Reinforcement learning with Human Feedback: Learning Dynamic Choices via Pessimism

Anonymous Authors[1]

Abstract

In this paper we study offline Reinforcement Learning with Human Feedback (RLHF) where we aim to learn the human's underlying reward and the MDP's optimal policy from a set of trajectories induced by human choices. We focus on the Dynamic Discrete Choice (DDC) model for modeling and understanding human choices, which is widely used to model a human decision-making process with forward-looking and bounded rationality. We propose a Dynamic-Choice-Pessimistic-Policy-Optimization (DCPPO) method and prove that the suboptimality of DCPPO *almost* matches the classical pessimistic offline RL algorithm in terms of suboptimality's dependency on distribution shift and dimension. To the best of our knowledge, this paper presents the first theoretical guarantees for off-policy offline RLHF with dynamic discrete choice model.

1. Introduction

Reinforcement Learning with Human Feedback (RLHF) is an area in machine learning research that incorporates human guidance or feedback to learn an optimal policy. In recent years, RLHF has achieved significant success in large language models, clinical trials, auto-driving, robotics, etc. (Ouyang et al., 2022; Gao et al., 2022; Glaese et al., 2022; Hussein et al., 2017; Jain et al., 2013; Kupcsik et al., 2018; Menick et al., 2022; Nakano et al., 2021; Novoseller et al., 2020). Unlike conventional offline reinforcement learning, where the learner aims to determine the optimal policy using observable reward data, in RLHF, the learner does not have direct access to the reward signal but instead can only observe a historical record of visited states and human-preferred actions. In such cases, the acquisition of reward knowledge becomes pivotal.

Dynamic Discrete Choice (DDC) model is a framework for studying learning for human choices from data, which has been extensively studied in econometrics literature. (Rust, 1987; Hotz & Miller, 1993; Hotz et al., 1994; Aguirregabiria & Mira, 2002; Kalouptsidi et al., 2021; Bajari et al., 2015; Chernozhukov et al., 2022). In a DDC model, the agent make decisions under unobservable perturbation, i.e. $\pi_h(a_h \mid s_h) = \mathrm{argmax}_a\{Q_h(s_h, a) + \epsilon_h(a)\}$, where ϵ_h is an unobservable random noise and Q_h is the agent's action value function.

In this work, we focus on RLHF within the context of a dynamic discrete choice model. Our challenges are three-folded: (i) The agent must first learn the human behavior policies from the feedback data. (ii) As the agent's objective is to maximize cumulative reward, the reward itself is not directly observable. We need to estimate the reward from the behavior policies. (iii) We face the challenge of insufficient dataset coverage and large state space.

With these coupled challenges, we ask the following question:

Without access to the reward function, can one learn the optimal pessimistic policy from merely human choices under the dynamic choice model?

Our Results. In this work, we propose the Dynamic-Choice-Pessimistic-Policy-Optimization (DCPPO) algorithm. By addressing challenges (i)-(iii), our contributions are three folds: (i) For learning behavior policies in large state spaces, we employ maximum likelihood estimation to estimate state/action value functions with function approximation. We establish estimation error bounds for general model class with low covering number. (ii) Leveraging the learned value functions, we minimize the Bellman mean squared error (BMSE) through linear regression. This allows us to recover the unobservable reward from the learned policy. Additionally, we demonstrate that the error of our estimated reward can be efficiently controlled by an uncertainty quantifier. (iii) To tackle the challenge of insufficient coverage, we follow *the principle of pessimism*, by incorporating a penalty into the value function during value iteration.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

We establish the suboptimality of our algorithm with high probability with only single-policy coverage.

Our result matches existing pessimistic offline RL algorithms in terms of suboptimality's dependence on distribution shift and dimension, even in the absence of an observable reward.To the best of our knowledge, our results offer the first theoretical guarantee for pessimistic RL under the human dynamic choice model.

1.1. Related Work

Reinforcement Learning with Human Feedback. In recent years RLHF and inverse reinforcement learning (IRL) has been widely applied to robotics, recommendation system, and large language model (Ouyang et al., 2022; Lindner et al., 2022; Menick et al., 2022; Jaques et al., 2020; Lee et al., 2021; Nakano et al., 2021). However, there are various ways to incorporate human preferences or expertise into the decision-making process of an agent. (Shah et al., 2015; Ouyang et al., 2022; Saha & Krishnamurthy, 2022) learn reward from pairwise comparison and ranking. (Pacchiano et al., 2021) study pairwise comparison with function approximation in pairwise comparison. (Zhu et al., 2023) study various cases of preference-based-comparison in contextual bandit problem with linear function approximation, however convergence of their algorithm relies on the implicit assumption of sufficient coverage. The majority of prior researches in RLHF only consider bandit cases and have not studied MDP case with transition dynamics. (Wang et al., 2018) study how to learn a uniformly better policy of an MDP from an offline dataset by learning the advantage function. However, they cannot guarantee the learned policy converges to the optimal policy.

Dynamic Discrete Choice Model. Dynamic Discrete Choice (DDC) model is a widely studied choice model in econometrics and is closely related to reward learning in IRL and RLHF. In the DDC model, the human agent is assumed to make decisions under the presence of Gumbel noise (Type I Extreme Error)(Aguirregabiria & Mira, 2002; Chernozhukov et al., 2022; Bajari et al., 2015; Kalouptsidi et al., 2021; Adusumilli & Eckardt, 2019), i.e. under bounded rationality, and the task is to infer the underlying utility. Most work in econometrics cares for asymptotic $\sqrt{n}$-convergence of estimated utility, and does not study finite sample estimation error. Moreover, their methods suffer from significant computation burdens from large or high dimensional state space (Zeng et al., 2022). In recent years, there has been work combining the dynamic discrete choice model and IRL. (Zeng et al., 2022) prove the equivalence between DDC estimation problem and maximum likelihood IRL problem and propose an online gradient method for reward estimation under ergodic dynamics assumption. (Zeng et al., 2023) reformulate the reward estimation in the DDC model into a bilevel optimization and propose a model-based approach by assuming an environment simulator.

Offline Reinforcement Learning and Pessimism. The idea of introducing pessimism for offline RL to deal with distribution shift has been studied in recent years (Jin et al., 2021; Uehara et al., 2021). (Jin et al., 2021) show that pessimism is sufficient to eliminate spurious correlation and intrinsic uncertainty when doing value iteration. (Uehara et al., 2021) show that with single-policy coverage, i.e. coverage over the optimal policy, pessimism is sufficient to guarantee a $\mathcal{O}(n^{-1/2})$ suboptimality. In this paper, we connect RLHF with offline RL and show our algorithm achieves pessimism by designing an uncertainty quantifier that can tackle error from estimating reward functions, which is crucial in pessimistic value iteration.

1.2. Notations and Preliminaries

For a positive-semidefinitematrix $A \in \mathbb{R}^{d\times d}$ and vector $x \in \mathbb{R}^d$, we use $\|x\|_A$ to denote $\sqrt{x^\top Ax}$. For an arbitrary space $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of all probability distribution on $\mathcal{X}$. For two vectors $x, y \in \mathbb{R}^d$, we denote $x \cdot y = \sum_i^d x_i y_i$ as the inner product of x, y. We denote the set of all probability measures on $\mathcal{X}$ as $\Delta(\mathcal{X})$. We use $[n]$ to represent the set of integers from 0 to $n-1$. For two matrices A and B, we write $A \succeq B$ if $A - B \succeq 0$. We define a finite horizon MDP model $M = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h\in[H]}, \{r_h\}_{h\in[H]})$, H is the horizon length, in each step $h \in [H]$, the agent starts from state s_h in the state space $\mathcal{S}$, chooses an action $a_h \in \mathcal{A}$ with probability $\pi_h(a_h \mid s_h)$, receives a reward of $r_h(s_h, a_h)$ and transits to the next state s' with probability $P_h(s' \mid s_h, a_h)$. Here $\mathcal{A}$ is a finite action set with $|\mathcal{A}|$ actions and $P_h(\cdot|s_h, a_h) \in \Delta(s_h, a_h)$ is the transition kernel condition on state action pair (s, a). For convenience we assume that $r_h(s, a) \in [0, 1]$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Without loss of generality, we assume that the initial state of each episode s_0 is fixed. Note that this will not add difficulty to our analysis. For any policy $\pi = \{\pi_h\}_{h\in[H]}$ the state value function is $V_h^\pi(s) = \mathbb{E}_\pi\big[\sum_{t=h}^H r_t(s_t, a_t) \bigm| s_h = s\big]$, and the action value function is $Q_h^\pi(s, a) = \mathbb{E}_\pi\big[\sum_{t=h}^H r_t(s_t, a_t) \bigm| s_h = s, a\big]$, here the expectation $\mathbb{E}_\pi$ is taken with respect to the randomness of the trajectory induced by π, i.e. is obtained by taking action $a_t \sim \pi_t(\cdot \mid s_t)$ and observing $s_{t+1} \sim P_h(\cdot \mid s_t, a_t)$. For any function $f : \mathcal{S} \to \mathbb{R}$, we define the transition operator $\mathbb{P}_h f(s, a) = \mathbb{E}[f(s_{h+1}) \mid s_h = s, a_h = a]$. We also define the Bellman equation for any policy π, $V_h^\pi(s) = \langle \pi_h(a \mid s), Q_h^{\pi_b}(s, a)\rangle$, $Q_h^\pi(s, a) = r_h(s, a) + \mathbb{P}_h V_{h+1}^\pi(s, a)$. For an MDP we denote its optimal policy as π^*, and define the performance metric for any policy π as $\mathrm{SubOpt}(\pi) = V_1^{\pi^*} - V_1^{\pi}$.

2. Problem Formulation

In this paper, we aim to learn from a dataset of human choices under dynamic discrete choice model. Suppose we are provided with dataset $\mathcal{D} = \{\mathcal{D}_h = \{s_h^i, a_h^i\}_{i\in[n]}\}_{h\in[H]}$, containing n trajectories collected by observing a single human behavior in a dynamic discrete choice model. Our goal is to learn the optimal policy π^* of the underlying MDP. We assume that the agent is bounded-rational and makes decisions according to the dynamic discrete choice model (Rust, 1987; Hotz & Miller, 1993; Chernozhukov et al., 2022; Zeng et al., 2023). In dynamic discrete choice model, the agent's policy has the following characterization (Rust, 1987; Aguirregabiria & Mira, 2002; Chernozhukov et al., 2022),

$$\pi_{b,h}(a \mid s) = \frac{\exp(Q_h^{\pi_b,\gamma}(s,a))}{\sum_{a'\in\mathcal{A}} \exp(Q_h^{\pi_b,\gamma}(s,a'))}, \tag{1}$$

here $Q_h^{\pi_b,\gamma}(\cdot,\cdot)$ works as the solution of the discounted Bellman equation,

$$V_h^{\pi_b,\gamma}(s) = \langle \pi_{b,h}(a \mid s), Q_h^{\pi_b,\gamma}(s,a)\rangle, \tag{2}$$

$$Q_h^{\pi_b,\gamma}(s,a) = r_h(s,a) + \gamma \cdot \mathbb{P}_h V_{h+1}^{\pi_b,\gamma}(s,a) \tag{3}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Note that (2) differs from the original Bellman equation due to the presence of γ, which is a discount factor in $[0,1]$, and measures the myopia of the agent. The case of $\gamma = 0$ corresponds to a *myopic* human agent. Such choice model comes from the perturbation of noises,

$$\pi_{b,h}(\cdot \mid s_h) = \operatorname{argmax}_{a\in\mathcal{A}} \left\{ r_h(s_h,a) + \epsilon_h(a) + \gamma \cdot \mathbb{P}_h V_{h+1}^{\pi_b,\gamma}(s_h,a) \right\},$$

where $\{\epsilon_h(a)\}_{a\in\mathcal{A}}$ are i.i.d Gumbel noises that is observed by the agent but not the learner, $\{V_h^{\gamma,\pi_b}\}_{h\in[H]}$ is the value function of the agent, and is widely used to model human decision. We also remark that the state value function defined in (2) corresponds to the *ex-ante* value function in econometric studies (Aguirregabiria & Mira, 2010; Arcidiacono & Ellickson, 2011; Bajari et al., 2015). When considering Gumbel noise as part of the reward, the value function may have a different form. However, such a difference does not add complexity to our analysis.

3. Reward Learning from Human Dynamic Choices

In this section, we present a general framework of an offline algorithm for learning the reward of the underlying MDP. Our algorithm consists of two steps: (i) The first step is to estimate the agent behavior policy from the pre-collected dataset $\mathcal{D}$ by maximum likelihood estimation (MLE). We recover the action value functions $\{Q_h^{\pi_b,\gamma}\}_{h\in[H]}$ from (1) and the state value functions $\{V_h^{\pi_b,\gamma}\}_{h\in[H]}$ from (2) using function approximation. In Section 3.1, we analyze the error of our estimation and prove that for any model class with a small covering number, the error from MLE estimation is of scale $\tilde{\mathcal{O}}(1/n)$ in dataset distribution. We also remark that our result does not need the dataset to be well-explored, which is implicitly assumed in previous works (Zhu et al., 2023; Chen et al., 2020). (ii) We recover the underlying reward from the model class by minimizing a penalized Bellman MSE with plugged-in value functions learned in step (i). In Section 3.2, we study linear model MDP as a concrete example. Theorem 3.5 shows that the error of estimated reward can be bounded by an elliptical potential term for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ in both settings. First, we make the following assumption for function approximation.

Assumption 3.1 (**Function Approximation Model Class**). We assume the existence of a model class $\mathcal{M} = \{\mathcal{M}_h\}_{h\in[H]}$ containing functions $f : \mathcal{S}\times\mathcal{A} \to [0,H]$ for every $h \in [H]$, and is rich enough to capture r_h and Q_h, i.e. $r_h \in \mathcal{M}_h$, $Q_h \in \mathcal{M}_h$.

In practice, $\mathcal{M}_h$ can be a (pre-trained) neural network or a random forest. We now present our algorithm for reward learning in RLHF.

Algorithm 1 DCPPO: Reward Learning for General Model Class

```
Require: Dataset $\{\mathcal{D}_h = \{s_h^i, a_h^i\}_{i\in[n]}\}_{h\in[H]}$, constant
    $\lambda > 0$, penalty function $\rho(\cdot)$, parameter $\beta$.
for step $h = H, \dots, 1$ do
    Set $\widehat{Q}_h$ by maximizing (4).
    Set $\widehat{\pi}_h(a_h|s_h)$ by (5).
    Set $\widehat{V}_h(s_h) = \langle \widehat{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot \mid s_h)\rangle_{\mathcal{A}}$.
    Set $\widehat{r}_h(s_h,a_h)$ by (6).
end for
Output: $\{\widehat{r}_h\}_{h\in[H]}$.
```

3.1. First Step: Recovering Human Policy and Human State-Action Values

For every step h, we use maximum liklihood estimaton (MLE) to estimate the behaviour policy $\pi_{b,h}$, corresponds to $Q_h^{\pi_b,\gamma}(s,a)$ in a general model class $\mathcal{M}_h$. For each step $h \in [H]$, we have the log-likelihood function

$$L_h(Q) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(Q(s_h^i, a_h^i))}{\sum_{a'\in\mathcal{A}} \exp(Q(s,a'))}\right) \tag{4}$$

for $Q \in \mathcal{M}_h$, and we estimate Q_h by maximizing (4). Then we recover the policy $\widehat{\pi}_h$ by

$$\widehat{\pi}_h(a_h \mid s_h) = \exp(\widehat{Q}_h(s_h, a_h)) / \sum_{a' \in \mathcal{A}} \exp(\widehat{Q}_h(s_h, a'). \tag{5}$$

Note that by Equation (1), adding a constant on $Q_h^{\pi_b,\gamma}$ will produce the same policy under dynamic discrete model, and thus the real behavior value function is unidentifiable in general. For identification, we have the following assumption.

Assumption 3.2 (Model Identification). We assume that there exists one $a_0 \in \mathcal{A}$, such that $Q(s, a_0) = 0$ for every $s \in \mathcal{S}$.

Note that this assumption does not affect our further analysis. Other identifications includes parameter constraint (Zhu et al., 2023) or utility constraints (Bajari et al., 2015). We can ensure the estimation of the underlying policy and corresponding value function is accurate in the states the agent has encountered. Formally, we have the following theorem,

Theorem 3.3 (Policy and Value Functions Recovery from Choice Model). *With Algorithm 1 , we have*

$$\mathbb{E}_{\mathcal{D}_h}\left[\|\widehat{\pi}_h(\cdot \mid s_h) - \pi_{b,h}(\cdot \mid s_h)\|_1^2\right] \le \mathcal{O}\left(\frac{\log\left(N(\mathcal{M}_h, \|\cdot\|_\infty, 1/n)/\delta\right)}{n}\right)$$

and

$$\mathbb{E}_{\mathcal{D}_h}\left[\|\widehat{Q}_h(s_h, \cdot) - Q_h^{\pi_b,\gamma}(s_h, \cdot)\|_1^2\right] \le \mathcal{O}\left(\frac{H^2 \cdot e^H \cdot \log\left(N(\mathcal{M}_h, \|\cdot\|_\infty, 1/n)/\delta\right)}{n}\right)$$

with probability at least $1 - \delta$. Here $\mathbb{E}_{\mathcal{D}_h}[\cdot]$ means the expectation is taken on collected dataset $\mathcal{D}_h$, i.e. the mean value taken with respect to $\{s_h^i\}_{i \in [n]}$.

Theorem 3.3 shows that we can efficiently learn $\pi_{b,h}$ from the dataset under identification assumption. As a result, we can provably recover the value functions by definition in Equation 1.

3.2. Reward Learning from Dynamic Choices

Bellman equation motivates the following estimate of the reward function:

$$\begin{aligned} \widehat{r}_h(s_h, a_h) = & \\ & \operatorname{argmin}_{r \in \mathcal{M}_h} \Big\{ \sum_{i=1}^n \big(r_h(s_h^i, a_h^i) + \gamma \cdot \widehat{V}_{h+1}(s_{h+1}^i) \\ & - \widehat{Q}_h(s_h^i, a_h^i)\big)^2 + \lambda \rho(r) \Big\}, \end{aligned} \tag{6}$$

i.e. we can recover the reward with previously learned $\widehat{V}_h, \widehat{Q}_h$ by minimizing Bellman MSE. As a concrete example, we study the instantiation of Algorithm 1 for the linear model class. We define the function class $\mathcal{M}_h = \{f(\cdot) = \phi(\cdot)^\top \theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \Theta\}$ for $h \in [H]$, where $\phi \in \mathbb{R}^d$ is the feature defined on $\mathcal{S} \times \mathcal{A}$, Θ is a subset of $\mathbb{R}^d$ which parameterizes the model class, and $d > 0$ is the dimension of the feature. Corresponding to Assumption 3.2, We also assume that $\phi(s, a_0) = 0$ for every $s \in \mathcal{S}$. Note that this model class contains the reward r_h and state action value function Q_h in tabular MDP where $\phi(s, a)$ is the one-hot vector of (s, a). The linear model class also contains linear MDP, which assumes both the transition $P(s_{h+1} \mid s_h, a_h)$ and the reward $r_h(s_h, a_h)$ are linear functions of feature $\phi(s_h, a_h)$ (Jin et al., 2020; Duan et al., 2020; Jin et al., 2021). In linear model case, our first step MLE in (4) turns into a logistic regression,

$$\widehat{\theta}_h = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \phi(s_h^i, a_h^i) \cdot \theta - \log\left(\sum_{a' \in \mathcal{A}} \exp(\phi(s_h^i, a') \cdot \theta)\right), \tag{7}$$

which can be efficiently solved by existing state-of-art optimization methods. We now have $\{\widehat{Q}_h\}_{h \in [H]}, \{\widehat{\pi}_h\}_{h \in [H]}$ and $\{\widehat{V}_h\}_{h \in [H]}$ in Algorithm 1 to be our estimations for $Q_h^{\pi_b,\gamma}, \pi_{b,h}$ and $V_h^{\pi_b,\gamma}$ correspondingly.

Note that in linear case, Line 5 has a closed form solution,

$$\widehat{w}_h = (\Lambda_h + \lambda I)^{-1}\left(\sum_{i=1}^n \phi(s_h^i, a_h^i)\big(\widehat{Q}_h(s_h^i, a_h^i) - \gamma \cdot \widehat{V}_{h+1}(s_{h+1}^i)\big)\right), \tag{8}$$

$$\text{where } \Lambda_h = \sum_{i=1}^n \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top,$$

and we set $\widehat{r}(s_h, a_h) = \phi(s_h, a_h) \cdot \widehat{w}_h$. We also make the following assumption on the model class Θ and the feature function.

Assumption 3.4 (Regular Conditions). We assume that: (i) For all $\theta \in \Theta$, we have $\|\theta\|_2 \le \sqrt{d}$; (ii) For all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, $\|\phi(s_h, a_h)\|_2 \le \sqrt{d}$. (iii)For all $n > 0$, $\log N(\Theta, \|\cdot\|_\infty, 1/n) \le c \cdot d \log n$ for some absolute constant c.

We are now prepared to highlight our main result:

Theorem 3.5 (Reward Estimation for Linear Model MDP). *With probability at least $1 - \delta$, we have the following estimation of our reward function for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\lambda > 0$,*

$$\begin{aligned} & |r_h(s, a) - \widehat{r}_h(s, a)| \\ & \le \|\phi(s, a)\|_{(\Lambda_h + \lambda I)^{-1}} \\ & \quad \cdot \mathcal{O}\left(\sqrt{\lambda \cdot d} + (1 + \gamma) H e^H \cdot d \sqrt{\log\left(nH/\delta\right)}\right). \end{aligned} \tag{9}$$

Note that the error can be bounded by the product of two terms, the elliptical potential term $\|\phi(s, a)\|_{(\Lambda + \lambda \cdot I)^{-1}}$ and the norm of a self normalizing term of scale $O(He^H \cdot$

$d\sqrt{\log(n/\delta)})$. Here the exponential dependency $\mathcal{O}(e^H)$ comes from estimating $Q_h^{\pi_b,\gamma}$ with logistic regression and also occurs in logistic bandit (Zhu et al., 2023; Fei et al., 2020). It remains an open question if this additional factor can be improved, and we leave it for future work.

Remark 3.6. We remark that except for the exponential term in H, Theorem 3.5 *almost* matches the result when doing linear regression on an observable reward dataset, in which case error of estimation is of scale $\tilde{\mathcal{O}}(\|\phi(s,a)\|_{(\Lambda+\lambda I)^{-1}} \cdot dH)$ (Ding et al., 2021; Jin et al., 2021). When the human behavior policy has sufficient coverage, i.e. the minimal eigenvalue of $\mathbb{E}_{\pi_b}[\phi\phi^\top]$, $\sigma_{\min}(\mathbb{E}_{\pi_b}[\phi\phi^\top]) \geq 0$, we have $\|\phi(s,a)\|_{(\Lambda_h+\lambda I)^{-1}} = \mathcal{O}(n^{-1/2})$ holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (Duan et al., 2020) and $\|r_h - \widehat{r}_h\|_\infty = \mathcal{O}(n^{-1/2})$.

4. Policy Learning from Dynamic Choices via Pessimistic Value Iteration

In this section, we describe the pessimistic value iteration algorithm, which minus a penalty function $\Gamma_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ from the value function when choosing the best action. Pessimism is achieved when Γ_h is a *uncertainty quantifier* for our learned value functions $\{\tilde{V}_h\}_{h\in[H]}$, i.e. $\left|\left(\widehat{r}_h + \widetilde{\mathbb{P}}_h\widetilde{V}_{h+1}\right)(s,a) - \left(r_h + \mathbb{P}_h\widetilde{V}_{h+1}\right)(s,a)\right| \leq \Gamma_h(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ with high probability. Then we use $\{\Gamma_h\}_{h\in[H]}$ as the penalty function for pessimistic planning, which leads to a conservative estimation of the value function. We formally describe our planning method in Algorithm 2. However, when doing pessimistic value iteration with $\{\widehat{r}_h\}_{h\in[H]}$ learned from human feedback, it is more difficult to design uncertainty quantifiers, since the estimation error from reward learning is inherited in pessimistic planning. In Section 4.1, we propose an efficient uncertainty quantifier and prove that with pessimistic value iteration, Algorithm 2 can achieve a $\mathcal{O}(n^{-1/2})$ suboptimality gap.

Algorithm 2 DCPPO: Pessimistic Value iteration

```
Require: Surrogate reward $\{\widehat{r}_h(\cdot,\cdot)\}_{h\in[H]}$, collected dataset $\{(s_h^i, a_h^i)\}_{i\in[n],h\in[H]}$, parameter $\beta$, penalty .
1: Set $\widetilde{V}_{H+1}(\cdot) = 0$.
2: for step $h = H, \ldots, 1$ do
3:     Set $\widetilde{\mathbb{P}}_h\widetilde{V}_{h+1}(s_h, a_h)$ by (10).
4:     Construct $\Gamma_h(s_h, a_h)$ based on $\mathcal{D}$.
5:     Set $\widetilde{Q}_h(s_h, a_h) = \min\left\{\widehat{r}_h(s_h, a_h) + \widetilde{P}_h\widetilde{V}_{h+1}(s_h, a_h) - \Gamma_h(s_h, a_h), H - h + 1\right\}_+$.
6:     Set $\widetilde{\pi}_h(\cdot \mid \cdot) = \mathrm{argmax}\langle\widetilde{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle$.
7:     Set $\widetilde{V}_h = \langle\widetilde{Q}_h(\cdot,\cdot), \tilde{\pi}_h(\cdot \mid \cdot)\rangle$.
8: end for
9: Output: $\{\widetilde{\pi}_h\}_{h\in[H]}$.
```

4.1. Suboptimality Gap of Pessimitic Optimal Policy

In Line (3) of Algorithm 2, we update $\mathbb{P}_h\hat{V}_{h+1}$ by solving the following minimization:

$$\widetilde{\mathbb{P}}_h\widetilde{V}_{h+1}(s_h, a_h) = \mathrm{argmin}_{f\in\mathcal{M}} \sum_{i\in[n]} \left(f(s_h^i, a_h^i) - \widetilde{V}_{h+1}(s_{h+1})\right)^2. \tag{10}$$

For linear model class defined in Section 3.2, we assume that we can capture the conditional expectation of value function in the next step with the known feature ϕ. In formal words, we make the following assumption.

Assumption 4.1 (Linear MDP). For the underlying MDP, we assume that for every $V_{h+1} : \mathcal{S} \to [0, H-h]$, there exists $u_h \in \mathbb{R}^d$ such that

$$\mathbb{P}_h V_{h+1}(s,a) = \phi(s,a) \cdot u_h$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We also assume that $\|u_h\| \leq (H - h + 1) \cdot \sqrt{d}$ for all $h \in [H]$.

Note that this assumption is directly satisfied by linear MDP class (Jin et al., 2021; 2020; Yang & Wang, 2019). For linear model MDP defined in Section 3.2, it suffices to have the parameter set Θ being closed under subtraction, i.e. if $x, y \in \Theta$ then $x - y \in \Theta$. Meanwhile, we construct Γ_h in Algorithm 2 based on dataset $\mathcal{D}$ as

$$\Gamma_h(s,a) = \beta \cdot \left(\phi(s,a)^\top(\Lambda_h + \lambda I)^{-1}\phi(s,a)\right)^{1/2} \tag{11}$$

for every $h \in [H]$. Here that Λ_h is defined in (8). To establish suboptimality for Algorithm 2, we assume that the trajectory induced by π^* is "covered" by $\mathcal{D}$ sufficiently well.

Assumption 4.2 (Single-Policy Coverage). Suppose there exists an absolute constant $c^\dagger > 0$ such that

$$\Lambda_h \geq c^\dagger \cdot n \cdot \mathbb{E}_{\pi^*}\left[\phi\left(s_h, a_h\right)\phi\left(s_h, a_h\right)^\top\right]$$

holds with probability at least $1 - \delta/2$.

We remark that Assumption 4.2 only assumes the human behavior policy can cover the optimal policy and is therefore weaker than assuming well-explored dataset, or sufficient coverage(Duan et al., 2020; Jin et al., 2021). With this assumption, we prove the following theorem.

Theorem 4.3 (Suboptimality Gap for DCPPO). *Suppose Assumption 3.2, 3.4, 4.1,4.2 holds. With $\lambda = 1$ and $\beta = \mathcal{O}(He^H \cdot d\sqrt{\log\left(nH/\delta\right)})$, we have (i) Γ_h defined in (11) being uncertainty quantifiers, and (ii)*

$$\mathrm{SubOpt}\left(\{\tilde{\pi}_h\}_{h\in[H]}\right) \leq c \cdot (1+\gamma)d^{3/2}H^2e^Hn^{-1/2}\sqrt{\xi}$$

holds with probability at least $1-\delta$, here $\xi = \log(dHn/\delta)$. In particular, if $\mathrm{rank}(\Sigma_h) \leq r$ at each step $h \in [H]$, then

$$\mathrm{SubOpt}\left(\{\tilde{\pi}_h\}_{h\in[H]}\right) \leq c \cdot (1+\gamma)r^{1/2}dH^2e^Hn^{-1/2}\sqrt{\xi},$$

here $\Sigma_h = \mathbb{E}_{\pi_b}[\phi(s_h, a_h)\phi(s_h, a_h)^\top]$.

Remark. It is worth highlighting that Theorem 4.3 nearly matches the standard result for pessimistic offline RL with observable rewards in terms of the dependence on data size and distribution, up to a constant factor of $\mathcal{O}(He^H)$ (Jin et al., 2020; Uehara & Sun, 2021), where their suboptimality is of $\tilde{\mathcal{O}}(dH^2n^{-1/2})$. Therefore, Algorithm 1 and 2 *almost* matches the suboptimality gap of standard pessimism planning with an observable reward, except for a $\mathcal{O}(e^H)$ factor inherited from reward estimation.

5. Conclusion

In this paper, we have developed a provably efficient online algorithm, Dynamic-Choice-Pessimistic-Policy-Optimization (DCPPO) for RLHF under dynamic discrete choice model. By maximizing log-likelihood function of the Q-value function and minimizing mean squared Bellman error for the reward, our algorithm learns the unobservable reward, and the optimal policy following the principle of pessimism. We prove that our algorithm is efficient in sample complexity for linear model MDP.

References

Adusumilli, K. and Eckardt, D. Temporal-difference estimation of dynamic discrete choice models. *arXiv preprint arXiv:1912.09509*, 2019.

Aguirregabiria, V. and Mira, P. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70(4):1519–1543, 2002.

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

Arcidiacono, P. and Ellickson, P. B. Practical methods for estimation of dynamic discrete choice models. *Annu. Rev. Econ.*, 3(1):363–394, 2011.

Bajari, P., Chernozhukov, V., Hong, H., and Nekipelov, D. Identification and efficient semiparametric estimation of a dynamic discrete game. Technical report, National Bureau of Economic Research, 2015.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment optimization with changing contextual information. *The Journal of Machine Learning Research*, 21(1):8918–8961, 2020.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization, 2022.

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

Hotz, V. J. and Miller, R. A. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529, 1993.

Hotz, V. J., Miller, R. A., Sanders, S., and Smith, J. A simulation estimator for dynamic models of discrete choice. *The Review of Economic Studies*, 61(2):265–289, 1994.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Jain, A., Wojcik, B., Joachims, T., and Saxena, A. Learning trajectory preferences for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

Jaques, N., Shen, J. H., Ghandeharioun, A., Ferguson, C., Lapedriza, A., Jones, N., Gu, S. S., and Picard, R. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*, 2020.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

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.

Kalouptsidi, M., Scott, P. T., and Souza-Rodrigues, E. Linear iv regression estimators for structural dynamic discrete choice models. *Journal of Econometrics*, 222(1): 778–804, 2021.

Kupcsik, A., Hsu, D., and Lee, W. S. Learning dynamic robot-to-human object handover from human feedback. *Robotics Research: Volume 1*, pp. 161–176, 2018.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Lindner, D., Tschiatschek, S., Hofmann, K., and Krause, A. Interactively learning preference constraints in linear bandits, 2022.

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

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

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

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.

Pacchiano, A., Saha, A., and Lee, J. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Rust, J. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pp. 999–1033, 1987.

Saha, A. and Krishnamurthy, A. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994. PMLR, 2022.

Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, pp. 856–865. PMLR, 2015.

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

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Wang, Q., Xiong, J., Han, L., Liu, H., Zhang, T., et al. Exponentially weighted imitation learning for batched historical data. *Advances in Neural Information Processing Systems*, 31, 2018.

Yang, L. F. and Wang, M. Sample-optimal parametric q-learning using linearly additive features, 2019.

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

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

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.