Reward Model Learning vs. Direct Policy Optimization: A Comparative Analysis of Learning from Human Preferences

Andi Nika [1] **Debmalya Mandal** [2] **Parameswaran Kamalaruban** [3] **Georgios Tzannetos** [1]
Goran Radanović [1] **Adish Singla** [1]

Abstract

In this paper, we take a step towards a deeper understanding of learning from human preferences by systematically comparing the paradigm of reinforcement learning from human feedback (RLHF) with the recently proposed paradigm of direct preference optimization (DPO). We focus our attention on the class of loglinear policy parametrization and linear reward functions. In order to compare the two paradigms, we first derive minimax statistical bounds on the suboptimality gap induced by both RLHF and DPO, assuming access to an oracle that exactly solves the optimization problems. We provide a detailed discussion on the relative comparison between the two paradigms, simultaneously taking into account the sample size, policy and reward class dimensions, and the regularization temperature. Moreover, we extend our analysis to the approximate optimization setting and derive exponentially decaying convergence rates for both RLHF and DPO. Next, we analyze the setting where the ground-truth reward is not realizable and find that, while RLHF incurs a constant additional error, DPO retains its asymptotically decaying gap by just tuning the temperature accordingly. Finally, we extend our comparison to the Markov decision process setting, where we generalize our results with exact optimization. To the best of our knowledge, we are the first to provide such a comparative analysis for RLHF and DPO.

[1]Max Planck Institute for Software Systems, Saarbrücken, Germany [2]University of Warwick, Coventry, UK [3]Independent Researcher, London, UK. Correspondence to: Andi Nika <andinika@mpi-sws.org>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Learning from human preferences has grown more prominent as we move closer to artificial general intelligence. One of the most effective ways to learn from preferences is through reinforcement learning from human feedback (RLHF), which involves a two-step process of reward learning and regularized policy optimization. The attractiveness of this paradigm lies in its ability to model the reward function based solely on preference data. This makes it highly applicable in numerous practical situations where rewards are not given *a priori* or are challenging to define accurately. Once the reward is modeled, RLHF solves a regularized value function maximization problem to obtain a fine-tuned policy. This paradigm has enjoyed a lot of applications varying from game-playing (Christiano et al., 2017; Warnell et al., 2018; Knox & Stone, 2008; MacGlashan et al., 2017), robotics (Shin et al., 2023; Brown et al., 2019), and training large language models (LLMs) (Ziegler et al., 2019; Nakano et al., 2021; Wu et al., 2021; Ouyang et al., 2022; Stiennon et al., 2020; Glaese et al., 2022; Ramamurthy et al., 2023; Menick et al., 2022; Ganguli et al., 2022; Bai et al., 2022; Gao et al., 2023).

As an alternative to RLHF, Rafailov et al. (2023) have recently proposed direct preference optimization (DPO), an RL-free paradigm to learning from preferences. DPO circumvents the reward modeling phase and directly optimizes the policy parameters based on the preference data. In certain LLM instances, DPO seems to be empirically superior to RLHF, due to its simple optimization framework.

That said, a statistical analysis of the differences between these paradigms is lacking. The sample complexity of RLHF in various settings has already been studied (Zhu et al., 2023; Zhan et al., 2023; Xiong et al., 2023), and there have been some initial attempts at theoretically understanding DPO and its variants (Azar et al., 2023). However, it is unclear when one of the paradigms is better and when these two paradigms are statistically comparable. Motivated by this observation, we initiate a thorough discussion of the theoretical comparison between RLHF and DPO. Specifically, the purpose of this paper is to address the following research questions:

		RLHF	DPO
Realizable Rewards	Exact Optimization	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \widetilde{\Theta}\left(\sqrt{\frac{d_R}{n}}\right)$	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \Theta\left(\frac{d_P}{\beta n}\right)$
	Approximate Optimization	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \widetilde{\Theta}\left(\sqrt{\frac{d_R}{n}}\right) + O\left((1-\frac{1}{n})^t + \frac{e^{-t}}{\beta}\right)$	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \Theta\left(\frac{d_P}{\beta n}\right) + O\left(\frac{1}{\beta}\left(1-\frac{\beta}{n}\right)^t\right)$
Non-realizable Rewards	Exact Optimization	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \widetilde{\Theta}\left(\sqrt{\frac{d_R}{n}}\right) + O(\epsilon_{\text{app}})$	$O\left(\beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}\|\|\mu)\right) + \Theta\left(\frac{d_P}{\beta n}\right) + O\left(\beta D_{\mathrm{KL}}\left(\pi_{\theta^*}\|\|\pi^*_{r^*}\right)\right)$

Table 1: A presentation of the bounds on the suboptimality gap for RLHF and DPO. The first two rows present bounds under the realizable reward assumption in the exact and approximate optimization frameworks; the last row presents the bounds when the ground-truth reward function is not realizable. Here, $\pi^{\mathrm{opt}}_{r^*}$ denotes an optimal policy with respect to the ground-truth reward function r^*, $\pi^*_{r^*}$ denotes an optimal regularized policy, and π_{θ^*} denotes an optimal loglinear regularized policy. Moreover, β denotes the regularization temperature, D_{KL} denotes the KL divergence, d_R denotes the reward dimension, and d_P denotes the policy dimension. Finally, n denotes the sample size, t denotes the optimization steps for the approximate setting, ϵ_{app} denotes the reward mismatch coefficient and $\widetilde{\Theta}$ hides any $\log$ factors.

What are the statistical guarantees of RLHF relative to those of DPO? What conditions benefit one as opposed to the other?

As DPO does not learn a reward model, but directly optimizes over the policy space, a dependence on the policy dimensionality d_P is expected. On the other hand, RLHF's performance evidently implies some dependence on reward dimensionality d_R due to its reward learning phase. Does this imply a discrepancy in the statistical bounds of these paradigms when the reward and policy dimensions are different? Moreover, what can be said about the dependency of the bounds on the sample size n or the regularization temperature β?

We address these questions in the following setting: finite spaces, Bradley-Terry preference model, linear rewards and loglinear policies. We first study the exact optimization setting and derive bounds on both RLHF and DPO. Then, we proceed to derive fast convergence rates for a modified version of the policy gradient for RLHF, and gradient descent for DPO. Next, we discuss some implications of our bounds when the reward function is not fully realizable. We close our paper by extending our comparative analysis to deterministic Markov decision processes. Our contributions are summarized below; see Table 1 for explicit bounds.

- First, we derive minimax bounds on the suboptimality gap induced by RLHF and DPO in the exact optimization setting by leveraging smoothness and strong convexity properties. We show that when the optimal regularized policy is loglinear and the reward function is linear, RLHF is $\widetilde{\Theta}(\sqrt{d_R/n})$-close to its objective, while DPO is $\widetilde{\Theta}(d_P/(\beta n))$-close. These results emphasize the comparison of the two paradigms in terms of the reward and policy dimensions when setting $\beta = \Theta(\sqrt{d_P/n})$ for DPO.
- Furthermore, we study the convergence rates of a version of the natural policy gradient for RLHF and gradient descent for DPO. Motivated by recent fast convergence results for entropy regularized RL with tabular softmax policies, we derive $O\left(e^{-t}/\beta\right)$ convergence rates in t iterations for a version of the natural policy gradient for RLHF. Moreover, for gradient descent we are able to show $O\left((1/\beta)(1-\beta/n)^t\right)$ convergence rates by using the fact that the DPO loss function satisfies the *PL condition* (Karimi et al., 2016). These results replicate the implications from the exact optimization setting on the difference in terms of reward and policy dimensions.
- We also consider the case where the ground-truth reward function is not realizable and its best linear fit is ϵ_{app}-close to it. We show that, while RLHF incurs an additional constant term on the suboptimality gap, DPO's dependence on the additional term can be controlled by setting the regularization temperature accordingly.
- Finally, we extend our comparison to deterministic Markov decision processes by proposing a new formulation of the DPO objective for this setting and then generalizing our results. The main motivation for this extension is that, arguably, the difference in reward and policy dimensions in this setting is higher.

2. Related Work

Learning from pairwise comparisons. In the context of RL, the problem of learning from pairwise comparisons has been studied thoroughly in the bandit setting, where the problem is known as the *dueling bandit* problem (Yue et al., 2009; Faury et al., 2020; Ailon et al., 2014; Gajane et al., 2015; Komiyama et al., 2015; Zoghi et al., 2014; Saha & Gopalan, 2019; Saha & Krishnamurthy, 2022). For the case

of dueling RL for linear MDPs, Saha et al. (2023) propose an algorithm that satisfies tight regret guarantees, while Chen et al. (2022) extend this formulation to the MDPs with general function approximation. Finally, Chatterji et al. (2021) consider a more general setting in which the trajectory-based feedback is generated from a generalized linear model, and they propose variants of optimistic algorithms for online RL.

In this paper, we consider the offline setting, where Zhu et al. (2023) and Zhan et al. (2023) have already provided statistical bounds on pessimistic RLHF for direct value maximization. Our focus, however, is on the regularized value maximization problem. While pessimism mitigates poor coverage for the setting considered by Zhu et al. (2023) and Zhan et al. (2023), for regularized RLHF and, as a consequence, for DPO, this issue remains and is captured by the coverage coefficients which we define with respect to both reward and policy features. The aim of this paper, however, is not to mitigate these issues, but to thoroughly analyse the statistical guarantees of the existing prominent paradigms of learning from human preferences and derive insightful results that shed light on their relative performances.

Direct preference optimization. In recent works, the RL-free fine-tuning paradigm of direct preference optimization (DPO) has gained popularity (Rafailov et al., 2023; An et al., 2023; Azar et al., 2023; Wang et al., 2023). Its original formulation was proposed for the contextual bandit setting. Hejna et al. (2023) propose an extension of DPO to MDPs under the assumption that the preferences depend on the advantage function of the optimal policy. While we also provide an extension of the DPO formulation for MDPs in Section 7, our primary focus is on a comparative analysis between the two paradigms in the contextual bandit setting.

Offline RL. In recent years, there has been a significant surge in interest towards offline RL, with an extensive literature both in the empirical front (Jaques et al., 2019; Laroche et al., 2019; Fujimoto et al., 2019; Kumar et al., 2020; Agarwal et al., 2020; Kidambi et al., 2020) and the theoretical one (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Uehara & Sun, 2021; Zanette et al., 2021). While the focus of this line of work is on the traditional reward-based offline RL, our problem is derived from a combination of reward-learning from pairwise feedback and KL-regularized offline RL based on it.

3. Formal Setting

This section presents the background material that will be used throughout the paper. We will use a notation similar to (Rafailov et al., 2023) and (Azar et al., 2023).

Notation. Let $\langle u, v\rangle = u^\top v$ denote the inner product between vectors u and v. The trace of a matrix A is denoted by $tr(A)$ and its pseudo-inverse by $A^\dagger$. Moreover, $\Delta(\mathcal{X})$ denotes the set of distributions over the finite set $\mathcal{X}$ and $\|v\|_M = \sqrt{v^\top M v}$ denotes the seminorm of vector v with respect to M. Finally, $\text{proj}_A(v)$ denotes the projection of vector v onto set A and $\widetilde{\Theta}(\cdot)$ hides any log factors.

3.1. Preliminaries

Let $\mathcal{X}$ be a finite set of contexts with cardinality X and $\mathcal{Y}$ be a finite set of actions with cardinality Y. Fix $\rho \in \Delta(\mathcal{X})$ as an initial distribution over contexts and let $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a reward function. We consider a class of linear reward functions defined below.

Definition 3.1 (*Linear reward function class*)**.** Let ϕ be a d_R-dimensional feature mapping with $\max_{x,y} \|\phi(x, y)\|_2 \leq 1$ and let $F > 0$. We consider the following class of linear reward functions:

$$\mathcal{F} = \Big\{ r_\omega \in [0, 1]^{XY} : r_\omega(x, y) = \omega^\top \phi(x, y), \\ \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \text{ where } \omega \in \mathbb{R}^{d_R} \text{ and } \|\omega\|_2 \leq F \Big\}.$$

Given $x \in \mathcal{X}$, a policy $\pi(\cdot|x) \in \Delta(\mathcal{Y})$ is a distribution over actions. Throughout the paper, we will consider the loglinear class of policies, defined as follows.

Definition 3.2 (*Loglinear policy class*)**.** Let ψ be a d_P-dimensional feature mapping with $\max_{x,y} \|\psi(x, y)\|_2 \leq 1$ and let $B > 0$. We consider the following class of loglinear policies:

$$\Pi = \Big\{ \pi_\theta : \pi_\theta(y|x) = \frac{\exp(\theta^\top \psi(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x, y'))}, \\ \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \text{ where } \theta \in \mathbb{R}^{d_P} \text{ and } \|\theta\|_2 \leq B \Big\}.$$

Given policy π, the value function of π with respect to reward function r and context distribution ρ is defined as

$$V_r^\pi(\rho) = \sum_x \rho(x) \sum_y \pi(y|x) r(x, y).$$

3.2. Offline Learning from Human Preferences

Let μ be a reference policy fixed throughout the paper, and $\beta > 0$ be a regularization parameter. Let us define the KL-regularized objective as

$$\mathcal{V}_r^\pi(\rho) = V_r^\pi(\rho) - \beta D_{\text{KL}}(\pi||\mu),$$

where $D_{\text{KL}}(\pi||\mu) = \sum_x \rho(x) \sum_y \pi(y|x) \log \frac{\pi(y|x)}{\mu(y|x)}$.

We assume access to the dataset $\mathcal{D}_n = \{(x_i, y_i^w, y_i^l)\}_{i=1}^n$, where y_i^w denotes the preferred action over y_i^l. In this paper,

we will assume that the distribution of human preferences follows the Bradley-Terry (BT) model (Bradley & Terry, 1952), which we formally state below.

Definition 3.3 (*Bradley-Terry preference model*)**.** There exists a latent reward function r^* and a probability law P^* such that, for every tuple (x, y^w, y^l), we have

$$P^*(y^w \succ y^l|x) = \frac{\exp\left(r^*(x, y^w)\right)}{\exp\left(r^*(x, y^w)\right) + \exp\left(r^*(x, y^l)\right)} ,$$

where $y^w \succ y^l$ denotes y^w being preferred over y^l.

The latent reward function r^* will be fixed throughout the paper as the ground-truth reward function.

3.3. Reinforcement Learning from Human Feedback

We consider the reinforcement learning from human feedback (RLHF) paradigm as formulated in (Ziegler et al., 2019). Having access to preference dataset $\mathcal{D}_n$ and a fixed reference policy μ, RLHF proceeds in two phases: the reward learning phase and the final KL-regularized reinforcement learning phase.

For the reward learning phase, RLHF estimates r^* by applying maximum likelihood estimation (MLE) to the dataset $\mathcal{D}_n$. The MLE optimization problem can be written as

$$\min_r \mathcal{L}^r_{\text{RLHF}}(\mathcal{D}_n) := -\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_n} \left[\log\left(\sigma\big(r(x^w, y^w) - r(x^l, y^l)\big)\right)\right] , \quad \text{(P1.1)}$$

where $\sigma(z) = 1/(1 + \exp(-z))$ denotes the sigmoid function. Let $\widehat{r}$ denote the solution of Problem (P1.1).

The final phase of RLHF consists of maximizing the KL-regularized objective with respect to $\widehat{r}$ by solving

$$\max_\pi \mathcal{V}^\pi_{\widehat{r}}(\mathcal{D}_n) := \mathbb{E}_{\substack{x\sim\mathcal{D}_n \\ y\sim\pi(\cdot|x)}} \left[\widehat{r}(x, y) - \beta \log \frac{\pi(y|x)}{\mu(y|x)}\right] . \quad \text{(P1.2)}$$

3.4. Direct Preference Optimization

Recently, alternative paradigms to RLHF have been studied. In particular, Rafailov et al. (2023) introduced direct preference optimization (DPO), a new fine-tuning paradigm that directly optimizes the policy parameters instead of going through the reward modeling phase. Their key observation is that the latent reward can be expressed in terms of its optimal policy and the reference policy. This yields a loss function that is directly defined in terms of the preference data.

Formally, Rafailov et al. (2023) show that there exists a policy π that maximizes the KL-regularized objective, for which we have

$$r^*(x, y) = \beta \log \frac{\pi(y|x)}{\mu(y|x)} + \beta \log Z(x) \quad (1)$$

for every (x, y), where $Z(x)$ denotes the partition function $\sum_y \mu(y|x) \exp(r^*(x, y)/\beta)$. A new objective is then derived, which directly depends on the policy. Given preference dataset $\mathcal{D}_n$, this objective leads to the following optimization problem:

$$\min_\pi \mathcal{L}^\pi_{\text{DPO}}(\mathcal{D}_n) := -\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_n} \left[\log\left(\sigma\Big(\beta \log \frac{\pi(y^w|x)}{\mu(y^w|x))} - \beta \log \frac{\pi(y^l|x)}{\mu(y^l|x))}\Big)\right)\right] . \quad \text{(P2)}$$

As it turns out, this elegant approach yields practical benefits. However, it is unclear whether these benefits can be theoretically justified. In this paper, we will provide a comparative analysis of both RLHF and DPO in different settings. Next, we define a unified metric of performance to compare these paradigms.

3.5. Performance Metric

Given a reward function r, let $\pi^{\text{opt}}_r \in \arg\max_\pi V^\pi_r(\rho)$ denote an optimal policy for $V^\pi_r(\rho)$ and $V^{\text{opt}}_r(\rho)$ the optimal value. For a given policy π, we define the suboptimality gap of π as

$$G(\pi) = V^{\text{opt}}_{r^*}(\rho) - V^\pi_{r^*}(\rho) .$$

$G(\pi)$ captures how well a policy is performing w.r.t. the ground-truth reward function r^*. In this paper, we will use the suboptimality gap G as our unified measure of performance when comparing the two paradigms.

We note that RLHF and DPO are designed to minimize regularlized objectives $\mathcal{V}$ instead of optimizing the value function V (see Sections 3.3 and 3.4). In order to rigorously analyze the differences between RLHF and DPO, we need to establish some additional notation. Let $\pi^*_r \in \arg\max_\pi \mathcal{V}^\pi_r(\rho)$ denote a regularized optimal policy with respect to r and $\mathcal{V}^*_r(\rho)$ denote the optimal regularized value. Analogous to G, we define the regularized suboptimality gap of π as

$$\mathcal{G}(\pi) = \mathcal{V}^*_{r^*}(\rho) - \mathcal{V}^\pi_{r^*}(\rho) = (V^{\pi^*_{r^*}}_{r^*}(\rho) - V^\pi_{r^*}(\rho)) - \beta(D_{KL}(\pi^*_{r^*}||\mu) - D_{KL}(\pi||\mu)) .$$

As RLHF and DPO are designed to minimize $\mathcal{G}(\pi)$, both will incur an additional term on their bounds when comparing their bounds w.r.t. $G(\pi)$. For this reason, we will formally define this discrepancy term as $D(\pi) = G(\pi) - \mathcal{G}(\pi)$ and discuss it further in Section 4. Next, we proceed to provide a comparative analysis for these paradigms, starting with the exact optimization setting when the ground-truth reward is realizable. We note that any $\log$ terms are omitted for clarity of presentation. All proofs and related discussions can be found in the Appendix.

4. Realizable Rewards: Exact Optimization

In this section, we analyze the statistical differences in performance between RLHF and DPO in the exact optimization setting. We assume throughout the section that the ground-truth reward function is linear and realizable, i.e. $r^* \in \mathcal{F}$. Moreover, we assume a loglinear regularized optimal policy exists, i.e. $\pi^*_{r^*} \in \Pi$. Note that, for linear reward function, $\mathcal{L}^r_{\text{RLHF}}(\mathcal{D}_n)$ can be equivalently written as

$$-\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_n}\left[\log\sigma\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)\right]. \tag{2}$$

Moreover, for loglinear policies, $\mathcal{L}^\pi_{\text{DPO}}(\mathcal{D}_n)$ can be equivalently written as

$$-\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_n} \left[\log\left(\sigma\left(\beta\theta^\top\left(\psi(x,y^w)-\psi(x,y^l)\right)-J(x,y^w,y^l)\right)\right)\right], \tag{3}$$

where $J(x,y^w,y^l)=\log(\mu(y^w|x)/\mu(y^l|x))$.

We will denote the losses for the RLHF reward learning phase and DPO as $\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)$ and $\mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$, respectively. Let $r_{\widehat{\omega}}$ denote the reward estimate and let $\pi_{\widehat{\theta}}$ denote the policy learned by RLHF. Moreover, let $\pi_{\widetilde{\theta}}$ denote the policy learned by DPO. Formally, we assume that RLHF has access to an oracle that exactly solves both optimization problems and returns $r_{\widehat{\omega}} \in \arg\min \mathcal{L}^\omega_{RLHF}(\mathcal{D}_n)$ and $\pi_{\widehat{\theta}} \in \arg\max_\theta \mathcal{V}^{\pi_\theta}_{r_{\widehat{\omega}}}(\mathcal{D}_n)$. Similarly, for DPO we assume that the oracle returns $\pi_{\widetilde{\theta}} \in \arg\min_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$.

4.1. Theoretical Results

Before stating our main results of this section, we will need to define the covering numbers with respect to $\mathcal{D}_n$. Let the sample covariance matrix with respect to the reward features be defined as

$$\Sigma_{\mathcal{D}_n,R}=\frac{1}{n}\sum_{(x,y^w,y^l)\in\mathcal{D}_n}\overline{\phi}(x,y^w,y^l)\overline{\phi}(x,y^w,y^l)^\top,$$

where $\overline{\phi}(x,y^w,y^l) = \left(\phi(x,y^w)-\phi(x,y^l)\right)$, for every $(x,y^w,y^l)\in\mathcal{D}_n$. Fix $\lambda>0$ and let $\Lambda_R = \left\|(\Sigma_{\mathcal{D}_n,R}+\lambda I)^{-1/2}\right\|_2$ be the reward covering number.

Similarly, let the sample covariance matrix of the policy features be defined as

$$\Sigma_{\mathcal{D}_n,P}=\frac{1}{n}\sum_{(x,y^w,y^l)\in\mathcal{D}_n}\overline{\psi}(x,y^w,y^l)\overline{\psi}(x,y^w,y^l)^\top,$$

where $\overline{\psi}(x,y^w,y^l) = \left(\psi(x,y^w)-\psi\left(x,y^l\right)\right)$, for every $(x,y^w,y^l)\in\mathcal{D}_n$. The policy covering number of $\mathcal{D}_n$ is $\Lambda_P = \left\|(\Sigma_{\mathcal{D}_n,P}+\lambda I)^{-1/2}\right\|_2$. Our bounds will depend on these quantities. We will start with minimax bounds on the suboptimality for RLHF in the above-mentioned setting.

Theorem 4.1. *Let $\delta>0$. Assume that $r^*\in\mathcal{F}$. Then, with probability at least $1-\delta$, the suboptimality gap incurred by RLHF is*

$$G\left(\pi_{\widehat{\theta}}\right)=\Theta\left(\Lambda_R\sqrt{\frac{d_R+\log(6/\delta)}{S_R^2 n}+\lambda F^2}\right)+D\left(\pi_{\widehat{\theta}}\right),$$

where $S_R = 1/\left(2+\exp\left(-2F\right)+\exp\left(2F\right)\right)$.

Proof sketch. The proof follows from splitting the gap into sub-gaps which we can bound directly, and results from (Zhu et al., 2023). □

Next, we will consider the suboptimality gap induced by DPO. First, note that Equation (3) for loglinear policies essentially becomes a logistic regression problem in d_P+1 dimensions, by adding a dummy variable to θ that corresponds to $J(x,y^w,y^l)$. Our goal is to use loglinearity to derive smoothness properties for logistic regression so that we can obtain minimax bounds – without this assumption on the policy class, $\mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$ may not satisfy these properties.[1]

Moreover, note that Equation (1) relates r^* to one of the regularized optimal policies $\pi^*_{r^*}$. Obviously, this does imply that $\pi^*_{r^*}$ is loglinear. Nevertheless, the following lemma states that that is the case for linear rewards. Let $\Phi\in\mathbb{R}^{d_R\times XY}=[\phi(x,y)]_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ and $\Psi\in\mathbb{R}^{d_P\times XY}=[\psi(x,y)]_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ denote the reward feature and policy feature matrices, respectively.

Lemma 4.1. *Assume that $r^*\in\mathcal{F}$ and $\mu\in\Pi$. Furthermore, assume that the column space of Φ is a subspace of the column space of Ψ. Then, there exists $\theta^*\in\Theta$, such that $\pi_{\theta^*}\in\arg\max_\pi\mathcal{V}^\pi_{r^*}(\rho)$ and $r^*(x,y)=\beta\log(\pi_{\theta^*}(y|x)/\mu(y|x))+\beta\log Z(x)$.*

With these observations in place, we are now ready to state the minimax bounds on the suboptimality for DPO.

Theorem 4.2. *Let $\delta>0$ and $\beta>0$. Assume that $r^*\in\mathcal{F}$, $\mu\in\Pi$, and that the condition of Lemma 4.1 is satisfied. Let $n\geq O\left(tr(\Sigma^\dagger_{\mathcal{D}_n,P})/(\beta B^2)\right)$. Then, with probability at least $1-\delta$, the suboptimality gap of DPO is*

$$G\left(\pi_{\widetilde{\theta}}\right)=D\left(\pi_{\widetilde{\theta}}\right)+\Theta\left(\frac{\Lambda_P(d_P+1)}{\beta n}+\beta\lambda\Lambda_P B^2\right).$$

Proof sketch. We start by splitting the suboptimality gap and focus on the $\mathcal{G}(\pi_{\widetilde{\theta}})$ term. Here, we utilize the expression for the ground-truth reward as $r^*(x,y)=\beta\log\pi_{\theta^*}(y|x)-\beta\log\mu(y|x)+Z_{\widetilde{\theta}}(x)$, where $Z_{\widetilde{\theta}}(x)$ denotes the partition function with respect to $\pi_{\widetilde{\theta}}$. Then, using the fact that the policies are loglinear we further expand and reduce the whole

[1] Lemma J.6 shows this loss is not smooth for tabular settings.

gap in terms of the log differences. Next, we utilize the smoothness and Lipschitzness of the log-sum-exp function to finally obtain the upper bounds. For the lower bound, we construct an example where the policy feature matrix Ψ is full rank, and show that the log-sum-exp function becomes strongly convex. This finally leads to the stated bounds. □

Before discussing the implications of our results, let us say a few words on the regularization gap $D(\pi)$ for RLHF and DPO. Let π_{θ^*} denote an optimal loglinear regularized policy. A characterization of $D(\pi)$ is given as follows.

Lemma 4.2. *For any θ, we have that $\beta D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) - \beta D_{\mathrm{KL}}\left(\pi_\theta||\mu\right) \leq D(\pi_\theta)$ and $D(\pi_\theta) \leq \beta D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*}||\mu\right) - \beta D_{\mathrm{KL}}\left(\pi_\theta||\mu\right)$.*

Furthermore, for DPO, this quantity can be upper bounded in terms of $\pi^{\mathrm{opt}}_{r^*}$ and π_{θ^*} as follows.

Lemma 4.3. *Given $\delta > 0$, with probability at least $1 - \delta$, we have $D(\pi_{\widehat{\theta}}) \leq \beta D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*}||\mu\right) - \beta D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) + \widetilde{O}\left(d_P/n^{3/2}\right)$.*

In general, it is known that the KL divergence may not be upper-bounded. However, assuming that optimal policy $\pi^{\mathrm{opt}}_{r^*}$, optimal regularized policy π_{θ^*}, and sampling policy μ are not far away from each other, then these $D_{\mathrm{KL}}(\cdot)$ quantities would not be too large.

4.2. Comparative Analysis

In this section, we will provide some insights into the implications of our theoretical results. For the purpose of this section, we will focus our attention on the problem-dependent parameters and ignore the quantity $D(\pi)$.

The role of dimensionality. Note that RLHF has $\widetilde{\Theta}(\sqrt{d_R})$ dependence on the reward dimension, while DPO has $\Theta(d_P)$ dependence on the policy dimension. When $d_R = d_P$ and the sample size is small, RLHF seems to statistically outperform DPO. Any setting where $d_R \ll d_P$ makes this difference more apparent. In Section 7, we will discuss an extension of our analysis to a setting where the reward dimension can be much smaller than the policy dimension in practice.

The role of sample size. Next, we take into consideration the sample size. Note that DPO's bounds depend on n being large enough (cf. Theorem 4.2). Assume everything else constant and $d = d_R = d_P$. If the $D(\pi)$ terms are similar for both paradigms, then, for large sample sizes such that $n \gg d$, DPO seems to outperform RLHF asymptotically. Whenever $n < d$ (which is usually the case for large language models), RLHF has a smaller suboptimality gap.

The role of β. Finally, we discuss the role of the temperature β on the bounds. First, note that RLHF can effectively set $\beta = 0$ to annihilate the effect that $D(\pi)$ has on its bounds. On the other hand, DPO cannot set β to 0 due to a disproportional dependence of its bounds on it. Thus, the optimal choice of β for DPO is $\beta = \Theta(\sqrt{d_P/n})$, yielding $\Theta(\sqrt{d_P/n})$ bounds and matching the order of n in the bounds of RLHF. For such a value of β, the same implications hold – the main difference between both settings is in terms of the differences between the reward and policy parameter dimensions.

5. Realizable Rewards: Approximate Optimization

In this section, we shift our focus to the approximate setting, where access to oracles is not given. Here, both paradigms have to approximately solve their estimation problems based on the given data. Similar to the previous section, we assume throughout this section that the ground-truth reward function r^* is linear and realizable in $\mathcal{F}$, and that there exists a loglinear regularized policy $\pi^*_{r^*} \in \Pi$. Moreover, we assume that, for every data tuple $(x, y^w, y^l) \in \mathcal{D}_n$, we have $\phi(x, y^w) \neq \phi(x, y^l)$ and $\psi(x, y^w) \neq \psi(x, y^l)$.

5.1. Theoretical Results

Let us start with the reward learning phase. Recall the definition of the loss $\mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)$ for MLE, as defined in Section 4. Let ω_0 be initialized randomly, and let

$$\omega_{t+1} = \operatorname*{proj}_{\omega:\|\omega\|_2 \leq F} \left(\omega_t - \eta \nabla_\omega \mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)\right) , \qquad (4)$$

for any iterate $t \geq 0$, where η denotes the learning rate. Let $\omega^*_{\mathcal{D}_n} \in \arg\max \mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)$. The first result of this section provides fast convergence rates of gradient descent for the reward learning phase of RLHF.

Theorem 5.1. *For every $t \geq 0$, the gradient descent procedure (4) with learning rate $\eta = 1/\exp(2F)$ satisfies*

$$\left\|\omega_t - \omega^*_{\mathcal{D}_n}\right\|^2_{\Sigma_{\mathcal{D}_n,R}} \leq O\left(1 - \frac{1}{n}\right)^t .$$

Proof sketch. We begin by showing Lipschitzness and smoothness of $\mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)$ with respect to ω. Then, we show that the PL condition (Karimi et al., 2016) is enough to guarantee fast convergence of projected gradient descent by showing that such a condition implies that $\mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)$ also satisfies the *proximal* PL condition (Karimi et al., 2016) when its domain is restricted to a ball. The result follows by applying the convexity of $\mathcal{L}^{\omega}_{\mathrm{RLHF}}(\mathcal{D}_n)$. □

Next, we discuss the policy optimization phase of RLHF. Let $r_{\widehat{\omega}}$ denote the reward estimated from the previous phase. Initialize $\theta_0 \in \mathbb{R}^{d_P}$ with $\|\theta_0\|_2 \leq B$. For any $t \geq 0$, let

$$\theta_{t+1} = \theta_t + \eta' \left(\Psi_n \Psi_n^\top\right)^\dagger \nabla_\theta \mathcal{V}^{\pi_\theta}_{r_{\widehat{\omega}}}(\mathcal{D}_n) , \qquad (5)$$

where $\eta' > 0$ is the learning rate and $\Psi_n = [\psi(x,y)]_{x\in\mathcal{D}_n, y\in\mathcal{Y}}$ denotes the sample feature matrix.[2] Then, the following result holds.

Theorem 5.2. *Let $\delta > 0$. Assume that Ψ_n has full column rank. Then, with probability at least $1-\delta$, for every $t \geq 1$, update rule (5) with learning rate $\eta' \leq n/\beta$ satisfies*

$$\mathcal{V}^*_{r_{\widehat{\omega}}}(\mathcal{D}_n) - \mathcal{V}^{\pi_{\theta_t}}_{r_{\widehat{\omega}}}(\mathcal{D}_n) \leq O\left(\frac{1}{\beta}\exp\left(-(t-1)\right)\right) .$$

Proof sketch. After deriving a naive gradient update rule in matrix notation, we examine the conditions needed to obtain fast convergence rates for our setting. As a consequence, we design our gradient update which resembles natural policy gradient for loglinear policies – the matrix $(\Psi_n\Psi_n^\top)^\dagger$ captures the gradient information for this case. Then, we use similar techniques to those in (Mei et al., 2020) and use the fact that Ψ_n is full rank, to finally obtain the desired bounds. □

Remark 5.1. It is important to emphasize that the assumption on Ψ_n is not restrictive. Indeed, given a preference dataset, it is always possible to construct alternative feature representations that satisfy the assumption with respect to the given data (e.g. different LLMs use different encoding methods while being fine-tuned using the same data).

Remark 5.2. Before going to our next result, it is important to clarify the double usage of $\mathcal{D}_n$ for both the reward learning and policy optimization phases. Our theoretical guarantees are based on the data being independently generated in these phases. Thus, the standard approach is to split the data into two batches for both purposes. Note that both batches would still be $O(n)$ in size and the dependence of the results on n would not change. We use the same $\mathcal{D}_n$ for both phases for simplicity of presentation.

Next, we provide convergence results of gradient descent for DPO with loglinear policies. Let θ_0 be initialized randomly, and let

$$\theta_{t+1} = \operatorname*{proj}_{\theta:\|\theta\|\leq B}\left(\theta_t - \eta''\nabla_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)\right) , \qquad (6)$$

for any iterate $t \geq 0$, where η'' denotes the learning rate. Let $\theta^*_{\mathcal{D}_n} \in \arg\max \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$. Then, we have the following result.

Theorem 5.3. *For every $t \geq 0$, the gradient descent procedure (6) with learning rate $\eta'' = O\left(1/\beta^2\right)$ satisfies*

$$\left\|\theta_t - \theta^*_{\mathcal{D}_n}\right\|^2_{\Sigma_{\mathcal{D}_n,P}} \leq O\left(\frac{1}{\beta}\left(1-\frac{\beta}{n}\right)^t\right) .$$

We have omitted the dependence on the absolute constants that are irrelevant to our discussion – see Appendix F for a detailed expression of the hidden constants.

[2] We only need $\|\theta_0\|_2 \leq B$ for RLHF, since its bounds do not depend on $\|\theta_t\|$, for $t \geq 1$. Thus, we do not need projection.

5.2. Comparative Analysis

The regularized suboptimality gap for RLHF is

$$\mathcal{G}(\pi_{\widehat{\theta}}) \leq \Theta\left(\sqrt{\frac{d_R}{n}}\right) + O\left(\left(1-\frac{1}{n}\right)^t + \frac{\exp(-t)}{\beta}\right)$$

and the regularized suboptimality gap for DPO is

$$\mathcal{G}\left(\pi_{\widetilde{\theta}}\right) \leq \Theta\left(\frac{d_P}{\beta n}\right) + O\left(\frac{1}{\beta}\left(1-\frac{\beta}{n}\right)^t\right) .$$

Both of these paradigms satisfy exponential convergence rates, thus, the main implications of the discussion of Section 4.2 hold in this setting as well if β is to be set as constant. If β is to be tuned for DPO, it cannot be made arbitrarily small or large as observed in Section 4.2 – DPO's overall bounds disproportionately depend on the parameter β. Even so, setting β to its optimal value of $\Theta(1/\sqrt{n})$ for the exact optimization setting would not affect the convergence rate of gradient descent for DPO.

6. Non-realizable Rewards: Exact Optimization

In this section, we consider the case when the ground-truth reward function r^* does not belong to the linear class $\mathcal{F}$ (see Definition 3.1). We again assume that there exists an optimal $\pi^*_{r^*}$ regularized policy that belongs to the loglinear class Π (see Definition 3.2) for some θ^*. We will capture the mismatch between the reward function and its best linear approximation in $\mathcal{F}$ by the following condition.

Assumption 6.1 (*Non-realizability of the ground-truth reward*)**.** There exists $r_{\omega^*} \in \mathcal{F}$ with parameter ω^* such that $\|r^* - r_{\omega^*}\|_\infty \leq \epsilon_{\text{app}}$, where $\epsilon_{\text{app}} > 0$ denotes the mismatch coefficient.

As we will see, the mismatch coefficient will appear linearly in the RLHF bounds on the gap as an additional constant that cannot be improved by increasing the data size.

6.1. Theoretical Results

We begin with the RLHF result, which can be derived from Theorem 4.1.

Theorem 6.1. *Let $\delta > 0$. Suppose that Assumption 6.1 holds. Then, with probability at least $1-\delta$, we have $G\left(\pi_{\widehat{\theta}}\right) \leq D\left(\pi_{\widehat{\theta}}\right) + \widetilde{\Theta}\left(\Lambda_R\sqrt{d_R/n}\right) + 2\epsilon_{\text{app}}$.*

We can directly obtain a similar dependence on ϵ_{app} for DPO. In addition to that, we can also obtain alternative bounds that can be controlled by β as follows.

Theorem 6.2. *Let $\delta > 0$. Suppose that Assumption 6.1 and the condition of Lemma 4.1 hold. Then, with probability at least $1-\delta$, we have $G\left(\pi_{\widetilde{\theta}}\right) \leq D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\Lambda_P d_P/(\beta n)\right) + \min\{2\epsilon_{\text{app}}, O\left(\beta D_{\text{KL}}\left(\pi_{\theta^*}||\pi^*\right)\right)\}$.*

6.2. Comparative Analysis

The key observation to be made in this section is the discrepancy of the bounds in terms of the reward mismatch coefficient. RLHF does not use MLE for the policy parameter estimation, but first learns a reward model. Thus, it cannot bypass the error coming from the reward unrealizability. For DPO, note that, if we set $\beta = O(1/\sqrt{n})$, its bounds improves asymptotically with n, assuming that $D_{\text{KL}}(\pi_{\theta^*}||\pi^*_{r^*})$ is bounded (recall that $\pi^*_{r^*}$ denotes the optimal regularized policy and π_{θ^*} denotes its best loglinear fit). This setting benefits DPO as it is designed to bypass the reward function and directly optimize over the policy space.

7. Realizable Rewards: Exact Optimization – An Extension to Deterministic MDPs

Up to this point, our discussion was concentrated on the contextual bandit setting, which has been used in the DPO literature for the KL-regularized problem (Rafailov et al., 2023). Now, we focus on a generalization of our comparative analysis to Markov decision processes (MDPs), where contexts are related to each other through transition dynamics.

As mentioned previously, the discrepancy between reward and policy dimensions plays a crucial role in the relative performances of RLHF and DPO. While these dimensions could arguably be similar (or have a small gap) for the contextual bandit setting, that is not necessarily the case in general when extending to MDPs, where the reward dimension can be smaller than the policy dimension. For this section, we assume that the ground-truth reward function r^* is linear and realizable in $\mathcal{F}$.

7.1. Preliminaries for Deterministic MDPs

For an MDP, $\mathcal{X}$ is the set of states and $\mathcal{Y}$ the set of actions. In particular, we consider deterministic MDPs, with a transition function $T : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ that provides the next state, given the current state-action.

The value function in infinite-horizon MDPs is given as $V^\pi_r(x) = \mathbb{E}[\sum_{t\geq 0}\gamma^t r(x_t, y_t)|\rho, \pi]$, where ρ denotes the initial state distribution. Given policy π, the occupancy measure of π is given by $d^\pi_\rho(x, y) = (1-\gamma)\sum_{t\geq 0}\gamma^t \mathbb{P}(x_t = x, y_t = y|\rho, \pi)$. We will consider the class of loglinear occupancy measures, as defined next.

Definition 7.1 (*Loglinear occupancy measures class*)**.** Let $\psi'(x, y) \in \mathbb{R}^{d_M}$ denote the feature vector of the pair (x, y) with $\max_{x,y} \|\psi'(x, y)\|_2 \leq 1$, and $B' > 0$. We consider the following class of loglinear occupancy measures:

$$\Pi' = \Big\{ d^\theta_\rho : d^\theta_\rho(x, y) = \frac{\exp(\theta^\top \psi'(x, y))}{\sum_{x',y'} \exp(\theta^\top \psi'(x', y'))}, \\ \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \text{ where } \theta \in \mathbb{R}^{d_M} \text{ and } \|\theta\|_2 \leq B' \Big\}.$$

In this section, we use the $\mathcal{V}^{d_\rho}_r(\rho)$ notation, instead of $\mathcal{V}^\pi_r(\rho)$. Similarly, we use $G(d_\rho)$ to denote the gap in terms of occupancy measure d_ρ, and $D(d_\rho)$ for the difference of the gaps. For a complete discussion, see Appendix H.

In this setting, we are given a dataset $\mathcal{D}_n = \{(x_{0,i}, \tau^w_i, \tau^l_i)\}_{i=1}^n$, where $x_{0,i}$ denotes the initial state of the ith sample and τ^w_i denotes the preferred trajectory $(x_{0,i}, y^w_{0,i}, x^w_{1,i}, \ldots)$ over τ^l_i. Analogous to Section 3, we define the Bradley-Terry preference model for two trajectories τ^w and τ^l as $P^*(\tau^w \succ \tau^l|x_0) = \sigma(R^*(\tau^w) - R^*(\tau^l))$, where $R^*(\tau) = \sum_{t\geq 0}\gamma^t r^*(x_t, y_t)$ is the discounted return, and $\tau^w \succ \tau^l$ denotes τ^w being preferred over τ^l.

7.2. RLHF and DPO for MDPs

Similar to the contextual bandit setting, the objective for the reward learning phase of RLHF in MDPs with linear rewards can be written as follows:

$$\min_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n) := -\mathbb{E}_{(x_0,\tau^w,\tau^l)\sim\mathcal{D}_n} \\ \left[\log\sigma\left(\omega^\top\left(\sum_{t\geq 0}\gamma^t\left(\phi(x^w_t, y^w_t) - \phi(x^l_t, y^l_t)\right)\right)\right)\right]. \tag{P3.1}$$

Once we have the estimated reward function $r_{\hat{\omega}}$, the objective is to solve the KL-regularized problem. Following previous literature on KL-regularized RL (Nachum et al., 2019; Lee et al., 2021), we formulate the objective in this setting as

$$\max_\pi V^\pi_{r_{\hat{\omega}}}(\rho) - \beta D_{\text{KL}}\left(d^\pi_\rho||d^\mu_\rho\right). \tag{P3.2}$$

We will assume throughout this section that we are given access to oracles that exactly solve Problem (P3.1) and (P3.2).

Remark 7.2. Note that the objective in Problem (P3.2) depends on ρ, while the objective in Problem (P1.2) depends on $\mathcal{D}_n$. This is due to considering occupancy measures instead of policies. We keep our current formulation for ease of presentation and leave its extension to a sample objective formulation for future work.

For the purposes of our comparative analysis, we also need an extension of DPO to MDPs, based on the preference model of Section 7.1. The key difficulty of extending DPO to the MDP setting is that the gradient has a non-linear dependence on the policy. To bypass this issue, we leverage the fact that transitions are deterministic to simplify cumulative differences of the optimal Lagrange multipliers for Problem (P3.2), and obtain the following loss function for DPO:

$$\mathcal{L}^{d_\rho}_{\text{DPO}}(\mathcal{D}_n) = -\mathbb{E}_{(x_0,\tau^w,\tau^l)\sim\mathcal{D}_n}\Bigg[\\ \log\sigma\left(\beta\sum_{t=0}^\infty\gamma^t\left(\log\frac{d_\rho(x^w_t, y^w_t)}{d^\mu_\rho(x^w_t, y^w_t)} - \log\frac{d_\rho(x^l_t, y^l_t)}{d^\mu_\rho(x^l_t, y^l_t)}\right)\right)\Bigg].$$

All derivations are in Appendix H. Next, we generalize the bounds from Section 4 for the above formulations.

7.3. Theoretical Results

Analogous to the previous sections, let us define

$$\Sigma'_{\mathcal{D}_n,R} = \frac{1}{n} \sum_{(x_0,\tau^w,\tau^l)\in\mathcal{D}_n} \overline{\phi}'(x_0,\tau^w,\tau^l)\overline{\phi}'(x_0,\tau^w,\tau^l)^\top ,$$

where $\overline{\phi}'(x_0,\tau^w,\tau^l) = \sum_{t\geq 0}\gamma^t(\phi(x_t,y_t^w)-\phi(x_t,y_t^l))$. For $\lambda>0$, define $\Lambda'_R = \left\|(\Sigma'_{\mathcal{D}_n,R}+\lambda I)^{-1/2}\right\|_2$. Similarly, let the sample covariance matrix with respect to occupancy measure features be defined as

$$\Sigma'_{\mathcal{D}_n,M} = \frac{1}{n} \sum_{(x_0,\tau^w,\tau^l)\in\mathcal{D}_n} \overline{\psi}'(x_0,\tau^w,\tau^l)\overline{\psi}'(x_0,\tau^w,\tau^l)^\top ,$$

where $\overline{\psi}'(x_0,\tau^w,\tau^l) = \sum_{t\geq 0}\gamma^t((\psi'(x_t,y_t^w)-\psi'(x_t,y_t^l))$. Let $\Lambda'_M = \left\|(\Sigma'_{\mathcal{D}_n,M}+\lambda I)^{-1/2}\right\|_2$. For RLHF with exact optimization, it is straightforward to extend our previous bounds as follows.

Theorem 7.1. *Let $\delta>0$. Assume that the policy learning phase yields $\widehat{\theta}\in\arg\max_\theta \mathcal{V}^{d^\theta_\rho}_{r_{\widehat{\omega}}}(\rho)$, where $r_{\widehat{\omega}}$ is the estimated reward. Then, with probability at least $1-\delta$, the suboptimality gap incurred by RLHF is $G(d^{\widehat{\theta}}_\rho) = D(d^{\widehat{\theta}}_\rho) + \widetilde{\Theta}(\Lambda'_R\sqrt{d_R/n})$.*

Now that we have a formulation for DPO, we can also extend the previous bounds for the MDP setting. Similar to Lemma 4.1, we now state an analogous result for this setting that guarantees the expression of the ground-truth reward in terms of optimal loglinear occupancy measures. Recall that Φ and Ψ denote the reward and policy feature matrices, respectively. Let $\pi^*_{r^*}\in\arg\max_\pi \mathcal{V}^\pi_{r^*}(\rho)$. Define $\Phi_{\pi^*_{r^*}}$ to be the $d_R\times XY$-dimensional matrix with columns defined as $\gamma\mathbb{E}[\sum_t\gamma^t\phi(x_t,y_t)|x_0=x,\pi^*_{r^*}]-\mathbb{E}[\sum_t\gamma^t\phi(x_t,y_t)|x_0=x,y_0=y,\pi^*_{r^*}]$. We have the following result.

Lemma 7.3. *Assume that $r^*\in\mathcal{F}$, $d^\mu_\rho\in\Pi'$ and $d^{\pi^*_{r^*}}_\rho\in\Pi'$, for some optimal $d^{\pi^*_{r^*}}_\rho$. Furthermore, assume that the column space of $\Phi+\Phi_{\pi^*_{r^*}}$ is contained in the column space of Ψ. Then, for finite MDPs with deterministic transitions, there exists θ^* such that $d^{\theta^*}_\rho\in\arg\max_d \mathcal{V}^d_{r^*}(\rho)$ and $d^{\theta^*}_\rho(x,y)\propto d^\mu_\rho(x,y)\exp(A^{\pi^*_{r^*}}_{r^*}(x,y)/\beta)$, where $A^{\pi^*_{r^*}}_{r^*}(x,y) = Q^{\pi^*_{r^*}}_{r^*}(x,y) - V^{\pi^*_{r^*}}_{r^*}(x)$.*

Now we are ready to state the DPO result for this section.

Theorem 7.2. *Let $\delta>0$. Let $d^{\widetilde{\theta}}_\rho$ denote the occupancy measure returned by DPO and assume that $d^{\pi^*}_\rho\in\Pi'$, for some $d^{\pi^*}_\rho\in\arg\max_{d_\rho}\mathcal{V}^{d_\rho}_{r^*}(\rho)$, and that the condition of Lemma 7.3 is satisfied. Then, for any $n\geq O\left(tr((\Sigma'_{\mathcal{D}_n,M})^\dagger)/(\beta(B')^2)\right)$, with probability at least $1-\delta$, we have $G(d^{\widetilde{\theta}}_\rho) = D(d^{\widetilde{\theta}}_\rho) + \Theta(\Lambda'_M(d_M+1)/(\beta n) + \beta\Lambda_M\lambda(B')^2)$.*

Proof sketch. We start by expressing the optimal discounted reward in terms of an optimal occupancy measure, which is also loginear, by using Lemma 7.3. Then, we equivalently express the value function in terms of occupancy measures. This allows us to cancel out some terms and express the whole gap in terms of the $D_{\mathrm{KL}}(d^{\widetilde{\theta}}_\rho||d^{\theta^*}_\rho)$. Finally, similar to the proof of Theorem 4.2, using loglinearity and properties of the log-sum-exp function, we obtain the desired bounds. □

7.4. Comparative Analysis

The main implication of the above results is that the observations made in Section 4 extend to deterministic MDPs, using our proposed formulation of RLHF and DPO. For the optimal value of β for DPO as discussed in Section 4.2, the RLHF and DPO bounds become directly comparable in terms of the dimension differences for deterministic MDPs. In MDPs with simple reward models (e.g., low-dimensional linear reward models), typically there is still a necessity for high-dimensional policy parameters to represent the value function effectively. This suggests that the complexity of the policy class exceeds that of the reward class and that RLHF outperforms DPO in such instances.

8. Concluding Discussion

In this paper, we provided a comparative analysis between reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO). We performed a thorough analysis under different settings, where we derived sample complexity bounds for both paradigms and drew conclusions on their statistical comparison, based on sample size, regularization temperature, and the dimensionality of their respective parametrizations. We believe these results will initiate a larger discussion on the differences between these two paradigms.

There are many interesting future directions to pursue. The first natural extension of this work is to relax the assumptions made on policy and reward classes and provide a comparative analysis of RLHF and DPO on more realistic settings (eg. general function approximation). Next, a systematic large-scale empirical investigation that would validate the theoretical insights of this paper would be of great importance. Finally, as our current extension of DPO for MDPs is limited to deterministic MDPs using a loglinear occupancy measure regularization, it would be interesting to see whether DPO can be extended to more general formulations.

Impact Statement

This paper focuses on the theoretical aspects of machine learning, providing a comparative analysis of different paradigms of learning from human preferences. We do not foresee any direct negative outcomes from the findings of this paper. On the contrary, we believe that our results might initiate a larger discussion on the statistical properties of learning from human preferences.

Acknowledgements

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

References

Agarwal, R., Schuurmans, D., and Norouzi, M. An Optimistic Perspective on Offline Reinforcement Learning. In *ICML*, 2020.

Ailon, N., Karnin, Z. S., and Joachims, T. Reducing Dueling Bandits to Cardinal Bandits. In *ICML*, 2014.

An, G., Lee, J., Zuo, X., Kosaka, N., Kim, K.-M., and Song, H. O. Direct Preference-based Policy Optimization without Reward Modeling. In *NeurIPS*, 2023.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A General Theoretical Paradigm to Understand Learning from Human Preferences. *CoRR*, abs/2310.12036, 2023.

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

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

Brown, D. S., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In *ICML*, 2019.

Chatterji, N., Pacchiano, A., Bartlett, P., and Jordan, M. On the Theory of Reinforcement Learning with Once-per-episode Feedback. In *NeurIPS*, 2021.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation. In *ICML*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep Reinforcement Learning from Human Preferences. In *NeurIPS*, 2017.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved Optimistic Algorithms for Logistic Bandits. In *ICML*, 2020.

Fujimoto, S., Meger, D., and Precup, D. Off-policy Deep Reinforcement Learning without Exploration. In *ICML*, 2019.

Gajane, P., Urvoy, T., and Clérot, F. A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits. In *ICML*, 2015.

Ganguli, D. et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *CoRR*, abs/2209.07858, 2022.

Gao, L., Schulman, J., and Hilton, J. Scaling Laws for Reward Model Overoptimization. In *ICML*, 2023.

Glaese, A. et al. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *CoRR*, abs/2209.14375, 2022.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., and Sadigh, D. Contrastive Prefence Learning: Learning from Human Feedback without RL. *CoRR*, abs/2310.13639, 2023.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way Off-policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *CoRR*, abs/1907.00456, 2019.

Jin, Y., Yang, Z., and Wang, Z. Is Pessimism Provably Efficient for Offline RL? In *ICML*, 2021.

Karimi, H., Nutini, J., and Schmidt, M. Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak-Łojasiewicz Condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, 2016.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based Offline Rinforcement Learning. In *NeurIPS*, 2020.

Knox, W. B. and Stone, P. Tamer: Training an Agent Manually via Evaluative Reinforcement. In *ICDL*, 2008.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem. In *COLT*, 2015.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for Offline Reinforcement Learning. In *NeurIPS*, 2020.

Laroche, R., Trichelair, P., and Des Combes, R. T. Safe Policy Improvement with Baseline Bootstrapping. In *ICML*, 2019.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. Optidice: Offline Policy Optimization via Stationary Distribution Correction Estimation. In *ICML*, 2021.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E., and Littman, M. L. Interactive Learning from Policy-dependent Human Feedback. In *ICML*, 2017.

Mei, J., Xiao, C., Szepesvári, C., and Schuurmans, D. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *ICML*, 2020.

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

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy Gradient from Arbitrary Experience. *CoRR*, abs/1912.02074, 2019.

Nakano, R. et al. Webgpt: Browser-assisted Question-answering with Human Feedback. *CoRR*, abs/2112.09332, 2021.

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

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*, 2023.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. Is Reinforcement Learning (not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. In *ICLR*, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. In *NeurIPS*, 2021.

Saha, A. and Gopalan, A. Active Ranking with Subset-wise Preferences. In *AISTATS*, 2019.

Saha, A. and Krishnamurthy, A. Efficient and Optimal Algorithms for Contextual Dueling Bandits under Realizability. In *ALT*, 2022.

Saha, A., Pacchiano, A., and Lee, J. Dueling RL: Reinforcement Learning with Trajectory Preferences. In *AISTATS*, 2023.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramch, K., Wainwright, M. J., et al. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. In *AISTATS*, 2016.

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

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to Summarize with Human Feedback. In *NeurIPS*, 2020.

Uehara, M. and Sun, W. Pessimistic Model-based Offline Reinforcement Learning Under Partial Coverage. *CoRR*, abs/2107.06226, 2021.

Wang, C. et al. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. *CoRR*, abs/2309.16240, 2023.

Warnell, G., Waytowich, N., Lawhern, V., and Stone, P. Deep TAMER: Interactive Agent Shaping in High-dimensional State Spaces. In *AAAI*, 2018.

Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., and Christiano, P. Recursively Summarizing Books with Human Feedback. *CoRR*, abs/2109.10862, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent Pessimism for Offline Reinforcement Learning. In *NeurIPS*, 2021.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-constraint. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The K-armed Dueling Bandits Problem. In *COLT*, 2009.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable Benefits of Actor-critic Methods for Offline Reinforcement Learning. In *NeurIPS*, 2021.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable Offline Reinforcement Learning with Human Feedback. *CoRR*, abs/:2305.14816, 2023.

Zhu, B., Jordan, M. I., and Jiao, J. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *ICML*, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593, 2019.

Zoghi, M., Whiteson, S. A., De Rijke, M., and Munos, R. Relative Confidence Sampling for Efficient On-line Ranker Evaluation. In *WSDM*, 2014.

Appendix

Table of Contents

A. Statistical Bounds for RLHF (Section 4 and 7)

In this section, we prove the main RLHF result, Theorem 4.1. We state the detailed version of it together with the necessary constants.

Theorem A.1. *Let $\delta > 0$. Assume that the preference data satisfies the BT model, and $r^* \in \mathcal{F}$. Denote by $\widehat{\omega}$ and $\widehat{\theta}$ the reward and policy parameters learned via RLHF, respectively. Furthermore, assume that*

$$\widehat{\omega} \in \arg\min_{\omega} \mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n)$$

and

$$\widehat{\theta} \in \arg\max_{\theta} \mathcal{V}^{\pi_\theta}_{r_{\widehat{\omega}}}(\mathcal{D}_n)\,.$$

Then, with probability at least $1-\delta$, for any $\lambda > 0$, the suboptimality gap incurred by RLHF is

$$G\left(\pi_{\widehat{\theta}}\right) = \Theta\left(\left\|\left(\Sigma_{\mathcal{D}_n,R} + \lambda I\right)^{-1/2}\right\|_2 \cdot \sqrt{\frac{d_R + \log(6/\delta)}{S_R^2 n} + \lambda F^2}\right) + D\left(\pi_{\widehat{\theta}}\right),$$

where $S_R = 1/\left(2 + \exp\left(-2F\right) + \exp\left(2F\right)\right)$.

Proof. Let $\Phi \in \mathbb{R}^{d_R \times XY}$ be the reward feature matrix. Then, for any $\lambda > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&G(\pi_{\widehat{\theta}}) = V^{\text{opt}}_{r^*}(\rho) - V^{\pi_{\widehat{\theta}}}_{r^*}(\rho) \\
&= D(\pi_{\widehat{\theta}}) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*}(\rho)\right) \\
&= D(\pi_{\widehat{\theta}}) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*}(\rho) - \mathcal{V}^{\pi^*_{r^*}}_{r^*_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*}(\rho)\right) \\
&\overset{(a)}{\leq} D(\pi_{\widehat{\theta}}) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*}(\rho) - \mathcal{V}^{\pi^*_{r^*}}_{r^*_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi^*_{r^*_{\widehat{\omega}}}}_{r^*_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*}(\rho)\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} D(\pi_{\widehat{\theta}}) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*}(\rho) - \mathcal{V}^{\pi^*_{r^*}}_{r_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi^*_{r_{\widehat{\omega}}}}_{r_{\widehat{\omega}}}(\mathcal{D}_n) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(\mathcal{D}_n)\right) + O\left(\sqrt{\frac{\log(6/\delta)}{n}}\right) + \left(\mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*}(\rho)\right) \\
&\overset{(c)}{\leq} D(\pi_{\widehat{\theta}}) + \left(\mathcal{V}^{\pi^*_{r^*}}_{r^*}(\rho) - \mathcal{V}^{\pi^*_{r^*}}_{r_{\widehat{\omega}}}(\rho)\right) + \left(\mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(\mathcal{D}_n) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(\mathcal{D}_n)\right) + O\left(\sqrt{\frac{\log(6/\delta)}{n}}\right) + \left(\mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(\rho) - \mathcal{V}^{\pi_{\widehat{\theta}}}_{r^*}(\rho)\right) \\
&= D(\pi_{\widehat{\theta}}) + \sum_{x,y} \rho(x) \cdot (\pi^*_{r^*}(y|x) - \pi_{\widehat{\theta}}(y|x)) \cdot (r^*(x,y) - r_{\widehat{\omega}}(x,y)) + O\left(\sqrt{\frac{\log(6/\delta)}{n}}\right) \\
&\overset{(d)}{=} D(\pi_{\widehat{\theta}}) + \left(d^*_\rho - d^{\pi_{\widehat{\theta}}}_\rho\right)^\top (\omega^* - \widehat{\omega})\Phi + O\left(\sqrt{\frac{\log(6/\delta)}{n}}\right) \\
&\overset{(e)}{\leq} D(\pi_{\widehat{\theta}}) + \left\|\Phi\left(d^*_\rho - d^{\pi_{\widehat{\theta}}}_\rho\right)\right\|_{(\Sigma_{\mathcal{D}_n,R} + \lambda I)^{-1}} \|\omega^* - \widehat{\omega}\|_{\Sigma_{\mathcal{D}_n,R} + \lambda I} + O\left(\sqrt{\frac{\log(6/\delta)}{n}}\right) \\
&\overset{(f)}{\leq} D(\pi_{\widehat{\theta}}) + O\left(\left\|(\Sigma_{\mathcal{D}_n,R} + \lambda I)^{-1/2}\right\|_2 \cdot \sqrt{\frac{d_R + \log(6/\delta)}{S_R^2 n} + \lambda F^2}\right) ,
\end{aligned}
$$

where (a) is due to the fact that $\pi^*_{r_{\widehat{\omega}}} \in \arg\max_\pi \mathcal{V}^\pi_{r_{\widehat{\omega}}}(\rho)$; (b) follows from Lemma A.2 and the union bound – if we have $\mathbb{P}(\mathcal{E}_i^c) \leq \delta_i$, for $i = 1, 2, 3$, where $\mathcal{E}^c$ denotes the complement of event $\mathcal{E}$, letting $\delta_i = \delta/3$, for all i, we have

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3\right) &= 1 - \mathbb{P}\left(\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c\right) \\
&\geq 1 - \left(\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_3)\right) \\
&\geq 1 - (\delta_1 + \delta_2 + \delta_3) \\
&= 1 - \delta \, .
\end{aligned}
$$

Next, (c) is due to the fact that $\pi_{\widehat{\theta}} \in \arg\max_{\pi \in \Pi} \mathcal{V}^\pi_{r_{\widehat{\omega}}}(\mathcal{D}_n)$ and $\pi^*_{r_{\widehat{\omega}}} \in \Pi$ (as per Lemma 4.1); (d) is due to $d^*_\rho(x,y) = \rho(x,y)\pi^*_{r^*}(y|x)$ and $d^{\pi_{\widehat{\theta}}}_\rho(x,y) = \rho(x)\pi_{\widehat{\theta}}(y|x)$; (e) is an application of the Cauchy-Schwarz inequality with respect to the semi-norm induced by matrix $\Sigma_{\mathcal{D}_n,R} + \lambda I$; and (f) is a direct application of Lemma 3.1 of (Zhu et al., 2023) for the discounted infinite-horizon setting.

The lower bound is an immediate application of Theorem 3.10 of (Zhu et al., 2023). Note that we are under the same conditions; our reward function is assumed to be linear, and we also assume a bounded covering number. For the lower bound construction, let $\mu = \pi^{\text{opt}}_{r^*}$, i.e., the reference policy is the actually optimal one. Let $\text{CB}(\Lambda)$ denote the set of bandit instances coupled with datasets with a covering number no more than Λ. Let $\mathcal{Q}$ denote such an instance. Under these assumptions, Theorem 3.10 of (Zhu et al., 2023) implies an information-theoretic lower bound of

$$
\inf_\pi \sup_{\mathcal{Q} \in \text{CB}(\Lambda_R)} \left(V^{\text{opt}}_{\mathcal{Q}}(\rho) - V^\pi_{\mathcal{Q}}(\rho)\right) \geq O\left(\Lambda_R \sqrt{\frac{d_R}{n}}\right) .
$$

□

Theorem 7.1. *Let $\delta > 0$. Assume that the policy learning phase yields $\widehat{\theta} \in \arg\max_\theta \mathcal{V}^{d^\theta_\rho}_{r_{\widehat{\omega}}}(\rho)$, where $r_{\widehat{\omega}}$ is the estimated reward. Then, with probability at least $1-\delta$, the suboptimality gap incurred by RLHF is $G(d^{\widehat{\theta}}_\rho) = D(d^{\widehat{\theta}}_\rho) + \widetilde{\Theta}(\Lambda'_R \sqrt{d_R/n})$.*

Proof. The proof of this result is an immediate application of the previous result with instead an application of Lemma 5.1 of (Zhu et al., 2023) with

$$
S_R = 1/\left(2 + \exp\left(-2F(1-\gamma)\right) + \exp\left(2F(1-\gamma)\right)\right) .
$$

□

Next, we connect the difference between the sample regularized gap and the expected regularized gap with respect to the context distribution, and obtain the following result.

Lemma A.2. *Let $\delta > 0$ and assume that the conditions of Theorem A.1 are satisfied. Then, we have that*

$$\left| \mathbb{E}_{x\sim\rho}\left[\mathcal{V}^{\pi^*_{r_{\widehat{\omega}}}}_{r_{\widehat{\omega}}}(x)\right] - \frac{1}{n}\sum_{x\in\mathcal{D}_n} \mathcal{V}^{\pi^*_{r_{\widehat{\omega}}}}_{r_{\widehat{\omega}}}(x) \right| \leq \sqrt{\frac{\log(4/\delta)}{n}} \,,$$

and

$$\left| \frac{1}{n}\sum_{x\in\mathcal{D}_n} \mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(x) - \mathbb{E}_{x\sim\rho}\left[\mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(x)\right] \right| \leq \sqrt{\frac{(1+\beta(2B+\log Y))\log(4/\delta)}{n}} \,.$$

with probability at least $1-\delta$.

Proof. Using the reward-to-policy mapping of Equation (1), we have that, for every (x, y),

$$\mu(y|x) = \pi^*_{r_{\widehat{\omega}}}(y|x)\exp\left(-\frac{1}{\beta}r_{\widehat{\omega}}(x,y)\right) \,.$$

Thus, note that, for every $x \in \mathcal{X}$,

$$\begin{aligned}
\left|\mathcal{V}^{\pi^*_{r_{\widehat{\omega}}}}_{r_{\widehat{\omega}}}(x)\right| &= \left|\sum_y \pi^*_{r_{\widehat{\omega}}}(y|x) r_{\widehat{\omega}}(x,y) - \beta\pi^*_{r_{\widehat{\omega}}}(y|x)\log\frac{\pi^*_{r_{\widehat{\omega}}}(y|x)}{\mu(y|x)}\right| \\
&\leq 1 + \beta\left|\sum_y \pi^*_{r_{\widehat{\omega}}}(y|x)\log\frac{\pi^*_{r_{\widehat{\omega}}}(y|x)}{\mu(y|x)}\right| \\
&= 1 + \beta\left|\sum_y \pi^*_{r_{\widehat{\omega}}}(y|x)\frac{1}{\beta}r_{\widehat{\omega}}(x,y)\right| \\
&\leq 2 \,,
\end{aligned}$$

where we have used that the reward lies in $[0, 1]$. On the other hand, we have

$$\begin{aligned}
\left|\mathcal{V}^{\pi_{\widehat{\theta}}}_{r_{\widehat{\omega}}}(x)\right| &= \left|\sum_y \pi_{\widehat{\theta}}(y|x) r_{\widehat{\omega}}(x,y) - \beta\pi_{\widehat{\theta}}(y|x)\log\frac{\pi_{\widehat{\theta}}(y|x)}{\mu(y|x)}\right| \\
&\leq 1 + \beta\left|\sum_y \pi_{\widehat{\theta}}(y|x)\left(\log\frac{\pi_{\widehat{\theta}}(y|x)}{\pi^*_{r_{\widehat{\omega}}}(y|x)} + \frac{1}{\beta}r_{\widehat{\omega}}(x,y)\right)\right| \\
&\leq 2 + \beta\max_y\left|\log\frac{\pi_{\widehat{\theta}}(y|x)}{\pi_{\theta^*}(y|x)}\right| \\
&\leq 2 + \beta\max_{x,y}\left(\left|\log\frac{\exp(\psi(x,y)^\top\widehat{\theta})}{\sum_{y'}\exp(\psi(x,y')^\top\widehat{\theta})}\right| + \left|\log\frac{\exp(\psi(x,y)^\top\theta^*)}{\sum y'\exp(\psi(x,y')^\top\theta^*)}\right|\right) \\
&\leq 2 + \beta\max_{x,y}\left(\left|\log\exp(\psi(x,y)^\top\widehat{\theta})\right| + \left|\log\sum_{y'}\exp(\psi(x,y')^\top\widehat{\theta})\right|\right. \\
&\qquad \left. + \left|\log\exp(\psi(x,y)^\top\theta^*)\right| + \left|\log\sum_{y'}\exp(\psi(x,y')^\top\theta^*)\right|\right) \\
&\leq 2 + \beta\left(2B + 2\log\left(Y\exp(B)\right)\right) \\
&\leq 2 + 2\beta(2B + \log Y) \,,
\end{aligned}$$

where we have used Lemma J.1 and the fact that

$$-B \leq \langle\psi(x,y),\theta\rangle \leq B \,.$$

The result then follows from Hoeffding's inequality. □

B. Statistical Bounds for DPO (Section 4)

In this section, we prove the main DPO result for Section 4.

Theorem 4.2. *Let $\delta > 0$ and $\beta > 0$. Assume that $r^* \in \mathcal{F}$, $\mu \in \Pi$, and that the condition of Lemma 4.1 is satisfied. Let $n \geq O\left(tr(\Sigma^{\dagger}_{\mathcal{D}_n,P})/(\beta B^2)\right)$. Then, with probability at least $1-\delta$, the suboptimality gap of DPO is*

$$G\left(\pi_{\widetilde{\theta}}\right) = D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda_P(d_P+1)}{\beta n} + \beta\lambda\Lambda_P B^2\right) .$$

Proof. Lemma 4.1 implies that there exists $\theta^* \in \mathbb{R}^d$, such that, for every (x,y),

$$r^*(x,y) = \beta \log \frac{\pi_{\theta^*}(y|x))}{\mu(y|x)} + \beta Z(x)$$

and $\mathcal{V}^*_{r^*}(\rho) = \mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho)$. Now, observe that

$$\begin{aligned}
G\left(\pi_{\widetilde{\theta}}\right) &= V^{\text{opt}}_{r^*}(\rho) - V^{\pi_{\widetilde{\theta}}}_{r^*}(\rho) \\
&= D(\pi_{\widetilde{\theta}}) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_{\widetilde{\theta}}}_{r^*}(\rho)\right) \\
&= D(\pi_{\widetilde{\theta}}) + \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\theta^*}(\cdot|x)}}\left[r^*(x,y) - \beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)}\right] - \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\widetilde{\theta}}(\cdot|x)}}\left[r^*(x,y) - \beta\log\frac{\pi_{\widetilde{\theta}}(y|x)}{\mu(y|x)}\right] \\
&= D(\pi_{\widetilde{\theta}}) + \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\theta^*}(\cdot|x)}}\left[\beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)} + \beta\log Z(x) - \beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)}\right] \\
&\qquad - \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\widetilde{\theta}}(\cdot|x)}}\left[\beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)} + \beta\log Z(x) - \beta\log\frac{\pi_{\widetilde{\theta}}(y|x))}{\mu(y|x)}\right] \\
&= D(\pi_{\widetilde{\theta}}) + \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\widetilde{\theta}}(\cdot|x)}}\left[\beta\log\pi_{\widetilde{\theta}}(y|x) - \beta\log\pi_{\theta^*}(y|x)\right] \\
&= D(\pi_{\widetilde{\theta}}) + \mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\widetilde{\theta}}}}\left[\beta\left\langle\psi(x,y), \widetilde{\theta} - \theta^*\right\rangle + \beta\log\frac{\sum_{y'}\exp(\psi(x,y')^\top\theta^*)}{\sum_{y'}\exp(\psi(x,y')^\top\widetilde{\theta})}\right] \\
&= D(\pi_{\widetilde{\theta}}) + \beta\mathbb{E}_{\substack{x\sim\rho \\ y\sim\pi_{\widetilde{\theta}}}}\left[\left\langle\psi(x,y), \widetilde{\theta} - \theta^*\right\rangle\right] + \beta(A(\theta^*) - A(\widetilde{\theta})) ,
\end{aligned}$$

where we have denoted by $A(\theta)$ the log-sum-exp function

$$A(\theta) = \sum_x \rho(x)\log\sum_{y'}\exp\left(\psi(x,y')^\top\theta\right) ,$$

At this point, some properties of the log-exp-sum function will be useful. The proof of the following result can be found in Appendix J.

Lemma B.1. *The function $A(\theta)$ is 1-Lipschitz and 2-smooth. Moreover, if the features are sampled from a 0-mean distribution and span R^{d_P}, then there exists $\kappa > 0$, such that $A(\theta)$ is κ-strongly convex.*

Since $A(\theta)$ is 2-smooth, we have

$$\begin{aligned}
A(\theta^*) - A(\widetilde{\theta}) &\leq \left\langle\nabla_\theta A(\widetilde{\theta}), \theta^* - \widetilde{\theta}\right\rangle + \left\|\theta^* - \widetilde{\theta}\right\|_2^2 \\
&= \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\left\langle\psi(x,y), \theta^* - \widetilde{\theta}\right\rangle\right] + \left\|\theta^* - \widetilde{\theta}\right\|_2^2 .
\end{aligned}$$

Substituting to the suboptimality gap equalities, we obtain

$$\begin{aligned}
G\left(\pi_{\widetilde{\theta}}\right) &\leq D(\pi_{\widetilde{\theta}}) + \left\|\theta^* - \widetilde{\theta}\right\|_2^2 \\
&\overset{(a)}{\leq} D(\pi_{\widetilde{\theta}}) + \beta\left\|(\Sigma_{\mathcal{D}_n,P} + \lambda I)^{-1}\right\|_2\left\|\theta^* - \widetilde{\theta}\right\|^2_{\Sigma_{\mathcal{D}_n,P}+\lambda I}
\end{aligned}$$

$$\overset{(b)}{\leq} D(\pi_{\widetilde{\theta}}) + \beta \left\|(\Sigma_{\mathcal{D}_n,P} + \lambda I)^{-1}\right\|_2 \left\|\theta^* - \widetilde{\theta}\right\|^2_{\Sigma_{\mathcal{D}_n,P}} + 4\beta\lambda\Lambda_P B^2$$
$$\overset{(c)}{\leq} D(\pi_{\widetilde{\theta}}) + O\left(\frac{\Lambda_P(d_P+1)}{\beta n}\right) + 4\beta\lambda\Lambda_P B^2 \,,$$

where for (a) we have used that $\langle x, Ax\rangle \leq \|A\|_2 \|x\|_2^2$; (b) is due to the fact that $\|\theta\|_2 \leq B$; (c) follows from Theorem J.7 and the assumption that $\widetilde{\theta} \in \arg\min_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$. The value of λ can be tuned accordingly. Note that, for fixed β, letting $\lambda = \Theta(1/n)$ yields the desired bound. If $\beta = \Theta(1/\sqrt{n})$, then any small value of λ works.

On the other hand, consider the following feature construction. Let Ψ be full rank with zero-mean columns. Then, there exists $\kappa > 0$ such that $A(\theta)$ is κ-strongly convex. This, in turn, implies that

$$\begin{aligned} A(\theta^*) - A(\widetilde{\theta}) &\geq \left\langle \nabla_\theta A(\widetilde{\theta}), \theta^* - \widetilde{\theta}\right\rangle + \frac{\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_2^2 \\ &= -\mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\left\langle \psi(x,y), \widetilde{\theta} - \theta^*\right\rangle\right] + \frac{\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_2^2 . \end{aligned}$$

Thus, substituting in the original gap expression, we obtain

$$\begin{aligned} G\left(\pi_{\widetilde{\theta}}\right) &\geq D(\pi_{\widetilde{\theta}}) + \frac{\beta\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_2^2 \\ &\geq D(\pi_{\widetilde{\theta}}) + \frac{\beta\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_2^2 \|\Sigma_{\mathcal{D}_n,P}\|_2 \\ &\geq D(\pi_{\widetilde{\theta}}) + \frac{\beta\kappa}{2}\left\langle \widetilde{\theta} - \theta^*, \Sigma_{\mathcal{D}_n,P}(\widetilde{\theta} - \theta^*)\right\rangle \\ &\geq D(\pi_{\widetilde{\theta}}) + \frac{\beta\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_{\Sigma_{\mathcal{D}_n,P}} \\ &\geq D(\pi_{\widetilde{\theta}}) + \frac{\beta\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|_{\Sigma_{\mathcal{D}_n,P}} \\ &\geq D(\pi_{\widetilde{\theta}}) + \Omega\left(\frac{(d_P+1)}{\beta n}\right) , \end{aligned}$$

for any $n \geq O\left(tr(\Sigma^\dagger_{\mathcal{D}_n,P})/(\beta B^2)\right)$, where the second inequality uses the fact that $\|\Sigma_{\mathcal{D}_n,P}\|_2 \leq 1$; the third inequality follows from Cauchy-Schwarz; the last inequality follows by Theorem J.7. □

C. Statistical Bounds for DPO for MDPs (Section 7)

In this section, we will prove Corollary 7.2, the statistical convergence rate of the DPO method in the MDP setting. We restate the result with all the quantities appearing in the bounds.

Theorem C.1. *Assume that $r^* \in \mathcal{F}$, the data $\mathcal{D}$ satisfies the BT preference model for trajectories, and that the feature matrix is full rank. Furthermore, assume that $d^*_\rho \in \Pi'$ and assume that we have 0 optimization error from gradient descent on the MLE loss. Let $\beta > 0$, and*

$$\begin{aligned} S_M &= (\exp(-B') + \exp(B') + 2)^{-1} , \\ U' &= \exp(-2B') + \exp(2B') + 2 , \\ \Lambda_M &= \left\|(\Sigma'_{\mathcal{D}_n,M} + \lambda I)^{-1/2}\right\|_2 . \end{aligned}$$

Then, DPO incurs the following minimax bounds on the suboptimality gap:

$$G\left(d^{\widetilde{\theta}}_\rho\right) = D\left(d^{\widetilde{\theta}}_\rho\right) + \Theta\left(\frac{\Lambda_M U'(d_M+1)}{\beta S_P n}\right) .$$

Proof. First, note that, under some assumptions on the feature space, Lemma J.2 implies that there exists θ^* for which we

can have $\mathcal{V}^*_{r^*}(\rho) = \mathcal{V}^{d^{\theta^*}_\rho}_{r^*}(\rho)$, with respect to some policy $\pi^*_{r^*}$ and that

$$\sum_{t\geq 0}\gamma^t r^*(x_t,y_t) = \sum_{t\geq 0}\gamma^t \beta \log \frac{d^{\theta^*}_\rho(x_t,y_t)}{d^\mu_\rho(x_t,y_t)} + \beta + \alpha^*(x_0)\,, \tag{7}$$

for any trajectory τ, where α^* is the optimal dual variable of Problem (P3.2'), and we have used Equation (15) to express the ground-truth reward in terms of an optimal regularized policy. Now, let us denote by $\widetilde{\pi}$ the policy corresponding to $d^{\widetilde{\theta}}_\rho$. Observe that

$$\begin{aligned}
G\left(d^{\widetilde{\theta}}_\rho\right) &= V^{\text{opt}}_{r^*}(\rho) - V^{d^{\widetilde{\theta}}_\rho}_{r^*}(\rho)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \left(\mathcal{V}^{d^{\theta^*}_\rho}_{r^*}(\rho) - \mathcal{V}^{d^{\widetilde{\theta}}_\rho}_{r^*}(\rho)\right)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \mathbb{E}_{x_0\sim\rho, y_t\sim\pi^*_{r^*}(\cdot|x_t)}\left[\sum_{t\geq 0}\gamma^t r^*(x_t,y_t)\right] - \beta\mathbb{E}_{(x,y)\sim d^{\theta^*}_\rho}\left[\log\frac{d^{\theta^*}_\rho(x,y)}{d^\mu_\rho(x,y)}\right] - \mathcal{V}^{d^{\widetilde{\theta}}_\rho}_{r^*}(\rho)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \mathbb{E}_{x_0\sim\rho, y_t\sim\pi^*_{r^*}(\cdot|x_t)}\left[\sum_{t\geq 0}\gamma^t \beta\log\frac{d^{\theta^*}_\rho(x_t,y_t)}{d^\mu_\rho(x_t,y_t)} + \beta + \alpha^*(x_0)\right] - \beta\mathbb{E}_{(x,y)\sim d^{\theta^*}_\rho}\left[\log\frac{d^{\theta^*}_\rho(x,y)}{d^\mu_\rho(x,y)}\right] - \mathcal{V}^{d^{\widetilde{\theta}}_\rho}_{r^*}(\rho) \qquad (8)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \mathbb{E}_{(x,y)\sim d^{\theta^*}_\mu}\left[\beta\log\frac{d^{\theta^*}_\rho(x,y)}{d^\mu_\rho(x,y)}\right] + \frac{\beta}{1-\gamma} + \sum_x \rho(x)\alpha^*(x) - \beta\mathbb{E}_{(x,y)\sim d^{\theta^*}_\rho}\left[\log\frac{d^{\theta^*}_\rho(x,y)}{d^\mu_\rho(x,y)}\right] - \mathcal{V}^{d^{\widetilde{\theta}}_\rho}_{r^*}(\rho) \qquad (9)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \frac{\beta}{1-\gamma}\\
&\quad + \sum_x \rho(x)\alpha^*(x) - \mathbb{E}_{x_0\sim\rho, y_t\sim\widetilde{\pi}(\cdot|x_t)}\left[\sum_{t\geq 0}\gamma^t\beta\log\frac{d^{\theta^*}_\rho(x_t,y_t)}{d^\mu_\rho(x_t,y_t)} + \beta + \alpha^*(x_0)\right] + \beta\mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\log\frac{d^{\widetilde{\theta}}_\rho(x,y)}{d^\mu_\rho(x,y)}\right]\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \beta\mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\log\frac{d^{\widetilde{\theta}}_\rho(x,y)}{d^\mu_\rho(x,y)} - \log\frac{d^{\theta^*}_\rho(x,y)}{d^\mu_\rho(x,y)}\right]\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\beta\log d^{\widetilde{\theta}}_\rho(x,y) - \beta\log d^{\theta^*}_\rho(x,y)\right]\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\beta\left\langle\psi'(x,y),\widetilde{\theta}-\theta^*\right\rangle + \beta\log\frac{\sum_{x',y'}\exp(\psi'(x',y')^\top\theta^*)}{\sum_{x',y'}\exp(\psi'(x',y')^\top\widetilde{\theta})}\right] \qquad (10)\\
&= D\left(d^{\widetilde{\theta}}_\rho\right) + \beta\mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\left\langle\psi'(x,y),\widetilde{\theta}-\theta^*\right\rangle + A(\theta^*) - A(\widetilde{\theta})\right]\,,
\end{aligned}$$

where we have denoted by

$$A(\theta) = \log\sum_{x',y'}\exp(\psi'(x',y')^\top\theta)$$

the log-sum-exp function. Above, in Equation (8) we have applied Equation (7); Equation (9) uses the fact that, for any policy π and function $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, we have

$$\mathbb{E}_{x\sim\rho, y_t\sim\pi(\cdot|x_t)}\left[\sum_{t\geq 0}\gamma^t f(x_t,y_t)\right] = \mathbb{E}_{(x,y)\sim d^\pi}\left[f(x,y)\right]\,.$$

Finally, for Equation (10) we have used loglinearity. Now, given $\theta\in\mathbb{R}^{d_P}$, note that

$$\nabla_\theta A(\theta) = \frac{\sum_{x,y}\exp(\psi'(x,y)^\top\theta)\cdot\psi'(x,y)}{\sum_{x',y'}\exp(\psi'(x',y')^\top\theta)} = \sum_{x,y} d^\theta_\rho(x,y)\psi'(x,y)\,.$$

On the other hand, the Hessian of $A(\theta)$ is

$$\begin{aligned}\nabla^2_\theta A(\theta) &= \sum_{x,y} \nabla_\theta d^\theta_\rho(x,y)\psi'(x,y) \\ &= \sum_{x,y} d^\theta_\rho(x,y)\left(\psi'(x,y) - \mathbb{E}_{(x',y')\sim d^\theta_\rho}[\psi'(x',y')]\right)\psi'(x,y)^\top \\ &= \mathbb{E}_{(x,y)\sim d^\theta_\rho}\left[\psi'(x,y)\psi'(x,y)^\top\right] - \mathbb{E}_{(x,y)\sim d^\theta_\rho}[\psi'(x,y)]\mathbb{E}_{(x,y)\sim d_\theta}[\psi'(x,y)]^\top \\ &= \mathbb{E}_{(x,y)\sim d^\theta_\rho}\left[\left(\psi'(x,y) - \mathbb{E}_\theta\left[\psi'(x,y)\right]\right)\left(\psi'(x,y) - \mathbb{E}_\theta\left[\psi'(x,y)\right]\right)^\top\right] .\end{aligned}$$

By assumption on the feature mapping, we have that

$$\begin{aligned}\left\|\nabla^2_\theta A(\theta)\right\|_2 &\le \max_{x,y}\left\|\left(\psi'(x,y) - \mathbb{E}_\theta\left[\psi'(x,y)\right]\right)\left(\psi'(x,y) - \mathbb{E}_\theta\left[\psi'(x,y)\right]\right)^\top\right\|_2 \\ &\le \max_{x,y}\left\|\psi'(x,y) - \mathbb{E}_\theta[\psi'(x,y)]\right\|_2 \\ &\le 2\max_{x,y}\left\|\psi'(x,y)\right\|_2 = 2 .\end{aligned}$$

Therefore, the function $A(\theta)$ is 2-smooth in θ, which implies that

$$\begin{aligned}A(\theta^*) - A(\widetilde{\theta}) &\le \left\langle \nabla_\theta A(\widetilde{\theta}), \theta^* - \widetilde{\theta}\right\rangle + \left\|\theta^* - \widetilde{\theta}\right\|^2_2 \\ &= \mathbb{E}_{(x,y)\sim d^{\widetilde{\theta}}_\rho}\left[\left\langle \psi'(x,y), \theta^* - \widetilde{\theta}\right\rangle\right] + \left\|\theta^* - \widetilde{\theta}\right\|^2_2 .\end{aligned}$$

Substituting to the suboptimality gap equalities, we obtain

$$\begin{aligned}G\left(d^{\widetilde{\theta}}_\rho\right) &\le D\left(d^{\widetilde{\theta}}_\rho\right) + \left\|\theta^* - \widetilde{\theta}\right\|^2_2 \\ &\le D\left(d^{\widetilde{\theta}}_\rho\right) + \beta\left\|(\Sigma'_{\mathcal{D}_n,M} + \lambda I)^{-1}\right\|_2\left\|\theta^* - \widetilde{\theta}\right\|^2_{\Sigma'_{\mathcal{D}_n,M}+\lambda I} \\ &\le D\left(d^{\widetilde{\theta}}_\rho\right) + \frac{\Lambda_M U' d_M}{S_M \beta n} + 4\beta\Lambda_M\lambda(B')^2 ,\end{aligned}$$

where the last inequality follows from Theorem J.7 and the assumption on exact optimization.

For the lower bound, let ψ' be sampled from a 0-mean bounded distribution. Note that, for any non-zero vector in $\mathbb{R}^{d_M}$, we have

$$\begin{aligned}z^\top\nabla^2_\theta A(\theta) z &= \mathbb{E}_{(x,y)\sim d^\theta_\rho}\left[z^\top\psi'(x,y)\psi'(x,y)^\top z\right] \\ &\ge \min_{\theta,x,y} d^\theta_\rho(x,y)\sum_{x,y}(\psi'(x,y)^\top z)^2 \\ &\ge C_3\sum_{x,y}(\psi'(x,y)^\top z)^2 ,\end{aligned}$$

for a positive C_3, since d^θ_ρ is in the loglinear class, for every θ. Now, note that, if z can be expressed as a linear combination of $\{\psi'(x,y)\}_{x,y}$, the summation cannot be zero for non-zero z. Thus, if $\{\psi'(x,y)\}_{x,y}$ spans $\mathbb{R}^{d_M}$, that is, the feature matrix is full rank, then there exists an absolute positive constant κ, such that we have

$$\left\|\nabla^2_\theta A(\theta)\right\|_2 \ge \kappa > 0 .$$

Thus, the function $A(\theta)$ is κ-strongly convex. This, in turn, implies that

$$\begin{aligned}A(\theta^*) - A(\widetilde{\theta}) &\ge \left\langle \nabla_\theta A(\widetilde{\theta}), \theta^* - \widetilde{\theta}\right\rangle + \frac{\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|^2_2 \\ &\ge \left\langle \nabla_\theta A(\widetilde{\theta}), \theta^* - \widetilde{\theta}\right\rangle + \frac{\kappa}{2}\left\|\widetilde{\theta} - \theta^*\right\|^2_{\Sigma'_{\mathcal{D}_n,M}}\end{aligned}$$

$$= -\mathbb{E}_{(x,y)\sim d_\rho^{\widetilde{\theta}}}\left[\left\langle \psi'(x,y), \widetilde{\theta}-\theta^*\right\rangle\right] + \frac{\kappa}{2}\left\|\widetilde{\theta}-\theta^*\right\|^2_{\Sigma'_{\mathcal{D}_n,M}} ,$$

using a similar argument as in the proof of Theorem 4.2. Thus, substituting in the original gap expression, we obtain

$$\begin{aligned} G\left(d_\rho^{\widetilde{\theta}}\right) &\geq D\left(d_\rho^{\widetilde{\theta}}\right) + \frac{\beta\kappa}{2}\left\|\widetilde{\theta}-\theta^*\right\|_2^2 \\ &\geq D\left(d_\rho^{\widetilde{\theta}}\right) + \Omega\left(\frac{d_M+1}{\beta n}\right) , \end{aligned}$$

for any $n \geq O\left(tr((\Sigma'_{\mathcal{D}})^\dagger)/(S_M(B')^2)\right)$, by Theorem J.7. □

D. Convergence of Gradient Descent for RLHF Reward Learning (Section 5)

In this section, we will prove convergence bounds for projected gradient descent for the RLHF reward learning phase. Recall that, the projected gradient update rule for the reward learning phase is given as

$$\omega_{t+1} = \operatorname*{proj}_{\omega:\|\omega\|_2\leq F}\left(\omega_t - \eta\nabla_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)\right) .$$

First, we show Lipschitzness and smoothness of the loss function.

Lemma D.1. *The RLHF reward learning objective $\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)$ is $2\exp(2F)$-Lipschitz and $2\exp(2F)$-smooth.*

Proof. Note that the gradient of $\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)$ satisfies

$$\begin{aligned} \left\|\nabla_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)\right\|_2 &= \left\|\frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \nabla_\omega \log\left(1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)\right)\right\|_2 \\ &\leq \frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \frac{\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}{1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}\left\|\left(\phi(x,y^w)-\phi(x,y^l)\right)\right\|_2 \\ &\leq \exp(2F)\left\|\phi(x,y^w)-\phi(x,y^l)\right\|_2 \\ &\leq 2\exp(2F) . \end{aligned}$$

Moreover, the Hessian of $\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)$ satisfies

$$\begin{aligned} &\left\|\nabla^2_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)\right\|_2 = \left\|\frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \nabla_\omega \frac{\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}{1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}\left(\phi(x,y^w)-\phi(x,y^l)\right)\right\|_2 \\ &\leq \frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \frac{\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}{\left(1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)\right)^2}\left\|\left(\phi(x,y^w)-\phi(x,y^l)\right)\left(\phi(x,y^w)-\phi(x,y^l)\right)^\top\right\|_2 \\ &\leq \exp(2F)\left\|\left(\phi(x,y^w)-\phi(x,y^l)\right)\right\|_2 \\ &\leq 2\exp(2F) . \end{aligned}$$

The result follows. □

Next, we show that $\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)$ satisfies the PL condition (Karimi et al., 2016) defined below.

Definition D.2. A function $\mathcal{L}$ is said to satisfy the PL condition with coefficient $C_{PL} > 0$ if, for every ω in the domain of $\mathcal{L}$, we have

$$\|\nabla_\omega \mathcal{L}(\omega)\|_2^2 \geq C_{PL}\left(\mathcal{L}(\omega) - \mathcal{L}^*\right) ,$$

where $\mathcal{L}^*$ is the minimum value of $\mathcal{L}$ in its domain.

Lemma D.3. *Let* $L_2 = 2\exp(2F)$ *and*

$$C_{PL} = \frac{\exp(-2F)\xi(1+\exp(-2F))}{n(1+\exp(2F)^2} \,,$$

where

$$0 < \xi = \min_{(x,y^w,y^l)\sim\mathcal{D}_n} \left\|\phi(x,y^w) - \phi(x,y^l)\right\|_2^2 \,.$$

Then, we have

$$\frac{1}{2}\left\|\nabla_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)\right\|^2 \geq C_{PL}\left(\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{RLHF}}(\mathcal{D}_n)\right) \,.$$

Proof. Due to the assumption on the features and parameter vectors, we have

$$\begin{aligned}
\frac{1}{2}\left\|\nabla_\omega \mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n)\right\|^2 &\geq \frac{1}{2n^2}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \frac{\exp(-2F)}{1+\exp(2F)} \min_{(x,y^w,y^l)\sim\mathcal{D}_n} \left\|\phi(x,y^w) - \phi(x,y^l)\right\|_2^2 \\
&\geq \frac{\exp(-2F)\xi}{n(1+\exp(2F))} \,.
\end{aligned}$$

On the other hand, note that, for some ω^* such that $\|\omega^*\|_2 \leq F$, we have

$$\begin{aligned}
\mathcal{L}^\omega_{\text{RLHF}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{RLHF}}(\mathcal{D}_n) &= \mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}_n}\Big[\log\left(1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)\right) \\
&\qquad - \log\left(1+\exp\left((\omega^*)^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)\right)\Big] \\
&= \frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \log\frac{1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}{1+\exp\left((\omega^*)^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)} \\
&\leq \frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \left(\frac{1+\exp\left(\omega^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)}{1+\exp\left((\omega^*)^\top\left(\phi(x,y^w)-\phi(x,y^l)\right)\right)} - 1\right) \\
&\leq \frac{1}{n}\sum_{(x,y^w,y^l)\sim\mathcal{D}_n} \frac{1+\exp(2F)}{1+\exp(-2F)} \\
&\leq \frac{1+\exp(2F)}{1+\exp(-2F)} \,,
\end{aligned}$$

where the third inequality follows from $\log x \leq x-1$, for $x > 0$. Solving for C_{PL} the equation

$$C_{PL}\frac{1+\exp(2F)}{1+\exp(-2F)} = \frac{\exp(-2F)\xi}{n(1+\exp(2F)} \,,$$

we obtain

$$C_{PL} = \frac{\exp(-2F)\xi(1+\exp(-2F))}{n(1+\exp(2F)^2} \,.$$

□

Now, we are ready to state the convergence result for gradient descent.

Theorem 5.1. *For every* $t \geq 0$*, the gradient descent procedure* (4) *with learning rate* $\eta = 1/\exp(2F)$ *satisfies*

$$\left\|\omega_t - \omega^*_{\mathcal{D}_n}\right\|^2_{\Sigma_{\mathcal{D}_n,R}} \leq O\left(1-\frac{1}{n}\right)^t \,.$$

Proof. The projected gradient descent rule is equivalent to the proximal gradient update (Karimi et al., 2016), given as

$$\omega_{t+1} = \arg\min_{\omega} \left(\langle \nabla_\omega \mathcal{L}^{\omega_t}_{\text{RLHF}}(\mathcal{D}_n), \omega - \omega_t \rangle + \frac{L_2}{2} \|\omega - \omega_t\|_2^2 + g(\omega) - g(\omega_t) \right) ,$$

where $g(\omega) = 0$, if $\|\omega\|_2 \leq F$ and ∞ otherwise. To see that, note that we can equivalently write the above as

$$\begin{aligned} \omega_{t+1} &= \arg\min_{\omega} \left\| \omega - \left(\omega_t - \frac{1}{L_2} \nabla_\omega \mathcal{L}^{\omega_t}_{\text{RLHF}}(\mathcal{D}_n) \right) \right\|_2 , \text{ s.t. } \|\omega\|_2 \leq F \\ &= \text{proj}_{\omega : \|\omega\|_2 \leq F} \left(\omega_t - \frac{1}{L_2} \nabla_\omega \mathcal{L}^{\omega_t}_{\text{RLHF}}(\mathcal{D}_n) \right) . \end{aligned}$$

Theorem 5 of (Karimi et al., 2016) gives us linear rates of convergence for projected gradient descent under the *proximal PL* condition. This condition is shown in Appendix G of (Karimi et al., 2016) to be equivalent to the following condition.

Definition D.4. A function F is said to satisfy the Kurdyka-Lojasiewicz condition with exponent $1/2$ if there exists $C > 0$ such that

$$\min_{s \in \partial F(\omega)} \|s\|_2^2 \geq C(F(\omega) - F_*) ,$$

where $\partial F(\omega)$ is the Frechet subdifferential of F at ω and F_* denotes the minimum value of F.

Note that, in our case we have $F(\omega) = \mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n) - g(\omega)$ and the Frechet subdifferential of this function in the domain $\{\omega : \|\omega\|_2 \leq F\}$ only contains $\nabla_\omega \mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n)$. Thus, the above condition is equivalent to the PL condition. As a consequence Theorem 5 of (Karimi et al., 2016) implies that

$$\mathcal{L}^{\omega_t}_{\text{RLHF}}(\mathcal{D}_n) - \mathcal{L}^{*}_{\text{RLHF}}(\mathcal{D}_n) \leq \left(1 - \frac{C_{PL}}{L_2}\right)^t \left(\mathcal{L}^{\omega_0}_{\text{RLHF}}(\mathcal{D}_n) - \mathcal{L}^{*}_{\text{RLHF}}(\mathcal{D}_n)\right) .$$

Now, recalling the Hessian of our loss, note that for any non-zero vector $v \in \mathbb{R}^{d_P}$, we have

$$\begin{aligned} & v^\top \nabla^2_\omega \mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n) v \\ &= v^\top \left(\frac{1}{n} \sum_{(x,y^w,y^l) \sim \mathcal{D}_n} \frac{\exp\left(\omega^\top \left(\phi(x,y^w) - \phi(x,y^l)\right)\right)}{\left(1 + \exp\left(\omega^\top \left(\phi(x,y^w) - \phi(x,y^l)\right)\right)\right)^2} \left(\phi(x,y^w) - \phi(x,y^l)\right) \left(\phi(x,y^w) - \phi(x,y^l)\right)^\top \right) v \\ &\geq \frac{\exp(-F)}{(1+\exp(2F))^2} \|v\|^2_{\Sigma_{\mathcal{D}_n,R}} . \end{aligned}$$

Thus, $\mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n)$ is $\frac{\exp(-F)}{(1+\exp(2F))^2}$-strongly convex with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}_n,R}}$ around ω^*. Therefore, for any ω we have

$$\begin{aligned} \mathcal{L}^{\omega}_{\text{RLHF}}(\mathcal{D}_n) - \mathcal{L}^{*}_{\text{RLHF}}(\mathcal{D}_n) &\geq \langle \nabla_\omega \mathcal{L}^{*}_{\text{RLHF}}(\mathcal{D}_n), \omega - \omega^* \rangle + \frac{\exp(-2F)}{(1+\exp(2F))^2} \|\omega - \omega^*\|^2_{\Sigma_{\mathcal{D}_n,R}} \\ &\geq \frac{\exp(-2F)}{(1+\exp(2F))^2} \|\omega - \omega^*\|^2_{\Sigma_{\mathcal{D}_n,R}} . \end{aligned}$$

Putting everything together, we obtain

$$\|\omega - \omega^*\|^2_{\Sigma_{\mathcal{D}_n,R}} \leq O\left(\left(1 - \frac{1}{n}\right)^t\right) .$$

□

E. Convergence of Natural Policy Gradient for RLHF (Section 5)

In this section, we prove fast convergence rates for the a version of the natural policy gradient algorithm with loglinear policy class. We begin by deriving the gradient of the KL-regularized objective. Throughout, let us fix reward function r and dataset $\mathcal{D}_n$.

Lemma E.1. *Given reward r and policy π_θ, the gradient of $\mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n)$, for the softmax policy class can be written as*

$$\nabla_\theta \mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n) = \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \overline{\psi}_\theta(x,y) \,,$$

where

$$\overline{\psi}_\theta(x,y) = \psi(x,y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \psi(x,y') \,.$$

Proof. First, note that, for softmax policies with linear action preferences, we have, for any given (x,y),

$$\begin{aligned}
\nabla_\theta \pi_\theta(y|x) &= \nabla_\theta \frac{\exp(\theta^\top \psi(x,y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y'))} \\
&= \frac{\exp(\theta^\top \psi(x,y)) \sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y'))}{(\sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y')))^2} \psi(x,y) - \frac{\exp(\theta^\top \psi(x,y)) \sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y')) \psi(x,y')}{(\sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y')))^2} \psi(x,y) \\
&= \pi_\theta(y|x) \left(\psi(x,y) - \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[\psi(x,y')] \right) \,.
\end{aligned}$$

On the other hand, for the regularizer, we have

$$\begin{aligned}
\nabla_\theta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) &= \nabla_\theta \left(\theta^\top \psi(x,y) - \log \overline{Z}_\theta(x) - \log \mu(y|x) \right) \\
&= \psi(x,y) - \frac{1}{\overline{Z}_\theta(x)} \sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x,y)) \psi(x,y) \\
&= \psi(x,y) - \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[\psi(x,y')] \,,
\end{aligned}$$

where

$$\overline{Z}_\theta(x) = \sum_{y \in \mathcal{Y}} \exp(\theta^\top \psi(x,y)) \,.$$

Using the definition of $\mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n)$ and the above derivations, we have

$$\begin{aligned}
\nabla_\theta \mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n) &= \frac{1}{n} \sum_{x \in \mathcal{D}_n} \nabla_\theta \left(\sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \right) \\
&= \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \left(\psi(x,y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \psi(x,y') \right) \\
&\quad - \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(\psi(x,y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y',x) \psi(x,y') \right) \\
&= \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \left(\psi(x,y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \psi(x,y') \right) \\
&\quad - \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \psi(x,y) + \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \psi(x,y') \sum_{y \in \mathcal{Y}} \pi_\theta(y|x)
\end{aligned}$$

$$= \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \left(\psi(x,y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \psi(x,y') \right)$$
$$= \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_\theta(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \right) \overline{\psi}_\theta(x,y) \, .$$

□

Now we consider the following gradient update rule. Let $\Psi_n \in \mathbb{R}^{d_P \times nY}$ denote the feature matrix corresponding to $\mathcal{D}_n$ with columns $\psi(x,y)$, for every $(x,y) \in \mathcal{D}_n \times \mathcal{Y}$. We will assume that Ψ_n is full column rank. For every $t \geq 0$, let

$$\theta_{t+1} = \theta_t + \eta' \left(\Psi_n \Psi_n^\top \right)^\dagger \nabla_\theta \mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n) \, . \tag{11}$$

First, we will expand this gradient expression in the following lemma.

Lemma E.2. *For every $t \geq 0$, the gradient update can be written as*

$$\theta_{t+1} = \theta_t + \frac{\eta'}{n} \left(\Psi_n \Psi_n^\top \right)^\dagger \Psi_n H(\pi_{\theta_t}) \alpha_t \, ,$$

where $\mathbb{R}^{nY \times nY} \ni H(\pi) = diag(\pi) - M(\pi)$, for any policy π, with $M(\pi)$ being a blog-diagonal matrix composed of n blocks $\pi(x)\pi(x)^\top \in \mathbb{R}^{Y \times Y}$, with $\pi(x) = [\pi(y|x)]_{y \in \mathcal{Y}}$, for every $x \in \mathcal{D}_n$, and

$$\alpha_t = \beta \Psi_n^\top \theta_t - r - \beta \log \mu - \frac{\left(\beta \Psi_n^\top \theta_t - r - \beta \log \mu \right)^\top \mathbf{1}}{Y} \cdot \mathbf{1} \, .$$

Here, $r = [r(x,y)]_{(x,y) \in \mathcal{D}_n \times \mathcal{Y}}$ denotes the reward vector, and $\log \mu = [\log \mu(y|x)]_{(x,y) \in \mathcal{D}_n \times \mathcal{Y}}$ denotes the vector of log values for te reference policy.

Proof. Using the gradient update and Lemma E.1, we have

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \eta' \left(\Psi_n \Psi_n^\top \right)^\dagger \nabla_\theta \mathcal{V}_r^{\pi_\theta}(\mathcal{D}_n) \\
&= \theta_t + \eta' \left(\Psi_n \Psi_n^\top \right)^\dagger \frac{1}{n} \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_{\theta_t}(y|x) \left(r(x,y) - \beta \log \left(\frac{\pi_{\theta_t}(y|x)}{\mu(y|x)} \right) \right) \overline{\psi}_{\theta_t}(x,y) \\
&= \theta_t + \frac{\eta'}{n} \left(\Psi_n \Psi_n^\top \right)^\dagger \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_{\theta_t}(y|x) \Big(r(x,y) - \beta \theta_t^\top \psi(x,y) + \beta \log \overline{Z}_{\theta_t}(x) + \beta \log \mu(y|x) \Big) \overline{\psi}_{\theta_t}(x,y) \\
&= \theta_t + \frac{\eta'}{n} \left(\Psi_n \Psi_n^\top \right)^\dagger \sum_{x \in \mathcal{D}_n} \sum_{y \in \mathcal{Y}} \pi_{\theta_t}(y|x) \Big(r(x,y) - \beta \theta_t^\top \psi(x,y) + \beta \log \mu(y|x) \Big) \overline{\psi}_{\theta_t}(x,y) \\
&= \theta_t + \frac{\eta'}{n} \left(\Psi_n \Psi_n^\top \right)^\dagger \Psi_n H(\pi_{\theta_t}) \left(r - \beta \Psi_n^\top \theta_t + \beta \log \mu \right) \\
&= \theta_t - \frac{\eta'}{n} \left(\Psi_n \Psi_n^\top \right)^\dagger \Psi_n H(\pi_{\theta_t}) \left(\beta \Psi_n^\top \theta_t - r - \beta \log \mu - \frac{\left(\beta \Psi_n^\top \theta_t - r - \beta \log \mu \right)^\top \mathbf{1}}{Y} \cdot \mathbf{1} \right) \, ,
\end{aligned}$$

where the fourth equality follows from the observation that

$$\beta \sum_y \pi_\theta(y|x) \log \overline{Z}_\theta(x) \overline{\psi}_\theta(x,y) = \beta \log \overline{Z}_\theta(x) \sum_y \pi_\theta(x,y) \left(\psi(x,y) - \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[\psi(x,y')] \right) = 0 \, ,$$

while the last equality follows from the fact that $H(\pi_\theta) c\mathbf{1} = \mathbf{0}$, for any constant c. □

Now we will express α_{t+1} in a different way using the above derivation.

Lemma E.3. *For every $t \geq 0$, we have*

$$\alpha_{t+1} = \left(I - (\eta' \beta / n) H(\pi_{\theta_t}) \right) \alpha_t$$

Proof. Observe that

$$\begin{aligned}
\alpha_{t+1} &= \beta\Psi_n^\top\theta_{t+1} - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_{t+1} - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1} \\
&= \beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1} + \beta\Psi_n^\top(\theta_{t+1}-\theta_t) - \frac{\beta(\Psi_n^\top(\theta_t-\theta_{t+1}))^\top\mathbf{1}}{Y}\cdot\mathbf{1} \\
&= \beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1} - \frac{\beta(\Psi_n^\top(\theta_t-\theta_{t+1}))^\top\mathbf{1}}{Y}\cdot\mathbf{1} \\
&\quad - \beta\Psi_n^\top\left(\frac{\eta'}{n}\left(\Psi_n\Psi_n^\top\right)^\dagger\Psi_n H(\pi_{\theta_t})\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right)\right) \\
&= \left(I - (\beta\eta'/n)\Psi_n^\top\left(\Psi_n\Psi_n^\top\right)^\dagger\Psi_n H(\pi_{\theta_t})\right)\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right) \\
&\quad - \frac{\beta(\Psi_n^\top(\theta_t-\theta_{t+1}))^\top\mathbf{1}}{Y}\cdot\mathbf{1} \\
&= \left(I - (\beta\eta'/n)H(\pi_{\theta_t})\right)\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right) \\
&\quad - \frac{\beta(\Psi_n^\top(\theta_t-\theta_{t+1}))^\top\mathbf{1}}{Y}\cdot\mathbf{1}\,,
\end{aligned}$$

where the third equality uses Lemma E.2 and the last equality follows from the fact that

$$\Psi_n^\top\left(\Psi_n\Psi_n^\top\right)^\dagger\Psi_n = \Psi_n^\top(\Psi_n^\dagger)^\top\Psi_n^\dagger\Psi_n = (\Psi_n^\dagger\Psi_n)^\top\Psi_n^\dagger\Psi_n = I\,,$$

since we assume Ψ_n to be full column rank. For the last term in the derivation above, we have

$$\begin{aligned}
&\frac{\beta(\Psi_n^\top(\theta_t-\theta_{t+1}))^\top\mathbf{1}}{Y}\cdot\mathbf{1} \\
&= \frac{\beta\eta'}{nY}\left(\Psi_n^\top\left(\Psi_n\Psi_n^\top\right)^\dagger\Psi_n H(\pi_{\theta_t})\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right)\right)^\top\mathbf{1}\cdot\mathbf{1} \\
&= \frac{\beta\eta'}{nY}\left(H(\pi_{\theta_t})\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right)\right)^\top\mathbf{1}\cdot\mathbf{1} \\
&= \frac{\beta\eta'}{nY}\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\cdot\mathbf{1}\right)^\top H(\pi_{\theta_t})^\top\mathbf{1}\cdot\mathbf{1} \\
&= \mathbf{0}\,,
\end{aligned}$$

where we have used the fact that $H(\pi_\theta)^\top\mathbf{1} = \mathbf{0}$. The result follows. □

Next, we will decompose the matrix $H(\pi_\theta)$ into simpler pieces and explore its structure.

Lemma E.4. *The eigenvalues of $H(\pi_\theta)$ satisfy the following. The lowest eigenvalue is $\lambda_1 = 0$ with multiplicity n with corresponding eigenvectors e_i, for each $i \in [n]$, where e_i are the vectors of ones in indices $Y(i-1)$ to Yi and zeros everywhere else. Furthermore, we have that $\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}}\pi(y|x) \le \lambda_2$ and $\lambda_{\max} \le \max_{x\in\mathcal{D}_n, y\in\mathcal{Y}}\pi(y|x)$.*

Proof. Again, given $x \in \mathcal{D}_n$, let us denote by $\pi(x)$ the vector $[\pi(y|x)]_{y\in\mathcal{Y}}$. Note that, for any policy π, we can write $H(\pi)$ as

$$H(\pi) = \begin{bmatrix} diag(\pi(x_1)) & 0 & \dots & 0 \\ 0 & diag(\pi(x_2)) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & diag(\pi(x_n)) \end{bmatrix}$$

$$- \begin{bmatrix} \pi(x_1)\pi(x_1)^\top & 0 & \dots & 0 \\ 0 & \pi(x_2)\pi(x_2)^\top & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \pi(x_n)\pi(x_n)^\top \end{bmatrix}$$
$$= \begin{bmatrix} H(\pi(x_1)) & 0 & \dots & 0 \\ 0 & H(\pi(x_2)) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & H(\pi(x_n)) \end{bmatrix}$$

Now, given $x \in \mathcal{D}_n$, Lemma 22 of (Mei et al., 2020) states that the spectrum of $H(\pi(x))$ satisfies $\lambda_1 = 0$ with corresponding eigenvector $\mathbf{1} \in \mathbb{R}^Y$, and

$$\pi(y_{i-1}|x) \leq \lambda_i \leq \pi(y_i|x) \,,$$

for each $2 \leq i \leq Y$, where $\lambda_1 \leq \ldots \leq \lambda_Y$ and $\pi(y_1|x) \leq \ldots \leq \pi(y_Y|x)$. Furthermore, it is known that the spectrum of a block diagonal matrix is composed of the eigenvalues of each block, counting multiplicities. Thus, we have that 0 is the lowest eigenvalue of $H(\pi)$ occurring with multiplicity n. The rest follows. □

Lemma E.5. *Let $v \in \mathbb{R}^{nY}$ be any given vector. Then, we have that*

$$\left\| (I - H(\pi)) \left(v - \frac{v^\top \mathbf{1}}{Y} \mathbf{1} \right) \right\|_2 \leq \left(1 - \min_{x \in \mathcal{D}_n, y \in \mathcal{Y}} \pi(y|x) \right) \left\| v - \frac{v^\top \mathbf{1}}{Y} \mathbf{1} \right\|_2$$

Proof. First, for every $i \in [n]$, let $e_i \in \mathbb{R}^{nY}$ denote a vector with entries 1 at the indices $Y(i-1)$ to Yi and 0 everywhere else. Note that

$$\sum_{i \leq n} e_i = \mathbf{1} \in \mathbb{R}^{nY} \,.$$

Next, let $v(j)$ denote an nY-dimensional vector with entries v_k for each $Y(j-1) \leq k \leq Yj$. Since $H(\pi)$ is diagonalizable, as a symmetric matrix, any vector can be represented as a linear combination of its eigenvectors. Since $H(\pi)$ is symmetric, this representation is unique. Now, by Lemma E.4, note that

$$v = \sum_{j \leq nY} a_j u_j = \sum_{j \leq n} a_j e_j + \sum_{k=n+1}^{nY} a_k u_k = \sum_{j \leq n} \frac{v(j)^\top \mathbf{1}_Y}{Y} e_j + \sum_{k=n+1}^{nY} a_k u_k = \frac{v^\top \mathbf{1}}{Y} \mathbf{1} + \sum_{k=n+1}^{nY} a_k u_k \,,$$

where $(u_{i,j})_{i \leq n, j \leq Y}$ is the eigenvector basis, with the first n eigenvectors being e_i, for $i \leq n$. Thus, we have that

$$v' = v - \frac{v^\top \mathbf{1}}{Y} \mathbf{1} = \sum_{k=n+1}^{nY} a_k u_k \,,$$

with $a_{n+1} > 0$, and that

$$\|v'\|_2 = \sum_{j=n+1}^{nY} a_j^2 \,.$$

From the above, we obtain

$$(I - H(\pi))\, v' = \sum_{j=n+1}^{nY} a_j (1 - \lambda_j) u_j \,,$$

and thus, by Lemma E.4,

$$\|(I - H(\pi))\, v'\|_2 = \sqrt{\sum_{j=n+1}^{nY} a_j^2 (1 - \lambda_j)^2}$$

$$\leq \sqrt{(1-\lambda_{n+1})\left(\sum_{j=n+1}^{nY} a_j^2\right)}$$
$$= (1-\lambda_{n+1})\left\|v'\right\|_2$$
$$\leq \left(1 - \min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi(y|x)\right)\left\|v'\right\|_2 .$$

□

Lemma E.6. *Suppose η' and β are such that $\eta'\beta/n \leq 1$. For the loglinear policy class, for every $t \geq 1$,*

$$\left\|\alpha_t\right\|_2 \leq \frac{2\left(\beta B + 1\right)\sqrt{Y}}{\exp\left(\eta'\beta \sum_{s=1}^{t-1} \min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_\theta(y|x)\right)} .$$

Proof. By Lemma E.3 and Lemma E.5, for all $t \geq 1$,

$$\left\|\left(I - (\eta'\beta/n)H(\pi_{\theta_{t+1}})\right)\alpha_{t+1}\right\|_2 \leq \left(1 - (\eta'\beta/n)\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_t}(y|x)\right)\left\|\alpha_t\right\|_2$$
$$\leq \frac{1}{\exp\left((\eta'\beta/n)\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_t}(y|x)\right)}\left\|\alpha_t\right\|_2$$
$$\leq \frac{1}{\exp\left((\eta'\beta/n)\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_t}(y|x)\right)}\left(1 - \min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_{t-1}}(y|x)\right)\left\|\alpha_{t-1}\right\|_2$$
$$\leq \frac{1}{\exp\left((\eta'\beta/n)\sum_{s=t-1}^{t}\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_s}(y|x)\right)}\left\|\alpha_{t-1}\right\|_2$$
$$\leq \frac{1}{\exp\left((\eta'\beta/n)\sum_{s=1}^{t}\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_s}(y|x)\right)}\left\|\alpha_1\right\|_2 .$$

For the first iteration, observe that

$$\left\|\alpha_1\right\|_2 = \left\|\beta\Psi_n^\top\theta_0 - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_0 - r - \beta\log\mu\right)^\top \mathbf{1}}{Y}\mathbf{1}\right\|_2$$
$$\leq \left\|\beta\Psi_n^\top\theta_0 - r - \beta\log\mu\right\|_2 + \frac{1}{\sqrt{Y}}\left\|\beta\Psi_n^\top\theta_0 - r - \beta\log\mu\right\|_2\left\|\mathbf{1}\right\|_2$$
$$= 2\left\|\beta\Psi_n^\top\theta_0 - r - \beta\log\mu\right\|_2$$
$$\leq 2\left(\beta\left\|\Psi_n^\top\theta_0\right\|_\infty + 1\right)\sqrt{Y}$$
$$\leq 2\left(\beta B + 1\right)\sqrt{Y} ,$$

where the second inequality follows from the triangle inequality, the Cauchy-Schwarz inequality and the fact that rewards lie in the unit ball, while the last follows from the fact that the features lie in a unit subspace of $\mathbb{R}^{d_P}$, while $\left\|\theta_0\right\|_\infty \leq B$. The result follows. □

Lemma E.7. *There exists a constant $C = C(\beta, Y, B) > 0$, such that, for all $t \geq 1$, we have $\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_t}(y|x) \geq C$.*

Proof. First, by Lemma E.5, note that, for any $t \geq 1$,

$$\left\|\alpha_{t+1}\right\|_2 \leq \left(1 - (\eta'\beta/n)\min_{x\in\mathcal{D}_n, y\in\mathcal{Y}} \pi_{\theta_t}(y|x)\right)\left\|\alpha_t\right\|_2 \leq \left\|\alpha_t\right\|_2 \leq \ldots \leq \left\|\alpha_1\right\|_2 \leq 2\left(\beta B + 1\right)\sqrt{Y} ,$$

where the second inequality follows from the fact that policies are probability distributions, and the last inequality follows from Lemma E.6. Next, observe that, for any $(x, y) \in \mathcal{D}_n \times \mathcal{Y}$,

$$\left|\psi(x,y)^\top\theta_t - \frac{1}{\beta}r(x,y) - \log\mu(y|x) - \frac{\left(\Psi_n^\top\theta_t - r/\beta - \log\mu\right)^\top\mathbf{1}}{Y}\right|$$

$$
\begin{aligned}
&\leq \frac{1}{\beta}\left|\beta\psi(x,y)^\top\theta_t - r(x,y) - \beta\log\mu(y|x) - \frac{\beta\left(\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top\mathbf{1}}{Y}\right| \\
&\leq \frac{1}{\beta}\left\|\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{\left(\beta\Psi_n^\top\theta_t - r - \beta\log\mu\right)^\top\mathbf{1}}{Y}\mathbf{1}\right\|_2 \\
&\leq \frac{1}{\beta}\left\|\alpha_t\right\|_2 \\
&\leq 2(B+1/\beta)\sqrt{Y}\,.
\end{aligned}
$$

Now, define $(x_1,y_1) = \arg\min_{x\in\mathcal{D}_n,y\in\mathcal{Y}}\psi(x,y)^\top\theta_t$ and $(x_2,y_2) = \arg\max_{x\in\mathcal{D}_n,y\in\mathcal{Y}}\psi(x,y)^\top\theta_t$. By the above, we have

$$
\begin{aligned}
\Psi_n(x_1,y_1)^\top\theta_t &\geq \frac{1}{\beta}r(x_1,y_1) + \log\mu(y_1|x_1) + \frac{\left(\Psi_n^\top\theta_t - r/\beta - \log\mu\right)^\top\mathbf{1}}{Y} - 2(B+1/\beta)\sqrt{Y}\,, \\
-\Psi_n(x_2,y_2)^\top\theta_t &\geq -\frac{1}{\beta}r(x_2,y_2) - \log\mu(y_2|x_2) - \frac{\left(\Psi_n^\top\theta_t - r/\beta - \log\mu\right)^\top\mathbf{1}}{Y} - 2(B+1/\beta)\sqrt{Y}\,,
\end{aligned}
$$

which imply

$$
\begin{aligned}
\min_{x\in\mathcal{D}_n,y\in\mathcal{Y}}\pi_{\theta_t}(y|x) &\geq \min_{x\in\mathcal{D}_n,y\in\mathcal{Y}}\frac{\exp(\psi(x,y)^\top\theta_t)}{\sum_{y'\in\mathcal{Y}}\exp(\psi(x,y)\top\theta_t)} \geq \frac{1}{Y}\exp\left(\left(\Psi_n(x_1,y_1) - \Psi_n(x_2,y_2)\right)^\top\theta_t\right) \\
&\geq \frac{1}{Y}\exp\left(\frac{1}{\beta}\left(r(x_1,y_1) - r(x_2,y_2)\right) + \log\frac{\mu(y_1|x_1)}{\mu(y_2|x_2)} - 4\left(B+1/\beta\right)\sqrt{Y}\right) \\
&\geq \frac{1}{Y}\exp\left(-\frac{1}{\beta} - 4\left(B+1/\beta\right)\sqrt{Y}\right) = C\,.
\end{aligned}
$$

□

Let us denote by $softmax(\Psi_n^\top v)$ the policy $\exp(\psi(x,y)^\top v)/\sum_{y'}\exp(\psi(x,y')^\top v)$, for any parameter vector v and pair $(x,y)\in\mathcal{D}_n\times\mathcal{Y}$. Now, we are ready to prove the main result of this section.

Theorem E.8. *Let $\pi_{\theta_t} = softmax(\Psi_n^\top\theta_t)$. Using update rule (11) with $\eta' \leq n/\beta$, for all $t\geq 1$,*

$$
\mathcal{V}_r^{\pi_{\theta^*}}(\mathcal{D}_n) - \mathcal{V}_r^{\pi_{\theta_t}}(\mathcal{D}_n) \leq \frac{2\sqrt{Y}(B+1/\beta)}{\exp((\beta\eta'/n)\cdot C\cdot(t-1))}\,,
$$

where

$$
C = \frac{1}{Y}\exp\left(-\frac{1}{\beta} - 4\left(B+1/\beta\right)\sqrt{Y}\right)\,.
$$

Proof. Observe that, since $\pi_{\theta^*}\propto\mu(y|x)\exp(r(x,y)/\beta)$, we have

$$
\begin{aligned}
&\mathcal{V}_r^{\pi_{\theta^*}}(\mathcal{D}_n) - \mathcal{V}_r^{\pi_{\theta_t}}(\mathcal{D}_n) = \frac{1}{n}\sum_{x\in\mathcal{D}_n}\sum_{y\in\mathcal{Y}}\left(\pi_{\theta_n^*}(y|x)r(x,y) - \beta D_{\mathrm{KL}}(\pi_{\theta^*}||\mu) - \pi_{\theta_t}(y|x)r(x,y) + \beta D_{\mathrm{KL}}(\pi_{\theta_t}||\mu)\right) \\
&\leq \frac{1}{n}\sum_{x\in\mathcal{D}_n}\left\|\pi_{\theta^*}(\cdot|x) - \pi_{\theta_t}(\cdot|x)\right\|_1 \\
&\quad + \frac{1}{n}\sum_{x\in\mathcal{D}_n} -\beta D_{\mathrm{KL}}(\pi_{\theta^*}||\pi_{\theta^*}) + \pi_{\theta^*}(y|x)r(x,y) + \beta D_{\mathrm{KL}}(\pi_{\theta_t}||\pi_{\theta^*}) - \pi_{\theta_t}(y|x)r(x,y) \\
&\leq \frac{2}{n}\sum_{x\in\mathcal{D}_n}\left\|\pi_{\theta^*}(\cdot|x) - \pi_{\theta_t}(\cdot|x)\right\|_1 + \beta D_{\mathrm{KL}}(\pi_{\theta_t}||\pi_{\theta^*}) \\
&\leq (2Y+\beta)D_{\mathrm{KL}}(\pi_{\theta_t}||\pi_{\theta^*})
\end{aligned}
$$

$$\leq (2Y+\beta)\left\|\Psi_n^\top\theta^* - \Psi_n\theta_t + \frac{(\Psi_n^\top(\theta_t-\theta^*))^\top\mathbf{1}}{Y}\mathbf{1}\right\|_\infty^2 ,$$

where the third inequality uses Pinsker's inequality and the last one follows from Lemma J.8. Now, note that the optimal softmax policy parameter θ^* satisfies, for each $(x,y)\in\mathcal{D}_n$,

$$\psi(x,y)^\top\theta^* = \frac{1}{\beta}\left(r(x,y)+\log\mu(y|x)\right) ,$$

by setting the gradient at (x,y) to 0. Its existence is guaranteed by the assumption that $r^*\in\mathcal{F}$ and Lemma J.1. Thus, we have

$$\begin{aligned}
\mathcal{V}_r^{\pi_{\theta^*}}(\mathcal{D}_n) - \mathcal{V}_r^{\pi_{\theta_t}}(\mathcal{D}_n) &\leq (2Y+\beta)\left\|\Psi_n^\top\theta^* - \Psi_n\theta_t + \frac{(\Psi_n^\top(\theta_t-\theta^*))^\top\mathbf{1}}{Y}\mathbf{1}\right\|_\infty^2 \\
&= (2Y+\beta)\left\|\frac{1}{\beta}(r+\log\mu) - \Psi_n^\top\theta_t + \frac{(\beta\Psi^\top\theta_t - r - \beta\log\mu)^\top\mathbf{1}}{\beta Y}\mathbf{1}\right\|_\infty^2 \\
&= \frac{(2Y+\beta)}{\beta}\left\|\beta\Psi_n^\top\theta_t - r - \beta\log\mu - \frac{(\beta\Psi^\top\theta_t - r - \beta\log\mu)^\top\mathbf{1}}{Y}\mathbf{1}\right\|_\infty^2 \\
&\leq \frac{2}{\beta}\frac{(2Y+\beta)2\sqrt{Y}(\beta B+1)}{\exp((\eta'\beta/n)C(t-1))} ,
\end{aligned}$$

where the last inequality follows from Lemma E.6 and Lemma E.7. □

F. Convergence of Gradient Descent for DPO (Section 5)

In this section, we will prove convergence bounds for the projected gradient descent procedure for DPO. Recall that the projected gradient descent is defined as

$$\theta_{t+1} = \operatorname*{proj}_{\theta:\|\theta\|_2\leq B}\left(\theta_t - \eta''\nabla_\theta\mathcal{L}_{\text{DPO}}^\theta(\mathcal{D}_n)\right) ,$$

We begin by showing that the DPO objective satisfies the PL condition (Karimi et al., 2016) stated in Definition D.2. We will show that the DPO objective satisfies this condition for the loglinear parametrization. First, we need to show that such an objective has Lipschitz gradients, which holds under the assumption that the parameter vectors θ have a length of no more than B.

Lemma F.1. *The DPO objective $\mathcal{L}_{\text{DPO}}^\theta(\mathcal{D}_n)$ is Lipschitz continuous with parameter $L_1' = \beta\exp(2\beta(B+J))$ and has Lipschitz gradients with parameter $L_2' = \beta^2\exp(2\beta(B+J))$, where*

$$J = \max_{(x,y^w,y^l)\in\mathcal{D}_n}\beta\left|\log\frac{\mu(y^w|x)}{\mu(y^l|x)}\right| .$$

Proof. Note that, in order to show L-Lipschitzness, it suffices to prove that the Hessian of $\mathcal{L}_{\text{DPO}}^\theta(\mathcal{D}_n)$ has bounded eigenvalues. Let us first compute the Hessian. Before doing that, we first simplify the gradient expression, when instantiated for the softmax parametrization. First, given parameter vector θ, corresponding to π_θ, we have

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}^\theta(\mathcal{D}_n) &= -\mathbb{E}_{(y^w,y^l,x)\sim\mathcal{D}_n}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y^w|x)}{\mu(y^w|x)} - \beta\log\frac{\pi_\theta(y^l|x)}{\mu(y^l|x)}\right)\right] \\
&= -\mathbb{E}_{(y^w,y^l,x)\sim\mathcal{D}_n}\left[\log\sigma\left(\beta\log\frac{\exp(\theta^\top\psi(x,y^w))}{\sum_{y\in\mathcal{Y}}\exp(\theta^\top\psi(x,y))} - \beta\log\frac{\exp(\theta^\top\psi(x,y^l))}{\sum_{y\in\mathcal{Y}}\exp(\theta^\top\psi(x,y))} - \beta\log\frac{\mu(y^w|x)}{\mu(y^l|x)}\right)\right] \\
&= -\mathbb{E}_{(y^w,y^l,x)\sim\mathcal{D}_n}\left[\log\sigma\left(\beta\theta^\top(\psi(x,y^w)-\psi(x,y^l)) - \beta\log\frac{\mu(y^w|x)}{\mu(y^l|x)}\right)\right] \\
&= \mathbb{E}_{(y^w,y^l,x)\sim\mathcal{D}_n}\left[\log\left(1+\exp\left(\beta\theta^\top\left(\psi(x,y^w)-\psi(x,y^l)\right) - J(x,y^w,y^l)\right)\right)\right] ,
\end{aligned}$$

where we let

$$J(x, y^w, y^l) = \beta \log \frac{\mu(y^w|x)}{\mu(y^l|x)} .$$

Based on the above, we have

$$\begin{aligned}
&\nabla_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) = \nabla_\theta \mathbb{E}_{(y^w, y^l, x)\sim\mathcal{D}_n} \left[\log\left(1 + \exp\left(\beta\theta^\top \left(\psi(x, y^w) - \psi(x, y^l)\right) - J(x, y^w, y^l)\right)\right)\right] \\
&= \frac{1}{n} \sum_{(x, y^w, y^l)\in\mathcal{D}_n} \frac{\beta \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)}{\left(1 + \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)\right)} \left(\psi(x, y^w) - \psi(x, y^l)\right) ,
\end{aligned}$$

and

$$\begin{aligned}
&\nabla^2_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) = \frac{1}{n} \sum_{(x, y^w, y^l)\in\mathcal{D}_n} \\
&\quad \nabla_\theta \frac{\beta \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)}{\left(1 + \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)\right)} \left(\psi(x, y^w) - \psi(x, y^l)\right) \\
&= \frac{1}{n} \sum_{(x, y^w, y^l)\in\mathcal{D}_n} \\
&\frac{\beta^2 \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)}{\left(1 + \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right)\right)^2} \left(\psi(x, y^w) - \psi(x, y^l)\right) \left(\psi(x, y^w) - \psi(x, y^l)\right)^\top .
\end{aligned}$$

Now, define

$$E(\theta, x, y) = \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - J(x, y^w, y^l)\right) .$$

Note that we have

$$\left\|\nabla \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)\right\|_2 \le \beta \exp(2\beta(B + J)) \left\|\psi(x, y^w) - \psi(x, y^l)\right\|_2 \le \beta \exp(2\beta(B + J)) ,$$

and

$$\begin{aligned}
\nabla^2_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) &= \beta^2 \sum_{(x, y^w, y^l)\in\mathcal{D}_n} \frac{E(\theta, x, y)}{n\left(1 + E(\theta, x, y)\right)^2} \psi(x)\psi(x)^\top \\
&\preceq \frac{\beta^2 \exp\left(2\beta(B + J)\right)}{n} \sum_{(x, y^w, y^l)\in\mathcal{D}_n} \psi(x)\psi(x)^\top \\
&\preceq \beta^2 \exp\left(2\beta(B + J)\right) I_d ,
\end{aligned}$$

where the last inequality follows from the fact that the feature norms are bounded by 1, and thus the maximum eigenvalue of the sample covariance matrix is no more than 1. □

Next, we show that the DPO objective satisfies the PL condition under some mild assumption on the data.

Lemma F.2. *Assume that, for each triple $(x, y^w, y^l) \in \mathcal{D}_n$, we have that $\psi(x, y^w) \neq \psi(x, y^l)$. Then, if we let*

$$C'_{PL} = \frac{\beta \exp(-2\beta(B + J))^3 \left(1 + \exp(-2\beta(B + J))\right)}{n\left(1 + \exp(2\beta(B + J))\right)^2} \min_{(x, y^w, y^l)\in\mathcal{D}_n} \left\|\psi(x, y^w) - \psi(x, y^l)\right\|^2 ,$$

we have

$$\frac{1}{2} \left\|\nabla \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)\right\|^2 \ge C'_{PL} \left(\mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n)\right)$$

where $\mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n) = \min_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)$ denotes the optimal loss value.

Proof. Using the notation $E(\theta, x, y) = \exp\left(\beta\theta^\top(\psi(x, y^w) - \psi(x, y^l)) - E(x)\right)$, and noting that every quantity in the expression below is non-negative, we have

$$\begin{aligned}\frac{1}{2}\left\|\nabla_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n)\right\|^2 &\geq \frac{\beta^2}{2n^2}\sum_{(x,y)\in\mathcal{D}_n}\frac{E(\theta,x,y)^2}{\left(1+E(\theta,x,y)\right)^2}\left\|\psi(x,y^w)-\psi(x,y^l)\right\|^2 \\ &\geq \frac{\beta^2\exp(-2\beta(B+J))^2}{n\left(1+\exp(2\beta(B+J))\right)^2}\min_{(x,y)\in\mathcal{D}_n}\left\|\psi(x,y^w)-\psi(x,y^l)\right\|^2\end{aligned}$$

since $-J \leq E(x) \leq J$, and thus, $\exp(-2\beta(B+J)) \leq E(\theta, x, y) \leq \exp(2\beta(B+J))$. On the other hand, observe that

$$\begin{aligned}\mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n) &= \mathbb{E}_{(x,y)\sim\mathcal{D}_n}\Big[\log\left(1+E\left(\theta,x,y\right)\right) - \log\left(1+E(\theta^*,x,y)\right)\Big] \\ &= \frac{1}{n}\sum_{i=1}^n \log\frac{1+E(\theta,x_i,y_i)}{1+E(\theta^*,x_i,y_i)} \\ &\leq \frac{1}{n}\sum_{i=1}^n\left(\frac{1+E(\theta,x_i,y_i)}{1+E(\theta^*,x_i,y_i)} - 1\right) \\ &= \frac{1}{n}\sum_{i=1}^n\frac{E(\theta,x_i,y_i)-E(\theta^*,x_i,y_i)}{1+E(\theta^*,x_i,y_i)} \\ &\leq \frac{1}{n}\sum_{i=1}^n\frac{E(\theta,x_i,y_i)}{1+E(\theta^*,x_i,y_i)} \\ &\leq \frac{\exp(2\beta(B+J))}{1+\exp(-2\beta(B+J))}\,.\end{aligned}$$

Now, the assumption on the data implies that, there exists ξ such that

$$0 < \xi' = \min_{(x,y^w,y^l)\in\mathcal{D}_n}\left\|\psi(x,y^w)-\psi(x,y^l)\right\|^2\,.$$

Using ξ' and solving the equation

$$C'_{PL}\cdot\frac{\exp(2\beta(B+J))}{1+\exp(-2\beta(B+J))} = \frac{\beta\exp\left(-2\beta(B+J)\right)^2}{n\left(1+\exp\left(2\beta(B+J)\right)\right)^2}\xi'$$

for C'_{PL}, we obtain

$$C'_{PL} = \frac{\beta\exp(-2\beta(B+J))^3\left(1+\exp(-2\beta(B+J))\right)}{n\left(1+\exp(2\beta(B+J))\right)^2}\xi'\,. \tag{12}$$

□

These two conditions are enough to obtain the following result.

Theorem 5.3. *For every $t \geq 0$, the gradient descent procedure (6) with learning rate $\eta'' = O\left(1/\beta^2\right)$ satisfies*

$$\left\|\theta_t - \theta^*_{\mathcal{D}_n}\right\|^2_{\Sigma_{\mathcal{D}_n,P}} \leq O\left(\frac{1}{\beta}\left(1-\frac{\beta}{n}\right)^t\right)\,.$$

Proof. A similar argument as the one in the proof of Theorem 5.1 implies that, for every $t \geq 1$ we have:

$$\mathcal{L}^{\theta_t}_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n) \leq \left(1-\frac{C'_{PL}}{L'_2}\right)^t\left(\mathcal{L}^{\theta_0}_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n)\right)\,,$$

where $L_1' = \beta \exp(2\beta(B+J))$ is the Lipschitz constant and $C'_{LPL} = BC_{PL}$. Using the expression for the Hessian derived in the proof of Lemma F.1 we have, for any non-zero vector v, that

$$v^\top \nabla^2_\theta \mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) v \geq \beta^2 \frac{\exp\left(-\beta\left(B+J\right)\right)}{1+\exp(\beta(B+J))} \|v\|^2_{\Sigma_{\mathcal{D}_n,P}} \ .$$

Thus, $\mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n)$ is $\beta^2 \frac{\exp(-\beta(B+J))}{1+\exp(\beta(B+J))}$-strongly convex with respect to the semi-norm $\|\cdot\|_{\Sigma_\mathcal{D}}$ around $\theta^*_{\mathcal{D}_n}$, where $\theta^*_{\mathcal{D}_n}$ is a parameter vector that achieves $\mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n)$. for any θ, we have

$$\begin{aligned}
\mathcal{L}^\theta_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\mathcal{D}_n} &\geq \left\langle \nabla_\theta \mathcal{L}^*_{\text{DPO}}(\mathcal{D}_n), \theta - \theta^*_{\mathcal{D}_n} \right\rangle + \beta^2 \frac{\exp\left(-\beta\left(B+J\right)\right)}{2(1+\exp(\beta(B+J)))} \|\theta - \theta^*\|^2_{\Sigma_{\mathcal{D}_n,P}} \\
&\geq \beta^2 \frac{\exp\left(-\beta\left(B+J\right)\right)}{2(1+\exp(\beta(B+J)))} \|\theta - \theta^*\|^2_{\Sigma_{\mathcal{D}_n,P}}
\end{aligned}$$

Therefore, using the upper bound on the loss, we finally obtain, for any iterate θ_t of GD,

$$\begin{aligned}
\left\|\theta_t - \theta^*_{\mathcal{D}_n}\right\|^2_{\Sigma_{\mathcal{D}_n,P}} &\leq O\left(\frac{\mathcal{L}^{\theta_0}_{\text{DPO}}(\mathcal{D}_n) - \mathcal{L}^*_{\mathcal{D}_n}}{\beta^2}\left(1-\frac{\beta}{n}\right)^t\right) \\
&\leq O\left(\frac{1}{\beta}\left(1-\frac{\beta}{n}\right)^t\right)
\end{aligned}$$

□

G. Non-realizable Rewards (Section 6)

In this section, we will derive the proofs of the two results from Section 6. We restate them for convenience.

Theorem 6.1. *Let $\delta > 0$. Suppose that Assumption 6.1 holds. Then, with probability at least $1-\delta$, we have $G\left(\pi_{\widehat{\theta}}\right) \leq D\left(\pi_{\widehat{\theta}}\right) + \widetilde{\Theta}\left(\Lambda_R \sqrt{d_R/n}\right) + 2\epsilon_{\text{app}}$.*

Proof. From Theorem 4.1, we have

$$\begin{aligned}
G(\pi_{\widehat{\theta}}) &= D\left(\pi_{\widehat{\theta}}\right) + \left\langle d^*_\rho - d^{\pi_{\widehat{\theta}}}_\rho, r^* - r_{\widehat{\omega}} \right\rangle \\
&= D\left(\pi_{\widehat{\theta}}\right) + \left\langle d^*_\rho - d^{\pi_{\widehat{\theta}}}_\rho, r^* - r_{\omega^*} \right\rangle + \left\langle d^*_\rho - d^{\pi_{\widehat{\theta}}}_\rho, r_{\omega^*} - r_{\widehat{\omega}} \right\rangle \\
&\leq D\left(\pi_{\widehat{\theta}}\right) + 2\max_{x,y} |r^*(x,y) - r_{\omega^*}(x,y)| + O\left(\Lambda_R \sqrt{\frac{d_R}{n}}\right) \\
&\leq D\left(\pi_{\widehat{\theta}}\right) + O\left(\Lambda_R \sqrt{\frac{d_R}{n}}\right) + 2\epsilon_{\text{app}} \ ,
\end{aligned}$$

where for the first inequality we have used Cauchy-Schwarz, while for the last inequality we have used Theorem 4.1 and Condition 6.1. □

Next, we prove the analogous result for DPO.

Theorem 6.2. *Let $\delta > 0$. Suppose that Assumption 6.1 and the condition of Lemma 4.1 hold. Then, with probability at least $1-\delta$, we have $G\left(\pi_{\widetilde{\theta}}\right) \leq D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\Lambda_P d_P/(\beta n)\right) + \min\{2\epsilon_{\text{app}}, O\left(\beta D_{\text{KL}}\left(\pi_{\theta^*} || \pi^*\right)\right)\}$.*

Proof. Since the ground-truth reward function is not linear, we are not guaranteed that the optimal policy representable in terms of the reward is loglinear. Let π^* denote the optimal policy for the KL-regularized problem with respect to r^*, and let π_{θ^*} be the loglinear approximation of π^*.

$$G\left(\pi_{\widetilde{\theta}}\right) = V^{\text{opt}}_{r^*}(\rho) - V^{\pi_{\widetilde{\theta}}}_{r^*}(\rho)$$

$$
\begin{aligned}
&= D\left(\pi_{\widetilde{\theta}}\right) + \left(\mathcal{V}_{r^*}^{\pi^*}(\rho) - \mathcal{V}_{r^*}^{\pi_{\widetilde{\theta}}}(\rho)\right) \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \mathbb{E}_{x\sim\rho, y\sim\pi^*(\cdot|x)}\left[r^*(x,y) - \beta\log\frac{\pi^*(y|x)}{\mu(y|x)}\right] - \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[r^*(x,y) - \beta\log\frac{\pi_{\widetilde{\theta}}(y|x)}{\mu(y|x)}\right] \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \mathbb{E}_{x\sim\rho, y\sim\pi^*(\cdot|x)}\left[\beta\log\frac{\pi^*(y|x)}{\mu(y|x)} + \beta\log Z(x) - \beta\log\frac{\pi^*(y|x)}{\mu(y|x)}\right] \\
&\quad - \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\beta\log\frac{\pi^*(y|x)}{\mu(y|x)} + \beta\log Z(x) - \beta\log\frac{\pi_{\widetilde{\theta}}(y|x)}{\mu(y|x)}\right] \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\beta\log\frac{\pi_{\widetilde{\theta}}(y|x)}{\mu(y|x)} - \beta\log\frac{\pi^*(y|x)}{\mu(y|x)}\right] \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\beta\log\frac{\pi_{\widetilde{\theta}}(y|x)}{\mu(y|x)} - \beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)}\right] \\
&\quad + \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\beta\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)} - \beta\log\frac{\pi^*(y|x)}{\mu(y|x)}\right] \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + \mathbb{E}_{x\sim\rho, y\sim\pi_{\widetilde{\theta}}(\cdot|x)}\left[\beta\log\pi_{\theta^*}(y|x) - \beta\log\pi^*(y|x)\right] \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + \beta\sum_x \rho(x)\sum_y \frac{\pi_{\widetilde{\theta}}(y|x)}{\pi_{\theta^*}(y|x)}\log\frac{\pi_{\theta^*}(y|x)}{\pi^*(y|x)} \\
&\le D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + \beta Y\exp(2B) D_{\mathrm{KL}}\left(\pi_{\theta^*}||\pi^*\right) \ .
\end{aligned}
$$

On the other hand, using the same idea as in the proof of Theorem 4.2, we have

$$
\begin{aligned}
G\left(\pi_{\widetilde{\theta}}\right) &= V_{r^*}^{\mathrm{opt}}(\rho) - V_{r^*}^{\pi_{\widetilde{\theta}}}(\rho) \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \left(\mathcal{V}_{r^*}^{\pi_{\theta^*}}(\rho) - \mathcal{V}_{r^*}^{\pi_{\widetilde{\theta}}}(\rho)\right) \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \left(\mathcal{V}_{r^*}^{\pi_{\theta^*}}(\rho) - \mathcal{V}_{r^*}^{\pi_{\theta^*}}(\rho)\right) + \left(\mathcal{V}_{r_{\omega^*}}^{\pi_{\theta^*}}(\rho) - \mathcal{V}_{r_{\omega^*}}^{\pi_{\widetilde{\theta}}}(\rho)\right) + \left(\mathcal{V}_{r_{\omega^*}}^{\pi_{\widetilde{\theta}}}(\rho) - \mathcal{V}_{r^*}^{\pi_{\widetilde{\theta}}}(\rho)\right) \\
&= D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + \mathbb{E}_{(x,y)\sim d_\rho^{\theta^*}}\left[r^*(x,y) - r_{\omega^*}(x,y)\right] + \mathbb{E}_{(x,y)\sim d_\rho^{\widetilde{\theta}}}\left[r_{\omega^*}(x,y) - r^*(x,y)\right] \\
&\le D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + \mathbb{E}_{(x,y)\sim d_\rho^{\theta^*}}\left[r^*(x,y) - r_{\omega^*}(x,y)\right] + 2\max_{x,y}|r^*(x,y) - r_{\omega^*}(x,y)| \\
&\le D\left(\pi_{\widetilde{\theta}}\right) + \Theta\left(\frac{\Lambda(d_P+1)}{\beta n}\right) + 2\epsilon_{\mathrm{app}} \ ,
\end{aligned}
$$

where the fourth equality follows from Theorem 4.2 and the last inequality from Condition 6.1. □

H. The DPO Extension to MDPs (Section 7)

First, we start with MDP setting preliminaries.

H.1. Deterministic Markov Decision Processes

An infinite-horizon discounted deterministic Markov decision process (MDP) is a mathematical object $\mathcal{M} = (\mathcal{X}, \mathcal{Y}, T, r^*, \gamma, \rho)$, where $\mathcal{X}$ denotes the state space, $\mathcal{Y}$ denotes the action space, both of which are assumed to be finite with cardinalities X and Y, respectively. $T : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}$ denotes the deterministic transition function, where $T(x,y)$ denotes the next state after taking action y in state x. The reward function is denoted by $r^* : \mathcal{X}\times\mathcal{Y} \to [0,1]$. Finally, $\gamma \in [0,1)$ denotes the discount factor, while $\rho \in \Delta(\mathcal{X})$ denotes the initial state distribution.

Policies π are mappings from states to distributions over actions, that is, $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$. Given policy π, the state occupancy measure of state x with respect to initial state x_0 is given as

$$
d_{x_0}^{\pi}(x) = (1-\gamma)\sum_{t\ge 0}\gamma^t \mathbb{P}\left(x_t = x | x_0, \pi\right) \ ,
$$

while the state-action occupancy measure is given as $d^\pi_{x_0}(x,y) = d^\pi_{x_0}(x)\pi(y|x)$. We also write $d^\pi_\rho(x,y) = \mathbb{E}_{x_0\sim\rho}[d_{x_0}(x,y)]$. Furthermore, given policy π and an arbitrary reward function r, the value function of policy π with respect to reward r is defined as

$$V^\pi_r(x) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(x_t,y_t)\Big|x_0 = x,\pi\right] ,$$

and the action-value function is defined as

$$Q^\pi_r(x,y) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(x_t,y_t)\Big|x_0 = x, y_0 = y,\pi\right] ,$$

for every state-action pair (x,y). We denote by $V^\pi_r(\rho) = \mathbb{E}_{x\sim\rho}[V^\pi_r(x)]$ the expected value function over the initial distribution.

H.2. DPO for MDPs

A direct extension of DPO to the MDP setting is not straightforward. To understand this, it is enough to see that the optimal policy-to-reward mapping in this case is not linear. Fix a reward function r. The gradient of the KL-regularized objective with respect to r is given as

$$\nabla_\theta \mathcal{V}^\theta_r(\rho) = \frac{1}{1-\gamma}\sum_x d^{\pi_\theta}_\rho(x)\sum_y \pi_\theta(y|x)\left(r(x,y) + \gamma\mathcal{V}^{\pi_\theta}(T(x,y)) - \beta\log\frac{\pi_\theta(y|x)}{\mu(y|x)}\right)\overline{\psi}_\theta(x,y) ,$$

where $\overline{\psi}_\theta(x,y) = \psi(x,y) - \mathbb{E}_{y'\sim\pi_\theta(\cdot|x)}[\psi(x,y')]$. See Appendix I for derivations. What complicates things is the occupancy measure d^π_ρ, which is non-linearly dependent on policy π, and the gradient of the occupancy measure. To allow for the change of variables to carry through in this case, we utilize the dual formulation of Problem (P3.2):

$$\begin{aligned} \max_{d_\rho}\ & \sum_{x,y} d_\rho(x,y)r(x,y) - \beta\sum_{x,y} d_\rho(x,y)\log\frac{d_\rho(x,y)}{d^\mu_\rho(x,y)} \\ \text{s.t.}\ & \sum_y d_\rho(x,y) = (1-\gamma)\rho(x) + \gamma\sum_{x',y'}\mathbb{1}\left(x = T\left(x',y'\right)\right)d_\rho(x',y'), \forall x\in\mathcal{X} , \end{aligned} \tag{P3.2'}$$

where we have used that

$$V^\pi_r(\rho) = \sum_{x,y} d^\pi_\rho(x,y)r(x,y) ,$$

and also taken the KL-divergence of the occupancy measures, instead of the actual policies. This is a convex program and thus any stationary points are optimal. The Lagrangian of the above problem can be written as

$$\begin{aligned} L(d_\rho,\alpha) &= \sum_{x,y} d_\rho(x,y)\left(r^*(x,y) - \beta\log\frac{d_\rho(x,y)}{d^\mu_\rho(x,y)}\right) \\ &\quad + \sum_x \alpha(x)\left(\sum_y d_\rho(x,y) - (1-\gamma)\rho(x) - \gamma\sum_{x',y'}\mathbb{1}(x = T(x',y'))d_\rho(x',y')\right) \\ &= -\beta\sum_{x,y} d_\rho(x,y)\log\frac{d_\rho(x,y)}{d^\mu_\rho(x,y)} - (1-\gamma)\sum_x \rho(x)\alpha(x) \\ &\quad + \sum_{x,y} d_\rho(x,y)\left(\underbrace{r^*(x,y) - \gamma\sum_{x'}\mathbb{1}(x = T(x',y'))\alpha(x') + \alpha(x)}_{e_\alpha(x,y)}\right) , \end{aligned}$$

Then, given (x, y), the gradient of the Lagrangian with respect to $d_\rho(x, y)$ is

$$\nabla_{d_\rho(x,y)} L(d_\rho, \alpha) = -\beta \left(\log \frac{d_\rho(x, y)}{d^\mu_\rho(x, y)} - \mathbf{1} \right) + e_\alpha(x, y) \, ,$$

which, when set to zero, yields

$$d_\rho(x, y) = d^\mu_\rho(x, y) \exp\left(\frac{1}{\beta} e_\alpha(x, y) \right) \exp(-1) \, .$$

Primal feasibility implies that $\sum_{x,y} d_\rho(x, y) = 1$, thus, our choice of α should satisfy such condition. Letting $Z = \exp(1)$, and α^* be the optimal Lagrange multiplier, we have

$$d^*_\rho(x, y) = \frac{1}{Z} d^\mu_\rho(x, y) \exp\left(\frac{1}{\beta} e_{\alpha^*}(x, y) \right) \, . \tag{13}$$

Writing the expression for the reward function, we get

$$r^*(x, y) = \beta \log \frac{d^*_\rho(x, y)}{d^\mu_\rho(x, y)} + \beta + \gamma \sum_{x'} \mathbb{1}(x = T(x', y')) \alpha^*(x') - \alpha^*(x) \, .$$

Now, observe that, given a trajectory $\tau = (x_0, y_0, x_1, \ldots)$, we can write the discounted return using the above expression and obtain

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t r^*(x_t, y_t) &= \sum_{t=0}^{\infty} \gamma^t \left(\beta \log \frac{d^*_\rho(x_t, y_t)}{d^\mu_\rho(x_t, y_t)} + \beta + \gamma \sum_{x} \mathbb{1}(x = T(x', y')) \alpha^*(x) - \alpha^*(x_t) \right) \\
&= \sum_{t=0}^{\infty} \gamma^t \left(\beta \log \frac{d^*_\rho(x_t, y_t)}{d^\mu_\rho(x_t, y_t)} + \beta + \gamma \alpha^*(x_{t+1}) - \alpha^*(x_t) \right) \qquad (14) \\
&= \sum_{t=0}^{\infty} \gamma^t \left(\beta \log \frac{d^*_\rho(x_t, y_t)}{d^\mu_\rho(x_t, y_t)} + \beta - \alpha^*(x_0) \right) \, , \qquad (15)
\end{aligned}$$

where for Equation (14) we have used the fact that the transitions are deterministic, and for Equation (15) note that the terms $\alpha^*(x)$ cancel each other out.

Now, let us get back to the BT preference model for MDPs. Given a dataset $\mathcal{D}_n$ of pairs of trajectories, each pair of which starts from the same initial state, we can express the MLE loss directly in terms of the occupancy measures using the above derivation as follows:

$$\mathcal{L}_{\text{DPO}}(d_\rho) = -\mathbb{E}_{(\tau^w, \tau^l) \sim \mathcal{D}_n} \left[\log \sigma \left(\sum_{t=0}^{\infty} \gamma^t \beta \log \frac{d_\rho(x^w_t, y^w_t)}{d^\mu_\rho(x^w_t, y^w_t)} - \sum_{t=0}^{\infty} \gamma^t \beta \log \frac{d_\rho(x^l_t, y^l_t)}{d^\mu_\rho(x^l_t, y^l_t)} \right) \right] \, ,$$

where we have used the fact that the terms β and $\alpha^*(x_0)$ cancel out.

Now, note that the minimizer to the above loss may not satisfy the Bellman flow constraints of Problem (P3.2'). Thus, we need to restrict the domain of the problem to the following set

$$\mathcal{B} = \left\{ d \in \Delta(\mathcal{X} \times \mathcal{Y}) : \sum_y d(x, y) = (1 - \gamma)\rho(x) + \gamma \sum_{x', y'} \mathbb{1}\left(x = T(x', y')\right) d(x', y'), \forall x \in \mathcal{X} \right\}$$

H.3. DPO for MDPs with Loglinear Occupancy Measures

Similar to the contextual bandit setting, we want to write the DPO loss such that it resembles logistic regression. For loglinear occupancy measures, as defined in Definition 7.1, with parameter set restricted to

$$\Theta' := \{\theta \in \mathbb{R}^{d_M} : d^{\pi_\theta}_\rho \in \mathcal{B}\} \, ,$$

we can write the loss so that it resembles logistic regression. Note that the domain of θ is restricted only to those parameters which imply that $d_\rho^{\pi_\theta}$ is an occupancy measure with respect to the underlying MDP. For this case, the DPO loss becomes

$$\mathcal{L}_{\mathcal{D}_n}(\theta) = -\mathbb{E}_{(\tau^w,\tau^l)\sim\mathcal{D}_n}\left[\log\sigma\left(\beta\theta^\top\left(\sum_{t=0}^{\infty}\gamma^t\left(\psi(x_t^w,y_t^w)-\psi(x_t^l,y_t^l)\right)\right)+K(\tau^w,\tau^l)\right)\right]$$

where

$$K(\tau^w,\tau^l) = \sum_{t=0}^{\infty}\gamma^t\log\frac{d_\rho^\mu(x_t^l,y_t^l)}{d_\rho^\mu(x_t^w,y_t^w)}\ .$$

Given the learned occupancy measure $d_\rho^{\pi_\theta}$, one can finally compute an optimal policy, for each state-action pair, as

$$\pi_\theta(y|x) = \frac{d_\rho^{\pi_\theta}(x,y)}{\sum_y d_\rho^{\pi_\theta}(x,y)}\ .$$

Note that, in general, the quantity $K(\tau^w,\tau^l)$ is not easy to compute as it requires access to the occupancy measure with respect to μ. However, in practice, $K(\tau^w,\tau^l)$ can be treated as a hyperparameter of the problem and tuned accordingly.

I. Gradient Expression for KL-regularized Objective in MDPs

In this section, we derive the gradient for the loglinear policy class. We rewrite the problem below for convenience.

$$\max_\theta \mathbb{E}_{x\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(x_t,y_t)-\beta D_{\text{KL}}\left(\pi_\theta(\cdot|x_t)||\mu(\cdot|x_t)\right)\right)\Big|y_t\sim\pi_\theta(\cdot|x_t)\right]$$

Lemma I.1. *Let*

$$\mathcal{V}_r^{\pi_\theta}(x) = \mathbb{E}_{x\sim\rho,y_t\sim\pi_\theta(\cdot|x_t)}\left[\sum_{t\geq 0}\gamma^t\left(r(x_t,y_t)-\beta\log\frac{\pi_\theta(y_t|x_t)}{\mu(y_t|x_t)}\right)\right]$$

and

$$\mathcal{Q}_r^{\pi_\theta}(x,y) = r(x,y)+\gamma\mathcal{V}_r^{\pi_\theta}(T(x,y))\ .$$

The gradient expression for $\mathcal{V}^{\pi_\theta}(\rho)$ is given by

$$\nabla_\theta\mathcal{V}_r^\theta(\rho) = \frac{1}{1-\gamma}\sum_x d_\rho^{\pi_\theta}(x)\sum_y\pi_\theta(y|x)\left(\mathcal{Q}_r^{\pi_\theta}(x,y)-\beta\log\frac{\pi_\theta(y|x)}{\mu(y|x)}\right)\overline{\psi}_\theta(x,y)\ .$$

Proof. Note that we have

$$\mathcal{V}_r^{\pi_\theta}(\rho) = \mathbb{E}_{x\sim\rho}\left[\sum_y\pi_\theta(y|x)\left(\mathcal{Q}_r^{\pi_\theta}(x,y)-\beta\log\frac{\pi_\theta(y|x)}{\mu(y|x)}\right)\right]\ .$$

Thus, we can write

$$\begin{aligned}\nabla_\theta\mathcal{V}_r^{\pi_\theta}(\rho) &= \sum_{x,y}\rho(x)\Bigg(\nabla_\theta\pi_\theta(y|x)\left(\mathcal{Q}_r^{\pi_\theta}(x,y)-\beta\log\frac{\pi_\theta(y|x)}{\mu(y|x)}\right)\\ &\qquad+\pi_\theta(y|x)\left(\nabla_\theta\mathcal{Q}_r^{\pi_\theta}(x,y)-\frac{\mu(y|x)}{\pi_\theta(y|x)}\nabla_\theta\pi_\theta(y|x)\right)\Bigg)\\ &= \sum_{x,y}\rho(x)\left(\pi_\theta(y|x)\left(\mathcal{Q}_r^{\pi_\theta}(x,y)-\beta\log\frac{\pi_\theta(y|x)}{\mu(y|x)}-1\right)\overline{\psi}_\theta(x,y)+\pi_\theta(y|x)\nabla_\theta\mathcal{Q}_r^{\pi_\theta}(x,y)\right)\end{aligned}$$

$$= \sum_{x,y} \rho(x) \left(\pi_\theta(y|x) \left(\mathcal{Q}_r^{\pi_\theta}(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \overline{\psi}_\theta(x,y) \right)$$
$$+ \gamma \sum_{x,y} \rho(x) \pi_\theta(y|x) \nabla_\theta \mathcal{V}_r^{\pi_\theta}(T(x,y))$$
$$= \frac{1}{1-\gamma} \sum_x d_\rho^{\pi_\theta}(x) \sum_y \pi_\theta(y|x) \left(\mathcal{Q}_r^{\pi_\theta}(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\mu(y|x)} \right) \overline{\psi}_\theta(x,y) \,,$$

where the second equality follows from the derivation of the gradient of loglinear policies (see the proof of Lemma E.1, while the third equality follows from the fact that $\mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[\overline{\psi}_\theta(x,y)] = 0$, for each $x \in \mathcal{X}$. □

J. Technical Lemmas

The purpose of this section is to present various technical results that are useful for our paper. Let us denote by $\Phi \in \mathbb{R}^{d_R \times XY}$ and $\Psi \in \mathbb{R}^{d_P \times XY}$ the reward and policy feature matrices with columns $\phi(x,y)$ and $\psi(x,y)$, respectively.

Lemma J.1. *Assume that $r^* \in \mathcal{F}$, $\pi^* \in \Pi$ and $\mu \in \Pi$, for some $\pi_{r^*}^* \in \arg\max_\pi \mathcal{V}^\pi(\rho)$. Furthermore, assume that the columns space of Φ is a subspace of the column space of Ψ. Then, there exists $\theta^* \in \Theta$, for which π_{θ^*} maximizes the objective of (P1.2) and that can be represented in terms of the ground-truth reward function, i.e. $\pi_{r^*}^*(y|x) = \pi_{\theta^*}(y|x) \propto \mu(y|x) \exp(r^*(x,y)/\beta)$, for all (x,y).*

Proof. Let $x \in \mathcal{X}$. From Equation (1) we have that

$$\pi_{r^*}^*(x,y) \propto \mu(y|x) \exp\left(\frac{1}{\beta} r^*(x,y) \right)$$
$$\propto \exp\left(\theta_\mu^\top \psi(x,y) + \phi(x,y)^\top \omega^* \right) \,,$$

for some $\pi_{r^*}^* \in \arg\max_\pi \mathcal{V}_{r^*}^\pi(\rho)$, where the second relation holds due to the assumptions on the policy class and reward class. Thus, if we can find a $\theta^* \in \mathbb{R}^{d_P}$ such that, for all (x,y), we have

$$\exp\left(\psi(x,y)^\top \theta^* \right) = \exp\left(\theta_\mu^\top \psi(x,y) + \phi(x,y)^\top \omega^* \right) \,,$$

then we have shown that the optimal regularized policy belongs to the loglinear class. For the above to hold, we equivalently need

$$\Psi^\top (\theta^* - \theta_\mu) - \Phi^\top \omega^* = \mathbf{0} \,.$$

The above equation has a solution for θ^* whenever the column space of Φ is contained in the column space of Ψ. □

Lemma J.2. *Assume that $r^* \in \mathcal{F}$, $d_\rho^\mu \in \Pi'$ and $d_\rho^{\pi_{r^*}^*} \in \Pi'$, for some optimal $d_\rho^{\pi_{r^*}^*}$. Furthermore, assume that the column space of $\Phi + \Phi_{\pi_{r^*}^*}$ is contained in the column space of Ψ, where $\Phi_{\pi_{r^*}^*} \in \mathbb{R}^{d_R \times XY}$ has columns*

$$\gamma \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \phi(x_t, y_t) \Big| x_0 = x, \pi_{r^*}^* \right] - \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \phi(x_t, y_t) \Big| x_0 = x, y_0 = T(x,y), \pi_{r^*}^* \right] \,.$$

Then, for finite MDPs with deterministic transitions, there exists θ^* such that $d_\rho^{\pi_{r^*}^*}(x,y)$.

Proof. From Equation (13) we have

$$d_\rho^*(x,y) = \frac{1}{Z} d_\rho^\mu(x,y) \exp\left(\frac{1}{\beta} e_{\alpha^*}(x,y) \right) \,,$$

where

$$e_\alpha(x,y) = r^*(x,y) + \gamma \alpha^*(T(x,y)) - \alpha^*(x) \,,$$

is the advantage function when using α^* and α^* denote the optimal dual variables for Problem (P3.2'). As shown in (Lee et al., 2021), these variables correspond to the optimal value function with respect to r^*. Thus, we have

$$\begin{aligned}
e_{\alpha^*}(x,y) &= r^*(x,y) + \gamma\alpha^*(T(x,y)) - \alpha^*(x) \\
&= \phi(x,y)^\top\omega^* + \gamma\mathbb{E}\left[\sum_{t\geq 0}\gamma^t r^*(x,y)\middle| x_0 = x, \pi^*_{r^*}\right] - \mathbb{E}\left[\sum_{t\geq 0}\gamma^t r^*(x,y)\middle| x_0 = x, y_0 = T(x,y), \pi^*_{r^*}\right] \\
&= \phi(x,y)^\top\omega^* + \gamma\mathbb{E}\left[\sum_{t\geq 0}\gamma^t \phi(x,y)^\top\omega^*\middle| x_0 = x, \pi^*_{r^*}\right] - \mathbb{E}\left[\sum_{t\geq 0}\gamma^t \phi(x,y)^\top\omega^*\middle| x_0 = x, y_0 = T(x,y), \pi^*_{r^*}\right] \\
&= \left(\phi(x,y) - \gamma\mathbb{E}\left[\sum_{t\geq 0}\gamma^t \phi(x,y)\middle| x_0 = x, \pi^*_{r^*}\right] - \mathbb{E}\left[\sum_{t\geq 0}\gamma^t \phi(x,y)\middle| x_0 = x, y_0 = T(x,y), \pi^*_{r^*}\right]\right)^\top \omega^* .
\end{aligned}$$

Thus, we have

$$\begin{aligned}
d^*_\rho(x,y) = \frac{1}{Z}d^\mu_\rho(x,y)\exp\left(\frac{1}{\beta}e_{\alpha^*}(x,y)\right) \propto \exp\Bigg(&\theta_\mu^\top\psi(x,y) \\
&+ (\omega^*)^\top\left(\phi(x,y) + \gamma\mathbb{E}\left[\sum_{t\geq 0}\gamma^t\phi(x,y)\middle| x_0 = x, \pi^*_{r^*}\right] - \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\phi(x,y)\middle| x_0 = x, y_0 = T(x,y), \pi^*_{r^*}\right]\right)\Bigg) .
\end{aligned}$$

For the above to hold, we equivalently need

$$\Psi^\top(\theta^* - \theta_\mu) - \left(\Phi + \Phi_{\pi^*_{r^*}}\right)^\top\omega^* = \mathbf{0} .$$

The above has a solution whenever the column space of $\Phi + \Phi_{\pi^*_{r^*}}$ is contained in the column space of Ψ . □

Next, we will prove a result that connects the suboptimality gap with the gap in terms of the KL-regularized objectives.

Lemma J.3. *For any θ, we have*

$$\beta D_{\text{KL}}\left(\pi_{\theta^*}||\mu\right) - \beta D_{KL}\left(\pi_\theta||\mu\right) \leq D(\pi_\theta) \leq \beta D_{\text{KL}}\left(\pi^{\text{opt}}_{r^*}||\mu\right) - \beta D_{KL}\left(\pi_\theta||\mu\right) ,$$

Proof. Note that, by definition,

$$\begin{aligned}
G\left(\pi_\theta\right) &= V^{\text{opt}}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho) \\
&= \left(V^{\text{opt}}_{r^*}(\rho) - \mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_\theta}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho)\right) \\
&\leq \left(V^{\text{opt}}_{r^*}(\rho) - \mathcal{V}^{\pi^{\text{opt}}_{r^*}}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_\theta}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho)\right) \\
&\leq \beta D_{\text{KL}}\left(\pi^{\text{opt}}_{r^*}||\mu\right) - \beta D_{\text{KL}}\left(\pi_\theta||\mu\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) .
\end{aligned}$$

Similarly,

$$\begin{aligned}
G\left(\pi_\theta\right) &= V^{\text{opt}}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho) \\
&= \left(V^{\text{opt}}_{r^*}(\rho) - \mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_\theta}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho)\right) \\
&\geq \left(V^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) + \left(\mathcal{V}^{\pi_\theta}_{r^*}(\rho) - V^{\pi_\theta}_{r^*}(\rho)\right) \\
&\geq \beta D_{\text{KL}}\left(\pi_{\theta^*}||\mu\right) - \beta D_{\text{KL}}\left(\pi_\theta||\mu\right) + \left(\mathcal{V}^{\pi_{\theta^*}}_{r^*}(\rho) - \mathcal{V}^{\pi_\theta}_{r^*}(\rho)\right) .
\end{aligned}$$

The result follows. □

Next, we will control the quantity $D(\pi_\theta)$ for the DPO setting.

Lemma J.4. *With probability at least $1-\delta$, we have*

$$D(\pi_{\widetilde{\theta}}) \leq \beta D_{\mathrm{KL}}(\pi^{\mathrm{opt}}_{r^*}, \pi_{\theta^*}) + \frac{\Lambda_P U d_P}{S_P n}\sqrt{\frac{\log(4/\delta)}{2n}}$$
$$= \beta\left(D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*}||\mu\right) - D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right)\right) + O\left(\frac{d_P}{n^{3/2}}\right) .$$

Proof. Recall that, for any θ, we have defined

$$D_{\mathrm{KL}}\left(\pi_\theta||\mu\right) = \sum_x \rho(x)\sum_{x,y}\pi_\theta(y|x)\log\frac{\pi_\theta(y|x)}{\mu(y|x)} .$$

First, note that

$$D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) - D_{\mathrm{KL}}\left(\pi_\theta||\mu\right) = \left(D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) - \frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}}||\mu\right)\right) + \left(\frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}}||\mu\right)\right.$$
$$\left. - \frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}}||\mu\right)\right) + \left(\frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}}||\mu\right) - D_{\mathrm{KL}}\left(\pi_\theta||\mu\right)\right)$$
$$= \left(D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) - \frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}}||\mu\right)\right) + \left(\frac{1}{n}\sum_{x\in\mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}}||\mu\right) - D_{\mathrm{KL}}\left(\pi_\theta||\mu\right)\right) ,$$

where the second equality follows from the exact optimization assumption. Note that the two summands above are deviations from means. If we can show that each individual quantity is bounded, then we can apply Hoeffding bounds. To that end, first, note that

$$D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) = \sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)} = \sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)\log\frac{\mu(y|x)\exp\left(\frac{1}{\beta}r^*(x,y)\right)}{\mu(y|x)}$$
$$= \frac{1}{\beta}\sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)r^*(x,y) \leq \frac{1}{\beta} ,$$

since the reward cannot be more than 1. Similarly, for every $x\in\mathcal{D}$, we have

$$D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}}||\mu\right) = D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}}||\mu\right) = \sum_x \rho(x)\sum_y \pi_{\theta^*_{\mathcal{D}_n}}(y|x)\log\frac{\mu(y|x)\exp\left(\frac{1}{\beta}r^*(x,y)\right)}{\mu(y|x)}$$
$$= \frac{1}{\beta}\sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)r^*(x,y) \leq \frac{1}{\beta} .$$

For other contexts $x\notin\mathcal{D}$, such a relation does not hold. Thus, we take another approach. We show that, for such points, the KL divergence between the learned policy and the sampling policy cannot be too far away from that between the optimal policy and the sampling policy.

$$\left|D_{\mathrm{KL}}\left(\pi_{\theta^*}||\mu\right) - D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}}||\mu\right)\right| = \left|\sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)\log\frac{\pi_{\theta^*}(y|x)}{\mu(y|x)} - \sum_{x,y}\pi_{\theta^*_{\mathcal{D}_n}}(y|x)\log\frac{\pi_{\theta^*_{\mathcal{D}_n}}(y|x)}{\mu(y|x)}\right|$$
$$= \left|\sum_x \rho(x)\sum_y \pi_{\theta^*}(y|x)\left(\log\frac{\pi_{\theta^*}(y|x)}{\pi_{\theta^*}(y|x)} + \log\exp\left(\frac{1}{\beta}r^*(x,y)\right)\right)\right.$$
$$\left. - \sum_{x,y}\pi_{\theta^*_{\mathcal{D}_n}}(y|x)\left(\log\frac{\pi_{\theta^*_{\mathcal{D}_n}}(y|x)}{\pi_{\theta^*}(y|x)} + \log\exp\left(\frac{1}{\beta}r^*(x,y)\right)\right)\right|$$
$$\leq \frac{1}{\beta}\left|\left(V^{\pi_{\theta^*}}_{r^*}(\rho) - V^{\pi_{\theta^*_{\mathcal{D}_n}}}(\rho)\right) - D_{\mathrm{KL}}\left(\pi_{\theta^*}||\pi_{\theta^*_{\mathcal{D}_n}}\right)\right|$$

$$\leq \frac{1}{\beta} + D_{\mathrm{KL}}\left(\pi_{\theta^*} || \pi_{\theta^*_{\mathcal{D}_n}}\right) .$$

Now, for the last term of the right-hand side, we have

$$\begin{aligned} D_{\mathrm{KL}}\left(\pi_{\theta^*} || \pi_{\theta^*_{\mathcal{D}_n}}\right) &= \sum_x \rho(x) \sum_y \pi_{\theta^*}(y|x) \left(\log \pi_{\theta^*}(y|x) - \log \pi_{\theta^*_{\mathcal{D}_n}}(y|x)\right) \\ &= \sum_x \rho(x) \sum_y \pi_{\theta^*}(y|x) \left(\left\langle \psi(x,y), \theta^* - \theta^*_{\mathcal{D}_n} \right\rangle + \log \frac{\sum_{x',y'} \exp\left(\psi(x',y')^\top \theta^*_{\mathcal{D}_n}\right)}{\sum_{x',y'} \exp\left(\psi(x',y')^\top \theta^*\right)}\right) \\ &\leq \frac{\Lambda_P U d_P}{S_P \beta n} , \end{aligned}$$

where the last inequality follows from the same arguments as in the proof of Theorem 4.2. Going back to the original expression, note that, for any given $x \in \mathcal{X}$, we have

$$0 \leq D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right) \leq \frac{1}{\beta} , \quad \text{and} \quad 0 \leq D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}} || \mu\right) \leq \frac{1}{\beta} + \frac{\Lambda_P U d_P}{S_P \beta n} .$$

Thus, by Hoeffding's inequality, for any $\delta \geq 0$, with probability at least $1 - \delta$, we have

$$\left| D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right) - \frac{1}{n} \sum_{x \in \mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\theta^*_{\mathcal{D}_n}} || \mu\right) \right| \leq \frac{1}{\beta} \sqrt{\frac{\log(4/\delta)}{2n}} ,$$

and

$$\left| \frac{1}{n} \sum_{x \in \mathcal{D}} D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}} || \mu\right) - D_{\mathrm{KL}}\left(\pi_\theta || \mu\right) \right| \leq \left(\frac{1}{\beta} + \frac{\Lambda_P U d_P}{S_P \beta n}\right) \sqrt{\frac{\log(4/\delta)}{2n}} ,$$

which implies that

$$-\left(\frac{2}{\beta} + \frac{\Lambda_P U d_P}{S_P \beta n}\right) \sqrt{\frac{\log(4/\delta)}{2n}} \leq D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right) - D_{\mathrm{KL}}\left(\pi_\theta || \mu\right) \leq \left(\frac{2}{\beta} + \frac{\Lambda_P U d_P}{S_P \beta n}\right) \sqrt{\frac{\log(4/\delta)}{2n}} .$$

On the other hand, note that

$$\begin{aligned} D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*} || \mu\right) - D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}} || \mu\right) &= \left(D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*} || \mu\right) - D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right)\right) + \left(D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right) - D_{\mathrm{KL}}\left(\pi_{\widetilde{\theta}} || \mu\right)\right) \\ &\leq \left(D_{\mathrm{KL}}\left(\pi^{\mathrm{opt}}_{r^*} || \mu\right) - D_{\mathrm{KL}}\left(\pi_{\theta^*} || \mu\right)\right) + \left(\frac{2}{\beta} + \frac{\Lambda_P U d_P}{S_P \beta n}\right) \sqrt{\frac{\log(4/\delta)}{2n}} . \end{aligned}$$

□

Next, we prove some useful properties of the log-exp-sum function.

Lemma J.5. *The function defined as*

$$A(\theta) = \sum_x \rho(x) \log \sum_{x,y} \exp\left(\theta^\top \psi(x,y)\right) .$$

is 1*-Lipschitz and* 2*-smooth. Moreover, if the features are sampled from a* 0*-mean distribution and span* R^{d_P}*, then there exists* $\kappa > 0$*, such that* $A(\theta)$ *is* κ*-strongly convex.*

Proof. Let $\theta \in \mathbb{R}^{d_P}$. Note that

$$\nabla_\theta A(\theta) = \sum_x \rho(x) \frac{\sum_y \exp(\psi(x,y)^\top \theta)}{\sum_{y'} \exp(\psi(x,y')^\top \theta)} \psi(x,y')$$

$$
\begin{aligned}
&= \sum_x \rho(x) \sum_y \pi_\theta(y|x)\psi(x,y) \\
&\leq \max_{x,y} \|\psi(x,y)\|_2 \\
&\leq 1 \,.
\end{aligned}
$$

On the other hand, the Hessian of $A(\theta)$ is

$$
\begin{aligned}
\nabla^2_\theta A(\theta) &= \sum_x \rho(x) \sum_y \nabla_\theta \pi_\theta(y|x)\psi(x,y) \\
&= \sum_x \rho(x) \sum_y \pi_\theta(y|x) \left(\psi(x,y) - \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[\psi(x,y')]\right) \psi(x,y)^\top \\
&= \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)} \left[\psi(x,y)\psi(x,y)^\top\right] - \mathbb{E}_{x\sim,\rho,y\sim\pi_\theta(\cdot|x)}[\psi(x,y)]\mathbb{E}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)}[\psi(x,y)]^\top \\
&= \mathbb{E}_{x\sim\rho,y\sim\pi_\theta(y|x)} \left[\left(\psi(x,y) - \mathbb{E}_\theta\left[\psi(x,y)\right]\right)\left(\psi(x,y) - \mathbb{E}_\theta\left[\psi(x,y)\right]\right)^\top\right] \,.
\end{aligned}
$$

By assumption on the feature mapping, we have that

$$
\begin{aligned}
\left\|\nabla^2_\theta A(\theta)\right\|_2 &\leq \max_{x,y} \left\|\left(\psi(x,y) - \mathbb{E}_\theta\left[\psi(x,y)\right]\right)\left(\psi(x,y) - \mathbb{E}_\theta\left[\psi(x,y)\right]\right)^\top\right\|_2 \\
&\leq \max_{x,y} \|\psi(x,y) - \mathbb{E}_\theta[\psi(x,y)]\|_2 \\
&\leq 2 \max_{x,y} \|\psi(x,y)\|_2 = 2 \,.
\end{aligned}
$$

Therefore, the function $A(\theta)$ is 2-smooth in θ. For strong convexity, let ψ be sampled from a 0-mean bounded distribution. Note that, for any non-zero vector in $\mathbb{R}^{d_P}$, we have

$$
\begin{aligned}
z^\top \nabla^2_\theta A(\theta) z &= \mathbb{E}_{x\sim\rho,y\sim\pi_\theta(\cdot|x)} \left[z^\top \psi(x,y)\psi(x,y)^\top z\right] \\
&\geq \min_{\theta,x,y} \pi_\theta(\cdot|x) \sum_{x,y} (\psi(x,y)^\top z)^2 \\
&\geq C_3 \sum_{x,y} (\psi(x,y)^\top z)^2 \,,
\end{aligned}
$$

for a positive C_3, since π_θ is in the loglinear class, for every θ, and using Lemma E.7. Now, note that, if z can be expressed as a linear combination of $\{\psi(x,y)\}_{x,y}$, the summation cannot be zero for non-zero z. Thus, if $\{\psi(x,y)\}_{x,y}$ spans $\mathbb{R}^{d_P}$, that is, the feature matrix is full rank, then there exists an absolute positive constant κ, such that we have

$$
\left\|\nabla^2_\theta A(\theta)\right\|_2 \geq \kappa > 0 \,.
$$

Thus, the function $A(\theta)$ is κ-strongly convex. □

Lemma J.6. *In general, the norms of the gradient and Hessian for the loss of tabular DPO are unbounded from above.*

Proof. Observe that, given policy π and $(x, y^w) \in \mathcal{D}$, we have

$$
\nabla_{\pi(y^w|x)} \mathcal{L}_\mathcal{D}(\pi) = \frac{\beta}{n} \left(1 - \sigma\left(\beta \log \frac{\pi(y^w|x)}{\mu(y^w|x)} - \beta \log \frac{\pi(y^l|x)}{\mu(y^l|x)}\right)\right) \frac{1}{\pi(y^w|x)} \,.
$$

On the other hand, for the second derivative with respect to $\pi(y^w|x)$, we have the following. First, let

$$
f(\pi(y^w|x)) = \beta \log \frac{\pi(y^w|x)}{\mu(y^w|x)} - \beta \log \frac{\pi(y^l|x)}{\mu(y^l|x)} \,.
$$

We have that $\nabla_{\pi(y^w|x)} f(\pi(y^w|x)) = \beta/\pi(y^w|x)$. Now, observe that

$$
\nabla^2_{\pi(y^w|x)} \mathcal{L}_\mathcal{D}(\pi) = \frac{\beta}{n} \nabla_{\pi(y^w|x)} \frac{\exp(f(\pi(y^w|x)))}{\pi(y^w|x)(1 + \exp(f(\pi(y^w|x))))}
$$

$$= \frac{\beta}{n}\left(\frac{\frac{\beta}{\pi(y^w|x)}\exp\left(f(\pi(y^w|x))\right)\pi(y^w|x)(1+\exp(f(\pi(y^w|x))))}{(\pi(y^w|x)(1+\exp(f(\pi(y^w|x)))))^2}\right)$$
$$- \frac{\beta}{n}\left(\frac{\exp\left(f(\pi(y^w|x))\right)\left((1+\exp(f(\pi(y^w|x))))+\pi(y^w|x)\frac{\beta}{\pi(y^w|x)}\exp\left(f(\pi(y^w|x))\right)\right)}{(\pi(y^w|x)(1+\exp(f(\pi(y^w|x)))))^2}\right)$$
$$= \frac{\beta\left((\beta-1)\exp(f(\pi(y^w|x)))(1+\exp(f(\pi(y^w|x))))-\beta\exp(f(\pi(y^w|x)))^2\right)}{n\left(\pi(y^w|x)(1+\exp(f(\pi(y^w|x))))\right)^2}\,.$$

The above numerator is not always non-negative, as solving for $\exp(f(\pi(y^w|x)))$ will show. Moreover, neither the norm of the gradient nor the operator norm of the Hessian can be upper-bounded in general, due to the presence of $\pi(y^w|x)$ in the denominator. □

Theorem J.7 (Theorem 1.(c) of (Shah et al., 2016))**.** *For the BT preference model, B-bounded weight vector and sample size $n \geq O\left(tr(\Sigma^\dagger)/\beta^2 B^2\right)$, where Σ denotes the Laplacian with respect to features, the maximum likelihood estimator satisfies the minimax bounds*

$$\Omega\left(\frac{d}{\beta n}\right) \leq \left\|\widetilde{\theta}-\theta^*\right\|_\Sigma^2 \leq O\left(\frac{d}{\beta n}\right)\,.$$

Lemma J.8 (Lemma 27 of (Mei et al., 2020))**.** *Let $\pi_\theta = softmax(\Psi\theta)$ and $\pi_{\theta'} = softmax(\Psi\theta')$. Then, for any constant c, we have*

$$D_{\mathrm{KL}}(\pi_\theta||\pi_{\theta'}) \leq \frac{1}{2}\left\|\Psi\theta-\Psi\theta'-c^\top\mathbf{1}\right\|^2\,.$$