Understanding Preference Fine-Tuning Through the Lens of Coverage

Anonymous authors
Paper under double-blind review

Abstract

Learning from human preference data has emerged as the dominant paradigm for fine-tuning large language models (LLMs). The two most common families of techniques – online reinforcement learning (RL) such as Proximal Policy Optimization (PPO) and offline contrastive methods such as Direct Preference Optimization (DPO) – were positioned as equivalent in prior work due to the fact that both have to start from the same offline preference dataset. To further expand our theoretical understanding of the similarities and differences between online and offline techniques for preference fine-tuning, we conduct a rigorous analysis through the lens of *dataset coverage*, a concept that captures how the training data covers the test distribution and is widely used in RL. We prove that a global coverage condition is both necessary and sufficient for offline contrastive methods to converge to the optimal policy, but a weaker partial coverage condition suffices for online RL methods. This separation provides one explanation of why online RL methods can perform better than offline methods, especially when the offline preference data is not diverse enough. Finally, motivated by our preceding theoretical observations, we derive a hybrid preference optimization (HyPO) algorithm that uses offline data for contrastive-based preference optimization and online data for KL regularization. Theoretically and empirically, we demonstrate that HyPO is more performant than its pure offline counterpart DPO, while still preserving its computation and memory efficiency.

1 Introduction

Due to the difficulty of manually specifying reward functions for complex tasks (Casper et al., 2023), preference-based learning has emerged as a critical component in the fine-tuning procedure for large language models (LLMs) (Stiennon et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Team et al., 2023). There are two predominant flavors of preference learning for LLMs: *online* reinforcement learning (RL) methods such as PPO (Christiano et al., 2017; Ouyang et al., 2022) and *offline* contrastive methods like Direct Preference Optimization (DPO) (Rafailov et al., 2024b) and Identity Preference Optimization (IPO) (Azar et al., 2024).

Online RL methods usually follow the two-stage procedure prescribed in Ouyang et al. (2022): one first trains a reward model (classifier) on a fixed offline preference dataset before using it to provide reward labels for on-policy generations, which are then fed to a downstream RL algorithm like Proximal Policy Optimization (PPO) (Schulman et al., 2017). Since the reward model is learned from static offline preference data, to avoid over-optimizing the reward model (Gao et al., 2023), one typically adds a reverse KL penalty to encourage the model to stay close to some reference policy. We will refer to this procedure as reinforcement learning from human feedback (RLHF) in this paper. While empirically performant, RLHF requires repeated querying of the reward model (which is often itself an LLM) as well as sampling from the current policy. In response to the computational expense and relatively complex nature of this procedure, purely offline methods like DPO (Rafailov et al., 2024b) and IPO (Azar et al., 2024) have been proposed as alternative methods for preference fine-tuning. These methods do not need to fit separate reward models, instead opting to simply train the policy directly on the offline preference dataset via a ranking loss.

Offline contrastive methods like DPO are usually derived via applying a reparameterization trick to the closed-form solution of the minimum relative entropy problem (Ziebart et al., 2008) that RLHF techniques attempt to approximate. Thus, several authors have described these methods as equivalent (at least in theory) to the standard RLHF procedure (Rafailov et al., 2024b; Azar et al., 2024). However, recent (mostly empirical) work has contradicted this perspective: Tang et al. (2024) find that online methods out-perform offline methods and attribute this fundamentally to on-policy sampling, Xu et al. (2024) argues that the online RL methods produce an often desirable subset of the possible DPO loss minimizers, and Tajwar et al. (2024) provide empirical support for the claim that online and contrastive training provide orthogonal benefits. However, a rigorous theoretical separation is still lacking in the pre-existing literature, which motivates our key questions:

What is the statistical separation between the online RLHF method and offline contrastive methods? What causes this separation and what does it imply?

To answer these questions, we focus on the coverage of the preference dataset, a key concept that is widely used in RL Kakade & Langford (2002); Bagnell et al. (2003); Song et al. (2022); Xie et al. (2023) for analyzing the impact of offline or exploratory data distributions. Through the lens of coverage of the offline preference dataset, we make the following contributions:

1. We prove that the global coverage condition, the strongest possible coverage condition in RL, is necessary for offline contrastive algorithms like DPO to converge to the optimal policy. In contrast, we identify a weaker local coverage condition that is sufficient for online RLHF algorithms, thus provably separating the two types of algorithms. The separation is due to the difference in reward modeling and on/offline regularization – in short, *there is no free lunch from bypassing explicit reward learning and online rollouts*. As global coverage is an unrealistic condition in practice, our separation result can perhaps explain why RLHF works better than offline methods (Tajwar et al., 2024; Tang et al., 2024; Yuan et al., 2024).

2. Although offline contrastive methods are derived from a reverse-KL objective, we prove that the policies trained via offline methods can still have *infinite* reverse-KL in the partial coverage setting. In contrast, we show that RLHF can always control the reverse KL via directly optimizing reverse KL using online samples. This means that on realistic problems, RLHF has stronger guarantees for remaining close to the reference policy than offline contrastive methods.

3. We propose Hybrid Preference Optimization (HyPO) to address the deficiencies of offline contrastive methods while maintaining some of their computational simplicity. HyPO is a *hybrid RL* algorithm (Song et al., 2022) where offline data is used for the DPO objective while online samples are used to explicitly control the reverse KL divergence to the reference policy. We empirically demonstrate that HyPO outperforms DPO, on the TL;DR summarization task Stiennon et al. (2020) on all metrics including both the GPT4 win-rate and the reverse KL divergence to the reference policy.

4. We provide a coverage-based explanation of why RLHF and offline contrastive methods decrease the probability of preferred responses. In particular, under our function approximation-based global coverage condition, we show that such behavior is actually desirable for DPO and RLHF policies to extrapolate and generalize to optimal actions that do not appear in the dataset. This establishes the importance of function approximation for the success of the algorithms such as DPO.

Take together, our results establish the critical role ***coverage*** plays in terms of convergence properties of preference learning algorithms as well as in the design of new, performant empirical approaches.

2 Preliminaries

Following a wide range of recent works (Rafailov et al., 2024b; Azar et al., 2024), we consider the RLHF problem in the contextual bandit formulation (Langford & Zhang, 2008). This is a reasonable

simplification, as one can consider the generated sequence of tokens as one single action, due to the fact that the states are the generated tokens, and the dynamics are deterministic. We denote the context (prompt) space as $\mathcal{X}$, and the action (response) space as $\mathcal{Y}$. Note that due to the finiteness of the possible tokens, the action space is finite but combinatorially large. We use $\rho \in \Delta(\mathcal{X})$ to denote the distribution of the prompts, and $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$ as policies (LLMs) that map prompts to a distribution of responses. We also consider the reward function class $\mathcal{R} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which assigns a reward to each context-response pair.

We assume access to a reference policy π_{ref}, which is usually referred to as the policy learnt using supervised data when training the LLM, that needs to be further fine-tuned to align with human values. An offline preference dataset is collected in the format of $\mathcal{D} = \{x, y^+, y^-\}$ triplets: given context $x \sim \rho$, the preference policy samples two responses $y^1, y^2 \sim \mu(\cdot \mid x)$, where μ is the offline response distribution. Previous works assume either μ to be the same distribution as π_{ref} (Rafailov et al., 2024b) or different offline distribution (Azar et al., 2024; Rosset et al., 2024; Gao et al., 2024). Then, y^1 is labelled as y^+ (thus y^2 as y^-) with probability $p^*(y^1 \succ y^2 \mid x)$, where p^* is defined by the Bradley-Terry model (Bradley & Terry, 1952):

$$p^*(y^1 \succ y^2 \mid x) = \frac{\exp(r^*(x, y^1))}{\exp(r^*(x, y^1)) + \exp(r^*(x, y^2))},$$

where r^* is the human's implicit reward function. Note that this rules out intransitive preferences (Swamy et al., 2024; Munos et al., 2023). Through out the paper we will make the following assumption on the reward function:

Assumption 2.1 (Boundedness of the reward)**.** $\|r^*\|_\infty \le R$.

In many previous works, this formulation has been the canonical way to model the preference data in the RLHF literature (Christiano et al., 2017; Rafailov et al., 2024b; Azar et al., 2024). The goal is to learn a policy π to maximize the objective $J(\pi)$, where

$$J(\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x)}[r^*(x, y)] - \beta\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(x)), \tag{1}$$

i.e., we want to both maximize the human implicit reward, and not deviate too much from the reference policy. We denote the optimal policy $\pi^* \in \operatorname{argmax}_{\pi\in\Pi} J(\pi)$. Here we call $\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(x))$ reverse KL because π – the policy to be optimized, appears first. We will call $\mathsf{KL}(\pi_{\mathsf{ref}}(x)||\pi(\cdot \mid x))$ forward KL. By the definition of KL, we have

$$\text{Definition of reverse KL:} \qquad \mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) := \mathbb{E}_{y\sim\pi(x)} \ln(\pi(y|x)/\pi_{\mathsf{ref}}(y|x)). \tag{2}$$

Note that the expectation in reverse KL is under π (highlighted by red in Eq. 2), indicating that evaluating and optimizing reverse KL requires drawing *online samples* from π. In contrast, evaluating forward KL only requires *offline samples* drawn from π_{ref}. As we will show, this key difference between reverse KL and forward KL plays an important role of separating online RLHF and offline contrastive methods such as DPO. In this paper, we consider two types of algorithms: online RL-based algorithms, and offline contrastive-based algorithms.

Online RLHF Algorithms. We consider algorithms such as Christiano et al. (2017); Ahmadian et al. (2024) as the online RL based methods. We abstract these algorithms as the following procedure: the algorithm performs the following two-stage procedure: one first trains a reward model $\widehat{r}$ that minimizes the Bradley-Terry loss [1]

$$\widehat{r} \in \operatorname*{argmax}_{r\in\mathcal{R}} \widehat{\mathbb{E}}_{x,y^+,y^-\sim\mathcal{D}} \log\left(\frac{\exp(r(x, y^+))}{\exp(r(x, y^+)) + \exp(r(x, y^-))}\right), \tag{3}$$

and perform policy optimization (such as PPO (Schulman et al., 2017)) to optimize the policy optimization problem with the reward model $\widehat{r}$:

$$\pi_{\mathsf{rlhf}} \in \operatorname*{argmax}_{\pi} \widehat{\mathbb{E}}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(\cdot|x)}[\widehat{r}(x, y)] - \beta\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(x)).$$

[1] We use $\widehat{\mathbb{E}}$ to denote the empirical expectation over the dataset.

However, this policy optimization step requires extensive online sampling, and training an additional critic model (e.g., PPO), in addition to the reward model and policy.

Offline Contrastive Algorithms. To circumvent the above-mentioned computational burden, several purely offline contrastive-based methods (i.e., without RL) have been proposed. In this paper, we focus on the following two most representative methods. The first is Direct Preference Optimization (DPO) (Rafailov et al., 2024b), where the objective is $\pi_{\mathsf{dpo}} \in \operatorname{argmax}_\pi \ell_{\mathsf{dpo}}(\pi)$ with

$$\ell_{\mathsf{dpo}}(\pi) = \widehat{\mathbb{E}}_{x,y^+,y^-\sim\mathcal{D}} \log\left(\frac{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right)\right)}{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right)\right) + \exp\left(\beta \log\left(\frac{\pi(y^-|x)}{\pi_{\mathsf{ref}}(y^-|x)}\right)\right)}\right). \tag{4}$$

Another offline contrastive method we will discuss in our paper is Identity Preference Optimization (Azar et al., 2024), but we will defer its technical details to the appendix.

3 Offline Contrastive Methods Require a Stronger Coverage Condition than Online RL Methods

We start by introducing the mathematical formulation of coverage framework. The strongest coverage condition is the following global coverage condition (Munos & Szepesvári, 2008): we say any offline distribution μ covers a policy π if we have $\max_{x,y:\rho(x)>0} \frac{\pi(y|x)}{\mu(y|x)} \le C_{\mathsf{glo}}$. Throughout this section, we will adopt the setting where $\mu = \pi_{\mathsf{ref}}$ (Rafailov et al., 2024b). Formally, we assume the following condition:

Assumption 3.1 (Global Coverage)**.** *For **all** π, we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \le C_{\mathsf{glo}}.$$

For the coverage terms, we always adopt the convention that $\frac{0}{0} = 0$. Note that one sufficient condition for this assumption is that, for any prompt x, and any token sequence y, we have $\pi_{\mathsf{ref}}(y \mid x) \ge 1/C_{\mathsf{glo}}$.

As been recognized in the offline RL literature, global coverage is a strong assumption, and efforts have been made to circumvent this assumption with more relaxed coverage conditions (Uehara & Sun, 2021; Zhan et al., 2022). In this paper, we will consider the following partial coverage assumption that is weaker than Assumption 3.1:

Assumption 3.2 (Local KL-ball Coverage)**.** *For any policy π such that $\mathbb{E}_{x\sim\rho}\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(\cdot \mid x)) \le \varepsilon_{\mathsf{kl}}$, we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \le C_{\varepsilon_{\mathsf{kl}}}.$$

This coverage notion is relatively new in the RL theory literature, but it appears in previous analysis for RLHF algorithms, e,g., Chang et al. (2024). We call this local coverage condition since it only requires π_{ref} to cover the policies that is within some KL-divergence ball centered at π_{ref}. The intuition of this assumption is, for any algorithm that can control the reverse KL of the output policy, we can leverage the coverage condition to relate the error under the output policy to its error under the offline distribution, and thus guarantee its performance. Finally, we note that since the policies with bounded KL is a subset of all policies, for a fixed π_{ref}, we always have $C_{\varepsilon_{\mathsf{kl}}} \le C_{\mathsf{glo}}$.

Taking a closer look at Assumption 3.2, we can see that this assumption is always true: for any policy with $\varepsilon_{\mathsf{kl}} < \infty$, $\max_{x,y:\rho(x)>0} \frac{\pi(y|x)}{\pi_{\mathsf{ref}}(y|x)} < \infty$. However, a simple calculation can show that $\max_{x,y:\rho(x)>0} \frac{\pi(y|x)}{\pi_{\mathsf{ref}}(y|x)}$ can be as large as $\max_{x,y:\pi(y|x)>0} \exp\left(\frac{\varepsilon_{\mathsf{kl}}}{\pi(y|x)}\right)$, even though bounded. This is undesirable because this suggests bounded reverse KL itself is not enough to guarantee optimality: the error can have an *exponential* amplification switching from π to π_{ref}. Thus this motivates Assumption 3.2, which assumes that $C_{\varepsilon_{\mathsf{kl}}}$ is reasonably small, but always bounded in the worst case.

In what follows, we will show that the global coverage assumption (Assumption 3.1) is necessary for offline contrastive-based algorithms such as DPO and IPO, but partial coverage assumption such as Assumption 3.2 is sufficient for online RL based algorithms. This establishes a separation between the two types of algorithms. We emphasize this theoretical separation explains why in practice online methods is less prone to problems such as reward hacking and producing out-of-distribution responses that are due to dataset with insufficient coverage.

3.1 Global Coverage is Necessary for Offline Contrastive Algorithms

Failure of DPO Under Partial Coverage. Now we show that if the strong coverage Assumption 3.1 breaks, then DPO can not guarantee any performance with respect to the objective function Eq. (1). The intuition is based on a rather common observation of the DPO algorithm: the DPO policy π_{dpo} may generate out of distribution responses, while in contrast, RLHF does not generate responses outside of the support of π_{ref} due to online reverse-KL constraint. For example, (Xu et al., 2024) provides a construction where π_{dpo} chooses a response where RLHF policy assigns 0 mass onto, thus proving that RLHF policies are a subset of DPO policies.

However, such construction assumes that the *reward learning* procedure of DPO makes arbitrarily large errors. Also, previous constructions assume deterministic preference, which is only true if the underlying reward function is unbounded. This violates the natural assumption of Assumption 2.1. In the following, we relax these constraints and thus show that DPO fails to guarantee any performance in a rather strong sense. Concretely, DPO constructs the following implicit reward class with the policy class Π: $\mathcal{R}_{\mathsf{dpo}} = \left\{\beta \log\left(\frac{\pi(y|x)}{\pi_{\mathsf{ref}}(y|x)Z(x)}\right) \mid \pi \in \Pi\right\}$, where $Z(x)$ is a partition function that maps context to a real number and is independent of y. Plugging this formulation into the BT loss (Eq. (3)) recovers exactly the DPO loss (Eq. (4)) as the partition functions are canceled. Now we can characterize the returned policy by DPO as exactly whose corresponding reward function is accurate *in distribution*:

Assumption 3.3 (In Distribution Reward Learning)**.** *We assume the DPO policy π_{dpo} satisfies that:*

$$\mathbb{E}_{x,y\sim\rho\circ\pi_{\mathsf{ref}}}\left(\beta\log\left(\frac{\pi_{\mathsf{dpo}}(y\mid x)}{\pi_{\mathsf{ref}}(y\mid x)Z(x)}\right) - r^*(x,y)\right)^2 \le \varepsilon_{\mathsf{dpo}}.$$

Note that this is a rather strong assumption for BT loss – by Lemma C.2, at best one can only hope: for any learned reward function $\widehat{r}$, for each context x, there exists a constant $c(x)$ such that

$$\mathbb{E}_{x,y\sim\rho\circ\pi_{\mathsf{ref}}}(\widehat{r}(x,y) - r^*(x,y) - c(x))^2 \le \varepsilon, \tag{5}$$

i.e., the reward model predicts the human reward up to a gap that is independent of y. This is due to the nature of BT loss only requires the reward function to capture the relative difference, or in the other word, any constant shift (with respect to context) in the reward will be cancelled in the BT loss. However, for the rest of the section, we will make the stronger learning assumption that the gap $c(x) = 0$ (such as in the case of Assumption 3.3). Previous counterexamples analysis violates this assumption, but we will show that even under this assumption, DPO can not guarantee any performance.

Proposition 3.1. *Denote π_{ref} as* any *reference policy such that Assumption 3.1 breaks. Let Π_{dpo} be the set of DPO returned policies such that Assumption 3.3 holds. Then there exists policy $\pi \in \Pi_{\mathsf{dpo}}$ such that $J(\pi) = -\infty$.*

Proof sketch. Without loss of generality, we consider a promptless setting, and assume that the response space is $\mathcal{Y} = \{y_1, y_2, y_3\}$. Again without loss of generality, we assume π_{ref} only covers y_1 and y_2, and thus Assumption 3.1 breaks. We assume partition function $Z = 1$ for all π but we will be rigorous in the formal proof. Then consider the following policy π such that

$$\beta\log\left(\frac{\pi(y_1)}{\pi_{\mathsf{ref}}(y_1)}\right) = r^*(y_1) - \sqrt{\varepsilon_{\mathsf{dpo}}}, \quad \text{and} \quad \beta\log\left(\frac{\pi(y_2)}{\pi_{\mathsf{ref}}(y_2)}\right) = r^*(y_2) - \sqrt{\varepsilon_{\mathsf{dpo}}},$$

One can check π satisfies Assumption 3.3. Now consider the optimal policy $\pi^*(y_i) = \pi_{\mathsf{ref}}(y_i)\exp\left(\frac{1}{\beta}r^*(y_i)\right)$, for $i \in \{1,2\}$, and $\pi^*(y_3) = 0$. Since $\pi^*(y_1) + \pi^*(y_2) = 1$, combining

everything we get $\pi(y_3) > 0$, which implies $\mathsf{KL}(\pi||\pi_{\mathsf{ref}})$ is unbounded, thus we complete the proof. □

One can first relate the above construction to the parital coverage assumption Assumption 3.2: since the policy π considered in the proof has unbounded reverse KL with respect to π_{ref}, thus it is not in the KL-ball of $\varepsilon_{\mathsf{kl}}$ around π_{ref}, which implies that Assumption 3.2 is not sufficient for DPO. Next we show that the global coverage is necessary for the IPO algorithm.

Failure of IPO Under Partial Coverage. To show that the global coverage is necessary for IPO, we can even assume a stronger in-distribution learning guarantee, that is, the returned policy achieves the smallest error on its population loss in distribution.

Proposition 3.2 (Informal)**.** *Denote π_{ref} as* any *reference policy such that Assumption 3.1 breaks. Let Π_{ipo} be the set of IPO returned policies such that it is the minimizer of in-distribution error on its population loss. Then there exists policy $\pi \in \Pi_{\mathit{ipo}}$ such that $J(\pi) = -\infty$.*

We defer the detailed setup and formal version to Appendix E, but the construction for the above proofs share the same intuition: the reverse KL term in the objective function can be unbounded. For offline contrastive-based algorithms, the KL regularization is only enforced under the data distribution, and thus the algorithm can not guarantee bounded reverse KL if the reference policy does not cover the response space well. Although we only showed counterexamples for DPO and IPO, we conjecture that the same intuition holds for other offline contrastive-based algorithms.

3.2 Global Coverage is Sufficient for Offline Contrastive Algorithms

After showing that global coverage is necessary for DPO to guarantee any performance, we now show that it is sufficient for the performance guarantee.

Theorem 3.1. *Let π_{ref} be any reference policy such that Assumption 3.1 holds. For any policy π_{dpo} such that the event in Assumption 3.3 holds, we have that*

$$J(\pi^*) - J(\pi_{\mathsf{dpo}}) = O(C_{\mathsf{glo}}\sqrt{\varepsilon_{\mathsf{dpo}}}).$$

Proof. By Lemma C.1, we have

$$\begin{aligned}
J(\pi^*) - J(\pi_{\mathsf{dpo}}) &\le \mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\pi_{\mathsf{dpo}}(\cdot|x)}\left[r^*(x,y^1) - \widehat{r_{\mathsf{dpo}}}(x,y^1) - r^*(x,y^2) + \widehat{r_{\mathsf{dpo}}}(x,y^2)\right] \\
&\le \sqrt{\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\pi_{\mathsf{dpo}}(\cdot|x)}\left[\left(r^*(x,y^1) - \widehat{r_{\mathsf{dpo}}}(x,y^1) - r^*(x,y^2) + \widehat{r_{\mathsf{dpo}}}(x,y^2)\right)^2\right]} \\
&\le \sqrt{C_{\mathsf{glo}}^2\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1,y^2\sim\pi_{\mathsf{ref}}(\cdot|x)}\left[\left(r^*(x,y^1) - \widehat{r_{\mathsf{dpo}}}(x,y^1) - r^*(x,y^2) + \widehat{r_{\mathsf{dpo}}}(x,y^2)\right)^2\right]},
\end{aligned}$$

and we can complete the proof by plugging in the error guarantee from Assumption 3.3. □

Note that as the proof suggests, the result holds with the more general reward learning guarantee as in Lemma C.2 – one only need to be accurate on predicting the relative rewards between response pairs.

3.3 Online RL method Under Partial Coverage

Finally, we contrast the previous negative results in Section 3.1 for offline contrastive-based algorithms to a positive result for online RL-based algorithms, under the partial coverage setting. We will show that in general global coverage is not necessary for RLHF, i.e., it can guarantee performance under partial coverage. In fact, one might still be able to show an impossibility result for RLHF under partial coverage, by reusing the same counterexample as in the previous section (c.r., Proposition 3.1). Concretely, as long as the learned reward $\widehat{r}(y_3) \to \infty$, $\pi_{\mathsf{rlhf}}(y_3)$ will be 1 and thus the reverse KL will be unbounded. However, this is a rather unrealistic scenario, as the construction requires a neural network to output an unbounded value. Thus this motivates the following assumption:

Assumption 3.4. *For any reward model $\widehat{r}$ in the reward model class, we have that $\|\widehat{r}\|_\infty \le R'$.*

At this point, one might argue why a similar assumption is missing for the offline contrastive-based analysis. The reason lies in the different construction of the model class $\widehat{r}$ for those algorithm: for DPO and IPO, the reward model is constructed as $\widehat{r_{\mathsf{dpo}}} = \beta \log\left(\frac{\pi}{\pi_{\mathsf{ref}} \cdot Z}\right)$, and there is no natural function class for π such that Assumption 3.4 holds. In contrast, on-the-fly normalization of rewards is standard in practice, which the policy will always witness bounded rewards (Gao et al., 2024; Chang et al., 2024; 2023; Ahmadian et al., 2024). As we will show in the following, the difference in the reward function (which is tied to the offline vs. online nature of the algorithms) can explain the different coverage requirement of the algorithms.

To relate to Assumption 3.2, we first show that the reverse KL divergence of the RLHF policy is always bounded under Assumption 3.4.

Lemma 3.1. *Suppose that Assumption 3.4 holds. Then for any RLHF policy π_{rlhf}, we have that*

$$\mathsf{KL}(\pi_{\mathsf{rlhf}} || \pi_{\mathsf{ref}}) := \mathbb{E}_{x \sim \rho} \mathbb{E}_{y \sim \pi_{\mathsf{rlhf}}(\cdot|x)} \left[\log \left(\frac{\pi_{\mathsf{rlhf}}(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \right) \right] \leq \frac{2R'}{\beta}.$$

Then we can show that the RLHF algorithm can guarantee performance under partial coverage:

Theorem 3.2. *Suppose that Assumption 3.4 holds. Then for any reference policy π_{ref} for which Assumption 3.2 holds with $\varepsilon_{\mathsf{kl}} = \frac{2R'}{\beta}$, and any RLHF policy π_{rlhf} with $\widehat{r}$ such that (c.r. Assumption 3.3) $\mathbb{E}_{x,y \sim \rho \circ \pi_{\mathsf{ref}}} \left[\left(r^*(x,y) - \widehat{r}(x,y) \right)^2 \right] \leq \varepsilon_{\mathsf{reward}}$, we have*

$$J(\pi^*) - J(\pi_{\mathsf{rlhf}}) \leq O(C_{\varepsilon_{\mathsf{kl}}} \sqrt{\varepsilon_{\mathsf{reward}}}).$$

Conditioned on Lemma 3.1, the proof of this theorem is similar to that of Theorem 3.1 so we defer it to Appendix E. Similar to Theorem 3.1, we note that Theorem 3.2 holds under a weaker reward learning guarantee as in Lemma C.2. We also remark that as long as $\varepsilon_{\mathsf{kl}}$ is finite, $C_{\varepsilon_{\mathsf{kl}}}$ is finite, so the bound is never vacuous. *Since $C_{\varepsilon_{\mathsf{kl}}} \leq C_{\mathsf{glo}}$ for all $\varepsilon_{\mathsf{kl}}$, it indicates the regret bound of RLHF is never worse and can be much better than the regret bound of DPO.* Combining Theorem 3.1 and Theorem 3.2, we complete the separation result between offline contrastive methods and online RL methods.

A natural question at this point could be: can we further relax the local KL-ball coverage condition in Assumption 3.2 to a single-policy coverage condition, i.e., just assuming $\max_{x,y} \pi^*(y|x)/\pi_{\mathsf{ref}}(y|x) \leq C$? Prior work Zhan et al. (2023) shows that with explicit pessimism, it is possible. However, using pessimism makes the algorithm from Zhan et al. (2023) not computationally tractable and hard to scale to LLM experiments. Our conjecture is that for the RLHF policy π_{rlhf}, it is not possible to achieve meaningful regret under the single policy coverage condition, due to KL not being strong enough to induce pessimism (i.e., bounded KL between π and π_{ref} can still imply exponentially large density ratio π/π_{ref}). Developing a lower bound for π_{rlhf} under single policy coverage in this case can be an interesting future work.

4 Hybrid Preference Optimization: Regularizing Offline Learning with Online Samples

In this section, we will provide a practical algorithm that bridges the gap between the offline contrastive-based algorithms and the online RL-based algorithms. As we see in the previous sections, the difference between the two types of algorithms is their reward model parametrization, and whether to perform online rollouts. In the following we will show that these two properties are in fact tightly intervened with each other.

Here we will focus on the DPO algorithm. One way to fix the issue of the unbounded reward model class for DPO is to consider the following ideal procedure: at the beginning of the algorithm, we first go through the policy class Π, and then we filter out all the policies such that $\mathsf{KL}(\pi || \pi_{\mathsf{ref}}) \geq \frac{2R'}{\beta}$, where R' is the boundedness of the reward function class for RLHF. Now applying the same analysis of Theorem 3.2, we can show that this revised DPO algorithm can guarantee performance under

Table 1: Results on TL;DR dataset. Winrate is evaluated by GPT4 and RM score is from the trained reward model. Experiments are repeated for 3 random seeds. Mean and standard deviation are reported.

Algorithm	Winrate (↑)	RM score (↑)	$\mathsf{KL}(\pi\|\|\pi_{\mathsf{ref}})(\downarrow)$
DPO	42.17% (2.5%)	0.16 (0.05)	44.90 (1.29)
HyPO	46.17% (0.17%)	0.56 (0.03)	25.23 (0.55)

the partial coverage assumption, because now the DPO implicit reward function is bounded by R', recovering Assumption 3.4. We defer the detailed statement and analysis to Appendix F.1.

However, such filtering procedure is not possible in practice, but we can instead consider the following constrained optimization problem: we call the definition of DPO loss in Eq. (4), we want to solve

$$\max_{\pi} \ell_{\mathsf{dpo}}(\pi) \quad \text{s.t.} \quad \mathsf{KL}(\pi||\pi_{\mathsf{ref}}) \leq \frac{2R'}{\beta}, \tag{6}$$

using the KKT conditions, we can show that the following Lagrangian form is equivalent to Eq. (6):

$$\max_{\pi} \ell_{\mathsf{dpo}}(\pi) - \lambda \mathsf{KL}(\pi||\pi_{\mathsf{ref}}), \tag{7}$$

where λ is the Lagrange multiplier. However, in reality, since we do not know the exact value of R', we can consider setting λ to be a hyperparameter. We present the pseudocode in Algorithm 1. Note that due to the reverse KL term, the Hybrid Preference Optimization (HyPO) algorithm optimizes Eq. (7) via both offline and online samples where the offline samples are used for constructing and optimizing ℓ_{dpo} (here σ denotes the sigmoid function), and the online samples $y \sim \pi(x)$ are for KL. Note that regularizing with reverse KL via online samples is widely used in online RLHF (e.g., PPO (Stiennon et al., 2020), APA (Zhu et al., 2023), REBEL (Gao et al., 2024)). Here sg refers to the stop gradient operation, which is a common practice in estimating KL in the LLM fine-tuning setting (Ouyang et al., 2022; von Werra et al., 2020).

Experimental Results. We perform experiments on TL;DR dataset (Stiennon et al., 2020). Our experiment setup mostly follows (Gao et al., 2024): we use a maximum context length of 512 and the maximum generation length of 53. We use Pythia 1.4B (Biderman et al., 2023) as the pre-trained model. For the supervised fine-tuning (SFT) model, we train it over 1 epoch of the dataset with human reference responses as labels. We train the reward model on top of the SFT over 1 epoch of preference data. Both HyPO and DPO are trained over 1 epoch of preference data with Low-rank Adaptation (LoRA) (Hu et al., 2021). We defer more experiment details in Appendix F.

We summarize the results in Table 1: HyPO outperforms DPO in all metrics, including GPT4 win-rage, reward model (RM) evaluation, and KL. However, compared with PPO (e.g., Table 1 in Gao et al. (2024)), HyPO is still lower in winrate and RM evaluation. However, we do preserve most of the benefit of DPO: we avoid training additional reward and critic models, and although we need to perform online generation, we only need to train for 1 epoch while PPO requires 4 epochs of online generation.

Discussion. There are a few limitations of our work: 1) our theoretical analysis only considers the statistical perspective of each algorithm, but we believe our result is complementary to the other work that considers the optimization perspectives (Tajwar et al., 2024). 2) we only conduct experiments on limited models and benchmarks. 3) The experiment result shows that HyPO is still a gap compared to the online RL method: this might suggest that our theory does not fully explain the benefit of all the component of online RL method. For example, one hypothesis is that the learn reward function may have better generalization ability. 4) It is not clear that the KL-ball coverage is necessary for online RL-based methods. However, as we discussed, since a bounded reverse KL might still induce exponentially error amplification, we conjecture that at least a single policy coverage Zhan et al. (2022) is not sufficient for online RL-based methods. We believe these limitations lead to several interesting further directions. Finally, our method may not explicitly address the potential hallucinations or toxic behavior of LLMs, which is a common shortcoming of general-purpose fine-tuning algorithms.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

James Bagnell, Sham M Kakade, Jeff Schneider, and Andrew Ng. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Jonathan D Chang, Kiante Brantley, Rajkumar Ramamurthy, Dipendra Misra, and Wen Sun. Learning to generate better than your llm. *arXiv preprint arXiv:2306.11816*, 2023.

Jonathan D Chang, Wenhao Shan, Owen Oertell, Kianté Brantley, Dipendra Misra, Jason D Lee, and Wen Sun. Dataset reset policy optimization for rlhf. *arXiv preprint arXiv:2404.08495*, 2024.

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

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Zhaolin Gao, Jonathan D Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J Andrew Bagnell, Jason D Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards. *arXiv preprint arXiv:2404.16767*, 2024.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback. 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824, 2008.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

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

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to $q*$: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024a.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024b.

Juntao Ren, Gokul Swamy, Zhiwei Steven Wu, J Andrew Bagnell, and Sanjiban Choudhury. Hybrid inverse reinforcement learning. *arXiv preprint arXiv:2402.08848*, 2024.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

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

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of ai feedback for aligning large language models. *arXiv preprint arXiv:2402.12366*, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

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

Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pp. 33299–33318. PMLR, 2023.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Banghua Zhu, Hiteshi Sharma, Felipe Vieira Frujeri, Shi Dong, Chenguang Zhu, Michael I Jordan, and Jiantao Jiao. Fine-tuning language models with advantage-induced policy alignment. *arXiv preprint arXiv:2306.02231*, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Algorithm 1 Hybrid Preference Optimization (HyPO)

require Pretrained LLM π_{θ_0}, reference policy π_{ref}, offline data $\mathcal{D}$, learning rate α, KL coefficient λ.

1: **for** $t = 1, \ldots, T$ **do**

2: Sample a minibatch of offline data $D_{\mathsf{off}} := \{x, y^+, y^-\} \sim \mathcal{D}$.

3: Compute DPO loss $l_{\mathsf{dpo}} := \sum_{x,y^+,y^- \in D_{\mathsf{off}}} \log\left(\sigma\left(\beta \log\left(\frac{\pi_{\theta_{t-1}}(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right) - \beta \log\left(\frac{\pi_{\theta_{t-1}}(y^-|x)}{\pi_{\mathsf{ref}}(y^-|x)}\right)\right)\right)$.

4: Sample online data $D_{\mathsf{on}} := \{x, y\}$ where $x \sim \mathcal{D}, y \sim \pi_{\theta_{t-1}}(x)$.

5: Compute $l_{\mathsf{kl}} := \sum_{x,y \in D_{\mathsf{on}}} \log(\pi_{\theta_{t-1}}(y|x)) \cdot \mathrm{sg}\left(\log\left(\frac{(\pi_{\theta_{t-1}}(y|x))}{(\pi_{\mathsf{ref}}(y|x))}\right)\right)$.

6: Update $\theta_t = \theta_{t-1} - \alpha \cdot \nabla_{\theta_{t-1}}(l_{\mathsf{dpo}} - \lambda l_{\mathsf{kl}})$.

return π_T.

A Related Work

Preference Fine-Tuning. As discussed in the introduction of our work, there are two major paradigms for preference fine-tuning of LLMs. The first one, online RL methods Ouyang et al. (2022), proposes to first train a reward model (classifier) to predict human preferences, followed by running an RL method to optimize this learned reward function. While PPO Schulman et al. (2017) is the most popular RL algorithm used in the online RLHF framework by far Ouyang et al. (2022); Stiennon et al. (2020); Touvron et al. (2023), more recent work by Ahmadian et al. (2024) shows that simpler online RL algorithms like REINFORCE Williams (1992) also work well. The second class of methods, offline contrastive techniques (Rafailov et al., 2024b; Zhao et al., 2023; Azar et al., 2024), avoid explicit reward modeling and directly optimize their objective on the offline preference dataset. Recently there are *hybrid* methods that combine offline preference data with online preference labels (Guo et al., 2024; Rosset et al., 2024; Azar et al., 2024) – we leave extending our analysis to this setting to future work. Throughout our paper, we assume for simplicity of analysis that preferences are generated by an underlying utility function and therefore contain no intransitivities (Swamy et al., 2024; Munos et al., 2023). Future work could also explore the effect of using more efficient *local exploration-based* RLHF algorithms (Chang et al., 2023; 2024; Swamy et al., 2023; Ren et al., 2024).

Understanding PFT. Prior work has studied different parts of the standard RLHF recipe (Gao et al., 2023; Kirk et al., 2023; Singhal et al., 2023; Eisenstein et al., 2023) and the impact of preference data quality (Sharma et al., 2024). In our work, we instead take a converge-based perspective on the relationship between online RL methods and offline contrastive methods. Although derived from the same minimum relative entropy objective (Ziebart et al., 2008) and perceived as equivalent by some early work (Rafailov et al., 2024b; Azar et al., 2024), more recent work has started to unravel the distinctions between these two classes of methods. Tang et al. (2024) repeatedly observe better performance from online rather than offline methods and after rigorously validating a variety of hypotheses, conclude that on-policy sampling is indispensable for ensuring a high quality policy. Tajwar et al. (2024) perform an in-depth study of the effects of preference data, contrastive losses, and on-policy sampling and conclude that a combination of contrastive losses and interactive training is most preferable in practice. (Xu et al., 2024) also observe better performance from online PPO than from offline DPO and argue this is because the former is able to eliminate a larger set of policies that are undesirable from the perspective of the rater. We supplement these mostly empirical observations with a rigorous theoretical explanation for the observed behavior through the lens of dataset coverage, as well as designing an algorithm that addresses the key weaknesses of offline contrastive approaches.

Recent work Yuan et al. (2024); Pal et al. (2024); Rafailov et al. (2024a) has observed an interesting effect of the DPO procedure: a simultaneously decreases in the likelihood of both preferred and rejected responses. This behavior is surprising at the first glance because one would expect that DPO will increase the likelihood of preferred responses and decrease the likelihood of rejected responses. We provide a rigorous statistical explanation of this behavior and show that this behavior is natural when the offline preference data only contains sub-optimal responses but the function approximation allows DPO to extrapolate and generalize to the correct optimal responses. This highlights the role of function approximation in the success of offline contrastive based methods.

Coverage. We analyze online RLHF and offline contrastive-based methods via the concept of *coverage*. Coverage measures how well an offline (data) distribution covers the support of the policy of interest, which has been the key technical tool in offline RL (Munos & Szepesvári, 2008; Uehara & Sun, 2021; Zhan et al., 2022), offline-online (I)RL (Xie et al., 2021; Song et al., 2022; Ren et al., 2024) and online RL (Bagnell et al., 2003; Kakade & Langford, 2002). The data coverage plays an important role in our analysis since both online RLHF and offline contrastive-based methods rely on an offline preference dataset for learning.

B Function Approximation Coverage: Can Fine-tuned Policies Extrapolate?

Our final result is a theoretical explanation of the extrapolation behavior of preference fine-tuning algorithms under the global coverage assumption in the function approximation setting. The extrapolation behavior refers to the phenomenon that policies assign decreasing likelihood to the preferred responses, even to the preferred samples during the training, and instead increase the likelihood outside the preference distribution data (Pal et al., 2024).

A previous attempt (Rafailov et al., 2024a) to explain this behavior is based on the assumption that the responses from the reference policy have the same distribution as the *preferred* responses from the dataset, i.e., $y^+ \sim \mu \overset{d}{=} y \sim \pi_{\mathsf{ref}}$. However, as mentioned in Section 2, more realistically, one can assume that $y \sim \mu \overset{d}{=} y \sim \pi_{\mathsf{ref}}$ since it is implied by using the reference policy to generate the dataset, including the not preferred responses; or even more generally by considering $\mathsf{supp}(\mathcal{D}) \subset \mathsf{supp}(\pi_{\mathsf{ref}})$. The latter is common in practice, for example, the dataset is precollected, or the reference policy might place a small mass on responses so they are not sampled during the data collection process.

In the following example, we consider the linear function approximation setting and an offline dataset that does not contain the optimal action. We show that DPO can correctly increase the model's likelihood of the optimal action by decreasing the likelihood of both the preferred and rejected actions from the offline data.

Example B.1. *Consider a promptless setting, where the response space is $\mathcal{Y} = \{y_1, y_2, y_3\}$. Consider the linear function approximation setting with feature map ϕ, where $\phi(y_1) = [1, 0], \phi(y_2) = [1/2, 1/2], \phi(y_3) = [0, 1]$. Suppose all policies are parametrized as softmax linear policies, i.e., $\pi(y) \propto \exp(w_\pi^\top \phi(y))$. Let $w_{\mathsf{ref}} = [1, 1]$, then we have $\pi_{\mathsf{ref}}(y_i) = 1/3, \forall i \in \{1, 2, 3\}$.*

Consider the ground truth reward function $r^*(y) = [10, 1]^\top \phi(y)$, and suppose $\mathsf{supp}(\mu) = \{y_1, y_2\}$, i.e., the data only covers y_1 and y_2. And as always, the preference is based on the ground truth reward function under the Bradley-Terry model.

We can first check that the data distribution indeed has global coverage in the linear function approximation case (Xiong et al., 2022), i.e., let $\Sigma_\mu = \mathbb{E}_{y\sim\mu}\phi(y)\phi(y)^\top$, then for all π,

$$\mathbb{E}_{y\sim\pi}\|\phi(y)\|^2_{\Sigma_\mu^{-1}} \le C_\pi.$$

If we parameterize $\widehat{r}(y) = \widehat{w}^\top \phi(y)$ (or in case of DPO, we can still check and see that $\widehat{r_{\mathsf{dpo}}}(y) = \widehat{w_{\mathsf{dpo}}}^\top \phi(y)$ because of the softmax linear parametrization of the policies), for either direct reward learning or DPO, we can have the learned reward function $\widehat{r}(y) = [10, 1]^\top \phi(y) + c$, where c is the constant reward shift (c.r. Eq. (5)). Then a simple calculation (by $\pi(y) \propto \pi_{\mathsf{ref}}(y) \exp(\widehat{r}(y)/\beta)$) shows that, as long as c is small enough, the policies will decrease the likelihood of y_1 and y_2 and increase the likelihood of y_3. ◁

B.1 Synthetic experiment for extrapolation

To validate our theory result, in this section we perform a synthetic experiment on global coverage with linear function approximation. As shown in Figure 1, the extrapolation behavior is observed in both online RL method and DPO. In addition, we show that without the linear function approximation, i.e., when each action is treated independently, DPO can erroneously assign a higher probability to unseen

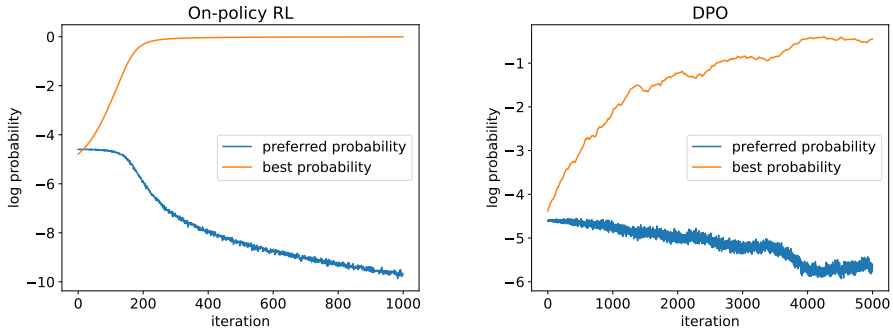

Figure 1: Extrapolation behavior of Online RL method and DPO under linear function approximation. We plot the mean log probability of the preferred responses and the log probability of the best response, which is unseen in the training data. We see that both algorithms correctly assigns increasing probability to the best response.

suboptimal responses, indicating DPO can fail to extrapolate and generalize. Our investigation identifies that function approximation plays an important role in the success of RLHF and DPO algorithms in terms of generalizing to optimal actions beyond the offline data.

B.2 Extrapolation with function approximation

We first describe our experiment setup. We consider linear function approximation setting where we have 100 responses ($|\mathcal{Y}| = 100$). We consider a 16-dimensional feature vector $\phi : \mathcal{Y} \to \mathbb{R}^{16}$, and we generate $\phi(y)$ by simply sampling 99 random 16-dimensional vectors where the ℓ_1 norm of each vector is 1. We add one final $\phi(y) = [1, 0, 0, \dots]$.

We construct the implicit human reward $r^*(y) = {w^*}^\top \phi(y)$, where $w^* = [5, ...]$, and the rest of the entries are sampled from Unif(-2,2).

We parametrize the policies as softmax linear policies, i.e., we parametrize each policy π with $w^\pi \in \mathbb{R}^{16}$ such that $\pi(y) = \frac{{w^\pi}^\top \phi(y)}{\sum_{y \in \mathcal{Y}} {w^\pi}^\top \phi(y)}$. One can check in this formulation the implicit reward in DPO ($\widehat{r_{\mathsf{dpo}}}$) is linear in ϕ.

We generate 10000 preference pairs, according to the BT model under r^*, for the first 50 responses. We checked that the first responses indeed span $\mathbb{R}^{16}$. Thus the offline data has global coverage in linear function approximation setting.

For on-policy RL methods, we first train a reward model. Then we simply perform gradient descent on the KL-regularized bandit loss (we assume π_{ref} is uniform). For DPO, we simply perform SGD on the offline preference dataset. We track two qualities over the training: the mean log probability of a random subset of preferred responses, and the log probability of best response $\phi(y) = [1, 0, 0, \dots]$. We plot the results in Figure 1. We observe that both methods have the extrapolation behavior – the probability of preferred responses decays but the probability of the optimal response goes up.

B.3 Extrapolation without function approximation

Now we describe the setting where function approximation fails, and this reduces to a Multi-arm bandit setting. We set $|\mathcal{Y}| = 500$, and the offline data only covers the first half of the responses. The $r^*(y)$ is set by sampling from Unif(-10,10), and we generate 10000 offline samples by uniformly sample pairs of responses from the first half of the response space, and then label them with BT model under r^*. We train DPO with 5000 iterations, and plot the mean probability of the responses *outside* of the data support in Figure 2: we observe that the mean probability of the out-of-distribution

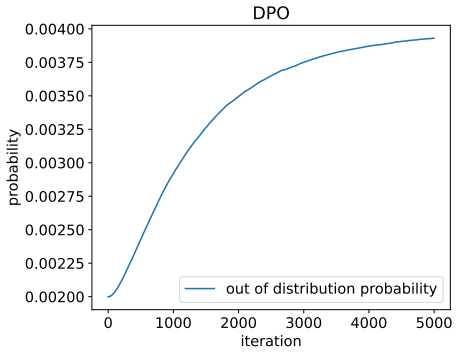

Figure 2: Extrapolation behavior of DPO without function approximation. We plot the average probability of out-of-distribution responses along the training and DPO assigns increasing probability to out-of-distribution responses.

responses are increasing, however, this could be an undesirable behavior because the reward of the out-of-distribution responses could be arbitrarily bad.

C Auxiliary Lemmas

Lemma C.1 (Objective decomposition)**.** *Let $J(\pi)$ be the objective function defined in (1), and for reward function $\hat{r}$, we let*

$$\hat{\pi} \in \operatorname*{argmax}_{\pi} \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(x)), \tag{8}$$

then we have

$$J(\pi^*) - J(\hat{\pi}) \le \mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\hat{\pi}(\cdot|x)}\big[r^*(x,y^1) - \hat{r}(x,y^1) - r^*(x,y^2) + \hat{r}(x,y^2)\big].$$

Proof. We have

$$\begin{aligned}
&J(\pi^*) - J(\hat{\pi})\\
=&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[r^*(x,y)] - \beta\mathsf{KL}(\pi^*(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) - \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] + \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x))\\
=&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[r^*(x,y)] - \beta\mathsf{KL}(\pi^*(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) - \big(\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x))\big)\\
&+ \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) - \big(\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x))\big)\\
\le&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[r^*(x,y)] - \beta\mathsf{KL}(\pi^*(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) - \big(\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\pi^*(\cdot \mid x)||\pi_{\mathsf{ref}}(x))\big)\\
&+ \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x)) - \big(\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[\hat{r}(x,y)] - \beta\mathsf{KL}(\hat{\pi}(\cdot \mid x)||\pi_{\mathsf{ref}}(x))\big)\\
=&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[r^*(x,y) - \hat{r}(x,y)] - \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r^*(x,y) - \hat{r}(x,y)],
\end{aligned}$$

where the inequality is due to Eq. (8). To complete the proof, note that

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x)}[r^*(x,y) - \hat{r}(x,y)] - \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\hat{\pi}(\cdot|x)}[r^*(x,y) - \hat{r}(x,y)]\\
=&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\hat{\pi}(\cdot|x)}[r^*(x,y^1) - \hat{r}(x,y^1)] - \mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\hat{\pi}(\cdot|x)}[r^*(x,y^2) - \hat{r}(x,y^2)]\\
=&\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\hat{\pi}(\cdot|x)}\big[r^*(x,y^1) - \hat{r}(x,y^1) - r^*(x,y^2) + \hat{r}(x,y^2)\big].
\end{aligned}$$

□

Lemma C.2 (Lemma C.2 from (Chang et al., 2024))**.** *Assume that r^* is bounded, let $\mathcal{R}$ be the reward function class, and Let*

$$\hat{r} = \operatorname*{argmin}_{r\in\mathcal{R}} \hat{\mathbb{E}}_{x,y^+,y^-\sim\mathcal{D}} \log\left(\frac{\exp(r(x,y^+))}{\exp(r(x,y^+)) + \exp(r(x,y^-))}\right),$$

then we have with probability at least $1-\delta$ *that*

$$\mathbb{E}_{x,y^1,y^2\sim\mu\circ\pi_{\mathsf{ref}}}\left[\left(r^*(x,y^1)-r^*(x,y^2)-\hat{r}(x,y^1)+\hat{r}(x,y^2)\right)^2\right]\leq\frac{c\kappa^2\log(|\mathcal{R}|/\delta)}{N},$$

where κ *measures the non-linearity of the link function, and* c *is a constant,* $N:=|\mathcal{D}|$ *is the size of the offline dataset.*

D Results for IPO

In this section we give detailed technical details for IPO, and the negative results for IPO under partial coverage. Recall that the empirical objective of IPO is is $\pi_{\mathsf{ipo}}\in\operatorname{argmin}_\pi\widehat{\ell_{\mathsf{ipo}}}(\pi)$, where

$$\widehat{\ell_{\mathsf{ipo}}}(\pi)=\widehat{\mathbb{E}}_{x,y^+,y^-\sim\mathcal{D}}\left[\left(\log\left(\frac{\pi(y^+\mid x)\pi_{\mathsf{ref}}(y^-\mid x)}{\pi(y^-\mid x)\pi_{\mathsf{ref}}(y^+\mid x)}\right)-\frac{\beta^{-1}}{2}\right)^2\right].$$

The empirical objective is derived from the following population loss

$$\ell_{\mathsf{ipo}}(\pi)=\mathbb{E}_{x,y^1,y^2\sim\rho\circ\pi_{\mathsf{ref}}}\left[\left(h_\pi(y^1,y^2)-I(y^1,y^2)/\beta\right)^2\right],\tag{9}$$

where

$$h_\pi(y^1,y^2)=\log\left(\frac{\pi(y^1)\pi_{\mathsf{ref}}(y_2)}{\pi(y^2)\pi_{\mathsf{ref}}(y_1)}\right),$$

and $I(y^1,y^2)$ is a Bernoulli random variable with parameter $p=p^*(y_1\succ y_2)$, where here p^* can be any underlying human preference (that is not necessarily parametrized by the Bradley Terry model). To show the negative result, we can make the following learning assumption:

Assumption D.1 (In distribution guarantee for IPO)**.** *We assume that the returned policy* π_{ipo} *satisfies that*

$$\pi_{\mathsf{ipo}}=\operatorname*{argmin}_{\pi\in\Pi}\ell_{\mathsf{ipo}}(\pi),$$

i.e., the returned policy π_{ipo} *induces the smallest possible in-distribution error on its population loss.*

With the setup, we can state and prove the formal version of the result:

Proposition D.1 (Formal version of of Proposition 3.2)**.** *Denote* π_{ref} *as* any *reference policy such that Assumption 3.1 breaks. Let* Π_{ipo} *be the set of IPO returned policies such that Assumption D.1 holds. Then there exists policy* $\pi\in\Pi_{\mathit{ipo}}$ *such that* $J(\pi)=-\infty$*.*

Proof. Without loss of generality, we consider a promptless setting, and assume that the response space is $\mathcal{Y}=\{y_1,y_2,y_3\}$. Again without loss of generality, we assume π_{ref} only covers y_1 and y_2, and thus Assumption 3.1 breaks. Specifically, let $\pi_{\mathsf{ref}}(y_1)=\pi_{\mathsf{ref}}(y_2)=1/2$. Then we have

$$\pi_{\mathsf{ipo}}=\operatorname*{argmin}_{\pi\in\Pi}\mathbb{E}_{y^1,y^2\sim\pi_{\mathsf{ref}}}\left[\left(\log\left(\frac{\pi(y^1)}{\pi(y^2)}\right)-I(y^1,y^2)/\beta\right)^2\right],$$

which gives

$$\log\left(\frac{\pi_{\mathsf{ipo}}(y_1)}{\pi_{\mathsf{ipo}}(y_2)}\right)=p^*(y_1\succ y_2)/\beta,$$

and thus we have the relation that

$$\pi_{\mathsf{ipo}}(y_1)=\pi_{\mathsf{ipo}}(y_2)\cdot\exp(p^*(y_1\succ y_2)/\beta).$$

Let $\pi_{\mathsf{ipo}}(y_2)=\alpha\in(0,1]$, then for any α such that $\pi_{\mathsf{ipo}}(y_3)=1-(1+\exp(p^*(y_1\succ y_2)/\beta))\alpha>0$, we will have that $\mathsf{KL}(\pi_{\mathsf{ipo}}||\pi_{\mathsf{ref}})$ is unbounded, and thus we complete the proof. □

E Missing Proofs

E.1 Proof of Proposition 3.1

Proposition E.1 (Restatement of Proposition 3.1)**.** *Denote* π_{ref} *as* any *reference policy such that Assumption 3.1 breaks. Let* Π_{dpo} *be the set of DPO returned policies such that Assumption 3.3 holds. Then there exists policy* $\pi \in \Pi_{dpo}$ *such that* $J(\pi) = -\infty$*.*

Proof. Again as in the proof sketch, without loss of generality, we consider a promptless setting, and assume that the response space is $\mathcal{Y} = \{y_1, y_2, y_3\}$. Again without loss of generality, we assume π_{ref} only covers y_1 and y_2, and thus Assumption 3.1 breaks. Now consider the optimal policy

$$\pi^*(y) = \frac{\pi_{\mathsf{ref}}(y)\exp(r^*(y)/\beta)}{Z^*(t)}, \forall y \in \mathcal{Y},$$

where $Z^* = \sum_{y\in\mathcal{Y}} \pi_{\mathsf{ref}}(y)\exp(r^*(y)/\beta)$, note that by construction $\pi^*(y_3) = 0$.

Then consider the following policy π such that

$$\beta\log\left(\frac{\pi(y_1)}{\pi_{\mathsf{ref}}(y_1)\cdot Z^*}\right) = r^*(y_1) - \sqrt{\varepsilon_{\mathsf{dpo}}}, \quad \text{and} \quad \beta\log\left(\frac{\pi(y_2)}{\pi_{\mathsf{ref}}(y_2)\cdot Z^*}\right) = r^*(y_2) - \sqrt{\varepsilon_{\mathsf{dpo}}},$$

Then we have

$$\mathbb{E}_{y\sim\pi_{\mathsf{ref}}}\left(\beta\log\left(\frac{\pi_{\mathsf{dpo}}(y)}{\pi_{\mathsf{ref}}(y)\cdot Z^*}\right) - r^*(x,y)\right)^2 = \varepsilon_{\mathsf{dpo}},$$

thus π satisfies Assumption 3.3. Rearranging we can see that $\pi(y_1) < \pi^*(y_1)$ and $\pi(y_2) < \pi^*(y_2)$.

Now since $\pi^* = 0$, we have

$$\pi^*(y_1) + \pi^*(y_2) = 1,$$

and combine we get $\pi(y_3) > 0$, which implies $\mathsf{KL}(\pi||\pi_{\mathsf{ref}})$ is unbounded, since $\pi_{\mathsf{ref}}(y_3) = 0$. □

E.2 Proof of Theorem 3.2

In this section we prove Theorem 3.2:

Theorem E.1 (Restatement of Theorem 3.2)**.** *Suppose that Assumption 3.4 holds. Then for any reference policy* π_{ref} *such that Assumption 3.2 holds with* $\varepsilon_{\mathsf{kl}} = \frac{2R'}{\beta}$*, for any RLHF policy* π_{rlhf} *with* $\widehat{r}$ *such that (c.r. Assumption 3.3),*

$$\mathbb{E}_{x,y\sim\rho\circ\pi_{\mathsf{ref}}}\left[(r^*(x,y) - \widehat{r}(x,y))^2\right] \le \varepsilon_{\mathsf{reward}},$$

or more generally, the event in Lemma C.2 holds for $\widehat{r}$*, we have*

$$J(\pi^*) - J(\pi_{\mathsf{rlhf}}) \le O(C_{\varepsilon_{\mathsf{kl}}}\sqrt{\varepsilon_{\mathsf{reward}}}).$$

To prove this we first prove the following lemma so we can leverage Assumption 3.2:

Lemma E.1 (Restatement of Lemma 3.1)**.** *Suppose that Assumption 3.4 holds. Then for any RLHF policy* π_{rlhf}*, we have that*

$$\mathsf{KL}(\pi_{\mathsf{rlhf}}||\pi_{\mathsf{ref}}) := \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\log\left(\frac{\pi_{\mathsf{rlhf}}(y\mid x)}{\pi_{\mathsf{ref}}(y\mid x)}\right)\right] \le \frac{2R'}{\beta}.$$

Proof. since we have that $\pi_{\mathsf{rlhf}}(y \mid x) = \frac{\pi_{\mathsf{ref}}(y|x)\exp(\hat{r}(x,y)/\beta)}{Z(x)}$ *for all* $x \in \mathsf{supp}(\rho), y \in \mathcal{Y}$, we have

$$\mathsf{KL}(\pi_{\mathsf{rlhf}}||\pi_{\mathsf{ref}}) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\log\left(\frac{\exp(\hat{r}(x,y))}{\beta Z(x)}\right)\right] = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\frac{\hat{r}(x,y)}{\beta} - \log(Z(x))\right].$$

Plugging in the definition of $Z(x)$ we get

$$\log(Z(x)) = \log\left(\mathbb{E}_{y\sim\pi_{\mathsf{ref}}(\cdot|x)}\left[\exp\left(\frac{\hat{r}(x,y)}{\beta}\right)\right]\right) \geq \mathbb{E}_{y\sim\pi_{\mathsf{ref}}(\cdot|x)}\left[\frac{\hat{r}(x,y)}{\beta}\right]$$

due to Jensen's inequality. Thus we have

$$\mathsf{KL}(\pi_{\mathsf{rlhf}}||\pi_{\mathsf{ref}}) \leq \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\frac{\hat{r}(x,y)}{\beta}\right] - \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\frac{\hat{r}(x,y)}{\beta}\right] \leq \frac{2R'}{\beta}.$$

□

Now with Lemma 3.1, we can prove Theorem 3.2:

Proof. By Lemma C.1, we have

$$\begin{aligned}
&J(\pi^*) - J(\pi_{\mathsf{rlhf}}) \\
&\leq \mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[r^*(x,y^1) - \widehat{r}(x,y^1) - r^*(x,y^2) + \widehat{r}(x,y^2)\right] \\
&\leq \sqrt{\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1\sim\pi^*(\cdot|x),y^2\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\left(r^*(x,y^1) - \widehat{r}(x,y^1) - r^*(x,y^2) + \widehat{r}(x,y^2)\right)^2\right]} \\
&\leq \sqrt{C_{\mathsf{glo}}^2\mathbb{E}_{x\sim\rho}\mathbb{E}_{y^1,y^2\sim\pi_{\mathsf{ref}}(\cdot|x)}\left[\left(r^*(x,y^1) - \widehat{r}(x,y^1) - r^*(x,y^2) + \widehat{r}(x,y^2)\right)^2\right]} \\
&\qquad\qquad \text{(Lemma 3.1 and Assumption 3.2)} \\
&\leq C\sqrt{\varepsilon_{\mathsf{reward}}}. \qquad\qquad \text{(Lemma C.2)}
\end{aligned}$$

□

F Details of Section 4

F.1 Theoretical guarantee

In this section, we consider the constrained optimization version of HyPO (Eq. (6)). Note that the reward function class is identical to DPO, i.e., $\mathcal{R}_{\mathsf{hypo}} = \left\{\beta\log\left(\frac{\pi(y|x)}{\pi_{\mathsf{ref}}(y|x)Z(x)}\right) \mid \pi \in \Pi\right\}$, where $Z(x)$ is the partition function. Then for each output policy π_{hypo}, we can denote its implicit reward function $\widehat{r_{\mathsf{hypo}}}(x,y) := \beta\frac{\pi_{\mathsf{hypo}}(y|x)}{\pi_{\mathsf{ref}}(y|x)\cdot Z(x)}$, and similarly to Theorem 3.2, we can obtain the following guarantee in the partial coverage condition:

Theorem F.1. *For any reference policy* π_{ref} *such that Assumption 3.2 holds with* $\varepsilon_{\mathsf{kl}} = \frac{2R'}{\beta}$*, for any HyPO policy* π_{hypo} *such that the event in Lemma C.2 holds, i.e.,*

$$\mathbb{E}_{x,y^1,y^2\sim\mu\circ\pi_{\mathsf{ref}}}\left[\left(r^*(x,y^1) - r^*(x,y^2) - \widehat{r_{\mathsf{hypo}}}(x,y^1) + \widehat{r_{\mathsf{hypo}}}(x,y^2)\right)^2\right] \leq \varepsilon_{\mathsf{hypo}},$$

we have

$$J(\pi^*) - J(\pi_{\mathsf{hypo}}) \leq O(C_{\varepsilon_{\mathsf{kl}}}\sqrt{\varepsilon_{\mathsf{hypo}}}).$$

The proof is identical to the proof of Theorem 3.2 and thus we omit the proof.

F.2 Experiment details

In this section, we provide more details of our experiment. We use the Pythia 1.4B model (Biderman et al., 2023) with hugging face model card: EleutherAI/pythia-1.4b-deduped. The TL;DR dataset is available at https://github.com/openai/summarize-from-feedback. The human reference dataset contains 117k training, 6.45K validation and 6.55K testing data. The preference dataset contains 92.9K training and 83.8K validation data. The reward evaluation and KL computation is performed on the whole validation data of the reference dataset. The GPT winrate is computed on a subset of 600 samples from the validation data. The GPT API checkpoint we use is gpt-4-0613. We follow the standard prompt for the winrate evaluation (e.g., see Appendix D.3 of Gao et al. (2024)). Below we provide the hyperparameter for HyPO and DPO.

For our experiment, we run on a cluster of mixture of Nvidia A6000 and L40 GPUs with 48 GB VRAM. We use 4 GPUs in parallel for training, and for DPO the experiment time varies from 1 hour to 2 hour to finish, and for HyPO the time varies between 4 hours to 5 hours.

Table 2: RM/SFT hyperparameters.

Learning rate	3e-6
Batch size	64
Learning rate scheduler	cosine
Optimizer	Adamw
LoRA	False

Table 3: DPO hyperparameters.

Learning rate	3e-6
Batch size	64
Learning rate scheduler	cosine
Optimizer	Adamw
β	0.05

Table 4: HyPO hyperparameters.

Learning rate	3e-6
Batch size	64
Learning rate scheduler	cosine
Optimizer	Adamw
β	0.05
λ	0.02

Table 5: Lora configurations.

r	1024
α	2048
Dropout	0