On the Exponential Convergence for Offline RLHF with Pairwise Comparisons

Zhirui Chen[1], **Vincent Y. F. Tan**[1]

[1]National University of Singapore
zhiruichen@u.nus.edu, vtan@nus.edu.sg

Abstract

We consider the problem of offline reinforcement learning from human feedback (RLHF) with pairwise comparisons, where the implicit reward is a linear function of an unknown parameter. Given an offline dataset, our objective is to identify the optimal action for each state, with the ultimate goal of minimizing the simple regret. We propose an algorithm, Reinforcement Learning with Locally Optimal Weights (RL-LOW), which achieves an exponential rate of simple regret that decays exponentially with the ratio of the number of data samples to an instance-dependent hardness parameter. This hardness parameter depends explicitly on the suboptimality gap of each action. Furthermore, we derive the first instance-dependent lower bound for offline RLHF with pairwise comparisons. Interestingly, the lower and upper bounds on the simple regret match in an order-wise sense in the exponent, demonstrating the order-wise optimality of RL-LOW. Motivated by privacy considerations in practical applications, we further extend RL-LOW to the setting of differential privacy and show, somewhat surprisingly, that the hardness parameter remains unchanged in the asymptotic regime as the number of data samples tends to infinity. This result highlights the inherent efficiency of RL-LOW in preserving the privacy of the observed rewards. By establishing instance-dependent bounds with exponential convergence rates, our work fills an important gap in the existing literature, which has primarily focused on worst-case regret bounds with inverse polynomial convergence rates for offline RLHF with pairwise comparisons.

1 Introduction

Reinforcement Learning (RL) (Sutton and Barto 2018) has been widely recognized for its capacity to facilitate agents in learning a sequence of optimal actions through iterative interactions with their environments. However, RL encounters significant hurdles in environments that are characterized by uncertainty or lacking explicit reward signals. To address these shortcomings, the concept of RL with human feedback (or RLHF) has emerged as a prominent paradigm. Preference-based RL (PbRL) (Christiano et al. 2017; Chen et al. 2022; Ibarz et al. 2018; Palan et al. 2019) has stood out as one of the most widely used frameworks for RLHF. In this regard, preference-based RL has achieved remarkable performances in practical applications, with particular importance lying in its ability to align large language models (LLMs) with human intent, thereby mitigating the output of toxic and dishonest information (Ouyang et al. 2022; Ziegler et al. 2019; Glaese et al. 2022; Bai et al. 2022; Liu, Sferrazza, and Abbeel 2023), and improving the quality of applying to the specific tasks (Stiennon et al. 2020; Wu et al. 2021; Nakano et al. 2021).

In this work, we tackle the problem of offline RLHF with pairwise comparisons, wherein the learning mechanism operates solely on pre-existing (or offline) data without dynamically engaging with the environment. Given the high cost associated with human interaction, offline RLHF has assumed particular importance in the context of incorporating human feedback. The significance of this offline framework has been justified by many previous prominent works (Shin, Dragan, and Brown 2023; Ouyang et al. 2022; Zhu, Jordan, and Jiao 2023; Kim et al. 2023). For instance, within the learning process of InstructGPT (Ouyang et al. 2022) or the training procedure of Ahmadian et al. (2024), a pivotal procedure involves the training of a reward model utilizing pre-trained LLM feature vectors, coupled with the utilization of pre-collected human pairwise comparisons as the training dataset. Conceptually, this procedure can be construed as treating the current prompt context as a state within a certain Markov Decision Process (MDP), while the responses generated by the LLM serve as actions within this process. Empirical findings presented by Ouyang et al. (2022) demonstrate the efficacy of this offline framework in effectively aligning human intent with the outputs of LLMs.

However, the literature concerning theoretical analyses within the domain of offline PbRL remains rather scant. Previous theoretical analyses (Zhu, Jordan, and Jiao 2023; Zhan et al. 2024) of offline PbRL predominantly focused on the worst-case (or minimax) regret, often resulting in the derivation of regret upper bounds for their algorithms of the form $\tilde{O}(n^{-1/2})$, where n is the size of the offline dataset. In this work, we adopt a different approach that is centered on instance-dependent guarantees. In other words, we wish to derive performance guarantees that are functions of the specific problem instance, thus elucidating the role of fundamental hardness parameters. This yields complementary insights to the existing worst-case analyses. To this end, we design and analyze RL-LOW, a preference-based RL algorithm. Our analysis of the performance RL-LOW unveils an instance-

dependent simple regret bound of $\exp(-\Omega(n/H))$, where H is a hardness parameter. This reveals that the simple regret decays exponentially fast in the size of the dataset n and the exponential rate of convergence has also been identified. Complementarily, by proving an instance-dependent lower bound, we show that any algorithm will suffer from a simple regret of at least $\exp(-O(n/H))$. Thus, the dependence of the problem on H is fundamental and cannot be improved upon, thereby demonstrating the efficacy of RL-LOW and the tightness of our analyses.

1.1 Related Works

Preference-Based RL: From the empirical viewpoint, Christiano et al. (2017) initially demonstrated that RL systems can effectively address complex tasks like Atari games and simulated robot locomotion by learning from human preferences between trajectory segments. Later, numerous researchers started to employ human pairwise comparisons to enhance the performance of LLMs, e.g., aligning the LLMs' behavior with human intent (Ziegler et al. 2019; Glaese et al. 2022; Bai et al. 2022; Liu, Sferrazza, and Abbeel 2023), and enhancing the efficacy of application to specific tasks (Stiennon et al. 2020; Wu et al. 2021; Nakano et al. 2021).

From the theoretical perspective, the existing literature remains sparse in offline RLHF with pairwise comparisons. Zhu, Jordan, and Jiao (2023) elucidated the failure of the maximum likelihood estimation (MLE) procedure in some scenarios. Motivated by this, they theoretically prove the (near) minimax optimality of the PESSIMISTIC MLE approach with a high probability guarantee. In addition, Zhan et al. (2024) introduced a novel paradigm for general reward functions, and they introduce ε-bracket approximations for reward models, accompanied by a rigorous theoretical analysis delineating sample complexity in terms of approximation error ε and the high-probability parameter δ. Recently, despite the significant contributions of Cen et al. (2024) and Liu et al. (2024) in advancing the integration of experimental findings and theoretical analysis in offline RLHF, the bounds they established are still characterized as worst-case bounds with inverse polynomial forms.

We observe that the above theoretical investigations, while invaluable, are not instance-dependent and do not exhibit exponential convergence. Typically, the above minimax or worst-case guarantees yield upper bounds in the form of $\tilde{O}(n^{-1/2})$ and do not depend on any problem-specific factors (such as suboptimality gaps). Our research stands out as a pioneering attempt in offering an instance-dependent examination that yields exponential convergence for offline RLHF with pairwise comparisons, thereby bridging a critical gap in the existing literature.

Label-Differential Privacy: In our study, we also consider the notion of *label privacy*, acknowledging that the labels in our offline dataset originate from users, thus highlighting the imperative to protect user privacy. Chaudhuri and Hsu (2011) were among the pioneers in exploring the concept of *label privacy* within the context of supervised learning for binary classification. Their foundational work posits that the sensitive information primarily resides in the labels, while considering the unlabeled attributes as non-sensitive. Later, the concept of label privacy has been investigated across various machine learning paradigms, including but not limited to PAC learning (Beimel, Nissim, and Stemmer 2013) and deep learning frameworks (Ghazi et al. 2021). This broadened examination underscores the significance and relevance of label privacy considerations across diverse areas of machine learning research and applications.

More recently, Chowdhury, Zhou, and Natarajan (2024) investigated the use of label differential privacy to protect the privacy of human labelers in the process of estimating rewards from preference-based feedback. Chowdhury, Zhou, and Natarajan (2024) derive an upper bound for their proposed algorithm on the estimation error. They show that it also decays as $O(n^{-1/2})$ and the implied constant here depends on (ε, δ), the parameters that define differential privacy. This bound only applies in the scenario of estimating the reward value and is not applicable if we want to understand how it depends on the simple regret of a specific instance. In our work, we consider the effect of (ε, δ)-DP on the simple regret.

1.2 Our Contributions

We summarize our main contributions as follows:

1. We establish the first-of-its-kind instance-dependent lower bound characterized by suboptimality gaps for a given problem instance. Our analysis reveals that this lower bound takes the form $\exp(-O(n/H))$, where H is a hardness parameter that is an explicit function of the suboptimality gaps. This finding furnishes a novel, and possibly generalizable, analytical approach for assessing algorithmic performance within the realm of preference-based RL.

2. We design a simple algorithm RL-LOW based on the novel concept of *locally optimal weights*. Our analysis demonstrates that its expected simple regret matches the aforementioned instance-dependent lower bound (in the exponential decay rate of the simple regret), thus revealing our algorithm's achievement of instance-dependent optimality.

3. We extend RL-LOW to be applicable to the (ε, δ)-differential privacy with labels by combining the Gaussian mechanism with the aforementioned locally optimal weights. Our analysis demonstrates that, for large datasets, this combination enables our algorithm to achieve differential privacy without weakening the bound on the simple regret, underscoring the superiority of the design and analysis of RL-LOW.

4. As a by-product of our analyses, we show that RL-LOW achieves a worst-case bound of the form $O(n^{-1/2})$ for the dependency of n. If we translate the high-probability upper bound in Zhu, Jordan, and Jiao (2023) to the same worst-case setting, we obtain a bound of the form $O(\sqrt{n^{-1}\log n})$. Thus, our work provides a noticeable (albeit small) improvement over the state-of-the-art theoretical result in Zhu, Jordan, and Jiao (2023) in terms of the dependency on the sample size n.

2 Preliminaries and Problem Setup

Let $\mathcal{S} = \{1, \ldots, S\}$ denote the state space, and $\mathcal{A} = \{1, \ldots, A\}$ denote the action set. The i-th action of state

k is associated with the feature vector $\phi(k,i) \in \mathbb{R}^d$, and its associated (unknown) reward is

$$r_{k,i} = \langle \phi(k,i), \theta \rangle, \tag{1}$$

where $\theta \in \mathbb{R}^d$ is an unknown parameter vector. The collection of all feature vectors is denoted as $\phi = \{\phi(k,i)\}_{k \in \mathcal{S}, i \in \mathcal{A}}$. For all $k \in \mathcal{S}$, we denote the suboptimaliy gap of action $i \in \mathcal{A}$ as $\Delta_{k,i} = \max_{j \in \mathcal{A}} r_{k,j} - r_{k,i}$. Let $(a^{(0)}, a^{(1)}) \in \mathcal{A}^2$ be a pair of comparisons and let $s \in \mathcal{S}$ be a state. Then, we define a stochastic label $\sigma \in \{0,1\}$, following the Bradley–Terry–Luce (BTL) model as

$$\mathbb{P}\big(\sigma = 1 \mid a^{(0)}, a^{(1)}, s\big) = \frac{\exp\big(r_{s,a^{(1)}}\big)}{\exp\big(r_{s,a^{(0)}}\big) + \exp\big(r_{s,a^{(1)}}\big)}. \tag{2}$$

Given this model, we assume throughout that we have access to an *offline dataset*, which we denote as $\mathcal{D} = \{(s_i, a_i^{(0)}, a_i^{(1)}, \sigma_i)\}_{i=1}^n$. Note that this dataset consists of n tupies of states, pairs of actions for comparison, and stochastic labels. Without loss of generality, we assume that the comparisons are arranged such that $a_i^{(0)} < a_i^{(1)}$ for all $i = 1, \ldots, n$, and $a_i^{(0)} < a_j^{(0)}$ (or $a_i^{(1)} \le a_j^{(1)}$ if $a_i^{(0)} = a_j^{(0)}$) for all $i < j$. For simplicity, we assume that the feature vectors satisfy $\phi(k,i) \neq \phi(k,j)$ for all states $k \in \mathcal{S}$ and all actions $i \neq j$. In addition, we assume that for each state k, the best action $i_k^* = \arg\max_{j \in \mathcal{A}} r_{k,j}$ is unique. Broadly speaking, our objective is to use the offline dataset $\mathcal{D}$ to estimate the best action i_k^* for each state $k \in \mathcal{S}$. Following Zhu, Jordan, and Jiao (2023), we aim to design a (possibly randomised) algorithm Π that uses the dataset $\mathcal{D}$ to output a set of actions $\{\hat{i}_k\}_{k \in \mathcal{S}}$ that minimizes the *simple regret*[1], defined as

$$R_n = \mathbb{E}_{k \sim \rho}\big[r_{k,i_k^*} - r_{k,\hat{i}_k}\big], \tag{3}$$

where $\rho = (\rho_1, \ldots, \rho_S)$ is an unknown static distribution over states. Without loss of generality, we assume $\rho_i > 0$ for $i \in \mathcal{S}$. We also consider a generalised version of the regret that is amenable to the MDP setting of RL in Section 5. by extending ρ to be dynamic. Let $N \in \mathbb{R}^{S \times A \times A}$ be a tensor that collects the proportions of each comparison in the dataset $\mathcal{D}$, which satisefies $\lceil nN_{k,i,j} \rceil = \sum_{\iota=1}^n \mathbf{1}\{s_\iota = k, a_\iota^{(1)} = i, a_\iota^{(2)} = j\}$, where $N_{k,i,j} \in [0,1]$ is the proportion of the number of times actions i and j have been compared under state k. In this paper, we consider the concept of problem instance, denoted as v, which is characterised by $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$.

In the following, we index instance-specific parameters with the instance v to indicate their dependence on v; this will be omitted when the instance is clear from the context. In addition, we write $\mathbb{P}_v^\Pi$ (resp. $\mathbb{E}_v^\Pi$) to denote the probability measure (reps. the expectation) induced under algorithm Π and under the instance v. For notational brevity in the rest of the paper, we assume $nN_{k,i,j} \in \mathbb{N}$ is an integer. [2]

[1]The term "simple regret" is referred to as "performance gap" in some existing works (e.g., Zhu, Jordan, and Jiao (2023)).

[2]For notational brevity, we assume that $nN_{k,i,j} \in \mathbb{N}$ is an integer. To be more precise, the sample count for (k,i,j) should be written as $\lceil nN_{k,i,j} \rceil$.

Assumption 2.1. (Bounded Reward) There exists a finite and known constant L such that for any $k \in \mathcal{S}$ and $i \in \mathcal{A}$, it holds that $|\langle \phi(k,i), \theta \rangle| \le L$.

In previous works (Zhu, Jordan, and Jiao 2023), the authors assume that the norms of the feature vectors $\phi(k,i)$ and parameter vector θ are separately bounded. This clearly implies that Assumption 2.1 is satisfied, but Assumption 2.1 is weaker as it is a bound on the rewards.

Definition 2.2. (Consistent Instance) A problem instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$ is *consistent* if for all $(k,i,j) \in \mathcal{S} \times \mathcal{A}^2$, it holds that $\phi(k,i) - \phi(k,j) \in \mathrm{Span}\{\phi(k',i') - \phi(k',j') : (k',i',j') \in \mathcal{S} \times \mathcal{A}^2 \text{ and } N_{k',i',j'} > 0\}$.

We say an instance v is *inconsistent* if it is not consistent. In the following, we will be only concerned with those instances that are consistent as the following result shows that it is impossible to design a algorithm that achieves vanishing simple regret for inconsistent instances.

Proposition 2.3. *(Impossibility Result) For any inconsistent instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$, there exists an instance $v' = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta')$ such that for all algorithms Π*

$$\liminf_{n \to \infty} \big\{\mathbb{E}_v^\Pi[R_n] + \mathbb{E}_{v'}^\Pi[R_n]\big\} > 0. \tag{4}$$

3 The Proposed Algorithm: RL-LOW

In this section, we describe our computationally and statistically efficient algorithm for offline RLHF with pairwise comparisons based on the novel idea of *locally optimal weights* for estimating the relative reward of each pair of actions. For clarity in exposition, this section is devoted to the setting of a static state distribution ρ; this is done as a foundational step before we extend the ideas to the MDP setting of RL in Section 5.

Our proposed algorithm, called RL-LOW, is simple and is presented formally in Algorithm 1. Before we describe its components, we introduce some notations.

Let $B_{k,i,j}$ be the empirical success rate with the comparison of i and j, i.e., for $k \in \mathcal{S}$ and $i,j \in \mathcal{A}$ with $N_{k,i,j} > 0$,

$$B_{k,i,j} \coloneqq \frac{1}{nN_{k,i,j}} \sum_{\iota=1}^n \sigma_\iota \mathbb{1}\{s_\iota = k, a_\iota^{(1)} = i, a_\iota^{(2)} = j\}, \tag{5}$$

and $B_{k,j,i} \coloneqq 1 - B_{k,i,j}$. If $N_{k,i,j} = N_{k,j,i} = 0$, we define $B_{k,i,j} = B_{k,j,i} = 0$. Subsequently, certain empirical success rates may exhibit magnitudes that are either excessively large or small. We clip them by means of the following operation: $B_{k,i,j}^{\mathrm{CLP}} = \mathrm{CLIP}_L(B_{k,i,j})$, where

$$\mathrm{CLIP}_L(a) = \begin{cases} \frac{\exp(2L)}{1+\exp(2L)} & a > \frac{\exp(2L)}{1+\exp(2L)} \\ \frac{1}{1+\exp(2L)} & a < \frac{1}{1+\exp(2L)} \\ a & \text{otherwise} \end{cases}. \tag{6}$$

Per Assumption 2.1, the implicit rewards are bounded by L. Consequently, within our BTL model framework, the success rate of each comparison necessarily falls within the interval $\big[\frac{1}{1+\exp(2L)}, \frac{\exp(2L)}{1+\exp(2L)}\big]$. We exploit this in Eqn. (6) to ensure that the implementation of our clip operation is consistent with the model's dynamics. We are now ready to introduce the notion of *locally optimal* weights, which plays a central role in the estimation of the rewards.

Algorithm 1: Reinforcement Learning with Locally Optimal Weights (RL-LOW)

Input: Dataset $\mathcal{D} = \{(s_i, a_i^{(1)}, a_i^{(2)}, \sigma_i)\}_{i=1}^n$ and feature maps $\phi = \{\phi(k,i)\}_{k\in\mathcal{S}, i\in\mathcal{A}}$.

Output: The estimated best action $\hat{i}_k \in \mathcal{A}$ for each $k \in \mathcal{S}$.

1: Compute the sample proportions $N_{k,i,j} \leftarrow \frac{1}{n}\sum_{\iota=1}^n \mathbb{1}\{s_\iota = k, a_\iota^{(1)} = i, a_\iota^{(2)} = j\}$.
2: For $k \in \mathcal{S}$ and $i, j \in \mathcal{A}$ such that $N_{k,i,j} > 0$, compute the success rate $B_{k,i,j}$ using Eqn. (5).
3: Compute $B_{k,i,j}^{\mathrm{CLP}}$ by clipping $B_{k,i,j}$ through Eqn. (6).
4: For each state $k \in \mathcal{S}$ and distinct actions $i, j \in \mathcal{A}$ with $i < j$, compute the locally optimal weights $(w_{k',i',j'}^{(k,i,j)})_{k'\in\mathcal{S},i',j'\in\mathcal{A}}$ using Eqn. (7).
5: Compute the empirical relative reward $\hat{r}_{k,i,j}$ for each $k \in \mathcal{S}, i, j \in \mathcal{A}$ using Eqn. (8).
6: RETURN for any $k \in \mathcal{S}$, let $\hat{i}_k \in \{i \in \mathcal{A} : \hat{r}_{k,i,j} \geq 0, \forall j \neq i\}$; resolve ties uniformly.

Definition 3.1. (Locally Optimal Weight) For an consistent instance v, let $\mathcal{U}_{k,i,j} = \{u \in \mathbb{R}^{S\times A\times A} : \phi(k,i) - \phi(k,j) = \sum_{k'\in\mathcal{S},i',j'\in\mathcal{A}} u_{k',i',j'}(\phi(k',i') - \phi(k',j'))$ and $u_{k',i',j'} = 0$ if $N_{k',i',j'} = 0\}$. We say that $w^{(k,i,j)} = (w_{k',i',j'}^{(k,i,j)})_{k'\in\mathcal{S},i',j'\in\mathcal{A}}$ is a set of *locally optimal weights* for $(k,i,j) \in \mathcal{S}\times\mathcal{A}^2$ with $i \neq j$ if

$$w^{(k,i,j)} \in \operatorname*{argmin}_{u\in\mathcal{U}_{k,i,j}} \left\{ \sum_{k'\in\mathcal{S},i',j'\in\mathcal{A}:N_{k',i',j'}>0} \frac{(u_{k',i',j'})^2}{N_{k',i',j'}} \right\}. \quad (7)$$

The weights in Eqn. (7) are described as "locally optimal" because they are customized to each (k,i,j) tuple. Hence, $w^{(k,i,j)}$ is *local* to (k,i,j). This is a novelty in the design of our algorithm.

By the definition of the consistency of an instance (cf. Definition 2.2), there exists a subset $\beta \subset \mathcal{S}\times\mathcal{A}^2$ such that $\phi(k,i) - \phi(k,j) \in \mathrm{Span}\{\phi(k',i') - \phi(k',j') : (k',i',j') \in \beta\}$. Hence, there exists a locally optimal weight for every pair of actions given a consistent instance. In addition, $w^{(k,i,j)}$ can be calculated efficiently by its analytic form (see details in Appendix F[3]).

Equipped with the definition of locally-optimal weights, we now provide an estimate of the relative reward for state $k \in \mathcal{S}$ and pair of action $(i,j) \in \mathcal{A}^2$ with $i \neq j$ as follows:

$$\hat{r}_{k,i,j} = \sum_{k'\in\mathcal{S},i'\in\mathcal{A},j'\in\mathcal{A}} w_{k',i',j'}^{(k,i,j)} \log\Big(\frac{B_{k',i',j'}^{\mathrm{CLP}}}{1 - B_{k',i',j'}^{\mathrm{CLP}}}\Big), \quad (8)$$

and we define $\hat{r}_{k,i,i} = 0$ for all $k \in \mathcal{S}$ and $i \in \mathcal{A}$. The term $\frac{(u_{k',i',j'})^2}{N_{k',i',j'}}$ in Eqn. (7) is a proxy for the variance introduced by the pair of actions (i',j') in state k' when associated with the coefficient $u_{k',i',j'}$ in the linear combination of the definition of $\mathcal{U}_{k,i,j}$. Our objective is to minimize the cumulative variance proxy for (k,i,j), thus enhancing the precision of the estimate of the relative reward for (k,i,j) for the purposes of establishing the tightest possible concentration result for subGaussian random variables (see Appendix H).

Finally, for any $k \in \mathcal{S}$, let $\hat{i}_k \in \hat{\mathcal{I}}_k := \{i \in \mathcal{A} : \hat{r}_{k,i,j} \geq 0, \forall j \neq i\}$ be any estimate of the best action under state k. It is natural to wonder whether $\hat{i}_k$ exists, i.e., whether the set $\hat{\mathcal{I}}_k$ is empty. The following proposition answers this in the affirmative.

Proposition 3.2. *For any consistent instance v and using estimate of the best action $\hat{i}_k$ under each state k as prescribed by* RL-LOW*, we have $|\hat{\mathcal{I}}_k| \geq 1$ and*

$$\operatorname*{argmax}_{i\in\mathcal{A}} \hat{r}_{k,i,j_1} = \operatorname*{argmax}_{i\in\mathcal{A}} \hat{r}_{k,i,j_2} = \hat{\mathcal{I}}_k \quad \textit{for any } j_1, j_2 \in \mathcal{A}.$$

Computational Complexity: Proposition 3.2 obviates the need to compute all values of $\hat{r}_{k,i,j}$ for each $(k,i,j) \in \mathcal{S}\times\mathcal{A}^2$. We demonstrate that the RL-LOW algorithm can be efficiently implemented with a computational complexity of $\mathcal{O}(SAd + nd^2 + d^3)$, as the term SAd corresponds to the natural process of scanning the feature vectors for all state-action pairs. The terms $nd^2 + d^3$ are typical in scenarios involving a linear reward structure, such as in linear regression. It is worth noting that the term SAd can be removed if we do not need to output $\hat{i}_k$ for each $k \in \mathcal{S}$, but rather a *parametric function* $\hat{i}(k;\vartheta)$ is to be learned, and the overall computational complexity becomes $\mathcal{O}(nd^2 + d^3)$; see details in Appendix F.2.

3.1 Upper Bound of RL-LOW

In this section, we provide an instance-dependent upper bound of the simple regret for Rl-LOW, and we also provide a worst-case upper bound as a by-product. First, we define an instance-dependent hardness parameter $H(v)$. Let

$$H(v) \coloneqq \max_{k\in\mathcal{S},i\in\mathcal{A}:i\neq i_k^*} \frac{\gamma_{k,i}}{\Delta_{k,i}^2}, \quad (9)$$

$$\text{where } \gamma_{k,i} \coloneqq \sum_{k'\in\mathcal{S},i',j'\in\mathcal{A}:N_{k',i',j'}>0} \frac{(w_{k',i',j'}^{(k,i,i_k^*)})^2}{N_{k',i',j'}}.$$

The parameter $\gamma_{k,i}$ exhibits a positive correlation with the variance proxy of the relative empirical reward $\hat{r}_{k,i,i_k^*}$. Consequently, the ratio $\frac{\gamma_{k,i}}{\Delta_{k,i}^2}$ in the definition of $H(v)$ serves as a quantitative measure of the difficulty that the empirical reward of a suboptimal action i surpasses that of i_k^* in state k; see more intuitive explanations in Appendix C.

Theorem 3.3. *(Instance-Dependent Upper Bound) For any consistent instance v, under* RL-LOW*, we have for all sufficiently large n,*

$$\mathbb{E}_v^{\text{RL-LOW}}[R_n] \leq \exp\Big(-\frac{n}{C_{\mathrm{up}}\cdot H(v)}\Big), \quad (10)$$

where C_{up} is a universal constant.[4]

[3] Following AAAI policy, the appendices are provided on the arxiv (Chen and Tan 2025).

[4] In this paper, our universal constants depend on L, which is known and fixed throughout.

From Theorem 3.3, it is evident that the upper bound decays exponentially fast and the exponent is a function of an instance-dependent hardness term $H(v)$. This is the first instance-dependent analysis in offline reinforcement learning with pairwise comparisons.

It is natural to wonder why we do not devise an instance-dependent analysis of or modification to the existing PESSIMISTIC MLE in Zhu, Jordan, and Jiao (2023). Note that PESSIMISTIC MLE is designed to perform well *with high probability* and not necessarily *in expectation*. In particular, the regret bound of PESSIMISTIC MLE holds with probability at least $1-\delta$. Hence, to ensure the regret is less than $\exp(-\Omega(n/H(v)))$, one should set the failure probability δ to be $\exp(-\Theta(n/H(v)))$, which is not possible as $H(v)$ is unknown to the algorithm (since θ is also unknown). We further provide a worst-case upper bound for RL-LOW as follows.

Proposition 3.4. *(Worst-Case Upper Bound) For any consistent instance v and for all $n \geq 1$,*

$$\mathbb{E}_v^{\text{RL-LOW}}[R_n] \leq \frac{\sum_{k,i:i\neq i_k^*} \rho_k(\sqrt{\gamma_{k,i}}+\tilde{\gamma}_{k,i})}{C_{\text{wup}}\sqrt{n}} \tag{11}$$

where $\tilde{\gamma}_{k,i} = \sum_{k',i',j':N_{k',i',j'}>0} \frac{|w_{k',i',j'}^{(k,i,i_k^*)}|}{\sqrt{N_{k',i',j'}}}$

and $C_{\text{wup}} > 0$ is a universal constant.

We note that in Zhu, Jordan, and Jiao (2023), the high probability upper bound is of the form $O\big(\sqrt{n^{-1}\log(1/\delta)}\big)$ for the dependency of n and δ. Hence, if we desire a bound in expectation, we obtain, through the law of total probability, a bound of the form $\mathbb{E}[R_n] = O\big(\sqrt{n^{-1}\log(1/\delta)}+\delta\big)$. Minimizing this bound over δ yields $\mathbb{E}[R_n] = O\big(\sqrt{n^{-1}\log n}\big)$. In terms of the dependence on n, it exhibits a performance that is slightly inferior to our established upper bound $O(n^{-1/2})$ of RL-LOW.

4 Instance-Dependent Lower Bound

In this section, we derive the first-of-its-kind instance-dependent lower bound on offline RLHF with pairwise comparisons. Before we present our bound, we present some auxiliary lemmas that are potentially instrumental in deriving lower bounds on other preference-based RL problems.

Given any instance v, we let $P_v^{(n)}$ denote the joint distribution of the associated labels $\{\sigma_i\}_{i=1}^n$. The following lemma provides an estimate of the Kullback–Leibler (KL) divergence between instances v and v' that share the same parameters except for the latent vector θ as in Eqn. (1).

Lemma 4.1. *For any instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$ and $v' = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta')$, it holds that*

$$2n\exp(-4R_{\max}) \leq \frac{D_{\text{KL}}(P_v^{(n)}\|P_{v'}^{(n)})}{\tilde{D}(v,v')} \leq 2n\exp(2R_{\max})$$

where $\tilde{D}(v,v') = \sum_{k\in\mathcal{S},i,j\in\mathcal{A}} N_{k,i,j}(\langle\phi(k,i)-\phi(k,j),\theta-\theta'\rangle)^2,$ (12)

$R_{\max} = \max_{k\in\mathcal{S},i\in\mathcal{A}} \max\{|\langle\phi(k,i),\theta\rangle|, |\langle\phi(k,i),\theta'\rangle|\}$ *is the maximum absolute reward in these two instances.*

This lemma demonstrates that when the rewards are bounded, the weighted sum of squared differences of the relative rewards can be used to approximate the KL divergence between the distributions of two instances. The approximation is precisely $\tilde{D}(v,v')$ defined in Eqn. (12). Furthermore, for any $\mathbf{z} \in \mathbb{R}^d$, $\eta \in \mathbb{R}$ and consistent instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$, we let $\text{Alt}(v, \mathbf{z}, \eta)$ be the set of instances that share the same instance parameters except for θ and satisfies $\langle \mathbf{z}, \theta' - \theta\rangle = \eta$ for all $v' = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta') \in \text{Alt}(v, \mathbf{z}, \eta)$. The following lemma states a useful property that relates the Alt set to the "approximate KL divergence" $\tilde{D}$.

Lemma 4.2. *Let $\mathcal{G}$ be an arbitrary orthonormal basis of $\text{Span}\{\phi(k',i') - \phi(k',j') : (k',i',j') \in \mathcal{S}\times\mathcal{A}^2$ and $N_{k',i',j'} > 0\}$. Also let $[\mathbf{w}]_{\mathcal{G}}$ denote the column vector that represents $\mathbf{w}$ under basis $\mathcal{G}$ (Meyer 2000, Chapter 4). Define the matrix*

$$V := \sum_{k\in\mathcal{S},i,j\in\mathcal{A}} N_{k,i,j}[\phi(k,i)-\phi(k,j)]_{\mathcal{G}}[\phi(k,i)-\phi(k,j)]_{\mathcal{G}}^\top. \tag{13}$$

Then for any consistent instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta)$, $\eta \in \mathbb{R}$, and $\mathbf{z} \in \text{Span}\{\phi(k',i') - \phi(k',j') : (k',i',j') \in \mathcal{S}\times\mathcal{A}^2$ and $N_{k',i',j'} > 0\}$,

$$\min_{v'\in\text{Alt}(v,\mathbf{z},\eta)} \tilde{D}(v,v') = \frac{\eta^2}{\|[\mathbf{z}]_{\mathcal{G}}\|_{V^{-1}}^2}. \tag{14}$$

Lemma 4.2 provides an estimate of the KL divergence between instance v and $v' \in \text{Alt}(v, \mathbf{z}, \eta)$. This, in turn, provides a convenient means to apply the ubiquitous *change of measure* technique to derive the lower bound.

In addition, let $(\bar{i}, \bar{k})$ be the state-action pair that attains maximum in the definition of hardness in Eqn. (3.1). Define the subset of instances

$$\mathcal{Q} = \left\{ v \text{ consistent} : \frac{\gamma_{\bar{k},\bar{i}}}{\Delta_{\bar{k},\bar{i}}^2} \geq \frac{4\gamma_{k,i}}{\Delta_{k,i}^2} \ \forall (k,i) \neq (\bar{k},\bar{i}),\ i \neq i_k^* \right\}. \tag{15}$$

We are now ready to state our lower bound.

Theorem 4.3. *(Instance-Dependent Lower Bound) For any instance $v = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta) \in \mathcal{Q}$, there exists another instance $v' = (\rho, \mathcal{S}, \mathcal{A}, \phi, N, \theta')$ with $H(v) \leq H(v') \leq 8H(v)$ such that for all sufficiently large n,*

$$\inf_{\Pi} \{\mathbb{E}_v^{\Pi}[R_n] + \mathbb{E}_{v'}^{\Pi}[R_n]\} \geq \exp\left(-\frac{n}{C_{\text{lo}}\cdot H(v)}\right),$$

where $C_{\text{lo}} > 0$ is a universal constant.

The alternative instance v' that appears in Theorem 4.3 is judiciously chosen to be $v' \in \text{Alt}(v, \phi(\bar{k}(v), \bar{i}(v)) - \phi(\bar{k}(v), i_{\bar{k}}^*(v)), 2\Delta_{\bar{k}(v),\bar{i}(v)}(v))$. In particular, it is designed so that the optimal action $i_{\bar{k}}^*$ under state k of instance v will become suboptimal under instance v', and its suboptimality gap is at least $\Delta_{\bar{k}(v),\bar{i}(v)}(v)$ under v'.

Theorem 4.3 is an instance-dependent lower bound for all instances in the set $\mathcal{Q}$. The condition that defines $\mathcal{Q}$ in Eqn. (15) ensures that the hardness quantities $H(v)$ and

$H(v')$ have the same order. Since instances in $\mathcal{Q}$ cover all possible hardness values $H(v)$ (i.e., for every hardness values $h > 0$, there exists an instance in $\mathcal{Q}$ of hardness h), we conclude that for any (small) $\epsilon \in (0, 1)$, there does not exist any algorithm Π that achieves for all consistent instance v,

$$\mathbb{E}_v^{\Pi}[R_n] = \exp\Big(-\Omega\Big(\frac{n}{H(v)^{1-\epsilon}}\Big)\Big). \tag{16}$$

In this sense, the exponential decay rate of the simple regret of RL-LOW presented in Theorem 3.3 is asymptotically tight (or optimal) and the exponential dependence on the hardness parameter $H(v)$ is necessary, fundamental, and cannot be improved upon.

5 Extension to the MDP Setting of RL

Similar to Zhu, Jordan, and Jiao (2023, Section 1), our definition of simple regret is based on the *static* state distribution ρ in the previous sections. In this section, we extend our results to the MDP setting of RL when the transition probabilities $P(k'|k, i)$ for $(k', k, i) \in \mathcal{S}^2 \times \mathcal{A}$ are assumed to be known. This assumption is consistent with Zhu, Jordan, and Jiao (2023, Section 5), and we also provide a motivational example for this assumption in Appendix A.2. Given the transition probabilities $P(k'|k, i)$ and an MDP policy π, we let d^π denote the state distribution (Sutton and Barto 2018, Section 9.2) under π. Without loss of generality, we assume the MDP policies are deterministic, and we denote $\pi(k) \in \mathcal{A}$ to be the output action of π under state k. Let π^* denote the optimal MDP policy that is assumed to be unique, i.e.,

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}_{k\sim d^\pi}\big[r_{k,\pi(k)}\big].$$

Then, we define the *simple regret* of any MDP policy π as

$$R^{\mathrm{MDP}}(\pi) = \mathbb{E}_{k\sim d^{\pi^*}}\big[r_{k,\pi^*(k)}\big] - \mathbb{E}_{k\sim d^{\pi}}\big[r_{k,\pi(k)}\big]. \tag{17}$$

We now adapt our RL-LOW to the MDP setting by redefining the output as an MDP policy:

$$\hat{\pi}_{\mathrm{out}} \in \operatorname*{argmax}_{\pi} \mathbb{E}_{k\sim d^\pi}[\hat{r}_{k,\pi(k),j^\dagger}],$$

where $j^\dagger \in \mathcal{A}$ is arbitrarily fixed (e.g., $j^\dagger = 1$). We simply call this adaptation RL-LOW-MDP. The upper bound on its simple regret is stated as follows.

Theorem 5.1. *(Instance-Dependent Upper Bound for* RL-LOW-MDP*) Given any consistent instance v, for all sufficiently large n,*

$$\mathbb{E}_v^{\text{RL-LOW-MDP}}\Big[R^{\mathrm{MDP}}(\hat{\pi}_{\mathrm{out}})\Big] \le \exp\Big(-\frac{n}{C_{\mathrm{MDP}} \cdot H_{\mathrm{MDP}}(v)}\Big)$$

where $C_{\mathrm{MDP}} > 0$ is a universal constant,

$$H_{\mathrm{MDP}}(v) \coloneqq \max_{\pi\neq\pi^*} \frac{\gamma^{\mathrm{MDP}}(\pi)}{(\mathbb{E}_{k\sim d^{\pi^*}}\big[r_{k,\pi^*(k)}\big] - \mathbb{E}_{k\sim d^{\pi}}\big[r_{k,\pi(k)}\big])^2},$$

$$\gamma^{\mathrm{MDP}}(\pi) \coloneqq \max_{k:\pi(k)\neq\pi^*(k)} \sum_{k',i',j':N_{k',i',j'}>0} \frac{(w^{(k,\pi(k),\pi^*(k)}_{k',i',j'})^2}{N_{k',i',j'}}.$$

In the presence of the MDP, $H_{\mathrm{MDP}}(v)$, which is a generalization of $H(v)$ in Eqn. (3.1), turns out to be the instance-dependence hardness parameter of the problem. The proof of Theorem 5.1 is provided in Appendix H. It is important to observe that there exist MDPs (e.g., $P(k|k, i) = 1$ or $S = 1$) such that Theorem 5.1 particularizes to Theorem 3.3. Moreover, the lower bound in Theorem 4.3 is also applicable to the present more general MDP setting when the transition probability kernel $P(k'|k, i)$ is independent of (k, i) and k' follows the distribution ρ. Admittedly, the complexity of the problem increases substantially when the transition probabilities are unknown; this aspect warrants further investigation in future studies. Our findings serve as an initial step in exploring instance-dependent bounds in offline RLHF.

6 Extension to (ε, δ)-Differential Privacy (DP)

In this section, we extend our algorithm RL-LOW to be amenable to (ε, δ)-differential privacy with labels, and we provide a motivational example of this extension in Appendix A. To formalize our results, we provide the definition of (ε, δ)-DP, following Dwork, Roth et al. (2014). We say that two sets of preference labels, $\sigma := \{\sigma_i\}_{i=1}^n$ and $\sigma' := \{\sigma'_i\}_{i=1}^n$ are *neighboring* if there exists $s \in [n]$ such that $\sigma_s \neq \sigma'_s$ and $\sigma_j = \sigma'_j$ for all $j \neq s$.

Definition 6.1. (Differential Privacy with labels) Fix any label-free dataset $\{(s_i, a_i^{(1)}, a_i^{(2)})\}_{i=1}^n$. A (randomized) algorithm $\mathcal{M} : \{0, 1\}^n \to \mathcal{A}^S$ (that takes as inputs a set of labels and outputs a set of actions, one for each state) satisfies (ε, δ)-DP if for all neighboring labels $\sigma := \{\sigma_i\}_{i=1}^n$ and $\sigma' := \{\sigma'_i\}_{i=1}^n$ it holds $\forall\, \mathcal{Z} \subset \mathcal{A}^S$

$$\mathbb{P}(\mathcal{M}(\sigma) \in \mathcal{Z}) \le e^{\varepsilon}\, \mathbb{P}(\mathcal{M}(\sigma') \in \mathcal{Z}) + \delta. \tag{18}$$

Note that Definition 6.1 primarily concerns protecting the privacy of users' *labels*. In particular, the DP mechanism guarantees that any alteration in a user's label must not substantially affect the output of our algorithm. Otherwise, there exists a risk that a user's label might be inferred through the algorithm's output. Our definition of differential privacy (DP) aligns with that of Chowdhury, Zhou, and Natarajan (2024).

We now adapt our RL-LOW to (ε, δ)-DP by using the Gaussian mechanism (Dwork, Roth et al. 2014). Firstly, we introduce the private version of the empirical success rate (analogous to $B_{k,i,j}$ in Eqn. (5)), i.e., for all $k \in \mathcal{S}$ and $i, j \in \mathcal{A}$ and $N_{k,i,j} > 0$,

$$\tilde{B}_{k,i,j} \coloneqq \frac{1}{nN_{k,i,j}} \sum_{\iota=1}^{n} \sigma_\iota \mathbb{1}\{s_\iota = k, a_\iota^{(1)} = i, a_\iota^{(2)} = j\} + \tilde{\xi}_{k,i,j},$$

where $\tilde{\xi}_{k,i,j}$ is an independent (across k, i, and j) Gaussian noise with zero mean and variance $\frac{2\log(1.25/\delta)}{(\varepsilon n N_{k,i,j})^2}$, and we let $\tilde{B}_{k,j,i} \coloneqq 1 - \tilde{B}_{k,i,j}$. If $N_{k,i,j} = N_{k,j,i} = 0$, we define $\tilde{B}_{k,i,j} = \tilde{B}_{k,j,i} = 0$. Again, analogously to the operation in Eqn. (6), we clip $\tilde{B}_{k,i,j}$ to form

$$\tilde{B}^{\mathrm{CLP}}_{k,i,j} = \mathrm{CLIP}_L(\tilde{B}_{k,i,j}) \tag{19}$$

Similarly to Eqn. (8), the perturbed estimated relative rewards are given as follows

$$\tilde{r}_{k,i,j} = \sum_{k'\in\mathcal{S},(i',j')\in\mathcal{A}^2} w^{(k,i,j)}_{k',i',j'} \log\Big(\frac{\tilde{B}^{\mathrm{CLP}}_{k',i',j'}}{1 - \tilde{B}^{\mathrm{CLP}}_{k',i',j'}}\Big), \tag{20}$$

where $w^{(k,i,j)}$ is defined in Definition 3.1. Finally, the empirical best action is $\hat{i}_k \in \tilde{\mathcal{I}}_k := \{i \in \mathcal{A} : \tilde{r}_{k,i,j} \geq 0, \forall\, j \neq i\}$. A similar argument as Proposition 3.2 shows that $\hat{i}_k$ exists; see details in Appendix F for the details. The algorithm described above is a differentially private version of RL-LOW and hence, it is named DP-RL-LOW.

DP-RL-LOW with the carefully chosen variance of $\xi_{k,i,j}$ fulfils the requirement of (ε, δ)-DP.

Proposition 6.2. *Given privacy parameters $\varepsilon, \delta \in (0, 1)$,* DP-RL-LOW *satisfies (ε, δ)-DP.*

The proof of Proposition 6.2 follows exactly along the lines of the proof of Dwork, Roth et al. (2014, Theorem A.1) and is omitted. Then, the upper bound is as follows.

Theorem 6.3. *(Instance-Dependent Upper Bound for* DP-RL-LOW*) Given any consistent instance v, for all sufficiently large n,*

$$\mathbb{E}_v^{\text{DP-RL-LOW}}[R_n] \leq \exp\left(-C_{\text{DP}} \cdot \Big(\frac{n}{H(v)} \wedge \Big(\frac{n}{H_{\text{DP}}^{(\varepsilon,\delta)}(v)}\Big)^2\Big)\right), \tag{21}$$

where $C_{\text{DP}} > 0$ is a universal constant, and

$$H_{\text{DP}}^{(\varepsilon,\delta)}(v) = \max_{k \in \mathcal{S}, i \in \mathcal{A}: i \neq i_k^*} \frac{\sqrt{\log(\frac{1.25}{\delta})\gamma_{k,i}^{\text{DP}}}}{\sqrt{\varepsilon}\Delta_{k,i}}, \quad \textit{and} \tag{22}$$

$$\gamma_{k,i}^{\text{DP}} = \sum_{k',i',j' \in \mathcal{A}: N_{k',i',j'} > 0} \left(\frac{w_{k',i',j'}^{(k,i,i_k^*)}}{N_{k',i',j'}}\right)^2. \tag{23}$$

Consequently,

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E}_v^{\text{DP-RL-LOW}}[R_n] \leq -\frac{C_{\text{DP}}}{H(v)}. \tag{24}$$

The limiting statement in Eqn. (24) implies that DP-RL-LOW has the same order of the exponential decay rate as its non-differentially privacy counterpart RL-LOW when n is sufficiently large; in particular, $n > (H_{\text{DP}}^{(\varepsilon,\delta)}(v))^2/H(v)$ suffices to nullify the effect of the privacy requirement. In other words, in the sense of the exponent, privacy comes "for free" for sufficiently large offline datasets.

In addtion, we derive the worst-case upper bound of DP-RL-LOW in Appendix I.2, which yields the form of $O(\frac{1}{\sqrt{n}} + \frac{\sqrt{\log(1/\delta)}}{\varepsilon n})$ for the dependency in n, ε and δ. This again implies that privacy comes for free for sufficiently large offline dataset.

7 Related Works Beyond PbRL

Offline RL without Human Feedback The domain of offline RL has been extensively researched over an extended period. Here, We focus on the recent works. Chen and Jiang (2019) revisits and provides theoretical insights into the essential but underexplored assumptions of mild distribution shift and strong representation conditions in value-function approximations, advancing their necessity and applicability. Xie et al. (2021) bridges the gap between online and offline reinforcement learning by introducing the policy finetuning problem, proposing algorithms that leverage a reference policy close to the optimal policy to achieve sample-efficient learning in episodic MDPs. Yin et al. (2022) investigates the statistical limits of offline reinforcement learning using linear models, introducing the variance-aware pessimistic value iteration method to improve learning bounds with offline data. More recently, Wang, Cui, and Du (2022) enhances the understanding of gap-dependent sample complexity in offline reinforcement learning, demonstrating improved rates under specific policy coverage conditions and providing algorithms nearly matching lower bounds. Similarly, Nguyen-Tang et al. (2023) investigated gap-dependent analysis for offline RL, providing both gap-dependent upper and lower bounds for performance with linear function approximation.

Overall, our study identifies a significant oversight in previous research: the absence of human feedback consideration. Consequently, our work represents the inaugural investigation into instance-dependent bounds within the context of offline reinforcement learning incorporating human feedback with pairwise comparisons.

Dueling Bandits The Dueling Bandits problem was first introduced by Yue and Joachims (2009), sparking a substantial body of subsequent research on the topic. In this section, we highlight some relevant works. Inspired by the classical contextual bandits problem, Dudík et al. (2015) extend the framework of duel bandits into a contextual setting, and they propose a new concept of von Neumann winner, a game-theoretic solution concept that addresses limitations of the Condorcet winner, along with three efficient algorithms for its online learning and approximation from data. In contrast, Saha (2021) explore a distinct aspect of contextual dueling bandits through their proposed Subsetwise-Preference Feedback Model, and the author presents two algorithms for pairwise preferences, achieving near-optimal regret bounds, and extending the analysis to general subsetwise preferences, demonstrating that the fundamental performance limits remain consistent regardless of the subset size. However, this study mainly focuses on the worst-case analysis. More recently, Di, He, and Gu (2024) addressed the contextual dueling bandits with adversarial feedback, proposing a robust algorithm using uncertainty-weighted maximum likelihood estimation. Nonetheless, this work focuses on the adversarial setting, whereas our work examines the stochastic setting.

8 Concluding Remarks

This paper studies offline RLHF with pairwise comparisons, aiming to minimize simple regret by identifying the optimal action per state. We propose new algorithms achieving simple regret of the form $\exp(-\Omega(n/H(v)))$, where n is the sample size and $H(v)$ captures instance-dependent hardness from suboptimality gaps. We also establish the first instance-dependent lower bound for this setting, matching our upper bound and proving exponential-rate optimality. To ensure privacy, we adapt our method to be (ε, δ)-differentially private, showing that $H(v)$ remains asymptotically unchanged as $n \to \infty$. By establishing instance-dependent bounds of exponential convergence, our results close a gap in prior works that focused on worst-case regret.

Acknowledgements

This research work was funded by two Singapore Ministry of Education Academic Research Fund Tier 1 grants (A-8000980-00-00 and A-8002934-00-00).

References

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Beimel, A.; Nissim, K.; and Stemmer, U. 2013. Private learning and sanitization: Pure vs. approximate differential privacy. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, 363–378. Springer.

Cen, S.; Mei, J.; Goshvadi, K.; Dai, H.; Yang, T.; Yang, S.; Schuurmans, D.; Chi, Y.; and Dai, B. 2024. Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF. *arXiv preprint arXiv:2405.19320*.

Chaudhuri, K.; and Hsu, D. 2011. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, 155–186. JMLR Workshop and Conference Proceedings.

Chen, J.; and Jiang, N. 2019. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 1042–1051. PMLR.

Chen, X.; Zhong, H.; Yang, Z.; Wang, Z.; and Wang, L. 2022. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, 3773–3793. PMLR.

Chen, Z.; and Tan, V. Y. 2025. On the Exponential Convergence for Offline RLHF with Pairwise Comparisons. *arXiv preprint arXiv:2406.12205*.

Chowdhury, S. R.; Zhou, X.; and Natarajan, N. 2024. Differentially private reward estimation with preference feedback. In *International Conference on Artificial Intelligence and Statistics*, 4843–4851. PMLR.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Di, Q.; He, J.; and Gu, Q. 2024. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. *arXiv preprint arXiv:2404.10776*.

Dudík, M.; Hofmann, K.; Schapire, R. E.; Slivkins, A.; and Zoghi, M. 2015. Contextual dueling bandits. In *Conference on Learning Theory*, 563–587. PMLR.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Ghazi, B.; Golowich, N.; Kumar, R.; Manurangsi, P.; and Zhang, C. 2021. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34: 27131–27145.

Glaese, A.; McAleese, N.; Trębacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Kim, C.; Park, J.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2023. Preference Transformer: Modeling Human Preferences using Transformers for RL. In *The Eleventh International Conference on Learning Representations*.

Liu, H.; Sferrazza, C.; and Abbeel, P. 2023. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*.

Liu, Z.; Lu, M.; Zhang, S.; Liu, B.; Guo, H.; Yang, Y.; Blanchet, J.; and Wang, Z. 2024. Provably Mitigating Overoptimization in RLHF: Your SFT Loss is Implicitly an Adversarial Regularizer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Meyer, C. D. 2000. *Matrix analysis and applied linear algebra*, volume 71. SIAM.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nguyen-Tang, T.; Yin, M.; Gupta, S.; Venkatesh, S.; and Arora, R. 2023. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9310–9318.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Palan, M.; Landolfi, N. C.; Shevchuk, G.; and Sadigh, D. 2019. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*.

Saha, A. 2021. Optimal Algorithms for Stochastic Contextual Preference Bandits. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

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

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Wang, X.; Cui, Q.; and Du, S. S. 2022. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 14865–14877.

Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Xie, T.; Jiang, N.; Wang, H.; Xiong, C.; and Bai, Y. 2021. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34: 27395–27407.

Yin, M.; Duan, Y.; Wang, M.; and Wang, Y.-X. 2022. Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism. In *International Conference on Learning Representation*.

Yue, Y.; and Joachims, T. 2009. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1201–1208.

Zhan, W.; Uehara, M.; Kallus, N.; Lee, J. D.; and Sun, W. 2024. Provable Offline Preference-Based Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Zhu, B.; Jordan, M.; and Jiao, J. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, 43037–43067. PMLR.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.