Policy Teaching via Data Poisoning in Learning from Human Preferences

Andi Nika
MPI-SWS

Jonathan Nöther
MPI-SWS

Debmalya Mandal
University of Warwick

Parameswaran Kamalaruban
Featurespace

Adish Singla
MPI-SWS

Goran Radanović
MPI-SWS

Abstract

We study data poisoning attacks in learning from human preferences. More specifically, we consider the problem of teaching/enforcing a target policy $\pi^\dagger$ by synthesizing preference data. We seek to understand the susceptibility of different preference-based learning paradigms to poisoned preference data by analyzing the number of samples required by the attacker to enforce $\pi^\dagger$. We first propose a general data poisoning formulation in learning from human preferences and then study it for two popular paradigms, namely: (a) reinforcement learning from human feedback (RLHF) that operates by learning a reward model using preferences; (b) direct preference optimization (DPO) that directly optimizes policy using preferences. We conduct a theoretical analysis of the effectiveness of data poisoning in a setting where the attacker is allowed to augment a pre-existing dataset and also study its special case where the attacker can synthesize the entire preference dataset from scratch. As our main results, we provide lower/upper bounds on the number of samples required to enforce $\pi^\dagger$. Finally, we discuss the implications of our results in terms of the susceptibility of these learning paradigms under such data poisoning attacks.

1 Introduction

Learning from human preferences has recently attracted considerable attention, largely due to its effectiveness in fine-tuning large language models (LLMs). Unlike the traditional approach, which uses training data labeled with absolute scores, this method relies on pairs of examples marked with binary signals indicating preference—essentially assigning a relative score to each example. As such, it has proven to be practically beneficial since comparative feedback between two examples is more easily accessible than their individual absolute scores.

Despite its practical advantages, learning from human preferences is susceptible to data poisoning attacks [Biggio et al., 2012], due to the potential presence of malicious human feedback in the training data. Malignant third parties could easily alter preference datasets to steer LLMs toward generating biased or harmful content which can lead to undesired model behaviours. This vulnerability is particularly concerning, given the rapid integration of LLMs into critical applications. It is thus essential to understand these attacks in order to design models with robust guarantees against them.

These concerns have motivated a lot of recent work, all of which has focused on empirical investigations of poisoning attacks. For example, Wang et al. [2023b] demonstrated the effectiveness of ranking poisoning attacks, where attackers manipulate preference labels without altering the underlying data. Shi et al. [2023] showed how an attacker could inject trigger words into training prompts, influencing the LLM's sentiment analysis. Rando and Tramèr [2023] proposed universal backdoor attacks, embedding hidden functionalities within LLMs, and Baumgärtner et al. [2024] investigated data augmentation attacks that inject entirely new preference pairs.

Despite these attempts, a strong theoretical foundation for understanding the robustness of learning from human preferences against data poisoning attacks remains elusive. Motivated by this, we initiate a theoretical study of data poisoning attacks in learning from human preferences. We seek to analyze these attacks from the

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

attacker's viewpoint which, in turn, would allow us to identify more robust settings and design effective defenses, ultimately ensuring the security and reliability of preference-based learning in real-world applications.

In particular, we focus on two of the most prominent techniques of learning from human preferences, namely, reinforcement learning from human feedback (RLHF) [Stiennon et al., 2020, Ouyang et al., 2022, Ziegler et al., 2019, Gao et al., 2023, Menick et al., 2022, Glaese et al., 2022, Bai et al., 2022, Brown et al., 2019, Shin et al., 2023] and direct preference optimization (DPO) [Rafailov et al., 2023]. We consider an attacker that aims to enforce a target policy $\pi^\dagger$ by generating new preference data $\widehat{D}$. The aim of the attacker is to ensure an RLHF/DPO learner trained on $\widehat{D}$ converges to a policy that is ϵ-close to $\pi^\dagger$. We analyze the number of samples the attacker requires to enforce $\pi^\dagger$ in settings where the attacker can synthesize the entire preference dataset from scratch and where it has to augment a pre-existing dataset $\overline{D}$. Our contributions are summarized below:

- **Attack problem formulation:** We propose a general data poisoning formulation in learning from human preferences and instantiate it for two popular paradigms of RLHF and DPO. In particular, we propose an attack problem formulation using the ℓ_1-norm as a constraint, motivated by previous formulations suited for reward-based RL with deterministic target policies. We consider two types of RLHF learners: unregularized RLHF and regularized RLHF, depending on whether the learner is restricted to remain close to a given reference policy μ or not.
- **Attacks on RLHF:** We analyze the sample complexity of the attack in unregularized and regularized RLHF settings. In the unregularized setting, our bounds depend on the state-action space cardinality, the attack granularity parameter ϵ, the covariance matrix of the pre-existing data, and its size $\overline{n}$. In the regularized RLHF setting, the dependence on the state-action space size is replaced with dependence on the regularization temperature β and the gap between $\pi^\dagger$ and μ.
- **Attacks on DPO:** Moreover, we provide lower and upper bounds on the attack sample complexity in the DPO setting, by showing that feasible regions for the surrogate problem act both as relaxations and restrictions of the original feasible region, for different values of attack granularity parameter. In this setting, the sample complexity depends on the squared norms of the parameters of $\pi^\dagger$ and μ and the pre-existing dataset size $\overline{n}$.
- **Comparison**: Finally, we derive conclusions on the susceptibility of DPO to attack relative to RLHF, both in the data augmentation and data synthesis setting. Our results suggest that, the farther away the target policy $\pi^\dagger$ is from the reference policy μ in the parameter space, the stronger the tendency of DPO to remain closer to μ under attacks, relative to RLHF.

2 Preliminaries and Background on Learning from Human Preferences

In this section, we provide the necessary technical background that will be used throughout the paper.

2.1 Preliminaries

Environment. Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma, \rho \rangle$ be an infinite-horizon discounted Markov decision process (MDP), where $\mathcal{S}$ denotes the state space and $\mathcal{A}$ denotes the action space, with cardinalities S and A, respectively; $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes the transition function, where $P(s, a, s')$ denotes the probability of transitioning to state s' when taking action a in state s; the reward is denoted by $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the discount factor by $\gamma \in [0, 1)$. Finally, we let ρ be the initial state distribution. A contextual bandit can be viewed as a special case of this MDP formalism where the state transitions are independent of the actions taken, and the discount factor γ is set to 0.

Policies and value functions. Stochastic policies are mappings from states to action simplices, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the probability simplex with support in $\mathcal{A}$. A deterministic policy is a mapping $\pi : \mathcal{S} \to \mathcal{A}$, a special case of stochastic policies. Let Π and Π^{det} denote the set of all stochastic and deterministic policies defined over $\mathcal{S}$ and $\mathcal{A}$, respectively. For a policy π and state-action pair (s, a), we define $d^\pi_{s,a}(s', a') = (1 - \gamma) \sum_{t=1}^\infty \gamma^t \mathbb{P}(s_t = s', a_t = a' | s_0 = s, a_0 = a, \pi)$, and $d^\pi_s(s', a') = (1-\gamma) \sum_{t=1}^\infty \gamma^t \mathbb{P}(s_t = s', a_t = a' | s_0 = s, \pi)$. Furthermore, we define $d^\pi_{s,a}(s') = \sum_{a'} d^\pi_{s,a}(s', a')\pi(a'|s')$, $d^\pi_s(s') = \sum_{a'} d^\pi_s(s', a')\pi(a'|s')$, and $d^\pi_\rho(s', a') = \mathbb{E}_{s\sim\rho}[d^\pi_s(s', a')]$. We focus on *ergodic* MDPs where every state is reachable under any policy and initial distribution. The value function of a policy $\pi \in \Pi$ with respect to a given reward function r is given as

$$V^\pi_r(s) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) | s_0 = s\right] .$$

Similarly, its action-value function is given as

$$Q^\pi_r(s, a) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right] .$$

We also define $V_r^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V_r^\pi(s)]$. A policy $\pi^\star \in \Pi$ is said to be optimal in Π with respect to r if $V_r^{\pi^\star}(\rho) \geq V_r^\pi(\rho)$, for all $\pi \in \Pi$. Moreover, $\pi^\star$ is said to be ϵ-robust optimal in Π with respect to r, for a given $\epsilon > 0$, if $V_r^{\pi^\star}(\rho) \geq V_r^\pi(\rho) + \epsilon$, for all $\pi \in \Pi \setminus \{\pi^\star\}$. We define the KL-divergence between two policies π and π' as $D_{\mathrm{KL}}(\pi||\pi') = \sum_{s,a} \rho(s)\pi(a|s)\left(\log \pi(s|a) - \log \pi'(a|s)\right)$. When $\pi, \pi' \in \Pi^{\mathrm{det}}$, we define $D_{\mathrm{KL}}(\pi||\pi') = 0$ iff $\pi = \pi'$, and ∞ otherwise. Later in our analysis, we consider the following classes.

Definition 2.1 (*Linear rewards*)**.** Let ϕ be a d-dimensional feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with $\max_{s,a} \|\phi(s,a)\|_2 \leq 1$. We consider the following class of linear reward functions:

$$\mathcal{R} = \{r_\omega : r_\omega(s,a) = \omega^\top \phi(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A} \text{ for } \omega \in \mathbb{R}^d\} .$$

Definition 2.2 (*Loglinear policies*)**.** Let ψ be a d'-dimensional feature mapping $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d'}$ with $\max_{s,a} \|\psi(s,a)\| \leq 1$. We consider the following class of loglinear policies:

$$\Pi^{\log} = \left\{ \pi_\theta : \pi_\theta(a|s) = \frac{\exp\left(\psi(s,a)^\top \theta\right)}{\sum_{a'} \exp\left(\psi(s,a')^\top \theta\right)}, \right. \\ \left. \forall (s,a) \in \mathcal{S} \times \mathcal{A} \text{ where } \theta \in \mathbb{R}^{d'} \right\} .$$

By overloading the notation, for a trajectory $\tau = (s_0, a_0, s_1, \ldots)$, we define $\phi(\tau) = \sum_{t=0}^\infty \gamma^t \phi(s_t, a_t)$ and $\psi(\tau) = \sum_{t=0}^\infty \gamma^t \psi(s_t, a_t)$. Furthermore, for a policy π and state-action pair (s,a), we define $\phi^\pi(s,a) = \sum_{s',a'} d^\pi_{s,a}(s',a')\phi(s',a')$ and $\psi^\pi(s,a) = \sum_{s',a'} d^\pi_{s,a}(s',a')\psi(s',a')$.

2.2 Background on Learning from Human Preferences

In learning from human preferences, a learner refines behavior by iteratively learning from human feedback. Given a preference dataset $D = \{(\tau, \tau', o)\}$, where $o = 1$ indicates that trajectory τ is preferred over τ', and $o = -1$ indicates the opposite, the learner L outputs a policy $\pi = L(D, \mu)$ that aligns better with human preferences. Here, μ represents a reference policy, which could be a pre-trained model. Examples of such learners include agents that use reinforcement learning from human feedback (RLHF) L_{RLHF} [Ziegler et al., 2019] or direct preference optimization (DPO) L_{DPO} [Rafailov et al., 2023]. Before introducing these methods, we first define the preference model.

Definition 2.3 (*Bradley-Terry preference model* [Bradley and Terry, 1952])**.** The Bradley-Terry preference model w.r.t. a reward function r is defined as follows: for every tuple (τ, τ', o), we have $\mathbb{P}(o = 1|\tau, \tau') = \sigma\left(\sum_{t=0}^\infty \gamma^t r(s_t, a_t) - \sum_{t=0}^\infty \gamma^t r(s'_t, a'_t)\right)$, where $\sigma(z) = 1/(1 + \exp(-z))$.

Reinforcement learning from human feedback. With access to the dataset D and reference policy μ, RLHF [Ziegler et al., 2019] proceeds in two phases. In the first phase, a reward function is learned from D using maximum likelihood estimation (MLE) based on the Bradley-Terry preference model. This involves solving the following regularized MLE problem (with regularization parameter $\lambda > 0$):

$$\min_\omega \ell^\omega_{\mathrm{RLHF}}(D) := -\sum_{(\tau,\tau',o)\in D} \log \\ \sigma\left(o \cdot \sum_{t\geq 0} \gamma^t \left(r_\omega(s_t, a_t) - r_\omega(s'_t, a'_t)\right)\right) + \frac{\lambda}{2}\|\omega\|^2 . \tag{P:RLHF.Reward}$$

Let $\widehat{\omega}$ denote the solution of Problem Equation P:RLHF.Reward. In the second phase, RLHF solves the following regularized policy optimization based on the learned reward $r_{\widehat{\omega}}$ (with regularization parameter $\beta > 0$) to obtain optimal solution $\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}$ in:

$$\arg\max_{\pi\in\Pi} \mathbb{E}_{\pi,P}\left[\sum_{t\geq 0} \gamma^t \left(r_{\widehat{\omega}}(s_t, a_t) - \beta \log \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right)\right] . \tag{P:RLHF.Policy}$$

We denote by $\mathcal{V}^\pi_{r_{\widehat{\omega}}}(\rho)$ the regularized objective above.

Direct preference optimization. DPO [Rafailov et al., 2023] leverages the relationship between a reward function r and its corresponding regularized optimal policy π^{reg}_r to bypass the reward-learning phase, and directly optimize the policy. However, DPO is currently limited to contextual bandit settings, where we have $\tau = (s,a)$ and $\tau' = (s,a')$. Given D and μ, DPO solves the following optimization problem (with regularization parameter $\lambda > 0$) to obtain optimal solution $\widehat{\theta}$ as:

$$\min_\theta \ell^\theta_{\mathrm{DPO}}(D) := -\sum_{(\tau,\tau',o)\in D} \log \\ \sigma\left(o \cdot \left(\beta \log \frac{\pi_\theta(a|s)}{\mu(a|s)} - \beta \log \frac{\pi_\theta(a'|s)}{\mu(a'|s)}\right)\right) + \frac{\lambda}{2}\|\theta - \theta_\mu\|^2 . \tag{P:DPO}$$

3 Preference Poisoning Attack Setup

In this section, we formulate the problem of data poisoning attacks on learning from human preferences. We consider an attacker aiming to impose a target policy $\pi^\dagger$ onto a learner. To achieve this, the attacker modifies the preference dataset so that the learner trained on the altered data produces a policy close to $\pi^\dagger$. Ideally, the attacker should make minimal modifications to achieve this outcome.

Attack goal and poisoned data. Let $\overline{D}$ represent the clean preference dataset originating from the environment $\mathcal{M}$, and let μ denote the reference policy. The learner, denoted by L, is trained on the preference dataset. The attacker is modeled as a mapping $\mathcal{P}$, which takes the clean dataset $\overline{D}$, environment $\mathcal{M}$, reference policy μ, and learner L as inputs, and outputs a poisoned dataset $\widehat{D} = \mathcal{P}(\overline{D}, \mu, \mathcal{M}, L)$. We assume the attacker has full knowledge of the environment. Given a margin parameter ϵ, the attacker poisons the dataset $\overline{D}$ so that the learner L trained on the poisoned dataset $\widehat{D}$ converges to a policy within an ϵ distance of $\pi^\dagger$, i.e., $\left\|\pi^\dagger - \pi^L\right\|_1 \leq \epsilon$, where $\pi^L = L(\widehat{D}, \mu)$ and $\|\pi - \pi'\|_1 = \sum_{s,a} \rho(s) |\pi(a|s) - \pi'(a|a)|$ is the ℓ_1 norm between given policies π and π'.

Attack cost and formulation. We consider a setting where the attacker can *augment* a pre-existing preference dataset $\overline{D}$. Moreover, we instantiate this setting to the case when $\overline{D}$ is empty. Since adding samples incurs a cost, the attacker aims to enforce $\pi^\dagger$ by minimally adding additional samples. The attacker's optimization problem can be formalized as

$$\min_D |D| \text{ such that } \left\|\pi^\dagger - \pi^L\right\|_1 \leq \epsilon \,, \quad \text{where } \pi^L = L(\overline{D} \cup D, \mu) \,. \tag{P:Attack}$$

Attack feasibility and synthesis. We consider poisoning attacks on two learning paradigms in learning from human preferences, RLHF and DPO, as introduced in Section 2.2. We identify the specific conditions within both learning paradigms that make such attacks feasible:

- We consider an attacker that has the synthesis capability to find a trajectory pair (τ, τ') for any given $z \in \mathbb{R}^d$ such that $\phi(\tau) - \phi(\tau') = z$ or $\psi(\tau) - \psi(\tau') = z$. In practice, this assumption translates to the following requirement. Given a prompt-response pair $(x, y) := \tau$ and a vector $z \in \mathbb{R}^d$, the attacker can find a response y' such that the difference between (x, y) and $(x, y') := \tau'$ in the embedding space is approximately equal to z, i.e., $\phi(\tau) - \phi(\tau') \approx z$ or $\psi(\tau) - \psi(\tau') \approx z$.
- For an unregularized RLHF attack to be feasible, we must have $\pi^\dagger \in \Pi^{\text{det}}$. For a regularized RLHF and DPO attack to be feasible, we must have $\pi^\dagger, \mu \in \Pi^{\log}$.

Notation. Next we introduce some notation that will be useful in the following sections. As usual, $[n] = \{1, 2, \ldots, n\}$ denotes the set of first n natural numbers. $\langle v, z \rangle = v^\top z$ denotes the inner product of two compatible vectors v and z. Further, we denote by Φ and Ψ the matrices with columns $\phi(s,a)^\top$ and $\psi(s,a)^\top$, respectively, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$. We assume throughout that Φ and Ψ are full rank. Unless otherwise specified, $\|v\|$ denotes the Euclidean norm, $\|M\|$ denotes the spectral norm for matrix M, and M^+ denotes its pseudoinverse. $\mathbf{0}$ and $\mathbf{1}$ denote the vectors of zeroes and ones, respectively, and I denotes the identity matrix. We denote by $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ the maximum and minimum eigenvalues (or singular values) of a given square (rectangular) matrix M, respectively. Let $\xi_{\max} := \max_x x/(1+\exp(x))$ and denote by $x^\star = \arg\max_x x/(1+\exp(x))$. Let $\xi^{-1}(a)$ denote the solution to $a = x/(1+\exp(x))$, for any $a < \xi_{\max}$, where the domain of $x/(1+\exp(x))$ is $(-\infty, x^\star]$. Moreover, we define $\xi_1(a) = \xi^{-1}(a \lceil a/\xi_{\max} \rceil^{-1})$ and $\xi_2(a) = \xi^{-1}(2a \lceil a/(2\xi_{\max}) \rceil^{-1})$.

4 Poisoning Attacks on RLHF

In this section, we study data poisoning attacks on RLHF, where the attacker synthesizes $\widehat{D}$ from $\overline{D}$. We start our discussion by formulating the general attack problem for RLHF. We do this by instantiating Problem P:Attack for this setting as:

$$\min_D |D| \text{ such that } \widehat{\omega} = \arg\min_\omega \ell^\omega_{\text{RLHF}}\left(\overline{D} \cup D\right) \text{ and } \left\|\pi^\dagger - \pi^{\text{reg}}_{r_{\widehat{\omega}}}\right\|_1^2 \leq \epsilon \,. \tag{P:Attack:RLHF.1}$$

In Appendix A, we show that Problem P:Attack:RLHF.1 is feasible whenever $\pi^\dagger \in \Pi^{\text{det}}$ or $\pi^\dagger, \mu \in \Pi^{\log}$. We study the following scenarios: (i) the general unregularized RLHF ($\beta = 0$) setting with data augmentation ($\overline{D} \neq \emptyset$) and data synthesis ($\overline{D} = \emptyset$), for deterministic $\pi^\dagger$; (ii) the regularized RLHF setting ($\beta > 0$) with data augmentation and data generation for general policies $\pi^\dagger$.

4.1 Unregularized RLHF

For $\beta = 0$, Problem P:RLHF.Policy reduces to a standard value maximization problem, for which optimal policies π^L are known to be deterministic. Thus, if we are given a pre-existing dataset $\overline{D}$ and can augment it into $\widehat{D}$ such that, when used to solve Problem P:RLHF.Reward, yields a reward function making $\pi^\dagger$ ϵ'-robust optimal, for some $\epsilon' > 0$, then π^L is guaranteed to converge to $\pi^\dagger$ as the unique optimal policy. Rakhsha et al. [2021] show that checking the ϵ'-robust optimality condition for neighboring policies $\pi^\dagger\{s,a\}$ of $\pi^\dagger$ for all (s,a) pairs is sufficient. Here, $\pi^\dagger\{s,a\}(s') = \pi^\dagger(s)$, if $s \neq s'$, and $\pi^\dagger\{s,a\}(s') = a$, otherwise. In the linear reward setting, these constraints are represented compactly by the polytope $M_{\pi^\dagger}^\top \omega \geq \boldsymbol{\epsilon}'$ (see Figure 1 for a geometric illustration), where $M_{\pi^\dagger}$ is

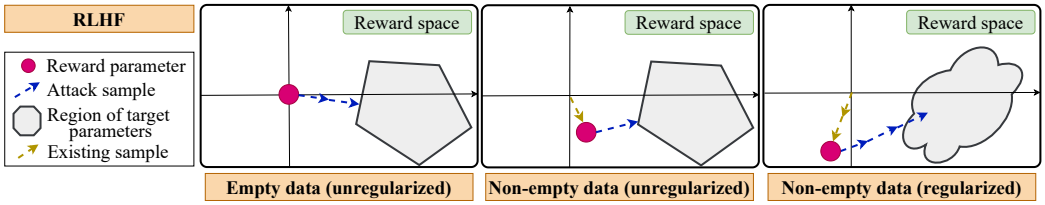

Figure 1: A geometric illustration of our attack model for RLHF. The shaded regions represent the reward parameter spaces where optimal policies are ϵ-close to $\pi^\dagger$. The blue arrows represent attack samples, while the yellow arrows represent the pre-existing data samples from $\overline{D}$. Finally, the red shape represents the optimal reward parameters with respect to the generated dataset $\widehat{D}$. Each added attack sample moves the optimal parameter closer to the shaded region. For unregularized RLHF with empty $\overline{D}$ (left), the attack problem is solved in the reward parameter space, and the target space is a polytope. For unregularized RLHF with non-empty $\overline{D}$ (middle), the required samples depend on the alignment of $\pi^\dagger$ with $\overline{D}$. For regularized RLHF (right), since the optimal policy is not necessarily deterministic, the geometry of the target space becomes non-linear.

a $d \times S(A-1)$-dimensional matrix.[1] Its columns are defined as $\sum_{s'} \phi(s', \pi^\dagger(s')) - \phi(s', \pi^\dagger\{s,a\}(s'))$, for all (s,a), and $\boldsymbol{\epsilon}'$ is the $S(A-1)$-dimensional vector with entries ϵ'. The columns of matrix $M_{\pi^\dagger}$ represent the differences in the average feature distribution between $\pi^\dagger$ and its neighbors.

Leveraging this polytope constraint, we instantiate the general preference poisoning attack problem P:Attack:RLHF.1 for the RLHF paradigm in the setting where the clean preference dataset $\overline{D}$ is non-empty (with size $\left|\overline{D}\right| = \overline{n}$), as follows:

$$\min_{D} |D| \text{ such that } \widehat{\omega} = \arg\min_{\omega} \ell^{\omega}_{\mathrm{RLHF}}\left(\overline{D} \cup D\right) \text{ and } M_{\pi^\dagger}^\top \widehat{\omega} \geq \boldsymbol{\epsilon}' \,. \tag{P:Attack:RLHF.2}$$

Note that the above is a surrogate of Problem P:Attack:RLHF.1, since, in unregularized RLHF $\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}$ is the optimal policy with respect to $r_{\widehat{\omega}}$ which, as we explained above, is enforced by the constraint $M_{\pi^\dagger}^\top \widehat{\omega} \geq \boldsymbol{\epsilon}'$. The solution to the above data augmentation attack problem depends on data-dependent quantities due to the pre-existing data. Let us define the covariance matrix with respect to $\overline{D}$ as

$$\Sigma^{\phi}_{\overline{D}} = (1/\overline{n}) \sum_{(\tau,\tau') \in \overline{D}} (\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^\top \,.$$

We are now ready to state our first result.

Theorem 4.1. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, let $\beta = 0$, $\epsilon' > 0$ and $\pi^\dagger \in \Pi^{\mathrm{det}}$. Furthermore, let $\overline{\omega}$ be optimal for $\ell^{\omega}_{\mathrm{RLHF}}(\overline{D})$, define $\omega^\dagger$ as*

$$\mathrm{proj}_{\omega : M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}'}(\overline{\omega}) = \overline{\omega} + M_{\pi^\dagger}\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^{+}\left(\boldsymbol{\epsilon}' - M_{\pi^\dagger}\overline{\omega}\right)$$

and let $\gamma \geq 1 - 2\left\|\omega^\dagger\right\| / (\xi_{\max} + 1)$. Then, the dataset of $\left\lceil \left| (\omega^\dagger)^\top \nabla_\omega \ell^{\omega^\dagger}_{\mathrm{RLHF}}(\overline{D}) \right| / \xi_{\max} \right\rceil$ identical samples satisfying $o = 1$ and

$$\phi(\tau) - \phi(\tau') = \xi_1 \left(\left(\omega^\dagger\right)^\top \nabla_\omega \ell^{\omega^\dagger}_{\mathrm{RLHF}}(\overline{D}) \right) \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2},$$

is a feasible solution to Problem P:Attack:RLHF.1. Furthermore, there exists an optimal solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\mathrm{RLHF}}$ identical samples such that

$$\widehat{n}_{\mathrm{RLHF}} \leq \left\lceil \frac{2\overline{n} + \lambda}{\xi_{\max}} \left(\frac{(\epsilon')^2 SA}{\sigma^2_{\min}(M_{\pi^\dagger})} + \|\overline{\omega}\| \frac{\epsilon' \sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})} \right) \right\rceil \,.$$

Sketch of proof. We start by considering the reward learning subproblem for the surrogate problem P:Attack:RLHF.2. We utilize the solution of the problem of machine teaching to logistic regression learners [Liu and Zhu, 2016] and relate it to our reward subproblem. Next, we proceed to solving the modified optimization problem with respect to the final constraint. In our setting, the samples from $\overline{D}$ are fixed, and thus cannot be treated as variable. Therefore, we need to design a new attack dataset $\widehat{D}$ with samples depending on the gradient of the loss with respect to $\overline{D}$, which captures how aligned $\overline{D}$ is with $\pi^\dagger$. In the best case, the optimal parameter with respect to $\overline{D}$ is already in the target polytope, meaning that the number of samples in this case is 0. In order to obtain closed-form bounds, we construct a feasible solution using the projection of the optimal solution with respect to $\overline{D}$ onto the polytope and make use of strong convexity and Lipschitzness of $\ell^{\omega}_{\mathrm{RLHF}}(\overline{D})$ to get our results. We also show in Appendix E that $M_{\pi^\dagger}$ is full rank under mild assumptions. This guarantees that our bounds are finite. □

Before going to the next section, we instantiate Problem P:Attack:RLHF.2 to the case when the pre-existing

[1] Note that, since the vector $\boldsymbol{\epsilon}'$ has all entries ϵ, the inequality condition can be satisfied even when the rank of $M_{\pi^\dagger}$ is less than SA. The minimal requirement is that all inequalities of the system are consistent, i.e., they yield intersecting regions.

dataset is empty. The following result is a corollary of Theorem 4.1.

Corollary 4.1. *Let $\overline{D} = \emptyset$, $\beta = 0$ and $\epsilon' > 0$, and let $\pi^\dagger \in \Pi^{\text{det}}$. Define $\omega^\dagger = M_{\pi^\dagger} \left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^+ \boldsymbol{\epsilon}'$. Then, the dataset of $\left\lceil \frac{\lambda \|\omega^\dagger\|^2}{\xi_{\max}} \right\rceil$ identical samples satisfying*

$$\phi(\tau) - \phi(\tau') = \xi_1 \left(\lambda \left\|\omega^\dagger\right\|\right) \cdot \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2}, \quad o = 1$$

is a feasible solution for Problem P:Attack:RLHF.1. Furthermore, there exists an optimal solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\text{RLHF}}$ samples such that

$$\widehat{n}_{\text{RLHF}} \leq \left\lceil \frac{(\epsilon')^2 \lambda S A}{\xi_{\max} \sigma_{\min}^2 \left(M_{\pi^\dagger}\right)} \right\rceil .$$

4.2 Regularized RLHF.

Now, we consider Problem P:Attack:RLHF.1 under a general setting, where the regularization parameter $\beta > 0$ and the preference dataset $\overline{D} \neq \emptyset$. In this setting, we are dealing with loglinear policies (see Section 3) which, since they are stochastic, lead to infinitely many constraints – hence, it is challenging to apply the polytope constraint idea from the above unregularized RLHF setting. Therefore, we consider a surrogate problem with constraints that are suited to stochastic policies. We will make use of the KL divergence for that purpose. We write the surrogate attack problem as

$$\begin{aligned} \min_{D} \ |D| \ \text{ such that } \ \widehat{\omega} = \arg\min_{\omega} \ell^{\omega}_{\text{RLHF}} \left(\overline{D} \cup D\right) \\ \text{and } \ D_{\text{KL}} \left(\pi^\dagger || \pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) \leq \epsilon' , \end{aligned} \qquad \text{(P:Attack:RLHF.3)}$$

where $\pi^{\text{reg}}_{r_{\widehat{\omega}}}$ is the optimal policy for the regularized objective P:RLHF.Policy. We obtain the following bounds for this setting:

Theorem 4.2. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, $\beta > 0$ and $0 < \epsilon' \leq \epsilon$. Moreover, define*

$$\Gamma^{\omega}_{\phi}(\pi^\dagger || \pi^{\text{reg}}_{r_\omega}) = \sum_{s,a} \rho(s)(\pi^\dagger(a|s) - \pi^{\text{reg}}_{r_\omega}(a|s))\phi^{\pi^{\text{reg}}_{r_\omega}}(s,a) ,$$

for any given ω. Then, there exists a feasible solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\text{RLHF}}$ samples which yields $\omega^\dagger$ when solving Problem P:RLHF.Reward on dataset $\widehat{D}$, such that $\Gamma^{\omega}_{\phi}(\pi^\dagger || \pi^{\text{reg}}_{r_{\omega^\dagger}}) \neq 0$ and

$$\widehat{n}_{\text{RLHF}} \leq O\left(\frac{\beta^2 \left(D_{\text{KL}}\left(\pi^\dagger || \mu\right) - \epsilon'\right)^2}{(1-\gamma)^2 \sigma^2_{\min}(\Sigma^{\phi}_{\overline{D}}) \left\| \Gamma^{\omega^\dagger}_{\phi}(\pi^\dagger || \pi^{\text{reg}}_{r_{\omega^\dagger}}) \right\|^2} + \frac{\overline{n}}{(1-\gamma)^4 \sigma^4_{\min}(\Sigma^{\phi}_{\overline{D}})} \right) .$$

Sketch of proof. We start by showing that solving the surrogate subproblem is enough to obtain upper bounds on the sample size for suitable ϵ'. Then, we solve the reward learning subproblem and obtain a dataset $\widehat{D}$ with $\widehat{n}_{\text{RLHF}}$ identical samples that satisfy $\phi(\tau) - \phi(\tau') = \xi^{-1}\left(\lambda \left|\nabla_\omega \ell^{\omega}_{\text{RLHF}}(\overline{D})^\top \omega\right| / \widehat{n}_{\text{RLHF}}\right) \omega / \|\omega\|^2$ with $o = 1$, where $\widehat{n}_{\text{RLHF}}$ is given in terms of $\left|\nabla_\omega \ell^{\omega}_{\text{RLHF}}(\overline{D})^\top \omega\right|$. Then, we solve the equivalent problem of minimizing $\left|\nabla_\omega \ell^{\omega}_{\text{RLHF}}(\overline{D})^\top \omega\right|$, subject to ω yielding a regularized optimal policy $\pi^{\text{reg}}_{r_\omega}$ that satisfies $D_{\text{KL}}(\pi^\dagger || \pi^{\text{reg}}_{r_\omega}) \leq \epsilon$. Using $\omega^\dagger$ as a feasible solution in the expression, we finalize the bounds.

Remark 4.1. Note that the term $\Gamma^{\omega^\dagger}_{\phi}(\pi^\dagger || \pi^{\text{reg}}_{r_{\omega^\dagger}})$ represents the average trajectory feature difference between $\pi^\dagger$ and $\pi^{\text{reg}}_{r_\omega}$ when rolling out trajectories using $\pi^{\text{reg}}_{r_{\omega^\dagger}}$. If this term is uniformly $\mathbf{0}$ at the problem solution $\omega^\dagger$, then this means that, either $\pi^\dagger = \pi^{\text{reg}}_{r_{\omega^\dagger}}$, or at least the two policies are identical in the average trajectory feature space. Therefore, less samples are needed to satisfy our objective. The following result provides bounds that depend on such a solution $\omega^\dagger$. Its proof follows immediately from the proof of Theorem 4.2.

Corollary 4.2. *There exists $\omega^\dagger$ such that $\pi^\dagger = \pi^{\text{reg}}_{r_{\omega^\dagger}}$. Moreover, we have that $\omega^\dagger$ is a feasible solution for Problem P:Attack:RLHF.1 and*

$$\widehat{n}_{\text{RLHF}} \leq \left\lceil \frac{1}{\xi_{\max}} \left(\lambda \left\|\omega^\dagger\right\|^2 + \left\|\omega^\dagger\right\| \cdot \frac{2\overline{n}}{(1-\gamma)^2} \right) \right\rceil .$$

Finally, we instantiate the above result in the case when the pre-existing data is empty.

Corollary 4.3. *Let $\overline{D} = \emptyset$, $\beta > 0$ and $0 < \epsilon' \leq \epsilon$. There exists a feasible solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\text{RLHF}}$ samples such that*

$$\widehat{n}_{\text{RLHF}} \leq \left\lceil \frac{\lambda \left\|\omega^\dagger\right\|^2}{\xi_{\max}} \right\rceil ,$$

where $\omega^\dagger$ is defined as in Corollary 4.2.

Remark 4.2. Note that, when $\overline{D} = \emptyset$, the number of samples required for an efficient attack on regularized RLHF depends on the norm of the reward parameter that makes $\pi^\dagger$ nearly-optimal.

5 Poisoning Attacks on DPO

As we have mentioned, the DPO objective is formulated only for the contextual bandit setting. Recall that trajectories here are defined in terms of context-action pairs $\tau = (s, a)$. We instantiate the general poisoning attack problem P:Attack for the DPO paradigm as follows:

$$\min_{D} |D| \text{ such that } \widehat{\theta} = \arg\min_{\theta} \ell^{\theta}_{\mathrm{DPO}}(D \cup \overline{D}) \text{ and } \left\|\pi^{\dagger} - \pi_{\widehat{\theta}}\right\|_1^2 \leq \epsilon \ . \quad \text{(P:Attack:DPO.1)}$$

In general, obtaining an intuitive form of attack construction for the above problem, as presented in Theorem 4.1, is challenging. To facilitate a more intuitive and efficient attack construction, we consider the following surrogate problem:

$$\min_{D} |D| \text{ such that } \widehat{\theta} = \arg\min_{\theta} \ell^{\theta}_{\mathrm{DPO}}\left(\overline{D} \cup D\right) \text{ and } \left\|\widehat{\theta} - \theta^{\dagger}\right\|^2 \leq \epsilon' \ , \quad \text{(P:Attack.DPO.2)}$$

where $\theta^{\dagger} \in \mathbb{R}^{d'}$ such that $\pi^{\dagger} = \pi_{\theta^{\dagger}}$ (see Section 3). Specifically, we have replaced the ℓ_1-norm-based constraint with an ℓ_2-norm-based constraint on the policy parameter space. The attack is successful if the learner's policy parameter converges closer to the target parameter $\pi_{\theta^{\dagger}}$. We derive the following result for this setting:

Theorem 5.1. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, let $\beta > 0$ and $0 < \epsilon' \leq \epsilon/2$. Furthermore, let $\overline{\theta}$ be the optimal point for $\ell^{\theta}_{\mathrm{DPO}}(\overline{D})$ and define*

$$\widetilde{\theta} = \mathrm{proj}_{\theta:\|\theta-\theta^{\dagger}\|\leq\epsilon'}\left(\overline{\theta}\right) = \theta^{\dagger} + \frac{(\epsilon')^2}{\left\|\overline{\theta} - \theta^{\dagger}\right\|^2}\left(\overline{\theta} - \theta^{\dagger}\right) \ .$$

Then, the dataset $\widehat{D}$ containing

$$2\left\lceil |(\nabla_{\theta}\ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D}))^{\top}(\widetilde{\theta} - \theta_{\mu})|/(2\xi_{\max})\right\rceil$$

identical samples satisfying

$$\beta\left(\widetilde{\theta} - \theta_{\mu}\right)^{\top}\left(\psi(s,a) - \psi(s,a')\right) = o \cdot \xi_2\left(\nabla_{\theta}\ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D}))^{\top}\left(\widetilde{\theta} - \theta_{\mu}\right)\right),$$

with $o = 1$ for half the samples and $o = -1$ for the remaining, is a feasible solution to Problem P:Attack:DPO.1. Furthermore, there exists an optimal solution $\widehat{D}$ to Problem P:Attack:DPO.1 with $\widehat{n}_{\mathrm{DPO}}$ identical samples such that

$$\widehat{n}_{\mathrm{DPO}} \leq 2 \cdot \left\lceil (\overline{n}\beta + \lambda)\frac{\left|\left\|\overline{\theta} - \theta^{\dagger}\right\|^2 - (\epsilon')^2\right|}{2\xi_{\max}\left\|\theta^{\dagger} - \overline{\theta}\right\|}\left(3\left\|\theta^{\dagger}\right\| + \|\theta_{\mu}\| + \sqrt{\epsilon'}\right)\right\rceil .$$

Sketch of proof. First, we show that, for suitable ϵ', any feasible solution to the surrogate problem is feasible for the original problem. The first main challenge in this setting for constructing $\widehat{D}$ is the presence of both $\overline{D}$ and θ_{μ} in the DPO objective. We first show that, given parameter θ, the sample size that makes π_{θ} optimal for $\ell^{\theta}_{\mathrm{DPO}}(\overline{D} \cup D)$ is a factor of $\left|\nabla_{\theta}\ell^{\theta}_{\mathrm{DPO}}(\overline{D})^{\top}(\theta - \theta_{\mu})\right|$. Using this, we redefine the objective of our problem and use the ℓ_2-ball centered at $\theta^{\dagger}$ with radius $\sqrt{\epsilon'}$ as constraint. The second main challenge consists of dealing with inverses of sums of matrices, due to the effect of the pre-existing data, for which we use the Woodbury inversion formula, and then proceed to solve a quadratic equation to obtain a fixed-point solution of our problem. Using the properties of that solution, we obtain the bounds on the norm of the optimal parameter $\widetilde{\theta}$. To obtain bounds on the norm of the gradient, we utilize Lipschitzness of $\ell^{\theta}_{\mathrm{DPO}}(\overline{D})$, and the geometrical relationship between $\overline{\theta}$ and $\theta^{\dagger}$. □

Next, we establish a lower bound on the attack sample complexity for the DPO setting.

Theorem 5.2. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples and let $\beta > 0$. Then, there exists $\eta_{\min} > 0$, such that for any $\epsilon' \geq \epsilon/\eta_{\min}$, we have*

$$\widehat{n}_{\mathrm{DPO}} \geq 2\left\lceil \frac{\lambda}{2\xi_{\max}}\left(\left\|\theta^{\dagger} - \theta_{\mu}\right\| - \sqrt{\epsilon'}\right)^2\right\rceil - \overline{n} \ .$$

Sketch of proof. Let $\Theta_1 = \{\theta : \|\pi_{\theta} - \pi_{\theta^{\dagger}}\|_1^2 \leq \epsilon$ and $\Theta_2 = \{\theta : \left\|\theta - \theta^{\dagger}\right\|^2 \leq \epsilon'\}$. First, we prove that there exists a positive constant $\eta_{\min}$ such that $\Theta_1 \subseteq \Theta_2$, for any $\epsilon' \geq \epsilon/\eta_{\min}$. This means that the feasible region of the surrogate problem is larger, which implies that a lower bound on the solution of Problem P:Attack.DPO.2 is also a lower bound on the solution of Problem P:Attack:DPO.1. Next, we focus on Problem P:Attack.DPO.2 and show that we can reduce it to a convex program. Using KKT conditions, we obtain an exact solution to the problem. We then use this solution as a lower bound for Problem P:Attack:DPO.1. □

Remark 5.1. Note that the upper bounds from Theorem 5.1 hold for different value of ϵ' than the one required for the lower bounds of Theorem 5.2. This is because we are essentially tuning the radius of the feasible region of Problem P:Attack.DPO.2 so that it is either contained in the feasible region of Problem P:Attack:DPO.1 (which is what we need for the upper bounds), or it contains the feasible region of Problem P:Attack:DPO.1 (which is what we need for lower bounds).

When there is no pre-existing data, the attacker can synthesise any poisoning dataset from scratch. An immediate solution to this problem is the instantiation of Theorem 5.1 when $\overline{D} = \emptyset$. However, below we also provide tight upper bounds that match the lower bounds of Theorem 5.2 for the empty data setting.

Theorem 5.3. *Let $\overline{D} = \emptyset$, let $\beta > 0$ and $0 < \epsilon' \leq \epsilon/2$. Furthermore, let $\pi^{\dagger}, \mu \in \Pi^{\log}$ be loglinear with*

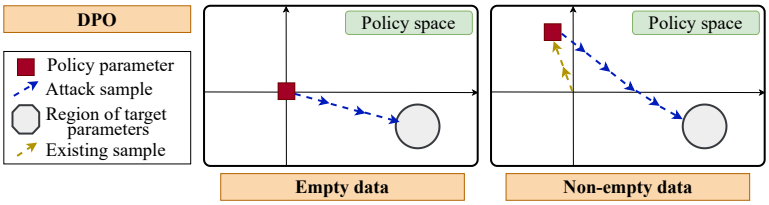

Figure 2: A geometric illustration of our attack model for DPO. Here, the distinction between empty and non-empty $\overline{D}$ is similar to Figure 1. In contrast to the RLHF setting, here the attacker operates directly in the policy parameter space and the target feasible region is a ball centered around $\theta^\dagger$ with radius ϵ as outlined in the formulation of Problem P:Attack.DPO.2.

parameters $\theta^\dagger$ and θ_μ, respectively. Define

$$\widetilde{\theta} = \theta^\dagger + e\sqrt{\epsilon'}(\theta_\mu - 2\theta^\dagger)/\left\|\theta_\mu - 2\theta^\dagger\right\|,$$

where $e = 1$, if ${\theta^\dagger}^\top\left(\theta^\dagger - \theta_\mu\right) \geq \sqrt{\epsilon'} - \epsilon'$, and $e = -1$, otherwise. Then, the dataset of $2\left\lceil \frac{\lambda|\lambda\widetilde{\theta}^\top(\widetilde{\theta}-\theta_\mu)|}{2\xi_{\max}}\right\rceil$ samples satisfying

$$\beta\left(\widetilde{\theta} - \theta_\mu\right)^\top(\psi(s,a) - \psi(s,a')) = o \cdot \xi_2\left(\lambda\left\|\widetilde{\theta} - \theta_\mu\right\|^2\right)$$

with $o = 1$ for half of the samples, and $o = -1$ for the remaining is a feasible solution to Problem P:Attack:DPO.1. Furthermore, there exists an optimal solution $\widehat{D}$ to Problem P:Attack:DPO.1 with $\widehat{n}_{\mathrm{DPO}}$ identical samples such that

$$\widehat{n}_{\mathrm{DPO}} \leq 2\left\lceil \frac{\lambda}{2\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2\right\rceil.$$

Finally, there exists $\eta_{\min} > 0$, such that, for any $\epsilon' \geq \epsilon/\eta_{\min}$, we have

$$\widehat{n}_{\mathrm{DPO}} \geq 2\left\lceil \frac{\lambda}{2\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2\right\rceil.$$

6 Comparison between RLHF and DPO for Attack Susceptibility

In this section, we present some interesting takeaways from the analysis of previous sections. We aim to provide a comparative analysis of the attack sample complexities between the RLHF and DPO paradigms. Specifically, we focus on the contextual bandit setting.

The following result establishes an explicit relationship between the sample complexities of data augmentation attacks on the RLHF and DPO paradigms. Specifically, we compare the sample complexities, $\widehat{n}_{\mathrm{RLHF}}$ and $\widehat{n}_{\mathrm{DPO}}$, required by the optimal solutions to Problems P:Attack:RLHF.1 and P:Attack:DPO.1, respectively (with $\left|\overline{D}\right| = \overline{n}$).

Theorem 6.1. *Let $\pi^\dagger, \mu \in \Pi^{\log}$ be loglinear with parameters $\theta^\dagger$ and θ_μ, respectively. Furthermore, let ϵ be such that $\epsilon \leq 1/(2\xi_{\max})$, $\epsilon' \geq \epsilon/\eta_{\min}$, where $\eta_{\min} > 0$ is an absolute constant, and let $\omega^\dagger$ be a feasible solution to Problem P:Attack:RLHF.1. Define κ_1 as*

$$\left(\frac{\lambda}{\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2 - \overline{n}\right) \cdot \left\lceil \frac{1}{\xi_{\max}}\left(\lambda\left\|\omega^\dagger\right\|^2 + \left\|\omega^\dagger\right\| \cdot \frac{2\overline{n}}{(1-\gamma)^2}\right)\right\rceil^{-1}$$

Then, we have $\widehat{n}_{\mathrm{DPO}} \geq \kappa_1 \cdot \widehat{n}_{\mathrm{RLHF}}$.

The value of κ_1 is proportional to the distance between $\theta^\dagger$ and θ_μ, which captures how far $\pi^\dagger$ is from the reference policy μ. We observe that, the greater the distance between $\pi^\dagger$ and μ, the less susceptible DPO becomes relative to RLHF. Lower susceptibility implies DPO has a stronger tendency to remain close to μ. Furthermore, note that if $\frac{\lambda}{\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2 \leq \overline{n}$, the lower bound in Theorem 5.2 becomes vacuous. Therefore, we should assume that $\pi^\dagger$ and μ are far enough, or that the size of the $\overline{D}$ is small, for this bound to be meaningful. This assumption is not restrictive, as in practice, the nature of $\pi^\dagger$ often differs significantly from μ, leading to large divergence terms.

7 Related Work

Adversarial attacks in machine learning (ML). The problem of adversarial attacks in ML has a long history [Szegedy et al., 2013, Biggio et al., 2013, Nguyen et al., 2015, Papernot et al., 2017, Biggio et al., 2012, Li et al., 2016, Xiao et al., 2012], where various types of attacks have been considered, including training-time attacks, test-time attacks, and backdoor attacks. The focus of the present study is on training-time attacks. The closest to our work in this domain is that of Liu and Zhu [2016], who consider the teaching problem (via data synthesis) to various types of learners, including logistic regression learners. Similar to [Liu and Zhu, 2016], we also consider logistic regression in our optimization problems. However, our attack problems include additional constraints, which necessitate the usage of additional technical machinery.

Data poisoning attacks and defenses in (multi-agent) reinforcement learning ((MA)RL). Adversarial attacks in RL have been explored extensively in the literature [Huang et al., 2017, Gleave et al., 2020, Lin et al., 2017, Sun et al., 2020, Rangi et al., 2022b, Ma et al., 2019, Rakhsha et al., 2020a, 2021], including training-time attacks [Rakhsha et al., 2020b, Xu et al., 2021], test-time attacks [Behzadan and Munir, 2017, Huang et al., 2017, Kos and Song, 2017, Sun et al., 2020], backdoor attacks [Kiourti et al., 2020, Wang et al., 2021, Yang et al., 2019] and attacks to MARL systems [Wu et al., 2024, Nika et al., 2024c, Mohammadi et al., 2023, Nika et al., 2024b]. Our research focuses on training-time poisoning attacks against single agents, where adversaries manipulate training data within certain constraints [Mei and Zhu, 2015, Xiao et al., 2015, Rakhsha et al., 2020a, 2021]. In the unregularized RLHF setting, our attack problem utilizes constraints that determine the target policy's strict optimality as in [Rakhsha et al., 2021]. Different from these works, our focus is on the studying attacks in RLHF.

Complementing this, significant research has also been conducted on robust RL methods designed to defend against poisoning attacks [Zhang et al., 2021, Lykouris et al., 2021, Kumar et al., 2021, Rangi et al., 2022a, Wu et al., 2022, Zhang et al., 2022, McMahan et al., 2024, Banihashem et al., 2023, Nika et al., 2023]. Our work diverges from these studies by being the first to theoretically investigate the inherent robustness of RLHF and DPO paradigms against poisoning attacks.

Data poisoning attacks and defenses in learning from human preferences. Recent studies have empirically explored the vulnerability of the RLHF paradigm to poisoning attacks [Wang et al., 2023b, Shi et al., 2023, Rando and Tramèr, 2023, Baumgärtner et al., 2024], addressing various attack types such as label flipping, backdoor, and data augmentation. Despite these empirical explorations, prior work has primarily focused on evaluating the effectiveness of these attacks, whereas our research aims to provide a theoretical understanding of these vulnerabilities. On the defense side, Mandal et al. [2024] and Chowdhury et al. [2024] have proposed robust RLHF and DPO algorithms to handle data corruption. However, our work aims to understand the natural robustness of these algorithms against structured data poisoning attacks and studies the problem from an attacker's perspective.

Theoretical analysis of learning from human preferences. Significant research has focused on theoretically understanding and improving the performance of RLHF and DPO [Zhu et al., 2023, Zhan et al., 2023, An et al., 2023, Azar et al., 2023, Wang et al., 2023a, Hejna et al., 2023, Nika et al., 2024a]. While in [Zhu et al., 2023, Zhan et al., 2023] the focus is on unregularized RLHF, Azar et al. [2023], Hejna et al. [2023], Nika et al. [2024a] study regularized RLHF and the effect of regularization. Our work also tries to further theoretical understanding of learning from human preferences, by undertaking a rigorous analysis of data poisoning attacks on these methods.

8 Concluding Discussion

We considered data poisoning attacks in learning from human preferences. Specifically, we studied data augmentation attacks on RLHF and DPO. Based on our findings, we compared these paradigms in terms of susceptibility to attacks. There are several directions for future work. First, it would be interesting to relax the (log)linearity assumptions and solve the problem for general parametrizations. It is currently not clear whether RLHF and DPO attacks are feasible for general formulations. Second, as it is not possible to obtain closed-form solutions to our problems whenever KL constraints are present, it would be useful to study this attack framework with alternate constraints, such as total variation distance. Third, it would also be important to study other forms of attacks, e.g., *label-flipping attacks* where the attacker is only allowed to flip a fraction of the preference labels in the dataset. Finally, it would be interesting to study attacks against more recent preference-based RL methods and understand the effectiveness of these attacks when a learner uses robust variants of these methods.

Acknowledgements

The work of Andi Nika and Goran Radanovic was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. Direct Preference-based Policy Optimization without Reward Modeling. In *NeurIPS*, 2023.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A General Theoretical Paradigm to Understand Learning from Human Preferences. *CoRR*, abs/2310.12036, 2023.

Yuntao Bai et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862, 2022.

Kiarash Banihashem, Adish Singla, and Goran Radanovic. Defense Against Reward Poisoning Attacks in Reinforcement Learning. *TMLR*, 2023.

Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. Best-of-Venom: Attacking RLHF by Injecting Poisoned Preference Data. *CoRR*, abs/2404.05530, 2024.

Vahid Behzadan and Arslan Munir. Whatever does not Kill Deep Reinforcement Learning, Makes it Stronger. *CoRR*, abs/1712.09344, 2017.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks against Support Vector Machines. In *ICML*, 2012.

Battista Biggio et al. Evasion Attacks Against Machine Learning at Test Time. In *ECML PKDD*, 2013.

Ralph Allan Bradley and Milton E Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4), 1952.

Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In *ICML*, 2019.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably Robust DPO: Aligning Language Models with Noisy Feedback. *CoRR*, abs/2403.00409, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *ICML*, 2023.

Amelia Glaese, Nat McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *CoRR*, abs/2209.14375, 2022.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial Policies: Attacking Deep Reinforcement Learning. In *ICLR*, 2020.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive Prefence Learning: Learning from Human Feedback without RL. *CoRR*, abs/2310.13639, 2023.

Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the Lambert W Function and Hyperpower Function. *J. Inequal. Pure and Appl. Math*, 9(2):5–9, 2008.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial Attacks on Neural Network Policies. *CoRR*, abs/1702.02284, 2017.

Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. Trojdrl: Evaluation of Backdoor Attacks on Deep Reinforcement Learning. In *ACM/IEEE (DAC)*, 2020.

Jernej Kos and Dawn Song. Delving into Adversarial Attacks on Deep Policies. *CoRR*, abs/1705.06452, 2017.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy Smoothing for Provably Robust Reinforcement Rearning. *CoRR*, abs/2106.11420, 2021.

Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data Poisoning Attacks on Factorization-based Collaborative Filtering. In *NeurIPS*, 2016.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In *IJCAI*, 2017.

Ji Liu and Xiaojin Zhu. The Teaching Dimension of Linear Learners. *Journal of Machine Learning Resesearch*, 17:162:1–162:25, 2016.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust Exploration in Episodic Reinforcement Learning. In *COLT*, 2021.

Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy Poisoning in Batch Reinforcement Learning and Control. In *NeurIPS*, 2019.

Debmalya Mandal, Andi Nika, Parameswaran Kamalaruban, Adish Singla, and Goran Radanović. Corruption Robust Offline Reinforcement Learning with Human Feedback. *CoRR*, abs/2402.06734, 2024.

Jeremy McMahan, Young Wu, Xiaojin Zhu, and Qiaomin Xie. Optimal Attack and Defense for Reinforcement Learning. In *AAAI*, 2024.

Shike Mei and Xiaojin Zhu. Using Machine Teaching to Identify Optimal Training-set Attacks on Machine Learners. In *AAAI*, 2015.

Jacob Menick et al. Teaching Language Models to Support Answers with Verified Quotes. *CoRR*, abs/2203.11147, 2022.

Mohammad Mohammadi, Jonathan Nöther, Debmalya Mandal, Adish Singla, and Goran Radanovic. Implicit Poisoning Attacks in Two-agent Reinforcement Learning: Adversarial Policies for Training-time Attacks. In *AAMAS*, 2023.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the Gap Between Value and Policy based Reinforcement Learning. In *NeurIPS*, 2017.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *CVPR*, 2015.

Andi Nika, Adish Singla, and Goran Radanovic. Online Defense Strategies for Reinforcement Learning Against Adaptive Reward Poisoning. In *AISTATS*, 2023.

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. In *ICML*, 2024a.

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanovic. Corruption-robust Offline Two-player Zero-sum Markov Games. In *AISTATS*, pages 1243–1251. PMLR, 2024b.

Andi Nika, Jonathan Nöther, Adish Singla, and Goran Radanovic. Defending Against Unknown Corrupted Agents: Reinforcement Learning of Adversarially Robust Nash Equilibria. 2024c.

Long Ouyang et al. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*, 2022.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical Black-box Attacks Against Machine Learning. In *ACM*, 2017.

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. In *NeurIPS*, 2023.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *ICML*, 2020a.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching via Environment poisoning: Training-time Adversarial Attacks Against Reinforcement Learning. In *ICML*, 2020b.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching in Reinforcement Learning via Environment Poisoning attacks. *Journal of Machine Learning Research*, 22: 210:1–210:45, 2021.

Javier Rando and Florian Tramèr. Universal Jailbreak Backdoors from Poisoned Human Feedback. In *ICLR*, 2023.

Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Saving Stochastic Bandits from Poisoning Attacks via Limited Data Verification. In *AAAI*, 2022a.

Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the Limits of Poisoning Attacks in Episodic Reinforcement Learning. In *IJCAI*, 2022b.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT. *CoRR*, abs/2304.12298, 2023.

Daniel Shin, Anca D. Dragan, and Daniel S. Brown. Benchmarks and Algorithms for Offline Preference-Based Reward Learning. *Transactions of Machine Learning Research*, 2023.

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. In *NeurIPS*, 2020.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and Efficient Adversarial Attacks Against Deep Reinforcement Learning. In *AAAI*, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. *CoRR*, abs/1312.6199, 2013.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. *CoRR*, abs/2309.16240, 2023a.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. On the Exploitability of Reinforcement Learning with Human Feedback for Large Language Models. *CoRR*, abs/2311.09641, 2023b.

Yue Wang, Esha Sarkar, Wenqing Li, Michail Maniatakos, and Saif Eddin Jabari. Stop-and-go: Exploring Backdoor Attacks on Deep Reinforcement Learning-based Traffic Congestion Control Systems. *IEEE Transactions on Information Forensics and Security*, 16, 2021.

Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. Copa: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks. *CoRR*, abs/2203.08398, 2022.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Data Poisoning to Fake a Nash Equilibria for Markov Games. In *AAAI*, 2024.

Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial Label Flips Attack on Support Vector Machines. In *ECAI*, 2012.

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is Feature Selection Secure Against Training Data Poisoning? In *ICML*, 2015.

Hang Xu, Rundong Wang, Lev Raizman, and Zinovi Rabinovich. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning. In *ICAAMS*, 2021.

Zhaoyuan Yang, Naresh Iyer, Johan Reimann, and Nurali Virani. Design of Intentional Backdoors in Sequential Models. *CoRR*, abs1902.09972, 2019.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable Offline Reinforcement Learning with Human Feedback. *CoRR*, abs/:2305.14816, 2023.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust Policy Gradient Against Strong Data Corruption. In *ICML*, 2021.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust Offline Reinforcement Learning. In *AISTATS*, 2022.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *ICML*, 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. *CoRR*, abs/1909.08593, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Table of Contents

A Feasibility of Problem P:Attack:RLHF.1.

In this section, we prove that Problem P:Attack:RLHF.1 is feasible for certain classes of $\pi^\dagger$.

Theorem A.1. *Let $\epsilon, \beta > 0$ and μ be with full support. Then, the following statements hold:*

- *Let $\pi^\dagger \in \Pi^{\text{det}}$ and assume that $\omega^\dagger_{\text{opt}}$ is such that $\pi^\dagger$ is ϵ-robust optimal with respect to $r_{\omega^\dagger_{\text{opt}}}$. Then, there exists $c > 0$ such that $c \cdot \omega^\dagger_{\text{opt}}$ is a feasible solution for Problem P:Attack:RLHF.1.*
- *Let $\pi^\dagger, \mu \in \Pi^{\log}$ and assume that the column space of Φ is a subspace of the column space of Ψ. Then the solution to $\Phi\omega = \beta(\log \widehat{\pi}^\dagger - \log \mu)$, where $\log \widehat{\pi}^\dagger - \log \mu$ is the vector with entries $\log \widehat{\pi}^\dagger(a|s) - \log \mu(a|s)$, for each (s, a), is a feasible solution for Problem P:Attack:RLHF.1, for any $\widehat{\pi}^\dagger$ such that $D_{\text{KL}}(\pi^\dagger || \widehat{\pi}^\dagger) \leq \epsilon'$.*

Proof. We start with the first statement. Given dataset D, logistic regression applied on D returns a reward function r. If D is to be feasible for Problem P:Attack:RLHF.1, this necessitates that the optimal regularized policy π^{reg}_r with respect to reward function r is ϵ-close to $\pi^\dagger$ in the sense that $\left\|\pi^\dagger - \pi^{\text{reg}}_r\right\|_1 \leq \epsilon$. Note that, if we can show that $\left\|\pi^\dagger - \pi^{\text{reg}}_r\right\|_\infty \leq \epsilon$, then we are done. Thus, for the rest of this proof, we will focus on $\|\cdot\|_\infty$.

Now, given reward function r with parameter ω, Theorem E.15 shows us that we can always synthesise a dataset D that makes r the outcome when logistic regression is applied on it. Therefore, we focus our attention on finding the right reward function r for which our constraint is satisfied. Note that our problem can be written as

$$\begin{aligned} &\arg\max_\pi V^\pi_r(\rho) - \beta D^\gamma_{\text{KL}}(\pi||\mu) := \pi^{\text{reg}}_r \\ &\text{s.t. } \left\|\pi^\dagger - \pi^{\text{reg}}_r\right\|_\infty \leq \epsilon \ , \end{aligned}$$

where

$$D^\gamma_{\text{KL}}(\pi||\mu) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \log \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \middle| \rho, \pi\right]$$

$$= \mathbb{E}\left[\sum_{t\geq 0}\gamma^t \sum_a \pi(a|s_t)\log\frac{\pi(a|s_t)}{\mu(a|s_t)}\middle|\rho,\pi\right]$$
$$= \mathbb{E}\left[\sum_{t\geq 0}\gamma^t D_{\mathrm{KL}}\left(\pi(\cdot|s_t)||\mu(\cdot|s_t)\right)\middle|\rho,\pi\right] . \tag{1}$$

Given $\kappa << \epsilon$, let us define the set $\Pi_\kappa = \{(1-\kappa)\cdot\pi + \kappa\cdot u | \pi\in\Pi\}$, where u denotes the policy that takes any action uniformly at random, at any given state. Furthermore, let $l = \min_{\pi,s} d^\pi_\rho(s)$. By the ergodicity assumption, we have that $l > 0$. Let $A^\pi_\rho(s,a)$ denote the advantage function of π at state-action (s,a).

By assumption, there exists an $\omega^\dagger$ such that $M^\top_{\pi^\dagger}\omega^\dagger \geq \boldsymbol{\epsilon}$, for a given ϵ. Obviously, the policy $\pi^\dagger$ is optimal with respect to the reward function $r_{\omega^\dagger}$. We will use the short-hand notation $r^\dagger = r_{\omega^\dagger}$.

Now, let us define $\Delta > 0$ such that $A^{\pi^\dagger}(s,a) \leq -\Delta$, for all $a \neq \pi^\dagger(s)$, let $c > 0$ be an arbitrary positive constant and let $r^\dagger_{\max}$ denote the maximum component of $r^\dagger$. Let $\pi_1 = (1-\kappa)\cdot\pi^\dagger + \kappa\cdot u$. For any policy π, we have

$$V^{\pi_1}_{c\cdot r^\dagger}(\rho) - V^{\pi}_{c\cdot r^\dagger}(\rho) = \left(V^{\pi_1}_{c\cdot r^\dagger}(\rho) - V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho)\right) + \left(V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho) - V^{\pi}_{c\cdot r^\dagger}(\rho)\right) .$$

We will bound each term separately. First, using vector notation for the reward and occupancy measures, note that

$$V^{\pi_1}_{c\cdot r^\dagger}(\rho) - V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho) = c\cdot\left(d^{\pi_1}_\rho - d^{\pi^\dagger}_\rho\right)^\top r^\dagger \tag{2}$$
$$\geq -c\cdot\left\|d^{\pi_1}_\rho - d^{\pi^\dagger}_\rho\right\|_1 \cdot \left\|r^\dagger\right\|_\infty \tag{3}$$
$$\geq -\frac{\gamma\cdot c\cdot\kappa\cdot r^\dagger_{\max}}{1-\gamma} , \tag{4}$$

where Equation 2 uses the occupancy measure expression of the discounted return; Equation 3 uses Cauchy-Schwarz; for Equation 4 we have used that $\left\|\pi_1 - \pi^\dagger\right\|_1 = \kappa$ and that *similar policies imply similar state visitations*, i.e., if $\|\pi_1 - \pi_2\|_1 \leq \zeta$, then $\left\|d^{\pi_1}_\rho - d^{\pi_2}_\rho\right\|_1 \leq \zeta\gamma/(1-\gamma)$, for given positive ζ (see the RL Theory book Lemma 14.1). For the second term, we have

$$V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho) - V^{\pi}_{c\cdot r^\dagger}(\rho) = -\left(V^{\pi}_{c\cdot r^\dagger}(\rho) - V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho)\right)$$
$$= -\left(\sum_{s'} d^\pi_\rho(s')\sum_{a'}\pi(a'|s')A^{\pi^\dagger}_{c\cdot r^\dagger}(s',a')\right) \tag{5}$$
$$= -\left(\sum_{s'} d^\pi_\rho(s')\sum_{a'\neq\pi^\dagger(s')}\pi(a'|s')A^{\pi^\dagger}_{c\cdot r^\dagger}(s',a')\right) \tag{6}$$
$$= \sum_{s'} d^\pi_\rho(s')\sum_{a'\neq\pi^\dagger(s')}\pi(a'|s')\cdot\left(-A^{\pi^\dagger}_{c\cdot r^\dagger}(s',a')\right)$$
$$\geq \sum_{s'} d^\pi_\rho(s')\sum_{a'\neq\pi^\dagger(s')}\pi(a'|s')\cdot c\cdot\Delta \tag{7}$$
$$\geq l\cdot\sum_{s',a'\neq\pi^\dagger(s')}\pi(a'|s')\cdot c\cdot\Delta , \tag{8}$$

where Equation 5 follows from the Performance Difference Lemma; Equation 6 uses the fact that $A^{\pi^\dagger}_{c\cdot r^\dagger}(s,\pi^\dagger(s)) = c\cdot A^{\pi^\dagger}_{r^\dagger}(s,\pi^\dagger(s)) = 0$; Equation 7 uses the fact that $A^{\pi^\dagger}_{c\cdot r^\dagger}(s,\pi^\dagger(s)) \leq -c\cdot\Delta$; Equation 8 uses the definition of l.

Now, let us define the set $\Pi^\epsilon_\kappa = \left\{\pi\in\Pi_\kappa : \left\|\pi - \pi^\dagger\right\|_\infty \leq \epsilon\right\}$. Equation 8 above implies that, for any $\pi\notin\Pi^\epsilon_\kappa$,

$$V^{\pi^\dagger}_{c\cdot r^\dagger}(\rho) - V^{\pi}_{c\cdot r^\dagger}(\rho) \geq l\cdot\sum_{s',a'\neq\pi^\dagger(s')}\pi(a'|s')\cdot c\cdot\Delta$$

$$\geq l \cdot \max_{s, a \neq \pi^\dagger(s)} \pi(a|s) \cdot c \cdot \Delta$$
$$> l \cdot \epsilon \cdot c \cdot \Delta \ , \tag{9}$$

where the third inequality follows from the fact that $\left\|\pi - \pi^\dagger\right\|_\infty > \epsilon$. Thus, combining Equation 4 and Equation 9 into the original suboptimality gap, for $\pi \notin \Pi^\epsilon_\kappa$, we obtain

$$V^{\pi_1}_{c\cdot r^\dagger}(\rho) - V^{\pi}_{c\cdot r^\dagger}(\rho) > l \cdot \epsilon \cdot c \cdot \Delta - \frac{\gamma \cdot c \cdot \kappa \cdot r^\dagger_{\max}}{1-\gamma} \ . \tag{10}$$

Now, for a given $\kappa > 0$, the exists $C^\kappa > 0$, such that $D^\gamma_{\mathrm{KL}}(\pi_1||\mu) < C^\kappa$, since both π_1 and μ have full support and thus $D^\gamma_{\mathrm{KL}}(\pi_1||\mu) \leq (1/(1-\gamma)) \max_s D_{\mathrm{KL}}(\pi_1(\cdot|s)||\mu(\cdot|s)) := C^\kappa$. Hence, we have that

$$\max_{\pi \in \Pi^\epsilon_\kappa} V^{\pi}_{c\cdot r^\dagger} - \beta D^\gamma_{\mathrm{KL}}(\pi||\mu) \geq V^{\pi_1}_{c\cdot r^\dagger} - \beta \cdot C^\kappa \ .$$

Moreover, using the fact that $D^\gamma_{\mathrm{KL}}(\pi||\mu) \geq 0$, we also have

$$\max_{\pi \in \Pi^\epsilon_\kappa} V^{\pi}_{c\cdot r^\dagger} - \beta D^\gamma_{\mathrm{KL}}(\pi||\mu) \leq \max_{\pi \in \Pi^\epsilon_\kappa} V^{\pi}_{c\cdot r^\dagger} < V^{\pi_1}_{c\cdot r^\dagger}(\rho) - c \cdot \left(l \cdot \epsilon \cdot \Delta - \frac{\gamma \cdot \kappa \cdot r^\dagger_{\max}}{1-\gamma} \right) \ ,$$

where the second inequality follows from Equation 10. Finally, let $\kappa << l \cdot \epsilon \cdot (1-\gamma)/(\gamma \cdot r^\dagger_{\max})$ and $c = \beta \cdot C^\kappa / (l \cdot \epsilon - (\gamma/(1-\gamma)) \cdot \kappa \cdot r^\dagger_{\max})$. Then, we obtain

$$\begin{aligned}
\max_{\pi \in \Pi^\epsilon_\kappa} V^{\pi}_{c\cdot r^\dagger} - \beta D^\gamma_{\mathrm{KL}}(\pi||\mu) &\geq V^{\pi_1}_{c\cdot r^\dagger} - \beta \cdot C^\kappa \\
&= V^{\pi_1}_{c\cdot r^\dagger} - c \cdot \left(l \cdot \epsilon \cdot \Delta - \frac{\gamma \cdot \kappa \cdot r^\dagger_{\max}}{1-\gamma} \right) \\
&> \max_{\pi \notin \Pi^\epsilon_\kappa} V^{\pi}_{c\cdot r^\dagger}(\rho) - \beta \cdot D^\gamma_{\mathrm{KL}}(\pi||\mu) \ .
\end{aligned}$$

Hence, we conclude that, for the choice of $r = c \cdot r^\dagger$ (and thus $\omega = c \cdot \omega^\dagger$), the solution to the regularized value maximization problem is necessarily in Π^ϵ_κ, and thus, ϵ-close to $\pi^\dagger$.

Next, we consider the second statement. Let $r = \beta(\log \pi^\dagger - \log \mu)$. Note that we have

$$\begin{aligned}
\arg\max_\pi V^\pi_r(\rho) - \beta D^\gamma_{\mathrm{KL}}(\pi||\mu) &= \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \left(r(s_t, a_t) - \beta \log \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \right) \middle| \pi, \rho \right] \\
&= \arg\max_\pi \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \left(\beta \log \frac{\pi^\dagger(a_t|s_t)}{\mu(a_t|s_t)} - \beta \log \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \right) \middle| \pi, \rho \right] \\
&= \arg\max_\pi \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \left(\beta \log \frac{\pi^\dagger(a_t|s_t)}{\pi(a_t|s_t)} \right) \middle| \pi, \rho \right] \\
&= \arg\max_\pi -\mathbb{E}\left[\sum_{t \geq 0} \gamma^t \left(\beta \log \frac{\pi(a_t|s_t)}{\pi^\dagger(a_t|s_t)} \right) \middle| \pi, \rho \right] \\
&= \arg\min_\pi D^\gamma_{\mathrm{KL}}\left(\pi||\pi^\dagger\right) \\
&= \pi^\dagger \ .
\end{aligned}$$

This implies that, as long as we can find a reward function as proposed, enforcing $\pi^\dagger$ is feasible. Since $\pi^\dagger, \mu \in \Pi^{\log}$, there exist parameters $\theta^\dagger, \theta_\mu$ that realize these policies in the loglinear space. Thus, to find a linear reward function that satisfies our equation, we can equivalently solve

$$\Phi\omega = \beta \cdot \Psi\left(\theta^\dagger - \theta_\mu\right) \ ,$$

which would automatically be a solution. Lemma 4.1 of [Nika et al., 2024a] guarantees that we can find such a vector ω whenever the column space of Φ is a subspace of the column space of Ψ. By assumption, this condition is satisfied, and thus there exists $\widehat{\omega}$ for which we have $r_{\widehat{\omega}} = \beta(\log \pi^\dagger - \log \mu)$. □

B Proofs of Section 4

In this section, we provide the full proofs of results from Section 4.

B.1 Unregularized RLHF with Nonempty Existing Data

We begin with the unregularized setting when the pre-existing data is non-empty and prove the following result.

Theorem 4.1. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, let $\beta = 0$, $\epsilon' > 0$ and $\pi^\dagger \in \Pi^{\text{det}}$. Furthermore, let $\overline{\omega}$ be optimal for $\ell^{\omega}_{\text{RLHF}}(\overline{D})$, define $\omega^\dagger$ as*

$$\text{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}'}(\overline{\omega}) = \overline{\omega} + M_{\pi^\dagger}\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^+ (\boldsymbol{\epsilon}' - M_{\pi^\dagger}\overline{\omega})$$

and let $\gamma \geq 1 - 2\left\|\omega^\dagger\right\|/(\xi_{\max}+1)$. Then, the dataset of $\left\lceil \left|(\omega^\dagger)^\top \nabla_\omega \ell^{\omega^\dagger}_{\text{RLHF}}(\overline{D})\right| / \xi_{\max} \right\rceil$ identical samples satisfying $o = 1$ and

$$\phi(\tau) - \phi(\tau') = \xi_1 \left(\left(\omega^\dagger\right)^\top \nabla_\omega \ell^{\omega^\dagger}_{\text{RLHF}}(\overline{D}) \right) \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2},$$

is a feasible solution to Problem P:Attack:RLHF.1. Furthermore, there exists an optimal solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\text{RLHF}}$ identical samples such that

$$\widehat{n}_{\text{RLHF}} \leq \left\lceil \frac{2\overline{n} + \lambda}{\xi_{\max}} \left(\frac{(\epsilon')^2 SA}{\sigma^2_{\min}\left(M_{\pi^\dagger}\right)} + \|\overline{\omega}\| \frac{\epsilon'\sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})} \right) \right\rceil .$$

Proof. Lemma E.3 implies that any feasible solution to Problem P:Attack:RLHF.2 is feasible for Problem P:Attack:RLHF.1. Thus, we focus on Problem P:Attack:RLHF.2.

First, note that the condition $\gamma \geq 1 - 2\left\|\omega^\dagger\right\|/(\xi_{\max}+1)$ is needed to ensure a well-defined feature construction for our dataset, based on the condition provided by Lemma E.10.

Now, let us consider the simpler problem of augmenting the data so that the solution to the logistic regression subproblem is a given $\widehat{\omega}$. The subproblem can be written as

$$\begin{aligned} &\min_{D} \ |D| \\ &\text{s.t } \ \widehat{\omega} = \arg\min_{\omega} \sum_{(\tau,\tau',o)\in\overline{D}\cup D} \log\left(1 + \exp\left(-o\cdot\omega^\top(\phi(\tau) - \phi(\tau'))\right)\right) + \frac{\lambda}{2}\|\omega\|^2 \ . \end{aligned}$$

Lemma E.6 implies that the solution to the above is the dataset of

$$\left\lceil \frac{\left|(\widehat{\omega})^\top \nabla_\omega \ell^{\widehat{\omega}}_{\text{RLHF}}(\overline{D})\right|}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau_j) - \phi(\tau_j') = \xi^{-1} \left((\widehat{\omega})^\top \nabla_\omega \ell^{\widehat{\omega}}_{\text{RLHF}}(\overline{D}) \left\lceil \frac{\left|(\widehat{\omega})^\top \nabla_\omega \ell^{\widehat{\omega}}_{\text{RLHF}}(\overline{D})\right|}{\xi_{\max}} \right\rceil^{-1} \right) \frac{\nabla_\omega \ell^{\widehat{\omega}}_{\text{RLHF}}(\overline{D})}{(\widehat{\omega})^\top \nabla_\omega \ell^{\omega^\dagger}_{\text{RLHF}}(\overline{D})} \ , o_j = 1 \ .$$

Given this solution, we can equivalently write Problem Equation P:Attack:RLHF.2 in terms of ω as

$$\begin{aligned} &\min_{\omega} \ \left|\omega^\top \nabla_\omega \ell^{\omega^\dagger}_{\text{RLHF}}(\overline{D})\right| \\ &\text{s.t. } \boldsymbol{\epsilon} - M_{\pi^\dagger}^\top \omega \leq \mathbf{0} \ . \end{aligned}$$

Now, before we go any further, it is important to note that, whenever we have

$$\nabla_\omega \ell^{\omega^\dagger}_{\text{RLHF}}(\overline{D}) = \mathbf{0} \ ,$$

the number of samples needed for the attack to succeed is 0 since $\omega^\dagger$ is already optimal with respect to $\overline{D}$. Recall that

$$\Sigma_{\overline{D}}^{\phi} = \frac{1}{\overline{n}} \sum_{(\tau,\tau')\in\overline{D}} \left(\phi(\tau) - \phi(\tau')\right)\left(\phi(\tau) - \phi(\tau')\right)^\top$$

denotes the sample covariance matrix with respect to $\overline{D}$ and let $C_{\overline{D}}^{\phi}$ denote its minimum eigenvalue. We will rewrite the problem using a different notation for simplicity in calculations. First, let

$$X_{\overline{D}}^{\omega} = \sum_{(\tau,\tau',o)\in\overline{D}} \frac{-o\left(\phi(\tau) - \phi(\tau')\right)}{1 + \exp\left(o\omega^\top\left(\phi(\tau) - \phi(\tau')\right)\right)} \ , \tag{11}$$

and

$$Y_{\overline{D}}^{\omega} = \nabla_\omega(X_{\overline{D}}^{\omega} + 2\lambda\omega) \tag{12}$$

$$= \sum_{(\tau,\tau',o)\in\overline{D}} \frac{\exp(o\omega^\top\left(\phi(\tau) - \phi(\tau')\right)}{\left(1 + \exp(o\omega^\top\left(\phi(\tau) - \phi(\tau')\right)\right)^2}\left(\phi(\tau) - \phi(\tau')\right)\left(\phi(\tau) - \phi(\tau')\right)^\top + 2\lambda I \ . \tag{13}$$

Note that

$$\begin{aligned}
\left\|X_{\overline{D}}^{\omega}\right\| &= \left\| \sum_{(\tau,\tau',o)\in\overline{D}} \frac{-o\left(\phi(\tau) - \phi(\tau')\right)}{1 + \exp\left(o\omega^\top\left(\phi(\tau) - \phi(\tau')\right)\right)} \right\| \\
&\leq \sum_{(\tau,\tau',o)\in\overline{D}} \frac{1}{1 + \exp\left(o\omega^\top\left(\phi(\tau) - \phi(\tau')\right)\right)} \left\|\phi(\tau) - \phi(\tau')\right\| &(14)\\
&\leq \sum_{(\tau,\tau',o)\in\overline{D}} \left(\|\phi(\tau)\| + \|\phi(\tau')\|\right) &(15)\\
&\leq \frac{2\overline{n}}{(1-\gamma)} \ , &(16)
\end{aligned}$$

where Equation 14 follows from the triangle inequality, while Equation 15 uses that fact that

$$\left\|\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)\right\| \leq \sum_{t=0}^{\infty} \gamma^t \left\|\phi(s_t, a_t)\right\| \leq \frac{1}{1-\gamma} \ ,$$

by assumption, and the fact that $1/(1 + \exp(x)) \leq 1$. Also, note that, for strictly positive λ, the matrix $Y_{\overline{D}}^{\omega}$ is symmetric positive definite. To obtain further information about the spectrum of $Y_{\overline{D}}^{\omega}$, which we will use for solving our problem, we prove the following intermediate result.

Going back to our optimization problem and using the above notation and result, the Lagrangian can be written as

$$\mathcal{L}(\omega, \alpha) = \left|\omega^\top\left(X_{\overline{D}}^{\omega} + \lambda\omega\right)\right| + (\boldsymbol{\epsilon} - M_{\pi^\dagger}^\top\omega)^\top\alpha \ ,$$

and its gradient is

$$\begin{aligned}
\nabla_\omega\mathcal{L}(\omega, \alpha) &= e_\omega \cdot \left(X_{\overline{D}}^{\omega} + \lambda\omega + \nabla_\omega(X_{\overline{D}}^{\omega} + \lambda\omega)\omega\right) - M_{\pi^\dagger}\alpha \\
&= e_\omega \cdot \left(X_{\overline{D}}^{\omega} + Y_{\overline{D}}^{\omega}\omega\right) - M_{\pi^\dagger}\alpha \ ,
\end{aligned}$$

where

$$e_\omega = \operatorname{sgn}\left(\omega^\top\left(X_{\overline{D}}^{\omega} + \lambda\omega\right)\right)$$

is the sign of the quantity inside the brackets. Thus, the first-order condition implies

$$\omega^\dagger = (Y_{\overline{D}}^{\omega^\dagger})^{-1}\left(e_\omega \cdot M_{\pi^\dagger}\alpha - X_{\overline{D}}^{\omega^\dagger}\right) \ .$$

Complementary slackness implies that

$$\text{either } \; M_{\pi^\dagger}^\top \omega^\dagger = \boldsymbol{\epsilon}, \;\; \text{or } \; \alpha = \mathbf{0} \; .$$

We consider the first case, since the goal of the attacker is to be able to enforce $\pi^\dagger$, thus we cannot have $\alpha = 0$ since this would disregard the constraint altogether. If $M_{\pi^\dagger}^\top \omega^\dagger = \boldsymbol{\epsilon}$, then this, together with the first-order condition, imply

$$M_{\pi^\dagger}^\top (Y_{\overline{D}}^{\omega^\dagger})^{-1} \left(e_\omega \cdot M_{\pi^\dagger} \alpha - X_{\overline{D}}^{\omega^\dagger} \right) = \boldsymbol{\epsilon} \; .$$

Using this and Cauchy-Schwarz, we have

$$\begin{aligned} \|M_{\pi^\dagger}\| \left\| (Y_{\overline{D}}^{\omega^\dagger})^{-1} \left(e \cdot M_{\pi^\dagger} \alpha - X_{\overline{D}}^{\omega^\dagger} \right) \right\| &\geq \left\| M_{\pi^\dagger}^\top (Y_{\overline{D}}^{\omega^\dagger})^{-1} \left(e_\omega \cdot M_{\pi^\dagger} \alpha - X_{\overline{D}}^{\omega^\dagger} \right) \right\| \\ &= \sqrt{SA}\epsilon \end{aligned}$$

which yields

$$\left\| \omega^\dagger \right\| = \left\| (Y_{\overline{D}}^{\omega^\dagger})^{-1} \left(e \cdot M_{\pi^\dagger} \alpha - X_{\overline{D}}^{\omega^\dagger} \right) \right\| \geq \frac{\sqrt{SA}\epsilon}{\|M_{\pi^\dagger}\|} \geq \frac{\sqrt{SA}\epsilon}{\sigma_{\max}\left(M_{\pi^\dagger}\right)} \; . \tag{17}$$

On the other hand, using again Cauchy-Schwarz on the complementary slackness condition, we get

$$\left\| \omega^\dagger \right\| \leq \frac{\sqrt{SA}\epsilon}{\sigma_{\min}(M_{\pi^\dagger})} \; .$$

Next, we will provide an upper bound on the norm of the gradient of $\omega^\dagger$ with respect to $\overline{D}$. Let $\overline{\omega}$ be the optimal point with respect to $\overline{D}$ and let $\text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega})$ denote the projection of $\overline{\omega}$ onto the ϵ-robust optimality polytope. To derive the upper bound, we use the fact that the projection of $\overline{\omega}$ onto the polytope is a feasible solution and thus provides an immediate upper bound. Note that

$$\begin{aligned} \left| \left(\omega^\dagger\right)^\top \left(X_{\overline{D}}^{\omega^\dagger} + \lambda \omega^\dagger \right) \right| &= \left| \left(\omega^\dagger\right)^\top \nabla_\omega \ell_{\text{RLHF}}^{\omega^\dagger}(\overline{D}) \right| \\ &\leq \left| \left(\text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega}) \right)^\top \nabla_\omega \ell_{\text{RLHF}}^{\text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega})}(\overline{D}) \right| \\ &\leq \left\| \text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega}) \right\| \left\| \nabla_\omega \ell_{\text{RLHF}}^{\text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega})}(\overline{D}) \right\| \\ &\leq (2\overline{n} + \lambda) \left\| \text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega}) \right\| \left\| \text{proj}_{\omega : M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon}} (\overline{\omega}) - \overline{\omega} \right\| \; , \end{aligned} \tag{18}$$

where the second inequality uses Cauchy-Schwarz and the third inequality uses the fact that the loss is Lipschitz (see Lemma E.14). Now, to compute the projection, we solve the following problem:

$$\min_{\omega} \frac{1}{2} \left\| \omega - \overline{\omega} \right\|^2 \;\; \text{such that } \; M_{\pi^\dagger} \omega \geq \boldsymbol{\epsilon} \; .$$

The Lagrangian of this problem is

$$\mathcal{L}(\omega, \nu) = \frac{1}{2} \left\| \omega - \overline{\omega} \right\|^2 - \nu \left(M_{\pi^\dagger} \omega - \boldsymbol{\epsilon} \right) \; ,$$

and the first order condition becomes

$$\omega - \overline{\omega} - M_{\pi^\dagger} \nu = \mathbf{0} \; ,$$

which gives us

$$\omega = \overline{\omega} + M_{\pi^\dagger} \nu \; .$$

To compute ν we use complementary slackness to obtain

$$M_{\pi^\dagger}\left(\overline{\omega} + M_{\pi^\dagger}\nu\right) = \boldsymbol{\epsilon}\ ,$$

which in turn gives us

$$\nu = \left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^+ \left(\boldsymbol{\epsilon} - M_{\pi^\dagger}\overline{\omega}\right)\ .$$

Hence, we get

$$\mathrm{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}}\left(\overline{\omega}\right) = \overline{\omega} + M_{\pi^\dagger}\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^+ \left(\boldsymbol{\epsilon} - M_{\pi^\dagger}\overline{\omega}\right)\ . \tag{19}$$

We can upper bound the projection norm by using complementary slackness as:

$$\epsilon\sqrt{SA} = \left\|M_{\pi^\dagger}\left(\overline{\omega} + M_{\pi^\dagger}\nu\right)\right\| \leq \left\|M_{\pi^\dagger}\right\| \left\|\overline{\omega} + M_{\pi^\dagger}\nu\right\| \leq \sigma_{\max}(M_{\pi^\dagger}) \left\|\mathrm{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}}\left(\overline{\omega}\right)\right\|\ ,$$

and

$$\epsilon\sqrt{SA} = \left\|M_{\pi^\dagger}\left(\overline{\omega} + M_{\pi^\dagger}\nu\right)\right\| \geq \sigma_{\min}\left(M_{\pi^\dagger}\right) \left\|\overline{\omega} + M_{\pi^\dagger}\nu\right\| = \sigma_{\min}\left(M_{\pi^\dagger}\right) \left\|\mathrm{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}}\left(\overline{\omega}\right)\right\|\ ,$$

where we use the fact that $\sigma_{\min}(M)\left\|x\right\| \leq \left\|Mx\right\| \leq \sigma_{\max}(M)\left\|x\right\|$. Using the overall case when $\alpha = 0$ and Equation 18, we have

$$\begin{aligned}
\left|\left(\omega^\dagger\right)^\top \left(X\frac{\omega^\dagger}{D} + \lambda\omega^\dagger\right)\right| &\leq (2\overline{n} + \lambda) \left\|\mathrm{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}}\left(\overline{\omega}\right)\right\| \left\|\mathrm{proj}_{\omega: M_{\pi^\dagger}\omega \geq \boldsymbol{\epsilon}}\left(\overline{\omega}\right) - \overline{\omega}\right\| \\
&\leq (2\overline{n} + \lambda)\frac{\epsilon\sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})}\left(\frac{\epsilon\sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})} + \left\|\overline{\omega}\right\|\right) \\
&\leq (2\overline{n} + \lambda)\left(\frac{\epsilon^2 SA}{\sigma_{\min}^2\left(M_{\pi^\dagger}\right)} + \left\|\overline{\omega}\right\|\frac{\epsilon\sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})}\right)\ .
\end{aligned}$$

Putting things together, we finally obtain

$$\widehat{n}_{\mathrm{RLHF}} \leq \left\lceil \frac{2\overline{n} + \lambda}{\xi_{\max}}\left(\frac{\epsilon^2 SA}{\sigma_{\min}^2\left(M_{\pi^\dagger}\right)} + \left\|\overline{\omega}\right\|\frac{\epsilon\sqrt{SA}}{\sigma_{\min}(M_{\pi^\dagger})}\right)\right\rceil$$

□

B.2 Unregularized RLHF with Empty Existing Data.

Next, we consider the unregularized setting when the pre-existing dataset $\overline{D}$ is empty.

Corollary 4.1. *Let $\overline{D} = \emptyset$, $\beta = 0$ and $\epsilon' > 0$, and let $\pi^\dagger \in \Pi^{\mathrm{det}}$. Define $\omega^\dagger = M_{\pi^\dagger}\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)^+ \boldsymbol{\epsilon}'$. Then, the dataset of $\left\lceil \frac{\lambda\left\|\omega^\dagger\right\|^2}{\xi_{\max}} \right\rceil$ identical samples satisfying*

$$\phi(\tau) - \phi(\tau') = \xi_1\left(\lambda\left\|\omega^\dagger\right\|\right) \cdot \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2}, \quad o = 1$$

is a feasible solution for Problem P:Attack:RLHF.1. Furthermore, there exists an optimal solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\mathrm{RLHF}}$ samples such that

$$\widehat{n}_{\mathrm{RLHF}} \leq \left\lceil \frac{(\epsilon')^2\lambda SA}{\xi_{\max}\sigma_{\min}^2\left(M_{\pi^\dagger}\right)} \right\rceil\ .$$

Proof. Lemma E.3 implies that any feasible solution to Problem P:Attack:RLHF.2 is feasible for Problem P:Attack:RLHF.1. Thus, we focus on Problem P:Attack:RLHF.2.

First, note that, for $\beta = 0$ the optimal policy is deterministic. Now, given $\omega^\dagger \neq \mathbf{0}$, Lemma E.6 implies that the solution to the problem

$$\begin{aligned}&\min_D |D| \\ &\text{s.t. } \omega^\dagger = \arg\min_\omega \sum_{(\tau,\tau',o)\in D} \log\left(1+\exp\left(-o\cdot\omega^\top\left(\phi(\tau)-\phi(\tau')\right)\right)\right) + \frac{\lambda}{2}\left\|\omega\right\|^2\end{aligned}$$

is the dataset of

$$\left\lceil \frac{\lambda \left\|\omega^\dagger\right\|^2}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau_i) - \phi(\tau_i') = \xi^{-1}\left(\lambda\left\|\omega^\dagger\right\|^2 \left\lceil \frac{\lambda \left\|\omega^\dagger\right\|^2}{\xi_{\max}} \right\rceil^{-1}\right) \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2},\ o_i = 1\ .$$

Since the optimal sample size to solve the logistic regression subproblem depends on the norm of the parameter $\omega^\dagger$, then, by using the construction above we can directly minimize $\left\|\omega^\dagger\right\|$ and constrain $\omega^\dagger$ to remain in the desired region. In this case, we can equivalently rewrite our original problem in terms of ω as

$$\begin{aligned}&\min_\omega \frac{1}{2}\left\|\omega\right\|^2 \\ &\text{s.t. } \left(\sum_{s'} d_\rho^{\pi^\dagger}(s')\phi(s',\pi^\dagger(s')) - \sum_{s'} d_\rho^{\pi^\dagger\{s,a\}}(s')\phi(s',\pi^\dagger\{s,a\}(s'))\right)^\top \omega \geq \epsilon,\ \forall s, a \neq \pi^\dagger(s)\ .\end{aligned}$$

Using vector notation, we have

$$\begin{aligned}&\min_\omega \frac{1}{2}\left\|\omega\right\|^2 \\ &\text{s.t. } \boldsymbol{\epsilon} - M_{\pi^\dagger}^\top \omega \leq \mathbf{0}\ ,\end{aligned}$$

where $\boldsymbol{\epsilon} = \epsilon\mathbf{1}$. This is a convex program and thus local minima are global. The Lagrangian of the above is

$$\mathcal{L}(\omega,\alpha) = \frac{1}{2}\left\|\omega\right\|^2 + \alpha^\top\left(\boldsymbol{\epsilon} - M_{\pi^\dagger}^\top\omega\right)\ ,$$

and setting its gradient to zero gives us

$$\omega^\dagger = M_{\pi^\dagger}\alpha\ .$$

Complementary slackness implies

$$M_{\pi^\dagger}^\top\omega^\dagger = \boldsymbol{\epsilon}\ ,$$

which, together with the first-order condition, imply

$$\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)\alpha = \boldsymbol{\epsilon}\ .$$

Note that

$$\left\|M_{\pi^\dagger}^\top M_{\pi^\dagger}\right\| = \left\|M_{\pi^\dagger} M_{\pi^\dagger}^\top\right\| \leq Tr\left(M_{\pi^\dagger} M_{\pi^\dagger}^\top\right) \leq 2SA\ ,$$

due to the fact that

$$\left\|\sum_{s'} d_\rho^{\pi^\dagger}(s')\phi(s',\pi^\dagger(s')) - \sum_{s'} d_\rho^{\pi^\dagger\{s,a\}}(s')\phi(s',\pi^\dagger\{s,a\}(s'))\right\|$$

$$
\begin{aligned}
&\leq \left\| \sum_{s'} d_\rho^{\pi^\dagger}(s')\phi(s', \pi^\dagger(s')) \right\| + \left\| \sum_{s'} d_\rho^{\pi^\dagger\{s,a\}}(s')\phi(s', \pi^\dagger\{s,a\}(s')) \right\| \\
&\leq \sum_{s'} d_\rho^{\pi^\dagger}(s') \left\| \phi(s', \pi^\dagger(s')) \right\| + \sum_{s'} d_\rho^{\pi^\dagger\{s,a\}}(s') \left\| \phi(s', \pi^\dagger\{s,a\}(s')) \right\| \\
&\leq 2 \ ,
\end{aligned}
$$

for every (s, a) pair. Thus, it follows that

$$
\left\|\omega^\dagger\right\| \left\|M_{\pi^\dagger}\right\| \geq \left\|M_{\pi^\dagger}\omega^\dagger\right\| = \left\|\left(M_{\pi^\dagger}^\top M_{\pi^\dagger}\right)\alpha\right\| = \|\boldsymbol{\epsilon}\| = \epsilon\sqrt{SA} \ ,
$$

and

$$
\sigma_{\min}\left(M_{\pi^\dagger}\right)\left\|\omega^\dagger\right\| \leq \left\|M_{\pi^\dagger}\omega^\dagger\right\| = \epsilon\sqrt{SA} \ .
$$

From this, we have

$$
\frac{\epsilon\sqrt{SA}}{\sigma_{\max}\left(M_{\pi^\dagger}\right)} \leq \left\|\omega^\dagger\right\| \leq \frac{\epsilon\sqrt{SA}}{\sigma_{\min}\left(M_{\pi^\dagger}\right)} \ ,
$$

which implies that

$$
\widehat{n}_{\mathrm{RLHF}} \leq \left\lceil \frac{\epsilon^2 \lambda SA}{\xi_{\max}\sigma_{\min}^2\left(M_{\pi^\dagger}\right)} \right\rceil \ .
$$

□

B.3 Regularized RLHF with Non-empty Existing Data

In this section, we provide the full proof of the following result for the regularized setting.

Theorem 4.2. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, $\beta > 0$ and $0 < \epsilon' \leq \epsilon$. Moreover, define*

$$
\Gamma_\phi^\omega(\pi^\dagger || \pi_{r_\omega}^{\mathrm{reg}}) = \sum_{s,a} \rho(s)(\pi^\dagger(a|s) - \pi_{r_\omega}^{\mathrm{reg}}(a|s))\phi^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a) \ ,
$$

for any given ω. Then, there exists a feasible solution $\widehat{D}$ for Problem P:Attack:RLHF.1 with $\widehat{n}_{\mathrm{RLHF}}$ samples which yields $\omega^\dagger$ when solving Problem P:RLHF.Reward on dataset $\widehat{D}$, such that $\Gamma_\phi^\omega(\pi^\dagger || \pi_{r_{\omega^\dagger}}^{\mathrm{reg}}) \neq 0$ and

$$
\widehat{n}_{\mathrm{RLHF}} \leq O\left(\frac{\beta^2 \left(D_{\mathrm{KL}}\left(\pi^\dagger || \mu\right) - \epsilon'\right)^2}{(1-\gamma)^2\sigma_{\min}^2(\Sigma_{\overline{D}}^\phi)\left\|\Gamma_\phi^{\omega^\dagger}(\pi^\dagger || \pi_{r_{\omega^\dagger}}^{\mathrm{reg}})\right\|^2} + \frac{\overline{n}}{(1-\gamma)^4\sigma_{\min}^4(\Sigma_{\overline{D}}^\phi)} \right) \ .
$$

Proof. Let $\epsilon' \leq \epsilon$. Lemma E.4 implies that any feasible solution to Problem P:Attack:RLHF.3 is feasible for Problem P:Attack:RLHF.1. Thus, we focus on Problem P:Attack:RLHF.3.

First, let us recall the KL-regularized objective in this setting. Given learned reward r, the objective is

$$
\max_\pi \mathcal{V}_r^\pi(\rho) := \mathbb{E}_{s\sim\rho,\pi}\left[\sum_{t=0}^\infty \gamma^t \left(r(s_t, a_t) - \beta \log \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \right) \Big| s_0 = s \right] \ .
$$

Let us use the shorthand notation $\mathcal{V}^r(s)$, $\mathcal{Q}^r(s,a)$ and $\mathcal{A}^r(s,a)$ to denote the functions $\mathcal{V}_r^{\pi_r^{\mathrm{reg}}}(s)$, $\mathcal{Q}_r^{\pi_r^{\mathrm{reg}}}(s,a)$ and $\mathcal{A}_r^{\pi_r^{\mathrm{reg}}}(s,a)$, respectively. Similarly, for parametrized reward function r_ω, Let us use the short-hand notation $\mathcal{V}^\omega(s), \mathcal{Q}^\omega(s,a), \mathcal{A}^\omega(s,a)$ to denote $\mathcal{V}_{r_\omega}^{\pi_{r_\omega}^{\mathrm{reg}}}(s)$, $\mathcal{Q}_{r_\omega}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a)$ and $\mathcal{A}_{r_\omega}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a)$, respectively.

Given reward r, Lemma E.1 implies that the regularized optimal policy in this case can be written as

$$\pi_r^{\text{reg}}(a|s) = \mu(a|s) \exp\left(\frac{1}{\beta}\mathcal{Q}^r(s,a) - \frac{1}{\beta}\mathcal{V}^r(s)\right) ,$$

with $\exp(\mathcal{V}^r(s)/\beta)$ as the partition constant, since it is action-independent. This implies the following relation between the two value functions holds:

$$\mathcal{V}^r(s) = \beta \log \sum_a \mu(a|s) \exp\left(\frac{1}{\beta}\mathcal{Q}^r(s,a)\right) . \tag{20}$$

Now, note that the second constraint of Problem P:Attack:RLHF.3 can be written as

$$\begin{aligned} D_{\text{KL}}\left(\pi^\dagger || \pi^{\text{reg}}_{r_{\omega^\dagger}}\right) &= \sum_{s,a} \rho(s)\pi^\dagger(a|s) \log \frac{\pi^\dagger(a|s)}{\pi^{\text{reg}}_{r_{\omega^\dagger}}(a|s)} \\ &= \sum_{s,a} \rho(s)\pi^\dagger(a|s) \left(\log \frac{\pi^\dagger(a|s)}{\mu(a|s)} - \frac{1}{\beta}\mathcal{Q}^{\omega^\dagger}(s,a) + \frac{1}{\beta}\mathcal{V}^{\omega^\dagger}(s)\right) \\ &= D_{\text{KL}}(\pi^\dagger||\mu) - \frac{1}{\beta}\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^{\omega^\dagger}(s,a)\right] . \end{aligned}$$

Thus, the attack optimization problem can be written as

$$\begin{aligned} &\min_D |D| \\ &\text{s.t } \omega^\dagger = \arg\min_\omega \sum_{(\tau,\tau',o)\in\overline{D}\cup D} \log\left(1+\exp\left(-o\cdot(\phi(\tau)-\phi(\tau'))\right)\right) + \frac{\lambda}{2}\|\omega\|^2 \\ &\quad D_{\text{KL}}(\pi^\dagger||\mu) - \frac{1}{\beta}\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^{\omega^\dagger}(s,a)\right] - \epsilon' \le 0 . \end{aligned}$$

Lemma E.6 implies that the solution to the subproblem

$$\begin{aligned} &\min_D |D| \\ &\text{s.t } \omega^\dagger = \arg\min_\omega \sum_{(\tau,\tau',o)\in\overline{D}\cup D} \log\left(1+\exp\left(-o\cdot(\phi(\tau)-\phi(\tau'))\right)\right) + \frac{\lambda}{2}\|\omega\|^2 \end{aligned}$$

is the dataset of

$$\left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau) - \phi(\tau') = \xi^{-1}\left(\frac{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})}{\left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil}\right) \frac{\omega^\dagger}{\|\omega^\dagger\|^2}, \quad o = 1 .$$

Thus, we can equivalently write the original problem in terms of ω as

$$\begin{aligned} \min_\omega \quad & \left|\omega^\top \nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})\right| \\ \text{s.t.} \quad & D_{\text{KL}}(\pi^\dagger||\mu) - \frac{1}{\beta}\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^{\omega}(s,a)\right] - \epsilon' \le 0 , \end{aligned}$$

with Lagrangian

$$\mathcal{L}(\omega,\alpha) = \left|\omega^\top \nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})\right| + \alpha\left(D_{\text{KL}}(\pi^\dagger||\mu) - \frac{1}{\beta}\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^{\omega}(s,a)\right] - \epsilon'\right) .$$

Before we consider the first-order conditions of the problem, we rewrite the expected advantage function in the constraint and derive its gradient below.

Lemma B.1. *Given any $\omega \in \mathbb{R}^d$, let us define*

$$\Gamma_\phi^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t,a_t) | a_0 = a, s_0 = s, \pi\right] , \tag{21}$$

for any policy π and feature mapping ϕ, and let

$$\Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}\left(\pi^\dagger || \pi_{r_\omega}^{\text{reg}}\right) = \mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}(s,a)\right] - \mathbb{E}_{s\sim\rho, a\sim\pi_{r_\omega}^{\text{reg}}(\cdot|s)}\left[\Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}(s,a)\right] . \tag{22}$$

Then, we have

$$\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^\omega(s,a)\right] = \omega^\top \Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}\left(\pi^\dagger || \pi_{r_\omega}^{\text{reg}}\right) ,$$

and

$$\nabla_\omega \mathcal{A}^\omega(s,a) = \Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}(s,a) - \sum_{a'} \pi_{r_\omega}^{\text{reg}}(a'|s) \Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}(s,a') .$$

Proof. We start by deriving the gradient of the action-value function. Given reward function r, policy π and trajectory τ, let

$$\mathbb{P}_s^\pi(\tau) = \pi(a_0|s) P(s,a_0,s_1) \pi(a_1|s_1) P(s_1,a_1,s_2) \pi(a_2|s_2) \dots$$

and

$$\mathbb{P}_{s,a}^\pi(\tau) = P(s,a,s_1) \pi(a_1|s_1) P(s_1,a_1,s_2) \pi(a_2|s_2) \dots .$$

Now, for any given state-action pair (s,a) and parameter ω, observe that

$$\begin{aligned}
\nabla_\omega \mathcal{Q}^\omega(s,a) &= \nabla_\omega \left(r_\omega(s,a) + \gamma \sum_{s'} P(s,a,s') \mathcal{V}^\omega(s') \right) \\
&= \phi(s,a) + \gamma \sum_{s'} P(s,a,s') \beta \nabla_\omega \log \sum_{a'} \mu(a'|s') \exp\left(\frac{1}{\beta} \mathcal{Q}^\omega(s',a')\right) \qquad (23) \\
&= \phi(s,a) + \gamma\beta \sum_{s'} P(s,a,s') \frac{1}{\sum_{a''} \mu(a''|s') \exp\left(\frac{1}{\beta}\mathcal{Q}^\omega(s',a'')\right)} \sum_{a'} \mu(a'|s') \exp\left(\frac{1}{\beta}\mathcal{Q}^\omega(s',a')\right) \\
&\quad \cdot \frac{1}{\beta} \nabla_\omega \mathcal{Q}^\omega(s',a') \\
&= \phi(s,a) + \gamma \sum_{s'} P(s,a,a') \sum_{a'} \frac{\mu(a'|s') \exp\left(\frac{1}{\beta}\mathcal{Q}^\omega(s',a')\right)}{\exp\left(\frac{1}{\beta}\mathcal{V}^\omega(s')\right)} \nabla_\omega \mathcal{Q}^\omega(s',a') \qquad (24) \\
&= \phi(s,a) + \gamma \sum_{s',a'} P(s,a,s') \pi_{r_\omega}^{\text{reg}}(a'|s') \nabla_\omega \mathcal{Q}^\omega(s',a') \\
&= \mathbb{E}_{\tau\sim\mathbb{P}_{s,a}^{\pi_{r_\omega}^{\text{reg}}}}\left[\phi(s,a)\right] + \gamma \mathbb{E}_{\tau\sim\mathbb{P}_{s,a}^{\pi_{r_\omega}^{\text{reg}}}}\left[\phi(s',a')\right] + \dots \qquad (25) \\
&= \mathbb{E}_{\tau\sim\mathbb{P}_{s,a}^{\pi_{r_\omega}^{\text{reg}}}}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t,a_t)\right] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}_{s,a}^{\pi_{r_\omega}^{\text{reg}}}}\left[\phi(\tau)\right] \\
&= \Gamma_\phi^{\pi_{r_\omega}^{\text{reg}}}(s,a) , \qquad (26)
\end{aligned}$$

where Equation 23 and Equation 24 follow from Equation 20, while Equation 25 follows from recursion. Similarly, for the gradient of the advantage function, we have

$$\begin{aligned}
\nabla_\omega \mathcal{A}^\omega(s,a) &= \nabla_\omega \left(\mathcal{Q}^\omega(s,a) - \mathcal{V}^\omega(s)\right) \\
&= \mathbb{E}_{\tau\sim\mathbb{P}_{s,a}^{\pi_{r_\omega}^{\text{reg}}}}\left[\phi(\tau)\right] - \nabla_\omega \beta \log \sum_{a'} \mu(a'|s) \exp\left(\frac{1}{\beta}\mathcal{Q}^\omega(s,a')\right) \frac{1}{\beta} \mathbb{E}_{\tau\sim\mathbb{P}_{s,a'}^{\pi_{r_\omega}^{\text{reg}}}}\left[\phi(\tau)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\tau \sim \mathbb{P}_{s,a}^{\pi_{r_\omega}^{\mathrm{reg}}}} [\phi(\tau)] - \mathbb{E}_{\tau \sim \mathbb{P}_{s}^{\pi_{r_\omega}^{\mathrm{reg}}}} [\phi(\tau)] \\
&= \Gamma_{\phi}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a) - \sum_{a'} \pi_{r_\omega}^{\mathrm{reg}}(a'|s) \Gamma_{\phi}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a') \ ,
\end{aligned}
$$

where, for the last equality, we have used the definition of the trajectory feature as a discounted sum of state-action feature vectors and rewritten the expectation. Now, let us use the short-hand notation $\Gamma_{\phi}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a)$ by $\Gamma_{\phi}^{\omega}(s,a)$, following our earlier convention, for brevity. For the final statement of our result, we rewrite the expected regularized advantage function with respect to $\pi^\dagger$ as

$$
\begin{aligned}
&\mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)} \left[\mathcal{A}^\omega(s,a)\right] = \mathbb{E}_{s\sim\rho, a\sim\pi^\dagger(\cdot|s)} \left[\mathcal{Q}^\omega(s,a) - \sum_{a'} \pi_{r_\omega}^{\mathrm{reg}}(a'|s) \mathcal{Q}^\omega(s,a')\right] \\
&= \sum_{s,a} \rho(s) \left(\pi^\dagger(s|a) - \pi_{r_\omega}^{\mathrm{reg}}(a|s)\right) \mathcal{Q}^\omega(s,a) \\
&= \sum_{s,a} \rho(s) \left(\pi^\dagger(s|a) - \pi_{r_\omega}^{\mathrm{reg}}(a|s)\right) \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \left(\omega^\top \phi(s_t,a_t) - \beta \log \frac{\pi_{r_\omega}^{\mathrm{reg}}(a_t|s_t)}{\mu(a_t|s_t)}\right) \Big| s_0 = s, a_0 = a, \pi_{r_\omega}^{\mathrm{reg}}\right] \\
&= \sum_{s,a} \rho(s) \left(\pi^\dagger(s|a) - \pi_{r_\omega}^{\mathrm{reg}}(a|s)\right) \\
&\qquad \cdot \left(\omega^\top \Gamma_{\phi}^{\omega}(s,a) - \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \beta \log \frac{\mu(a_t|s_t) \exp\left(\frac{1}{\beta}\mathcal{A}^\omega(s_t,a_t)\right)}{\mu(a_t|s_t)} \Big| s_0 = s, a_0 = a, \pi_{r_\omega}^{\mathrm{reg}}\right]\right) \\
&= \sum_{s,a} \rho(s) \left(\pi^\dagger(s|a) - \pi_{r_\omega}^{\mathrm{reg}}(a|s)\right) \left(\omega^\top \Gamma_{\phi}^{\omega}(s,a) - \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \mathcal{A}^\omega(s_t,a_t) \Big| s_0 = s, a_0 = a, \pi_{r_\omega}^{\mathrm{reg}}\right]\right) \\
&= \omega^\top \Gamma_{\phi}^{\omega} \left(\pi^\dagger || \pi_{r_\omega}^{\mathrm{reg}}\right) \\
&\qquad - \sum_{s,a} \rho(s) \left(\pi^\dagger(s|a) - \pi_{r_\omega}^{\mathrm{reg}}(a|s)\right) \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \sum_{a} \pi_{r_\omega}^{\mathrm{reg}}(a|s_t) \mathcal{A}^\omega(s_t,a) \Big| s_0 = s, a_0 = a, \pi_{r_\omega}^{\mathrm{reg}}\right] \\
&= \omega^\top \Gamma_{\phi}^{\omega} \left(\pi^\dagger || \pi_{r_\omega}^{\mathrm{reg}}\right) \ , \qquad (27)
\end{aligned}
$$

where the Equation 27 follows from the fact that

$$
\sum_a \pi(a|s) \mathcal{A}^\pi(s,a) = \sum_a \pi(a|s) \mathcal{Q}^\pi(s,a) - \sum_a \pi(a|s) \mathcal{Q}^\pi(s,a) = 0 \ .
$$

□

Now, let us return to the proof of our main result. First, note that Lemma B.1 implies that the constraint can be written as

$$
D_{\mathrm{KL}}(\pi^\dagger || \mu) - \frac{1}{\beta} \omega^\top \Gamma_{\phi}^{\omega}(\pi^\dagger || \pi_{r_\omega}^{\mathrm{reg}}) - \epsilon' \leq 0 \ .
$$

As in the proof of Lemma B.1, let us denote $\Gamma_{\phi}^{\pi_{r_\omega}^{\mathrm{reg}}}(s,a)$ by $\Gamma_{\phi}^{\omega}(s,a)$. Moreover, let us reuse some definitions from the proof of Theorem 4.1. Recall that

$$
X_{\overline{D}}^{\omega} = \sum_{(\tau,\tau',o)\in\overline{D}} \frac{-o\left(\overline{\phi}(\tau) - \overline{\phi}(\tau')\right)}{1 + \exp\left(o\omega^\top \left(\overline{\phi}(\tau) - \overline{\phi}(\tau')\right)\right)} \ ,
$$

and

$$
Y_{\overline{D}}^{\omega} = \nabla_\omega (X_{\overline{D}}^{\omega} + 2\lambda\omega)
$$

$$= \sum_{(\tau,\tau',o)\in\overline{D}} \frac{\exp(o\omega^\top\left(\overline{\phi}(\tau)-\overline{\phi}(\tau')\right)}{\left(1+\exp(o\omega^\top\left(\overline{\phi}(\tau)-\overline{\phi}(\tau')\right)\right)^2}\left(\overline{\phi}(\tau)-\overline{\phi}(\tau')\right)\left(\overline{\phi}(\tau)-\overline{\phi}(\tau')\right)^\top + 2\lambda I \ .$$

Then, we can equivalently write the Lagrangian as

$$\mathcal{L}(\omega,\alpha) = \left|\omega^\top\left(X^\omega_{\overline{D}}+\lambda\omega\right)\right| + \alpha\left(D_{\mathrm{KL}}(\pi^\dagger||\mu) - \frac{1}{\beta}\mathbb{E}_{s\sim\rho,a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^\omega(s,a)\right]-\epsilon'\right)$$

Using this notation, the first-order condition then implies

$$\nabla_\omega\mathcal{L}(\omega,\alpha) = e_\omega\cdot\left(X^\omega_{\overline{D}}+Y^\omega_{\overline{D}}\omega\right) - \frac{\alpha}{\beta}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right) = \mathbf{0},$$

where

$$e_\omega = \mathrm{sgn}\left(\omega^\top\left(X^\omega_{\overline{D}}+\lambda\omega\right)\right)$$

denotes the sign of the quantity inside the brackets. In the above, we have used the fact that

$$\mathbb{E}_{s\sim\rho,a\sim\pi^\dagger(\cdot|s)}\left[\mathcal{A}^\omega(s,a)\right] = \omega^\top\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right) \ ,$$

from Lemma B.1. Assuming $\alpha\neq 0$, we can write the primal solution in terms of the fixed-point equation

$$\omega = \left(Y^\omega_{\overline{D}}\right)^{-1}\left(\frac{e_\omega\alpha}{\beta}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right) - X^\omega_{\overline{D}}\right) \ . \tag{28}$$

Note that, if $\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}\right)=\mathbf{0}$, at the solution of the fixed-point Equation 28, then the primal condition for this case reduces to the following solution, for any α:

$$\widehat{\omega} = -\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}X^{\widehat{\omega}}_{\overline{D}} \ ,$$

in which case we obtain

$$\begin{aligned}
\widehat{n}_{\mathrm{RLHF}} = \left|(\widehat{\omega})^\top\nabla_\omega\ell^\omega_{\mathrm{RLHF}}(\overline{D})\right| &= \left|\left(X^{\widehat{\omega}}_{\overline{D}}+\lambda\widehat{\omega}\right)^\top\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}X^{\widehat{\omega}}_{\overline{D}}\right| \\
&= \left|\left(X^{\widehat{\omega}}_{\overline{D}}-\lambda\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}X^{\widehat{\omega}}_{\overline{D}}\right)^\top\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}X^{\widehat{\omega}}_{\overline{D}}\right| \\
&\leq \left\|X^{\widehat{\omega}}_{\overline{D}}\right\|^2\left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|\left\|I-\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\| \\
&\leq \left(\frac{2\overline{n}}{1-\gamma}\right)^2\frac{1}{\overline{n}\sigma_{\min}(\Sigma^\phi_{\overline{D}})+2\lambda} \qquad (29) \\
&\leq O\left(\frac{\overline{n}}{(1-\gamma)^2\sigma_{\min}\left(\Sigma^\phi_{\overline{D}}\right)}\right) \ , \qquad (30)
\end{aligned}$$

where Equation 29 follows from Equation 16 and Lemma E.12. Having taken care of this case, let us now assume that the solution to the fixed point equation implies $\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}\right)\neq\mathbf{0}$. Using Lemma B.1, complementary slackness corresponding to the KL constraint can be written as

$$\omega^\top\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right) = \beta\left(D_{\mathrm{KL}}\left(\pi^\dagger||\mu\right)-\epsilon'\right) \ .$$

Using Equation 28 above and expanding, we obtain the following result:

$$\beta\left(D_{\mathrm{KL}}\left(\pi^\dagger||\mu\right)-\epsilon'\right) = \left(\left(Y^\omega_{\overline{D}}\right)^{-1}\left(\frac{e_\omega\alpha}{\beta}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right)-X^\omega_{\overline{D}}\right)\right)^\top\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_\omega}\right)$$

$$= \left(\frac{e_\omega \alpha}{\beta}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_\omega}\right) - \left(X^\omega_{\overline{D}}\right)\right)^\top \left(\left(Y^\omega_{\overline{D}}\right)^{-1}\right)^\top \Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_\omega}\right)$$
$$= \frac{e_\omega \alpha}{\beta}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_\omega}\right)^\top \left(Y^\omega_{\overline{D}}\right)^{-1}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_\omega}\right) - \left(X^\omega_{\overline{D}}\right)^\top \left(Y^\omega_{\overline{D}}\right)^{-1}\Gamma^\omega_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_\omega}\right) \ .$$

This further implies that

$$\alpha = e_\omega \cdot \frac{\beta^2\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right) + \beta\left(X^\omega_{\overline{D}}\right)^\top \left(Y^\omega_{\overline{D}}\right)^{-1}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}{\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)^\top \left(Y^\omega_{\overline{D}}\right)^{-1}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)} \ ,$$

where we have used the fact that $Y^\omega_{\overline{D}}$ is positive definite from Lemma E.12, so the denominator is well-defined for any non-zero $\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)$. Putting things together we get

$$\widehat{\omega} = \left(Y^\omega_{\overline{D}}\right)^{-1} \cdot \left(e_{\widehat{\omega}} \cdot \frac{\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right) + \left(X^\omega_{\overline{D}}\right)^\top \Gamma^{\widehat{\omega}}_{\phi,\overline{D}}\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}{\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)^\top \Gamma^{\widehat{\omega}}_{\phi,\overline{D}}\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) - X^\omega_{\overline{D}}\right) \ ,$$

where

$$\Gamma^{\widehat{\omega}}_{\phi,\overline{D}}\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) = \left(Y^\omega_{\overline{D}}\right)^{-1}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) \ .$$

Note that, as long as we choose γ so that

$$1 - \gamma \leq \frac{2\left\|\widehat{\omega}\right\|}{\xi_{\max} + 1} \ ,$$

where $\widehat{\omega}$ is the solution to the above fixed-point equation, then Lemma E.10 guarantees that the dataset construction formula satisfies the feature boundedness condition.

We will now derive upper bounds on the norm of the parameter $\widehat{\omega}$ as given by the fixed-point equation above – we will use this bound later for the final result. Observe that

$$\left\|\widehat{\omega}\right\| =$$
$$\left\|\left(Y^{\widehat{\omega}}_D\right)^{-1}\left(e_{\widehat{\omega}} \cdot \frac{\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right) + \left(X^{\widehat{\omega}}_{\overline{D}}\right)^\top \Gamma^{\widehat{\omega}}_{\phi,\overline{D}}\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}{\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)^\top \Gamma^{\widehat{\omega}}_{\phi,\overline{D}}\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) - X^\omega_{\overline{D}}\right)\right\|$$
$$\leq \left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|\left\|e_{\widehat{\omega}} \cdot \frac{\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right) + \left(X^{\widehat{\omega}}_{\overline{D}}\right)^\top \left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}{\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)^\top \left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)}\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right) - X^{\widehat{\omega}}_{\overline{D}}\right\| \tag{31}$$
$$\leq \left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|$$
$$\cdot \left(\frac{\left|\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right)\right| + \left\|X^{\widehat{\omega}}_{\overline{D}}\right\|\left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|\left\|\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)\right\|}{\left\|\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)\right\|^2 \sigma_{\min}\left(\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right)}\left\|\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)\right\| + \left\|X^{\widehat{\omega}}_{\overline{D}}\right\|\right) \tag{32}$$
$$= \frac{\left|\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right)\right|\left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|}{\left\|\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)\right\|\sigma_{\min}\left(\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right)} + \frac{\left\|X^{\widehat{\omega}}_{\overline{D}}\right\|\left\|(Y^{\widehat{\omega}}_{\overline{D}})^{-1}\right\|^2}{\sigma_{\min}\left(\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right)} + \left\|X^{\widehat{\omega}}_{\overline{D}}\right\|\left\|(Y^{\widehat{\omega}}_{\overline{D}})^{-1}\right\| \tag{33}$$
$$\leq \frac{\left|\beta\left(D_{\text{KL}}\left(\pi^\dagger||\mu\right) - \epsilon'\right)\right|\left\|\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right\|}{\left\|\Gamma^{\widehat{\omega}}_\phi\left(\pi^\dagger||\pi^{\text{reg}}_{r_{\widehat{\omega}}}\right)\right\|\sigma_{\min}\left(\left(Y^{\widehat{\omega}}_{\overline{D}}\right)^{-1}\right)} + \frac{2\overline{n}(\overline{n} + 2(1-\gamma)\lambda)^2}{(1-\gamma)^2(\overline{n}\widetilde{C}^\phi_{\overline{D}}C^\phi_{\overline{D}} + 2\lambda)^2} + \frac{2\overline{n}}{(1-\gamma)(\overline{n}\widetilde{C}^\phi_{\overline{D}}C^\phi_{\overline{D}} + 2\lambda)} \tag{34}$$

$$\leq \frac{\left|\beta\left(D_{\mathrm{KL}}\left(\pi^\dagger||\mu\right)-\epsilon'\right)\right|}{\left\|\Gamma_\phi^{\widehat{\omega}}\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}\right)\right\|} O\left(\frac{\overline{n}+2(1-\gamma)\lambda}{(1-\gamma)\left(\overline{n}\sigma_{\min}(\Sigma^\phi_{\overline{D}})+2\lambda\right)}\right)+O\left(\frac{\overline{n}}{\left((1-\gamma)\sigma_{\min}(\Sigma^\phi_{\overline{D}})\right)^2}\right) \tag{35}$$

$$\leq O\left(\frac{\left|\beta\left(D_{\mathrm{KL}}\left(\pi^\dagger||\mu\right)-\epsilon'\right)\right|}{(1-\gamma)\sigma_{\min}(\Sigma^\phi_{\overline{D}})\left\|\Gamma_\phi^{\widehat{\omega}}\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\widehat{\omega}}}\right)\right\|}+\frac{\overline{n}}{\left((1-\gamma)\sigma_{\min}(\Sigma^\phi_{\overline{D}})\right)^2}\right), \tag{36}$$

where Equation 31 follows from the Cauchy-Schwarz inequality; Equation 32 follows from the triangle inequality and Cauchy-Schwarz applied on the spectral norm; in Equation 33 we expand and cancel out equal terms; finally, for Equation 34 and Equation 35 we have used Equation 16, and Equation 51 and Equation 52 from the proof of Lemma E.12.

Now, in order to obtain upper bounds, we need to deal with the term in the denominator, which depends on the optimal solution $\widehat{\omega}$ itself. To address this, we will use a feasible solution to the problem.

Now, let $\widehat{\pi}^\dagger$ be a policy that is at most ϵ-close to $\pi^\dagger$ and that the solution $\omega^\dagger$ to $\Phi\omega = \beta\log(\widehat{\pi}^\dagger - \mu)$ yields $\Gamma^\omega_\phi(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\omega^\dagger}}) \neq 0$. Theorem A.1 shows that $\omega^\dagger$ makes $\widehat{\pi}^\dagger$ optimal for the regularized problem, and thus, $\omega^\dagger$ is feasible. Using this, and the above derivations, we finally obtain:

$$\begin{aligned}\widehat{n}_{\mathrm{RLHF}}(\overline{D}) &= \left|(\widehat{\omega})^\top\nabla_\omega\ell^{\widehat{\omega}}_{\mathrm{RLHF}}(\overline{D})\right| \leq \|\widehat{\omega}\|^2+\|\widehat{\omega}\|\left\|X^{\widehat{\omega}}_{\overline{D}}\right\| \\ &\leq O\left(\frac{\beta^2\left(D_{\mathrm{KL}}\left(\pi^\dagger||\mu\right)-\epsilon\right)^2}{(1-\gamma)^2\sigma^2_{\min}(\Sigma^\phi_{\overline{D}})\left\|\Gamma^{\omega^\dagger}_\phi\left(\pi^\dagger||\pi^{\mathrm{reg}}_{r_{\omega^\dagger}}\right)\right\|^2}+\frac{\overline{n}^2}{(1-\gamma)^4\sigma^4_{\min}(\Sigma^\phi_{\overline{D}})}\right),\end{aligned}$$

where the first inequality uses Cauchy-Schwarz and the last inequality uses Equation 16 and Equation 36.

□

C Proofs of Section 5

In this section, we provide the full proofs of the results from Section 5.

C.1 DPO with Non-empty Existing Data

In this section, we provide the full proof of Theorem 5.1.

Theorem 5.1. *Let $\overline{D}$ be a given preference dataset of $\overline{n}$ samples, let $\beta > 0$ and $0 < \epsilon' \leq \epsilon/2$. Furthermore, let $\overline{\theta}$ be the optimal point for $\ell^\theta_{\mathrm{DPO}}(\overline{D})$ and define*

$$\widetilde{\theta} = \mathrm{proj}_{\theta:\|\theta-\theta^\dagger\|\leq\epsilon'}\left(\overline{\theta}\right) = \theta^\dagger + \frac{(\epsilon')^2}{\left\|\overline{\theta}-\theta^\dagger\right\|^2}\left(\overline{\theta}-\theta^\dagger\right) .$$

Then, the dataset $\widehat{D}$ containing

$$2\left\lceil|(\nabla_\theta\ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D}))^\top(\widetilde{\theta}-\theta_\mu)|/(2\xi_{\max})\right\rceil$$

identical samples satisfying

$$\begin{aligned}&\beta\left(\widetilde{\theta}-\theta_\mu\right)^\top(\psi(s,a)-\psi(s,a')) \\ &\quad= o\cdot\xi_2\left(\nabla_\theta\ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right)^\top\left(\widetilde{\theta}-\theta_\mu\right),\end{aligned}$$

with $o = 1$ for half the samples and $o = -1$ for the remaining, is a feasible solution to Problem P:Attack:DPO.1. Furthermore, there exists an optimal solution $\widehat{D}$ to Problem P:Attack:DPO.1 with $\widehat{n}_{\mathrm{DPO}}$ identical samples such that

$$\widehat{n}_{\mathrm{DPO}} \leq 2\cdot$$

$$\left\lceil (\overline{n}\beta+\lambda)\frac{\left|\left\|\overline{\theta}-\theta^\dagger\right\|^2-(\epsilon')^2\right|}{2\xi_{\max}\left\|\theta^\dagger-\overline{\theta}\right\|}\left(3\left\|\theta^\dagger\right\|+\|\theta_\mu\|+\sqrt{\epsilon'}\right)\right\rceil .$$

Proof. Lemma E.5 implies that, for any $\epsilon' \leq \epsilon/2$, any feasible solution for Problem P:Attack.DPO.2 is feasible for Problem P:Attack:DPO.1. Thus, we focus on Problem P:Attack.DPO.2.

First, note that, since μ is loglinear, Lemma E.8 implies that Problem Equation P:Attack.DPO.2 can be written as

$$\begin{aligned}&\min_D |D|\\ &\text{s.t } \widetilde{\theta}=\arg\min_\theta \sum_{(s,a,a',o)\in\overline{D}\cup D}\log\left(1+\exp\left(-o\cdot\beta(\theta-\theta_\mu)^\top(\psi(s,a)-\psi(s,a'))\right)\right)+\frac{\lambda}{2}\|\theta-\theta_\mu\|^2\\ &\quad\left\|\widetilde{\theta}-\theta^\dagger\right\|^2\leq\epsilon' .\end{aligned}$$

Given fixed $\widetilde{\theta}\neq\mathbf{0}$, Lemma E.7 implies that the solution to the subproblem

$$\begin{aligned}&\min_D |D|\\ &\text{s.t } \widetilde{\theta}=\arg\min_\theta \sum_{(s,a,a',o)\in\overline{D}\cup D}\log\left(1+\exp\left(-o\cdot\beta(\theta-\theta_\mu)^\top(\psi(s,a)-\psi(s,a'))\right)\right)+\frac{\lambda}{2}\|\theta-\theta_\mu\|^2\end{aligned}$$

is the set of

$$2\left\lceil\frac{\left|(\nabla_\theta\ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top(\theta^\dagger-\theta_\mu)\right|}{2\xi_{\max}}\right\rceil$$

identical samples satisfying

$$\beta\left(\theta^\dagger-\theta_\mu\right)^\top(\psi(s,a)-\psi(s,a'))=o\cdot\xi^{-1}\left(\frac{(\nabla_\theta\ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top\left(\theta^\dagger-\theta_\mu\right)}{2\left\lceil\frac{\left|(\nabla_\theta\ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top(\theta^\dagger-\theta_\mu)\right|}{2\xi_{\max}}\right\rceil}\right),$$

with $o=1$ for half the samples and $o=-1$ for the remaining samples. Thus, the attack problem can be equivalently written in terms of θ as:

$$\begin{aligned}\min\ &\left|\left(\nabla_\theta\ell^\theta_{\mathrm{DPO}}(\overline{D})\right)^\top(\theta-\theta_\mu)\right|\\ \text{s.t }&\left\|\theta-\theta^\dagger\right\|^2\leq\epsilon' .\end{aligned}$$

The Lagrangian of the above can be written as

$$\mathcal{L}(\theta,\alpha)=\left|\left(\nabla_\theta\ell^\theta_{\mathrm{DPO}}(\overline{D})\right)^\top(\theta-\theta_\mu)\right|+\alpha\left(\left\|\theta-\theta^\dagger\right\|^2-\epsilon''\right),$$

and the first-order condition is

$$\nabla_\theta\mathcal{L}(\theta,\alpha)=e_\theta\cdot\left(\nabla^2_\theta\ell^\theta_{\mathrm{DPO}}(\overline{D})(\theta-\theta_\mu)+\nabla_\theta\ell^\theta_{\mathrm{DPO}}(\overline{D})\right)+2\alpha\left(\theta-\theta^\dagger\right)=\mathbf{0},$$

where

$$e_\theta=\operatorname{sgn}\left(\left(\nabla_\theta\ell^\theta_{\mathrm{DPO}}(\overline{D})\right)^\top(\theta-\theta_\mu)\right)$$

denotes the sign of the quantity inside the brackets. This yields

$$\theta^*=\left(\nabla^2_\theta\ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})+2\alpha I\right)^{-1}\left(e_{\theta^*}\nabla^2_\theta\ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\theta_\mu-e_{\theta^*}\nabla_\theta\ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})+2\alpha\theta^\dagger\right),$$

and

$$\theta^* - \theta_\mu = \left(\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\right)^{-1} \left(2e_{\theta^*}\alpha\left(\theta^* - \theta^\dagger\right) - \nabla_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\right) \ .$$

For non-zero α, we use complementary slackness to obtain

$$\left\| \left(\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) + 2\alpha I\right)^{-1} \left(e_{\theta^*}\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\theta_\mu - e_{\theta^*}\nabla_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) + 2\alpha\theta^\dagger\right) - \theta^\dagger \right\|^2 = \epsilon' \ .$$

Before we proceed any further, note that, since the Hessian of the loss is symmetric positive definite, then

$$\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) = U\Sigma U \ ,$$

where U is an orthonormal matrix and Σ is the diagonal matrix with the eigenvalues of the Hessian. Using the Woodbury inversion formula, we have

$$\begin{aligned}
\left(\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) + 2\alpha I\right)^{-1} &= \frac{1}{2\alpha} I - \frac{1}{4\alpha^2} U \left(\Sigma^{-1} + \frac{1}{2\alpha} U^\top U I\right)^{-1} U \\
&= \frac{1}{2\alpha} I - \frac{1}{4\alpha^2} U \left(\Sigma^{-1} + \frac{1}{2\alpha} I\right)^{-1} U \\
&= \frac{1}{2\alpha}\left(I - U\Sigma_\alpha U\right) \ ,
\end{aligned}$$

where Σ_α is a diagonal matrix with entries $\sigma_i/(\sigma_i + 2\alpha)$, where σ_i are the entries of Σ, for every $1 \le i \le d$. Let $M_\alpha = I - U\Sigma_\alpha U$, and

$$g = e_{\theta^*}\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\theta_\mu - e_{\theta^*}\nabla_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) \ .$$

Then, we have

$$\begin{aligned}
4\alpha^2\epsilon' &= \left\|(I - U\Sigma_\alpha U)\left(e_{\theta^*}\nabla^2_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\theta_\mu - e_{\theta^*}\nabla_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D}) + 2\alpha\theta^\dagger\right) - 2\alpha\theta^\dagger\right\|^2 \\
&= \left\|M_\alpha g + 2\alpha\left(M_\alpha\theta^\dagger - \theta^\dagger\right)\right\|^2 \\
&= \|M_\alpha g\|^2 + 4\alpha\left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle + 4\alpha^2\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 \ .
\end{aligned}$$

We can write the above as a quadratic equation of α as

$$4\left(\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - \epsilon'\right)\alpha^2 + 4\left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle\alpha + \|M_\alpha g\|^2 = 0 \ .$$

Note that we are treating as constants some terms that involve α non-linearly. This implies fixed-point solutions in terms of α. We follow this route since we are only interested in the final bounds, which will be independent of such components. The discriminant of this equation is

$$\Delta = 16\|M_\alpha g\|^2\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - 16\left(\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - \epsilon'\right)\|M_\alpha g\|^2 = 16\|M_\alpha g\|^2\epsilon' \ .$$

This implies the following fixed-point equations:

$$\alpha_{1,2} = \frac{-\left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle \pm \|M_\alpha g\|\sqrt{\epsilon'}}{2\left(\|M_\alpha\theta^\dagger - \theta^\dagger\|^2 - \epsilon'\right)} \ .$$

Note that we have

$$\frac{-\left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle - \|M_\alpha g\|\sqrt{\epsilon'}}{2\left(\|M_\alpha\theta^\dagger - \theta^\dagger\|^2 - \epsilon'\right)} \ge \frac{-\|M_\alpha g\|\left(\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\| + \sqrt{\epsilon'}\right)}{2\left(\|M_\alpha\theta^\dagger - \theta^\dagger\|^2 - \epsilon'\right)} = \frac{\|M_\alpha g\|}{2\left(\sqrt{\epsilon'} - \|M_\alpha\theta^\dagger - \theta^\dagger\|\right)}$$

which, since the maximum eigenvalue of M_α is bounded by 1, is positive whenever

$$\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\| \leq 2\left\|\theta^\dagger\right\| \leq \sqrt{\epsilon'} \ .$$

This condition would guarantee a positive root of α, which is required as α is a Lagrange multiplier. If there is none, we do not take into consideration complementary slackness and solve for $\alpha = 0$. For now, let us assume that there exists a positive root and consider the $\alpha = 0$ case later. Since the final bounds do not depend on which of the solutions we pick, let us pick the one with $+\left\|M_\alpha g\right\|\sqrt{\epsilon'}$ without loss of generality.

Plugging it into the first-order solution, we get

$$\theta^* = \frac{1}{2\alpha}M_\alpha\left(g + 2\alpha\theta^\dagger\right) = M_\alpha\theta^\dagger + \frac{1}{2\alpha}M_\alpha g = M_\alpha\theta^\dagger + \frac{\left(\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - \epsilon'\right)}{\left\|M_\alpha g\right\|\sqrt{\epsilon'} - \left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle}M_\alpha g \ . \tag{37}$$

This implies that

$$\begin{aligned}
\left\|\theta^*\right\| &\leq \left\|M_\alpha\theta^\dagger\right\| + \left|\frac{\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - \epsilon'}{\sqrt{\epsilon'} - \left\langle M_\alpha g, M_\alpha\theta^\dagger - \theta^\dagger\right\rangle / \left\|M_\alpha g\right\|}\right| \\
&\leq \left\|M_\alpha\theta^\dagger\right\| + \frac{\left|\left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|^2 - \epsilon'\right|}{\left|\sqrt{\epsilon'} - \left\|M_\alpha\theta^\dagger - \theta^\dagger\right\|\right|} \\
&= \left\|M_\alpha\theta^\dagger\right\| + \left\|M_\alpha\theta^\dagger - \theta^\dagger\right\| + \sqrt{\epsilon'} \\
&\leq \left\|\theta^\dagger\right\| + 2\left\|\theta^\dagger\right\| + \sqrt{\epsilon'} \\
&= 3\left\|\theta^\dagger\right\| + \sqrt{\epsilon'} \ ,
\end{aligned}$$

where we have used the triangle inequality, Cauchy Schwarz and the fact that the maximum eigenvalue of M_α is upper-bounded by 1. Now we provide upper bounds on the gradient with respect to the pre-existing data for θ^*, similar to the proof of Theorem 4.1. Let $\overline{\theta}$ be the optimal parameter with respect to the loss on the pre-existing dataset. Note that we have

$$\begin{aligned}
\left\|\nabla_\theta \ell^{\theta^*}_{\mathrm{DPO}}(\overline{D})\right\| &\leq (\overline{n}\beta + \lambda)\left\|\theta^* - \overline{\theta}\right\| \\
&\leq (\overline{n}\beta + \lambda)\left(\left\|\theta^* - \mathrm{proj}_{\theta:\|\theta-\theta^\dagger\|\leq\epsilon'}\left(\overline{\theta}\right)\right\| + \left\|\mathrm{proj}_{\theta:\|\theta-\theta^\dagger\|\leq\epsilon'}\left(\overline{\theta}\right) - \overline{\theta}\right\|\right) \\
&\leq (\overline{n}\beta + \lambda)\left(\epsilon' + \left\|\mathrm{proj}_{\theta:\|\theta-\theta^\dagger\|\leq\epsilon'}\left(\overline{\theta}\right) - \overline{\theta}\right\|\right) \ ,
\end{aligned}$$

where the first inequality follows from Lemma E.14; the second inequality uses triangle inequality and the third inequality uses the fact that any two points in the ϵ'-ball are no farther than ϵ' away from each-other. To compute the projection onto the ϵ'-ball, we solve the following problem:

$$\min_\theta \left\|\theta - \overline{\theta}\right\|^2, \ \text{ such that } \ \left\|\theta - \theta^\dagger\right\|^2 \leq (\epsilon')^2 \ .$$

The first-order of the problem with respect to dual variable α of the Lagrangian

$$\mathcal{L}(\theta, \alpha) = \left\|\theta - \overline{\theta}\right\|^2 + \alpha\left(\left\|\theta - \theta^\dagger\right\|^2 - (\epsilon')^2\right)$$

gives us

$$\theta = \frac{1}{1+\alpha}\left(\alpha\theta^\dagger + \overline{\theta}\right) \ .$$

Complementary slackness implies that

$$(1+\alpha)(\epsilon')^2 = \left\|\overline{\theta} - \theta^\dagger\right\|^2$$

which in turn implies that we have

$$\widetilde{\theta} = \theta^\dagger + \frac{(\epsilon')^2}{\left\|\overline{\theta} - \theta^\dagger\right\|^2}\left(\overline{\theta} - \theta^\dagger\right) \ .$$

Thus, we have

$$\begin{aligned}
\widehat{n}_{\mathrm{DPO}} &= \left|\left(\nabla_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right)^\top \left(\widetilde{\theta} - \theta_\mu\right)\right| \\
&\leq \left\|\nabla_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right\| \left(3\left\|\theta^\dagger\right\| + \|\theta_\mu\| + \sqrt{\epsilon'}\right) \\
&= (\overline{n}\beta + \lambda)\left\|\theta^\dagger + \frac{(\epsilon')^2}{\left\|\overline{\theta} - \theta^\dagger\right\|^2}\left(\overline{\theta} - \theta^\dagger\right) - \overline{\theta}\right\| \left(3\left\|\theta^\dagger\right\| + \|\theta_\mu\| + \sqrt{\epsilon'}\right) \\
&\leq (\overline{n}\beta + \lambda)\frac{\left|\left\|\overline{\theta} - \theta^\dagger\right\|^2 - (\epsilon')^2\right|}{\left\|\theta^\dagger - \overline{\theta}\right\|} \left(3\left\|\theta^\dagger\right\| + \|\theta_\mu\| + \sqrt{\epsilon'}\right) \ .
\end{aligned}$$

If $\alpha = 0$, then the first-order condition yields

$$\widetilde{\theta} - \theta_\mu = -\left(\nabla^2_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right)^{-1} \nabla_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D}) \ ,$$

which implies

$$\begin{aligned}
\widehat{n}_{\mathrm{DPO}} &= \left|\nabla_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})^\top \left(\nabla^2_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right)^{-1} \nabla_\theta \ell^{\widetilde{\theta}}_{\mathrm{DPO}}(\overline{D})\right| \\
&\leq \frac{(\overline{n}\beta + \lambda)}{\sigma_{\min}(\Sigma^{\psi}_{\overline{D}})} \cdot \frac{\left|\left\|\overline{\theta} - \theta^\dagger\right\|^2 - (\epsilon')^2\right|}{\left\|\theta^\dagger - \overline{\theta}\right\|} \ ,
\end{aligned}$$

using Lemma E.14 and the projection derivations above.

For the lower bound, note that, in the best case scenario for the attacker, $\overline{D}$ is a subset of an optimal solution $\widehat{D}$ for Problem P:Attack:DPO.1 in the case when $\overline{D} = \emptyset$. Thus, the attack sample size in this case is the lower bound of Theorem C.1 without the size $\overline{n}$ of $\overline{D}$. □

C.2 DPO with Empty Existing Data

In this section, we provide additional upper bounds on the sample complexity for the DPO setting when $\overline{D} = \emptyset$. Our aim is to obtain bounds that are tighter than the ones obtained by directly instantiating the bounds of Theorem 5.1. Moreover, we also provide lower bounds for this setting and use them for the general lower bounds of Theorem 5.1.

Theorem C.1. *Let $\overline{D} = \emptyset$, let $\beta > 0$ and $0 < \epsilon' \leq \epsilon/2$. Furthermore, let $\pi^\dagger, \mu \in \Pi^{\log}$ be loglinear with parameters $\theta^\dagger$ and θ_μ, respectively. Define*

$$\widetilde{\theta} = \theta^\dagger + e\sqrt{\epsilon'}(\theta_\mu - 2\theta^\dagger)/\left\|\theta_\mu - 2\theta^\dagger\right\| ,$$

where $e = 1$, if ${\theta^\dagger}^\top\left(\theta^\dagger - \theta_\mu\right) \geq \sqrt{\epsilon'} - \epsilon'$, and $e = -1$, otherwise. Then, the dataset of $2\left\lceil \frac{\lambda\left|\lambda\widetilde{\theta}^\top\left(\widetilde{\theta} - \theta_\mu\right)\right|}{2\xi_{\max}}\right\rceil$ samples satisfying

$$\beta\left(\widetilde{\theta} - \theta_\mu\right)^\top \left(\psi(s,a) - \psi(s,a')\right) = o \cdot \xi_2\left(\lambda\left\|\widetilde{\theta} - \theta_\mu\right\|^2\right)$$

with $o = 1$ for half of the samples, and $o = -1$ for the remaining is a feasible solution to Problem P:Attack:DPO.1. Furthermore, there exists an optimal solution $\widehat{D}$ to Problem P:Attack:DPO.1 with $\widehat{n}_{\mathrm{DPO}}$ identical samples such that

$$\widehat{n}_{\mathrm{DPO}} \leq 2\left\lceil \frac{\lambda}{2\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2\right\rceil \ .$$

Finally, there exists $\eta_{\min} > 0$, such that, for any $\epsilon' \geq \epsilon/\eta_{\min}$, we have

$$\widehat{n}_{\mathrm{DPO}} \geq 2\left\lceil \frac{\lambda}{2\xi_{\max}} \left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'} \right)^2 \right\rceil \ .$$

Proof. First, note that, since μ is loglinear, Lemma E.8 implies that Problem Equation P:Attack.DPO.2 can be written as

$$\begin{aligned} &\min_D |D| \\ &\text{s.t } \widetilde{\theta} = \arg\min_\theta \sum_{(s,a,a',o)\in D} \log\left(1 + \exp\left(-o \cdot \beta(\theta - \theta_\mu)^\top \left(\psi(s,a) - \psi(s,a')\right)\right)\right) + \frac{\lambda}{2}\left\|\theta - \theta_\mu\right\|^2 \\ &\quad \left\|\widetilde{\theta} - \theta^\dagger\right\|^2 \leq \epsilon' \ . \end{aligned}$$

Now, given $\widetilde{\theta} \neq \mathbf{0}$, Lemma E.7 implies that the solution to the problem

$$\begin{aligned} &\min_D |D| \\ &\text{s.t } \widetilde{\theta} = \arg\min_\theta \sum_{(s,a,a',o)\in D} \log\left(1 + \exp\left(-o \cdot \beta(\theta - \theta_\mu)^\top \left(\psi(s,a) - \psi(s,a')\right)\right)\right) + \frac{\lambda}{2}\left\|\theta - \theta_\mu\right\|^2 \end{aligned}$$

is the set of

$$2\left\lceil \frac{\lambda\left\|\widetilde{\theta} - \theta_\mu\right\|^2}{2\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\beta\left(\widetilde{\theta} - \theta_\mu\right)^\top \left(\psi(s,a) - \psi(s,a')\right) = o \cdot \xi^{-1}\left(\frac{\lambda\left\|\widetilde{\theta} - \theta_\mu\right\|^2}{2\left\lceil \frac{\lambda\left\|\widetilde{\theta} - \theta_\mu\right\|^2}{2\xi_{\max}} \right\rceil} \right) \ ,$$

with $o = 1$ for half the samples and $o = -1$ for the remaining samples.

Next, for the upper bound, we will consider the surrogate problem. Using Lemma E.5, for any $\epsilon' \leq \epsilon/(2\sqrt{d'})$, any feasible solution for Problem P:Attack.DPO.2 is feasible for Problem P:Attack:DPO.1. Thus, we focus on the former. We can rewrite the problem directly in terms of a variable θ:

$$\min_\theta \ \left\|\theta - \theta_\mu\right\|^2 \ \text{ such that } \ \left\|\theta - \theta^\dagger\right\|^2 \leq \epsilon' \ .$$

The Lagrangian of the above can be written as

$$\mathcal{L}(\theta, \alpha) = \left\|\theta - \theta_\mu\right\|^2 + \alpha\left(\left\|\theta - \theta^\dagger\right\|^2 - \epsilon'\right) \ .$$

The first-order condition can be written as

$$\nabla_\theta \mathcal{L}(\theta, \alpha) = (\theta - \theta_\mu) + \alpha\left(\theta - \theta^\dagger\right) = \mathbf{0} \ ,$$

which implies that

$$\widetilde{\theta} = \frac{1}{1+\alpha}\left(\alpha\theta^\dagger + \theta_\mu\right) \ .$$

Complementary slackness implies

$$\left\| \frac{1}{1+\alpha}\left(\alpha\theta^\dagger + \theta_\mu\right) - \theta^\dagger \right\| = \sqrt{\epsilon'} \ .$$

Equivalently,

$$\left\|\theta^\dagger - \theta_\mu\right\| = |1+\alpha|\sqrt{\epsilon'}\ ,$$

which, because all three terms are non-negative, implies that

$$\alpha = \frac{\left\|\theta^\dagger - \theta_\mu\right\|}{\sqrt{\epsilon'}} - 1\ ,$$

Plugging this into the first-order condition and subtracting both sides by θ_μ, we get

$$\begin{aligned}\widetilde{\theta} - \theta_\mu &= \frac{\sqrt{\epsilon'}}{\|\theta^\dagger - \theta_\mu\|}\left(\left(\frac{\left\|\theta^\dagger - \theta_\mu\right\|}{\sqrt{\epsilon'}} - 1\right)\theta^\dagger + \theta_\mu\right) - \theta_\mu = \theta^\dagger - \theta_\mu - \frac{\sqrt{\epsilon'}}{\|\theta^\dagger - \theta_\mu\|}\left(\theta^\dagger - \theta_\mu\right)\\ &= \left(\frac{\sqrt{\epsilon'}}{\|\theta^\dagger - \theta_\mu\|} - 1\right)\left(\theta^\dagger - \theta_\mu\right)\end{aligned}$$

Using the above and some algebraic manipulations, we can write

$$\left\|\widetilde{\theta} - \theta_\mu\right\|^2 = \left\|\left(\frac{\sqrt{\epsilon'}}{\|\theta^\dagger - \theta_\mu\|} - 1\right)\left(\theta^\dagger - \theta_\mu\right)\right\|^2 = \left\|\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right\|^2$$

The desired bounds follow as an immediate consequence of the above. □

D Proofs of Section 6

In this section, we provide the proof of the following result.

Theorem 6.1. *Let $\pi^\dagger, \mu \in \Pi^{\log}$ be loglinear with parameters $\theta^\dagger$ and θ_μ, respectively. Furthermore, let ϵ be such that $\epsilon \leq 1/(2\xi_{\max})$, $\epsilon' \geq \epsilon/\eta_{\min}$, where $\eta_{\min} > 0$ is an absolute constant, and let $\omega^\dagger$ be a feasible solution to Problem P:Attack:RLHF.1. Define κ_1 as*

$$\begin{aligned}&\left(\frac{\lambda}{\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2 - \overline{n}\right)\\ &\qquad\cdot\left\lceil\frac{1}{\xi_{\max}}\left(\lambda\left\|\omega^\dagger\right\|^2 + \left\|\omega^\dagger\right\|\cdot\frac{2\overline{n}}{(1-\gamma)^2}\right)\right\rceil^{-1}\end{aligned}$$

Then, we have $\widehat{n}_{\mathrm{DPO}} \geq \kappa_1 \cdot \widehat{n}_{\mathrm{RLHF}}$.

Proof. Note that we have

$$\begin{aligned}\widehat{n}_{\mathrm{DPO}} &\geq 2\left\lceil\frac{\lambda}{2\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2\right\rceil - \overline{n}\\ &\geq \frac{\lambda}{\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2 - \overline{n}\\ &= \left(\frac{\lambda}{\xi_{\max}}\left(\left\|\theta^\dagger - \theta_\mu\right\| - \sqrt{\epsilon'}\right)^2 - \overline{n}\right)\cdot\left\lceil\frac{1}{\xi_{\max}}\left(\lambda\left\|\omega^\dagger\right\|^2 + \left\|\omega^\dagger\right\|\cdot\frac{2\overline{n}}{(1-\gamma)^2}\right)\right\rceil^{-1}\\ &\qquad\cdot\left\lceil\frac{1}{\xi_{\max}}\left(\lambda\left\|\omega^\dagger\right\|^2 + \left\|\omega^\dagger\right\|\cdot\frac{2\overline{n}}{(1-\gamma)^2}\right)\right\rceil\\ &\geq \kappa_1\cdot\widehat{n}_{\mathrm{RLHF}}\ ,\end{aligned}$$

where the first inequality follows from Theorem 5.2 and the first inequality follows from Corollary 4.2. □

E Technical Lemmas

This section includes miscellaneous technical results used throughout the paper. We begin by stating a result about the structure of the regularized optimal policy.

Lemma E.1. *Given policy π and reward function r, let*

$$\mathcal{V}_r^\pi(s) = \mathbb{E}_{s\sim\rho, a_t\sim\pi(\cdot|s_t)}\left[\sum_{t\geq 0}\gamma^t\left(r(s_t,a_t) - \beta\log\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right)\right]$$

and

$$\mathcal{Q}_r^\pi(s,a) = r(s,a) + \gamma\sum_{s'}P(s,a,s')\mathcal{V}_r^\pi(s')$$

denote the regularized value and action value functions with respect to π and r, respectively. Moreover, let the regularized advantage function be defined as

$$\mathcal{A}_r^\pi(s,a) = \mathcal{Q}_r^\pi(s,a) - \mathcal{V}_r^\pi(s)\ .$$

Then, the unique optimal regularized policy can be written, for every (s,a), as

$$\pi_r^{\text{reg}}(a|s) = \mu(a|s)\exp\left(\frac{1}{\beta}\mathcal{A}^{\pi_r^{\text{reg}}(s,a)}\right)\ .$$

Proof. The proof of this result is a straightforward application of the results from Appendix C of [Nachum et al., 2017] from the entropy-based to KL divergence-based regularization. □

Now we prove that $M_{\pi^\dagger}$ is full rank.

Lemma E.2. *Let $\pi^\dagger \in \Pi^{\text{det}}$ and Φ be of rank d. Assume that the set $\{\phi(s,a) : s\in\mathcal{S}, a\in\mathcal{A}\setminus\text{supp}(\pi^\dagger)\}$, where $\text{supp}(\pi)$ denotes the set of actions chosen by deterministic policy π, contains d linearly independent vectors. Then, the matrix $M_{\pi^\dagger}$ is full rank.*

Proof. Let v be an arbitrary column of $M_{\pi^\dagger}$. Then, there exists (s,a) and coefficients $\alpha_{s',a'}$ such that

$$v = \sum_{s'} d^{\pi^\dagger}(s')\phi(s',\pi^\dagger(s')) - \sum_{s'} d^{\pi^\dagger\{s,a\}}(s')\phi(s',\pi^\dagger\{s,a\}(s')) = \sum_{s',a'}\alpha_{s',a'}\phi(s',a')\ .$$

By definition of the neighbors of $\pi^\dagger$, we have that $a = \pi^\dagger\{s,a\}(s) \neq \pi^\dagger(s) = \pi(s)$, for all $\pi\in\mathcal{N}(\pi^\dagger)\setminus\{\pi\{s,a\}\}$. This is because all neighboring policies are deterministic and change only in one state from $\pi^\dagger$, and, consequently, from each-other.

This implies that, there is no column v' in $M_{\pi^\dagger}$, such that $\phi(s,a)$ appears in the decomposition of v'. There are $S(A-1)$ such vectors, since there are $S(A-1)$ neighbors of $\pi^\dagger$. Assuming that they contain all vectors that span the column space of Φ, this means that the rank of $M_{\pi^\dagger}$ is equal to the rank of Φ. □

Next, we provide three results that connect the solutions of the surrogate problems to their original problems throughout the paper. We start with the unregularized RLHF setting.

Lemma E.3. *Let $\epsilon' > 0$. Then, any feasible solution to Problem P:Attack:RLHF.2 is a feasible solution to Problem P:Attack:RLHF.1.*

Proof. First, note that, when $\beta = 0$, given reward r, then $\pi_r^{\text{reg}}(\cdot|s)\in\arg\max_\pi V_r^\pi(s)$, for all s. Such optimal policies are known to be deterministic. Now, given $\widehat{\omega}$, note that, if $\epsilon' > 0$, then $V_{r_{\widehat{\omega}}}^{\pi^\dagger}(s) > V_{r_{\widehat{\omega}}}^{\pi}(s)$, for any state s. This means that $\pi^\dagger$ is the unique optimal policy under $r_{\widehat{\omega}}$. This further implies that $\pi_{r_{\widehat{\omega}}}^{\text{reg}} = \pi^\dagger$, and thus, $D_{\text{KL}}(\pi^\dagger||\pi_{r_{\widehat{\omega}}}^{\text{reg}}) = 0 < \epsilon$, for any $\epsilon > 0$. □

Next, we consider the regularized RLHF setting.

Lemma E.4. *Let $\epsilon' \leq (2\ln 2)\epsilon$. Then, any feasible solution to Problem P:Attack:RLHF.3 is a feasible solution to Problem P:Attack:RLHF.1.*

Proof. Given two policies π and π', we have

$$\begin{aligned}\|\pi - \pi'\|_1^2 &= \sum_{s,a} \rho(s) \left|\pi(\cdot|s) - \pi'(\cdot|s)\right|^2 \\ &\leq \sum_{s} \rho(s) 2\ln 2 D_{\mathrm{KL}}\left(\pi(\cdot|s)||\pi'(\cdot|s)\right) \\ &= (2\ln 2) D_{\mathrm{KL}}(\pi||\pi') \ ,\end{aligned}$$

where the first inequality follows from Pinsker's inequality. This implies that, as long as $D_{\mathrm{KL}}(\pi||\pi') \leq \epsilon'$, then $\|\pi - \pi'\|_1^2 \leq (2\ln 2)\epsilon' \leq \epsilon$. □

Finally, we consider the DPO setting.

Lemma E.5. *Let $\epsilon' \leq \epsilon/(2\sqrt{d'})$. Then, any feasible solution to Problem P:Attack.DPO.2 is a feasible solution to Problem P:Attack:DPO.1. Moreover, there exists $\eta_{\min} > 0$ such that, for all $\epsilon' \geq \epsilon/\eta_{\min}$, any feasible solution to Problem P:Attack:DPO.1 is also a feasible solution to Problem P:Attack.DPO.2.*

Proof. First, it can be easily shown that the gradient of a loglinear policy is

$$\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s)\left(\psi(s,a) - \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\psi(s,a')\right]\right).$$

We can bound this gradient by $\|\nabla_\theta \pi_\theta(a|s)\| \leq 2$. Thus, loglinear policies are 2-Lipschitz in their parameters θ. Now, given policies π_θ and $\pi_{\theta'}$, due to the above argument, we can write

$$\|\pi_\theta - \pi_{\theta'}\|_1 \leq 2\|\theta - \theta'\| \leq 2\epsilon' \leq \epsilon.$$

This implies that, for any $\epsilon' \leq \epsilon/2$, any feasible solution to Problem P:Attack.DPO.2 will be feasible for Problem P:Attack:DPO.1.

For the second statement, we argue as follows. Since the function π_θ is continuously differentiable in θ, the Mean Value Theorem implies that, for any θ and (s,a), there exists $\theta_M(s,a)$ such that

$$\pi_\theta(a|s) - \pi_{\theta^\dagger}(a|s) = \nabla_\theta \pi_{\theta_M(s,a)}(a|s)^\top\left(\theta - \theta^\dagger\right) = \pi_{\theta_M(s,a)}(a|s)\overline{\psi}_{\theta_M(s,a)}(s,a)^\top\left(\theta - \theta^\dagger\right) \ , \tag{38}$$

where

$$\overline{\psi}_\theta(s,a) = \psi(s,a) - \sum_{a'} \pi_\theta(a'|s)\psi(s,a') \ .$$

Let $\nabla_\theta \pi_{\theta_M}$ be the $S\cdot A$-dimensional matrix with columns $\rho(s)\pi_{\theta_M(s,a)}(a|s)\overline{\psi}_{\theta_M(s,a)}(s,a)$, and let π denote the $S\cdot A$-dimensional vector with entries $\rho(s)\pi(a|s)$, for each (s,a). Then, we have

$$\|\pi_\theta - \pi_{\theta^\dagger}\|_1 \geq \|\pi_\theta - \pi_{\theta^\dagger}\| = \left\|\nabla_\theta \pi_{\theta_M}\left(\theta - \theta^\dagger\right)\right\| \geq \sigma_{\min}\left(\nabla_\theta \pi_{\theta_M}\right)\left\|\theta - \theta^\dagger\right\| \ ,$$

where the first inequality follows from the relationship between ℓ_1 and ℓ_2 norms; the first equality follows from Equation 38 and the second inequality follows from the fact that $\|Ax\|_2 \geq \sigma_{\min}(A)\|x\|_2$, for compatible matrix A and vector x, where $\sigma_{\min}(A)$ denotes the minimum singular value of A. Now, observe that

$$\nabla_\theta \pi_{\theta_M}^\top \nabla_\theta \pi_{\theta_M} = \sum_{s,a} \rho(s)^2 \pi_{\theta_M(s,a)}(a|s)^2 \overline{\psi}_{\theta_M(s,a)}(s,a)\overline{\psi}_{\theta_M(s,a)}(s,a)^\top \ .$$

We will show that this matrix is positive definite, whenever the vectors $\psi(s,a)$ span $\mathbb{R}^{d'}$. First, the ergodicity assumption and the loglinearity of the policies imply that $\rho(s)\pi_\theta(a|s) > 0$, for any (s,a) and θ.

Next, observe that, since vectors $\psi(s,a)$ span $\mathbb{R}^{d'}$, then vectors $\overline{\psi}_{\theta_M(s,a)}(s,a)$ also do, as translations of basis vectors via mean vectors. It is clear that $\nabla_\theta \pi_{\theta_M}^\top \nabla_\theta \pi_{\theta_M}$ is positive semi-definite, meaning that, for every non-zero vector v,

$$v^\top\left(\nabla_\theta \pi_{\theta_M}^\top \nabla_\theta \pi_{\theta_M}\right) v \geq 0 \ .$$

The only way the above can be zero is if, for every (s, a), we have

$$v^\top \overline{\psi}_{\theta_M(s,a)}(s,a) = 0 \ .$$

But since $\overline{\psi}_{\theta_M(s,a)}(s,a)$ span $\mathbb{R}^{d'}$, then there exist coefficients $\alpha_{s,a}$ such that

$$v = \sum_{s,a} \alpha_{s,a} \overline{\psi}_{\theta_M(s,a)}(s,a) \ ,$$

which implies that

$$v^\top v = \sum_{s,a} \alpha_{s,a} \overline{\psi}_{\theta_M(s,a)}(s,a)^\top v = 0 \ ,$$

which is a contradiction, since v is assumed to be non-zero. Thus, the matrix $\nabla_\theta \pi_{\theta_M}^\top \nabla_\theta \pi_{\theta_M}$ is positive definite, for any given θ. This means that

$$\sigma_{\min}\left(\nabla_\theta \pi_{\theta_M}\right) = \sqrt{\lambda_{\min}\left(\nabla_\theta \pi_{\theta_M}^\top \nabla_\theta \pi_{\theta_M}\right)} = \eta_{\min}(\theta) \geq \min_\theta \eta_{\min}(\theta) := \eta_{\min} > 0 \ ,$$

which further implies that

$$\left\|\theta - \theta^\dagger\right\| \leq \frac{1}{\eta_{\min}} \left\|\pi_\theta - \pi_{\theta^\dagger}\right\| \leq \frac{1}{\eta_{\min}} \left\|\pi_\theta - \pi_{\theta^\dagger}\right\|_1 \leq \frac{1}{\eta_{\min}} \epsilon \leq \epsilon' \ .$$

□

We now prove two important lemmas which we use in most of the results of the paper. They provide solutions to the attack subproblems for RLHF and DPO. We start with the RLHF result.

Lemma E.6. *Let $\omega^\dagger \in \mathbb{R}^d$ and let $\overline{D} = \{(\tau, \tau', o)\}$ be a given preference dataset of n samples. Consider the problem*

$$\min_D |D| \ \text{ such that } \ \omega^\dagger = \arg\min_\omega \ell^\omega_{\text{RLHF}}\left(\overline{D} \cup D\right) \ .$$

Then, the solution to the above problem is the dataset of

$$\left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau) - \phi(\tau') = \xi^{-1} \left(\frac{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})}{\left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\text{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil} \right) \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2}, \ \ o = 1 \ .$$

Moreover, if $\overline{D} = \emptyset$, the solution of the problem is the set of

$$\left\lceil \frac{\lambda \left\|\omega^\dagger\right\|^2}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau) - \phi(\tau') = \xi^{-1} \left(\frac{\lambda \left\|\omega^\dagger\right\|^2}{\left\lceil \frac{\lambda \|\omega^\dagger\|^2}{\xi_{\max}} \right\rceil} \right) \frac{\omega^\dagger}{\left\|\omega^\dagger\right\|^2}, \ \ o = 1 \ .$$

Proof. First, note that the second case, when $\overline{D} = \emptyset$, is a direct consequence of Theorem E.15. We thus focus on the general case when $\overline{D} \neq \emptyset$.

The first-order condition of our problem can be written as

$$\sum_{(\tau,\tau',o)\in\overline{D}} \frac{-o\left(\phi(\tau)-\phi(\tau')\right)}{1+\exp\left(o(\omega^\dagger)^\top\left(\phi(\tau)-\phi(\tau')\right)\right)} + \sum_{(\tau,\tau',o)\in D} \frac{o\left(\phi(\tau)-\phi(\tau')\right)}{1+\exp\left(o(\omega^\dagger)^\top\left(\phi(\tau)-\phi(\tau')\right)\right)} + \lambda\omega = \mathbf{0}\ .$$

We can equivalently write the above as

$$\sum_{(\tau,\tau',o)\in D} \frac{o(\phi(\tau)-\phi(\tau'))}{1+\exp\left(o(\omega^\dagger)^\top(\phi(\tau)-\phi(\tau'))\right)} + \nabla_\omega \ell^{\omega^\dagger}_{\mathrm{RLHF}}(\overline{D}) = \mathbf{0}\ .$$

Recall that we have defined

$$\xi_{\max} := \max_t \frac{t}{1+\exp(t)}\ .$$

Let us denote by $\xi^{-1}(a)$ the solution to $a = x/(1+\exp(x))$, for $a \leq \xi_{\max}$. Such a solution exists and can be written in closed form [Liu and Zhu, 2016] as

$$\xi^{-1}(a) = a - W_{\mathrm{Lam}}(-a\exp(a))\ ,$$

for every $a \leq \xi_{\max}$, where W_{Lam} denotes the Lambert W function. Now, let $\widehat{D}$ be a preference dataset of

$$\widehat{n} := \left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\phi(\tau)-\phi(\tau') = \xi^{-1}\left(\frac{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})}{\left\lceil \frac{\left|\nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})^\top \omega^\dagger\right|}{\xi_{\max}} \right\rceil}\right) \frac{\omega^\dagger}{\|\omega^\dagger\|^2},\quad o = 1\ .$$

Note that the first-order condition with respect to $\widehat{D}$ yields

$$\begin{aligned}
&\sum_{(\tau,\tau',o)\in\widehat{D}} \frac{-o(\phi(\tau)-\phi(\tau'))}{1+\exp\left(o(\omega^\dagger)^\top(\phi(\tau)-\phi(\tau'))\right)} + \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \\
&= -\widehat{n}\frac{\xi^{-1}\left((\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \left\lceil \frac{\left|(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})\right|}{\xi_{\max}} \right\rceil^{-1}\right)}{1+\exp\left(\xi^{-1}\left((\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \left\lceil \frac{\left|(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})\right|}{\xi_{\max}} \right\rceil^{-1}\right)\right)} \cdot \frac{\nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})}{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})} + \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \\
&= -\widehat{n}(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \left\lceil \frac{\left|(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})\right|}{\xi_{\max}} \right\rceil^{-1} \frac{\nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})}{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})} + \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \\
&= \mathbf{0}\ ,
\end{aligned}$$

where the penultimate equality is due to the fact that

$$\omega^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D}) \left\lceil \frac{(\omega^\dagger)^\top \nabla_\omega \ell^\omega_{\mathrm{RLHF}}(\overline{D})}{\xi_{\max}} \right\rceil^{-1} \leq \xi_{\max}$$

and the property of $\xi^{-1}(\cdot)$; the last equality follows by the definition of $\widehat{n}$. Since the function $\ell^\omega_{\mathrm{RLHF}}(D)$ is strongly convex, the first-order condition is enough to determine the optimal solution. □

Similarly, we prove an analogous result for DPO. The difference here is that the problem of interest is not a homogeneous logistic regression anymore, due to the presence of θ_μ.

Lemma E.7. *Let $\theta^\dagger, \theta_\mu \in \mathbb{R}^{d'}$ for some reference policy μ and let $\overline{D} = \{(\tau, \tau', o)\}$ be a given preference dataset of n samples. Consider the problem*

$$\min_D |D| \text{ such that } \theta^\dagger = \arg\min_\theta \ell^\theta_{\mathrm{DPO}}\left(\overline{D} \cup D\right) .$$

Then, the solution to the above problem is the dataset of

$$2\left\lceil \frac{\left|(\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top(\theta^\dagger - \theta_\mu)\right|}{2\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\beta\left(\theta^\dagger - \theta_\mu\right)^\top (\psi(s,a) - \psi(s,a')) = o \cdot \xi^{-1}\left(\frac{(\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top\left(\theta^\dagger - \theta_\mu\right)}{2\left\lceil \frac{\left|(\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}))^\top(\theta^\dagger - \theta_\mu)\right|}{2\xi_{\max}} \right\rceil} \right) ,$$

with $o = 1$ for half the samples and $o = -1$ for the remaining samples. Moreover, if $\overline{D} \neq \emptyset$, the solution of the problem is the dataset of

$$2\left\lceil \frac{\lambda \left\|\theta^\dagger - \theta_\mu\right\|^2}{2\xi_{\max}} \right\rceil$$

identical samples satisfying

$$\beta\left(\theta^\dagger - \theta_\mu\right)^\top (\psi(s,a) - \psi(s,a')) = o \cdot \xi^{-1}\left(\frac{\lambda \left\|\theta^\dagger - \theta_\mu\right\|^2}{2\left\lceil \frac{\lambda\|\theta^\dagger - \theta_\mu\|^2}{2\xi_{\max}} \right\rceil} \right) ,$$

with $o = 1$ for half the samples and $o = -1$ for the remaining samples.

Proof. We prove the general case. The second case follows directly from the fact that, if $\overline{D} = \emptyset$, we have

$$\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) = \lambda \left\|\theta^\dagger - \theta_\mu\right\|^2 .$$

First, note that, for loglinear μ with parameter θ_μ, the optimization problem of interest becomes

$$\begin{aligned} &\min_D |D| \\ &\text{s.t } \theta^\dagger = \arg\min_\theta \sum_{(s,a,a',o)\in D\cup\overline{D}} \log\left(1 + \exp\left(-o \cdot \beta(\theta - \theta_\mu)^\top(\psi(s,a) - \psi(s,a'))\right)\right) + \frac{\lambda}{2}\left\|\theta - \theta_\mu\right\|^2 . \end{aligned}$$

The first-order condition of the above can be written as

$$\sum_{(s,a,a',o)\in D} \frac{-\beta\left(\psi(s,a) - \psi(s,a')\right)}{1 + \exp\left(o \cdot \beta(\theta^\dagger - \theta_\mu)^\top(\psi(s,a) - \psi(s,a'))\right)} + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) = \mathbf{0} .$$

Now let us consider the following construction. Let

$$\widehat{n} = 2\left\lceil \frac{\left|\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top\left(\theta^\dagger - \theta_\mu\right)\right|}{2\xi_{\max}} \right\rceil .$$

For every $1 \leq i \leq \widehat{n}/2$, let ψ_+ be such that

$$\beta \left(\theta^\dagger - \theta_\mu\right)^\top \psi_+ = z \ ,$$

where

$$z = \xi^{-1}\left(\frac{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)}{\widehat{n}} \right) \ ,$$

and for every $\widehat{n}/2 + 1 \leq j \leq \widehat{n}$, let

$$\psi_- = \psi_+ - \frac{2z}{\beta \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)} \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \ .$$

Note that

$$\beta \left(\theta^\dagger - \theta_\mu\right)^\top \psi_- = -z \ .$$

Using dataset $\widehat{D} = \{(s_i, a_i, a_i', o_i)\}_{i=1}^{\widehat{n}}$ such that $\psi(s_i, a_i) - \psi(s_i, a_i') = \psi_+$ and $o_i = 1$, for all $1 \leq i \leq \widehat{n}/2$, and $\psi(s_i, a_i) - \psi(a_i, s_i') = \psi_-$ and $o_i = -1$, for all $\widehat{n}/2 + 1 \leq j \leq \widehat{n}$, we consider the first-order condition of our problem:

$$\begin{aligned}
&\frac{-\beta\widehat{n}}{2} \cdot \frac{1}{1 + \exp\left(\beta \left(\theta^\dagger - \theta_\mu\right)^\top \psi_+\right)} \psi_+ + \frac{\beta\widehat{n}}{2} \cdot \frac{1}{1 + \exp\left(-\beta \left(\theta^\dagger - \theta_\mu\right)^\top \psi_-\right)} \psi_- + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \\
&= \frac{-\beta\widehat{n}}{2} \cdot \frac{1}{1 + \exp\left(z\right)} \psi_+ + \frac{\beta\widehat{n}}{2} \cdot \frac{1}{1 + \exp\left(z\right)} \left(\psi_+ - \frac{2z}{\beta \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)} \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \right) + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \\
&= -\widehat{n} \cdot \frac{z}{1 + \exp(z)} \left(\frac{1}{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)} \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \right) + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \\
&= -\widehat{n} \cdot \frac{\xi^{-1}\left(\frac{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)}{\widehat{n}} \right)}{1 + \exp\left(\xi^{-1}\left(\frac{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)}{\widehat{n}} \right) \right)} \left(\frac{1}{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)} \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \right) + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \\
&= -\widehat{n} \cdot \frac{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)}{\widehat{n}} \left(\frac{1}{\nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D})^\top \left(\theta^\dagger - \theta_\mu\right)} \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \right) + \nabla_\theta \ell^{\theta^\dagger}_{\mathrm{DPO}}(\overline{D}) \\
&= \mathbf{0} \ .
\end{aligned}$$

Strong convexity of $\ell^\theta_{\mathrm{DPO}}(D)$ (see Lemma E.14) implies that the first-order condition is enough to guarantee optimality. □

Next, we derive the DPO loss for loglinear policies and reference policy.

Lemma E.8. *The DPO loss for loglinear policy parametrization and loglinear reference policy μ can be written as*

$$\ell^\theta_{\mathrm{DPO}}(D) = \sum_{(s,a,a',o) \in D} \log\left(1 + \exp\left(-o\beta \left(\theta - \theta_\mu\right)^\top \left(\psi(s,a) - \psi(s,a')\right)\right)\right) + \frac{\lambda}{2} \left\|\theta - \theta_\mu\right\|^2 \ .$$

Proof. Note that

$$\begin{aligned}
\ell^\theta_{\mathrm{DPO}}\left(D\right) &= -\sum_{(\tau,\tau',o) \in D} \log \sigma\left(o \cdot \left(\beta \log \frac{\pi_\theta(a|s)}{\mu(a|s)} - \beta \log \frac{\pi_\theta(a'|s)}{\mu(a'|s)} \right) \right) + \frac{\lambda}{2} \left\|\theta - \theta_\mu\right\|^2 \\
&= \sum_{(\tau,\tau',o) \in D} \log\left(1 + \exp\left(-o \cdot \left(\beta \log \frac{\pi_\theta(a|s)}{\mu(a|s)} - \beta \log \frac{\pi_\theta(a'|s)}{\mu(a'|s)} \right) \right)\right) + \frac{\lambda}{2} \left\|\theta - \theta_\mu\right\|^2 \\
&= \sum_{(s,a,a',o) \in D} \log\left(1 + \exp\left(-o\beta \log \frac{\pi_\theta(a|s)\mu(a'|s)}{\mu(a|s)\pi_\theta(a'|s)} \right)\right) + \frac{\lambda}{2} \left\|\theta - \theta_\mu\right\|^2
\end{aligned}$$

$$= \sum_{(s,a,a',o)\in D} \log\left(1+\exp\left(-o\beta \log\frac{\exp(\theta^\top\psi(s,a))\exp(\theta_\mu^\top\psi(s,a'))\sum_{a''}\exp(\theta_\mu^\top\psi(s,a''))\sum_{a''}\exp(\theta^\top\psi(s,a''))}{\exp(\theta_\mu^\top\psi(s,a'))\exp(\theta^\top\psi(s,a'))\sum_{a''}\exp(\theta_\mu^\top\psi(s,a''))\sum_{a''}\exp(\theta^\top\psi(s,a''))}\right)\right) + \frac{\lambda}{2}\left\|\theta-\theta_\mu\right\|^2$$

$$= \sum_{(s,a,a',o)\in D} \log\left(1+\exp\left(-o\beta\left(\theta-\theta_\mu\right)^\top\left(\psi(s,a)-\psi(s,a')\right)\right)\right) + \frac{\lambda}{2}\left\|\theta-\theta_\mu\right\|^2 .$$

□

Next, we prove that the KL constraints are convex for loglinear policies.

Lemma E.9. *Let $\pi_\theta \in \Pi^{\log}$ be a loglinear policy with respect to feature mapping ψ. Then, $D_{\mathrm{KL}}(\pi^\dagger||\pi_\theta)$ is convex.*

Proof. Using the gradient of π_θ from the proof of Lemma E.1, we have

$$\begin{aligned}\nabla_\theta D_{\mathrm{KL}}(\pi^\dagger||\pi_\theta) &= \nabla_\theta \sum_{s,a}\rho(s)\pi^\dagger(a|s)\left(\log\pi^\dagger(a|s)-\log\pi_\theta(a|s)\right)\\ &= \nabla_\theta\left(\sum_s \rho(s)\log\sum_a\exp\left(\theta^\top\psi(s,a)\right)-\sum_{s,a}\rho(s)\pi^\dagger(a|s)\theta^\top\psi(s,a)\right)\\ &= \sum_{s,a}\rho(s)\pi_\theta(a|s)\psi(s,a)-\sum_{s,a}\rho(s)\pi^\dagger(a|s)\psi(s,a) .\end{aligned}$$

Furthermore, note that for the Hessian we have

$$\begin{aligned}\nabla_\theta^2 D_{\mathrm{KL}}(\pi^\dagger||\pi_\theta) &= \nabla_\theta\left(\sum_{s,a}\rho(s)\pi_\theta(a|s)\psi(s,a)-\sum_{s,a}\rho(s)\pi^\dagger(a|s)\psi(s,a)\right)\\ &= \sum_{s,a}\rho(s)\pi_\theta(a|s)\left(\psi(s,a)-\sum_{a'}\pi_\theta(a'|s)\psi(s,a')\right)\psi(s,a)^\top\\ &= \sum_s\rho(s)\left(\sum_a\pi_\theta(a|s)\psi(s,a)\psi(s,a)^\top-\left(\sum_a\pi_\theta(a|s)\psi(s,a)\right)\left(\sum_a\pi_\theta(a|s)\psi(s,a)\right)^\top\right)\\ &= \sum_s\rho(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\left(\psi(s,a)-\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\psi(s,a')\right]\right)\left(\psi(s,a)-\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\psi(s,a')\right]\right)^\top\right]\\ &\succeq 0 .\end{aligned}$$

□

Next, we provide the necessary condition for the constructed dataset to satisfy the feature conditions.

Lemma E.10. *Let ω_0 be a given parameter and f be an arbitrary function of ω_0. Assume that γ is such that*

$$1-\gamma \le \frac{2\left\|\omega_0\right\|}{\xi_{\max}+1} .$$

Then, if there exist trajectory pairs (τ,τ'), for which

$$\phi(\tau)-\phi(\tau') = \xi^{-1}\left(f(\omega_0)\left\lceil\frac{f(\omega_0)}{\xi_{\max}}\right\rceil^{-1}\right)\frac{\omega_0}{\left\|\omega_0\right\|} ,$$

they are feasible in the feature space, in the sense that they satisfy $\left\|\phi(\tau)\right\|_2, \left\|\phi(\tau')\right\|_2 \le 1$.

Proof. First, we will consider the case when the range of f is non-negative, i.e., $f(\omega_0) \geq 0$. For this case, note that

$$0 \leq f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \leq \xi_{\max} = W_{\text{Lam}}(1/e) < 0.3 \ ,$$

where we recall that

$$\xi_{\max} = \max_t \frac{t}{1+\exp(t)} \ .$$

Now, note that we can write

$$\xi^{-1}(a) = a - W_{\text{Lam}}(-a\exp(a)) \ , \tag{39}$$

for any $a \geq -1/e$, since, if we let $t^* = a - W_{\text{Lam}}(-a\exp(a))$, we obtain

$$\frac{t^*}{1+\exp(t^*)} = \frac{a - W_{\text{Lam}}(-a\exp(a))}{1+\exp(a - W_{\text{Lam}}(-a\exp(a)))} = \frac{a + a\exp(a)/\exp(a - W_{\text{Lam}}(-a\exp(a)))}{1+\exp(a - W_{\text{Lam}}(-a\exp(a)))} = a \ .$$

Our aim is to show that we can find pairs (τ, τ') that satisfy

$$\phi(\tau) - \phi(\tau') = \xi^{-1}\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \frac{\omega_0}{\|\omega_0\|} \ .$$

This amounts to showing the right-hand side of the above equation satisfies the feature boundedness, i.e., that it is consistent with the assumption that $\|\phi(\tau)\| \leq 1/(1-\gamma)$, for all τ.

To that end, we will derive upper bounds on the norm of the quantity on the right hand side of the equation above. First, we consider the $\xi^{-1}(\cdot)$ term. We have

$$\begin{aligned}
&\left| \xi^{-1}\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right| \\
&= \left| f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} - W_{\text{Lam}}\left(-f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \exp\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right) \right| && (40) \\
&\leq \left| f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right| + \left| W_{\text{Lam}}\left(-f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \exp\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right) \right| && (41) \\
&\leq \xi_{\max} + \left| \log\left(1 - f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \exp\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right) \right| && (42) \\
&\leq \xi_{\max} + \left| \log\left(1 - \xi_{\max}\exp\left(\xi_{\max}\right)\right) \right| \ , && (43)
\end{aligned}$$

where Equation 40 follows from Equation 39; Equation 41 follows from the triangle inequality; Equation 42 follows from Theorem 2.3 of [Hoorfar and Hassani, 2008], where we let $y = 1$; Equation 43 follows from the fact that the function $\log(1 - xe^x)$ is negative and decreasing in the range $(0, \xi_{\max})$.

Since we should have

$$\|\phi(\tau) - \phi(\tau')\| \leq \|\phi(\tau)\| + \|\phi(\tau')\| \leq \frac{2}{1-\gamma} \ ,$$

for the construction to be feasible in this case, we need

$$\left\| \xi^{-1}\left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \frac{\omega_0}{\|\omega_0\|} \right\| \leq \frac{\xi_{\max} + \left| \log\left(1 - \xi_{\max}\exp\left(\xi_{\max}\right)\right) \right|}{\|\omega_0\|} \leq \frac{2}{1-\gamma} \ .$$

Next, let us now consider the case when $f(\omega_0) \leq 0$. In this case, the dataset construction uses the term

$$f(\omega_0) \left\lceil \frac{|f(\omega_0)|}{\xi_{\max}} \right\rceil^{-1} \ .$$

Note that, since $f(\omega_0) \leq 0$, we have

$$-\xi_{\max} \leq f(\omega_0) \left\lceil \frac{|f(\omega_0)|}{\xi_{\max}} \right\rceil^{-1} \leq -\frac{|f(\omega_0)|\,\xi_{\max}}{|f(\omega_0)| + \xi_{\max}} \leq 0 \ .$$

In this case, we obtain

$$\left| \xi^{-1} \left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right| \leq \xi_{\max} + \left| \log \left(1 - f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \exp \left(f(\omega_0) \left\lceil \frac{f(\omega_0)}{\xi_{\max}} \right\rceil^{-1} \right) \right) \right| \tag{44}$$

$$\leq \xi_{\max} + 1 \ , \tag{45}$$

where Equation 44 follows directly from Equation 42, while Equation 45 follows from the fact that the maximum value of $\log(1 - xe^x)$ in the interval $(-\infty, 0)$ is upper bounded by 1.

Thus, combining both cases, we obtain the condition

$$\max \left\{ \frac{\xi_{\max} + |\log\left(1 - \xi_{\max} \exp\left(\xi_{\max}\right)\right)|}{\|\omega_0\|}, \frac{\xi_{\max} + 1}{\|\omega_0\|} \right\} = \frac{\xi_{\max} + 1}{\|\omega_0\|} \leq \frac{2}{1-\gamma} \ .$$

This finally implies that we should pick γ such that

$$1 - \gamma \leq \frac{2\,\|\omega_0\|}{\xi_{\max} + 1} \ .$$

□

Next, we provide results characterizing the spectra of matrices relevant to our setting.

Lemma E.11. *We have*

$$\sigma_{\max}\left(M_{\pi^\dagger}\right) \leq \sqrt{SA}$$

and

$$\sigma_{\max}\left(\Sigma_D\right) \leq \frac{1}{1-\gamma} \ .$$

Proof. Note that

$$\begin{aligned} \|M_{\pi^\dagger}\| &= \sqrt{\left\|M_{\pi^\dagger}^\top M_{\pi^\dagger}\right\|} = \sqrt{\left\|M_{\pi^\dagger} M_{\pi^\dagger}^\top\right\|} \\ &\leq \sqrt{2\min\{SA, d\} \max_{s,a} \|\phi(s,a)\|} \\ &\leq \sqrt{2\min\{SA, d\}} \ . \end{aligned}$$

Moreover, we have that

$$\begin{aligned} \sigma_{\max}\left(\Sigma_D\right) &= \sigma_{\max}\left(\frac{1}{n} \sum_{(\tau,\tau') \in D} \left(\phi(\tau) - \phi(\tau')\right) \left(\phi(\tau) - \phi(\tau')\right)^\top \right) \\ &\leq \frac{1}{n} \sum_{(\tau,\tau') \in D} \sigma_{\max}\left(\left(\phi(\tau) - \phi(\tau')\right) \left(\phi(\tau) - \phi(\tau')\right)^\top \right) \\ &\leq \frac{1}{n} \sum_{(\tau,\tau') \in D} Tr\left(\left(\phi(\tau) - \phi(\tau')\right) \left(\phi(\tau) - \phi(\tau')\right)^\top \right) \\ &= \frac{1}{n} \sum_{(\tau,\tau') \in D} \left(\phi(\tau) - \phi(\tau')\right)^\top \left(\phi(\tau) - \phi(\tau')\right)^\top \end{aligned}$$

$$\leq \frac{1}{n} \sum_{(\tau,\tau') \in D} (\|\phi(\tau)\| + \|\phi(\tau')\|)$$
$$\leq \frac{2}{1-\gamma} ,$$

where $Tr(M)$ denotes the trace of matrix M. □

Next, we provide a characterization of the spectral properties of a matrix of interest for the RLHF setting.

Lemma E.12. *Given parameter ω, let $Y_{\overline{D}}^{\omega}$ be defined as in Equation 12. Then, the following inequalities hold:*

$$\overline{n}\widetilde{C}_{\omega,\overline{D}}^{\phi} C_{\overline{D}}^{\phi} + 2\lambda \leq \sigma_{\min}\left(Y_{\overline{D}}^{\omega}\right) \leq \sigma_{\max}\left(Y_{\overline{D}}^{\omega}\right) \leq \overline{n}\sigma_{\max}\left(\Sigma_{\overline{D}}^{\phi}\right) + 2\lambda \leq \frac{\overline{n}}{1-\gamma} + 2\lambda , \tag{46}$$

and

$$\frac{1-\gamma}{\overline{n} + 2(1-\gamma)\lambda} \leq \sigma_{\min}\left((Y_{\overline{D}}^{\omega})^{-1}\right) \leq \sigma_{\max}((Y_{\overline{D}}^{\omega})^{-1}) \leq \frac{1}{\overline{n}\widetilde{C}_{\omega,\overline{D}}^{\phi} C_{\overline{D}}^{\phi} + 2\lambda} ,$$

where

$$\widetilde{C}_{\omega,\overline{D}}^{\phi} = \min_{(\tau,\tau',o) \in \overline{D}} \frac{\exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))}{(1 + \exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))^2} .$$

Proof. The first inequality in Equation 46 follows from

$$\sigma_{\min}(Y_{\overline{D}}^{\omega}) \geq \sigma_{\min}\left(\sum_{(\tau,\tau',o) \in \overline{D}} \frac{\exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))}{(1 + \exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))^2} (\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^{\top} + 2\lambda I\right)$$
$$\geq \widetilde{C}_{\omega,\overline{D}}^{\phi} \sigma_{\min}\left(\sum_{(\tau,\tau',o) \in \overline{D}} (\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^{\top}\right) + 2\lambda \tag{47}$$
$$\geq \overline{n}\widetilde{C}_{\omega,\overline{D}}^{\phi} C_{\overline{D}}^{\phi} + 2\lambda , \tag{48}$$

where Equation 47 uses the definition of $\widetilde{C}_{\omega,\overline{D}}^{\phi}$ and Equation 48 follows by the assumption on Σ_D^{ϕ}. The upper bound in Equation 46 follows from

$$\left\|Y_{\overline{D}}^{\omega}\right\| = \sigma_{\max}\left(Y_{\overline{D}}^{\omega}\right)$$
$$= \sigma_{\max}\left(\sum_{(\tau,\tau',o) \in \overline{D}} \frac{\exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))}{(1 + \exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))^2} (\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^{\top} + 2\lambda I\right)$$
$$\leq \sum_{(\tau,\tau',o) \in \overline{D}} \frac{\exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))}{(1 + \exp(o\omega^{\top}(\phi(\tau) - \phi(\tau')))^2} \sigma_{\max}\left((\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^{\top}\right) + 2\lambda$$
$$\leq \sum_{(\tau,\tau',o) \in \overline{D}} \frac{1}{2} Tr\left((\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^{\top}\right) + 2\lambda \tag{49}$$
$$= \sum_{(\tau,\tau',o) \in \overline{D}} \frac{1}{2} (\phi(\tau) - \phi(\tau'))^{\top}(\phi(\tau) - \phi(\tau')) + 2\lambda$$
$$\leq \frac{\overline{n}}{1-\gamma} + 2\lambda , \tag{50}$$

where Equation 49 follows from the fact that

$$\sigma_{\max}(vv^{\top}) \leq Tr\left(vv^{\top}\right) = v^{\top}v ,$$

for any nonzero vector v, and the fact that

$$\frac{x}{(1+x)^2} \leq \frac{1}{2}$$

for all positive x; Equation 50 follows from

$$\left(\phi(\tau) - \phi(\tau')\right)^\top \left(\phi(\tau) - \phi(\tau')\right) \leq 2 \max_\tau \|\phi(\tau)\| = \frac{2}{1-\gamma} \ .$$

This concludes the first part of our result.

For the characterization of the eigenspectrum of the inverse of $Y^\omega_{\overline{D}}$, we make use of the above derivations and immediately observe that

$$\sigma_{\min}\left((Y^\omega_{\overline{D}})^{-1}\right) \geq \frac{1}{\sigma_{\max}(Y^\omega_{\overline{D}})} \geq \frac{1-\gamma}{\overline{n} + 2(1-\gamma)\lambda} \ , \tag{51}$$

and

$$\sigma_{\max}((Y^\omega_{\overline{D}})^{-1}) \leq \frac{1}{\overline{n}\widetilde{C}^\phi_{\omega,\overline{D}} C^\phi_{\overline{D}} + 2\lambda} \ . \tag{52}$$

□

Next, we show that the KL divergence of two loglinear policies is Lipshcitz with respect to their respective parameters.

Lemma E.13. *Let π_θ and $\pi_{\theta'}$ be two loglinear policies with respect to feature mapping ψ. Then,*

$$D_{\mathrm{KL}}\left(\pi_\theta || \pi_{\theta'}\right) \leq 2 \left\|\theta - \theta'\right\| \ .$$

Proof. Note that

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi_\theta || \pi_{\theta'}\right) = \sum_{s,a} \pi_\theta(a|s) \left(\log \pi_\theta(a|s) - \log \pi_{\theta'}(a|s)\right) \\
&= \sum_{s,a} \pi_\theta(a|s) \left(\theta - \theta'\right)^\top \psi(s,a) + \sum_s \rho(s) \left(\log \sum_{a'} \exp\left((\theta')^\top \psi(s,a')\right) - \log \sum_{a'} \exp\left(\theta^\top \psi(s,a')\right) \right) \\
&\leq \sum_{s,a} \rho(s) \pi_\theta(a|s) \left\|\theta - \theta'\right\| \left\|\psi(s,a)\right\| + \sum_s \rho(s) \left\|\theta - \theta'\right\| \left\|\psi(s,a)\right\| \\
&\leq 2 \left\|\theta - \theta'\right\| \ ,
\end{aligned}$$

where the third inequality uses Cauchy-Schwarz and the fact that the log-sum-exp function is Lipshcitz with parameter 1, since

$$\left\| \nabla_\theta \sum_s \rho(s) \log \sum_a \exp\left(\theta^\top \psi(s,a)\right) \right\| = \left\| \sum_{s,a} \rho(s) \pi_\theta(a|s) \psi(s,a) \right\| \leq \max_{s,a} \|\psi(s,a)\| \leq 1 \ .$$

□

Our next result states some nice properties of the regularized logistic regression loss. These implications are easy to prove. Nevertheless, we provide the full proofs for completion.

Lemma E.14. *Given dataset $D = \{(x_i, y_i)\}_{i=1}^n$, let*

$$\ell^\upsilon(D) = \sum_{(x,y)\in D} \log\left(1 + \exp\left(\beta \cdot y \upsilon^\top x + b\right)\right) + \frac{\lambda}{2} \left\|\upsilon - \zeta\right\|^2$$

denote a regularized logistic regression loss with respect to $\upsilon \in \mathbb{R}^d$, for some $d \in \mathbb{N}$, where $b \in \mathbb{R}$, $\zeta \in \mathbb{R}^d$, and $\lambda > 0$. Moreover, let

$$\Sigma_D = \frac{1}{n} \sum_{x \in D} x x^\top$$

be the data covariance matrix with minimum eigenvalue σ, and let υ^* denote an optimal point for $\ell^\upsilon(D)$. Then, the following hold:

1. *The function $\ell^\upsilon(D)$ is strongly convex with parameter $n\beta C_\upsilon \sigma + \lambda$.*
2. *We have $\|\nabla_\upsilon \ell_\upsilon(D)\| \geq 2(n\beta C_\upsilon \sigma + \lambda) \|\upsilon - \upsilon^*\|$.*
3. *The function*

$$X_D^\upsilon = \sum_{(x,y) \in D} \frac{\beta \cdot y}{1 + \exp\left(\beta \cdot y \upsilon^\top x + b\right)} x$$

   is Lipschitz with parameter n.
4. *We have $\|\nabla_\upsilon \ell^\upsilon(D)\| \leq (n\beta + \lambda) \|\upsilon - \upsilon^*\|$.*

Proof. First, note that the Hessian of $\ell^\upsilon(D)$ can be written as

$$\begin{aligned}
\nabla_\upsilon^2 \ell^\upsilon(D) &= \sum_{(x,y) \in D} \frac{\beta \exp(\beta \cdot y \upsilon^\top x + b)}{\left(1 + \exp(\beta \cdot y \upsilon^\top x + b)\right)^2} x x^\top + \lambda I \\
&\succeq \beta n C_\upsilon \frac{1}{n} \sum_{(x,y) \in D} x x^\top + \lambda I \\
&\succeq (n\beta C_\upsilon \sigma + \lambda) I \ ,
\end{aligned}$$

where we have only used the assumptions of the statement. This means that $\ell^\upsilon(D)$ is strongly convex with parameter $n\beta C_\upsilon \sigma + \lambda$. The second statement is a direct implication of this.

For the third statement, first note that we have

$$\begin{aligned}
\|X_D^\upsilon\| &= \left\| \sum_{(x,y) \in D} \frac{-\beta x}{1 + \exp\left(\beta \cdot y \upsilon^\top x + b\right)} \right\| \\
&\leq \beta n.
\end{aligned}$$

Then, we can write

$$\begin{aligned}
\|\nabla_\upsilon \ell^\upsilon(D)\| &= \left\| \nabla_\upsilon \ell^\upsilon(D) - \nabla^\upsilon \ell^{\upsilon^*}(D) \right\| \\
&= \left\| X_D^\upsilon + \lambda(\upsilon - \zeta) - X_D^{\upsilon^*} + \lambda(\upsilon^* - \zeta) \right\| \\
&\leq n\beta \|\upsilon - \upsilon^*\| + \lambda \|\upsilon - \upsilon^*\| \\
&\leq (n\beta + \lambda) \|\upsilon - \upsilon^*\| \ ,
\end{aligned}$$

where the penultimate inequality uses Lipschitzness of X_D^υ. □

The next result is used for the solution of the logistic regression subproblems.

Theorem E.15 (Proposition 3 of [Liu and Zhu, 2016])**.** *Given any target model $\omega^\dagger \neq \mathbf{0}$, the following is a teaching set for the logistic regression problem*

$$\min_{D'} |D'| \ \text{ such that } \ \omega^\dagger \in \arg\min_{\omega} \sum_{(x_i, y_i) \in D'} \log\left(1 + \exp\left(-y_i x_i^\top \omega\right)\right) + \frac{\lambda}{2} \|\omega\|^2 \ .$$

There are $\widehat{n} = \left\lceil \frac{\lambda \left\| \omega^\dagger \right\|^2}{\xi_{\max}} \right\rceil$ *identical training samples, each taking the form*

$$x_i = \xi^{-1}\left(\frac{\lambda \left\| \omega^\dagger \right\|^2}{\left\lceil \frac{\lambda \|\omega^\dagger\|^2}{\xi_{\max}} \right\rceil} \right) \frac{\omega^\dagger}{\left\| \omega^\dagger \right\|^2}, \quad y_i = 1 \ .$$