsilon_n) - \mathcal{L}_n^D(\theta;\varepsilon_n)\right| &\le \left|\mathcal{L}^D(\theta;\varepsilon_n) - M(\theta)\right| + |M(\theta) - M_n(\theta)| + \left|M_n(\theta) - \mathcal{L}_n^D(\theta;\varepsilon_n)\right| \\ &\le 2\Delta_n + \sup_{\|\theta-\theta^*\|_2\le r}|M(\theta) - M_n(\theta)|\end{aligned}$$

where the first inequality holds from triangle inequality and the second holds from the assumption that $\left|\mathcal{L}^D(\theta) - \mathbb{E}_{\mathbb{P}}\big[\ell(z;\theta)\big]\right| \le \Delta_n$ and $\left|\mathcal{L}_n^D(\theta;\varepsilon_n) - \mathbb{E}_{\mathbb{P}_n^\circ}\big[\ell(z;\theta)\big]\right| \le \Delta_n$. Now let us define the following function class

$$\mathcal{F}_r = \{f_\theta(z) = \ell(z;\theta) - \ell(z;\theta^*) \mid \|\theta - \theta^*\|_2 \le r\}$$

Then bounding $\sup_{\|\theta-\theta^*\|_2\le r}|M(\theta) - M_n(\theta)|$ is equivalent to bounding $\sup_{f\in\mathcal{F}_r}|P_nf - Pf|$ where $P_nf - Pf$ corresponds to the empirical process $M(\theta) - M_n(\theta)$. Now by Lemma 10 (Symmetrization), using the notation $\mathcal{R}_n(\mathcal{F}_r) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_r}\frac{1}{n}\sum_{i=1}^n\sigma_if(z_i)\right]$ called the Rademacher complexity, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_r}|P_nf - Pf|\right] \le 2\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}_r)\right]$$

where the expectation is taken with respect to the data z in the expression above whereas in the Rademacher complexity, we condition on the data z and take the expectation over $\sigma_i \overset{\text{i.i.d}}{\sim} \text{Unif}\{-1, 1\}$. Assume that $\ell(z;\theta)$ is L_g-Lipschitz in θ. A standard covering argument gives us

$$N(\epsilon, \mathcal{F}_r, L_2(P)) \leq \left(\frac{3L_g r}{\epsilon}\right)^d$$

Corollary 5 (Dudley's entropy integral) then gives us

$$\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}_r)\right] \leq \frac{12}{\sqrt{n}} \int_0^{L_g r} \sqrt{\log N(\epsilon, \mathcal{F}_r, L_2(P))} d\epsilon \leq c_0 L_g r \sqrt{\frac{d}{n}}$$

where $c_0 = \int_0^1 \sqrt{\log(3/u)} du < \infty$ is an absolute constant. Thus we have that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_r} |P_n f - Pf|\right] \leq 2c_0 L_g r \sqrt{\frac{d}{n}}$$

To upgrade the above expectation bound into a high-probability bound, we will apply Bousquet's inequality (Theorem 9). First notice that since $\ell(z;\theta)$ is L_g-Lipschitz in θ, we have that

$$|f_\theta(z)| = |\ell(z;\theta) - \ell(z;\theta^*)| \leq L_g \|\theta - \theta^*\|_2 \leq L_g r$$

Thus by Popoviciu's inequality on variances, we have that $\text{Var}_{z \sim \mathbb{P}}\left(f_\theta(z)\right) \leq L_g^2 r^2$. Since each $f_\theta \in \mathcal{F}_r$ is uniformly bounded by K_ℓ by assumption and has bounded variance, then by applying Theorem 9 to Rademacher averages of a class, we find that there exists constants $c_1, c_2 > 0$ such that for any $\delta' \in (0,1)$, with probability at least $1 - \delta'$

$$\sup_{f \in \mathcal{F}_r} |P_n f - Pf| \leq 2\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}_r)\right] + c_1 L_g r \sqrt{\frac{\log(1/\delta')}{n}} + c_2 \frac{K_\ell}{n} \log(1/\delta')$$

Plugging in the Dudley integral bound, we get

$$\sup_{f \in \mathcal{F}_r} |P_n f - Pf| \leq 2c_0 L_g r \sqrt{\frac{d}{n}} + c_1 L_g r \sqrt{\frac{\log(1/\delta')}{n}} + c_2 \frac{K_\ell}{n} \log(1/\delta')$$

We use a **localization argument** to avoid assuming $\hat{\theta}_n$ is in a fixed ball from the start. Let $R_{\max} = 2B$ be the maximum possible distance in the parameter space (since $\theta \in \Theta$ and $\|\theta\|_2 \leq B$). Define a sequence of radii $r_k = 2^k r_0$ for $k = 0, 1, \ldots, K_{\max} = \lceil \log_2(R_{\max}/r_0) \rceil$, where r_0 is a small positive constant and $2^{K_{\max}} r_0 \geq R_{\max}$. This creates a series of nested balls $B(\theta^*, r_1) \subset \cdots \subset B(\theta^*, r_{K_{\max}})$ covering Θ. For each k, define an event $\mathcal{E}_k$ that the uniform convergence bound holds for the ball $B(\theta^*, r_k)$:

$$\mathcal{E}_k = \left\{ \sup_{\|\theta - \theta^*\|_2 \leq r_k} \left| \mathcal{L}_n^D(\theta; \varepsilon_n) - \mathcal{L}^D(\theta; \varepsilon_n) \right| \leq \psi_n(r_k, \delta_k) \right\}$$

where

$$\psi_n(r, \delta) = 2\Delta_n + 2c_0 L_g r \sqrt{\frac{d}{n}} + c_1 L_g r \sqrt{\frac{\log(1/\delta)}{n}} + c_2 \frac{K_\ell}{n} \log(1/\delta)$$

We set $\delta_k = \delta/(2^{k+1})$, so $\sum_{k=0}^{K_{\max}} \delta_k < \delta$. By a union bound, the event $\mathcal{E} = \bigcap_{k=0}^{K_{\max}} \mathcal{E}_k$ holds with probability at least $1 - \sum \delta_k > 1 - \delta$. On this event $\mathcal{E}$, the uniform convergence bound holds for *any* θ within *any* $B(\theta^*, r_k)$. Let $\widetilde{r} = \|\hat{\theta}_n - \theta^*\|_2$. Since $\hat{\theta}_n \in \Theta$, we know $\widetilde{r} \leq R_{\max}$. Let k^* be the smallest integer such that $\widetilde{r} \leq r_{k^*}$. By this definition, $\hat{\theta}_n \in B(\theta^*, r_{k^*})$. By the construction of dyadic radii, if $k^* > 0$, then $r_{k^*-1} < \widetilde{r} \leq r_{k^*}$, which implies $r_{k^*} = 2r_{k^*-1} < 2\widetilde{r}$. If $k^* = 0$, then $\widetilde{r} \leq r_0$. Thus, we can generally state $r_{k^*} \leq \max(r_0, 2\widetilde{r})$. The term $\log(1/\delta_{k^*})$ can be written as $\log(2^{k^*+1}/\delta) = (k^*+1)\log 2 + \log(1/\delta)$. This k^*-dependent term is bounded by $(K_{\max}+1)\log 2 + \log(1/\delta)$, where $K_{\max} = \lceil \log_2(R_{\max}/r_0) \rceil$. We can denote this as $\log(C_{\text{loc}}/\delta)$ where C_{loc} is a constant related to $R_{\max}$ and r_0. For simplicity in the final bound, we will write $\log(1/\delta)$ and assume that the universal constants c_1, c_2 implicitly absorb any logarithmic factors arising from the localization argument's union bound (e.g., $\log(C_{\text{loc}})$). Since $\hat{\theta}_n$ is the empirical minimizer:

$$\mathcal{L}_n^D(\hat{\theta}_n; \varepsilon_n) \leq \mathcal{L}_n^D(\theta^*; \varepsilon_n)$$

Now we use the uniform convergence bound on the event $\mathcal{E}$. Since $\hat{\theta}_n \in B(\theta^*, r_{k^*})$, and $\theta^* \in B(\theta^*, r_{k^*})$, we can apply the uniform convergence bound with radius r_{k^*} (and total error probability δ):

$$\mathcal{L}_n^D(\hat{\theta}_n; \varepsilon_n) \geq \mathcal{L}^D(\hat{\theta}_n; \varepsilon_n) - \psi_n(r_{k^*}, \delta_{k^*})$$
$$\mathcal{L}_n^D(\theta^*; \varepsilon_n) \leq \mathcal{L}^D(\theta^*; \varepsilon_n) + \psi_n(r_{k^*}, \delta_{k^*})$$

Combining these with the strong convexity of $\mathcal{L}^D(\theta; \varepsilon_n)$ around θ^*:

$$\begin{aligned}
\mathcal{L}^D(\theta^*; \varepsilon_n) + \frac{\alpha}{2}\|\hat{\theta}_n - \theta^*\|_2^2 - \psi_n(r_{k^*}, \delta_{k^*}) &\leq \mathcal{L}_n^D(\hat{\theta}_n; \varepsilon_n) \\
&\leq \mathcal{L}_n^D(\theta^*; \varepsilon_n) \\
&\leq \mathcal{L}^D(\theta^*; \varepsilon_n) + \psi_n(r_{k^*}, \delta_{k^*})
\end{aligned}$$

Subtracting $\mathcal{L}^D(\theta^*; \varepsilon_n)$ from all parts and simplifying, we get:

$$\frac{\alpha}{2}\|\hat{\theta}_n - \theta^*\|_2^2 \leq 2\psi_n(r_{k^*}, \delta_{k^*})$$

Substituting $r_{k^*} \leq 2\widetilde{r}$ (where $\widetilde{r} = \|\hat{\theta}_n - \theta^*\|_2$) into the definition of $\psi_n(r_{k^*}, \delta')$ for the linear term in r:

$$\frac{\alpha}{2}\widetilde{r}^2 \leq 2\left(2\Delta_n + \left(2c_0 L_g\sqrt{\frac{d}{n}} + c_1 L_g\sqrt{\frac{\log(1/\delta)}{n}}\right)(2\widetilde{r}) + c_2\frac{K_\ell}{n}\log(1/\delta)\right).$$

Expanding and rearranging into a quadratic inequality of the form $A\widetilde{r}^2 - B\widetilde{r} - C \leq 0$:

$$\frac{\alpha}{2}\widetilde{r}^2 - \left(8c_0 L_g\sqrt{\frac{d}{n}} + 4c_1 L_g\sqrt{\frac{\log(1/\delta)}{n}}\right)\widetilde{r} - \left(4\Delta_n + 2c_2\frac{K_\ell}{n}\log(1/\delta)\right) \leq 0.$$

Let $A = \alpha/2$, $B' = \left(8c_0 L_g\sqrt{\frac{d}{n}} + 4c_1 L_g\sqrt{\frac{\log(1/\delta)}{n}}\right)$, and $C' = \left(4\Delta_n + 2c_2\frac{K_\ell}{n}\log(1/\delta)\right)$. For such a quadratic $Ax^2 - B'x - C' \leq 0$ with $A, B', C' > 0$, the solutions are bounded by $x \leq \frac{B' + \sqrt{(B')^2 + 4AC'}}{2A}$. We use the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ twice. First, for the general quadratic solution, we apply the loose bound $\sqrt{(B')^2 + 4AC'} \leq \sqrt{(B')^2} + \sqrt{4AC'} = B' + 2\sqrt{AC'}$:

$$\begin{aligned}
\widetilde{r} &\leq \frac{B' + (B' + 2\sqrt{AC'})}{2A} \\
&= \frac{2B'}{2A} + \frac{2\sqrt{AC'}}{2A} \\
&= \frac{B'}{A} + \sqrt{\frac{AC'}{A^2}} \\
&= \frac{B'}{A} + \sqrt{\frac{C'}{A}}.
\end{aligned}$$

Now, substitute back A, B', C':

$$\begin{aligned}
\widetilde{r} &\leq \frac{1}{\alpha/2}\left(8c_0 L_g\sqrt{\frac{d}{n}} + 4c_1 L_g\sqrt{\frac{\log(1/\delta)}{n}}\right) + \sqrt{\frac{4\Delta_n + 2c_2\frac{K_\ell}{n}\log(1/\delta)}{\alpha/2}} \\
&= \frac{16c_0 L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{8c_1 L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8\Delta_n}{\alpha} + \frac{4c_2 K_\ell\log(1/\delta)}{\alpha n}}.
\end{aligned}$$

Finally, apply $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to the last square root term:

$$\|\hat{\theta}_n - \theta^*\|_2 \lesssim \frac{16c_0 L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{8c_1 L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8\Delta_n}{\alpha}} + \sqrt{\frac{4c_2 K_\ell\log(1/\delta)}{\alpha n}}.$$

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. $\square$

For all proofs that follow, note that by the Master Theorem (Theorem 4), it suffices to show for the metric we are dealing with that we have a dual "remainder" term as follows:

$$\left|\mathcal{L}^D(\theta;\varepsilon_n) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z;\theta)\right]\right| \;\leq\; \Delta_n.$$

Furthermore, the exact same argument applies to the empirical counterpart, i.e.,

$$\left|\mathcal{L}_n^D(\theta;\varepsilon_n) \;-\; \mathbb{E}_{\mathbb{P}_n^\circ}\left[\ell(z;\theta)\right]\right| \;\leq\; \Delta_n,$$

by simply replacing $\mathbb{P}^\circ$ with $\mathbb{P}_n^\circ$ in the derivations. This holds because the properties of the loss function and the definitions of the divergences are universally applicable to any probability measure. In all cases, we will show that $\Delta_n = O(n^{-1})$ by choosing ε_n appropriately.

G Proof of "Fast Rate" Wasserstein-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 "remainder" term holds:

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

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

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 $\mathbb{P}^\circ$:

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

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 apply Hölder's inequality (or the generalized mean inequality for distances):

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

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

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

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. ☐

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

Proof of Corollary 2. First, recall Lemma 11 (Gibbs variational principle characterization of the KL divergence) for probability measures $\mathbb{P}, \mathbb{Q}$:

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

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 $D_{\mathrm{KL}}(\mathbb{P} \mid\mid \mathbb{Q})$:

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

Now, suppose that $f = \ell(z;\theta) \in [0, K_\ell]$ almost surely (from Lemma 14). By Hoeffding's Lemma (Lemma 12), 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:

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

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. ☐

I Proof of "Fast Rate" χ^2-DRO-REBEL

Proof of Corollary 3. By Theorem 10 (Hammersley-Chapman-Robbins (HCR) lower bound), we immediately have:

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

Since $\ell(z;\theta) \in [0, K_\ell]$ almost surely, by Popoviciu's inequality on variances, 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. ☐

J 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 four 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) We show that the pointwise DPO loss, $\ell(z;\theta) = -y \log \sigma(\beta h_\theta) - (1-y)\log\sigma(-\beta h_\theta)$, is Lipschitz in θ. The gradient with respect to θ is $\nabla_\theta \ell(z;\theta) = \partial\ell/\partial h_\theta \cdot \nabla_\theta h_\theta$.

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

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

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.

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

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

$$
\|\nabla_\theta \ell(z;\theta)\|_2 = \left|\frac{\partial \ell}{\partial h_\theta}\right| \cdot \|\nabla_\theta h_\theta\|_2 \le 2\beta
$$

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

3. Verification of Bounded Loss From the derivation in Step 2 and the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
|h_\theta| &= |\langle \theta - \theta_{\text{ref}}, \psi(s,a^1) - \psi(s,a^2)\rangle| \\
&\le \|\theta - \theta_{\text{ref}}\|_2 \|\psi(s,a^1) - \psi(s,a^2)\|_2 \\
&\le 4B
\end{aligned}
$$

This implies the argument βh_θ is in $[-4\beta B, 4\beta B]$, and the loss is uniformly bounded by:

$$
K_\ell = \log(1 + e^{4\beta B})
$$

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

$$
\|\hat{\theta}_n^{\mathcal{W}_p} - \theta^{\mathcal{W}_p}\|_2 \le \frac{32 c_0 \beta}{\gamma\lambda}\sqrt{\frac{d}{n}} + \frac{16 c_1 \beta}{\gamma\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8 L_{\ell,z}}{\gamma\lambda n}} + \sqrt{\frac{4 c_2 \log(1+e^{4\beta B})\log(1/\delta)}{\gamma\lambda n}}
$$

□

K 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. Since we already proved that ℓ_{DPO} is Lipschitz in θ and uniformly bounded by $K_\ell = \log(1+e^{4\beta B})$, we must just verify local strong convexity. From Appendix C, Lemma 14 of Xu et al. [2025], the KL-DPO loss, $\mathcal{L}^{KL}(\theta)$, 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.

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

$$\|\hat{\theta}_n^{\mathrm{KL}} - \theta^{\mathrm{KL}}\|_2 \leq \frac{32 c_0 \beta}{\gamma\lambda}\sqrt{\frac{d}{n}} + \frac{16 c_1 \beta}{\gamma\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{8\log(1+e^{4\beta B})}{\gamma\lambda n}} + \sqrt{\frac{4c_2 \log(1+e^{4\beta B})\log(1/\delta)}{\gamma\lambda n}}$$

□

L Proof of Tractable χ^2-DRO-REBEL

Proof. 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 9 [Duchi and Namkoong, 2020], the dual form of the f-divergence based DRO problem is:

$$\sup_{\mathbb{P}:\, D_f(\mathbb{P}\|\mathbb{P}_n)\leq\rho} \mathbb{E}_{\mathbb{P}}[\ell(z;\theta)] \;=\; \inf_{\substack{\lambda\geq 0\\ \eta\in\mathbb{R}}} \Big\{ \lambda\, \mathbb{E}_{\mathbb{P}_n}\left[f^*\!\left(\tfrac{\ell(z;\theta)-\eta}{\lambda}\right)\right] \;+\; \lambda\,\rho \;+\; \eta \Big\}.$$

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 $P(z_i) \geq 0$ in the primal problem implies $w_i \geq 0$, which means $1 + \frac{\ell_i - \eta}{\lambda} \geq 0$. 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 $1 + \frac{\ell_i-\eta}{\lambda} < 0$, the optimal $w_i = 0$, which essentially means that 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 first part of the proposition. For the worst-case weights, the optimal density ratio w_i is found by $w_i = \max\left(0, 1 + \frac{\ell_i - \eta^*}{\lambda^*}\right)$. This can be rewritten using the specific structure of χ^2-divergence as $w_i = \frac{(\ell_i - \eta^*)_+}{n\,\lambda^*}$ where λ^* is the optimal λ for the given η^*. This form ensures that $\sum w_i = n$ (as demonstrated in Namkoong and Duchi [2017a]).

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.

1. Convexity of the Objective Function We show 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 $v(\eta) = [v_1(\eta), \dots, v_n(\eta)]^\top$ 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 L2-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.

2. Characterization of the Optimum 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.

3. Bounded Search Space 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 $i = 1, \dots, n$. 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.

4. Algorithm and Complexity 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 $O(n \cdot \log(1/\epsilon))$. For Algorithm 4, if we assume that $\mathrm{Card}\left(\{\ell_i\}_{i=1}^n\right) = n$, then the runtime will be $O(n \log n)$. $\square$