ROBUST RLHF FOR LLM ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Human Feedback (RLHF) aligns large language models (LLMs) with human preferences, but its reliance on static, offline preference data leaves aligned models vulnerable to shifts in deployment contexts and user preferences. We introduce two distributionally robust (DR) alignment formulations under shifts measured by KL divergence, along with tractable PPO-based algorithms for each: JORAL (joint context–preference shift robust alignment) and CORAL, a simpler context-shift formulation in which the preference model remains unchanged. By directly connecting the log-linear policy class commonly used in theoretical analyses to modern LLMs, we prove that both CORAL and JORAL achieve $O(n^{-1/2})$ sample complexity for the frozen-backbone, last-layer LLM policy class. Experiments on LLAMA-3.2-1B-INSTRUCT show that JORAL provides the most consistent and superior performance improvements across shifted evaluation environments, generally outperforming non-robust PPO and DPO and comparing favorably with existing robust-alignment methods, while CORAL exhibits effective robustness to context shifts.

1 INTRODUCTION

Aligning large language models (LLMs) with human intentions has become a central objective in modern machine learning, as a crucial step toward producing outputs that are useful, safe, and consistent with human intent. Reinforcement Learning from Human Feedback (RLHF) has emerged as a widely adopted framework for this purpose (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022), and has demonstrated empirical success (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021; Nakano et al., 2021; Ouyang et al., 2022).

A typical deployment of RLHF for LLMs includes the following steps (Ouyang et al., 2022).

1^{st} *Step: Supervised Pre-training*. Pre-train an LLM π_0 via supervised learning.

2^{nd} *Step: Reward Modeling.* Train a reward model based on the pre-trained LLM using human feedback. Let $\mathcal{S}$ be the set of states (contexts) and $\mathcal{A}$ be the set of actions (responses). Given any context $s \in \mathcal{S}$, two responses $a^0, a^1 \in \mathcal{A}$ are sampled independently from a reference policy (typically the pre-trained π_0 from the 1^{st} step). A human labeler then indicates a preference between a^0 and a^1 according to the Bradley-Terry (BT) model (Bradley & Terry, 1952):

$$P_{\phi^*}(a^1 \succ a^0 \mid s) = \frac{\exp(r_{\phi^*}(s, a^1))}{\exp(r_{\phi^*}(s, a^1)) + \exp(r_{\phi^*}(s, a^0))}, \tag{1}$$

where $a^1 \succ a^0$ denotes that response a^1 is preferred over a^0, and the reward function r_{ϕ^*} is parametrized as $r_{\phi^*}(s, a) = \langle \phi^*, \psi(s, a) \rangle$ for some known and fixed feature function $\psi(s, a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, which can be viewed as a pre-trained LLM with its last linear head removed.[1]

We use the label $w \in \{0, 1\}$ to indicate the human labeler's preference: $w = 0$ if $a^0 \succ a^1$ and $w = 1$ otherwise. All samples are collected in an offline dataset $\mathcal{D} := \{(s_i, a_i^0, a_i^1, w_i)\}_{i \in [n]}$, where, for each sample $i \in [n]$, the context $s_i \overset{\text{i.i.d.}}{\sim} P_S$, $a_i^0, a_i^1 \overset{\text{i.i.d.}}{\sim} \pi_0(\cdot \mid s_i)$, and the preference w_i is drawn from the BT model P_{ϕ^*}. From $\mathcal{D}$, we learn the reward model by computing the maximum likelihood estimator of the true parameter ϕ^*.

[1] The interpretation of ψ is detailed in Section 2.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

3^{rd} *Step: RL Alignment Fine-Tuning.* Fine-tune the pre-trained LLM based on the learned reward model within the LLM policy class Π, under the defined reward function:

$$\max_{\pi\in\Pi} \mathbb{E}_{P_S}\big[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}[r_{\phi^*}(S,A)] - \beta D_{\mathrm{KL}}[\pi(\cdot \mid S)\|\pi_0(\cdot \mid S)]\big] \tag{2}$$

using Proximal Policy Optimization (PPO) (Schulman et al., 2017; Wang et al., 2020).

Despite the empirical success of RLHF in LLM alignment, existing approaches remain fragile, largely due to the diversity of usage contexts and human preferences. Alignment methods typically rely on static, offline preference datasets, whereas real-world user preferences vary across geography, demographics, language, and evolving social norms (Tien et al., 2022). Human preferences and language habits also drift over time (Son et al., 2024), and such shifts can degrade the performance of fine-tuned models outside the distribution on which they were aligned. A growing literature therefore studies distributionally robust (DR) alignment for RLHF (Wu et al., 2025; Xu et al., 2025; Mandal et al., 2025). While these works establish the value of distributional robustness in LLM alignment, several questions remain.

First, several recent DR alignment methods focus on direct preference optimization (DPO) (Rafailov et al., 2023) based objectives (Wu et al., 2025; Xu et al., 2025). Reward-model-based PPO and DPO have different statistical and computational tradeoffs, and empirical studies show that PPO can outperform DPO in some settings and can benefit from improvements to the reward model (Ivison et al., 2024). This motivates studying how distributional robustness can be incorporated directly into the classical PPO-based RLHF pipeline.

Second, prior DR alignment methods address joint distribution shifts of context and preference, which is appropriate when both may change. However, these formulations do not explicitly separate context shift from preference shift within a common RLHF LLM alignment framework.

Third, theoretical analyses of DR LLM alignment commonly adopt log-linear policy classes (Xu et al., 2025; Mandal et al., 2025). Although analytically convenient, their precise relationship to practical autoregressive LLM policies is often left implicit, leaving it unclear how the resulting guarantees should be interpreted for LLM alignment.

Our contributions. We develop a DR RLHF framework for LLM alignment, yielding scalable PPO-based algorithms that are robust within KL-divergence uncertainty sets. We first introduce *context-shift* robust alignment, CoRAL, as a simple context-only instance of the framework. We then introduce JoRAL for *joint context-preference shift*. Strong duality reduces both robust objectives to one-dimensional dual optimizations, which induce adversarial sample weights that can be incorporated directly into standard PPO updates.

Theoretically, we make explicit the connection between the log-linear policy class commonly used in analyses and the softmax output layer of a modern LLM. The equivalence is exact at the next-token level when the backbone is frozen and only the output head is trained; the resulting response-level autoregressive policy is a product of such conditionals, and can be interpreted as a structured extension of the log-linear model. We establish high-probability robust policy-value regret bounds of $\tilde{O}(\sqrt{d/n})$ for CoRAL and JoRAL.

We draw two distinctions between our results and related work. The existing DR alignment literature (Xu et al., 2025; Mandal et al., 2025) obtains similar upper bounds, but their proofs rely on the log-linear policy class assumption, leaving it unclear how these results should be interpreted for modern LLM alignment. Classical DR policy-learning work, on the other hand, obtains related $\sqrt{1/n}$-type regret rates when rewards are directly observed (Si et al., 2023; Kallus et al., 2022; Mo et al., 2021; Wang et al., 2024b). Our contribution is a DR policy-learning analysis specific to RLHF-based LLM alignment, where reward is estimated from pairwise BT observations and its estimation error propagate during training.

On the empirical side, we fine-tune LLaMA-3.2-1B-Instruct (Meta, 2024) using CoRAL and JoRAL, and compare them with classic PPO, DPO, as well as existing robust-alignment baselines (Schulman et al., 2017; Rafailov et al., 2023; Xu et al., 2025). Across shifted evaluation environments, JoRAL greatly improves on the non-robust baselines and compares favorably with existing robust alignment methods. CoRAL excels when context shift presents.

1.1 Related works

Our framework is related to the DR RLHF literature (Huang et al., 2024; Rafailov et al., 2023; Ramesh et al., 2024; Wu et al., 2025). Among these, the closest is the work of Xu et al. (2025); Mandal et al. (2025), which focuses on joint-shift DR RLHF with DPO (Rafailov et al., 2023) under total variation (TV) and Kullback–Leibler (KL) divergence, and Wasserstein distance, respectively, on a given preference dataset. Mandal et al. (2025) additionally studies reward-model-based DR RLHF. Our focus is complementary: we develop a KL-DRO formulation tailored to the classical reward-model-based PPO pipeline, in which distributional robustness is applied to the alignment value while the standard KL-to-reference regularizer remains averaged under the nominal context distribution. This separation also permits a finite-sample analysis that explicitly propagates the error from reward-model estimation based on pairwise preferences.

Our research is also closely related to the following streams of literature.

Robust RLHF. This line of work has explored robust reward learning as a means of achieving robust RLHF (Padmakumar et al., 2024; Yan et al., 2024; Bukharin et al., 2024). For example, Bai et al. (2022) proposes adjusting the weighting of multiple loss components associated with different topics to enable robust reward learning. Chakraborty et al. (2024) proposes learning multiple reward functions for different subpopulations via an expectation–maximization approach, and subsequently deriving a robust policy through a max–min optimization over these rewards. Chowdhury et al. (2024) considers robust learning under an ε-fraction of corrupted preference labels and provides an algorithm robust to label noise, assuming prior knowledge of ε.

DR Offline RL. At a coarse glance, our framework appears closely related to DR offline RL, which has been studied extensively (Liu et al., 2019; Mo et al., 2021; Kido, 2022; Si et al., 2023; Kallus et al., 2022; Wang et al., 2024b; Mu et al., 2022). The goal of robust offline RL is to enhance model performance under distributional shift between training and testing data by adopting a minimax statistical learning framework (Namkoong & Duchi, 2016; Chen & Paschalidis, 2018; Mohajerin Esfahani & Kuhn, 2018; Kuhn et al., 2019; Levy et al., 2020; Duchi & Namkoong, 2021; Gao et al., 2024). In this framework, an adversary perturbs the testing distribution within a prescribed uncertainty set around the training distribution (typically measured via an f-divergence) to minimize the expected reward, while the learner seeks to maximize the resulting worst-case reward, learning from observed rewards. Our DR-PPO formulation is fundamentally different, however, as it learns from preference datasets consisting of pairwise comparisons, which introduces unique challenges: the reward is not directly observable, and preference estimation error propagates into reward estimation. A further challenge arises from the stochastic policy class, which renders classic DR RL sample complexity analyses based on policy class complexity measures ineffective. To address this, we leverage the Lipschitzness of the robust policy values under mild regularity conditions and adopt the chaining method (Zhou et al., 2023).

2 Distributionally robust alignment

The DR alignment problem in interest focuses on the 2^{nd} and 3^{rd} step of the RLHF. Consider the product space $\mathcal{P} = \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times \{0, 1\}$. The data-generating distribution P in the 2^{nd} step can be defined as $P := P_S \times \pi_0 \times \pi_0 \times P_{\phi*}$ with density

$$P(s, a^1, a^0, w) := P_S(s) \cdot \pi_0(a^0 \mid s) \cdot \pi_0(a^1 \mid s) \cdot P_{\phi*}(a^w \succ a^{1-w} \mid s).$$

Throughout the paper, we make the following regularity assumption on the parameterization of the BT reward model in Equation (1).

Assumption 1. *The fixed feature function ψ is known with $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\psi(s, a)\| \leqslant U$. We also let $\phi^* \in \Phi_B$ where $\Phi_B = \{\phi \in \mathbb{R}^d : \|\phi\| \leqslant B\}$.*

Bounded linear reward representations of this form are standard in the theoretical analysis of reinforcement learning and learning from human preferences (Agarwal et al., 2021; Modi et al., 2020; Zhu et al., 2023; Xu et al., 2025; Nika et al., 2024; Chowdhury et al., 2024).

Our goal is to learn a DR optimal policy from a given policy class Π from $\mathcal{D}$ collected in P. A policy class particularly convenient for theoretical analysis is the log-linear policy class.

Example 1 (Log-linear policy class)**.** *Define* $\Theta := \{\theta \in \mathbb{R}^d : \langle 1, \theta \rangle = 0,\ \|\theta\| \leqslant B\}$*. The log-linear policy class defined on* Θ *is*

$$\Pi_\Theta = \left\{ \pi_\theta : \pi_\theta(a \mid s) = \frac{\exp(\theta^\top \psi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \psi(s,a))},\ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \theta \in \Theta \right\}. \tag{3}$$

Log-linear policies have been widely adopted in theoretical analyses of RLHF and preference optimization (Zhu et al., 2023; Nika et al., 2024; Chowdhury et al., 2024; Xu et al., 2025; Mandal et al., 2025). Their connection with modern language models is often invoked informally in these literature: the policy is viewed as a softmax applied to a learned representation. The relationship can in fact be made exact for a frozen LLM backbone with a trainable output head.

Modern autoregressive LLMs may be decomposed into (i) an embedding and transformer backbone, (ii) a final normalization layer, and (iii) a linear language-modeling head followed by a softmax (Raschka, 2024). Suppose that the backbone and normalization layers of a pre-trained LLM are frozen. Let $H(s) \in \mathbb{R}^h$ denote the resulting hidden representation of the current tokenized context s, and let $W = (w_1, \cdots, w_D)^\top \in \mathbb{R}^{D \times h}$ be the trainable output head, where D is the vocabulary size.

Example 2 (Frozen-backbone, last-layer LLM policy class)**.** *Let* $\mathrm{Dict} = \{1, \cdots, D\}$ *denote the vocabulary dictionary. When only the output head* W *is fine-tuned, the next-token distribution is*

$$\pi_\theta(a \mid s) = \frac{\exp(w_a^\top H(s))}{\sum_{a' \in \mathrm{Dict}} \exp(w_{a'}^\top H(s))} = \frac{\exp(\theta^\top \psi_{\mathrm{LLM}}(s,a))}{\sum_{a' \in \mathrm{Dict}} \exp(\theta^\top \psi_{\mathrm{LLM}}(s,a'))} =, \quad a \in \mathrm{Dict}, \tag{4}$$

where we define $\theta = \mathrm{vec}(W) \in \mathbb{R}^{Dh}$ *and* $\psi_{\mathrm{LLM}}(s,a) = H(s) \otimes e_a$ *with* $e_a \in \mathbb{R}^D$ *being the* a*-th standard basis vector. Note that* $\theta^\top \psi_{\mathrm{LLM}}(s,a) = w_a^\top H(s)$*, and hence* (4) *is exactly a log-linear policy, with parameter dimension* $d = Dh$*.*

The softmax parameterization θ *is invariant to a common shift of all rows of the output head. In particular, for every* $c \in \mathbb{R}^h$*,* $\pi_{\mathrm{vec}(W + \mathbf{1}_D c^\top)}(a \mid s) = \pi_{\mathrm{vec}(W)}(a \mid s)$*. Thus,* W *is identifiable only modulo in the subspace* $\{\mathbf{1}_D c^\top : c \in \mathbb{R}^h\}$ *where* $\mathbf{1}_D := (1, \cdots, 1)^\top \in \mathbb{R}^D$*, and the same policy may be represented by* W *with arbitrarily large* $\|W\|_F$*. Without loss of generality, we select the centered representative of each equivalence class and impose* $\mathbf{1}_D^\top W = \mathbf{0}_h^\top$ *(with* $\mathbf{0}_h = (0, \cdots, 0)^\top \in \mathbb{R}^h$*), and define the frozen-backbone, last-layer LLM policy class as* $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$ *where* $\Theta = \{\theta = \mathrm{vec}(W) : W \in \mathbb{R}^{D \times h},\ \mathbf{1}_D^\top W = \mathbf{0}_h^\top,\ \|W\|_F \leqslant B\}$*.*

The centered representation of Θ serves as an identification convention for the softmax quotient space. An analogous zero-sum restriction is used by Zhu et al. (2023).

We note that Example 2 satisfies Assumption 1. The bounded-feature condition admits a direct interpretation, because $H(s)$ is known from the frozen-backbone pre-trained model, and $\|\psi_{\mathrm{LLM}}(s,a)\| = \|H(s)\|$, it suffices to uniformly bound the normalized hidden state. For a LayerNorm output of the form $H(s) = \nu \odot F(s) + \xi$ where $F(s)$ is the input neuron of s and ν, ξ are the learnable scale (weight) and shift (bias) parameters, $U = \|\nu\|_\infty \sqrt{h} + \|\xi\|_\infty$.

This observation interprets the log-linear policy class as the next-token prediction when the transformer representation is frozen and only the final linear output head is trained. We emphasize, however, that the exact equivalence above is a *next-token* statement. In reality, an action may instead represent an entire response $a = (a_1, \ldots, a_T)$, where a_T is the termination token. A frozen-backbone autoregressive LLM then induces the response-level policy

$$\pi_\theta(a \mid s) = \prod_{t=1}^{T} \frac{\exp(\theta^\top \psi_{\mathrm{LLM}}((s, a_{<t}), a_t))}{\sum_{a' \in \mathrm{Dict}} \exp(\theta^\top \psi_{\mathrm{LLM}}((s, a_{<t}), a'))} = \prod_{t=1}^{T} \frac{\exp(w_{a_t}^\top H(s, a_{<t}))}{\sum_{a' \in \mathrm{Dict}} \exp(w_{a'}^\top H(s, a_{<t}))}. \tag{5}$$

Thus the full response policy is a product of log-linear next-token conditionals, but is generally not itself a single log-linear softmax over complete responses with a fixed feature map. Accordingly, Example 2 should be viewed as an exact model of last-layer LLM tuning at the token level, and as a tractable surrogate for the induced response-level autoregressive LLM policy.

2.1 Distributionally robust alignment formulation

In this section, we introduce a DR-PPO framework at two levels of uncertainty, under the KL-divergence. We first develop the context-shift only formulation as a simple instance of DR-PPO, and then introduce the joint-shift robust formulation.

Definition 1 (KL Divergence). *The KL divergence between two distributions P, Q satisfying $Q \ll P$ is defined as $D_{\mathrm{KL}}(Q\|P) := \mathbb{E}_Q[\log \frac{dQ}{dP}]$, where $\frac{dQ}{dP}$ is the Radon-Nikodym derivative.*

Context shift. Let $\mathcal{U}_\delta(P_S)$ denote an uncertainty set centered at the context distribution P_S observed during data collection, with radius $\delta > 0$: $\mathcal{U}_\delta(P_S) = \{Q_S : D_{\mathrm{KL}}(Q_S\|P_S) \leqslant \delta\}$. Following DRO principles, we formulate the DR alignment objective as the expected reward collected by policy $\pi_\theta \in \Pi_\Theta$ under the worst-case shifted distribution in $\mathcal{U}_\delta(P_S)$:

$$\max_{\pi_\theta \in \Pi_\Theta} \min_{Q_S \in \mathcal{U}_\delta(P_S)} \mathbb{E}_{Q_S}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[r_{\phi*}(S,A) \mid S\big]\right] - \beta \mathbb{E}_{P_S}\left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\right]. \quad (6)$$

For brevity, we henceforth denote $X_\theta(s) := \mathbb{E}_{A\sim\pi_\theta(\cdot|S)}[r_{\phi*}(s,A) \mid S = s]$.

Remark 1. *We note that the objective in* (6) *is* not *the exact DR counterpart of* (2)*, because the KL term in* (6) *is not subject to distributional shift within the uncertainty set $\mathcal{U}_\delta(P_S)$. This formulation therefore differs fundamentally from the current DR alignment literature (Xu et al., 2025; Mandal et al., 2025), which allows the worst-case deployment distribution to simultaneously reweight the reward and the KL penalty. We apply distributional robustness only to the alignment value and retain the standard KL-to-reference penalty under P_S. This separation gives the KL term its usual interpretation as a policy regularizer, avoids imposing worst-case reference-policy closeness over every distribution in the uncertainty set, and leads to a simpler dual robust objective.*

For a fixed policy $\pi_\theta \in \Pi_\Theta$, we define the context-shift policy value as

$$\mathcal{V}(\pi_\theta) = \min_{Q_S \in \mathcal{U}_\delta(P_S)} \mathbb{E}_{Q_S}\big[X_\theta(S)\big] - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big]. \quad (7)$$

The following strong duality result transforms the infinite-dimensional DRO problem in (7) into a tractable one-dimensional optimization problem. Its proof is deferred to Appendix F.1.

Lemma 1 (Strong Duality of (7)). *Given any $\pi_\theta \in \Pi_\Theta$, for some $\bar{\alpha} > \underline{\alpha} > 0$,*

$$\mathcal{V}(\pi_\theta) = \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \left\{ -\alpha \log \mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha}\right)\right] - \alpha\delta \right\} - \beta \mathbb{E}_{P_S}\left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\right],$$

where the maximizer is $\alpha^*_{\pi_\theta} \in \arg\max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \{-\alpha \log \mathbb{E}_{P_S}[\exp(-X_\theta(S)/\alpha)] - \alpha\delta\}$.

The policy robust to context shifts within $\mathcal{U}_\delta(P_S)$ is $\pi_\theta^{\mathcal{V}*} := \arg\max_{\pi_\theta\in\Pi_\Theta} \mathcal{V}(\pi_\theta)$. We measure the performance of any policy $\pi_\theta \in \Pi_\Theta$ by its *regret*, defined as the sub-optimality gap in policy robust value $\mathcal{V}$: $\mathrm{Regret}(\pi_\theta; \mathcal{V}) := \mathcal{V}(\pi_\theta^{\mathcal{V}*}) - \mathcal{V}(\pi_\theta)$.

Joint context-preference shift. The specific challenge of a a joint-shift robust alignment formulation is that the classic RLHF alignment objective (2), does not account for the randomness introduced by the choice probability under the BT model $P_{\phi*}$. Accordingly, we define the reward $Y_\theta(s;\phi)$ as the probability that a response from policy π_θ wins over a response from π_{θ_0} under P_ϕ:

$$Y_\theta(s;\phi) = \mathbb{E}_{A\sim\pi_\theta(\cdot|s);A'\sim\pi_{\theta_0}(\cdot|s)}\big[P_\phi(A \succ A' \mid A, A', S)\big].$$

By construction, the reference policy satisfies $Y_{\theta_0}(s;\phi) = \frac{1}{2}$ for any $\phi \in \Phi_B$ and any context $s \in \mathcal{S}$; any policy π_θ that outperforms π_{θ_0} under P_ϕ therefore satisfies $Y_\theta(s;\phi) \geqslant \frac{1}{2}$, and conversely for a policy that underperforms. For a fixed policy $\pi_\theta \in \Pi_\Theta$, we define the joint-shift robust policy value

$$\mathcal{Q}(\pi_\theta) = \min_{Q\in\mathcal{U}_\delta(P)} \mathbb{E}_Q[Y_\theta(S;\phi)] - \beta \mathbb{E}_{P_S}\left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\right], \quad (8)$$

where the uncertainty set is $\mathcal{U}_\delta(P) := \{Q = Q_S \times \pi_{\theta_0} \times \pi_{\theta_0} \times P_\phi : D_{\mathrm{KL}}(Q\|P) \leqslant \delta\}$. For any $Q \in \mathcal{U}_\delta(P)$, the KL divergence can be decomposed by the chain rule as

$$D_{\mathrm{KL}}(Q\|P) = D_{\mathrm{KL}}(Q_S\|P_S) + \mathbb{E}_{S\sim Q_S, A^0\sim\pi_{\theta_0}, A^1\sim\pi_{\theta_0}}\big[D_{\mathrm{KL}}(P_\phi(\cdot \mid S, A^0, A^1)\|P_{\phi*}(\cdot \mid S, A^0, A^1))\big].$$

Hence, any shift $Q \in \mathcal{U}_\delta(P)$ from the data-generating environment P decomposes into two parts: a shift in the context distribution and a shift in the preference model. Solving

$$\max_{\pi\in\Pi} \min_{Q\in\mathcal{U}_\delta(P)} \mathbb{E}_Q\big[Y_\theta(S;\phi)\big] - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big] \quad (9)$$

thus yields a policy robust to joint context-preference shift.

Lemma 2 (Strong Duality of (8)). *Given any $\pi_\theta \in \Pi_\Theta$, for some $\bar{\lambda} > \underline{\lambda} > 0$,*

$$\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]} \left\{ -\lambda \log \mathbb{E}_P\left[\exp\left(\frac{-Y_\theta(S;\phi^*)}{\lambda}\right)\right] - \lambda\delta \right\} - \beta \mathbb{E}_{P_S}\Big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\Big],$$

where the maximizer is $\lambda^*_{\pi_\theta} \in \arg\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]} \{-\lambda \log \mathbb{E}_P[\exp(\frac{-Y_\theta(S;\phi^*)}{\lambda})] - \lambda\delta\}$.

The proof of the above lemma can be found in Appendix G.1. We similarly define the optimal policy $\pi_\theta^{\mathcal{Q}*} := \arg\max_{\pi_\theta\in\Pi_\Theta} \mathcal{Q}(\pi_\theta)$ and the performance measure: $\mathrm{Regret}(\pi_\theta; \mathcal{Q}) := \mathcal{Q}(\pi_\theta^{\mathcal{Q}*}) - \mathcal{Q}(\pi_\theta)$.

3 Robust alignment algorithms

Our alignment algorithm involves two main steps, similar to the RLHF paradigm: (i) reward model learning and (ii) alignment based on the learned reward model.

3.1 Reward model estimation via MLE

To recover the true BT model parameter ϕ^*, the common practice is to find its maximum likelihood estimator (MLE) $\hat{\phi}$ by minimizing the negative log-likelihood function $\ell_\mathcal{D}$ of the collected dataset $\mathcal{D}$. Details of the MLE derivation are deferred to Appendix E.1. The following result from Zhou et al. (2021, Lemma 3.2) generalizes the upper bound on the estimation error of $\hat{\phi}$.

Lemma 3 (Zhu et al. (2023, Lemma 3.2)). *For any $\rho > 0$, with probability at least $1 - \Delta$,*

$$\|\hat{\phi} - \phi^*\|_{\Sigma_\mathcal{D}+\rho I} \leqslant \sqrt{\frac{d + \log(1/\Delta)}{\gamma^2 n} + \rho B^2},$$

where $\Sigma_\mathcal{D} = \frac{1}{n}\sum_{i=1}^n (\phi(s_i, a_i^1) - \phi(s_i, a_i^0))(\phi(s_i, a_i^0) - \phi(s_i, a_i^1))^\top$, and $\gamma = 1/(2 + e^{-BU} + e^{BU})$.

If $\Sigma_\mathcal{D}$ is invertible, or if every $\phi \in \Phi_B$ is orthogonal to the nullspace of $\Sigma_\mathcal{D}$, the upper bound in Lemma 3 holds with $\beta = 0$. Otherwise, one can set $\beta = O(1/(B^2\gamma^2 n))$ to achieve an estimation error rate of $O(\sqrt{(d + \log(1/\Delta))/(\gamma^2 n)})$ with probability at least $1 - \Delta$ (Zhu et al., 2023). Achieving the parametric estimation rate is particularly important for our robust policy learning task, as it yields a regret comparable to the $O(n^{-1/2})$ rate of classic off-policy learning.

3.2 Distributionally robust alignment

With the learned $\hat{\phi}$, we are ready to proceed with DR alignment. For any policy $\pi_\theta \in \Pi_\Theta$, one can use $\hat{\phi}$ to estimate $\mathcal{V}(\pi_\theta)$ and $\mathcal{Q}(\pi_\theta)$. Hence, DR alignment reduces to maximizing the estimated policy values $\mathcal{V}(\pi_\theta)$, $\mathcal{Q}(\pi_\theta)$ using the offline data $\mathcal{D}$. The detailed algorithms for context-shift and joint-shift robust alignment (CoRAL and JoRAL) are given below.

CoRAL. Fix any $\pi_\theta \in \Pi_\Theta$. The inner expectation $\mathbb{E}_{P_S}[\exp(\frac{-X_\theta(S)}{\alpha})]$ can be estimated by $\hat{h}^n(\theta, \alpha) := \frac{1}{n}\sum_{i=1}^n \exp(-\hat{X}_\theta(s_i)/\alpha)$, where $\hat{X}_\theta(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \cdot r_{\hat{\phi}}(s, a)$. We also denote the sample average of the KL divergence term by $K^n(\theta) = \frac{1}{n}\sum_{i=1}^n K_\theta(s_i)$, where $K_\theta(s) = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s)\|\pi_{\theta_0}(\cdot \mid s))$. The policy value $\mathcal{V}(\pi_\theta)$ can then be estimated by

$$\hat{\mathcal{V}}^n(\pi_\theta) := \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} -\alpha \log \hat{h}^n(\theta, \alpha) - \alpha\delta - \beta K^n(\theta), \tag{10}$$

and alignment is thus equivalent to solving $\max_{\pi\in\Pi} \hat{\mathcal{V}}^n(\pi_\theta)$.

JoRAL. Following the same reasoning, the joint shift robust alignment problem is equivalent to $\max_{\pi_\theta\in\Pi_\Theta} \hat{\mathcal{Q}}^n(\pi_\theta)$, where

$$\hat{\mathcal{Q}}^n(\pi_\theta) := \max_{\lambda\in[\underline{\lambda},\bar{\lambda}]} -\lambda \log \hat{g}^n(\theta, \lambda) - \lambda\delta - \beta K^n(\theta) \tag{11}$$

is the estimator of the robust policy value $\mathcal{Q}(\pi_\theta)$. Here $\hat{g}^n(\theta, \lambda) := \frac{1}{n}\sum_{i=1}^n \exp(-\hat{Y}_\theta(s_i; \hat{\phi})/\lambda)$, where $\hat{Y}_\theta(s; \hat{\phi}) = \sum_{a,a'\in\mathcal{A}} \pi_\theta(a \mid s)\pi_{\theta_0}(a' \mid s)\sigma(\langle\hat{\phi}, \psi(s, a) - \psi(s, a')\rangle)$.

Algorithm 1 CoRAL

1: **Input:** Data $\mathcal{D}$, robust level δ
2: $\hat{\phi} \leftarrow \arg\min_{\phi \in \Phi_B} \ell_{\mathcal{D}}(\phi)$
3: Initialize $\pi_{\hat{\theta}} \in \Pi_\Theta$
4: **while** $\pi_{\hat{\theta}}$ does not converge **do**
5: $\quad \hat{\alpha} \leftarrow \underset{\alpha \in [0,\bar{\alpha}]}{\text{argmax}} \; -\alpha \log \hat{h}^n(\hat{\theta}, \alpha) - \alpha\delta$
6: $\quad \pi_{\hat{\theta}} \leftarrow \underset{\pi_{\theta'} \in \Pi_\Theta}{\text{argmax}} \hat{\mathcal{V}}^n(\pi_{\theta'})$ (10), via PPO
7: **end while**
8: **Output:** $\pi_\theta^{\text{CoRAL}} \leftarrow \pi_{\hat{\theta}}$

Algorithm 2 JoRAL

1: **Input:** Data $\mathcal{D}$, robust level δ
2: $\hat{\phi} \leftarrow \arg\min_{\phi \in \Phi_B} \ell_{\mathcal{D}}(\phi)$
3: Initialize $\pi_{\hat{\theta}} \in \Pi_\Theta$
4: **while** $\pi_{\hat{\theta}}$ does not converge **do**
5: $\quad \hat{\lambda} \leftarrow \underset{\lambda \in [\underline{\lambda}, \bar{\lambda}]}{\text{argmax}} \; -\lambda \log \hat{g}^n(\hat{\theta}, \lambda) - \lambda\delta$
6: $\quad \pi_{\hat{\theta}} \leftarrow \underset{\pi_{\theta'} \in \Pi_\Theta}{\text{argmax}} \hat{\mathcal{Q}}^n(\pi_{\theta'})$ (11), via PPO
7: **end while**
8: **Output:** $\pi_\theta^{\text{JoRAL}} \leftarrow \pi_{\hat{\theta}}$

Remark 2 (PPO implementation of Line 6 in Algorithms 1 and 2)**.** *Strong duality makes the policy optimization step, Line 6, in Algorithms 1 and 2 directly compatible with standard policy-gradient methods (detailed analytical expressions are given in Appendix E.3 and E.4). After the maximizer search in Line 5, the usual PPO machinery, including the advantage estimation and clipped policy update, can be readily applied.*

Remark 3. *In the context of LLM alignment, the action space may be too large to directly estimate $\hat{X}_\theta(s)$ and $\hat{Y}_\theta(s;\hat{\phi})$. In this case, we can instead adopt m-sample Monte Carlo estimators $\hat{X}_\theta^m(s) = \frac{1}{m}\sum_{j=1}^m r_{\hat{\phi}}(s,a_j)$ and $\hat{Y}_\theta^m(s;\hat{\phi}) = \frac{1}{m}\sum_{j=1}^m \sigma\big(\langle\hat{\phi}, \psi(s,a_j) - \psi(s,a_j')\rangle\big)$, where, for each query s, we draw m samples $a_j \sim \pi(s)$ and $a_j' \sim \pi_{\theta_0}(s)$, for $j = 1,\dots,m$. Hence the PPO update itself remains on-policy: for each offline prompt in $\mathcal{D}$, responses are freshly sampled from the current policy and evaluated using the learned reward model.*

Remark 4 (Computationally tractable CoRAL and JoRAL)**.** *Line 5 of Algorithms 1 and 2 solves a concave optimization problem, as shown in Appendix E.2. Convergence of Line 6 to the global optimum, however, is difficult to verify, as is also the case for PPO-type algorithms. We therefore claim only that Line 6 of Algorithms 1 and 2 converges to a stationary point, a guarantee comparable to that of classic PPO (Jin et al., 2024).*

3.3 Theoretical results

In this section, we present theoretical guarantees on the sample complexity of CoRAL and JoRAL. The core challenge in the regret analysis lies in taking a union bound over the stochastic policy class Π_Θ in Example 1 and 2, which renders classic sample complexity techniques that rely on deterministic policy class complexity measures, such as the Hamming entropy integral (Wainwright, 2019), inapplicable. To address this, we leverage the Lipschitzness of the robust policy values under Assumption 1 and adopt the chaining method (Zhou et al., 2023). Detailed proofs of the results below are given in Appendix F.2 and G.2, respectively.

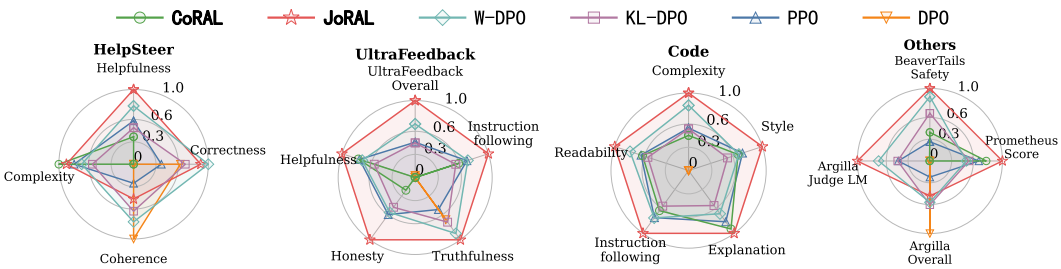
Theorem 1 (Sample Complexity of CoRAL)**.** *Set $\rho = O(1/(B^2\gamma^2 n))$, and let Π_Θ be as in Example 1 or 2. Then under Assumption 1, with probability at least $1-\Delta$,*

$$\textit{Regret}(\pi_\theta^{\text{CoRAL}}; \mathcal{V}) \leqslant 20BU\left(\frac{C_1(1+4BU)}{\underline{\alpha}} + \beta BU\right)\sqrt{\frac{d\log 2}{n}} + \left(\frac{BUC_1}{2\underline{\alpha}} + 2\beta B^2U^2 + C_2\right)\sqrt{\frac{2\log(4/\Delta)}{n}} + \frac{C_1 + 2\beta B^2U^2 + C_2}{2\sqrt{n}} + \frac{C_1 U}{\underline{\alpha}}\|(\Sigma_{\mathcal{D}} + \rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+1+\log(4/\Delta)}{\gamma^2 n}},$$

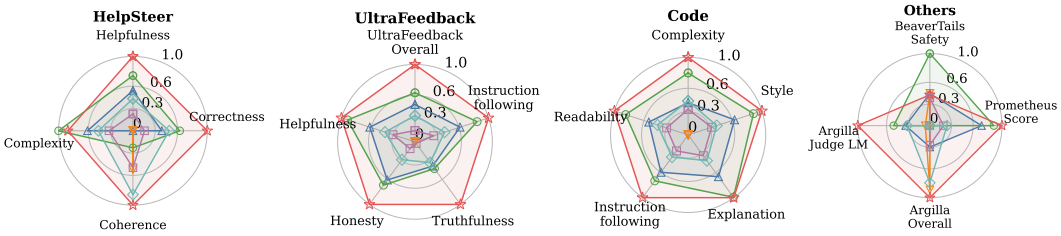
where we write $C_1 = \bar{\alpha}e^{2BU/\underline{\alpha}}$ and $C_2 = 2\sinh(2BU/\underline{\alpha})$.

Theorem 2 (Sample Complexity of JoRAL)**.** *Set $\rho = O(1/(B^2\gamma^2 n))$, and let Π_Θ be as in Example 1 or 2. Then under Assumption 1, with probability at least $1-\Delta$,*

$$\textit{Regret}(\pi_\theta^{\text{JoRAL}}; \mathcal{Q}) \leqslant \left(\frac{25\bar{\lambda}e^{1/\underline{\lambda}}BU}{\underline{\lambda}} + 20\beta B^2U^2\right)\sqrt{\frac{d\log 2}{n}} + \left(\bar{\lambda}e^{1/\underline{\lambda}} + 2\beta B^2U^2\right)\sqrt{\frac{2\log(4/\Delta)}{n}} + \frac{\beta B^2U^2 + 2.5\bar{\lambda}e^{1/\underline{\lambda}}}{\sqrt{n}} + \frac{\bar{\lambda}e^{1/\underline{\lambda}}U^2}{2\underline{\lambda}}\|(\Sigma_{\mathcal{D}} + \rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+1+\log(4/\Delta)}{\gamma^2 n}}.$$

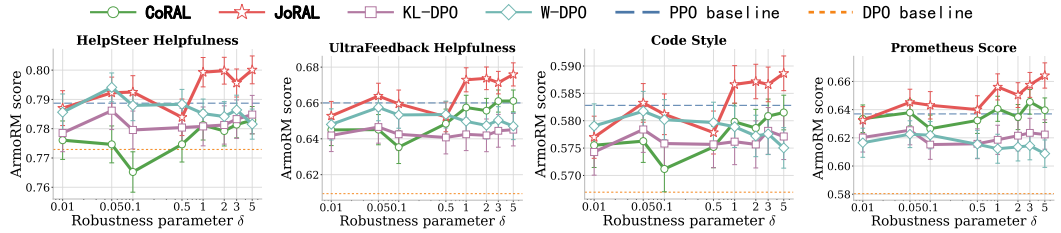


(a) Setting (i), tested on HELPSTEER2 prompts.

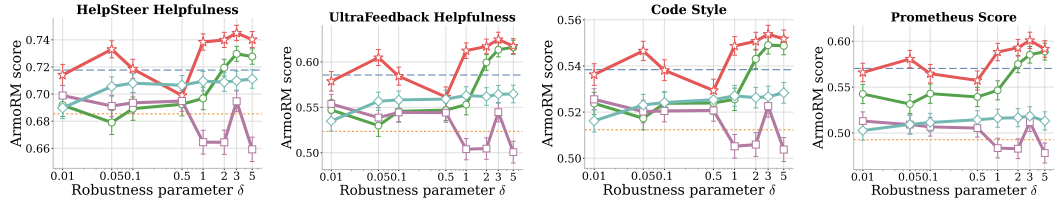


(b) Setting (ii), tested on HELPSTEER1 prompts.

Figure 1: Normalized ARMORM scores of alignment algorithms under their best δ's.



(a) Setting (i), tested on HELPSTEER2 prompts.



(b) Setting (ii), tested on HELPSTEER1 prompts.

Figure 2: ARMORM scores of DR alignment algorithms across different choices of δ. Error bars measure $\pm$2SE, across all tested prompts.

Theorems 1 and 2 apply to policy class in both Example 1 and 2, provided that one has access to $X_\theta(s), Y_\theta(s;\phi)$ for any context $s \in \mathcal{S}$ and $\phi \in \Phi_B$. If not, these results can be extended by adopting the m-sample Monte Carlo estimator in Remark 3, in which case the regret rate of both CORAL and JORAL becomes $O(n^{-1/2} + (nm)^{-1/2})$, due to the additional Monte Carlo sampling error.

4 EXPERIMENT

We evaluate CORAL and JORAL against four baselines on robust LLM alignment tasks: classic PPO (Schulman et al., 2017), classic DPO (Rafailov et al., 2023), and the DR alignemnt baselines W-DPO and KL-DPO (Xu et al., 2025), all post-trained from LLAMA-3.2-1B-INSTRUCT. Further experimental setup details and results are provided in Appendix C.

LLM alignment. We use LLAMA-3.2-1B-INSTRUCT as the pre-trained reference model π_{θ_0} for all methods. The preference dataset $\mathcal{D}$ is built by collecting π_{θ_0}'s responses to prompts sampled from HELPSTEER2 (Wang et al., 2024c), with the Absolute-Rating Multi-Objective Reward Model (ARMORM) (Wang et al., 2024a) serving as the preference model and selecting the preferred response according to its *correctness* dimension. The detailed descriptions of all 18 evaluation dimensions of ARMORM are provided in Table 4. For the PPO-based methods, $\mathcal{D}$ is used to train a reward model that then drives policy optimization; for the DPO-based methods, $\mathcal{D}$ is used to fine-tune π_{θ_0} directly. For each DR method, we sweep the robustness parameter over $\delta \in \{0.01, 0.05, 0.1, 0.5, 1, 2, 3, 5\}$ and report the best-performing δ unless noted otherwise. Model alignemnt training details are deferred to Appendix C.1.

Alignment method evaluation. We evaluate the trained models under three settings. (i) *Preference shift only*: test prompts are sampled from the same HELPSTEER2 distribution, but model responses are scored across all ARMORM dimensions, exposing the model to preference shift alone. (ii) *Joint context and preference shift*: test prompts are instead sampled from HELPSTEER1 (Wang et al., 2023), unseen during training, and again scored across all ARMORM dimensions. (iii) *Stress test*: we use Length-Controlled (LC) ALPACAEVAL (Dubois et al., 2024) to evaluate general instruction-following ability while controlling for verbosity; aligned models generate responses to ALPACAEVAL prompts, and GPT-5.6-LUNA judges them against the GPT-4 reference responses provided by ALPACAEVAL.

Results on setting (i) & (ii). Figure 1 reports normalized (by the best-performing model on each dimension) ARMORM scores on the full set of 18 dimensions. Figure 2 shows the raw ARMORM scores on 4 representative dimensions over a range of robust parameter δ selections, and it complete version is deferred to Appendix C.4. JORAL attains the best performance among the compared methods on most evaluated dimensions. CORAL, by contrast, performs better when tested in setting (ii) than in setting (i). CORAL's advantage in setting (ii) should be credited to its designed robustness to context shift, which is present in that setting. The performance gap between CORAL and JORAL in setting (ii) precisely characterizes how the two sources of shift each contribute to the joint shift: compared to CORAL, JORAL is also robust to additional preference shift. In setting (i), where only preference shift is present, the worst-case DRO construction underlying CORAL instead makes it more conservative than the non-DR PPO baseline, leading to underperformance. Overall, JORAL is the safer default when the source of shift is unknown.

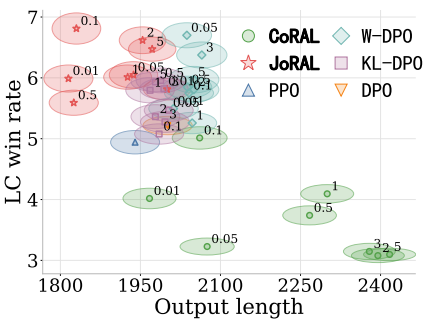


Figure 3: Trade-off between LC win rate and output length in setting (iii). The horizontal and vertical semi-axes of ellipses indicates the $\pm$2SE error bars of output length and LC win rate respectively, computed across all tested prompts.

Results on setting (iii). Figure 3 shows the trade-off between LC win rate (against GPT-4, judged by GPT-5.6-LUNA) and output length on ALPACAEVAL. JORAL again leads, achieving the highest LC win rate while producing the most concise responses. W-DPO is the closest competitor, matching JORAL's win rate but at the cost of noticeably more verbose outputs. CORAL performs relatively poorly here, as expected, since it is designed to hedge only against context shift and offers little protection against the environment shift that ALPACAEVAL induces.

Choice of robustness parameter δ. Figure 2 shows model performance across robustness parameters δ. Selecting δ is a known challenge in the DRO literature (Si et al., 2023), and this challenge is compounded in DR alignment because the true magnitude of context or preference shift is rarely known in advance and, even when observed, is hard to measure (Xu et al., 2025). In setting (i), where only preference shift is present, JORAL performs best at $\delta \approx 1$. In setting (ii), where the joint context-preference shift is large, both CORAL and JORAL benefit from a comparatively large radius, $\delta \geqslant 1$; for JORAL, the best performance in setting (ii) instead requires $\delta \approx 3.0$ to account for the additional preference shift. The best δ also depends on the f-divergence underlying each method: W-DPO, which measures shift via the Wasserstein distance, achieves its best performance at a much smaller $\delta \approx 0.1$, and at that same δ, it in fact outperforms JORAL on some ARMORM dimensions.

AI USE STATEMENT

In this work, we only used generative AI tools for polish writing, double checking technical statements and drafting experiment code. All AI-assisted code are double-checked by authors and experiments are implemented by authors.

REPRODUCIBILITY STATEMENT

All of our experiment results can be reproduced by executing the code package in the supplementary materials. The datasets used are correctly cited, and the training hyperparameters are reported for reproducibility.

REFERENCES

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, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

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

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. *arXiv preprint arXiv:2402.08925*, 2024.

Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

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. *arXiv preprint arXiv:2407.13399*, 2024.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633, 2024.

Ruinan Jin, Shuai Li, and Baoxiang Wang. On stationary point convergence of ppo-clip. In *International Conference on Learning Representations*, volume 2024, pp. 11594–11611, 2024.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pp. 10598–10632. PMLR, 2022.

Daido Kido. Distributionally robust policy learning with wasserstein distance. *arXiv preprint arXiv:2205.04637*, 2022.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in neural information processing systems*, 33:8847–8860, 2020.

Anqi Liu, Hao Liu, Anima Anandkumar, and Yisong Yue. Triply robust off-policy evaluation. *arXiv preprint arXiv:1911.05811*, 2019.

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

Meta. Llama 3.2 model card: Llama-3.2-1b-instruct. `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`, 2024. Accessed: 2026-05-06.

Weibin Mo, Zhengling Qi, and Yufeng Liu. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674, 2021.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Tong Mu, Yash Chandak, Tatsunori B Hashimoto, and Emma Brunskill. Factored dro: Factored distributionally robust policies for contextual bandits. *Advances in Neural Information Processing Systems*, 35:8318–8331, 2022.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

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. *arXiv preprint arXiv:2403.01857*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. Beyond the binary: Capturing diverse preferences with reward regularization. *arXiv preprint arXiv:2412.03822*, 2024.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

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

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

Sebastian Raschka. *Build a large language model (from scratch)*. Simon and Schuster, 2024.

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

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, 2023.

Seongho Son, William Bankes, Sayak Ray Chowdhury, Brooks Paige, and Ilija Bogunovic. Right now, wrong then: Non-stationary direct preference optimization under preference drift. *arXiv preprint arXiv:2407.18676*, 2024.

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

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. *arXiv preprint arXiv:2204.06601*, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024a.

Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. *arXiv preprint arXiv:2412.14297*, 2024b.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pp. 113–122. PMLR, 2020.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023.

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

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

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. *The Thirteenth International Conference on Learning Representations*, 2025.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *arXiv preprint arXiv:2502.01930*, 2025.

Yuzi Yan, Xingzhou Lou, Jialian Li, Yiping Zhang, Jian Xie, Chao Yu, Yu Wang, Dong Yan, and Yuan Shen. Reward-robust rlhf in llms. *arXiv preprint arXiv:2409.15360*, 2024.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

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

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. *arXiv preprint arXiv:1909.08593*, 2019.

A NOTATIONS

We denote the usual p-norm as $\|\cdot\|_p$, and use $\|\cdot\|$ to denote the commonly used 2-norm. When Σ is a positive semi-definite matrix, the semi-norm of a vector x is written as $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$. Denote P to be any probability measure defined on the probability space $(\Omega, \sigma(\Omega), P)$ and P^n to be the empirical distribution of P with n i.i.d. samples $\{x_i\}_{i\in[n]}$. For a random variable/vector X, we use $\mathbb{E}_X[\cdot], \mathbb{P}_X[\cdot]$ to indicate the expectation and the probability taken over the distribution of X, respectively. We use σ to denote the sigmoid function $\sigma(x) = 1/(1+e^{-x})$ for $x \in \mathbb{R}$. For any d-dimensional vector $x, y \in \mathbb{R}^d$, their inner product is denoted as $\langle x, y\rangle = x^\top y$; and their Hadamard product, i.e., element-wise multiplication is denoted as $x \odot y$. For any matrix $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$, their Kronecker product is denoted as $A \otimes B \in \mathbb{R}^{mp\times nq}$. For a policy that maps $\pi : \mathcal{S} \to \mathcal{A}$, we write $\mathbb{P}(\pi(s) = a \mid s) =: \pi(a \mid s)$.

B CONCLUSION

We developed a DR PPO-based RLHF framework for LLM alignment under distribution shifts measured by KL divergence, and proposed two PPO-based alignment algorithms. CoRAL is the simpler formulation targeted only at context-shifts, while JoRAL addresses joint context and preference shifts. Theoretically, by connecting the log-linear policy class to frozen-backbone LLMs with a trainable output head, we established high-probability robust policy-value regret bounds of $\tilde{O}(\sqrt{d/n})$ for both CoRAL and JoRAL. Beyond the KL-divergence setting developed in the main text, we further extend both formulations to uncertainty sets defined by general f-divergences in Appendix D.

A practical limitation of JoRAL is the computational cost introduced by its comparison-based reward structure. Estimating its policy value requires responses from both the current policy and the reference policy, and naively sampling both policies at every training step would incur additional generation cost. One remedy is to collect responses from the reference policy, and their corresponding reward scores, once before training, and reuse them throughout all epochs. This preprocessing removes the additional reference-policy sampling overhead from the online training loop entirely, at the expense of a one-time generation and storage cost. CoRAL has a different practical limitation: despite its clean formulation and robustness when context shift exists, it may lead to verbose response, as observed in our experiments.

Our finite-sample analysis currently covers log-linear policies and, at the token level, frozen-backbone LLMs with a trainable final output head, whereas our empirical implementation performs full-network LLM fine-tuning. Closing this gap between theory and modern post-training practice is an important direction for future work. In particular, it would be valuable to develop finite-sample or convergence guarantees for full-model fine-tuning that jointly account for autoregressive sequence generation, reward-model estimation, and Monte Carlo completion sampling.

C EXPERIMENTAL DETAILS

C.1 TRAINING DETAILS

Our alignment training follows the classic RLHF alignment pipeline on the pre-trained LLaMA-3.2-1B-Instruct.

Step 2: reward model training. Before training, we remove samples with overly long prompts and samples whose generated responses are truncated. The reward model is trained for one epoch on the pairwise preference data collected as described in Section 4, under the sigmoid preference loss specified in Appendix E.1. The training batch size is 256, with a micro-batch size of 2 and a learning rate of 3×10^{-6}.

Step 3: model alignment training. All PPO-based methods initialize the actor and critic from LLaMA-3.2-1B-Instruct, and are trained on prompts from HelpSteer2 for one episode and one epoch. The rollout batch size is 512, with a micro-batch size of 16; the training batch size is 256, with a micro-batch size of 8. The actor and critic learning rates are 5×10^{-7} and 5×10^{-6},

respectively. Both the policy and value clipping coefficients are set to 0.2, and the GAE parameter is set to 0.95. For CoRAL, α is selected from a logarithmic grid of 50 values between 0.01 and 5.0; for JoRAL, λ is selected from a logarithmic grid of 50 values between 0.01 and $1/\delta$, following the theoretical guideline in Appendix G.1. Both methods use early stopping: at each training step, the average policy value ($\hat{\mathcal{V}}^n(\pi_\theta)$ in (10) for CoRAL, $\hat{\mathcal{Q}}^n(\pi_\theta)$ in (11) for JoRAL) over the latest 10 steps is compared with the best average observed so far, and training stops once the improvement fails to exceed 10^{-4} for 10 consecutive checks.

All DPO-based methods are trained using the same hyperparameter settings as in Xu et al. (2025), on the same cleaned HelpSteer2 preference dataset used in Step 2.

Computation resources. All experiments are conducted on a single NVIDIA A100 GPU with 80 GB of memory. Training JoRAL requires the reward difference between responses generated by the current model and by the initial model, which incurs substantial sampling overhead at each step. Since the initial model remains fixed, we precompute its responses and their reward scores in advance; during training, only responses from the current model are generated online, substantially reducing the sampling cost and accelerating training.

C.2 Prompt datasets

This section describes the prompt datasets used in our experiments. Tables 1, 2, and 3 provide example prompts from HelpSteer1, HelpSteer2, and AlpacaEval, respectively.

No.	Example prompt
1	What is the function of the heart, and what is the role of the heartbeat in maintaining human health?
2	Read this for context: \n Hand me this hand The sail is bent The trail is sold and the wind tied in knots Answer the phone The tune is left The stones uncoil in a wide-angle lens A search for clues encounters bliss A lip on the edge of the flowering dark A shell falls with hail Hidden and hides The following sky The descending kiss.\n\n Your task is to:\n From the text given, explain the following phrases in plain English:\n\n - Hand me this hand\n - The sail is bent\n - The trail is sold\n - The wind tied in knots\n - A search for clues encounters bliss\n - A lip on the edge of the flowering dark\n - The following sky\n - The descending kiss.\n \n Use the following format: [Phrase]: [Explanation]
3	What is the most congested traffic intersection in the world? How bad is the traffic there?
4	What is a verb? How can you identify one in a sentence? What are some examples of verbs that can help you understand their definition better? Explore these questions and provide an overview of verbs in language, including their importance to sentence structure and meaning. Consider the different types of verbs and their uses in language. Keep this explanation to under 300 words.
5	You are a nomadic Mongolian farmer. Write a description of the types of crops you grow and livestock you raise. In addition, describe what the environment is like in your home of the Mongolian Steppe and the types of resources available to you there.

Table 1: Example prompts from HelpSteer1.

No.	Example prompt
1	How to plan user properties for membership status which includes monthly subscription, fixed length membership such as membership valid for 6 months, or write a content to get the membership for certain amount of time. the user properties will be used for amplitude and braze
2	Please make a list of independent Fertility coaching and consulting services in USA\n\n
3	I'm looking for some specific marketing advice. I'm starting a life coaching business and I use a method called ËFT Tappingïo help clients. I am also interested in providing relationship coaching. I'm looking to start getting traffic to my site and considering a search strategy, optimising for Google. I'm based in Melbourne Australia so we have a smaller population than other countries, and search volumes are smaller. I've tried doing some keyword research, and am considering optimising for \"Life Coach Melbourne\" and \"EFT Tapping\" but not sure how to validate these choices.
4	Hi, I need you to take on the role of an elite sprinting coach, you specialise in turning athletes into elite 100m sprinters. Your name will be 'Coach'. You use the latest, peer reviewed research to support your recommendations and you always ask the athlete for more information if you need it, to provide the most efficient route to increased performance results. Whenever I start a new message, always start by reminding yourself about these constraints.\n\n Do you understand?
5	what is meant by the term singularity

Table 2: Example prompts from HelpSteer2.

C.3 ArmoRM evaluation details

Table 4 lists the ArmoRM evaluation dimensions and their descriptions.

No.	Example prompt
1	What are the names of some famous actors that started their careers on Broadway?
2	How did US states get their names?
3	Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?
4	What is some cool music from the 1920s?
5	I like to host guests at my home from time to time, and I am gathering recipes of different dishes and drinks to keep things interesting. I am interested in trying some Georgian dishes. Can you give me a recipe for Kubdari?

Table 3: Example prompts from ALPACAEVAL.

Index	Evaluation dimension	Description
1	`helpsteer-helpfulness`	Satisfy the user's request.
2	`helpsteer-correctness`	Provide accurate and relevant information.
3	`helpsteer-coherence`	Clear, logical, and self-consistent response
4	`helpsteer-complexity`	The sophistication of the vocabulary and sentence structure.
5	`ultrafeedback-overall_score`	The overall quality of the response.
6	`ultrafeedback-instruction_following`	How well the response follows the user's instructions.
7	`ultrafeedback-truthfulness`	Whether the response is factually accurate.
8	`ultrafeedback-honesty`	Whether the response acknowledges uncertainty.
9	`ultrafeedback-helpfulness`	Relevant and useful response.
10	`beavertails-is_safe`	Whether the response is labeled as safe in the BeaverTails dataset.
11	`prometheus-score`	Response quality evaluated by Prometheus.
12	`argilla-overall_quality`	The overall response quality rated in the Argilla-Capybara dataset.
13	`argilla-judge_lm`	Response quality rated by the LLM in the Argilla-OpenOrca dataset.
14	`code-complexity`	Whether the code balances implementation complexity and efficiency.
15	`code-style`	Whether the code follows appropriate conventions and best practices.
16	`code-explanation`	The clarity of the code explanation.
17	`code-instruction-following`	Whether the code satisfies the user's instructions.
18	`code-readability`	Clear and well-structured code.

Table 4: Descriptions of the 18 valuation dimensions in ARMORM.

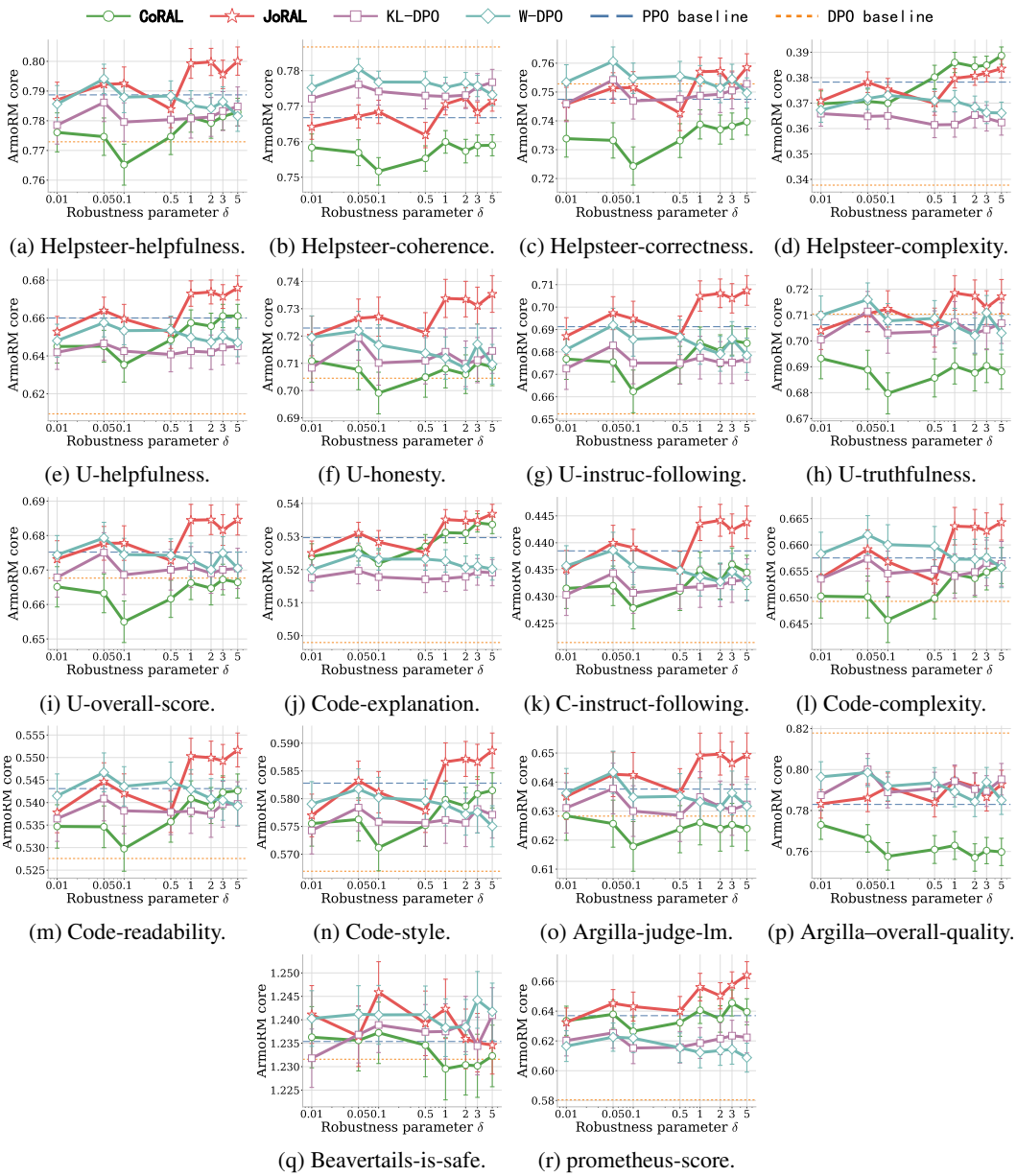
C.4 COMPLETE RESULTS ON SETTING (I) & (II)

Figures 4 and 5 present ARMORM evaluations across all dimensions in settings (i) and (ii), respectively, for different robustness parameters.

Figure 5c provides the cleanest context-shift evaluation: the prompt source changes from HELPSTEER2 to HELPSTEER1, while the evaluation criterion remains ARMORM correctness. In this panel, CORAL outperforms all tested alignment algorithms except JORAL at $\delta = 3.0$, supporting the effectiveness of CORAL in face of context shifts.

It is worth noting that, in both settings (i) and (ii), CORAL's responses are the most elaborate (as shown in Figures 4d and 5d), particularly when the robustness parameter δ is large. The same pattern appears in the ALPACAEVAL setting (iii) (detailed in Appendix C.5), where CORAL produces more verbose responses at larger δ.

For most evaluation dimensions other than correctness in both settings, JORAL remains strong across most test cases, demonstrating robustness to joint context-preference shifts. Its performance generally improves as the robustness parameter δ increases in these two settings.



(a) Helpsteer-helpfulness. (b) Helpsteer-coherence. (c) Helpsteer-correctness. (d) Helpsteer-complexity.

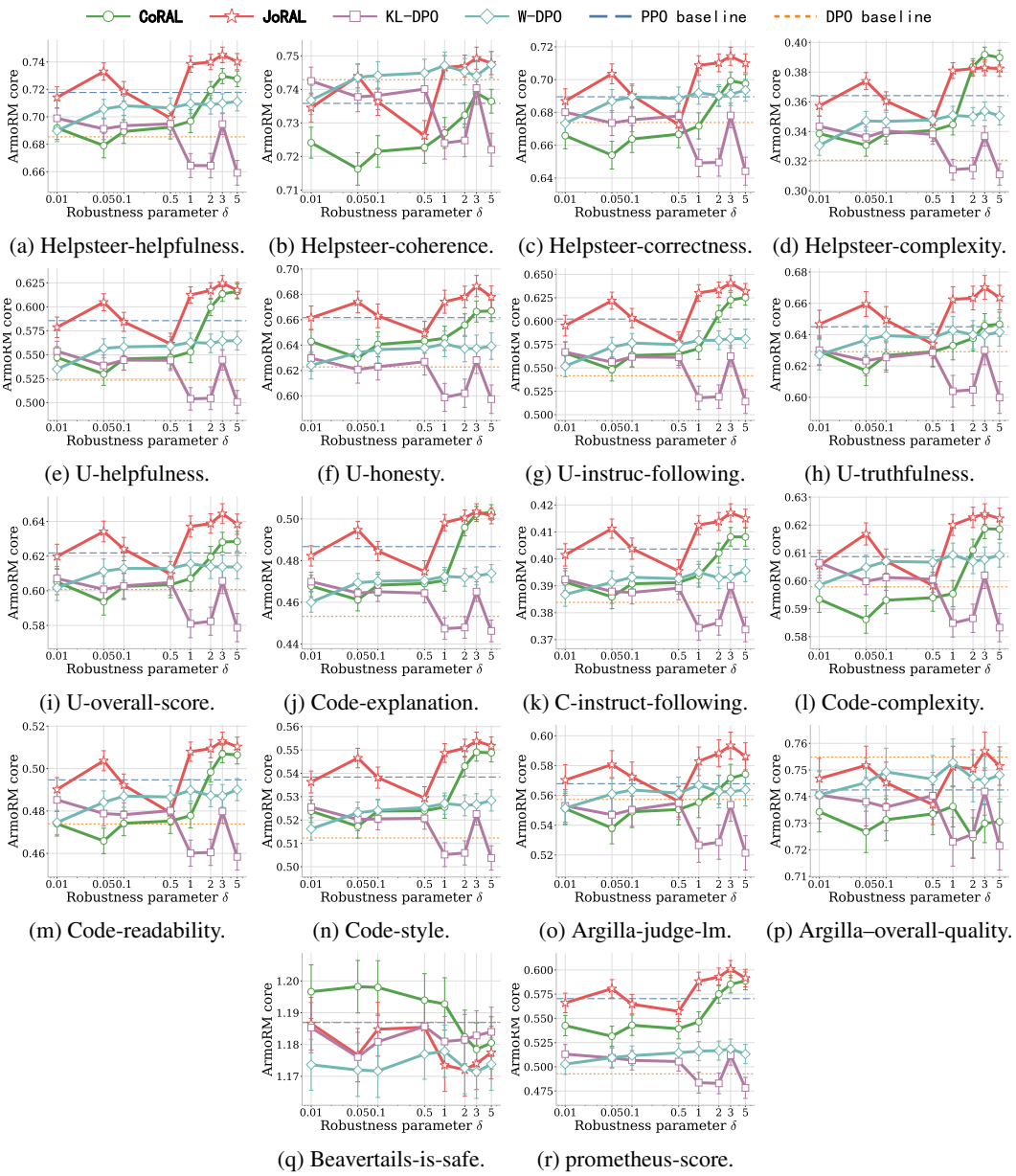
(e) U-helpfulness. (f) U-honesty. (g) U-instruc-following. (h) U-truthfulness.

(i) U-overall-score. (j) Code-explanation. (k) C-instruct-following. (l) Code-complexity.

(m) Code-readability. (n) Code-style. (o) Argilla-judge-lm. (p) Argilla–overall-quality.

(q) Beavertails-is-safe. (r) prometheus-score.

Figure 4: ARMORM scores of alignment algorithms with different robustness parameters in setting (i), tested on HELPSTEER2. "UltraFeedback-" and "instruction-" are abbreviated as "U-" and "instruc-" respectively.



(a) Helpsteer-helpfulness. (b) Helpsteer-coherence. (c) Helpsteer-correctness. (d) Helpsteer-complexity.

(e) U-helpfulness. (f) U-honesty. (g) U-instruc-following. (h) U-truthfulness.

(i) U-overall-score. (j) Code-explanation. (k) C-instruct-following. (l) Code-complexity.

(m) Code-readability. (n) Code-style. (o) Argilla-judge-lm. (p) Argilla–overall-quality.

(q) Beavertails-is-safe. (r) prometheus-score.

Figure 5: ARMORM scores of alignment algorithms with different robustness parameters in setting (ii), tested on HELPSTEER1. "UltraFeedback-" and "instruction-" are abbreviated as "U-" and "instruc-" respectively.

Policy	δ	LC Win Rate	Output Length	Policy	δ	LC Win Rate	Output Length
CoRAL	0.01	4.02 ± 0.17	1967 ± 49.68	WDPO	0.01	5.50 ± 0.17	2012 ± 46.37
	0.05	3.23 ± 0.14	2075 ± 50.75		0.05	6.70 ± 0.21	2037 ± 46.26
	0.1	5.01 ± 0.18	2061 ± 51.37		0.1	5.75 ± 0.16	2041 ± 45.98
	0.5	3.74 ± 0.16	2267 ± 50.24		0.5	5.85 ± 0.18	2038 ± 45.60
	1	4.09 ± 0.15	2300 ± 49.61		1	5.26 ± 0.18	2047 ± 45.85
	2	3.08 ± 0.11	2395 ± 49.56		2	5.81 ± 0.20	2050 ± 46.95
	3	3.15 ± 0.13	2379 ± 48.52		3	6.38 ± 0.21	2065 ± 47.12
	5	3.10 ± 0.10	2417 ± 48.78		5	5.97 ± 0.22	2050 ± 46.96
JoRAL	0.01	5.99 ± 0.22	1815 ± 45.96	KLDPO	0.01	5.82 ± 0.17	1994 ± 45.37
	0.05	6.05 ± 0.22	1937 ± 45.34		0.05	5.47 ± 0.18	2003 ± 45.82
	0.1	$\mathbf{6.81 \pm 0.26}$	1830 ± 45.57		0.1	5.08 ± 0.17	1985 ± 45.96
	0.5	5.59 ± 0.21	1825 ± 45.98		0.5	5.96 ± 0.20	1988 ± 46.14
	1	6.01 ± 0.20	1926 ± 44.84		1	5.80 ± 0.21	1968 ± 45.42
	2	6.62 ± 0.22	1954 ± 43.65		2	5.36 ± 0.20	1978 ± 45.68
	3	5.81 ± 0.18	2000 ± 44.51		3	5.28 ± 0.17	1996 ± 45.84
	5	6.47 ± 0.20	1972 ± 44.41		5	5.93 ± 0.16	1978 ± 45.58
PPO	–	4.94 ± 0.19	1940 ± 45.85	DPO	–	5.22 ± 0.15	2001 ± 46.12

Table 5: Length-controlled win rates and average output lengths of algorithms with different robustness parameters on AlpacaEval.

C.5 Detailed results on setting (iii)

Table 5 shows the LC win rates and output lengths of the alignment algorithms under different robustness parameters. Green and red shading highlight the best- and worst-performing configurations by LC win rate, respectively. Overall, JoRAL again achieves the highest LC win rate while producing relatively short responses. In this setting, LC instruction-following ability appears more closely aligned with the ArmoRM correctness dimension, so a smaller $\delta = 0.1$ helps JoRAL achieve its best performance. W-DPO with $\delta = 0.05$ also performs well, but its outputs are approximately 200 characters longer than those of the best-performing JoRAL configuration. KL-DPO, by contrast, produces moderately short outputs but achieves lower LC win rates.

D Extension to DR alignment under general f-divergences

Now, we generalize our formulation and methodology to uncertainty sets generated by f-divergence D_f different than D_{KL}.

Definition 2 (f-divergence and conjugate)**.** *For any convex function $f : \mathbb{R}^+ \to \mathbb{R}$ satisfying $f(1) = 0$: (i) the f-divergence of distributions $Q \ll P$ is $D_f(Q\|P) := \mathbb{E}_Q[f(dQ/dP)]$; (ii) the Fenchel conjugate of f is $f^*(z) := \sup_{x>0}\langle z, x\rangle - f(x)$.*

Strong duality gives a policy DR value for different distributional shifts, with a second dual variable η, ζ Namkoong & Duchi (2016). Specifically, define $\mathcal{U}_\delta(P; f) := \{Q : D_f(Q\|P) \leqslant \delta\}$ be the uncertainty set centered around P under D_f. The context shift robust policy value is

$$\mathcal{V}(\pi_\theta; f) = \min_{Q_S \in \mathcal{U}_\delta(P;f)} \mathbb{E}_{Q_S}\big[X_\theta(S)\big] - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big],$$

and the strong duality result shows the equivalence:

$$\mathcal{V}(\pi_\theta; f) = \max_{\alpha \geqslant 0, \eta \in \mathbb{R}} \left\{ -\alpha \mathbb{E}_{P_S}\left[f^*\left(\frac{-X_\theta(S) - \eta}{\alpha}\right)\right] - \alpha\delta - \eta \right\} - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big].$$

Similarly the joint shift robust policy value is

$$\mathcal{Q}(\pi_\theta; f) = \min_{Q \in \mathcal{U}_\delta(P;f)} \mathbb{E}_Q[Y_\theta(S;\phi))] - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big]$$

with strong duality equivalence

$$\mathcal{Q}_f(\pi, \delta) = \max_{\lambda \geqslant 0, \zeta \in \mathbb{R}} \left\{ -\alpha \mathbb{E}_P\left[f^*\left(\frac{-Y_\theta(S;\phi^*) - \zeta}{\lambda}\right)\right] - \lambda\delta - \zeta \right\} - \beta \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S)\|\pi_{\theta_0}(\cdot \mid S)\big)\big].$$

The above optimization problems have unique solutions $(\alpha^*_{\pi_\theta}, \eta^*_{\pi_\theta})$ and $(\lambda^*_{\pi_\theta}, \zeta^*_{\pi_\theta})$, respectively. By similar methods as in Algorithm 1 and 2, one can solve the maxmin problems

$$\max_{\pi_\theta \in \Pi_\Theta} \mathcal{V}(\pi_\theta; f), \quad \max_{\pi_\theta \in \Pi_\theta} \mathcal{Q}(\pi_\theta; f)$$

empirically through maximizing $\hat{\mathcal{V}}^n(\pi_\theta; f)$, $\hat{\mathcal{Q}}^n(\pi_\theta; f)$.

This general formulation also subsumes the D_{KL}subproblem, noting that the D_{KL} corresponds to the choice $f(x) = x \log x$ in Definition 2. However, under KL-divergence, the minimization over η, ζ admits a closed-form solution that recovers Lemmas 1 and 2. As a result, our KL-based methodology yields improved empirical performance, since it avoids the need to explicitly optimize over η, ζ.

For general f-divergence defined in Definition 2, the following strong duality holds.

Lemma 4 (Duchi & Namkoong (2021, Proposition 1)). *Let D_f be the f-divergence with some appropriate f in Definition 2. Then*

$$\min_{Q_Z \in \mathcal{U}_\delta(P_Z; f)} \mathbb{E}_{Q_Z}[R_\theta(Z)] = \max_{\alpha \geqslant 0, \eta \in \mathbb{R}} \mathbb{E}_{P_Z}\bigg[- \alpha f^*\bigg(\frac{-R_\theta(Z) - \eta}{\alpha}\bigg) - \alpha\delta - \eta\bigg].$$

Note that one can recover the KL-divergence DR policy value, by plugin $f^*(x) = e^{x-1}$ and notice that

$$\begin{aligned}\min_{Q_Z \in \mathcal{U}_\delta(P_Z; f)} \mathbb{E}_{Q_Z}[R_\theta(Z)] &= \max_{\alpha \geqslant 0, \eta \in \mathbb{R}} \mathbb{E}_{P_Z}\bigg[- \alpha \exp\bigg(\frac{-R_\theta(Z) - \eta}{\alpha} - 1\bigg) - \alpha\delta - \eta\bigg] \\ &= \max_{\alpha \geqslant 0} -\alpha \log \mathbb{E}_{P_Z}\bigg[\exp\bigg(\frac{-R_\theta(Z)}{\alpha}\bigg)\bigg] - \alpha\delta,\end{aligned}$$

where the last equality is due to the closed form solution $\eta^*_{\pi_\theta} = \alpha \log \mathbb{E}_P[\exp(-R_\theta(Z)/\alpha - 1)]$.

E Technical details of CoRAL and JoRAL

E.1 Reward model MLE

For a dataset $\mathcal{D}$ collected in P, an MLE $\hat{\phi}$ aims to find the underlying true BT model parameter ϕ^* by minimizing the negative log likelihood function:

$$\begin{aligned}\ell_{\mathcal{D}}(\phi) &= -\sum_{i=1}^n \log\bigg(\frac{\mathbb{1}\{w_i = 0\} \cdot \exp(r_\phi(s_i, a_i^0))}{\exp(r_\phi(s_i, a_i^0)) + \exp(r_\phi(s_i, a_i^1))} + \frac{\mathbb{1}\{w_i = 1\} \cdot \exp(r_\phi(s_i, a_i^1))}{\exp(r_\phi(s_i, a_i^0)) + \exp(r_\phi(s_i, a_i^1))}\bigg) \\ &= -\sum_{i=1}^n \log\bigg(\mathbb{1}\{w_i = 0\}\sigma\big(\langle \phi, \psi(s_i, a_i^0) - \psi(s_i, a_i^1)\rangle\big) + \mathbb{1}\{w_i = 1\}\sigma\big(\langle \phi, \psi(s_i, a_i^1) - \psi(s_i, a_i^0)\rangle\big)\bigg).\end{aligned}$$

Then the reward model MLE $\hat{\phi} \in \arg\min_{\phi \in \Phi_B} \ell_{\mathcal{D}}(\phi)$ estimates the true underlying ϕ^* BT model parameter, conditioned computed using dataset $\mathcal{D}$.

E.2 Line 5: search of maximizers in CoRAL and JoRAL

Let $R_\theta(Z) \in \mathbb{R}$ be some reward random variable received by $\pi_\theta \in \Pi_\Theta$ under a sample $Z \in \mathcal{Z}$, where $Z \sim P_Z$ for some distribution P_Z supported on $\mathcal{Z}$. The following lemma from Kallus et al. (2022) shows that Line 5 of Algorithm 1 and 2 are strictly convex in α and λ.

Lemma 5 (Strictly concavity, Kallus et al. (2022, Lemma A.1)). *For $\alpha > 0$, the 2^{nd} derivative*

$$\frac{\partial^2}{\partial \alpha^2} - \alpha \log \mathbb{E}[\exp(-R_\theta(Z)/\alpha)] - \alpha\delta \leqslant 0$$

with strict inequality if and only if $R_\theta(Z)$ is not almost surely a constant.

E.3 PPO IN LINE 6: GRADIENT STEP OF CORAL

The policy optimization step in Line 6 of CORAL Algorithm 1 can be done by taking gradient steps to minimize $\hat{\mathcal{L}}^n_{\text{CoRAL}}(\theta, \hat{\alpha}) := \hat{\alpha} \log \hat{h}^n(\theta, \hat{\alpha}) + \beta K^n(\theta)$. At each gradient step t,

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \nabla_\theta \hat{\mathcal{L}}^n_{\text{CoRAL}}(\theta^{(t)}, \hat{\alpha}),$$

where η_t is the learning rate.

We now give the direct computation of $\nabla_\theta \hat{\mathcal{L}}^n_{\text{CoRAL}}(\theta, \hat{\alpha})$. Firstly direct differentiation gives

$$\nabla_\theta \, \alpha \log \hat{h}^n(\theta, \alpha) = -\sum_{i=1}^n \underbrace{\frac{\exp(-\hat{X}_\theta(s_i)/\alpha)}{\sum_{j=1}^n \exp(-\hat{X}_\theta(s_i)/\alpha)}}_{\text{DR exponential weights } w_i^{\text{CoRAL}}} \nabla_\theta \hat{X}_\theta(s_i),$$

where

$$\nabla_\theta \hat{X}_\theta(s_i) = \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \Big[r_{\hat{\phi}}(s_i, A) \nabla_\theta \log \pi_\theta(A \mid s_i) \Big].$$

For the KL sample average term $K^n(\theta)$,

$$\begin{aligned}
\nabla_\theta K_\theta(s_i) =& \nabla_\theta \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \left[\log \frac{\pi_\theta(A \mid s_i)}{\pi_{\theta_0}(A \mid s_i)} \right] \\
=& \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \left[\left(\log \frac{\pi_\theta(A \mid s_i)}{\pi_{\theta_0}(A \mid s_i)} + 1 \right) \nabla_\theta \log \pi_\theta(A \mid s_i) \right] \\
=& \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \left[\log \frac{\pi_\theta(A \mid s_i)}{\pi_{\theta_0}(A \mid s_i)} \nabla_\theta \log \pi_\theta(A \mid s_i) \right],
\end{aligned}$$

where the last step is due to the fact that $\mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)}[\nabla_\theta \log \pi_\theta(A \mid s_i)] = 0$. Hence

$$\nabla_\theta \beta K^n(\theta) = \frac{\beta}{n} \sum_{i=1}^n \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \left[\log \frac{\pi_\theta(A \mid s_i)}{\pi_{\theta_0}(A \mid s_i)} \nabla_\theta \log \pi_\theta(A \mid s_i) \right].$$

Putting everything together,

$$\nabla_\theta \hat{\mathcal{L}}^n_{\text{CoRAL}} = \sum_{i=1}^n \mathbb{E}_{A \sim \pi_\theta(\cdot|s_i)} \left[\left(\frac{\beta}{n} \log \frac{\pi_\theta(A \mid s_i)}{\pi_{\theta_0}(A \mid s_i)} - \frac{\exp(-\hat{X}_\theta(s_i)/\alpha)}{\sum_{j=1}^n \exp(-\hat{X}_\theta(s_j)/\alpha)} r_{\hat{\phi}}(s_i, A) \right) \nabla_\theta \log \pi_\theta(A \mid s_i) \right].$$

Finally, the gradient

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(a \mid s) =& \psi(s, a) - \sum_{a' \in \mathcal{A}} \frac{\exp(\theta^\top \psi(s, a'))}{\sum_{b \in \mathcal{A}} \exp(\theta^\top \psi(s, b))} \psi(s, a') \\
=& \psi(s, a) - \sum_{a' \in \mathcal{A}} \pi_\theta(a' \mid s) \psi(s, a') = \psi(s, a) - \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\psi(s, a')].
\end{aligned}$$

E.4 PPO IN LINE 6: GRADIENT STEP OF JORAL

Similarly, the gradient step in Line 6 of JORAL Algorithm 2 is equivalent to minimize $\hat{\mathcal{L}}^n_{\text{JoRAL}}(\theta, \hat{\lambda}) := \hat{\lambda} \log \hat{g}^n(\theta, \hat{\lambda}) + \beta K^n(\theta)$ via gradient descent. At some gradient step t,

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \nabla_\theta \hat{\mathcal{L}}^n_{\text{JoRAL}}(\theta^{(t)}, \hat{\lambda}).$$

We proceed to compute $\nabla_\theta \hat{\mathcal{L}}^n_{\text{JoRAL}}(\theta, \hat{\lambda})$. Direct differentiation gives

$$\nabla_\theta \, \lambda \log \hat{g}^n(\theta, \lambda) = -\sum_{i=1}^n \underbrace{\frac{\exp(-\hat{Y}_\theta(s_i; \phi^*)/\lambda)}{\sum_{j=1}^n \exp(-\hat{Y}_\theta(s_j; \phi^*)/\lambda)}}_{\text{DR exponential weights } w_i^{\text{JoRAL}}} \nabla_\theta \hat{Y}_\theta(s_i),$$

where

$$\nabla_\theta \hat{Y}_\theta(s_i) = \mathbb{E}_{A\sim\pi_\theta(\cdot|s_i),A'\sim\pi_0(\cdot|s_i)}\Big[\sigma\big(\langle\hat{\phi},\psi(s_i,A)-\psi(s_i,A')\rangle\big)\nabla_\theta\log\pi_\theta(A\mid s_i)\Big].$$

Putting everything together and $\nabla_\theta\beta K^n(\theta)$ from previous Appendix E.3,

$$\nabla_\theta\hat{\mathcal{L}}^n_{\text{JoRAL}} = -\sum_{i=1}^n \mathbb{E}_{A\sim\pi_\theta(\cdot|s_i),A'\sim\pi_0(\cdot|s_i)}\left[\frac{\exp(-\hat{Y}_\theta(z_i)/\lambda)}{\sum_{j=1}^n\exp(-\hat{Y}_\theta(s_j;\pi)/\lambda)}\sigma\big(\langle\hat{\phi},\psi(s_i,a)-\psi(s_i,a')\rangle\big)\nabla_\theta\log\pi_\theta(A\mid s_i)\right]$$
$$+\frac{\beta}{n}\sum_{i=1}^n\mathbb{E}_{A\sim\pi_\theta(\cdot|s_i)}\left[\log\frac{\pi_\theta(A\mid s_i)}{\pi_{\theta_0}(A\mid s_i)}\nabla_\theta\log\pi_\theta(A\mid s_i)\right].$$

F Deferred proofs on context shift robust alignment

F.1 Proof of Lemma 1

Proof of Lemma 1. Fix a $\pi_\theta\in\Pi_\Theta$, from (Hu & Hong, 2013, Theorem 1), there is

$$\min_{Q_S\in\mathcal{U}_\delta(P_S)}\mathbb{E}_{Q_S}\big[X_\theta(S)\big] = \max_{\alpha\geqslant 0}\ -\alpha\log\mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha}\right)\right]-\alpha\delta.$$

We now proceed to prove the upper and lower bound of the maximizer $\alpha^*_{\pi_\theta}$.

Fix some $\alpha\geqslant 0$. By Assumption 1, the expectation

$$\mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha}\right)\right]\geqslant\exp(-BU/\alpha).$$

Hence we can upper bound the function

$$-\alpha\log\mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha}\right)\right]-\alpha\delta\leqslant-\alpha\log\exp(-BU/\alpha)-\alpha\delta = BU-\alpha\delta.$$

The RHS is a strictly decreasing function in α that upper bounds the objective on the LHS. We note that the LHS$\to-\infty$ as $\alpha\to\infty$, and that when $\alpha\to 0$, we have LHS$\to\operatorname{ess\,inf}_{P_S}X_\theta(S)\geqslant -BU$ the objective on the LHS cannot achieve its maximum at ∞. Hence

$$\max_{\alpha\geqslant 0}-\alpha\log\mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha}\right)\right]-\alpha\delta\geqslant\operatorname*{ess\,inf}_{P_S}X_\theta(S)\geqslant -BU$$

Furthermore, $BU-\alpha\delta\geqslant -BU$ gives the upper bound $\alpha^*_{\pi_\theta}\leqslant\frac{2BU}{\delta}$.

To show that $\alpha^*_{\pi_\theta}$ is lower bounded, we follow the argument in Zhou et al. (2021); Panaganti & Kalathil (2022); Xu et al. (2025).

Proposition 1 (Hu & Hong (2013, Proposition 2))**.** *Define the essential supremum of a random variable $R_\theta(z)$ under measure P_Z:*

$$\operatorname{ess\,sup}R_\theta(z) = \inf\{t\in\mathbb{R}:\mathbb{P}(R_\theta(z)>t)=0\},$$

and also define the mass of the measure P_Z on the essential supremum of $R_\theta(z)$: $\kappa_{\pi_\theta}=\mathbb{P}(R_\theta(z)=\operatorname{ess\,sup}R_\theta(z))$. The DRO objective with respect to the reward random variable $R_\theta(z)$ based on KL-divergence is

$$\min_{\alpha\geqslant 0}\alpha\log\mathbb{E}_{P_Z}\left[\exp\left(-\frac{R_\theta(Z)}{\alpha}\right)\right]+\alpha\delta,$$

with the solution

$$\alpha^*_{\pi_\theta}\in\left\{\alpha\log\mathbb{E}_{P_Z}\left[\exp\left(-\frac{R_\theta(Z)}{\alpha}\right)\right]+\alpha\delta\right\}.$$

Then $\alpha^*_{\pi_\theta}=0$ if and only if $R_\theta(Z)<\infty$, $\kappa_{\pi_\theta}>0$ and $\log\kappa_{\pi_\theta}+\delta\geqslant 0$.

By Proposition 1, the condition $\log\kappa_{\pi_\theta}+\delta\geqslant 0$ is problem-dependent due to the uncertainty set $\mathcal{U}_\delta(P_S)$ parameter δ. When κ_{π_θ} is close to zero, the condition $\log\kappa_{\pi_\theta}+\delta\geqslant 0$ is almost never true for a reasonable δ choice. Hence, of any reasonable choice of δ, the three necessary and sufficient conditions in Proposition 1 cannot be all satisfied, and it suffices to ignore the case where $\alpha^*_{\pi_\theta}=0$. Without loss of generality, we find the optimal $\alpha^*_{\pi_\theta}\geqslant\underline{\alpha}>0$. □

F.2 Proof of CoRAL sample complexity Theorem 1

The regret can be decomposed as

$$
\begin{aligned}
&\mathcal{V}(\pi_\theta^{\mathcal{V}*}) - \mathcal{V}(\pi_\theta^{\mathrm{CoRAL}}) \\
=&\mathcal{V}(\pi_\theta^{\mathcal{V}*}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathcal{V}*}) + \hat{\mathcal{V}}^n(\pi_\theta^{\mathcal{V}*}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathrm{CoRAL}}) + \hat{\mathcal{V}}^n(\pi_\theta^{\mathrm{CoRAL}}) - \mathcal{V}(\pi_\theta^{\mathrm{CoRAL}}) \\
\leqslant&|\mathcal{V}(\pi_\theta^{\mathcal{V}*}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathcal{V}*})| + |\mathcal{V}(\pi_\theta^{\mathrm{CoRAL}}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathrm{CoRAL}})| \\
\leqslant&|\mathcal{V}(\pi_\theta^{\mathcal{V}*}) - \mathcal{V}^n(\pi_\theta^{\mathcal{V}*})| + |\mathcal{V}^n(\pi_\theta^{\mathcal{V}*}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathcal{V}*})| + |\mathcal{V}(\pi_\theta^{\mathrm{CoRAL}}) - \mathcal{V}^n(\pi_\theta^{\mathrm{CoRAL}})| + |\mathcal{V}^n(\pi_\theta^{\mathrm{CoRAL}}) - \hat{\mathcal{V}}^n(\pi_\theta^{\mathrm{CoRAL}})| \\
\leqslant&2 \sup_{\pi_\theta \in \Pi_\Theta} |\mathcal{V}(\pi_\theta) - \mathcal{V}^n(\pi_\theta)| + 2 \sup_{\pi_\theta \in \Pi_\Theta} |\mathcal{V}^n(\pi_\theta) - \hat{\mathcal{V}}^n(\pi_\theta)|,
\end{aligned}
$$

where the first inequality is due to the definition of $\hat{\pi}_\theta$. By definition,

$$
\begin{aligned}
&|\mathcal{V}(\pi_\theta) - \mathcal{V}^n(\pi_\theta)| \\
=&\underbrace{\left| \max_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ -\alpha \log \mathbb{E}_{P_S}\left[\exp\left(\frac{-X_\theta(S)}{\alpha} \right) \right] - \alpha\delta \right\} - \max_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ -\alpha \log \left[\frac{1}{n} \sum_{i=1}^n \exp\left(-\frac{X_\theta(s_i)}{\alpha} \right) \right] - \alpha\delta \right\} \right|}_{\mathcal{R}_1(\theta)} \\
&+ \underbrace{\left| \beta\left(\frac{1}{n} \sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_i) \| \pi_{\theta_0}(\cdot \mid s_i)) \right] - \mathbb{E}_{P_S}\big[D_{\mathrm{KL}}(\pi_\theta(\cdot \mid S) \| \pi_{\theta_0}(\cdot \mid S)) \big] \right)}_{\mathcal{R}_2(\theta)} \right|.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&|\mathcal{V}^n(\pi_\theta) - \hat{\mathcal{V}}^n(\pi_\theta)| \\
=&\underbrace{\left| \max_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ -\alpha \log \left[\frac{1}{n} \sum_{i=1}^n \exp\left(-\frac{X_\theta(s_i)}{\alpha} \right) \right] - \alpha\delta \right\} - \max_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ -\alpha \log \left[\frac{1}{n} \sum_{i=1}^n \exp\left(-\frac{\hat{X}_\theta(s_i)}{\alpha} \right) \right] - \alpha\delta \right\} \right|}_{\mathcal{R}_3(\theta)} \\
&+ \underbrace{\left| \beta\left(\frac{1}{n} \sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_i) \| \pi_{\theta_0}(\cdot \mid s_i)) - \frac{1}{n} \sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_i) \| \pi_{\theta_0}(\cdot \mid s_i)) \right) \right|}_{=0}.
\end{aligned}
$$

We proceed to bound term $\sup_\theta |\mathcal{R}_1(\theta)|, \sup_\theta |\mathcal{R}_2(\theta)|, \sup_\theta |\mathcal{R}_3(\theta)|$ separately. To this end, we need the following helper result, and the proof of which is given in Appendix H.

Lemma 6 (Union bound with L-Lipschitz reward)**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. from some population P. Suppose we have a class of function $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\| \leqslant B\}$, and a function $f_\theta : Z \to \mathbb{R}$ such that*

$$
|f_\theta(z) - f_{\theta'}(z)| \leqslant L\|\theta - \theta'\|,
$$

for some constant $L > 0$. Also suppose that all function $f_\theta \in \mathcal{F}$ take values $\underline{f} \leqslant f_\theta(z) \leqslant \bar{f}$ a.s.. Then

$$
\mathbb{E}\left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n f_\theta(z_i) - \mathbb{E}[f_\theta(Z)] \right| \right] \leqslant 10LB\sqrt{\frac{d \log 2}{n}} + \frac{\bar{f} - \underline{f}}{2\sqrt{n}}.
$$

Lemma 7. *With probability at least $1 - \Delta$,*

$$
\sup_{\theta \in \Theta} |\mathcal{R}_1(\theta)| \leqslant \frac{20BU\bar{\alpha}e^{2BU/\underline{\alpha}}}{\underline{\alpha}} \sqrt{\frac{d \log 2}{n}} + \frac{\bar{\alpha}e^{2BU/\underline{\alpha}}}{2\sqrt{n}} + \frac{BU\bar{\alpha}e^{2BU/\underline{\alpha}}\sqrt{\log(1/\Delta)}}{\underline{\alpha}\sqrt{2n}}.
$$

Proof. Define $f_1(s;\theta) = \exp(\frac{-X_\theta(S)}{\alpha})$, and let $F_1(s_1,\cdots,s_n;\theta) = \frac{1}{n} f_1(s_i;\theta)$. We can bound

$$\begin{aligned}
\mathcal{R}_1(\theta) =& \Big| \min_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \big\{\alpha \log \mathbb{E}_{P_S}[f_1(S;\theta)] + \alpha\delta\big\} - \min_{\alpha\in[0,\bar{\alpha}]} \big\{\alpha \log(F_1(s_1,\cdots,s_n;\theta)) + \alpha\delta\big\}\Big| \\
&\overset{(a)}{\leqslant} \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \Big|\alpha \log \mathbb{E}_{P_S}[f_1(S;\theta)] - \alpha \log(F_1(s_1,\cdots,s_n;\theta))\Big| \\
&\leqslant \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \alpha \left| \log\left(\frac{F_1(s_1,\cdots,s_n;\theta)}{\mathbb{E}_{P_S}[f_1(S;\theta)]}\right)\right| \\
&= \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \alpha \left| \log\left(\frac{F_1(s_1,\cdots,s_n;\theta) - \mathbb{E}_{P_S}[f_1(S;\theta)]}{\mathbb{E}_{P_S}[f_1(S;\theta)]} + 1\right)\right| \\
&\overset{(b)}{\leqslant} \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \alpha \left|\frac{F_1(s_1,\cdots,s_n;\theta) - \mathbb{E}_{P_S}[f_1(S;\theta)]}{\exp(-BU/\alpha)}\right| \\
&\leqslant \bar{\alpha} e^{BU/\underline{\alpha}} \max_{\alpha\in[\underline{\alpha},\bar{\alpha}]} \Big|F_1(s_1,\cdots,s_n;\theta) - \mathbb{E}_{P_S}[f_1(S;\theta)]\Big|,
\end{aligned}$$

where step (a) is because $|\inf_x f(x) - \inf_x g(x)| \leqslant \sup_x |f(x) - g(x)|$ for functions f, g; step (b) is due to the fact that $|\log(1+x)| \leqslant |x|$ for $x \geqslant 0$.

Now, fix some $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, the function $f_1(s;\theta) =$ is $\frac{2Ue^{BU/\alpha}}{\alpha}$-Lipschitz in θ, with a fixed context sample $S \in \mathcal{S}$. We also note that $f_1(s;\theta) \in [0, e^{BU/\alpha}]$ a.s.. Applying Lemma 6,

$$\mathbb{E}\left[\sup_{\theta\in\Theta} \big|F_1(s_1,\cdots,s_n;\theta) - \mathbb{E}_{P_S}[f_1(S;\theta)]\big|\right] \leqslant \frac{20BUe^{BU/\alpha}}{\alpha}\sqrt{\frac{d\log 2}{n}} + \frac{e^{BU/\alpha}}{2\sqrt{n}}.$$

Consider two dataset $\{s_i\}_{i\in[n]}$ and $\{s_i'\}_{i\in[n]}$. For any $\theta, \theta' \in \Theta$ and $j \in [n]$,

$$\begin{aligned}
&|F_1(s_1,\cdots,s_j,\cdots,s_n;\theta)| - \sup_{\theta'\in\Theta} |F_1(s_1,\cdots,s_j',\cdots,s_n;\theta')| \\
\leqslant& |F_1(s_1,\cdots,s_j,\cdots,s_n;\theta)| - |F_1(s_1,\cdots,s_j',\cdots,s_n;\theta)| \\
\leqslant& \sup_{\theta\in\Theta} |F_1(s_1,\cdots,s_j,\cdots,s_n;\theta)| - |F_1(s_1,\cdots,s_j',\cdots,s_n;\theta)| \\
=& \sup_{\theta\in\Theta} \frac{1}{n}\left|\exp\left(-\frac{\sum_{a\in\mathcal{A}} \pi_\theta(a \mid s_i)\cdot r_{\phi^*}(s_j,a)}{\alpha}\right) - \exp\left(-\frac{\sum_{a\in\mathcal{A}} \pi_\theta(a \mid s_i)\cdot r_{\phi^*}(s_j',a)}{\alpha}\right)\right| \\
\leqslant& \frac{BUe^{BU/\alpha}}{\alpha n}.
\end{aligned}$$

Taking the supremum of all $\pi_\theta \in \Pi_\Theta$, we have that

$$\sup_{\theta\in\Theta} |F_1(s_1,\cdots,s_j,\cdots,s_n;\theta)| - \sup_{\theta'\in\Theta} |F_1(s_1,\cdots,s_j',\cdots,s_n;\theta')| \leqslant \frac{BUe^{BU/\alpha}}{\alpha n}.$$

By the bounded difference inequality Wainwright (2019, Corollary 2.21), for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\Bigg(&\sup_{\theta\in\Theta} \big|F_1(s_1,\cdots,S_n;\theta) - \mathbb{E}_{P_S}\big[F_1(s_1,\cdots,S_n;\theta)\big]\big| \\
&- \mathbb{E}\big[\sup_{\theta\in\Theta} F_1(s_1,\cdots,S_n;\theta) - \mathbb{E}_{P_S}\big[F_1(s_1,\cdots,S_n;\theta)\big]\big] \geqslant \epsilon\Bigg) \leqslant \exp\left(-\frac{2n\epsilon^2\alpha^2}{B^2U^2e^{2BU/\alpha}}\right).
\end{aligned}$$

Setting $\epsilon = \frac{BUe^{BU/\alpha}\sqrt{\log(1/\Delta)}}{\alpha\sqrt{2n}}$, then with probability at least $1-\Delta$,

$$\begin{aligned}
&\sup_{\theta\in\Theta} \big|F_1(s_1,\cdots,S_n;\theta) - \mathbb{E}_{P_S}\big[f_1(S;\theta)\big]\big| - \mathbb{E}\big[\sup_{\theta\in\Theta} F_1(s_1,\cdots,S_n;\theta) - \mathbb{E}_{P_S}\big[f_1(S;\theta)\big]\big] \\
\leqslant& \frac{BUe^{BU/\alpha}\sqrt{\log(1/\Delta)}}{\alpha\sqrt{2n}}.
\end{aligned}$$

Putting everything together and taking the maximum over $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta} |\mathcal{R}_1(\theta)| \leqslant \frac{20BU\bar{\alpha}e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\sqrt{\frac{d\log 2}{n}} + \frac{\bar{\alpha}e^{2BU/\underline{\alpha}}}{2\sqrt{n}} + \frac{BU\bar{\alpha}e^{2BU/\underline{\alpha}}\sqrt{\log(1/\Delta)}}{\underline{\alpha}\sqrt{2n}}.$$

□

Lemma 8. *With probability at least* $1-\Delta$*,*

$$\sup_{\theta\in\Theta}|\mathcal{R}_2(\theta)| \leqslant 20\beta B^2U^2\sqrt{\frac{d\log 2}{n}} + \frac{\beta(1+2\sqrt{2\log(1/\Delta)})B^2U^2}{\sqrt{n}}.$$

Proof. To apply Lemma 6 in this case, we need the below two results.

Lemma 9. *Under Assumption 1 and a reference policy* $\pi_{\theta_0}\in\Pi_\Theta$*, the KL divergence between* π_θ *and* π_{θ_0} *is bounded by* $0\leqslant D_{\mathrm{KL}}\big(\pi_\theta(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big)\leqslant 2B^2U^2$*, under Assumption 1.*

Lemma 10. *Under Assumption 1 and a reference policy* $\pi_{\theta_0}\in\Pi_\Theta$*,*

$$\sup_{s\in\mathcal{S}}\big|D_{\mathrm{KL}}\big(\pi_\theta(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big) - D_{\mathrm{KL}}\big(\pi_{\theta'}(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big)\big| \leqslant 2BU^2\|\theta-\theta'\|,$$

for any two policies $\pi_\theta,\pi_{\theta'}\in\Pi_\Theta$*.*

Define $f_2(s;\theta)=D_{\mathrm{KL}}\big(\pi_\theta(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big)$, and let $F_2(s_1,\cdots,s_n;\theta)=\frac{1}{n}\sum_{i=1}^n f_2(s_i;\theta)$. We are now ready to apply Lemma 6,

$$\mathbb{E}\bigg[\sup_{\theta\in\Theta}\bigg|\frac{1}{n}\sum_{i=1}^n f_2(s_i;\pi_\theta)-\mathbb{E}_{P_S}[f_2(S;\pi_\theta)]\bigg|\bigg] \leqslant 20B^2U^2\sqrt{\frac{d\log 2}{n}}+\frac{B^2U^2}{\sqrt{n}}.$$

Due to Lemma 9, applying the bounded difference inequality Wainwright (2019, Corollary 2.21), for any $\epsilon>0$,

$$\mathbb{P}\bigg(\sup_{\theta\in\Theta}\big|F_2(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_2(S;\theta)\big]\big| - \mathbb{E}\big[\sup_{\theta\in\Theta}F_1(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_2(S;\theta)\big]\big]\geqslant\epsilon\bigg)$$
$$\leqslant\exp\bigg(-\frac{n\epsilon^2}{8B^4U^4}\bigg).$$

Setting $\epsilon=\frac{2B^2U^2\sqrt{2\log(1/\Delta)}}{\sqrt{n}}$, then with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta}\big|F_2(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_2(S;\theta)\big]\big| - \mathbb{E}\big[\sup_{\theta\in\Theta}F_2(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_2(S;\theta)\big]\big]$$
$$\leqslant\frac{2B^2U^2\sqrt{2\log(1/\Delta)}}{\sqrt{n}}.$$

Putting everything together, with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta}|\mathcal{R}_2(\theta)| \leqslant 20B^2U^2\sqrt{\frac{d\log 2}{n}} + \frac{(1+2\sqrt{2\log(1/\Delta)})B^2U^2}{\sqrt{n}}.$$

□

Lemma 11. *With probability at least* $1-\Delta$*,*

$$\sup_{\theta\in\Theta}|\mathcal{R}_2(\theta)| \leqslant \frac{\bar\alpha e^{2BU/\underline\alpha}U}{\underline\alpha}\|(\Sigma_{\mathcal{D}}+\rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+\log(2/\Delta)}{\gamma^2 n}+\rho B^2} + \frac{80B^2U^2\bar\alpha e^{2BU/\underline\alpha}}{\underline\alpha}\sqrt{\frac{d\log 2}{n}}$$
$$+\frac{(1+\sqrt{2\log(2/\Delta)})2\sinh(2BU/\underline\alpha)}{\sqrt{n}}.$$

Proof. As before, we can bound

$$|\mathcal{R}_3(\theta)| \leqslant \frac{\bar\alpha}{e^{-BU/\underline\alpha}}\bigg|\frac{1}{n}\sum_{i=1}^n\exp\bigg(-\frac{\hat X_\theta(s_i)}{\alpha}\bigg)-\exp\bigg(-\frac{X_\theta(s_i)}{\alpha}\bigg)\bigg|.$$

Define $f_3(s;\theta) = \exp(-\hat{X}_\theta(s)/\alpha) - \exp(X_\theta(s)/\alpha)$, and let $F_3(s_1, \cdots, s_n;\theta) = \frac{1}{n}\sum_{i=1}^n f_3(s_i;\theta)$. If we let $\bar{f}_3(s;\theta) = f_3(s;\theta) - \mathbb{E}_{P_S}[f_3(S;\theta)]$, then $\mathbb{E}_{P_S}[\bar{f}_3(s;\theta)] = 0$, and we can bound

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n f_3(s_i;\theta)\right| &= \sup_{\theta\in\Theta}\left|\mathbb{E}_{P_S}[f_3(S;\theta)]\right| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n f_3(s_i;\theta) - \mathbb{E}_{P_S}[f_3(S;\theta)]\right| \\ &= \sup_{\theta\in\Theta}\left|\mathbb{E}_{P_S}[f_3(S;\theta)]\right| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}_3(s_i;\theta)\right|\end{aligned}$$

We will bound the above two term separately.

The derivative of the function $f_\alpha(x) = e^{-x/\alpha}$ for $x \in [-BU, BU]$

$$\left|\frac{df_\alpha(x)}{dx}\right| = \left|-\frac{1}{\alpha}e^{-x/\alpha}\right| \leqslant \frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}, \forall \alpha \in [\underline{\alpha}, \bar{\alpha}] \text{ and } x \in [-BU, BU].$$

Hence f_α is $\frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}$-Lipschitz. The expectation

$$\begin{aligned}\left|\mathbb{E}_{P_S}[f_3(S;\theta)]\right| &\leqslant \frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\left|\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_\theta(a \mid s)\cdot\left(r_{\hat{\phi}}(s,a) - r_{\phi^*}(s,a)\right)dP_S\right| \\ &\leqslant \frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_\theta(a \mid s)\|\hat{\phi} - \phi^*\|\|\psi(s,a)\|\,dP_S.\end{aligned}$$

Taking supremum of both sides,

$$\begin{aligned}\sup_{\pi_\theta\in\Pi_\Theta}\left|\mathbb{E}_{P_S}[f_3(S;\theta)]\right| &\leqslant \frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\int_{\mathcal{S}}\max_a\|\psi(s,a)\|\cdot\|\hat{\phi} - \phi^*\|\,dP_S \\ &\leqslant \frac{e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\max_{(s,a)}\|\psi(s,a)\|\cdot\|\hat{\phi} - \phi^*\| \\ &\leqslant \frac{e^{2BU/\underline{\alpha}}U}{\underline{\alpha}}\|(\Sigma_{\mathcal{D}} + \rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d + \log(1/\Delta)}{\gamma^2 n} + \rho B^2},\end{aligned}$$

on the event of Lemma 3.

On the other hand, by definition, we have

$$|\bar{f}_3(s;\theta)| \leqslant 2\left(e^{2BU/\alpha} - e^{-2BU/\alpha}\right) = 4\sinh(2BU/\alpha),$$

and simple calculation gives

$$|\bar{f}_3(s;\theta) - \bar{f}_3(s;\theta')| \leqslant \frac{8BU^2e^{2BU/\alpha}}{\alpha}\|\theta - \theta'\|.$$

Then, applying Lemma 6,

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}_3(s_i;\theta)\right| \leqslant \frac{80B^2U^2e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\sqrt{\frac{d\log 2}{n}} + \frac{2\sinh(2BU/\underline{\alpha})}{\sqrt{n}}.$$

By the bounded difference inequality Wainwright (2019, Corollary 2.21), due to the boundedness of $\bar{f}_3$ for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}_3(s_i;\theta)\right| - \mathbb{E}_{P_S}\left[\sup_{\theta\in\Theta}\left|\bar{f}_3(S;\theta)\right|\right] \geqslant \epsilon\right) \leqslant \exp\left(-\frac{n\epsilon^2}{8\sinh^2(2BU/\alpha)}\right).$$

Setting $\epsilon = \frac{2\sinh(2BU/\alpha)\sqrt{2\log(1/\Delta)}}{\sqrt{n}}$, then with probability at least $1 - \Delta$,

$$\begin{aligned}&\sup_{\theta\in\Theta}\left|F_2(s_1, \cdots, s_n;\theta) - \mathbb{E}_{P_S}\left[f_2(S;\theta)\right]\right| - \mathbb{E}\left[\sup_{\theta\in\Theta}F_2(s_1, \cdots, s_n;\theta) - \mathbb{E}_{P_S}\left[f_2(S;\theta)\right]\right] \\ \leqslant &\frac{2\sinh(2BU/\alpha)\sqrt{2\log(1/\Delta)}}{\sqrt{n}}.\end{aligned}$$

Putting everything together, with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta}|\mathcal{R}_2(\theta)| \leqslant \frac{\bar{\alpha}e^{2BU/\underline{\alpha}}U}{\underline{\alpha}}\|(\Sigma_{\mathcal{D}}+\rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+\log(2/\Delta)}{\gamma^2 n}+\rho B^2}+\frac{80B^2U^2\bar{\alpha}e^{2BU/\underline{\alpha}}}{\underline{\alpha}}\sqrt{\frac{d\log 2}{n}} + \frac{(1+\sqrt{2\log(2/\Delta)})2\sinh(2BU/\underline{\alpha})}{\sqrt{n}}.$$

□

We are now ready to prove Theorem 1.

Proof of Theorem 1. The proof is completed by combining the results of Lemma 7, 8 and 11. □

G Deferred proofs on joint shift robust alignment

G.1 Proof of Lemma 2

Proof of Lemma 2. Similarly as before, fix a $\pi_\theta \in \Pi_\Theta$, from (Hu & Hong, 2013, Theorem 1), there is

$$\min_{Q\in\mathcal{U}_\delta(P)}\mathbb{E}_Q[Y_\theta(S;\phi)] = \max_{\lambda\geqslant 0} -\lambda\log\mathbb{E}_P\left[\exp\left(\frac{-Y_\theta(S;\phi^*)}{\lambda}\right)\right]-\lambda\delta.$$

We follow an analogous argument as before to derive the upper and lower bound of $\lambda^*_{\pi_\theta}$.

Note that by definition,

$$Y_\theta(s;\phi)=\sum_{a,a'\in\mathcal{A}}\pi(a\mid s)\pi_0(a'\mid s)P_\phi(a\succ a'\mid s)\in[0,1].$$

Hence $0<\underline{\lambda}\leqslant\lambda^*_{\pi_\theta}\leqslant\bar{\lambda}=1/\delta$. □

G.2 Proof of JoRAL sample complexity Theorem 2

As in CoRAL, the regret can be decomposed as

$$\mathcal{Q}(\pi_\theta^{\mathcal{Q}*})-\mathcal{Q}(\pi_\theta^{\text{JoRAL}})\leqslant 2\sup_{\pi_\theta\in\Pi_\Theta}|\mathcal{Q}(\pi_\theta)-\mathcal{Q}^n(\pi_\theta)|+2\sup_{\pi_\theta\in\Pi_\Theta}|Q^n(\pi_\theta)-\hat{\mathcal{Q}}^n(\pi_\theta)|.$$

Again, by definition,

$$\begin{aligned}&|\mathcal{Q}(\pi_\theta)-\mathcal{Q}^n(\pi_\theta)|\\ =&\underbrace{\left|\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{-\lambda\log\mathbb{E}_P\left[\exp\left(\frac{-Y_\theta(S;\phi^*)}{\lambda}\right)\right]-\lambda\delta\right\}-\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{-\lambda\log\left[\frac{1}{n}\sum_{i=1}^n\exp\left(-\frac{Y_\theta(s_i;\phi^*)}{\lambda}\right)\right]-\lambda\delta\right\}\right|}_{\mathcal{R}_4(\theta)}\\ &+\underbrace{\left|\beta\left(\frac{1}{n}\sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot\mid s_i)\|\pi_{\theta_0}(\cdot\mid s_i))\right]-\mathbb{E}_{P_S}\left[D_{\mathrm{KL}}(\pi_\theta(\cdot\mid S)\|\pi_{\theta_0}(\cdot\mid S))\right]\right)\right|}_{\mathcal{R}_2(\theta)},\end{aligned}$$

and,

$$\begin{aligned}&|\mathcal{Q}^n(\pi_\theta)-\hat{\mathcal{Q}}^n(\pi_\theta)|\\ =&\underbrace{\left|\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{-\lambda\log\left[\frac{1}{n}\sum_{i=1}^n\exp\left(-\frac{Y_\theta(s_i;\phi^*)}{\lambda}\right)\right]-\lambda\delta\right\}-\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left\{-\lambda\log\left[\frac{1}{n}\sum_{i=1}^n\exp\left(-\frac{\hat{Y}_\theta(s_i;\hat{\phi})}{\lambda}\right)\right]-\lambda\delta\right\}\right|}_{\mathcal{R}_5(\theta)}\\ &+\underbrace{\left|\beta\left(\frac{1}{n}\sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot\mid s_i)\|\pi_{\theta_0}(\cdot\mid s_i))-\frac{1}{n}\sum_{i=1}^n D_{\mathrm{KL}}(\pi_\theta(\cdot\mid s_i)\|\pi_{\theta_0}(\cdot\mid s_i))\right)\right|}_{=0}.\end{aligned}$$

We note that $\sup_\theta |\mathcal{R}_2(\theta)|$ is analyzed before, so we proceed to bound term $\sup_\theta |\mathcal{R}_4(\theta)|$ and $\sup_\theta |\mathcal{R}_5(\theta)|$ separately.

Lemma 12. *With probability at least* $1-\Delta$,

$$\sup_{\theta\in\Theta}|\mathcal{R}_4(\theta)| \leqslant \bar{\lambda}e^{1/\underline{\lambda}}\left(\frac{5BU}{\underline{\lambda}}\sqrt{\frac{d\log 2}{n}}+\frac{1+8\sqrt{\log(1/\Delta)}}{2\sqrt{n}}\right).$$

Proof. As in Lemma 7,

$$\begin{aligned}&|\mathcal{R}_4(\theta)|\\ \leqslant&\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\lambda\left|\frac{\frac{1}{n}\sum_{j=1}^n\exp(-Y_\theta(s_i;\phi^*)/\lambda)-\mathbb{E}_P[\exp(-Y_\theta(s;\phi^*)/\lambda)]}{\mathbb{E}_P[\exp(-Y_\theta(s;\phi^*)/\lambda)]}\right|\\ \leqslant&\frac{\bar{\lambda}}{e^{-1/\underline{\lambda}}}\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\left|\frac{1}{n}\sum_{j=1}^n\exp(-Y_\theta(s_i;\phi^*)/\lambda)-\mathbb{E}_P[\exp(-Y_\theta(s;\phi^*)/\lambda)]\right|.\end{aligned}$$

Now fix a $\lambda\in[\underline{\lambda},\bar{\lambda}]$. Define $f_4(s;\theta)=\exp(-Y_\theta(s;\phi^*)/\lambda)$ and define $F_4(s_1,\cdots,s_n;\theta)=\frac{1}{n}\sum_{i=1}^n f_4(s_i;\theta)$. First note that $e^{-1/\lambda}\leqslant f_4\leqslant e^0=1$, and that $f_4(s;\theta)$ is $\frac{U}{2\lambda}$-Lipschitz. To check the Lipschitz constant, apply the chain rule:

$$\nabla_\theta f_4(s;\theta)=-\frac{1}{\lambda}\exp\left(-\frac{Y_\theta(s;\phi^*)}{\lambda}\right)\nabla_\theta Y_\theta(s;\phi^*).$$

Because $0\leqslant Y_\theta(s;\phi^*)\leqslant 1$ and $\lambda>0$, $\exp(-Y_\theta(s;\phi^*)/\lambda)\leqslant 1$. Therefore, $\|\nabla_\theta f_4(s;\theta)\|\leqslant\frac{1}{\lambda}\|\nabla_\theta Y_\theta(s;\phi^*)\|$. It now remains to bound $\|Y_\theta(s;\phi^*)\|$. Compute

$$\nabla_\theta Y_\theta(s;\phi^*)=\sum_{a\in\mathcal{A}}\pi_\theta(a\mid s)(\mathbb{E}_{A'\sim\pi_{\theta_0}(\cdot|s)}[P_{\phi^*}(a\succ A'\mid s,a,A')]-Y_\theta(s;\phi^*))\psi(s,a).$$

Equivalently, for $A\sim\pi_\theta(\cdot\mid s)$,

$$\begin{aligned}\nabla_\theta Y_\theta(s;\phi^*)=&\mathrm{Cov}_\theta\big(\mathbb{E}_{A'\sim\pi_{\theta_0}(\cdot|s)}[P_{\phi^*}(A\succ A'\mid s,A,A')],\psi(s,A)\big)\\ \|\nabla_\theta Y_\theta(s;\phi^*)\|\leqslant&U\sqrt{\mathrm{Var}_{\pi_\theta}\big(\mathbb{E}_{A'\sim\pi_{\theta_0}(\cdot|s)}[P_{\phi^*}(A\succ A'\mid s,A,A')]\big)}\end{aligned}$$

A random variable taking values in an interval of length 1 has variance at most $1/4$. Hence $\|\nabla_\theta Y_\theta(s;\phi^*)\|\leqslant\frac{U}{2}$.

Now, follow a similar argument in Lemma 10, for $\theta_t=\theta'+t(\theta-\theta')$, convexity of Θ gives

$$f_4(s;\theta)-f_4(s;\theta')=\int_0^1\langle\nabla_\theta f_4(s;\theta_t),\theta-\theta'\rangle\,dt.$$

Thus, $\|f_4(s;\theta)-f_4(s;\theta')\|\leqslant\frac{U}{2\lambda}\|\theta-\theta'\|$.

We are now ready to apply Lemma 6,

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n f_4(s_i;\theta)-\mathbb{E}[f_4(S;\theta)]\right|\right]\leqslant\frac{5BU}{\lambda}\sqrt{\frac{d\log 2}{n}}+\frac{1}{2\sqrt{n}}.$$

By the bounded difference inequality Wainwright (2019, Corollary 2.21), for any $\epsilon>0$,

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\theta\in\Theta}\big|F_4(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_4(S;\theta)\big]\big|-\mathbb{E}\big[\sup_{\theta\in\Theta}\big|F_4(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}[f_4(S;\theta)]\big|\big]\geqslant\epsilon\bigg)\\ \leqslant&\exp\left(-\frac{n\epsilon^2}{4}\right).\end{aligned}$$

Setting $\epsilon=4\sqrt{\frac{\log(1/\Delta)}{n}}$, then with probability at least $1-\Delta$,

$$\begin{aligned}&\sup_{\theta\in\Theta}\big|F_4(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}\big[f_4(S;\theta)\big]\big|-\mathbb{E}\big[\sup_{\theta\in\Theta}\big|F_4(s_1,\cdots,s_n;\theta)-\mathbb{E}_{P_S}[f_4(S;\theta)]\big|\big]\\ \leqslant&4\sqrt{\frac{\log(1/\Delta)}{n}}\end{aligned}$$

Putting everything together and taking the maximum over $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta}|\mathcal{R}_4(\theta)| \leqslant \bar{\lambda}e^{1/\underline{\lambda}}\bigg(\frac{5BU}{\underline{\lambda}}\sqrt{\frac{d\log 2}{n}} + \frac{1+8\sqrt{\log(1/\Delta)}}{2\sqrt{n}}\bigg).$$

□

Lemma 13. *With probability at least* $1-\Delta$*,*

$$\sup_{\theta\in\Theta}|\mathcal{R}_5(\theta)| \leqslant \bar{\lambda}e^{1/\underline{\lambda}}\bigg(\frac{U^2}{2\underline{\lambda}}\|(\Sigma_{\mathcal{D}}+\rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+\log(2/\Delta)}{\gamma^2 n}+\rho B^2} + \frac{20BU}{\underline{\lambda}}\sqrt{\frac{d\log 2}{n}} + \frac{2}{\sqrt{n}} + 8\sqrt{\frac{\log(2/\Delta)}{n}}\bigg).$$

Proof. First decompose

$$|\mathcal{R}_5(\theta)| \leqslant \frac{\bar{\lambda}}{e^{-1/\bar{\lambda}}}\max_{\lambda\in[\underline{\lambda},\bar{\lambda}]}\bigg|\frac{1}{n}\sum_{i=1}^n \exp\bigg(-\frac{Y_\theta(s_i;\phi^*)}{\lambda}\bigg) - \exp\bigg(-\frac{\hat{Y}_\theta(s_i;\hat{\phi})}{\lambda}\bigg)\bigg|.$$

Let $f_5(s;\theta) = \exp(-Y_\theta(s;\phi^*)/\lambda) - \exp(-\hat{Y}_\theta(s;\hat{\phi})/\lambda)$ and $F_5(s_1,\cdots,s_n;\theta) = \frac{1}{n}\sum_{i=1}^n f_5(s_i;\theta)$. We proceed to bound the following two terms:

$$\sup_{\theta\in\Theta}\bigg|\frac{1}{n}\sum_{i=1}^n f_5(s_i;\theta)\bigg| = \sup_{\theta\in\Theta}\big|\mathbb{E}_{P_S}[f_5(S;\theta)]\big| + \sup_{\theta\in\Theta}\bigg|\frac{1}{n}\sum_{i=1}^n \bar{f}_5(s_i;\theta)\bigg|,$$

where $\bar{f}_5(s;\theta) = f_5(s;\theta) - \mathbb{E}_P[f_5(S;\theta)]$.

Recall that, fix $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, the function $f_5(s;\theta)$ is $\frac{U}{\underline{\lambda}}$-Lipschitz. Hence the expectation

$$\begin{aligned}
&\big|\mathbb{E}_P[f_5(S;\theta)]\big| \\
\leqslant&\frac{U}{\underline{\lambda}}\bigg|\int_{\mathcal{S}}\sum_{a,a'\in\mathcal{A}}\pi_\theta(a\mid s)\pi_{\theta_0}(a'\mid s)\Big(\sigma\big(\langle\phi^*,\psi(s,a)-\psi(s,a')\rangle\big)-\sigma\big(\langle\hat{\phi},\psi(s,a)-\psi(s,a')\rangle\big)\Big)\,dP_S\bigg| \\
\leqslant&\frac{U}{4\underline{\lambda}}\int_{\mathcal{S}}\sum_{a,a'\in\mathcal{A}}\pi_\theta(a\mid s)\pi_{\theta_0}(a'\mid s)\|\phi^*-\hat{\phi}\|\|\psi(s,a)-\psi(s,a')\|\,dP_S,
\end{aligned}$$

where the second inequality used the fact that the sigmoid function is $\frac{1}{4}$-Lipschitz. Taking supremum of both sides,

$$\begin{aligned}
\sup_{\theta\in\Theta}\big|\mathbb{E}_P[f_5(S;\theta)]\big| \leqslant&\frac{U}{2\underline{\lambda}}\int_{\mathcal{S}}\sum_{a'\in\mathcal{A}}\pi_{\theta_0}(a'\mid s)\|\phi^*-\hat{\phi}\|\max_a\|\psi(s,a)\|\,dP_S \\
\leqslant&\frac{U}{2\underline{\lambda}}\max_{(s,a)}\|\psi(s,a)\|\cdot\|\phi^*-\hat{\phi}\| \\
\leqslant&\frac{U^2}{2\underline{\lambda}}\|(\Sigma_{\mathcal{D}}+\rho I)^{-\frac{1}{2}}\|\sqrt{\frac{d+\log(1/\Delta)}{\gamma^2 n}+\rho B^2}.
\end{aligned}$$

Now we now on to the second term. Recall that $\bar{f}_5(s;\theta)$ is $\frac{2U}{\underline{\lambda}}$-Lipschitz, and it is upper and lower bounded by $0 \leqslant \bar{f}_5 \leqslant 4$. Applying Lemma 6,

$$\mathbb{E}\bigg[\sup_{\theta\in\Theta}\bigg|\frac{1}{n}\sum_{i=1}^n \bar{f}_5(s_i;\theta)\bigg|\bigg] \leqslant \frac{20BU}{\lambda}\sqrt{\frac{d\log 2}{n}} + \frac{2}{\sqrt{n}}.$$

By the boundedness of $\bar{f}_5$, directly invoking the bounded difference inequality, for any $\epsilon > 0$,

$$\mathbb{P}\bigg(\sup_{\theta\in\Theta}\bigg|\frac{1}{n}\sum_{i=1}^n \bar{f}_5(s_i;\theta) - \mathbb{E}\big[\sup_{\theta\in\Theta}\big|\frac{1}{n}\sum_{i=1}^n \bar{f}_5(s_i;\theta)\big|\big] \geqslant \epsilon\bigg) \leqslant \exp\bigg(-\frac{n\epsilon^2}{8}\bigg).$$

Setting $\epsilon = 8\sqrt{\frac{\log(1/\Delta)}{n}}$, then with probability at least $1-\Delta$,

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^{n}\bar{f}_5(s_i;\theta) - \mathbb{E}\big[\sup_{\theta\in\Theta}\big|\frac{1}{n}\sum_{i=1}^{n}\bar{f}_5(s_i;\theta)\big|\big] \leqslant 8\sqrt{\frac{\log(1/\Delta)}{n}}.$$

Putting everything together and taking the maximum over $\lambda \in [\underline{\lambda}, \bar{\lambda}]$ concludes the proof. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. The proof is completed by combining the results of Lemma 12, 8 and 13. □

H Proof of helper Lemmas

H.1 Proof of Lemma 6

Proof of Lemma 6. Let $\epsilon_k = B2^{-k}$ for $k = 0, 1, 2, \cdots$, and for $k \geqslant 1$, choose a deterministic ϵ_k-net $\mathcal{N}_k := \mathcal{N}(\epsilon_k)$ of Θ, where we set $\mathcal{N}_0 = \{0\}$ without loss of generality. The Euclidean covering bound gives

$$|\mathcal{N}_k| \leqslant \left(1 + \frac{2B}{\epsilon_k}\right)^d = (1+2^{k+1})^d \leqslant 2^{d(k+2)}.$$

For every θ. choose $p_k(\theta) \in \mathcal{N}_k$, which by definition, $\|\theta - p_k(\theta)\| \leqslant \epsilon_k$. Then

$$\|p_k(\theta) - p_{k-1}(\theta)\| \leqslant \|p_k(\theta) - \theta\| + \|\theta - p_{k-1}(\theta)\| \leqslant \epsilon_k + \epsilon_{k-1} = 3B2^{-k}.$$

Now, define $D_\theta := \frac{1}{n}\sum_{i=1}^{n} f_\theta(Z_i) - \mathbb{E}[f_\theta(Z)]$. For a finite K, there is

$$D_\theta - D_0 = (D_\theta - D_{p_K(\theta)}) + \sum_{k=1}^{K}(D_{p_k(\theta)} - D_{p_{k-1}(\theta)}), \tag{12}$$

and

$$|D_\theta - D_{p_K(\theta)}| = \left|\frac{1}{n}\sum_{i=1}^{n} f_\theta(Z_i) - \mathbb{E}[f_\theta(Z)] - f_{p_K(\theta)}(Z_i) + \mathbb{E}[f_{p_K(\theta)}(Z)]\right| \leqslant 2L\|\theta - p_K(\theta)\| \leqslant 2L\epsilon_K.$$

To control the increments, define the set of possible nearby pairs at level k: $\mathcal{E}_k = \{(u,v) \in \mathcal{N}_k \times \mathcal{N}_{k-1} : \|u - v\|_2 \leqslant 3B2^{-k}\}$ and let $M_k = \max_{(u,v)\in\mathcal{E}_k} |D_u - D_v|$. Note that, by definition every pair $(p_k(\theta), p_{k-1}(\theta))$ belongs to $\mathcal{E}_k$. Hence Equation (12) implies that

$$\sup_{\theta\in\Theta}|D_\theta - D_0| \leqslant 2LB2^{-K} + \sum_{k=1}^{K} M_k. \tag{13}$$

We proceed to bound the expectation of M_k at each level k.

Fix any $\theta, \theta' \in \Theta$, we can bound

$$-L\|\theta - \theta'\| \leqslant f_\theta(z) - f_{\theta'}(z) \leqslant L\|\theta - \theta'\|,$$

by the fact that $f_\theta(z)$ is L-Lipschitz in θ. Recall that,

$$D_\theta - D_{\theta'} = \frac{1}{n}\sum_{i=1}^{n} f_\theta(z) - f_{\theta'}(z) - \mathbb{E}[f_\theta(Z) - f_{\theta'}(Z)].$$

By Hoeffding's lemma,

$$\mathbb{E}\left[\exp\left(\frac{t}{n}\Big(\big(f_\theta(z_i) - f_{\theta'}(z_i) - \mathbb{E}[f_\theta(Z) - f_{\theta'}(Z)]\big)\Big)\right)\right] \leqslant \exp\left(\frac{t^2L^2\|\theta - \theta'\|^2}{2n^2}\right).$$

Since the samples $Z_1, Z_2, \cdots$, are independent of each other,

$$\mathbb{E}\big[\exp(t(D_\theta - D_{\theta'}))\big] = \prod_{i=1}^{n} \mathbb{E}\bigg[\exp\bigg(\frac{t}{n}\Big(\big(f_\theta(z_i) - f_{\theta'}(z_i) - \mathbb{E}[f_\theta(Z) - f_{\theta'}(Z)]\big)\Big)\bigg] \leqslant \exp\bigg(\frac{t^2 L^2 \|\theta - \theta'\|^2}{2n}\bigg).$$

We conclude that

$$\mathbb{E}\big[\exp(t(D_\theta - D_{\theta'}))\big] \leqslant \exp\bigg(\frac{t^2 \tau_n^2(\theta, \theta')}{2}\bigg), \quad \tau_n(\theta, \theta') = \frac{L}{\sqrt{n}}\|\theta - \theta'\|,$$

implying that $D_\theta - D_{\theta'}$ is $\tau_n(\theta, \theta')$-sub-Gaussian. By the definition of $\mathcal{E}_k$ and M_k, at each level k, any $D_\theta - D_{\theta'}$ induced by $(\theta, \theta') in \mathcal{E}_k$ is sub-Gaussian with parameter at most $\sigma_k = \frac{3LB2^{-k}}{\sqrt{n}}$.

To proceed, we need the following result.

Lemma 14. *Suppose $Z_1, \cdots, Z_m$ are σ-sub-Gaussian, i.e., $\mathbb{E}e^{tZ_j} \leqslant e^{t^2\sigma^2/2}$ for any $t \in \mathbb{R}$. Then*

$$\mathbb{E}\Big[\max_{1 \leqslant j \leqslant m} |Z_j|\Big] \leqslant \sigma\sqrt{2\log(2m)}.$$

Proof of Lemma 14. For any $t > 0$, according to Jensen's inequality,

$$\exp\bigg(t\mathbb{E}\Big[\max_{1 \leqslant j \leqslant m} |Z_j|\Big]\bigg) \leqslant \mathbb{E}\Big[\exp\big(t \max_{1 \leqslant j \leqslant m} |Z_j|\big)\Big] \leqslant \sum_{j=1}^{m} \mathbb{E}\big(e^{tZ_j} + e^{-tZ_j}\big) \leqslant 2me^{t^2\sigma^2/2}.$$

Taking logritham of both sides gives

$$\mathbb{E}\Big[\max_{1 \leqslant j \leqslant m} |Z_j|\Big] \leqslant \frac{\log 2m}{t} + \frac{t\sigma^2}{2}.$$

Setting $t = \sqrt{\frac{2\log 2m}{\sigma^2}}$ concludes the proof. □

Applying Lemma 14,

$$\mathbb{E}[M_k] = \mathbb{E}\Big[\max_{u,v \in \mathcal{E}_k} |D_u - D_v|\Big] \leqslant 3LB2^{-k}\sqrt{\frac{2\log(2|\mathcal{E}_k|)}{n}}. \tag{14}$$

Now, since $|\mathcal{E}_k| \leqslant |\mathcal{N}_k||\mathcal{N}_{k-1}|$, we can bound

$$\log(2|\mathcal{E}_k|) \leqslant \log(2|\mathcal{N}_k||\mathcal{N}_{k-1}|) = \log(2 \cdot 2^{d(k+2)} \cdot 2^{d(k+1)}) = (1 + d(2k+3))\log 2 \leqslant 2d(k+2)\log 2.$$

Plug the above result to Equation (14),

$$\mathbb{E}[M_k] \leqslant 6LB2^{-k}\sqrt{\frac{d(k+2)\log 2}{n}}$$

We are now ready to bound $\mathbb{E}\sup_{\theta \in \Theta} |D_\theta - D_0|$ in Equation (13). Taking expectation of its LHS and taking $K \to \infty$,

$$\mathbb{E}\Big[\sup_{\theta \in \Theta} |D_\theta - D_0|\Big] \leqslant \sum_{k=1}^{\infty} \mathbb{E}[M_k] \leqslant 6LB \sum_{k=1}^{\infty} 2^{-k}\sqrt{\frac{d(k+2)\log 2}{n}}.$$

Since $\sum_{k=1}^{\infty} 2^{-k} = 1$ and $\sum_{k=1}^{\infty} 2^{-k}(k+2) = 4$, Cauchy-Schewarz gives

$$\sum_{k=1}^{\infty} 2^{-k}\sqrt{k+2} \leqslant \sqrt{\sum_{k=1}^{\infty} 2^{-k}(k+2)} = 2,$$

therefore,

$$\mathbb{E}\Big[\sup_{\theta\in\Theta}|D_\theta - D_0|\Big] \leqslant 12LB\sqrt{\frac{d\log 2}{n}} \leqslant 10LB\sqrt{\frac{d\log 2}{n}}.$$

Finally, observe that $\sup_{\theta\in\Theta}|D_\theta| \leqslant |D_0| + \sup_{\theta\in\Theta}|D_\theta - D_0|$, and it suffices to bound $|D_0|$ to conclude the proof. Since f_0 is bounded a.s., the variance $\mathrm{Var}(f_0(Z)) \leqslant (\bar{f} - \underline{f})^2/4$, and

$$\mathbb{E}|D_0| \leqslant \sqrt{\mathbb{E}D_0^2} = \sqrt{\frac{\mathrm{Var}(f_0(Z))}{n}} \leqslant \frac{\bar{f} - \underline{f}}{2\sqrt{n}}.$$

Combining everything, we have that

$$\mathbb{E}\Big[\sup_{\theta\in\Theta}|D_\theta|\Big] \leqslant 10LB\sqrt{\frac{d\log 2}{n}} + \frac{\bar{f} - \underline{f}}{2\sqrt{n}}.$$

□

H.2 Proof of Lemma 9

Proof of Lemma 9. Define

$$\chi_s(\theta) := \log\sum_{a\in\mathcal{A}}\exp(\theta^\top\psi(s,a)),\ \nabla_\theta\chi_s(\theta) = \sum_{a\in\mathcal{A}}\frac{\exp(\theta^\top\psi(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\theta^\top\psi(s,a'))}\psi(s,a) = \sum_{a\in\mathcal{A}}\pi_\theta(a\mid s)\psi(s,a),$$

we can write the KL divergence

$$D_{\mathrm{KL}}\big(\pi_\theta(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big) = \chi_s(\theta) - \chi_s(\theta_0) - (\theta-\theta_0)^\top\nabla_\theta\chi_s(\theta_0) = D_{\chi_s}(\theta,\theta_0),$$

where D_{χ_s} is the Bregman divergence under χ_s. Given any $s\in\mathcal{S}$, the function χ_s is twice-differentiable with the Hessian

$$\nabla^2_\theta\chi_s(\theta) = \mathrm{Cov}_{a\sim\pi_\theta(\cdot|s)}[\psi(s,a)] \preceq U^2 I.$$

Hence, by Taylor's Theorem with an integral remainder, there is

$$\begin{aligned}
\chi_s(\theta) &= \chi_s(\theta_0) + \nabla_\theta\chi_s(\theta_0)^\top(\theta-\theta_0) + \int_0^1 (1-t)(\theta-\theta_0)^\top\nabla^2_\theta\chi_s(\theta_0 + t(\theta-\theta_0))(\theta-\theta_0)\,dt,\\
\chi_s(\theta) - \chi_s(\theta_0) - \nabla_\theta\chi_s(\theta_0)^\top(\theta-\theta_0) &= \int_0^1 (1-t)(\theta-\theta_0)^\top\nabla^2_\theta\chi_s(\theta_0 + t(\theta-\theta_0))(\theta-\theta_0)\,dt\\
&\leqslant \int_0^1 (1-t)U^2\|\theta-\theta_0\|^2\,dt,
\end{aligned}$$

and we conclude

$$0 \leqslant D_{\mathrm{KL}}\big(\pi_\theta(\cdot\mid s)\|\pi_{\theta_0}(\cdot\mid s)\big) \leqslant \frac{U^2}{2}\|\theta-\theta_0\|^2 = 2B^2U^2.$$

□

H.3 Proof of Lemma 10

Proof of Lemma 10. For every $s\in\mathcal{S}$ and $a, a'\in\mathcal{A}$,

$$\log\pi_{\theta_0}(a\mid s) - \log\pi_{\theta_0}(a'\mid s) = \langle\theta_0, \psi(s,a) - \psi(s,a')\rangle \leqslant \|\theta_0\|\|\psi(s,a) - \psi(s,a')\| \leqslant 2BU.$$

Define

$$\Gamma_0 := \sup_{s\in\mathcal{S}}\left\{\max_{a\in\mathcal{A}}\log\pi_{\theta_0}(a\mid s) - \min_{a\in\mathcal{A}}\log\pi_{\theta_0}(a\mid s)\right\}.$$

It follows that $\Gamma_0 \leqslant 2BU$.

We now prove the Lipschitz bound. Fix an arbitrary $s \in \mathcal{S}$. For any policy $\pi_\theta \in \Pi_\Theta$, direct computation gives

$$\begin{aligned}\nabla_\theta \pi_\theta(a \mid s) =& \pi_\theta(a \mid s)(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}[\psi(s,a)]) \\ \nabla_\theta D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) =& \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a \mid s)\left(\log \frac{\pi_\theta(a \mid s)}{\pi_{\theta_0}(a \mid s)} + 1\right) \\ =& \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s)\Big(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}\big[\psi(s,a)\big]\Big) \log\left(\frac{\pi_\theta(a \mid s)}{\pi_{\theta_0}(a \mid s)}\right)\end{aligned} \tag{15}$$

where the last equality is due to the fact that $\sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a \mid s) = \nabla_\theta \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s) = 0$, and the term involving 1 vanishes. Since $\log \pi_\theta(a \mid s) = \langle \theta, \psi(s,a)\rangle - \log(\sum_{b\in\mathcal{A}} \exp(\langle \theta, \psi(s,b)\rangle))$,

$$\begin{aligned}(15) &= \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s)\Big(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}\big[\psi(s,a)\big]\Big)\big(\log \pi_\theta(a \mid s) - \log \pi_{\theta_0}(a \mid s)\big) \\ &= \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s)\Big(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}\big[\psi(s,a)\big]\Big)\big(\langle \theta, \psi(s,a)\rangle - \log \sum_{b\in\mathcal{A}} \exp(\langle \theta, \psi(s,b)\rangle) - \log \pi_{\theta_0}(a \mid s)\big) \\ &= -\log \sum_{b\in\mathcal{A}} \exp(\langle \theta, \psi(s,b)\rangle) \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s)\Big(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}\big[\psi(s,a)\big]\Big) \\ &\quad + \sum_{a\in\mathcal{A}} \pi_\theta(a \mid s)\Big(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}\big[\psi(s,a)\big]\Big)\big(\langle \theta, \psi(s,a)\rangle - \log \pi_{\theta_0}(a \mid s)\big);\end{aligned}$$

and $\mathbb{E}_{a\sim\pi_\theta}[\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}[\psi(s,a)]] = 0$, the log-partition term is independent of a and therefore disappears from (15). It follows that

$$\begin{aligned}\nabla_\theta D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) &= \mathbb{E}_{a\sim\pi_\theta}\big[(\psi(s,a) - \mathbb{E}_{a\sim\pi_\theta}[\psi(s,a)])(\langle \theta, \psi(s,a)\rangle - \log \pi_{\theta_0}(a \mid s))\big] \\ &= \mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \langle \theta, \psi(s,a)\rangle - \log \pi_{\theta_0}(a \mid s)) \\ &= \mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a))^\top \theta - \mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \log \pi_{\theta_0}(a \mid s)).\end{aligned} \tag{16}$$

We proceed to bound the two terms in (16) separately. First, for any unit vector $v \in \mathbb{R}^d$,

$$v^\top \mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a))v = \mathrm{Var}_{a\sim\pi_\theta}(v^\top \psi(s,a)) \leqslant \mathbb{E}_{a\sim\pi_\theta}\big[(v^\top \psi(s,a))^2\big] \leqslant U^2.$$

Therefore $\|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a))\|_{\mathrm{op}} \leqslant U^2$, and by definition of the operation norm,

$$\|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a))^\top \theta\| \leqslant \|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a))\|_{\mathrm{op}} \|\theta\| \leqslant U^2\|\theta\| \leqslant BU^2. \tag{17}$$

For the second term, again let $v \in \mathbb{R}^d$ be any unit vector. By the Cauchy-Schwarz inequality for covariance,

$$\begin{aligned}\big|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \log \pi_{\theta_0}(a \mid s))v\big| =& \big|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a)^\top v, \log \pi_{\theta_0}(a \mid s))\big| \\ \leqslant& \sqrt{\mathrm{Var}_{a\sim\pi_\theta}(\psi(s,a)^\top v)}\sqrt{\mathrm{Var}_{a\sim\pi_\theta}(\log \pi_{\theta_0}(a \mid s))}.\end{aligned}$$

Note that the random variable $\log \pi_{\theta_0}(A \mid s)$ takes values in an interval of length at most Γ_0, hence the variance $\mathrm{Var}_{a\sim\pi_\theta}(\log \pi_{\theta_0}(a \mid s))$ is at most $\frac{\Gamma_0^2}{4}$. For the random variable $\psi(s,A)^\top v$, the same argument follows and the variance $\mathrm{Var}_{a\sim\pi_\theta}(\psi(s,a)^\top v)$ is at most U^2. Consequently,

$$\big|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \log \pi_{\theta_0}(a \mid s))v\big| \leqslant \frac{U\Gamma_0}{2}.$$

Taking the supremum over all unit vectors v gives

$$\big\|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \log \pi_{\theta_0}(a \mid s))\big\| = \sup_{\|v\|=1} \big\|\mathrm{Cov}_{a\sim\pi_\theta}(\psi(s,a), \log \pi_{\theta_0}(a \mid s))v\big\| \leqslant \frac{U\Gamma_0}{2}. \tag{18}$$

Combining (16), (17), and (18), we conclude that

$$\|\nabla_\theta D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big)\| \leqslant BU^2 + \frac{U\Gamma_0}{2}.$$

Now for θ, θ', and define the line segment $\theta_t := \theta' + t(\theta - \theta')$ for $t \in [0, 1]$. Since the Euclidean ball is convex, we have that $\|\theta_t\| \leqslant B$. By the fundamental theorem of calculus,

$$\begin{aligned} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) =& \frac{\partial}{\partial t} \int_0^1 D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big)\, dt \\ =& \int_0^1 \frac{\partial}{\partial \theta} D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) \cdot \frac{\partial \theta}{\partial t} D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big)\, dt \\ =& \int_0^1 \big\langle \nabla_\theta D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big), \theta \big\rangle\, dt, \end{aligned}$$

where the second equality is due to chain rule. Hence

$$\begin{aligned} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) - D_{\mathrm{KL}}\big(\pi_{\theta'}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) &= \int_0^1 \big\langle \nabla_\theta D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big), \theta - \theta' \big\rangle\, dt \\ \big|D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) - D_{\mathrm{KL}}\big(\pi_{\theta'}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big)\big| &\leqslant \int_0^1 \big\| \nabla_\theta D_{\mathrm{KL}}\big(\pi_{\theta_t}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) \big\| \|\theta - \theta'\|\, dt \\ &\leqslant \left(BU^2 + \frac{U\Gamma_0}{2} \right) \|\theta - \theta'\|. \end{aligned}$$

Since the bound holds for every $s \in \mathcal{S}$, taking the supremum over s gives

$$\sup_{s \in \mathcal{S}} \big|D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big) - D_{\mathrm{KL}}\big(\pi_{\theta'}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s)\big)\big| \leqslant 2BU^2 \|\theta - \theta'\|,$$

due to the fact that $\Gamma_0 \leqslant 2BU$. □