DRO–REBEL: Fast and Robust Policy Optimization for RLHF via Distributional Regression

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 and rarely consider diverse human preferences [Mandal et al., 2025, Xu et al., 2025, Chakraborty et al., 2024]. In this work, we propose *DRO-REBEL*, a family of distributionally-robust variants of the efficient REBEL framework [Gao et al., 2024], instantiating ambiguity sets via Kullback–Leibler (KL), Chi-squared (χ^2), and type-p Wasserstein distances. We derive novel dual reformulations that reduce each robust policy update to a simple regression problem, and we prove both "slow" $O(n^{-1/4})$ and "fast" $O(n^{-1/2})$ estimation-error rates under standard linear-policy and data-coverage assumptions.

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-distributon detection to reject bad bahavior [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 variants of REBEL for uncertainty sets defined via type-p Wasserstein, KL, and χ^2 divergences.

- **Extends REBEL to multiple ambiguity sets.** We introduce *DRO–REBEL*, a family of robust REBEL updates instantiated under type-p Wasserstein, KL, and χ^2 uncertainty sets. Using Fenchel duality (or strong duality for Wasserstein), each robust policy update reduces to a simple relative-reward regression, inheriting REBEL's scalability.
- **Provides sharper slow-rate guarantees.** Under standard linear reward function class, log-linear policy, and data-coverage assumptions, we prove that every DRO–REBEL variant achieves an $O(n^{-1/4})$ "slow" estimation-error rate. Our analysis tightens the constants compared to prior DRO-DPO [Xu et al., 2025, Mandal et al., 2025] results by eliminating hidden logistic/exponential factors by the use of simple linear regression rather than logistic regression.
- **Recovers optimal fast-rate guarantees.** By developing a localized Rademacher-complexity argument for our regression-based loss, we show that each DRO–REBEL variant attains the minimax-optimal $O(n^{-1/2})$ "fast" parametric rate, even under distributional shifts, closing the gap between robust and non-robust RLHF methods.
- **Improves curvature and stability.** We demonstrate that DRO–REBEL's strong-convexity modulus scales purely with the data-coverage constant λ (and our step-size η), avoiding the exponential sensitivity to reward-link curvature found in WDPO/KLDPO [Xu et al., 2025]. This yields uniformly smaller excess-risk bounds and more stable updates.

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 w.r.t. 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, 2017, 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 $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ where δ is the Dirac measure. We denote $l(z;\theta)$ 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 $A \in \mathbb{S}^n$.

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)**.**

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

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:

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

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. A high-level description is given 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 $x \sim \rho$, $a^1 \sim \pi_t(\cdot \mid x)$, $a^2 \sim \mu(\cdot \mid x)$
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:

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

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.

3 Distributionally Robust REBEL

In this section, we will formally define the DRO variants of REBEL under type-p Wasserstein, KL, and χ^2 divergence ambiguity sets. We must first define the nominal data-generating distribution. Our definitions follow those stated by Xu et al. [2025]. Recall the sampling procedure mentioned in Section 2.3: We have some initial prompt $x \in \mathcal{S}$ that we will assume is sampled from some prompt distribution ρ. We will sample two responses $a^1, a^2 \overset{\text{i.i.d}}{\sim} \pi_{\text{SFT}}(\cdot \mid x)$. 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*

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

where $x \sim \rho$ and $y \sim \text{Ber}\left(\mathcal{P}^*\left(a^1 \succ a^2 \mid \cdot\right)\right)$. We will denote $z = (x, a^1, a^2, y) \in \mathcal{Z}$ and $\mathbb{P}^\circ(z) = \mathbb{P}^\circ(x, a^1, a^2, y)$. We will also assume $\mathbb{P}^\circ$ generates dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ used for learning i.e. $z_i \sim \mathbb{P}^\circ$.

3.1 Distributionally Robust REBEL

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

$$\ell(z; \theta) = \frac{1}{\eta}\left(\left[\log\frac{\pi_\theta(y \mid x)}{\pi_t(y \mid x)} - \log\frac{\pi_\theta(y' \mid x)}{\pi_t(y' \mid x)}\right] - \left[r(x, y) - r(x, y')\right]\right)^2$$

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:

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

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:

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

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.

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:*

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

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 $\max_{x,a} \|\psi(x,a)\|_2 \leq 1$. Assume a bounded policy parameter set $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$. 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 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 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 $\mathcal{B}_\varepsilon\left(\mathbb{P}^\circ; D\right)$:

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 2. *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 $\lambda \leq \lambda_{\min}(\Sigma_{\mathbb{P}^\circ})$, 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*

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

is $2\lambda/\eta$-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 atleast $1-\delta$*

$$||\theta^{\mathcal{W}_p} - \hat{\theta}_n^{\mathcal{W}_p}||_2^2 \leq \frac{1}{\lambda} K_g^2 \sqrt{\frac{2\log(2/\delta)}{n}}$$

where λ is from the regularity condition in Assumption 3 and $K_g = 4B + 2F$ 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 $K_g = 4B + 2F$. Using this, we can prove that $\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ 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 $2/\eta$-strongly convex in $||\cdot||_2$. The detailed proof for strong convexity can be found in Lemma 19.

Strong duality of Wasserstein DRO [Gao and Kleywegt, 2022] (Corollary 2) allows us to reduce the difference $|\mathcal{L}^{\mathcal{W}_p}(\theta;\varepsilon) - \mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)|$ to the concentration $\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[\ell_\Delta(z;\theta)\right] - \mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell_\Delta(z;\theta)\right]\right|$ where $\ell_\Delta(z;\theta) = \inf_{z'\in\mathcal{Z}}\left\{\Delta d^p(z,z') - \ell(z';\theta)\right\}$ 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 $\mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right) - \mathcal{L}^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)$ into $\mathcal{L}^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right) - \mathcal{L}_n^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right)$, $\mathcal{L}_n^{\mathcal{W}_p}\left(\theta^{\mathcal{W}_p};\varepsilon\right) - \mathcal{L}_n^{\mathcal{W}_p}\left(\hat{\theta}_n^{\mathcal{W}_p};\varepsilon\right)$, 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 detaled proof for the "slow rate" estimation error can be found at C □

We prove similar results for a KL and χ^2 ambiguity set 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 atleast $1-\delta$*

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

where λ is from the regularity condition in Assumption 3 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 atleast $1-\delta$*

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

where λ is from the regularity condition in Assumption 3 and $K_g = 4B + 2F$ 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 3 (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 19 shows that the DRO-REBEL loss $\mathcal{L}^{\mathcal{W}_p}$ is $(2\lambda/\eta)$-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

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

where $\bar{\lambda}$ bounds the dual multiplier and L bounds the loss. In prior KLDPO analyses, the dependence on $\exp(L/\bar{\lambda})$ 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 robust setting, one cannot exchange the supremum over distributions with the gradient operator on the empirical robust loss. As a result, the closed-form concentration argument—key to the non-robust analysis—fails. 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 Wasserstein ambiguity.

4.2 "Fast Rate" Estimation Errors

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 $\mathcal{L}^D(\theta) - \mathbb{E}_P[\ell(z;\theta)] = O(\Delta_n)$ 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*

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

and assume the usual strong-convexity and Lipschitz conditions, plus the "dual remainder"

$$\mathcal{L}^D(\theta;\varepsilon_n) \;-\; \mathbb{E}_{\mathbb{P}^\circ}\big[\ell(z;\theta)\big] \;\le\; \Delta_n$$

Then for any fixed $\delta > 0$, with probability atleast $1-\delta$

$$\|\hat{\theta}_n - \theta^\star\|_2 \;=\; \frac{4c_0 L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{2c_1 L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{2c_2 K_\ell}{\alpha n}\log(1/\delta)} + 2\sqrt{\frac{\Delta_n}{\alpha}} + 2\sqrt{\frac{K_\ell}{\alpha}}$$

and in particular if $\Delta_n = O(n^{-1})$ then $\|\hat{\theta}_n - \theta^\star\|_2 = O(n^{-1/2})$.

Proof Sketch of Theorem 4. For the full proof, we refer readers to Appendix F. At a high level, for any $\theta \in \Theta$, by using the triangle-inequality and the dual-remainder bound, we get $\big|\mathcal{L}^D(\theta) - \mathcal{L}_n^D(\theta)\big| \;\le\; 2\,\Delta_n + \big|M(\theta) - M_n(\theta)\big|$. 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 8), Dudley entropy-integral (Corollary 5), and Bousquet's inequality (Theorem 6) argument, one shows with probability atleast $1-\delta$, $\sup_{f\in\mathcal{F}_r}\big|P_n f - Pf\big| = O\big(\Delta_n + L_g\, r\sqrt{(d+\ln(1/\delta))/n} + K_\ell\,\ln(1/\delta)/n\big)$.

Letting $\mathcal{E}_r \;=\; \Big\{\sup_{\|\theta-\theta^*\|\le r}|\mathcal{L}^D(\theta;\varepsilon_n) - \mathcal{L}_n^D(\theta;\varepsilon_n)| \le \psi_n(r,\delta) \;\mid\; \|\theta-\theta^*\|_2 \le r\Big\}$ where $\psi_n(r,\delta) \;=\; O\big(\Delta_n + L_g\, r\sqrt{(d+\ln(1/\delta))/n} + K_\ell\,\ln(1/\delta)/n\big)$. So far we have proved that if $\theta \in B(\theta^*, r)$ (a closed ball centered at θ^* with radius of r), then $\mathbb{P}(\mathcal{E}_r) \ge 1-\delta$. Now we can argue that $\hat{\theta}_n \in B(\theta^*, r)$ for some sufficiently "nice" radius which we will denote $\widetilde{r}$ using proof by contradiction. Choosing $\widetilde{r} = O\big((L_g/\alpha)\sqrt{(d+\ln(1/\delta))/n} \;+\; \sqrt{\Delta_n/\alpha} \;+\; \sqrt{K_\ell/\alpha}\big)$ and taking $\Delta_n = O(n^{-1})$ yields the claimed $O(n^{-1/2})$-rate. □

Using this theorem, we can now state the following "fast rate" 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 atleast $1-\delta$*

$$||\theta^{\mathcal{W}_p} - \hat{\theta}_n^{\mathcal{W}_p}||_2 \le \frac{8c_0 K_g}{\lambda}\sqrt{\frac{d}{n}} + \frac{4c_1 K_g}{\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + K_g\sqrt{\frac{c_2\log(1/\delta)}{\lambda n}} + \sqrt{\frac{8K_g}{\lambda n}} + K_g\sqrt{\frac{2}{\lambda}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, λ is from the regularity condition in Assumption 3, and $K_g = 4B+2F$ 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 atleast $1-\delta$*

$$\|\theta^{\mathrm{KL}} - \hat{\theta}_n^{\mathrm{KL}}\|_2 \le \frac{8c_0 K_g}{\lambda}\sqrt{\frac{d}{n}} + \frac{4c_1 K_g}{\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + K_g\sqrt{\frac{c_2\log(1/\delta)}{\lambda n}} + K_g\sqrt{\frac{1}{\lambda n}} + K_g\sqrt{\frac{2}{\lambda}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, λ is from the regularity condition in Assumption 3, and $K_g = 4B+2F$ 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 atleast $1-\delta$*

$$\|\theta^{\chi^2} - \hat{\theta}_n^{\chi^2}\|_2 \le \frac{8c_0 K_g}{\lambda}\sqrt{\frac{d}{n}} + \frac{4c_1 K_g}{\lambda}\sqrt{\frac{\log(1/\delta)}{n}} + K_g\sqrt{\frac{c_2\log(1/\delta)}{\lambda n}} + K_g\sqrt{\frac{2}{\lambda n}} + K_g\sqrt{\frac{2}{\lambda}}$$

where $c_0, c_1, c_2 > 0$ are some absolute constants, λ is from the regularity condition in Assumption 3, and $K_g = 4B+2F$ 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 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. 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.

χ^2-Based Preference Learning. Very recently, Huang et al. [2025] introduced χ^2-PO, 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 χ^2-PO achieves a sample-complexity guarantee scaling as

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

where $C_{\pi^*} = 1 + 2D_{\chi^2}(\pi^* \| \pi_{\mathrm{ref}})$ is the single-policy concentrability coefficient (cf. Theorem 3.1 of Huang et al. [2025]). Our fast-rate analysis for the χ^2-robust REBEL update (Corollary 3) recovers exactly the same $O(n^{-1/2})$ parametric rate under analogous linear-policy, data-coverage, and strong-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 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. Due to time constraints, we were unable to complete these experiments before submission. However, we are actively working on them over the summer and expect the empirical results to align closely with our theory.

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 *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

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=0CenCW9qmp4C.

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 2 (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

$$v_p = \sup_{\mu \in \mathcal{P}(\Xi)} \Big\{ \int_\Xi \Psi(\xi)\, \mu(d\xi) \colon W_p(\mu, \nu) \leq \rho \Big\}, \quad \textit{(Primal)} \tag{2}$$

$$v_D = \inf_{\eta \geq 0} \Big\{ \eta\, \rho^p \;-\; \int_\Xi \inf_{\xi \in \Xi} \big[\eta\, d^p(\xi, \zeta) - \Psi(\xi) \big]\, \nu(d\zeta) \Big\}. \quad \textit{(Dual)} \tag{3}$$

Lemma 3 (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 3 by choosing $L = \operatorname{diam}(\Xi)^p$ and $M = \sup_{\xi \in \Xi} |\Psi(\xi)|$. □

A.2 Optimization

Lemma 4 (Beck [2023], Theorem 1.24; Linear Approximation Theorem)**.** *Let $f \colon U \to \mathbb{R}$ be twice continuously differentiable on an open set $U \subseteq \mathbb{R}^n$, and let $x, y \in U$ satisfy $[x, y] \subset U$. Then there exists $\xi \in [x, y]$ such that*

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

Lemma 5 (Beck [2017], Theorem 5.24; First-order characterizations of strong convexity)**.** *Let $f \colon E \to (-\infty, \infty]$ be a proper, closed, convex function, and let $\sigma > 0$. The following are equivalent:*

1. *For all $x, y \in \operatorname{dom}(f)$ and $\lambda \in [0, 1]$,*

$$f\big(\lambda x + (1 - \lambda) y\big) \leq \lambda f(x) + (1 - \lambda) f(y) - \tfrac{\sigma}{2}\, \lambda (1 - \lambda)\, \|x - y\|^2.$$

2. *For all $x \in \operatorname{dom}(\partial f)$, $y \in \operatorname{dom}(f)$ and $g \in \partial f(x)$,*

$$f(y) \;\geq\; f(x) \;+\; \langle g, y - x \rangle \;+\; \tfrac{\sigma}{2}\, \|y - x\|^2.$$

Lemma 6 (Beck [2017], Theorem 5.25; Existence and uniqueness of minimizer)**.** *Let $f \colon E \to (-\infty, \infty]$ be proper, closed, and σ-strongly convex with $\sigma > 0$. Then:*

1. *f has a unique minimizer x^*.*
2. *For all $x \in \operatorname{dom}(f)$,*

$$f(x) - f(x^*) \;\geq\; \tfrac{\sigma}{2}\, \|x - x^*\|^2.$$

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 $\ell \colon \mathcal{X} \to \mathbb{R}$:

Lemma 7 (Duchi and Namkoong [2020], Proposition 1)**.** *Let D_f be as in* (15)*. Then*

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

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

A.4 Empirical Process Theory

Lemma 8 (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.

Theorem 5 (Boucheron et al. [2013], Theorem 11.6; Contraction Principle)**.** *Let $x_1, \dots, x_n$ be vectors whose real-valued components are indexed by a set $\mathcal{T}$; that is,*

$$x_i = \big(x_{i,s}\big)_{s\in\mathcal{T}}, \quad i = 1, \dots, n.$$

For each $i = 1, \dots, n$, let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function with $\varphi_i(0) = 0$. Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \to \mathbb{R}$ be a nondecreasing, convex function. Then

$$\mathbb{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}} \sum_{i=1}^n \epsilon_i\, \varphi_i(x_{i,s})\Big)\Big] \;\le\; \mathbb{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}} \sum_{i=1}^n \epsilon_i\, x_{i,s}\Big)\Big]$$

and

$$\mathbb{E}\Big[\Psi\Big(\tfrac{1}{2}\sup_{s\in\mathcal{T}} \sum_{i=1}^n \epsilon_i\, \varphi_i(x_{i,s})\Big)\Big] \;\le\; \mathbb{E}\Big[\Psi\Big(\sup_{s\in\mathcal{T}} \sum_{i=1}^n \epsilon_i\, x_{i,s}\Big)\Big].$$

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 6 (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 9 (van Handel [2016], Lemma 4.10; Gibbs Variational Principle)**.** *Let $\mu, \nu \in \mathcal{P}(\Xi)$ be Borel probability measures supported on Ξ. Then*

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

Theorem 7 ([Polyanskiy, 2017, Lehmann and Casella, 1998]; Hammersley-Chapman-Robbins (HCR) lower bound)**.** *Let Θ be the set of parameters for a family of probability distributions $\{\mu_\theta : \theta \in \Theta\}$ on a sample space Ω. For any $\theta, \theta' \in \Theta$, let $\chi^2\big(\mu_{\theta'};\, \mu_\theta\big)$ denote the χ^2–divergence from μ_θ to $\mu_{\theta'}$. For any scalar random variable $\hat{g}\colon \Omega \to \mathbb{R}$ and any $\theta, \theta' \in \Theta$, we have*

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

A.6 Concentration Inequalities

Lemma 10 (Boucheron et al. [2013], Lemma 2.2; Hoeffding's Lemma)**.** *Let Y be a random variable with $\mathbb{E}[Y] = 0$ and almost surely $Y \in [a, b]$. Define $\psi_Y(\lambda) = \log \mathbb{E}\big[e^{\lambda Y}\big]$. Then for all $\lambda \in \mathbb{R}$, $\psi_Y''(\lambda) \leq \frac{(b-a)^2}{4}$ and consequently Y is sub-Gaussian with proxy variance $(b-a)^2/4$, i.e. $Y \sim \mathcal{SG}\left(\frac{b-a}{2}\right)$.*

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

Lemma 11 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent with $X_i \in [a_i, b_i]$ almost surely, and define*

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

Then for every $t > 0$,

$$\mathbb{P}(S \geq t) \leq \exp\Big(-\tfrac{2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

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

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

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

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

Lemma 12 (Uniform bound on $\ell(z;\theta)$)**.** *Let $K_g = \sup_{z,\theta} |g(z;\theta)| \leq 4B + 2F$ where $\ell(z;\theta) = \frac{1}{\eta} g(z;\theta)^2$ with $z = (x, a^1, a^2) \sim \mathbb{P}^\circ$. Then $\sup_{z,\theta} |\ell(z;\theta)| = K_\ell = \frac{1}{\eta} K_g^2$*

Proof of Lemma 12. Since we have that $\pi_\theta, \pi_{\theta_t} \in \Pi$, notice that

$$\begin{aligned}
\log\left(\frac{\pi_\theta(a \mid x)}{\pi_{\theta_t}(a \mid x)}\right) &= \log \pi_\theta(a \mid x) - \log \pi_{\theta_t}(a \mid x) \\
&= \log\left(\frac{\exp\left(\theta^\top \psi(x,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x,a')\right)}\right) - \log\left(\frac{\exp\left(\theta_t^\top \psi(x,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x,a')\right)}\right) \\
&= \log\left(\exp\left(\left(\theta - \theta_t\right)^\top \psi(x,a)\right)\right) + \log\left(\frac{\sum_{a' \in \mathcal{A}} \exp\left(\theta_t^\top \psi(x,a')\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta^\top \psi(x,a')\right)}\right) \\
&\leq \left(\theta - \theta_t\right)^\top \psi(x,a) + \log\left(\frac{\sum_{a' \in \mathcal{A}} \exp\left(||\theta_t||_2 ||\psi(x,a')||_2\right)}{\sum_{a' \in \mathcal{A}} \exp\left(||\theta||_2 ||\psi(x,a')||_2\right)}\right) \\
&\leq \left(\theta - \theta_t\right)^\top \psi(x,a) \\
&\leq \left(||\theta||_2 + ||\theta_t||_2\right) ||\psi(x,a)||_2 \\
&\leq 2B
\end{aligned}$$

where the first inequality holds from Cauchy-Schwartz, the second inequality holds from upper bounding θ, θ_t, ψ from Assumption 2 and noticing that the second term equates to $\log(1) = 0$. The third inequality from Cauchy-Schwartz and the Triangle inequality, and the last inequality holds again from upper bounding θ, θ_t, ψ. Now, we also have that $r \in \mathcal{F}$. Thus,

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

where the first inequality holds from Cauchy-Schwartz and Triangle inequality. Now recall the REBEL update 1. Using these facts we have that $|g(z;\theta)| \leq 4B + 2F$ so $K_g = \sup_{z,\theta} |g(z;\theta)| \leq 4B + 2F$. Since $\ell(z;\theta) = \frac{1}{\eta} g(z;\theta)^2$, $K_\ell = \frac{1}{\eta} K_g^2$. □

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

Lemma 13 (Lipschitz bound on $\ell(z;\theta)$)**.** *$\ell(z;\theta)$ is $\frac{4K_g}{\eta}$-Lispchitz in θ.*

Proof of Lemma 13. 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:

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

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

$$\nabla_\theta g(z;\theta) = \psi(x,a) - \psi(x,a')$$

Thus by Cauchy-Schwartz, $||\nabla_\theta g(z;\theta)||_2 \leq 2$. Now since $\ell(z;\theta) = \frac{1}{\eta} g(z;\theta)^2$, we have that $\nabla_\theta \ell(z;\theta) = \frac{2}{\eta} g(z;\theta) \nabla_\theta g(z;\theta)$. From Lemma 12, we know that $K_g = \sup_{z,\theta} |g(z;\theta)|$ so we see that $||\nabla_\theta \ell(z;\theta)||_2 \leq 4K_g/\eta$. 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 $h(\theta;\mathbb{P}) = \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is strongly convex for any $\mathbb{P}$.

Lemma 14 (Strong convexity of h)**.** *Let $\ell(z;\theta)$ be the REBEL loss function. Assume that Assumption 3 holds. Then $h(\theta;\mathbb{P}) = \mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2/\eta$-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 14. From Lemma 13, we know that $\nabla_\theta \ell(z;\theta) = \frac{2}{\eta} g(z;\theta) \nabla_\theta g(z;\theta)$ so we have

$$\nabla_\theta^2 \ell(z;\theta) = \frac{2}{\eta} \nabla_\theta g(z;\theta) \nabla_\theta g(z;\theta)^\top$$

We also know from Lemma 13 that $\nabla_\theta g(z;\theta) = \psi(x,a) - \psi(x,a')$. Thus we find that

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

Then taking an expectation under $\mathbb{P}$, we find that

$$\nabla_\theta^2 h(\theta;\mathbb{P}) = \frac{2}{\eta} \mathbb{E}_{z\sim\mathbb{P}}\left[\left(\psi(x,a^1) - \psi(x,a^2)\right)\left(\psi(x,a^1) - \psi(x,a^2)\right)^\top\right] = \frac{2}{\eta}\Sigma_\mathbb{P}$$

Now fix $\theta, \theta' \in \Theta$. Let $\Delta = \theta - \theta'$. By the second-order Taylor expansion, there exists $\widetilde{\theta}$ on the line segment between θ and θ' such that

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

where $\mu = \frac{2}{\eta}$ and $\Sigma_z = \big(\psi(x,a^1) - \psi(x,a^2)\big)\big(\psi(x,a^1) - \psi(x,a^2)\big)^\top$. Taking expectations we find that

$$h(\theta';\mathbb{P}) \geq h(\theta;\mathbb{P}) + \langle \nabla_\theta g(z;\theta), \Delta\rangle + \frac{\mu}{2}||\Delta||_{\Sigma_\mathbb{P}}$$

Thus, 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 15 (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 $2\lambda/\eta$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 19. In Lemma 14, we proved strong convexity of h. By Lemma 5, 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

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

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 $\mathcal{L}^{\mathcal{W}_p}$ 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 1. By strong duality for Wasserstein DRO 2, for fixed θ we have

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

where $\ell_\Delta(z;\theta) = \inf_{z'\in\mathcal{Z}}\{\Delta d^p(z,z') - \ell(z';\theta)\}$ where d is the metric used to define the type-p Wasserstein distance. Then notice that

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

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

$$l_\Delta(z;\theta) = \inf_{z' \in \mathcal{Z}} \{\Delta d^p(z,z') - \ell(z';\theta)\} \leq \inf_{z' \in \mathcal{Z}} \{\Delta d^p(z,z')\} = 0$$
$$\ell_\Delta(z;\theta) = \inf_{z' \in \mathcal{Z}} \{\Delta d^p(z,z') - \ell(z';\theta)\} \geq \inf_{z' \in \mathcal{Z}} \{\Delta d^p(z,z') - K_\ell\} \geq -K_\ell$$

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 have by By Hoeffding's inequality (by Lemma 11)

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

□

Since K_ℓ does not depend on Δ, this bound is uniform and thus does not require any uniform covering argument. By picking δ to be the right hand side, we find that with probability atleast $1-\delta$

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

Now we have that

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

where the first inequality holds from the fact that $\hat{\theta}_n^{\mathcal{W}_p} \in \arg\min_{\theta\in\Theta} \mathcal{L}_n^{\mathcal{W}_p}(\theta;\varepsilon)$. Now from Lemma 6 and Lemma 19, we have that

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

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

$$\|\theta^{\mathcal{W}_p} - \hat{\theta}_n^{\mathcal{W}_p}\|^2 \leq \frac{1}{\lambda}K_g^2\sqrt{\frac{2\log(2/\delta)}{n}}$$

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 3. 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:

Lemma 16 (Dual reformulation of KL-DRO-REBEL)**.** *Let $\ell(z;\theta)$ be the REBEL loss. The KL-DRO-REBEL loss function has the following dual reformulation:*

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

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

We will now establish strong convexity of $\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$. This proof will essentially be the same as Lemma 19

Lemma 17 (Strong convexity of $\mathcal{L}^{\mathrm{KL}}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\mathrm{KL}}(\theta;\varepsilon) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\mathrm{KL})}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2\lambda/\eta$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 19. In Lemma 14, we proved strong convexity of h. By Lemma 5, for $\theta,\theta'\in\Theta$ and $\alpha\in[0,1]$, this is equivalent to

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

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

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

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

Proof of Theorem 2. By the strong duality result for KL-DRO 16, we have for fixed θ

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

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

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

where the first equality holds from strong duality, the first inequality holds from $\inf_x f(x)-\inf_x g(x)\leq\sup_x|f(x)-g(x)|$, the seecond inequality holds from $|\log(1+x)|\leq|x|\ \forall x\geq 0$, and the last inequality holds from $l(z;\theta)\geq 0$. By Hoeffding's inequality (by Lemma 11)

$$
\mathbb{P}\left(\left|\mathbb{E}_{z\sim\mathbb{P}_n^\circ}\left[j(z,\lambda;\theta)\right]-\mathbb{E}_{z\sim\mathbb{P}^\circ}\left[j(z,\lambda;\theta)\right]\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{2n\epsilon^2}{\exp\left(L/\bar{\lambda}\right)}\right)
$$

Again since L and $\bar{\lambda}$ are independent of λ, we do not need a uniform covering argument over λ. Picking δ to be the right side, we find that with probability atleast $1-\delta$

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

Now we have that

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

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 6 and Lemma 16, we have that

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

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

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

□

E Proof of "Slow Rate" χ^2-DRO-REBEL

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

Lemma 18 (Dual reformulation of χ^2-DRO-REBEL)**.** *Let $\ell(z;\theta)$ be the REBEL loss. The χ^2-DRO-REBEL objective admits the dual form*

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

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

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

We again will prove strong convexity for χ^2

Lemma 19 (Strong convexity of $\mathcal{L}^{\chi^2}$)**.** *Let $l(z;\theta)$ be the REBEL loss function. Then $\mathcal{L}^{\chi^2}(\theta;\varepsilon)=\sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\mathbb{P}^\circ;\chi^2)}\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right]$ is $2\lambda/\eta$-strongly convex with respect to Euclidean norm $||\cdot||_2$ where λ is the regularity parameter from Assumption 3.*

Proof of Lemma 19. In Lemma 14, we proved strong convexity of h. By Lemma 5, for $\theta,\theta'\in\Theta$ and $\alpha\in[0,1]$, this is equivalent to

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

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

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

where the second inequality holds from $\sup_x(f(x)+g(x))\le\sup_x f(x)+\sup_x g(x)$, 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 $\hat{\theta}_n^{\chi^2}\in\arg\min_\theta\mathcal{L}_n^{\chi^2}(\theta;\varepsilon)$. By the dual reformulation (Lemma 18), for any fixed θ

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

and similarly

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

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

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

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 second second inequality in the argument above follows by the triangle inequality . From Lemma 12, we have that $\ell(z;\theta) \in [0, K_\ell]$. Then by Hoeffding's inequality (Lemma 11), we have that

$$\Pr\big(|\mu - \mu_n| \ge \delta\big) \;\le\; 2\exp\Big(-\tfrac{2n\delta^2}{K_\ell^2}\Big),$$

and

$$\Pr\big(|\sigma^2 - \sigma_n^2| \ge \delta'\big) \;\le\; 2\exp\Big(-\tfrac{2n\delta'^2}{K_\ell^4}\Big).$$

Thus with probability at least $1 - 2\exp(-2n\alpha^2/K_\ell^4)$, choosing $\delta, \delta' = K_\ell^2\sqrt{\frac{\log(2/\delta)}{2n}}$, we get

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

and

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

Therefore

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

and overall

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

Now set $\theta = \theta^{\chi^2}$ and $\theta = \hat{\theta}_n^{\chi^2}$ in turn, and use the "three-term" decomposition (as in the Wasserstein and KL proof) to conclude

$$\mathcal{L}^{\chi^2}(\theta^{\chi^2};\varepsilon) - \mathcal{L}^{\chi^2}(\hat{\theta}_n^{\chi^2};\varepsilon) \;\le\; 2\big|\mathcal{L}^{\chi^2}(\theta) - \mathcal{L}_n^{\chi^2}(\theta)\big| \le K_\ell\Big(1 + \tfrac{K_\ell}{4\bar{\lambda}}\Big)\sqrt{\frac{2\log(2/\delta)}{n}}$$

Finally, by strong convexity of $\mathcal{L}^{\chi^2}$ (cf. Lemma 6),

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

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

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

as claimed. □

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 $\mathcal{L}^D(\theta) - \mathbb{E}_{\mathbb{P}}\big[\ell(z;\theta)\big] \;\le\; \Delta_n$. 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 $\sup_{\|\theta-\theta^*\|_2\leq r}|M(\theta)-M_n(\theta)|$ 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 8 (Symmetrization) taking $\Phi=I$ and 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

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

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. Then by Theorem 5 (Contraction principle)

$$\mathcal{R}_n(\mathcal{F}_r)\leq L_g\mathcal{R}_n(\{\theta:\|\theta-\theta^*\|\leq r\})$$

But the latter is a class of points in a Euclidean ball in $\mathbb{R}^d$. A standard covering argument gives us

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

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

$$\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}_r)\right]\leq\frac{12}{\sqrt{n}}\int_0^{L_gr}\sqrt{\log N(\epsilon,\mathcal{F}_r,L_2(P))}d\epsilon\leq c_0L_gr\sqrt{\frac{d}{n}}$$

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

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_r}|P_nf-Pf|\right]\leq 2c_0L_gr\sqrt{\frac{d}{n}}$$

To upgrade the above expectation bound into a high-probability bound, we will apply Bousquet's inequality (Theorem 6). First notice that since $\ell(z;\theta)$ is L_g-Lipschitz in θ, we have that

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

Thus $f_\theta^2(z)\leq L_g^2r^2$. Thus we have that $\text{Var}_{z\sim\mathbb{P}}\left(f_\theta(z)\right)\leq L_g^2r^2$. Since each $f_\theta\in\mathcal{F}_r$ is uniformly bounded by $2K_\ell$ by assumption and has bounded variance, then by applying Theorem 6 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 atleast $1-\delta$

$$\sup_{f\in\mathcal{F}_r}|P_nf-Pf|\leq 2\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}_r)\right]+c_1L_gr\sqrt{\frac{\log(1/\delta)}{n}}+c_2\frac{K_\ell}{n}\log(1/\delta)$$

Plugging in the Dudley integral bound, we get

$$\sup_{f\in\mathcal{F}_r}|P_nf-Pf|\leq 2c_0L_gr\sqrt{\frac{d}{n}}+c_1L_gr\sqrt{\frac{\log(1/\delta)}{n}}+c_2\frac{K_\ell}{n}\log(1/\delta)$$

Let us denote $\mathcal{E}_r$ as follows

$$\mathcal{E}_r=\left\{\sup_{\|\theta-\theta^*\|\leq r}|\mathcal{L}^D(\theta;\varepsilon_n)-\mathcal{L}_n^D(\theta;\varepsilon_n)|\leq\psi_n(r,\delta)\ \mid\ \|\theta-\theta^*\|_2\leq r\right\}$$

where

$$\psi_n(r,\delta)=2\Delta_n+2c_0L_gr\sqrt{\frac{d}{n}}+c_1L_gr\sqrt{\frac{\log(1/\delta)}{n}}+c_2\frac{K_\ell}{n}\log(1/\delta)$$

So far we have proved that if $\theta\in B(\theta^*,r)$ (a closed ball centered at θ^* with radius of r), then $\mathbb{P}\left(\mathcal{E}_r\right)\geq 1-\delta$. Now we will argue that $\hat{\theta}_n\in B(\theta^*,r)$ for some sufficiently "nice" radius which we will denote $\widetilde{r}$. We argue by contradiction. That is, suppose that $\hat{\theta}_n\notin B(\theta^*,\widetilde{r})$. We can do the following three term decomposition

$$\mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right)=\mathcal{L}^D\left(\hat{\theta}_n\,;\varepsilon_n\right)-\left[\mathcal{L}^D\left(\hat{\theta}_n\,;\varepsilon_n\right)-\mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right)\right]$$

First notice that by uniform boundedness of ℓ, we have that $\left[\mathcal{L}^D\left(\hat{\theta}_n\,;\varepsilon_n\right) - \mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right)\right] \leq 2K_\ell$. By α-strong convexity of $\mathcal{L}^D$ around θ^*, we have

$$\begin{aligned}\mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right) &= \mathcal{L}^D\left(\hat{\theta}_n\,;\varepsilon_n\right) - \left[\mathcal{L}^D\left(\hat{\theta}_n\,;\varepsilon_n\right) - \mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right)\right] \\ &\geq \mathcal{L}^D\left(\theta^*\,;\varepsilon_n\right) + \frac{\alpha}{2}\|\hat{\theta}_n - \theta^*\|_2^2 - 2K_l \\ &\geq \mathcal{L}^D\left(\theta^*\,;\varepsilon_n\right) + \frac{\alpha}{2}r^2 - 2K_l\end{aligned}$$

Trivially $\theta^* \in B(\theta^*, r)$ for any $r > 0$. Thus, $\theta^* \in B(\theta^*, \widetilde{r})$ so with probability atleast $1-\delta$,

$$\mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right) \geq \mathcal{L}_n^D\left(\theta^*\,;\varepsilon_n\right) + \frac{\alpha}{2}\widetilde{r}^2 - \psi_n(\widetilde{r}, \delta) - 2K_\ell$$

Suppose that $\frac{\alpha}{2}\widetilde{r}^2 - \psi_n(\widetilde{r}, \delta) - 2K_\ell > 0$. Notice that $\psi_n(\widetilde{r}, \delta)$ is a function of $\widetilde{r}$ so we can form the following quadratic:

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

Solving for $\widetilde{r}$ gives us

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

where the inequality holds from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. If take this $\widetilde{r}$, then we conclude that

$$\begin{aligned}\mathcal{L}_n^D\left(\hat{\theta}_n\,;\varepsilon_n\right) &\geq \mathcal{L}_n^D\left(\theta^*\,;\varepsilon_n\right) + \frac{\alpha}{2}\widetilde{r}^2 - \psi_n(\widetilde{r}, \delta) - 2K_\ell \\ &> \mathcal{L}_n^D\left(\theta^*\,;\varepsilon_n\right)\end{aligned}$$

This is a contradiction of the fact that $\hat{\theta}_n = \arg\min_{\theta\in\Theta}\mathcal{L}_n^D\big(\theta;\varepsilon_n\big)$. Thus it must be the case that $\hat{\theta}_n \in B(\theta^*, \widetilde{r})$. Taking $r = \min\{\widetilde{r}, 2B\}$, we have that with probability atleast $1-\delta$

$$\|\hat{\theta}_n - \theta^*\|_2 \leq \frac{4c_0 L_g}{\alpha}\sqrt{\frac{d}{n}} + \frac{2c_1 L_g}{\alpha}\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{2c_2 K_\ell}{\alpha n}\log(1/\delta)} + 2\sqrt{\frac{\Delta_n}{\alpha}} + 2\sqrt{\frac{K_\ell}{\alpha}}$$

Taking $\Delta_n = O(n^{-1})$, we are done.

□

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

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

where $\Delta_n = O(n^{-1})$.

G Proof of "Fast Rate" Wasserstein-DRO-REBEL

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 $\mathcal{Z}$.

We note the following

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

where π is the coupling between the marginal distributions $\xi \sim \mathbb{P}$ and $\eta \sim \mathbb{P}^\circ$. This equality holds just from the marginal property of a coupling. Now for a function $\ell(z; \theta)$ that is L_z-Lipschitz, notice the following

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

where the first inequality holds from $\ell(z; \theta)$ being Lipschitz and the last inequality holds from the monotonicity of type-p Wasserstein Distance. In Lemma 13, we showed that $\ell(z; \theta)$ is $4K_g/\eta$-Lipschitz in θ. Thus putting everything together we find

$$\begin{aligned}
\mathcal{L}^{\mathcal{W}_p}(\theta) \; - \; \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z; \theta)\right] &= \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathcal{W}_p)} \left(\mathbb{E}_{z \sim \mathbb{P}}[\ell(z; \theta)] - \mathbb{E}_{\mathbb{P}^\circ}\left[\ell(z; \theta)\right] \right) \\
&\leq \frac{4K_g}{\eta} \sup_{\mathbb{P} \in \mathcal{B}_{\varepsilon_n}(\mathbb{P}^\circ; \mathcal{W}_p)} \left(\int d(\xi, \eta)^p \pi(d\xi, d\eta) \right)^{1/p} \\
&\leq \frac{4K_g}{\eta} \varepsilon_n
\end{aligned}$$

Taking $\varepsilon_n \asymp n^{-1}$, we get $\Delta_n = O(n^{-1})$. □

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

Proof of Corollary 2. First recall Lemma 9 (Gibbs variational principle characterization of the KL divergence). For probability measures $\mathbb{P}, \mathbb{Q}$

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

Now, for $\lambda \geq 0$, take $g(x) = \lambda(f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta)$ where f_θ is parameterized by some $\theta \in \Theta$. Then since an affine estimator will not necessarily achieve the supremum, we have that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{Q}) &= \sup_{g: \mathcal{Z} \to \mathbb{R}} \{\mathbb{E}_{\mathbb{P}}[g] - \log \mathbb{E}_{\mathbb{Q}}[e^g]\} \\
&\geq \sup_{\lambda \geq 0} \left\{ \lambda(\mathbb{E}_{\mathbb{P}} f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta) - \log \mathbb{E}_{\mathbb{Q}}\left[e^{\lambda(f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta)} \right] \right\}
\end{aligned}$$

Now suppose that $f_\theta \in [0, K_\ell]$ a.s. Then, by Lemma 10 (Hoeffding's Lemma), we know that $f_\theta \sim \mathcal{SG}\left(\frac{K_\ell}{2}\right)$ so we have that

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

Thus we have

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{Q}\right) &\geq \sup_{\lambda \geq 0}\left\{\lambda\left(\mathbb{E}_{\mathbb{P}} f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta\right) - \log \mathbb{E}_{\mathbb{Q}}\left[e^{\lambda(f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta)}\right]\right\} \\ &\geq \sup_{\lambda \geq 0}\left\{\lambda\left(\mathbb{E}_{\mathbb{P}} f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta\right) - \frac{\lambda^2 K_\ell^2}{8}\right\} \end{aligned}$$

Note that the objective is strictly concave in λ and so the supremum will be attained at a unique maximizer. That maximizer is as follows:

$$\lambda^* = \frac{4\left(\mathbb{E}_{\mathbb{P}} f_\theta - \mathbb{E}_{\mathbb{Q}} f_\theta\right)}{K_\ell^2}$$

Plugging this in, we find that

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

Rearranging terms, we find

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

Now take $\mathbb{Q} = \mathbb{P}^\circ$ and $f_\theta = \ell(z;\theta)$. From Lemma 12, we showed that $\ell(z;\theta) \in [0, K_\ell]$ a.s. Thus,

$$\mathbb{E}_{z\sim\mathbb{P}}\left[\ell(z;\theta)\right] - \mathbb{E}_{z\sim\mathbb{P}^\circ}\left[\ell(z;\theta)\right] \leq \frac{1}{2} K_l \sqrt{D_{\mathrm{KL}}\left(\mathbb{P} \mid\mid \mathbb{P}^\circ\right)}$$

Putting everything together we get

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

Taking $\varepsilon_n \asymp n^{-2}$, we get $\Delta_n = O(n^{-1})$ □

I Proof of "Fast Rate" χ^2-DRO-REBEL

Proof of Corollary 3. By Theorem 7 (Hammersley-Chapman-Robbins (HCR) lower bound), we immediately have

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

Since $\ell(z;\theta) \in [0, K_\ell]$ a.s., we have $\mathrm{Var}\left(\ell(z;\theta)\right) \leq K_l^2$. Thus we have

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

Taking $\varepsilon_n \asymp n^{-2}$, we get $\Delta_n = O(n^{-1})$ □