A Unified Linear Programming Framework for Offline Reward Learning from Human Demonstrations and Feedback

Kihyun Kim [* 1]   **Jiawei Zhang** [* 1]   **Asuman Ozdaglar** [1]   **Pablo Parrilo** [1]

Abstract

Inverse Reinforcement Learning (IRL) and Reinforcement Learning from Human Feedback (RLHF) are pivotal methodologies in reward learning, which involve inferring and shaping the underlying reward function of sequential decision-making problems based on observed human demonstrations and feedback. Most prior work in reward learning has relied on prior knowledge or assumptions about decision or preference models, potentially leading to robustness issues. In response, this paper introduces a novel linear programming (LP) framework tailored for offline reward learning. Utilizing pre-collected trajectories without online exploration, this framework estimates a feasible reward set from the primal-dual optimality conditions of a suitably designed LP, and offers an optimality guarantee with provable sample efficiency. Our LP framework also enables aligning the reward functions with human feedback, such as pairwise trajectory comparison data, while maintaining computational tractability and sample efficiency. We demonstrate that our framework potentially achieves better performance compared to the conventional maximum likelihood estimation (MLE) approach through analytical examples and numerical experiments.

1. Introduction

Reward learning involves inferring and shaping the underlying reward function from observed human demonstrations and feedback. Inverse reinforcement learning (IRL) and reinforcement learning from human feedback (RLHF, also known as preference-based reinforcement learning) are key

---
[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Kihyun Kim <kihyun@mit.edu>.

methodologies in reward learning, applied in various sequential decision-making tasks such as games (MacGlashan et al., 2017; Christiano et al., 2017; Ibarz et al., 2018), robotics (Finn et al., 2016; Brown et al., 2019; Shin et al., 2023), and language models (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021; Ouyang et al., 2022; Liu et al., 2023).s Particularly in the recent drastic development of large language models (LLMs), RLHF has played a crucial role in fine-tuning models to better align with human preferences (Ouyang et al., 2022). However, despite the notable empirical success of these algorithms, a significant gap remains in the theoretical analysis of IRL and RLHF, limiting us to guarantee their reliability. This work aims to bridge this gap by proposing a novel theoretical framework for offline IRL and RLHF.

IRL aims to infer a reward function that aligns with an expert behavior from demonstrations (Ng & Russell, 2000; Abbeel & Ng, 2004). Typical IRL algorithms employ a bi-level optimization framework within the context of maximum likelihood estimation (MLE). In this framework, the inner optimization evaluates the policy based on the current reward parameters, while the outer optimization updates these parameters to better match observed expert behavior. These algorithms have been extensively explored in the literature (Ziebart et al., 2008; Wulfmeier et al., 2015; Zhou et al., 2017; Zeng et al., 2022; 2023), and their convergence is studied in both online settings (Zeng et al., 2022) and offline settings (Zeng et al., 2023).

Building upon IRL, RLHF recovers a reward function from human feedback data, which is commonly given by pairwise or K-wise trajectory comparisons. Despite growing interest in RLHF, only a few algorithms provide optimality guarantees with sample complexity bound. Most existing approaches adopt an MLE framework, assuming that human evaluators follow a presupposed preference model, such as the Bradley-Terry-Luce (BTL) model. The reward parameters are then fine-tuned to maximize the log-likelihood of the collected offline preference data (Christiano et al., 2017). Recently, (Zhu et al., 2023; Zhan et al., 2023; Li et al., 2023) have proposed offline RLHF algorithms with optimality guarantees by adopting a pessimistic mechanism from offline reinforcement learning (RL) theory. (Park et al., 2024) further suggested a personalized setting that learns

multiple reward models for heterogeneous preferences.

While the MLE-based approach offers a solid theoretical framework for both IRL and RLHF, it comes with unavoidable limitations. Particularly in IRL, bi-level optimization algorithms face computational challenges due to their nested-loop structures. In addition, MLE-based algorithms rely on a specific decision-making or preference model they employ. For example, the IRL algorithm proposed by (Zeng et al., 2022) learns the reward function that aligns with an expert policy, which is assumed to be an optimal softmax policy. Furthermore, RLHF algorithms (e.g. (Zhu et al., 2023; Zhan et al., 2023)) assume a preference model for human evaluator, which might not fully capture the complex and diverse nature of real-world human preferences. Consequently, their optimality guarantees might be compromised if there exists a mismatch between actual human preferences and the model in use.

In response to these challenges inherent in MLE frameworks, recent research has shifted focus from estimating a single reward function (under a presupposed model) to recovering a *feasible reward set*, a set of rewards where the expert policy is (near) optimal. Notably, (Metelli et al., 2021; 2023) estimated the feasible reward set from finite-horizon Bellman equations and provided sample complexity bounds associated with estimation errors. However, their algorithm requires a generative model of state transition probabilities. (Lindner et al., 2022) mitigated this requirement by adopting an efficient exploration policy for sampling trajectories, though it remains in an online setting. More recently, and concurrent with our work, (Zhao et al., 2023) introduced the first offline algorithm with a theoretical guarantee. They introduce a pessimistic mechanism to address the issue of non-uniform data coverage, penalizing state-action pairs with low visitation frequency. Nevertheless, these penalty functions are nonlinear and non-convex, resulting in a non-convex reward set. This could limit flexibility for applications, especially when selecting a specific reward function within the set.

In line with this evolving perspective, we aim to obtain a *convex estimate* of a feasible reward set in an offline setting. To achieve this, we leverage recent advancements in the Linear Programming (LP) framework within the domain of offline RL. A fundamental challenge in offline RL is the non-uniform coverage of offline data (Fujimoto et al., 2019; Kumar et al., 2020). To address this issue, recent algorithms have employed a pessimistic mechanism that conservatively selects the value function or model within an uncertainty set (Liu et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021; Uehara & Sun, 2021; Chen & Jiang, 2022). However, these pessimistic approaches often introduce non-convex optimization problems, which can be intractable. In the latest research, a series of works (Zhan et al., 2022; Rashidinejad et al., 2022; Ozdaglar et al., 2023) have introduced LP-based methods that relax data coverage assumptions and provide tractable algorithms suitable for function approximation by introducing convex formulations.

Given the success of LP-based approaches in offline RL, investigating how it could address non-convexity and non-uniform data coverage issues in offline IRL presents a promising research direction. One notable advantage of LP is its flexibility in addressing intrinsic challenges in reward learning, such as avoiding undesirable degenerate solutions like $r = \mathbf{0}$. We demonstrate that a polyhedral estimate of the feasible reward set, provided by LP, offers efficient ways to identify a non-degenerate reward function. For example, it allows to select a reward function that maximizes the reward gap between the expert policy and suboptimal policies (e.g., uniform policy) over the solution set. We also highlight LP's suitability for function approximation, primarily due to its linear structure, which can further reduce the solution set and computational complexity. Furthermore, the LP framework enables the integration of extra information. As a notable example, we show that RLHF data can be incorporated by simply adding linear constraints, maintaining computational tractability and sample efficiency.

Our main contributions can be summarized as follows:

- We present an LP formulation for offline IRL that directly estimates the feasible reward set by the primal-dual optimality conditions in an empirical LP formulation of Markov decision process (MDP) (Section 3.1).

- In Theorem 3.3, the optimality of the estimated reward set is provided such that any reward function within this set ensures the expert policy is $\tilde{O}(\sqrt{|S||A|/N})$-suboptimal, under appropriate data coverage assumption.

- In offline RLHF, we align reward functions with pairwise trajectory comparison data using linear constraints (Section 4.1). In Theorem 4.1, we provide the generalization guarantee of the estimated reward function for unseen trajectory pairs.

- We address the potential degeneracy issue in reward learning (Section 3.3) and propose a unified framework that effectively combine IRL and RLHF to mitigate the degeneracy (Section 4.3).

- The proposed LP algorithm and the MLE algorithms in the literature are compared through numerical experiments (Section 5). We also provide an analytical example in offline RLHF, where MLE algorithm fails, while our LP approach succeeds to identify the optimal policy (Appendix F)

Additional related literature is discussed in Appendix A.

2. Preliminaries

2.1. Markov Decision Process (MDP)

We first revisit the standard notations for a tabular infinite-horizon discounted Markov decision process (MDP). An MDP $\mathcal{M}$ is represented as a tuple $\mathcal{M} = (S, A, P, \gamma, \mu_0, r)$, where S and A represent finite state and action spaces, $P : (S, A) \mapsto \Delta(S)$ denotes the transition probability function, $\gamma \in (0, 1)$ represents the discount factor, and $\mu_0 \in \Delta(S)$ is the initial state distribution. The reward function $r(s, a) : S \times A \mapsto [-1, 1]$ indicates the reward received for taking the action a in state s.

The primary objective in MDP is to identify a stochastic policy $\pi : S \mapsto \Delta(A)$ that maximizes the expected cumulative reward: $\mathbb{E}^\pi[\sum_{h=0}^\infty \gamma^h r(s_h, a_h)|s_0 \sim \mu_0]$. We define $v^\pi(s)$ as the expected total discounted reward received when initiating from state s and following π, such that $v^\pi(s) := \mathbb{E}^\pi \left[\sum_{h=0}^\infty \gamma^h r(s_h, a_h) \mid s_0 = s \right]$. Then, the optimal policy π^* maximizing the expected reward and its corresponding value function $v^* := v^{\pi^*}$ are related by the Bellman equation

$$v^*(s) = \max_{a \in A} \left\{ r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a)v^*(s') \right\}, \quad (1)$$

which holds for any $s \in S$. It is well-established that v^* can be determined by solving the following linear programming (LP) as outlined in (Puterman, 2014):

$$\min_{v \in \mathbb{R}^{|S|}} (1 - \gamma)\mu_0^\top v \quad \text{s.t.} \quad M^\top v - r \geq 0. \quad (2)$$

The matrix $M \in \mathbb{R}^{|S|^2|A|}$ is defined as $M(s', (s, a)) := \mathbf{1}_{\{s'=s\}} - \gamma P(s'|s, a)$, where $\mathbf{1}_{\{s'=s\}}$ denotes the indicator function for the case $\{s' = s\}$. Throughout the paper, we treat the above LP as the primal LP, and v as the primal optimization variable. Then, the dual LP is expressed as

$$\max_{d \in \mathbb{R}^{|S||A|}} r^\top d \quad \text{s.t.} \quad Md = (1 - \gamma)\mu_0, \quad d \geq 0. \quad (3)$$

The dual variable d, often interpreted as an *occupancy measure* or a discounted state-action visitation frequency in RL literature, is related to a policy π by

$$d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^\infty \gamma^h P_{\mu_0}^\pi(s_h = s, a_h = a). \quad (4)$$

Here, $P_{\mu_0}^\pi(s_h = s, a_h = a)$ represents the probability of $(s_h, a_h) = (s, a)$, given $s_0 \sim \mu_0$ and $a_{h'} \sim \pi(s_{h'})$ for all $h' \geq 0$. The dependence on γ and μ_0 is omitted in the notation d^π for simplicity. A more detailed relationship between π and d^π can be found in (Puterman, 2014).

2.2. Offline IRL: Learning from Expert Trajectories

While RL learns a policy to maximize rewards in a given environment, IRL aims to infer the underlying reward function

that drives observed behavior. In the standard IRL setting, a single expert agent collects trajectory (roll-out) samples, and the reward function is recovered from these samples. We denote the true expert policy and corresponding occupancy measure as π_e and d_e, respectively, where d_e is defined as $d_e := d^{\pi_e}$. Our objective is to learn a reward function r such that the occupancy measure d_e is (near) optimal, utilizing the offline dataset gathered from the expert policy π_e.

In offline setting, the true values of the expert policy π_e and the transition probability P are unknown. Instead, we have access to a static, pre-collected dataset $\mathcal{D}_{\text{IRL}}$ composed of N independent and identically distributed (i.i.d.) trajectory samples:

$$\mathcal{D}_{\text{IRL}} = \{\tau^n = (s_0^n, a_0^n, s_1^n, \ldots, s_{H-1}^n, a_{H-1}^n, s_H^n)\}_{n=1}^N. \quad (5)$$

Note that the sampling distribution is fully determined by μ_0, P, and π_e. Let $N_h(s, a)$ and $N_h(s, a, s')$ be the counts of n satisfying $(s_h^n, a_h^n) = (s, a)$ and $(s_h^n, a_h^n, s_{h+1}^n) = (s, a, s')$ in the dataset, respectively. Using these counts, we estimate the occupancy measure d_e as follows:

$$\hat{d}_e(s, a) = (1 - \gamma)\frac{1}{N} \sum_{h=0}^{H-1} \gamma^h N_h(s, a) \quad \forall (s, a) \in S \times A. \quad (6)$$

Using this empirical estimate $\hat{d}_e$, we aim to develop an LP formulation that identifies a reward function which ensures the optimality of d_e with an acceptable level of error.

2.3. Offline RLHF: Learning from Pairwise Trajectory Comparisons

We extend our LP framework to address the offline RLHF problem. Our primary objective is to derive a reward function that is aligned with pairwise trajectory comparison data, provided by human evaluators. We denote each comparison query as q_n, with the index n ranging from 1 to N_q. Each query comprises a pair of trajectories, such that $\tau^{n,1} = (s_0^{n,1}, a_0^{n,1}, \ldots, s_{H-1}^{n,1}, a_{H-1}^{n,1}, s_H^{n,1})$ and $\tau^{n,2} = (s_0^{n,2}, a_0^{n,2}, \ldots, s_{H-1}^{n,2}, a_{H-1}^{n,2}, s_H^{n,2})$. We assume $\tau^{n,1}$ and $\tau^{n,2}$ are sampled i.i.d. according to the sampling distribution μ_{HF}. In each query, a human evaluator is presented with both trajectories and asked to select the one they prefer. We denote the event where trajectory $\tau^{n,1}$ is preferred over $\tau^{n,2}$ by the variable $y^n = 1$, and conversely, $y^n = 2$ indicates the event where $\tau^{n,2}$ is favored over $\tau^{n,1}$. The human feedback dataset is then represented by $\mathcal{D}_{\text{HF}} = \{(\tau^{n,1}, \tau^{n,2}, y^n)\}_{n=1}^{N_q}$.

Given the dataset $\mathcal{D}_{\text{HF}}$, we design an LP formulation to identify a reward function r that aligns well with $\mathcal{D}_{\text{HF}}$. Notably, this LP approach is purely data-driven, relying solely on the observed comparisons without assuming any specific preference model associated with human evaluators. This

aspect distinguishes it from previous MLE algorithms for offline RLHF. The detailed comparison will be elaborated in a later section of the paper.

3. Offline Inverse Reinforcement Learning

3.1. LP Formulation of Offline IRL

Recall the dual LP formulation of MDP presented in the previous section:

$$\max_{d \in \mathbb{R}^{|S||A|}} r^\top d \quad \text{s.t.} \quad Md = (1-\gamma)\mu_0, \quad d \geq 0. \quad (7)$$

IRL aims to find a feasible reward function r such that the expert occupancy measure d_e is an optimal solution to the above dual LP. As the feasible reward function is not unique, we define the feasible reward set as follows:

$$\mathcal{R} = \{r \in [-1,1]^{|S||A|} \mid d_e \text{ is optimal to (7)}\} \quad (8)$$

In the offline setting, recovering the ground-truth $\mathcal{R}$ is challenging since we only have access to the empirical estimate of d_e and the transition matrix P. Consequently, our goal is to recover an estimate of $\mathcal{R}$ in which d_e is a near-optimal solution to (7) for any r in this estimated set.

Marginal importance sampling. The primary challenge in offline RL and IRL is the non-uniform coverage of the offline dataset. To address this issue, we adopt the marginal importance sampling (MIS) framework in the literature (Nachum et al., 2019; Lee et al., 2021), which considers the scaled version of LP. First, we define the optimization variable $w_d \in \mathbb{R}^{|S||A|}$ as

$$w_d(s,a) := \begin{cases} \frac{d(s,a)}{d_e(s,a)} & \text{if} \quad d_e(s,a) > 0, \\ 0 & \text{if} \quad d_e(s,a) = 0. \end{cases} \quad (9)$$

w_d is a scaled dual variable, which represents the ratio between the target d and the expert d_e. The expert optimization variable is denoted by $w_e := w_{d_e}$, which satisfies $w_e(s,a) = \mathbf{1}_{\{d_e(s,a)>0\}}$ for all $(s,a) \in S \times A$ by the definition of w_d. Note that our algorithm will not require information about which state-action pairs have zero visitation frequency under π_e (i.e., $d_e(s,a) = 0$), since it will automatically set the reward to zero, i.e. $r(s,a) = 0$, if $\hat{d}_e(s,a) = 0$.

Next, we define $u \in \mathbb{R}^{|S||A|}$ and $K \in \mathbb{R}^{|S|^2|A|}$ as

$$\begin{aligned} u(s,a) &:= r(s,a)d_e(s,a), \\ K(s',(s,a)) &:= d_e(s,a)\mathbf{1}_{\{s=s'\}} - \gamma d'_e(s,a,s'), \end{aligned} \quad (10)$$

where $d'_e(s,a,s') := d_e(s,a)P(s'|s,a)$ for any (s,a,s'). In this MIS framework, u and K correspond to r and P, respectively. The following lemma shows this relationship clearly (see Lemma 1 in (Ozdaglar et al., 2023) for the proof).

Lemma 3.1. $r^\top d = u^\top w_d$ and $Md = Kw_d$ hold for any $d \in \mathbb{R}^{|S||A|}$.

Empirical LP formulation. By Lemma 3.1, the dual LP can be written with u, w, and K as follows:

$$\max_{w \in \mathbb{R}^{|S||A|}} u^\top w \quad \text{s.t.} \quad Kw = (1-\gamma)\mu_0, \quad w \geq 0. \quad (11)$$

We omit the subscript d in w_d for ease of notation. In the above LP formulation, our objective is to identify the set of u for which w_e is optimal. However, in the offline setting, the true values of K and d_e remain unknown. Therefore, our goal shifts towards constructing an empirical version of this LP. We first define the empirical estimate of u as $u_{\mathcal{D}}(s,a) := r(s,a)\hat{d}_e(s,a)$ for all $(s,a) \in S \times A$, and replace the objective function $u^\top w$ with $u_{\mathcal{D}}^\top w$. Next, we introduce $K_{\mathcal{D}} \in \mathbb{R}^{|S|^2|A|}$, an empirical estimate of K, defined as:

$$K_{\mathcal{D}}(s',(s,a)) := \hat{d}_e(s,a)\mathbf{1}_{\{s=s'\}} - \gamma\hat{d}'_e(s,a,s'), \quad (12)$$

where for any $(s,a,s') \in S \times A \times S$,

$$\hat{d}'_e(s,a,s') := (1-\gamma)\frac{1}{N}\sum_{h=0}^{H-1}\gamma^h N_h(s,a,s'). \quad (13)$$

However, directly substituting the empirical estimate $K_{\mathcal{D}}$ for K in the equality constraint $Kw = (1-\gamma)\mu_0$ can be problematic, as it may cause the target variable w_e being infeasible. Therefore, we opt to relax the equality constraint to an inequality constraint.

Let $X = [x_1, \cdots, x_{N_X}] \in \mathbb{R}^{|S| \times N_X}$ be a coefficient matrix for the relaxation, where $\|x_i\|_\infty \leq 1$ for all $i \in \{1, \ldots, N_X\}$. Let $\epsilon_x \in \mathbb{R}^{N_x}$ be a parameter that controls the level of relaxation. Then, we replace the equality constraint $Kw = (1-\gamma)\mu_0$ with the relaxed inequality constraint $X^\top(K_{\mathcal{D}}w - (1-\gamma)\mu_0) \leq \epsilon_x$. One applicable choice of the coefficient matrix X would be a matrix that contains all $2^{|S|}$ binary (sign) vectors $[\pm 1, \pm 1, \ldots, \pm 1]$ in its columns. Then, the inequality constraint is equivalent to the $\mathcal{L}^1$ norm constraint, i.e. $\|K_{\mathcal{D}}w - (1-\gamma)\mu_0\|_1 \leq \epsilon_x$.

With this relaxation, the empirical version of the dual LP can be expressed as

$$\begin{aligned} \max_{w \in \mathbb{R}^{|S||A|}} \quad & u_{\mathcal{D}}^\top w \\ \text{s.t.} \quad & X^\top(K_{\mathcal{D}}w - (1-\gamma)\mu_0) \leq \epsilon_x, \quad w \geq 0. \end{aligned} \quad (14)$$

Additionally, the dual of (14) can be expressed as

$$\begin{aligned} \min_{v \in \mathbb{R}^{N_x}} \quad & (1-\gamma)\mu_0^\top Xv + \epsilon_x^\top v \\ \text{s.t.} \quad & K_{\mathcal{D}}^\top Xv \geq u_{\mathcal{D}}, \quad v \geq 0, \end{aligned} \quad (15)$$

where v is an optimization variable.

Feasible reward set estimation. Under the empirical LP formulations, our goal is to estimate the set of u such that w_e is (near) optimal to (14). Consider the primal-dual optimality conditions of (v, w) under a reward function u:

(Primal feasibility) : $K_{\mathcal{D}}^\top X v \geq u, \ v \geq 0,$

(Dual feasibility) : $X^\top(K_{\mathcal{D}} w - (1-\gamma)\mu_0) \leq \epsilon_x, \ w \geq 0,$

(Zero duality gap) : $(1-\gamma)\mu_0^\top X v + \epsilon_x^\top v = u^\top w.$
$$(16)$$

w is dual-optimal under u if and only if the above optimality conditions hold with some v.

Consequently, the feasible reward set can be estimated by identifying (u, v) pairs for which (u, v, w_e) satisfies (16). Here, we further relax the zero duality gap condition with the slack parameter $\epsilon_g \geq 0$ for two reasons. First, the true reward might not satisfy this equality constraint due to errors in empirical estimation of d_e and K. Second, we are also interested in the reward such that π_e is near-optimal. Therefore, we consider the following polyhedron as an estimate of the feasible reward set:

$$\hat{\mathcal{R}}_{\mathrm{IRL}}(\epsilon_g) := \{(u, v) \mid \underbrace{(1-\gamma)\mu_0^\top X v + \epsilon_x^\top v - u^\top \mathbf{1} \leq \epsilon_g}_{(i)},$$
$$\underbrace{K_{\mathcal{D}}^\top X v \geq u, \ v \geq 0}_{(ii)}, \ \underbrace{-\hat{d}_e \leq u \leq \hat{d}_e}_{(iii)}\}.$$
$$(17)$$

The constraint (i) denotes the upper bound on the duality gap, where ϵ_g is used as the parameter. (ii) represents primal feasibility condition and (iii) bounds the reward r to the range $[-1, 1]$. The vector $\mathbf{1}$ (vector of all ones) is used instead of w_e in (i), since $u^\top \mathbf{1} = u^\top w_e$ holds by the definition of u and w_e. Note that the dual feasibility condition is not required in $\hat{\mathcal{R}}_{\mathrm{IRL}}(\epsilon_g)$ because it is a condition for the constant value w_e.

3.2. Optimality Guarantee for Offline IRL

In this section, we analyze the statistical error involved in the estimate $\hat{\mathcal{R}}_{\mathrm{IRL}}$ of the feasible reward set $\mathcal{R}$. Before presenting the main results, we address the data coverage issue in our setting. In offline RL, distribution mismatch between the target policy and the behavior policy causes the inaccurate policy evaluation, and the concentrability-type assumption is required for an optimality guarantee. In offline IRL, since the behavior policy is identical to the expert policy, the reward estimation can be inaccurate for state-actions pairs where the occupancy measure $d_e(s, a)$ is small. We address this issue by defining the confident set of occupancy measures.

Confidence set. The confidence set if defined as the intersection of a set of valid occupancy measure (under MDP

$\mathcal{M}$) and an $\mathcal{L}^\infty$ norm ball with radius B:

$$D_B := \left\{ d \in \mathbb{R}_+^{|S||A|} \mid Md = (1-\gamma)\mu_0, \ \|w_d\|_\infty \leq B \right\}.$$
$$(18)$$

The radius $B \geq 1$ is a parameter that controls the conservativeness of the algorithm.

The set D_B includes all possible occupancy measures d if $B \geq d_{\min}^{-1}$, where

$$d_{\min} := \min_{(s,a) \in S \times A: \ d_e(s,a) \neq 0} d_e(s, a), \quad (19)$$

since $w_d(s, a) \leq d_{\min}^{-1}$ for any $d \in \Delta(S \times A)$ and $(s, a) \in S \times A$ by the definition of w_d (9). In this case, optimality over the set D_B implies global optimality.

It is worth highlighting that setting $B = d_{\min}^{-1}$ yields results comparable to those in recent works (Metelli et al., 2023; Zhao et al., 2023). The error bounds in these works depend on the constant $\pi_{\min}^{-1}$, defined as

$$\pi_{\min} := \min_{(s,a) \in S \times A: \ \pi_e(a|s) \neq 0} \pi_e(a|s). \quad (20)$$

Under a fixed μ_0, the value of $B = d_{\min}^{-1}$ is upper bounded by (constant) $\times \pi_{\min}^{-1}$, by the following inequality:

$$d_{\min} = d_e(s', a') = \pi_e(a'|s') \sum_{a \in A} d_e(s', a)$$
$$\geq (1-\gamma)\mu_0(s')\pi_{\min}, \quad (21)$$

where $(s', a') \in S \times A$ is a state-action pair that achieves the minimum in $d_{\min}$. Thus, setting $B = d_{\min}^{-1}$ can provide error bounds comparable to those in other works while ensuring global optimality.

Our goal is to establish the optimality of d_e within the confidence set D_B. The proof comprises two distinct steps. Firstly, we establish that $w_{\tilde{d}}$ is feasible to the dual empirical LP (14) with high probability for any $\tilde{d} \in D_B$, under appropriate level of relaxation ϵ_x. Next, we show that w_e has a (nearly) higher objective than $w_{\tilde{d}}$ with high probability since w_e has a small duality gap. In the following lemma, we prove that for any $\tilde{d} \in D_B$, corresponding $w_{\tilde{d}}$ is a feasible solution to the empirical LP with high probability under appropriate relaxation level ϵ_x in the constraint.

Lemma 3.2. *In dual empirical LP* (14), *let*

$$\epsilon_x = \left(B(1+\gamma)\gamma^H + B(1+\gamma)(1-\gamma^H) \right.$$
$$\left. \times \sqrt{\frac{2|S||A|}{N} \log \frac{2N_x}{\delta}} \right) \mathbf{1}, \quad (22)$$

where $\delta > 0$. Then, for any $\tilde{d} \in D_B$, $w_{\tilde{d}}$ is feasible to (14) *with probability at least $1 - \frac{\delta}{2^{|S||A|}}$.*

Proof. See Appendix B. □

From the above Lemma, we establish the optimality guarantee in the following theorem. Specifically, in words, under the reward function r recovered from the set (17), we show that d_e is an $\tilde{O}(\sqrt{|S||A|/N})$-suboptimal solution over the confidence set D_B with high probability.

Theorem 3.3. *Suppose $(u_{\mathcal{D}}, v_{\mathcal{D}}) \in \hat{\mathcal{R}}_{IRL}(\epsilon_g)$, with the relaxation level ϵ_x specified in Lemma 3.2. Let r satisfy*

$$r(s,a) = \frac{u_{\mathcal{D}}(s,a)}{\hat{d}_e(s,a)} \tag{23}$$

for all $(s,a) \in S \times A$, following the convention $0/0 = 0$. Then, we have

$$\mathbb{P}(r^\top d_e \geq r^\top \tilde{d} - \epsilon, \ \ \forall \tilde{d} \in D_B) \geq 1 - 3\delta, \tag{24}$$

where

$$\begin{aligned}
\epsilon = \epsilon_g + (1+B)\gamma^H + (1-\gamma^H)\sqrt{\frac{2}{N}\log\frac{1}{\delta}} \\
+ B(1-\gamma^H)\sqrt{\frac{2|S||A|}{N}\log\frac{2}{\delta}}.
\end{aligned} \tag{25}$$

Proof. See Appendix C. □

Sample complexity analysis. The proposed solution set achieves the $\tilde{O}(B(1-\gamma^H)\sqrt{|S||A|/N})$ sample complexity bound with additional error terms ϵ_g and $(1+B)\gamma^H$. Note that the parameter ϵ_g can be set to $\epsilon_g = \tilde{O}(1/\sqrt{N})$ to match the sample complexity. The term $(1+B)\gamma^H$ diminishes exponentially with the horizon of the collected trajectory data, which underscores the requirement for long-horizon data to ensure accurate estimation. To the best of our knowledge, besides our work, (Zeng et al., 2023) is the only other study that offers an optimality guarantee for the offline IRL problem under a discounted MDP, with a sample complexity of $\tilde{O}(1/\sqrt{N})$. However, as their algorithm is based on a bi-level optimization approach and their error bound is given for the log-likelihood function, a direct comparison of their result with ours is not feasible. The concurrent work by (Zhao et al., 2023) presents a comparable sample complexity for an episodic MDP. We provide a detailed comparison in Appendix D.

Trade-off between optimality and feasibility. In our formulation, there exists a trade-off between the optimality of the policy and the feasibility of the reward function, which is modulated by the parameter ϵ_g. The duality gap bound, denoted as ϵ_g, adjusts the size of the solution set $\hat{\mathcal{R}}_{IRL}(\epsilon_g)$; this set expands with an increase in ϵ_g. ϵ_g can be reduced to 0 without causing infeasibility, as the set $\hat{\mathcal{R}}_{IRL}(\epsilon_g)$ is always non-empty due to the trivial solution $(u,v) = (0,0)$. A smaller value of ϵ_g enhances the optimality of the expert policy π_e, as stated in Theorem 3.3. However, excessively reducing ϵ_g can lead to overly greedy choices, resulting in trivial or degenerate solutions. The impact of varying ϵ_g is demonstrated through numerical experiments in Section 5.

Function approximation. The proposed LP formulation is well-suited for function approximation (parameterization), which allows us to reduce both the computational cost and the size (dimension) of the solution set. Consider the parameterization of the variable (u,v) as (u_θ, v_θ), where θ represents a parameter within the parameter space $\Theta \subset \mathbb{R}^k$, which we aim to explore. If there exists a $\theta \in \Theta$ such that $(u_\theta, v_\theta) \in \hat{\mathcal{R}}_{IRL}(\epsilon_g)$, then the optimality guarantee provided in Theorem 3.3 is preserved for the reward function recovered from u_θ, while the computational complexity of the LP can be reduced to a polynomial in k, down from $|S||A|$. It is important to note that this formulation remains a linear (or convex) program under linear (or convex) parameterization. If non-convex function approximation is employed for high-dimensional or continuous state-action spaces, an efficient algorithm for solving the proposed optimization may be required; however, such an extension is beyond the scope of this work, and we defer this to future research.

3.3. Degeneracy Issue in Reward Learning

In the practical applications of reward learning, estimating the feasible reward set is not enough; we need to select a single reward function within the estimated feasible reward set to use. This is not a trivial problem due to existence of degenerate reward functions in the feasible reward set. Degenerate reward functions (e.g. $r = \mathbf{0}$), though theoretically feasible, are practically undesirable as they fail to separate the expert policy π_e from others. In our solution set $\hat{\mathcal{R}}_{IRL}(\epsilon_g)$ (17), degeneracy in the feasibility constraint $K_{\mathcal{D}}^\top X v \geq u$ is critical. If equality holds for some state-action pairs such that $(K_{\mathcal{D}}^\top X v - u)(s,a) = 0$, then the complementary slackness condition will not be violated by changing the value of $w_e(s,a)$, meaning that w_e may not be uniquely optimal. We suggest a simple and tractable method to obtain a non-degenerate reward function in $\hat{\mathcal{R}}_{IRL}(\epsilon_g)$.

Utilizing suboptimal trajectory samples. A straightforward approach to obtain a non-degenerate solution is utilizing a suboptimal policy π_{sub}. To be specific, we directly maximize the expected reward gap between the expert policy π_e and the suboptimal π_{sub}. A viable example of π_{sub} is a uniformly random policy such as $\pi_{\text{sub}}(a|s) = \frac{1}{|A|}$ $\forall(s,a) \in S \times A$, because this policy is unlikely to be optimal unless the expected rewards are uniform over all actions. To maximize the reward gap, we sample suboptimal trajectories with π_{sub} and estimate the occupancy measure of π_{sub} as $\hat{d}_{\text{sub}}$, using the sampling and estimation methods

discussed previously. Then, we maximize the empirical mean of the reward gap as per the following LP:

$$\max_{r,u,v} \quad r^\top(\hat{d}_e - \hat{d}_{\text{sub}})$$
$$\text{s.t.} \quad (u,v) \in \hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g), \quad u = \hat{d}_e \circ r. \quad (26)$$

Here, $\circ$ denotes the element-wise (Hadamard) product. The numerical experiments in Section 5 demonstrate that the above formulation efficiently recovers a non-degenerate reward function with only a small number of suboptimal trajectory samples.

4. Offline Reinforcement Learning from Human Feedback

4.1. LP Formulation of Offline RLHF

In this section, we extend our LP framework to address offline RLHF problem. As discussed in Section 2.3, our focus is on minimizing the error associated with the reward r and the human feedback data $\mathcal{D}_{\text{HF}}$. We begin by representing the cumulative reward of each trajectory $\tau^{n,i}$ ($i = 1, 2$) in the dataset $\mathcal{D}_{\text{HF}}$ as a linear function of the reward r. Specifically, the cumulative reward from the trajectory $\tau^{n,i}$ can be expressed as $r(\tau^{n,i}) = r^\top \psi^{n,i}$, where each vector $\psi^{n,i} \in \mathbb{R}^{|S||A|}$ can be mapped from the trajectory $\tau^{n,i}$ by

$$\psi^{n,i}(s,a) := \sum_{h=0}^{H-1} \gamma^h \mathbf{1}_{\{s_h^{n,i}=s, a_h^{n,i}=a\}}, \quad (27)$$

for any $(s,a) \in S \times A$. Following this, we define the error in the single data point $(\tau^{n,1}, \tau^{n,2}, y^n)$ associated with the reward function r as

$$\mathcal{L}(\tau^{n,1}, \tau^{n,2}, y^n; r) := r^\top(\psi^{n,2} - \psi^{n,1})\mathbf{1}_{\{y^n=1\}}$$
$$+ r^\top(\psi^{n,1} - \psi^{n,2})\mathbf{1}_{\{y^n=2\}}. \quad (28)$$

Note that naively minimizing the average or maximum error over queries might often lead to degenerate reward functions. This is because human evaluators sometimes provide conflicting feedback, such as $y^n = 1$ when $r^\top \psi^{n,2} > r^\top \psi^{n,1}$, due to their stochasticity. Under conflicting comparison data, minimizing $\mathcal{L}$ may result in degenerate solutions such as $r = \mathbf{0}$. To address this issue, we allows for a slack in the error $\mathcal{L}$ by introducing a parameter $\epsilon_r \in \mathbb{R}$, which controls the size of the solution set. Specifically, we define the solution set $\hat{\mathcal{R}}_{\text{HF}}$ as follows:

$$\hat{\mathcal{R}}_{\text{HF}}(\epsilon_r) := \{r \mid \mathcal{L}(\tau^{n,1}, \tau^{n,2}, y^n; r) \le \epsilon_r$$
$$\forall n = 1, 2, \dots, N_q, \quad r \in [-1,1]^{|S||A|}\}. \quad (29)$$

Under this adjustable solution set, if we have additional information, we could also apply the strategy discussed in the previous section for identifying non-degenerate solutions: maximizing the reward gap between the expert trajectories and the suboptimal trajectories.

One advantage of the proposed LP method is its robustness to different human evaluator preference models, in contrast to MLE-based algorithms. When the human evaluator deviates from the preference model assumed in MLE, the true reward parameter may diverge from the parameter space. However, the LP approach is not subject to this limitation. We illustrate this point with a bandit example in Appendix F, where our LP algorithm successfully finds an optimal policy, whereas the recent pessimistic MLE algorithm fail to do so.

4.2. Generalization Guarantee for Offline RLHF

Recent works in offline RLHF, such as (Zhu et al., 2023; Zhan et al., 2023), have proposed a pessimistic MLE algorithm and provided an error bound between the estimated and the true reward function of the supposed preference model. Our LP method does not offer an optimality guarantee in the same way as these works, as it obtains a set of reward functions without assuming a specific preference model. Instead, we analyze the generalization property of the obtained reward functions by examining how $r \in \hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$ aligns with unseen trajectory pairs sampled from μ_{HF}.

We first introduce the probabilistic preference model for a human evaluator generating feedback data. We emphasize that our proposed method is not dependent on this model; we introduce it to analyze a generalization property. Suppose that $y \in \{1, 2\}$ is sampled from a Bernoulli distribution with the probabilistic model $\mathbb{P}(y = 1 \mid \tau^1, \tau^2) = \Phi(r_{\text{true}}^\top(\psi^1 - \psi^2))$, where $\Phi : \mathbb{R} \mapsto [0, 1]$ is a monotonically non-decreasing function satisfying $\Phi(x) + \Phi(-x) = 1$ for all $x \in \mathbb{R}$. Φ represents the preference model of the evaluator, based on their personal reward function r_{true}. For example, if Φ is a sigmoid function, i.e. $\Phi(x) = 1/(1 + e^{-x})$, then the above probabilistic model is reduced to the Bradley-Terry-Luce (BTL) model (Christiano et al., 2017). In the following theorem, we provide a generalization guarantee of any reward functions r contained in the estimated solution set $\hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$. Specifically, for a random (unseen) trajectory pair (τ^1, τ^2) sampled from the sampling distribution μ_{HF} and the human feedback y sampled from the preference model Φ, we prove that the error $\mathcal{L}(\tau^1, \tau^2, y; r)$ is bounded by ϵ_r with high probability.

Theorem 4.1. *Suppose $r \in \hat{\mathcal{R}}_{HF}(\epsilon_r)$ and the human feedback data (τ^1, τ^2, y) is sampled i.i.d. from the joint distribution (μ_{HF}, Φ). Then, for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\mathcal{L}(\tau^1, \tau^2, y; r) \ge \epsilon_r\right) \le \sqrt{\frac{1}{2N_q}\log\frac{1}{\delta}} \quad (30)$$

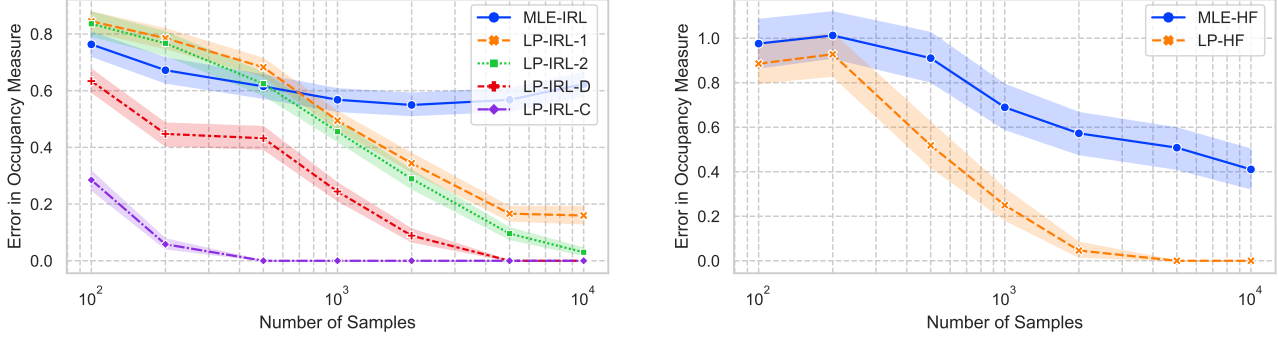*holds with probability at least $1 - \delta$.*

Figure 1. $\mathcal{L}^1$ error in the optimal occupancy measure under an estimated reward function. Left: Offline IRL algorithms; Right: Offline RLHF algorithms.

Proof. See Appendix E. ☐

Note that a trade-off between optimality and feasibility exists in $\hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$ with respect to the parameter ϵ_r, similar to $\hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g)$ in offline IRL. Specifically, while it is preferable to set the parameter ϵ_r as small as possible to avoid overly relaxing the RLHF constraint, excessively reducing ϵ_r can lead to feasibility issues. In practice, if any prior knowledge about the preference model Φ is available, this information can guide the selection of ϵ_r. If no such information is available, ϵ_r can be determined experimentally by starting with a large value and gradually reducing it until the reward set becomes trivial or infeasible.

4.3. Integration of IRL and RLHF

Finally, our LP framework facilitates the integration of two types of expert data: IRL (trajectories collected from the expert policy) and RLHF (pairwise trajectory comparisons). This is a unique feature of our LP framework, one that remains unexplored in the MLE framework. We propose to recover the reward function r from the intersection of two sets $\hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g)$ and $\hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$ such that

$$\hat{\mathcal{R}}_{\text{IRL-HF}}(\epsilon_g, \epsilon_r) =$$
$$\{(r, u, v) \mid (u, v) \in \hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g), \ r \in \hat{\mathcal{R}}_{\text{HF}}(\epsilon_r), \ u = \hat{d}_e \circ r\}. \quad (31)$$

In this combined formulation, the IRL constraint $(u, v) \in \hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g)$ provides the optimality guarantee of the expert policy, while the RLHF constraint $r \in \hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$ reduces the solution set and mitigates degeneracy by imposing additional constraints.

Strict trajectory comparison. If we can impose a strict reward gap between two trajectories, then the RLHF constraint can mitigate degeneracy more effectively. For instance, the constraint $r(\tau^1) \geq r(\tau^2) + \delta$ eliminates degenerate solutions that satisfy $r(\tau^1) = r(\tau^2)$ from the solution

set, if we can set a strict reward gap $\delta > 0$ based on human feedback data. To formalize this, we extend our approach to include the continuous feedback case, wherein the feedback is given as a continuous variable, $y \in [-1, 1]$, instead of the discrete variable used in previous sections. In Appendix G, we specify this continuous feedback setting, and suggest how we can extend the definition of the solution set $\hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$ as well as the generalization guarantee provided in Theorem 4.1. Additionally, in the next section, we compare the effects of discrete and continuous human feedback through numerical experiments.

5. Numerical Experiments

In this section, we demonstrate the performance of our LP algorithms through numerical experiments, comparing them to MLE algorithms in the literature. We consider an MDP with $|S| = 10$, $|A| = 2$, and $\gamma = 0.95$. In each experimental run, P and μ_0 are randomly selected. To introduce additional complexity to the problem, we have set the true rewards to have similar values: $r_{\text{true}}(s, a_1) = 1.0$ and $r_{\text{true}}(s, a_2) = 0.9$ for all states $s \in S$. The performance of each algorithm is assessed by measuring the proximity of an optimal occupancy measure under the true reward r_{true} and the estimated reward function $\hat{r}$. Specifically, we report $\|d^*(r_{\text{true}}) - d^*(\hat{r})\|_1$, which represents the $\mathcal{L}^1$ error between the optimal occupancy measures under r_{true} and $\hat{r}$. In each experiment, we sample N trajectories with a horizon of $H = 20$ according to π_e in IRL, and μ_{HF} in RLHF. For each sample size N, we conducted 200 experiments and reported the mean and standard deviation of the error. See Appendix H for detailed parameters and algorithms used in the experiments.

Offline IRL. The left side of Figure 1 compares the errors associated with each IRL algorithm. The results indicate that our LP-based algorithms generally outperform the bi-level optimization-based MLE algorithm (Zeng et al., 2023), demonstrating that LP is more sample-efficient in address-

ing ambiguity in the dynamics and the expert policy. The solution set with a smaller relaxation level $\epsilon_g = 0.001/\sqrt{N}$ (LP-IRL-2) exhibits better performance than that with a greater relaxation level $\epsilon_g = 0.01/\sqrt{N}$ (LP-IRL-1). This is consistent with the optimality-feasibility trade-off discussed in Section 3.2. Additionally, the integration of IRL and RLHF data leads to improved performance, as predicted. The use of continuous feedback (LP-IRL-C) is even more effective than discrete feedback (LP-IRL-D) by facilitating stricter constraints.

Offline RLHF. In numerical experiments for offline RLHF, the human feedback data is generated following the greedy model. The right side of Figure 1 compares the reward function obtained from LP (29) and the pessimistic MLE algorithm proposed by (Zhu et al., 2023) under the BTL model. In the LP algorithm, the error decreases rapidly as the number of samples increases, whereas the error in the MLE algorithm decreases more slowly. This result is consistent with the discussion in Appendix F, suggesting that the MLE algorithm might be inefficient or even fail if the human evaluator deviates from the assumed model, whereas LP does not.

6. Conclusion

In this paper, we have introduced a novel LP framework designed for offline reward learning. Our framework possesses several salient features, including (i) tractability and sample efficiency with an optimality guarantee, (ii) flexibility for extension due to its convex solution set, and (iii) robustness against diverse decision models. We believe our study opens a new avenue of research in the theories of offline reward learning. In the future, we aim to extend our framework to broader datasets, including those involving arbitrary sampling policies in IRL and K-wise comparisons in RLHF. Additionally, we plan to investigate the transferability of the estimated reward functions to similar environments.

Acknowledgments

Impact Statement

This paper aims to advance the understanding and efficiency of reward learning. We believe our work has the potential to benefit sequential decision-making systems in fields such as robotics, autonomous driving, and language models, though no immediate specific societal concerns are directly raised by our findings.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2004.

Brown, D., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, pp. 783–792. PMLR, 2019.

Chen, J. and Jiang, N. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pp. 378–388. PMLR, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58. PMLR, 2016.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34: 4028–4039, 2021.

Hoshino, H., Ota, K., Kanezaki, A., and Yokota, R. Opirl: Sample efficient off-policy inverse reinforcement learning via distribution matching. In *International Conference on Robotics and Automation*, pp. 448–454. IEEE, 2022.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. *Advances in Neural Information Processing Systems*, 31, 2018.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kostrikov, I., Nachum, O., and Tompson, J. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Li, Z., Yang, Z., and Wang, M. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, 2022.

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

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in Neural Information Processing Systems*, 33:1264–1274, 2020.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E., and Littman, M. L. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pp. 2285–2294. PMLR, 2017.

Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pp. 7665–7676. PMLR, 2021.

Metelli, A. M., Lazzati, F., and Restelli, M. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 24555–24591. PMLR, 2023.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 663–670. PMLR, 2000.

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

Ozdaglar, A. E., Pattathil, S., Zhang, J., and Zhang, K. Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning*, pp. 26769–26791. PMLR, 2023.

Park, C., Liu, M., Zhang, K., and Ozdaglar, A. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rashidinejad, P., Zhu, H., Yang, K., Russell, S., and Jiao, J. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *International Conference on Machine Learning*, pp. 729–736. PMLR, 2006.

Ratliff, N. D., Silver, D., and Bagnell, J. A. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 27:25–53, 2009.

Shin, D., Dragan, A. D., and Brown, D. S. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

Sikchi, H., Saran, A., Goo, W., and Niekum, S. A ranking game for imitation learning. *Transactions on machine learning research*, 2023a.

Sikchi, H., Zheng, Q., Zhang, A., and Niekum, S. Dual rl: Unification and new methods for reinforcement and imitation learning. In *International Conference on Learning Representations*, 2023b.

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

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

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

Wulfmeier, M., Ondruska, P., and Posner, I. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

Zeng, S., Hong, M., and Garcia, A. Structural estimation of markov decision processes in high-dimensional state space with finite-time guarantees. *arXiv preprint arXiv:2210.01282*, 2022.

Zeng, S., Li, C., Garcia, A., and Hong, M. Understanding expertise through demonstrations: A maximum likelihood framework for offline inverse reinforcement learning. *arXiv preprint arXiv:2302.07457*, 2023.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

Zhao, L., Wang, M., and Bai, Y. Is inverse reinforcement learning harder than standard reinforcement learning? a theoretical perspective. *arXiv preprint arXiv:2312.00054*, 2023.

Zhou, Z., Bloem, M., and Bambos, N. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 63(9):2787–2802, 2017.

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

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008.

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

A. Related Work

LP and Duality Approach in IRL. One of the foundational works in IRL (Ng & Russell, 2000) introduced the concept of characterizing a set of reward functions for which a given policy is optimal using an LP formulation. This idea has been further developed in subsequent literature (Metelli et al., 2021; Lindner et al., 2022; Metelli et al., 2023; Zhao et al., 2023), as outlined in the introduction. Recently proposed practical offline imitation learning (IL) algorithms, including ValueDICE (Kostrikov et al., 2019), IQ-Learn (Garg et al., 2021), OPIRL (Hoshino et al., 2022), and ReCOIL (Sikchi et al., 2023b), address an occupancy matching problem that minimizes the statistical divergence between the learner and the expert distribution. These algorithms exploit the duality of the LP formulation to obtain tractable algorithms, as extensively discussed in (Sikchi et al., 2023b). Despite the practical success of these algorithms, the resulting reward function and policy depend on the model in use, and they lack theoretical performance guarantees, such as provable sample efficiency.

RLHF without Preference Model Assumption. In offline RLHF, we impose a margin-based constraint on the solution set, which allows for the alignment of reward functions with preference data without assuming any preference models of human evaluators. The concept of employing a margin constraint originated in the early imitation learning literature. Specifically, maximum margin planning (MMP) (Ratliff et al., 2006; 2009) estimates the reward function such that the expert policy achieves a higher expected reward than all other policies by imposing a margin constraint in the reward optimization problem. Recently, (Sikchi et al., 2023a) introduced Rank-Game, a two-player game formulation between a policy agent, which optimizes the policy given a reward function, and a reward agent, which aligns the reward function with offline pairwise preference data. Their algorithm is model-free, as the reward agent minimizes ranking loss without relying on a specific preference model. A unification of demonstration and preference data is also proposed in their work, similar to our approach in the LP framework.

B. Proof of Lemma 3.2

For ease of notation, let $\delta' = \frac{\delta}{2|S||A|}$. To show

$$\mathbb{P}(X^\top(K_{\mathcal{D}}w_{\tilde{d}} - (1-\gamma)\mu_0) \leq \epsilon_x) \geq 1 - \delta', \tag{32}$$

we divide our proof into two parts. First, for any column x_i of X, we show that

$$x_i^\top(K_H - K)w_{\tilde{d}} \leq (1+\gamma)\gamma^H B \tag{33}$$

holds for the matrix K_H, which will be defined later. Next, we will prove that

$$x_i^\top(K_{\mathcal{D}} - K_H)w_{\tilde{d}} \geq \epsilon_{xi} - (1+\gamma)\gamma^H B \tag{34}$$

holds with probability less than δ'/N_x. Then, combining both inequalities yields $x_i^\top(K_{\mathcal{D}} - K)w_{\tilde{d}} = x_i^\top(K_{\mathcal{D}}w_{\tilde{d}} - (1-\gamma)\mu_0) \geq \epsilon_{xi}$ holds with probability less than δ'/N_x, since $Kw_{\tilde{d}} = M\tilde{d} = (1-\gamma)\mu_0$ holds by $\tilde{d} \in D_B$. Applying union bound to all columns x_i of X will lead to the conclusion.

To prove the first part, we first introduce the vector $d_e^H \in \mathbb{R}^{|S||A|}$, representing a finite-horizon truncation of d_e up to the horizon $H-1$:

$$d_e^H(s,a) := (1-\gamma)\sum_{h=0}^{H-1}\gamma^h P_{\mu_0}^{\pi_e}(s_h = s, a_h = a) \quad \forall(s,a) \in S \times A. \tag{35}$$

We also define the vector $d_e^{P,H} \in \mathbb{R}^{|S| \times |A| \times |S|}$ as the truncation of d_e' as follows:

$$d_e^{P,H}(s,a,s') := (1-\gamma)\sum_{h=0}^{H-1}\gamma^h P_{\mu_0}^{\pi_e}(s_h = s, a_h = a, s_{h+1} = s') \quad \forall(s,a,s') \in S \times A \times S. \tag{36}$$

Then, we define the matrix $K_H \in \mathbb{R}^{|S|^2|A|}$ using d_e^H and $d_e^{P,H}$ as follows:

$$K_H(s',(s,a)) := d_e^H(s,a)\mathbf{1}_{\{s=s'\}} - \gamma d_e^{P,H}(s,a,s') \quad \forall(s,a,s') \in S \times A \times S. \tag{37}$$

Since K_H can be considered as a finite-horizon truncation of K by its definition, only the terms from $h = H$ to ∞ remain in the matrix $K - K_H$. Consequently, we get the following inequalities that prove the first part:

$$
\begin{aligned}
|x_i^\top (K - K_H) w_{\tilde{d}}| &\le \sum_{s' \in S} |x_i(s')| \sum_{h=H}^{\infty} (1 - \gamma) \gamma^h \sum_{a \in A} \mathbb{P}_{\mu_0}^{\pi_e}(s_h = s', a_h = a) w_{\tilde{d}}(s', a) \\
&\quad + \sum_{s' \in S} |x_i(s')| \sum_{h=H}^{\infty} (1 - \gamma) \gamma^{h+1} \sum_{(s,a) \in S \times A} \mathbb{P}_{\mu_0}^{\pi_e}(s_h = s, a_h = a, s_{h+1} = s') w_{\tilde{d}}(s, a) \\
&\le \sum_{s' \in S} \sum_{h=H}^{\infty} (1 - \gamma) \gamma^h \mathbb{P}_{\mu_0}^{\pi_e}(s_h = s') B + \sum_{s' \in S} \sum_{h=H}^{\infty} (1 - \gamma) \gamma^{h+1} \mathbb{P}_{\mu_0}^{\pi_e}(s_{h+1} = s') B \\
&= \gamma^H B + \gamma^{H+1} B = (1 + \gamma) \gamma^H B.
\end{aligned}
\tag{38}
$$

The first inequality holds directly from the definitions of K and K_H, and the second inequality results from the assumptions $\|x_i\|_\infty \le 1$ and $\|w_{\tilde{d}}\|_\infty \le B$. Then, we have $x_i^\top (K_H - K) w_{\tilde{d}} \le (1 + \gamma) \gamma^H B$.

For the second step, consider the following definition of the random variable $z(\tau) \in \mathbb{R}$, where τ represents the finite-horizon trajectory sample:

$$
z(\tau) := \sum_{s' \in S} x_i(s') \sum_{(s,a) \in S \times A} w_{\tilde{d}}(s, a)(1 - \gamma) \times \sum_{h=0}^{H-1} \left[\gamma^h \mathbf{1}_{\{s_h = s, a_h = a\}} - \gamma^{h+1} \mathbf{1}_{\{s_h = s, a_h = a, s_{h+1} = s'\}} \right].
\tag{39}
$$

Then, $x_i^\top K_{\mathcal{D}} w_{\tilde{d}} = \frac{1}{N} \sum_{n=1}^{N} z(\tau^n)$ holds, implying that $x_i^\top K_{\mathcal{D}} w_{\tilde{d}}$ is the empirical mean of the random variable $z(\tau)$, derived from N trajectory samples τ^n. Meanwhile, $x_i^\top K_H w_{\tilde{d}}$ represents the expected value of $z(\tau)$ over τ. Moreover, we can show that the random variable $z(\tau)$ has a bounded range as follows:

$$
\begin{aligned}
|z(\tau)| &\le B(1 - \gamma) \sum_{s' \in S} \sum_{(s,a) \in S \times A} \sum_{h=0}^{H-1} \left[\gamma^h \mathbf{1}_{\{s_h = s, a_h = a\}} + \gamma^{h+1} \mathbf{1}_{\{s_h = s, a_h = a, s_{h+1} = s'\}} \right] \\
&\le B(1 - \gamma) \left(\frac{1 - \gamma^H}{1 - \gamma} + \frac{\gamma - \gamma^{H+1}}{1 - \gamma} \right) = B(1 + \gamma)(1 - \gamma^H),
\end{aligned}
\tag{40}
$$

where we used the assumptions $\|x_i\|_\infty \le 1$ and $\|w_{\tilde{d}}\|_\infty \le B$ in the first inequality. Therefore, we can apply Hoeffding's inequality as follows:

$$
\mathbb{P}\left(x_i^\top (K_{\mathcal{D}} - K_H) w_{\tilde{d}} \ge \epsilon \right) \le \exp\left(-\frac{N\epsilon^2}{2B^2(1 + \gamma)^2(1 - \gamma^H)^2} \right) \quad \forall \epsilon \ge 0.
\tag{41}
$$

Let $\epsilon = \sqrt{\frac{2B^2(1+\gamma)^2(1-\gamma^H)^2}{N} \log \frac{N_x}{\delta'}}$ and $\epsilon_{xi} = \epsilon + (1 + \gamma)\gamma^H B$. Then, the above inequality is equivalent to

$$
\mathbb{P}\left(x_i^\top (K_{\mathcal{D}} - K_H) w_{\tilde{d}} \ge \epsilon_{xi} - (1 + \gamma)\gamma^H B \right) \le \frac{\delta'}{N_x}.
\tag{42}
$$

Plugging the inequality $x_i^\top (K_H - K) w_{\tilde{d}} \le (1 + \gamma)\gamma^H B$ derived in the first part, we get

$$
\mathbb{P}\left(x_i^\top (K_{\mathcal{D}} - K_H) w_{\tilde{d}} + x_i^\top (K_H - K) w_{\tilde{d}} \ge \epsilon_{xi} \right) \le \mathbb{P}\left(x_i^\top (K_{\mathcal{D}} - K_H) w_{\tilde{d}} + (1 + \gamma)\gamma^H B \ge \epsilon_{xi} \right) \le \frac{\delta'}{N_x}.
\tag{43}
$$

Thus, $\mathbb{P}\left(x_i^\top (K_{\mathcal{D}} - K) w_{\tilde{d}} \ge \epsilon_{xi} \right) \le \delta'/N_x$. Taking the union bound to all events over $i = 1, 2, \ldots, N_x$,

$$
\mathbb{P}\left(X^\top (K_{\mathcal{D}} - K) w_{\tilde{d}} \ge \epsilon_x \right) \le \delta'.
\tag{44}
$$

Since $K w_{\tilde{d}} = M\tilde{d} = (1 - \gamma)\mu_0$ by $\tilde{d} \in D_B$, the above inequality is equivalent to

$$
\mathbb{P}\left(X^\top (K_{\mathcal{D}} w_{\tilde{d}} - (1 - \gamma)\mu_0) \ge \epsilon_x \right) \le \delta',
\tag{45}
$$

which completes the proof.

C. Proof of Theorem 3.3

The proof is comprised of three steps. In the first step, we employ a concentration bound to establish a limit on the difference between $u^\top \mathbf{1}$ and $u_\mathcal{D}^\top \mathbf{1}$, ensuring that $\mathbb{P}(u^\top \mathbf{1} - u_\mathcal{D}^\top \mathbf{1} \geq -\epsilon_{u1}) \geq 1 - \delta$ for a certain ϵ_{u1}. Next, from the feasibility of $w_{\tilde{d}}$ proven by Lemma 3.2 and using the optimality conditions in (17), we can show that $\mathbb{P}(u_\mathcal{D}^\top \mathbf{1} \geq u_\mathcal{D}^\top w_{\tilde{d}} - \epsilon_g, \ \forall \tilde{d} \in D_B) \geq 1 - \delta$ holds. The final step is to bound the difference between $u_\mathcal{D}^\top w_{\tilde{d}}$ and $u^\top w_{\tilde{d}}$, showing that $\mathbb{P}(u_\mathcal{D}^\top w_{\tilde{d}} - u^\top w_{\tilde{d}} \geq -\epsilon_{u2}, \ \forall \tilde{d} \in D_B) \geq 1 - \delta$ for a specific ϵ_{u2}. Combining these three results with the union bound completes the proof.

We first prove that $\mathbb{P}(u^\top \mathbf{1} - u_\mathcal{D}^\top \mathbf{1} \geq -\epsilon_{u1}) \geq 1 - \delta$ holds if we let $\epsilon_{u1} = \sqrt{\frac{2(1-\gamma^H)^2}{N} \log \frac{1}{\delta}} + \gamma^H$. Since $u^\top \mathbf{1} - u_\mathcal{D}^\top \mathbf{1} = r^\top d_e - r^\top \hat{d}_e = r^\top (d_e - d_e^H) + r^\top (d_e^H - \hat{d}_e)$, we bound two terms $r^\top (d_e - d_e^H)$ and $r^\top (d_e^H - \hat{d}_e)$ separately. First, $r^\top (d_e - d_e^H)$ can be bounded as

$$
\begin{aligned}
|r^\top (d_e - d_e^H)| &\leq (1-\gamma) \sum_{(s,a) \in S \times A} |r(s,a)| \sum_{h=H}^\infty \gamma^h P_{\mu_0}^\pi (s_h = s, a_h = a) \\
&\leq (1-\gamma) \sum_{(s,a) \in S \times A} \sum_{h=H}^\infty \gamma^h P_{\mu_0}^\pi (s_h = s, a_h = a) = \gamma^H.
\end{aligned}
\tag{46}
$$

Next, consider the random variable $z'(\tau)$ defined as

$$
z'(\tau) := (1-\gamma) \sum_{(s,a) \in S \times A} r(s,a) \sum_{h=0}^{H-1} \gamma^h \mathbf{1}_{\{s_h=s, a_h=a\}},
\tag{47}
$$

which represents the cumulative reward of τ multiplied by the constant $(1-\gamma)$. According to its definition, $r^\top d_e^H$ is the expected value of $z'(\tau)$ over τ, while $r^\top \hat{d}_e$ is the empirical mean of $z'(\tau)$ derived from the samples $(\tau^1, \tau^2, \ldots, \tau^N)$. Moreover, from its definition, we can easily show that $z'(\tau)$ lies in the interval $[-(1-\gamma^H), 1-\gamma^H]$. Thus, we can apply Hoeffding's inequality to bound the term $r^\top (d_e^H - \hat{d}_e)$ as follows:

$$
\mathbb{P}(r^\top (d_e^H - \hat{d}_e) \leq -\epsilon) \leq \exp\left(\frac{-N\epsilon^2}{2(1-\gamma^H)^2} \right) \quad \forall \epsilon \geq 0.
\tag{48}
$$

Letting $\epsilon = \epsilon_{u1} - \gamma^H$ yields $\mathbb{P}(r^\top (d_e^H - \hat{d}_e) \leq -\epsilon_{u1} + \gamma^H) \leq \delta$. Then, adding two results completes the first steps follows:

$$
\begin{aligned}
\mathbb{P}(u^\top \mathbf{1} - u_\mathcal{D}^\top \mathbf{1} \leq -\epsilon_{u1}) &= \mathbb{P}(r^\top (d_e^H - \hat{d}_e) + r^\top (d_e - d_H) \leq -\epsilon_{u1}) \\
&\leq \mathbb{P}(r^\top (d_e^H - \hat{d}_e) \leq -\epsilon_{u1} + \gamma^H) \leq \delta.
\end{aligned}
\tag{49}
$$

Next, by Lemma 3.2, $w_{\tilde{d}}$ is a feasible solution to (14) with probability at least $1 - \frac{\delta}{2^{|S||A|}}$. If $w_{\tilde{d}}$ is a feasible solution to (14), then the following inequalities hold:

$$
\begin{aligned}
u_\mathcal{D}^\top \mathbf{1} &\overset{(i)}{\geq} (1-\gamma)\mu_0^\top X v_\mathcal{D} + \epsilon_x^\top v_\mathcal{D} - \epsilon_g \\
&\overset{(ii)}{\geq} w_{\tilde{d}}^\top K_\mathcal{D}^\top X v_\mathcal{D} - \epsilon_g \\
&\overset{(iii)}{\geq} w_{\tilde{d}}^\top u_\mathcal{D} - \epsilon_g
\end{aligned}
\tag{50}
$$

The inequality (i) holds by the duality gap constraint, (ii) holds because $w_{\tilde{d}}$ is feasible to (14), and (iii) holds by the feasibility constraint $K_\mathcal{D}^\top X v_\mathcal{D} \geq u_\mathcal{D}$ and $w_{\tilde{d}} \geq 0$. Therefore, we get

$$
\mathbb{P}(u_\mathcal{D}^\top \mathbf{1} \geq u_\mathcal{D}^\top w_{\tilde{d}} - \epsilon_g) \geq 1 - \frac{\delta}{2^{|S||A|}}.
\tag{51}
$$

We then take union bound over all extreme points of D_B. Since D_B has at most $2^{|S||A|}$ extreme points, we get

$$
\mathbb{P}(u_\mathcal{D}^\top \mathbf{1} \geq u_\mathcal{D}^\top w_{\tilde{d}} - \epsilon_g, \ \forall \tilde{d} \in D_B) \geq 1 - \delta.
\tag{52}
$$

In addition, by similar steps to the first part of the proof, we can show that

$$\mathbb{P}(u_{\mathcal{D}}^\top w_{\tilde{d}} - u^\top w_{\tilde{d}} \geq -\epsilon_{u2}) \geq 1 - \frac{\delta}{2^{|S||A|}}, \tag{53}$$

if we let $\epsilon_{u2} = B\sqrt{\frac{2(1-\gamma^H)^2|S||A|}{N}\log\frac{2}{\delta}} + B\gamma^H$. Taking union bound over all extreme points of $\tilde{d} \in D_B$ yields

$$\mathbb{P}(u_{\mathcal{D}}^\top w_{\tilde{d}} - u^\top w_{\tilde{d}} \geq -\epsilon_{u2}, \ \forall\tilde{d} \in D_B) \geq 1 - \delta, \tag{54}$$

Taking union bound to the above three cases and using $u^\top \mathbf{1} = r^\top d_e$ and $u^\top w_{\tilde{d}} = r^\top \tilde{d}$ from Lemma 3.1, the conclusion holds as

$$\mathbb{P}(r^\top d_e \geq r^\top \tilde{d} - \epsilon, \ \forall\tilde{d} \in D_B) = \mathbb{P}(u^\top \mathbf{1} \geq u^\top w_{\tilde{d}} - \epsilon_g - \epsilon_{u1} - \epsilon_{u2}, \ \forall\tilde{d} \in D_B) \geq 1 - 3\delta. \tag{55}$$

D. Comparison to Pessimism-based Approach

The concurrent work by (Zhao et al., 2023) proposed an offline IRL algorithm for finite-horizon MDPs with comparable sample complexity, based on pessimistic value iteration. To be specific, they recover the mapping from the value and advantage functions to reward functions through Bellman iterations under estimated state transition probabilities. Though a direct comparison is limited since our work is developed for infinite-horizon discounted MDPs, there are some common structures between their algorithm and ours. Specifically, the value and advantage functions in their work can be considered as the primal optimization variable and the slack in the primal feasibility constraint in our formulation.

Nevertheless, there are differences between the resulting reward functions from their algorithm and ours. To address the uncertainty caused by non-uniform data coverage in the offline setting, they penalize the reward on uncertain state-action pairs that are less visited in the dataset. Such a pessimism-based reward estimation framework provides strong theoretical optimality guarantees, such as finite sample complexity bounds, similar to our approach. However, in contrast to our solution set, which is a polyhedron, the use of a nonlinear and non-convex penalty function in their reward model leads to a solution set that is also nonlinear and non-convex. This distinction makes our algorithm more flexible for any extension, such as function approximation and the integration of additional information.

E. Proof of Theorem 4.1

We first define the random variable

$$g(\tau^1, \tau^2, y) := \mathbf{1}_{\{\mathcal{L}(\tau^1, \tau^2, y; r) \geq \epsilon_r\}}(\tau^1, \tau^2, y), \tag{56}$$

where $\mathbf{1}_{\{\mathcal{L}(\tau^1, \tau^2, y; r) \geq \epsilon_r\}}(\tau^1, \tau^2, y)$ is the indicator function for the event that an error exceeds ϵ_r, i.e. $\mathcal{L}(\tau^1, \tau^2, y; r) \geq \epsilon_r$. The expected value of g can be expressed as

$$\bar{g} = \mathbb{E}_{(\tau^1, \tau^2, y)\sim(\mu_{\text{HF}}, \Phi)}[g(\tau^1, \tau^2, y)] = \mathbb{P}(\mathcal{L}(\tau^1, \tau^2, y; r) \geq \epsilon_r). \tag{57}$$

From the assumption that $r \in \hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)$, the empirical mean of g is given by 0:

$$\hat{g} = \frac{1}{N_q}\sum_{n=1}^{N_q} g(\tau^{n,1}, \tau^{n,2}, y^n) = 0. \tag{58}$$

From Hoeffding's inequality, we have $\mathbb{P}(\hat{g} - \bar{g} \leq -\epsilon) \leq e^{-2N_q\epsilon^2} = \delta$ if $\epsilon = \sqrt{\frac{1}{2N_q}\log\frac{1}{\delta}}$. Therefore, $\bar{g} \leq \epsilon$ holds with probability at least $1 - \delta$, which completes the proof.

F. Robustness of LP Framework

In this section, we discuss the robustness of the proposed LP framework compared to MLE framework, with respect to the different preference models of human evaluators. We first introduce the MLE framework in offline RLHF. Consider the

reward parameterization r_θ, where $\theta \in \Theta$ is a parameter and $\Theta \subset \mathbb{R}^k$ is a parameter space we aim to search the optimal parameter. Then, the standard MLE framework can be illustrated as maximizing the log-likelihood function as follows:

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \sum_{n=1}^{N_q} \log \left(\Phi \left(-\mathcal{L}(\tau^{n,1}, \tau^{n,2}, y^n; r_\theta) \right) \right). \tag{59}$$

Our LP framework finds the reward parameter in the solution set, such that $\hat{\theta}_{\text{LP}} \in \{\theta \in \Theta \mid r_\theta \in \hat{\mathcal{R}}_{\text{HF}}(\epsilon_r)\}$. Under the estimated reward parameters, we obtain the policy maximizing the reward function as follows:

$$\hat{\pi}_{\text{LP}} \in \arg\max_{\pi} \mathbb{E}_{s \sim d^\pi}[r_{\hat{\theta}_{\text{LP}}}(s, \pi(s))], \quad \hat{\pi}_{\text{MLE}} \in \arg\max_{\pi} \mathbb{E}_{s \sim d^\pi}[r_{\hat{\theta}_{\text{MLE}}}(s, \pi(s))]. \tag{60}$$

Pessimistic MLE. Recent works in offline RLHF, such as those by (Zhan et al., 2023; Zhu et al., 2023) have adapted the concept of pessimism from offline RL theory to address the data coverage issue. Specifically, these studies define a confidence set for the reward function and solve robust optimization problem to identify the policy that maximizes the worst-case reward within this set. For example, (Zhu et al., 2023) uses a semi-norm $\|\cdot\|_{\Sigma + \lambda I}$ as a metric for constructing the confidence set, where Σ represents the covariance of the comparison data and $\lambda > 0$ is a conservativeness parameter, such that

$$\mathcal{D}_{\text{PE}} = \left\{ \theta \in \Theta \mid \|\theta - \hat{\theta}_{\text{MLE}}\|_{\Sigma + \lambda I} \leq f(N, k, \delta, \lambda) \right\}. \tag{61}$$

Then, the policy is optimized for the worst-case parameter in $\mathcal{D}_{\text{PE}}$, such that

$$\hat{\pi}_{\text{PE}} \in \arg\max_{\pi} \min_{\theta \in \mathcal{D}_{\text{PE}}} \mathbb{E}_{s \sim d^\pi}[r_\theta(s, \pi(s))]. \tag{62}$$

(Zhu et al., 2023) prove that the true parameter θ^* exists in this confidence set with high probability as the number of sample increases, which enables them to provide an optimality guarantee.

However, the MLE-based algorithms require an assumption that a human evaluator follows a specific preference model, and the true reward parameter corresponding to the model should lie in the parameter space, such that $\theta^* \in \Theta$, where Θ must be bounded. Such realizability assumption can easily be violated in practice, when the true preference model deviates from the model used in algorithm. For instance, if the BTL model is assumed but the human evaluator follows the greedy policy, the true parameters diverge to $+\infty$ or $-\infty$, which violates the assumption that $\theta^* \in \Theta$. To illustrate these points, we provide a simple bandit problem that both MLE and pessimistic MLE fail but LP succeeds to recover the optimal policy.

Proposition F.1. *For any $\delta > 0$, there exists a linear bandit and a sampling distribution μ_{HF} such that*

$$\hat{\pi}_{LP} = \pi^*, \quad \hat{\pi}_{MLE} \neq \pi^*, \quad \text{and} \quad \hat{\pi}_{PE} \neq \pi^*$$

hold with probability at least $1 - \delta$.

Proof. Consider a linear bandit with a single state s and three actions a_1, a_2, and a_3. We consider the tabular setting such that $r_\theta = [\theta_1, \theta_2, \theta_3]$, where θ_i denotes the reward for the action a_i. Suppose that human evaluators follow the deterministic (greedy) model, and the preference order is given by $a_3 > a_2 > a_1$, i.e. a_3 is the most preferable and a_1 is the least preferable action.

We construct a sampling distribution such that both MLE and pessimistic MLE algorithms in (Zhu et al., 2023) returns a wrong policy with high probability, while LP succeeds to find an optimal policy. Specifically, if the pair (a_1, a_2) is sampled with a significantly higher probability compared to the pair (a_2, a_3) in the queries, we show that $\{\hat{\theta}_{\text{MLE}}\}_2 > \{\hat{\theta}_{\text{MLE}}\}_3$ holds under the BTL model and the greedy evaluator. The MLE algorithm proposed in (Zhu et al., 2023) estimate the reward parameter by solving

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \frac{N_{12}}{N} \log \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}} + \frac{N_{23}}{N} \log \frac{e^{\theta_3}}{e^{\theta_2} + e^{\theta_3}} + \frac{N_{31}}{N} \log \frac{e^{\theta_3}}{e^{\theta_3} + e^{\theta_1}}, \tag{63}$$

where N_{ij} denotes the number of queries (a_i, a_j), $N = N_{12} + N_{23} + N_{31}$, and $\Theta = \{\theta \mid \mathbf{1}^\top \theta = 0, \|\theta\|_2 \leq 1\}$. We first prove the following lemma:

Lemma F.2. *Let $\theta^* = [\theta_1^*, \theta_2^*, \theta_3^*]$ be an optimal solution to the following optimization problem:*

$$\max_{\theta \in \Theta} J(\theta_1, \theta_2, \theta_3) = \alpha \log \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}} + \beta \log \frac{e^{\theta_3}}{e^{\theta_2} + e^{\theta_3}} \tag{64}$$

where $\alpha, \beta \in (0, 1)$. If $\alpha > 2e^3\beta$, then $\theta_2^* > \theta_3^*$.

Proof. Define the Lagrangian function $\mathcal{L}(\theta_1, \theta_2, \theta_3, \lambda_1, \lambda_2) = J(\theta_1, \theta_2, \theta_3) + \lambda_1(1 - \theta_1^2 - \theta_2^2 - \theta_3^2) + \lambda_2(\theta_1 + \theta_2 + \theta_3)$. From the KKT conditions,

$$\begin{aligned}
\left[\frac{\partial \mathcal{L}}{\partial \theta_2}\right]_{(\theta^*, \lambda^*)} &= \alpha \frac{e^{\theta_1^*}}{e^{\theta_1^*} + e^{\theta_2^*}} - \beta \frac{e^{\theta_2^*}}{e^{\theta_2^*} + e^{\theta_3^*}} - 2\lambda_1^*\theta_2^* + \lambda_2^* = 0, \\
\left[\frac{\partial \mathcal{L}}{\partial \theta_3}\right]_{(\theta^*, \lambda^*)} &= \beta \frac{e^{\theta_2^*}}{e^{\theta_2^*} + e^{\theta_3^*}} - 2\lambda_1^*\theta_3^* + \lambda_2^* = 0.
\end{aligned} \tag{65}$$

Subtracting both equations yields

$$\alpha \frac{e^{\theta_1^*}}{e^{\theta_1^*} + e^{\theta_2^*}} - 2\beta \frac{e^{\theta_2^*}}{e^{\theta_2^*} + e^{\theta_3^*}} = 2\lambda_1^*(\theta_2^* - \theta_3^*). \tag{66}$$

If $\alpha > 2e^3\beta$, we can show that the left hand side of the above equality must be greater than 0 as follows:

$$\begin{aligned}
\alpha \frac{e^{\theta_1^*}}{e^{\theta_1^*} + e^{\theta_2^*}} - 2\beta \frac{e^{\theta_2^*}}{e^{\theta_2^*} + e^{\theta_3^*}} &> 2\beta \left(\frac{e^{\theta_1^* + 3}}{e^{\theta_1^*} + e^{\theta_2^*}} - \frac{e^{\theta_2^*}}{e^{\theta_2^*} + e^{\theta_3^*}} \right) = 2\beta \frac{e^{\theta_1^* + \theta_2^* + 3} + e^{\theta_1^* + \theta_3^* + 3} - e^{\theta_1^* + \theta_2^*} - e^{2\theta_2^*}}{(e^{\theta_1^*} + e^{\theta_2^*})(e^{\theta_2^*} + e^{\theta_3^*})} \\
&= 2\beta \frac{(e^{\theta_1^* + \theta_2^* + 3} - e^{\theta_1^* + \theta_2^*}) + (e^{3 - \theta_2^*} - e^{2\theta_2^*})}{(e^{\theta_1^*} + e^{\theta_2^*})(e^{\theta_2^*} + e^{\theta_3^*})} > 0,
\end{aligned} \tag{67}$$

where the last inequality comes from $\theta_2^* \leq 1$. Then, the right hand side $2\lambda_1^*(\theta_2^* - \theta_3^*)$ must be greater than zero as well. Since $\lambda_1^* \geq 0$, we get $\theta_2^* > \theta_3^*$. $\square$

By Lemma F.2, there exist $\alpha, \beta \in (0, 1)$ such that if $\frac{N_{12}}{N} \geq \alpha$, $\frac{N_{23}}{N} \leq \beta$, and $N_{31} = 0$, then $\{\hat{\theta}_{\text{MLE}}\}_2 > \{\hat{\theta}_{\text{MLE}}\}_3$. For any $\delta > 0$, there exists a sampling distribution μ_{HF} satisfying

$$\mathbb{P}(N_{12} \geq \alpha N, \ 1 \leq N_{23} \leq \beta N, \ N_{31} = 0) \geq 1 - \delta. \tag{68}$$

Then, under this sampling distribution μ_{HF}, $\hat{\pi}_{\text{MLE}}(s) = a_2$ with probability at least $1 - \delta$, while $\pi^*(s) = a_3$.

Next, we consider the pessimistic MLE under μ_{HF}. The pessimistic MLE imposes higher penalty on the reward function of state-action pairs that have less support in the data. Therefore, intuitively, the penalty for the state a_3 will be higher than a_2, and thus, $\hat{\pi}_{\text{PE}}(s) = a_2$ will hold. We use the penalty function proposed in (Zhu et al., 2023) to confirm this. Consider the covariance matrix

$$\Sigma = \frac{1}{N} \begin{bmatrix} N_{12} & -N_{12} & 0 \\ -N_{12} & N_{12} + N_{23} & -N_{23} \\ 0 & -N_{23} & N_{23} \end{bmatrix}. \tag{69}$$

Then, the penalty function for a_2 and a_3 are computed as

$$\phi_2 = \|[0, 1, 0]\|_{(\Sigma + \lambda I)^{-1}}^2 = \frac{(N_{12} + \lambda)(N_{23} + \lambda)}{|\Sigma + \lambda I|}, \quad \phi_3 = \|[0, 0, 1]\|_{(\Sigma + \lambda I)^{-1}}^2 = \frac{(N_{12} + \lambda)(N_{12} + N_{23} + \lambda) - N_{12}^2}{|\Sigma + \lambda I|}. \tag{70}$$

It is easy to show that $\phi_3 \geq \phi_2$ for any $\lambda \geq 0$. Then, the inequality

$$\{\hat{\theta}_{\text{MLE}}\}_2 - c\|\phi_2\|_{(\Sigma + \lambda I)^{-1}} > \{\hat{\theta}_{\text{MLE}}\}_3 - c\|\phi_3\|_{(\Sigma + \lambda I)^{-1}} \tag{71}$$

holds for any constant $c > 0$, and thus, $\hat{\pi}_{\text{PE}}$ chooses a_2 as the best action. Therefore, $\hat{\pi}_{\text{PE}} \neq \pi^*$. Finally, since there exists at least one query of (a_2, a_3) (by $N_{23} \geq 1$), we have $\{\hat{\theta}_{\text{LP}}\}_2 \leq \{\hat{\theta}_{\text{LP}}\}_3 + \epsilon_r$. Let $\epsilon_r \leq 0$, we have $\hat{\pi}_{\text{LP}} = \pi^*$, which completes the proof. $\square$

G. Extension to Continuous Feedback

Suppose that the cumulative distribution function (CDF) for y, given a query pair (τ^1, τ^2), can be expressed as:

$$\mathbb{P}(y \leq \alpha \mid (\tau^1, \tau^2)) = \Phi(\alpha; r_{\text{true}}^\top(\psi^1 - \psi^2)) \quad \forall \alpha \in [-1, 1]. \tag{72}$$

Specifically, $\Phi(\cdot; r) : [-1, 1] \mapsto [0, 1]$ is assumed to be a CDF for any $r \in \mathbb{R}$, i.e. right-continuous, monotonically non-decreasing, $\Phi(-1; r) = 0$, and $\Phi(1; r) = 1$. Furthermore, we assume that $\Phi(\alpha; \cdot) : \mathbb{R} \mapsto [0, 1]$ is monotonically non-increasing with respect to r to reflect a preference in the pairwise comparison. Then, we enforce the reward gap to be greater than $c|y|$ by defining an error as

$$\mathcal{L}'(\tau^{n,1}, \tau^{n,2}, y^n; r) := \left(cy^n + r^\top(\psi^{n,2} - \psi^{n,1})\right)\mathbf{1}_{\{y^n \geq 0\}} + \left(-cy^n + r^\top(\psi^{n,1} - \psi^{n,2})\right)\mathbf{1}_{\{y^n \leq 0\}}, \tag{73}$$

where $c > 0$ is a scaling parameter. The solution set $\hat{\mathcal{R}}_{\text{CHF}}(\epsilon_r)$ is then defined in the same way with (29) using the error $\mathcal{L}'$.

$$\hat{\mathcal{R}}_{\text{CHF}}(\epsilon_r) := \{r \mid \mathcal{L}'(\tau^{n,1}, \tau^{n,2}, y^n; r) \leq \epsilon_r \quad \forall n = 1, 2, \ldots, N_q, \quad r \in [-1, 1]^{|S||A|}\}. \tag{74}$$

The reward function r is recovered within the intersection of two sets $\hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g)$ and $\hat{\mathcal{R}}_{\text{CHF}}(\epsilon_r)$:

$$\hat{\mathcal{R}}_{\text{IRL-CHF}}(\epsilon_g, \epsilon_r) = \{(r, u, v) \mid (u, v) \in \hat{\mathcal{R}}_{\text{IRL}}(\epsilon_g), \ r \in \hat{\mathcal{R}}_{\text{CHF}}(\epsilon_r), \ u = \hat{d}_e \circ r\}. \tag{75}$$

It is important to note that $\hat{\mathcal{R}}_{\text{IRL-CHF}}(\epsilon_g, \epsilon_r)$ can become infeasible if ϵ_g or ϵ_r is set too small, due to the strict reward gap. Therefore, choosing proper values for ϵ_g and ϵ_r is crucial to ensure the feasibility of the LP. The generalization guarantee provided in Theorem 4.1 also holds with $\mathcal{L}'$, under a similar proof.

H. Detailed Experimental Setup

Table 1. Algorithm Details

Algorithms	Description	Parameters
MLE-IRL	Bi-level optimization algorithm for offline IRL (Zeng et al., 2023)	Step size $= 0.01$
LP-IRL-1	LP formulation of IRL (26) with a moderate ϵ_g	$\epsilon_g = 0.01/\sqrt{N}$
LP-IRL-2	LP formulation of IRL (26) with a tighter ϵ_g	$\epsilon_g = 0.001/\sqrt{N}$
LP-IRL-D	Integration of IRL and RLHF with discrete feedback (31)	$\epsilon_g = 0.01/\sqrt{N}, \epsilon_r = 0.01/\sqrt{N}$
LP-IRL-C	Integration of IRL and RLHF with continuous feedback (75)	$\epsilon_g = 0.1/\sqrt{N}, \epsilon_r = 0.01/\sqrt{N}$
MLE-HF	Pessimistic MLE under the BTL model (Zhu et al., 2023)	$\lambda = 0.1, B = 1$
LP-HF	LP formulation of RLHF (29)	$\epsilon_r = -0.01$

Environment setting and dataset. We consider an MDP with $|S| = 10$, $|A| = 2$, and $\gamma = 0.95$. The initial state distribution μ_0 and state transition probabilities P are randomly selected for each experiment. Specifically, each element of μ_0 and P is generated from a uniform distribution in the range of $[0, 1]$, and then scaled to form probability distributions. In each experimental run, we sample N trajectories with a horizon of $H = 20$. π_e is used for sampling trajectories in IRL, and the uniform policy ($\pi(a|s) = 1/|A| \quad \forall(s, a) \in S \times A$) is employed for sampling queries (trajectory pairs) in RLHF.

Performance criteria. To introduce additional complexity to the problem, we have set the true rewards to have similar values: $r_{\text{true}}(s, a_1) = 1.0$ and $r_{\text{true}}(s, a_2) = 0.9$ for all states $s \in S$. The performance of each algorithm is then assessed by measuring the proximity of an optimal occupancy measure under the true reward r_{true} and the obtained reward function $\hat{r}$. Specifically, we report $\|d^*(r_{\text{true}}) - d^*(\hat{r})\|_1$, the $\mathcal{L}^1$ error between the true optimal occupancy measure $d^*(r_{\text{true}})$ and the optimal occupancy measure $d^*(\hat{r})$ under the estimated reward $\hat{r}$. This error falls within the range of 0 to 2, with a value of 0 indicating that the estimated optimal policy is equivalent to the true optimal policy. For each sample size N, we conducted 200 experiments and reported the mean and standard deviation of the error.

Expert setting. In offline IRL, the expert policy for sampling trajectories is set to $\pi_e = 0.52 \times \pi^* + 0.48 \times \pi_r$, where π_r denotes a greedy policy that selects a suboptimal action. This setting reflects that π_e can deviate from π^*, particularly when all state-action pairs have similar rewards. In offline RLHF, we consider two different types of human feedback: discrete feedback $y \in \{1, 2\}$ and continuous feedback $y \in [-1, 1]$. The discrete feedback is generated according to the BTL model under the reward r_{true}. The continuous feedback is generated from the uniform distribution, in the range between 0 and $0.2 \times r_{\text{true}}^\top (\psi^1 - \psi^2)$.

Algorithm details. Table 1 provides a detailed description of the algorithms used in experiments. We employ a tabular setting for the reward function without any function approximation. In LP-IRL-1 and LP-IRL-2, we assume that $2/3$ of trajectories are sampled from π_e, and the remaining samples are obtained from the uniform policy to estimate $\hat{d}_{\text{sub}}$. In LP-IRL-D and LP-IRL-C, we assume that $2/3$ of trajectories are sampled from π_e, and the the remaining trajectories are sampled from the uniform policy to generate human feedback. In all LP-IRL algorithms, we use the $\mathcal{L}^1$ norm constraint for X, and we set $\delta = 0.1$ and $B = 100$. In MLE-HF and LP-HF, human feedback is given as discrete value following the greedy model. In LP-HF, a reward function is selected from (29) by optimizing a dummy objective function. For algorithm details of MLE-IRL and MLE-RLHF, please refer to (Zeng et al., 2023) and (Zhu et al., 2023), respectively.