Preference Elicitation for Offline Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Applying reinforcement learning (RL) to real-world problems is often made challenging by the inability to interact with the environment and the difficulty of designing reward functions. Offline RL addresses the first challenge by considering access to an offline dataset of environment interactions labeled by the reward function. In contrast, Preference-based RL does not assume access to the reward function and learns it from preferences, but typically requires an online interaction with the environment. We bridge the gap between these frameworks by exploring efficient methods for acquiring preference feedback in a fully offline setup. We propose Sim-OPRL, an offline preference-based reinforcement learning algorithm, which leverages a learned environment model to elicit preference feedback on simulated rollouts. Drawing on insights from both the offline RL and the preference-based RL literature, our algorithm employs a pessimistic approach for out-of-distribution data, and an optimistic approach for acquiring informative preferences about the optimal policy. We provide theoretical guarantees regarding the sample complexity of our approach, dependent on how well the offline data covers the optimal policy. Finally, we demonstrate the empirical performance of Sim-OPRL in different environments.

1 Introduction

While reinforcement learning (RL) [Sutton and Barto, 2018] achieves excellent performance in various decision-making tasks [Mirhoseini et al., 2020, Degrave et al., 2022], its practical deployment remains limited by the requirement of direct interaction with the environment. This can be impractical or unsafe in real-world scenarios. For example, patient management and treatment in intensive care units involve complex decision-making that has often been framed as a reinforcement learning problem [Raghu et al., 2017]. However, the timing, dosage, and combination of treatments required are critical to patient safety, and incorrect decisions can lead to severe complications or death, making the use of traditional RL algorithms unfeasible [Tang and Wiens, 2021]. Offline RL emerges as a promising solution, allowing policy learning from entirely observational data [Levine et al., 2020].

Still, a challenge with Offline RL is its requirement for an explicit reward function. Quantifying the numerical value of taking a certain action in a given environment state is often challenging [Yu et al., 2021]. Preference-based RL offers a compelling alternative, relying on comparisons between different trajectories, and being often easier for humans to provide [Wirth et al., 2017]. In medical settings, for instance, clinicians may be queried for feedback on which trajectories lead to favorable outcomes. Unfortunately, most algorithms for preference acquisition require environment interaction [Saha et al., 2023, Chen et al., 2022] and are therefore not applicable to the offline setting.

Lack of environment interaction and reward learning are thus two critical challenges for real-world RL deployment that are rarely tackled jointly. In this work, we address the problem of preference elicitation for offline reinforcement learning by asking: *What trajectories should we sample*

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

to minimize the number of human queries required to learn the best offline policy? This presents a challenging problem as it combines learning from offline data and active feedback acquisition, two frameworks that require opposing inductive biases for conservatism and exploration, respectively.

To the best of our knowledge, the only strategy proposed in prior work is to acquire feedback directly over samples within an offline dataset of trajectories [Shin et al., 2022, Offline Preference-based Reward Learning (OPRL)]. We propose an alternative solution that queries feedback on *simulated rollouts* by leveraging a learned environment model. Our offline preference-based reinforcement learning algorithm, Sim-OPRL, strikes a balance between conservatism and exploration by combining pessimism when handling states out-of-distribution from the observational data [Jin et al., 2021, Zhan et al., 2023a], and optimism in acquiring informative preferences about the optimal policy [Saha et al., 2023, Chen et al., 2022]. We validate our approach through both theoretical and empirical analysis, demonstrating the superior performance of Sim-OPRL across various environments.

Our contributions are the following: (1) In Section 3, we first formalize the new problem setting of preference elicitation for offline reinforcement learning, which allows for **complementing offline data with preference feedback**. This framework is crucial for real-world applications where direct environment interaction is infeasible and reward functions are challenging to design manually, yet experts can be queried for their knowledge. (2) In Section 4, we propose a novel offline preference-based RL algorithm that is independent of the specific preference elicitation strategy and recovers a robust policy from an offline dataset and preference feedback. (3) Next, in Section 5, we provide theoretical guarantees on eliciting preferences over samples from the offline dataset, complementing work from Shin et al. [2022]. (4) Then, in Section 6 we propose our own **efficient preference elicitation algorithm** based on simulated rollouts in a learned environment model. (5) Finally, we establish the **theoretical guarantees** of our algorithm and demonstrate its **empirical efficiency** and scalability in different decision-making environments.

2 Related Work

Our problem setting shares similarities with Offline RL and Preference-based RL, which we summarize below. We position ourselves relative to our closest related works in Table 1.

Offline RL. Offline Reinforcement Learning has gained significant traction in recent years, as the practicality of training RL agents without environment interaction makes it relevant to real-world applications [Levine et al., 2020]. However, learning from observational data only is a source of bias in the model, as the data may not cover the entire state-action space. Offline RL algorithms therefore output pessimistic policies, which has been shown to minimize suboptimality Jin et al. [2021]. Model-based approaches show particular promise for their sample efficiency [Yu et al., 2020, Kidambi et al., 2020, Uehara and Sun, 2021]. In this work, we study the setting where reward signals are unavailable and must be estimated by actively querying preference feedback.

Preference-based RL. Rather than accessing numerical reward values for each state-action pair as in traditional online RL, preference-based RL learns the reward model through collecting pairwise preferences over trajectories [Wirth et al., 2017]. Different preference elicitation strategies have been proposed for this framework, generally based on knowing the transition model exactly or on having access to the environment for rollouts [Christiano et al., 2017, Saha et al., 2023, Chen et al., 2022, Lindner et al., 2021, Zhan et al., 2023b, Sadigh et al., 2018, Brown et al., 2020].

Offline Preference-based RL. The development of preference-based RL algorithms based on offline data only is critical to settings where environment interaction is not feasible for safety and efficiency reasons. Still, this framework remains largely unexplored in the literature. While Zhu et al. [2023], Zhan et al. [2023a] demonstrate the value of pessimism in offline preference-based reinforcement learning, they do not consider how to query feedback actively. On the other hand, Shin

Table 1: **Comparison of related work on preference elicitation.**

Framework	Offline	Efficient Sampling	Robustness Guarantees	Practical Implementation
PbOP [Chen et al., 2022]	✗	✓	✓	✗
REGIME [Zhan et al., 2023b]	✗	✓	✓	✗
FREEHAND [Zhan et al., 2023a]	✓	✗	✓	✗
OPRL [Shin et al., 2022]	✓	✓[1]	✗	✓
Sim-OPRL (Ours)	✓	✓	✓	✓

[1]We demonstrate this in the present work.

et al. [2022] propose an empirical comparison of different preference sampling trajectories from an offline trajectories buffer. In Section 5, we provide a theoretical analysis of their approach, then propose an alternative sampling strategy based on simulated trajectory rollouts in Section 6, which enjoys both theoretical and empirical motivation.

3 Problem formulation

3.1 Preliminaries

Markov Decision Process. We consider the episodic Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, T, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the episode length, $T : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. We assume an initial state s_0, but our analysis could be easily generalized to a fixed initial state distribution. At time t, the environment is at state $s_t \in \mathcal{S}$ and an agent selects an action $a_t \in \mathcal{A}$. The agent then receives a reward $R(s_t, a_t)$ and the environment transitions to state $s_{t+1} \sim T(\cdot \mid s_t, a_t)$. We describe an agent's behavior through a policy function $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, such that $\pi(a|s)$ is the probability of taking action a in state s. Let $\tau = (s_0, a_0, \dots s_H, a_H)$ denote the trajectory of state-action pairs of an interaction episode with the environment. With an abuse of notation, we also write $R(\tau) = \sum_t R(s_t, a_t)$. Let d_T^π denote the distribution of trajectories induced by rolling out policy π in transition model T. We denote the expected return of policy π as $V_{T,R}^\pi = \mathbb{E}_{\tau \sim d_T^\pi}[R(\tau)]$, and $\pi^* = \operatorname{argmax}_\pi V_{T,R}^\pi$ denotes the optimal policy in $\mathcal{M}$.

Preference-based Reinforcement Learning. Rather than observing numerical rewards at each state and action, we receive preference feedback over trajectories. For a pair of trajectories (τ_1, τ_2), we obtain binary feedback $o \in \{0, 1\}$ about whether τ_1 is preferred to τ_2. We assume that preference labels follow the Bradley-Terry model [Bradley and Terry, 1952]:

$$P_R(\tau_1 \succ \tau_2) := P(o = 1 | \tau_1, \tau_2) = \frac{\exp(R(\tau_1))}{\exp(R(\tau_1)) + \exp(R(\tau_2))} = \sigma(R(\tau_1) - R(\tau_2)), \tag{1}$$

where $\succ$ denotes a preference relationship and σ is the sigmoid function. Within this framework, *preference elicitation* refers to the process of sampling preferences to obtain information about both the preference function and the system dynamics [Wirth et al., 2017].

3.2 Offline Preference Elicitation

We assume access to an observational dataset of trajectories $\mathcal{D}_{offline} = \{\tau : \tau \sim d_T^{\pi_\beta}\}$, where π_β is an unknown behavioural policy in $\mathcal{M}$. As in Offline RL, we do *not* have access to the decision-making environment to observe transition dynamics or rewards under alternative action choices. We assume *not* to have access to the reward function, but we can query preference feedback from a human to generate a dataset of preferences $\mathcal{D}_{pref} = \{(\tau_1, \tau_2, o)\}$.

Optimality Criterion. Based only on our offline dataset $\mathcal{D}_{offline}$, our goal is to recover a policy $\hat{\pi}^*$ that minimizes suboptimality in the true environment with as few human preference queries as possible. Let $\pi^*_{offline}$ denote the *optimal offline policy* estimated based on the offline data, with access to the true reward function R, and let ϵ_T denote its suboptimality. Since preference elicitation only allows us to estimate the reward function, we do not aim to achieve a suboptimality less than ϵ_T.[2] Our objective is then formalized as follows.

Definition 3.1 (Optimality Criterion of Offline Preference Elicitation)**.** *Let π^* be the optimal policy in $\mathcal{M}$ and $\hat{\pi}^*$ be the estimated optimal policy based on an offline dataset $\mathcal{D}_{offline}$ and $N_p > 0$ preference queries. Let ϵ_T be the inherent suboptimality assuming access to the true reward function. We say that a sampling strategy is (ϵ, δ, N_p)-correct if for every $\epsilon \geq \epsilon_T$, with probability at least $(1 - \delta)$, it holds that $V_{T,R}^{\pi^*} - V_{T,R}^{\hat{\pi}^*} \leq \epsilon$.*

Our work is the first to formalize this important problem, which faces the challenge of balancing exploration when actively acquiring feedback and bias mitigation in learning from offline data.

Function classes. We estimate the reward function and transition kernel with general function approximation; let $\mathcal{F}_R$ and $\mathcal{F}_T$ denote the classes of functions considered respectively. We also assume a policy class Π. Our theoretical analysis also requires the following assumptions and definitions, which are standard in preference-based RL [Chen et al., 2022, Zhan et al., 2023a].

[2]However, ϵ_T is not formally a lower bound for our problem, as shown in Appendix A.3.

Algorithm 1 Offline Preference-based Reinforcement Learning with Preference Elicitation

Input: Observational trajectories dataset $\mathcal{D}_{offline}$. Significance $\delta \in (0, 1)$, preference budget N_p.
Output: $\hat{\pi}^*$

```
Estimate T̂ and u_T via maximum likelihood over the observational data D_offline.
D_pref ← ∅.
for k = 1, ...N_p do
    Generate trajectory pairs (τ_1, τ_2).          ▷ Preference Elicitation: Sections 5 and 6
    Collect preference label o for (τ_1, τ_2).
    D_pref ← D_pref ∪ {(τ_1, τ_2, o)}.
    Estimate R̂ and u_R via maximum likelihood over the preference data D_pref.
end for
π̂* ← argmax_{π∈Π} E_{τ∼d^π_T̂} [R̂(τ) − u_R(τ) − u_T(τ)]
```

Assumption 3.1 (Realizability)**.** *The true reward function belongs to the reward class: $R \in \mathcal{F}_R$. The true transition function belongs to the transition class: $T \in \mathcal{F}_T$. The optimal policy belongs to the policy class: $\pi^* \in \Pi$.*

Assumption 3.2 (Boundedness)**.** *The reward function is bounded: $0 \leq \tilde{R}(\tau) \leq R_{max}$ for all $\tilde{R} \in \mathcal{F}_R$ and all trajectories τ.*

Definition 3.2 (ϵ-bracketing number)**.** *Let $\mathcal{F}$ be a class of real functions $f : \mathcal{X} \to \mathbb{R}$. We say (l, u) is an ϵ-bracket if $l(x) \leq u(x)$ and $\|u(x) - l(x)\|_1 \leq \epsilon$ for all $x \in \mathcal{X}$. The ϵ-bracketing number of $\mathcal{F}$, denoted $\mathcal{N}_{\mathcal{F}}(\epsilon)$, is the minimal number of ϵ-brackets $(l^n, u^n)_{n=1}^N$ needed so that for any $f \in \mathcal{F}$, there is a bracket $i \in [N]$ containing it, meaning $l^i(x) \leq f(x) \leq u^i(x)$ for all $x \in \mathcal{X}$.*

Let $\mathcal{N}_{\mathcal{F}_R}(\epsilon)$ and $\mathcal{N}_{\mathcal{F}_T}(\epsilon)$ denote the ϵ-bracketing numbers of $\mathcal{F}_R$ and $\mathcal{F}_T$ respectively. This measures the complexity of the function classes [Geer, 2000].

Definition 3.3 (Transition concentrability coefficient, Zhan et al. [2023a])**.** *The concentrability coefficient w.r.t. transition classes $\mathcal{F}_T$ and the optimal policy π^* is defined as:*

$$C_T(\mathcal{F}_T, \pi^*) = \sup_{\tilde{T} \in \mathcal{F}_T} \left[\frac{\mathbb{E}_{(s,a) \sim d_T^{\pi^*}}[|T(\cdot|s,a) - \tilde{T}(\cdot|s,a)|]}{\sqrt{\mathbb{E}_{(s,a) \sim \mathcal{D}_{offline}}[|T(\cdot|s,a) - \tilde{T}(\cdot|s,a)|^2]}} \right]$$

The concentrability coefficient measures the coverage of the optimal policy in the offline dataset. Note that C_T is upper-bounded by the density-ratio coefficient: $C_T(\mathcal{F}_T, \pi^*) \leq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_T^{\pi^*}(s,a)/d_T^{\pi_\beta}(s,a)$, where π_β is the behavioural policy underlying $\mathcal{D}_{offline}$.

4 Offline Preference-based RL with Preference Elicitation

In this section, we propose a general framework for offline preference-based reinforcement learning. The next two sections propose two different preference elicitation strategies. As learning must be carried out in two stages, with environment dynamics based on $\mathcal{D}_{offline}$ and reward learning on $\mathcal{D}_{pref}$, we adopt a model-based approach which we summarize in Algorithm 1.

Model Learning. We first leverage the offline data to learn a model of the environment dynamics, fitting a transition model $\hat{T}$ and an uncertainty function u_T through maximum likelihood:

$$\begin{aligned} \hat{T} \quad &= \text{argmax}_{\tilde{T} \in \mathcal{F}_T} \mathbb{E}_{(s,a,s') \sim \mathcal{D}_{offline}} \left[\log \tilde{T}(s'|s,a) \right], \\ u_T(s,a) &= \max_{\tilde{T}_1, \tilde{T}_2 \in \mathcal{T}} |\tilde{T}_1(\cdot|s,a) - \tilde{T}_2(\cdot|s,a)| \cdot R_{max}, \end{aligned}$$

where $\mathcal{T} = \{\tilde{T} \in \mathcal{F}_T \mid \mathbb{E}_{(s,a,s') \sim \mathcal{D}_{offline}} \left[\log \hat{T}(s'|s,a)/\tilde{T}(s'|s,a) \right] \leq \beta_T\}$, defining a confidence set over the MLE estimate, and β_T is a hyperparameter. In a practical implementation, this can be achieved by training an ensemble of models on different data bootstraps [Lakshminarayanan et al., 2017].

Iterative Preference Elicitation and Reward Learning. As with the transition model, our algorithm estimates the reward function $\hat{R}$ and its uncertainty function through maximum likelihood over iteratively collected preference data $\mathcal{D}_{pref}$:

$$\hat{R} = \text{argmax}_{\tilde{R}\in\mathcal{F}_R}\mathbb{E}_{(\tau_1,\tau_2,o)\sim\mathcal{D}_{pref}}\left[o\log P_{\tilde{R}}(\tau_1\succ\tau_2)+(1-o)\log P_{\tilde{R}}(\tau_2\succ\tau_1)\right],$$
$$u_R(\tau) = \max_{\tilde{R}_1,\tilde{R}_2\in\mathcal{R}}|\tilde{R}_1(\tau)-\tilde{R}_2(\tau)|,$$

where $\mathcal{R}=\{\tilde{R}\in\mathcal{F}_R \mid \mathbb{E}_{(\tau_1,\tau_2,o)\sim\mathcal{D}_{pref}}\left[\log P_{\hat{R}}(\tau_1\succ\tau_2)/P_{\tilde{R}}(\tau_1\succ\tau_2)\right]\leq\beta_R\}$ defines the confidence set and β_R is a hyperparameter. We also define preference uncertainty between two trajectories τ_1,τ_2 as $u_{P_R}(\tau_1,\tau_2)=\max_{\tilde{R}_1,\tilde{R}_2\in\mathcal{R}}|P_{\tilde{R}_1}(\tau_1\succ\tau_2)-P_{\tilde{R}_2}(\tau_1\succ\tau_2)|$.

The choice of trajectory sampling strategy for preference elicitation in line 4 is critical to efficiently obtaining an ϵ-optimal policy. We present two possible strategies in Sections 5 and 6. Note that by focusing on sample efficiency as in prior work on preference elicitation [Chen et al., 2022], we do not necessarily optimize for computational efficiency; this could be improved by collecting preferences in batches to reduce the number of reward training loops.

Pessimistic Policy Optimization. Finally, our algorithm outputs a policy $\hat{\pi}^*$ that is optimal while ensuring robustness to modeling error. This means optimizing for the worst-case value function over the remaining transition and reward uncertainties [Levine et al., 2020]:

$$\hat{\pi}^* = \text{argmax}_{\pi\in\Pi}\min_{\tilde{T}\in\mathcal{T},\tilde{R}\in\mathcal{R}}V^{\pi}_{\tilde{T},\tilde{R}}.$$

This analysis provides a worst-case robustness guarantee when considering well-calibrated confidence intervals, as detailed in Sections 5.1 and 6.1. For a practical implementation of our algorithm, we penalize the reward function by the uncertainty as in model-based offline RL methods [Yu et al., 2020, Chang et al., 2021]. Our optimal robust policy therefore maximizes the following objective:

$$\hat{\pi}^* = \text{argmax}_{\pi\in\Pi}\mathbb{E}_{\tau\sim d^{\pi}_{\hat{T}}}[\hat{R}(\tau)-u_R(\tau)-u_T(\tau)]. \tag{2}$$

We show in Appendix A.2 that this is indeed a lower bound of the true value function. This objective allows for controlling the degree of conservatism in practice through the width of the confidence intervals used to determine u_R and u_T.

5 Preference Elicitation from Offline Trajectories

A first strategy for preference elicitation without environment interaction is to sample trajectories directly from the offline dataset. Shin et al. [2022] propose this approach as Offline Preference-based Reward Learning (OPRL), and design a uniform and uncertainty-sampling variant:

OPRL Uniform Sampling: $\tau_1,\tau_2\sim\mathcal{D}_{offline}$

OPRL Uncertainty Sampling: $\tau_1,\tau_2=\text{argmax}_{\tau_1,\tau_2\in\mathcal{D}_{offline}}u_{P_R}(\tau_1,\tau_2)$

5.1 Theoretical Guarantees.

We obtain the following result, demonstrated in Appendix A.4. The suboptimality of the estimated policy $\hat{\pi}^*$ is bounded by the policy evaluation error for the optimal policy π^*. This error decomposes into a term depending on transition model estimation, and one on reward model estimation.

Theorem 5.1. *For any $\delta\in(0,1]$, let $\beta_T=c'_T\log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)/N_o$ and $\beta_R=c'_R\log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)/N_p$, where $N_o=H|\mathcal{D}_{offline}|$ is the number of observed transitions in the observational dataset and c'_T,c'_R are universal constants. The policy $\hat{\pi}^*$ estimated by Algorithm 1, with preference elicitation based on offline trajectories, achieves the following suboptimality with probability $1-\delta$:*

$$V^{\pi^*}-V^{\hat{\pi}^*}\leq\underbrace{HR_{max}C_T(\mathcal{F}_T,\pi^*)\sqrt{\frac{c_T\log(H\mathcal{N}_{\mathcal{F}_T})(1/N_o)/\delta)}{N_o}}}_{\text{transition term }\epsilon_T}+\underbrace{2\alpha\kappa C_R(\mathcal{F}_R,\pi^*)\sqrt{\frac{c_R\log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}}}_{\text{reward term}},$$

where $\alpha=1$ for uniform sampling or $\alpha\leq1$ for uncertainty sampling, C_R is a concentrability measure for the reward function, $\kappa=\sup_{r\in[-R_{max},R_{max}]}\frac{1}{\sigma'(r)}$ measures the degree of non-linearity of the sigmoid function, and c_T,c_R are universal constants.

In the special case where both the transition and reward functions are learned on a fixed initial preference dataset (no preference elicitation), we recover Theorem 1 from Zhan et al. [2023a]. Importantly, parameter α allows us to motivate the superior efficiency of uncertainty sampling over uniform sampling, observed empirically in Shin et al. [2022] and in Section 7.

6 Preference Elicitation from Simulated Trajectories

We now propose our alternative strategy for generating trajectories for offline preference elicitation: **Simulated Offline Preference-based Reward Learning (Sim-OPRL)**. This method simulates trajectories (τ_1, τ_2) by leveraging the *learned environment model*. This overcomes a limitation of OPRL, which is only designed to reduce uncertainty about the reward functions in $\mathcal{R}$, by instead reducing uncertainty about which policies are plausibly optimal. Our approach is inspired by efficient online preference elicitation algorithms [Saha et al., 2023, Chen et al., 2022], which we modify for practical implementation. We account for the offline nature of our problem by avoiding regions out of the distribution of the data: the sampling strategy is optimistic with respect to uncertainty in rewards, but pessimistic with respect to uncertainty in transitions.

We summarize our approach to generating simulated trajectories for preference elicitation in Algorithm 2 and refer the reader to Appendix B for practical implementation details. First, we construct a set of **candidate optimal policies** $\Pi_{offline}$, containing policy $\pi^*_{offline}$ (optimal under the pessimistic model and the true reward function) with high probability – as demonstrated in Appendix A.5.2. Next, within this set of candidate policies, we identify the two most **exploratory policies** π_1, π_2, chosen to maximize preference uncertainty u_{P_R}. Finally, we roll out these policies within our learned transition model to generate a trajectory pair (τ_1, τ_2) for preference feedback.

Algorithm 2 Preference Elicitation through Simulated Trajectory Sampling.

Input: Pessimistic transition model $\hat{T}_{inf}$. Reward confidence set $\mathcal{R}$ and preference uncertainty function u_{P_R}.
Output: (τ_1, τ_2)

1: Estimate optimal offline policy set: $\Pi_{offline} = \{\pi \mid \pi = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{\tau \sim d^{\pi}_{\hat{T}_{inf}}} \left[\tilde{R}(\tau)\right] \forall \tilde{R} \in \mathcal{R}\}$.

2: Identify exploratory policies: $\pi_1, \pi_2 = \text{argmax}_{\pi_1, \pi_2 \in \Pi_{offline}} \mathbb{E}_{\tau_1 \sim d^{\pi_1}_{\hat{T}_{inf}}, \tau_2 \sim d^{\pi_2}_{\hat{T}_{inf}}} \left[u_{P_R}(\tau_1, \tau_2)\right]$

3: Rollouts in model: $\tau_1 \sim d^{\pi_1}_{\hat{T}_{inf}}, \tau_2 \sim d^{\pi_2}_{\hat{T}_{inf}}$.

6.1 Theoretical Guarantees

We decompose suboptimality in a similar way to Section 5.1, but obtain a reward suboptimality term that depends on the learned dynamics model instead of the true one, and on $\pi^*_{offline}$ instead of π^*:

$$V^{\pi^*} - V^{\hat{\pi}^*} \leq \underbrace{(V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf},R})}_{\text{transition term } \epsilon_T} + \underbrace{(V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}})}_{\text{reward term}}. \tag{3}$$

Analysis of the suboptimality due to transition error is identical to above, but the reward term is thus significantly different. By design, our sampling strategy ensures good coverage of preferences over $\pi^*_{offline}$ within the learned environment model, which **eliminates the concentrability term for the reward** C_R. We refer the reader to Appendix A.5 for the proof of Theorem 6.1.

Theorem 6.1. *For any $\delta \in (0, 1]$, let $\beta_T = c'_T \log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)/N_o$ and $\beta_R = c'_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)/N_p$, where $N_o = H|\mathcal{D}_{offline}|$ is the number of observed transitions in the observational dataset and c'_T, c'_R are universal constants. The policy $\hat{\pi}^*$ estimated by Algorithm 1, with a preference sampling strategy based on rollouts in the learned transition model, achieves the following suboptimality with probability $1 - \delta$:*

$$V^{\pi^*} - V^{\hat{\pi}^*} \leq HR_{max}C_T(\mathcal{F}_T, \pi^*)\sqrt{\frac{c_T \log(H\mathcal{N}_{\mathcal{F}_T})(1/N_o)/\delta)}{N_o}} + 2\kappa\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}}.$$

6.2 Discussion

Our theoretical results demonstrate that the learned policy can achieve performance comparable to the optimal policy, and thus satisfy our optimality criterion in Definition 3.1, provided it is covered by the offline data $(C_T(\mathcal{F}_T, \pi^*), C_R(\mathcal{F}_R, \pi^*) < \infty)$. Empirical results in Section 7 confirm that performance is poor when the behavioral policy is suboptimal, inducing a large C_T or C_R.

Offline Trajectories vs. Simulated Rollouts. While both OPRL and Sim-OPRL depend on the offline dataset for estimating environment dynamics, they induce different suboptimality in modeling preference feedback. Simulated rollouts are designed to achieve good coverage of the optimal offline policy $\pi^*_{offline}$, which avoids wasting preference budget on trajectories with low rewards or high transition uncertainty. In contrast, as shown in Zhan et al. [2023a], due to the dependence of preferences on full trajectories, the reward concentrability term C_R in Theorem 5.1 can be large.

Transition vs. Preference Model Quality. Our theoretical analysis also suggests an interesting trade-off in the sample efficiency of our approach, depending on the accuracy of the transition model. The width of the confidence interval reduces as significance parameter δ or dataset size increase, or as function class complexity $\mathcal{N}_{\mathcal{F}_T}$ decreases. For a target suboptimality gap ϵ, provided the optimal offline policy $\pi^*_{offline}$ has a gap $\epsilon_T < \epsilon$, then the number of preferences required is of the order of $\mathcal{O}(\log(1/\delta)/(\epsilon - \epsilon_T)^2)$. A more accurate transition model should therefore require fewer preference samples to achieve a given suboptimality, which we again confirm empirically.

7 Experimental results

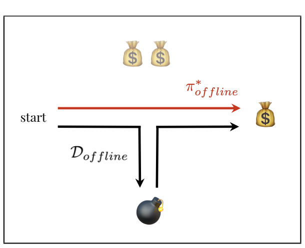
We demonstrate the effectiveness of preference elicitation for offline reinforcement learning in practice and compare the different sampling strategies introduced in Sections 5 and 6: OPRL with uniform and uncertainty-sampling, and Sim-OPRL.

Baselines. For comparison, we also propose a practical implementation of Preference-based Optimistic Planning (PbOP), an uncertainty-based preference elicitation approach over trajectory rollouts in the *true environment* [Chen et al., 2022]. Finally, we report the performance of $\pi^*_{offline}$ and π^* as upper bounds for the performance of our algorithm: the former is trained in the learned transition model with access to the true reward, and the latter has full knowledge of both transition and reward function. We refer the reader to Appendix B for implementation details.

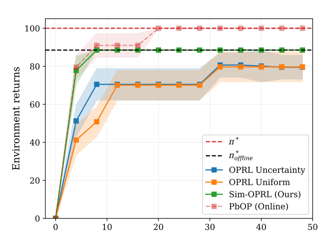
Star MDP. First, consider the tabular MDP in Figure 1a (we defer transition and reward details to Appendix C). Preferences collected over offline trajectories learn slowly about the negative reward in the bottom state, as it is always included in the comparison. Instead, simulated rollouts can directly query the optimal path. We thus find in Figure 1 that our preference elicitation strategy based on simulated rollouts achieves better returns than OPRL approaches, with fewer preference queries.

This example also illustrates the importance of pessimism with respect to the transition model. Even with access to true rewards, $\pi^*_{offline}$ avoids the out-of-distribution state, as it is unclear how to reach it. Thus, in Figure 1c, performance drops if pessimism is not applied to the output policy (purple lines). This confirms theoretical insights from Zhu et al. [2023], who demonstrate the importance of pessimism in offline preference-based RL. Pessimism is also crucial in simulated rollouts, to avoid wasting preference budget on regions of low confidence – as value estimates are in any case inaccurate. This is reflected in the lower efficiency of rollouts without pessimism over $\hat{T}$ in Figure 1c (brown line). We also note the importance of optimism against reward uncertainty, both in OPRL in Figure 1b and in our model-based rollouts in Figure 1c.

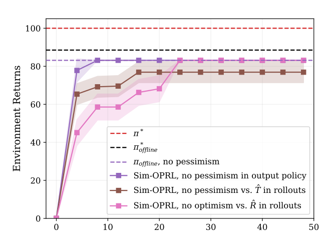
Finally, as an upper bound for the performance of our algorithm, we include baselines that have access to the environment in Figure 1b: the optimal policy π^*, as well as an algorithm querying feedback over real environment rollouts [Chen et al., 2022, PbOP]. Final environment returns are higher than with Sim-OPRL, as they do not suffer from the limited coverage of the transition model. As supported by our theoretical analysis, this result stresses the importance of having a high-quality transition model to make our method effective. We explore this in more detail in the following.



(a) **Star MDP.**



(b) Comparison against baselines.



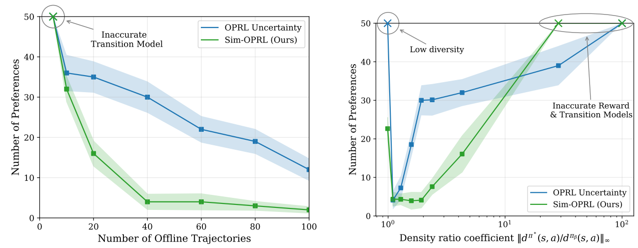
(c) Algorithm Ablations.

Figure 1: **Empirical results on the Star MDP.** Mean and 95% confidence interval over 20 experiments. Environment returns are normalized between 0 and 100. Only OPRL and Sim-OPRL are fully offline, all other methods have access to either environment interaction and/or to the true reward function.

Table 2: **Comparison of preference sample complexity N_p with different sampling methods**, to reach a suboptimality gap of $\epsilon = 20$ over normalized returns. Mean and 95% confidence interval over 20 experiments. The best-performing offline method is highlighted in bold.

Environment	OPRL Uniform	OPRL Uncertainty	**Sim-OPRL (Ours)**	PbOP (Online)
Star MDP (Figure 1a)	32 ± 4	30 ± 4	$\mathbf{4 \pm 2}$	4 ± 2
Gridworld	105 ± 11	66 ± 7	$\mathbf{49 \pm 7}$	32 ± 4
Sepsis Simulation	18,856 $\pm$ 427	2,246 $\pm$ 143	**830 $\pm$ 88**	261 ± 59

Transition vs. Preference Model Quality. Next, we study the trade-off between transition and preference model performance in our problem setting. In the low-data regime, evaluation error due to the misspecification of the transition model is large. As dictated by our theoretical analysis and as visualized in Figure 2a, this increases the number of preferences N_p required to achieve good final performance. Inversely, fewer preferences are needed if the offline dataset is large and the transition model is accurate. We observe a similar trend for both Sim-OPRL and OPRL.



(a) Offline dataset size.

(b) Optimal policy coverage.

Figure 2: **Preference sample complexity N_p as function of the properties of the observational data**, to reach a suboptimality gap of $\epsilon = 20$ over normalized environment returns (Star MDP). Mean and 95% confidence intervals over 20 experiments. $\times$ marks when the target suboptimality could not be achieved.

We also measure how coverage of the optimal policy affects performance. In Figure 2b, we vary the behavioral policy π_β underlying the offline data, ranging from optimal (density ratio 1) to suboptimal (large density ratio). We report the accuracy of transition and reward models in Appendix D. We observe that preference elicitation methods perform best when the data is close to optimal (except for a fully optimal, non-diverse dataset making reward learning from preferences challenging). More preference samples are required if the dataset has poor coverage of the optimal policy (large $C_T(\mathcal{F}_T, \pi^*)$), as transition and reward models become less accurate over the distribution of interest.

Gridworld and Sepsis Simulation. Finally, we validate our findings on more complex environments detailed in Appendix C: a gridworld experiment and a simulation of sepsis management in intensive care [Oberst and Sontag, 2019]. This example highlights another important advantage of Sim-OPRL over OPRL. In a sensitive setting such as healthcare where access is carefully controlled, it may be attractive to query experts about *synthetic* trajectories rather than real samples. Sample complexity results are given in Table 2, with similar conclusions: Sim-OPRL affords a higher preference sampling efficiency than OPRL baselines. For the sepsis environment, we note the number of preference samples needed to achieve our target suboptimality is large, likely due to the sparse nature of the reward function. In a real-world application, we could potentially warm-start the reward model by leveraging proxy rewards signals in the offline data [Yu et al., 2021].

8 Conclusion

Our work shows the potential of integrating human feedback within the framework of offline RL. We address the challenges of preference elicitation in a fully offline setup by exploring two key methods: sampling from the offline dataset [Shin et al., 2022, OPRL] and generating model rollouts (Sim-OPRL). By employing a pessimistic approach to handle out-of-distribution data and an optimistic strategy to acquire informative preferences, Sim-OPRL balances the need for robustness and informativeness in learning an optimal policy. We provide theoretical guarantees on the sample complexity of both approaches, demonstrating that performance depends on how well the offline data covers the optimal policy. Empirical evaluations in various environments confirm the effectiveness of our algorithm, as Sim-OPRL consistently outperforms baselines across different environments.

Overall, our approach not only advances the state-of-the-art in offline preference-based RL but also takes a significant step toward improving the practical utility of offline RL. This opens up new avenues for real-world applications of RL in healthcare, robotics, and manufacturing, where interaction with the environment is challenging but domain experts can be queried for feedback.

References

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

Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast Bayesian reward inference from preferences. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1165–1177. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/brown20a.html`.

Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems*, 34:965–979, 2021.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

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

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. Information directed reward learning for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3850–3862, 2021.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Sungmin Bae, et al. Chip placement with deep reinforcement learning. *arXiv preprint arXiv:2004.10746*, 2020.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. 2018.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 6263–6289. PMLR, 2023.

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

Daniel Shin, Anca Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *Transactions on Machine Learning Research*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

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

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023a.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. How to query human feedback efficiently in rl? In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023b.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

A Theoretical Details

This appendix provides proofs for the presented theorems and lemmas. In subsection A.1, we provide details on how we define the maximum likelihood estimators and confidence intervals of the preference and transition models. In subsection A.2 we provide the proof that our uncertainty-penalized objective in Equation (2) lower bounds the true value function and thus forms a valid pessimistic framework. In Appendix A.3, we show that the suboptimality of our offline preference elicitation framework is not lower-bounded by the performance of the optimal offline policy. In Appendix A.4, we provide our proof of theorem 5.1, analyzing the suboptimality of preferences sampled from an offline dataset. Finally, in Appendix A.5, we prove Theorem 6.1, which analyzes the suboptimality of preference sampling over simulated rollouts.

A.1 Maximum Likelihood and Confidence Intervals

Let $\mathcal{F}_g$ denote a function class over $\mathcal{X} \to \Delta_{\mathcal{Y}}$, where $\mathcal{X}, \mathcal{Y}$ are measurable sets, and $g \in \mathcal{F}_g$ denotes a function to be estimated.

Let $\hat{g}$ denote the maximum likelihood estimator (MLE) of g based on a dataset $\mathcal{D} = \{(x^n, y^n)\}_{n=1}^N$: $\hat{g} = \text{argmax}_{\tilde{g} \in \mathcal{F}_g} \mathbb{E}_{(x,y)\sim\mathcal{D}} \log(\tilde{g}(y|x))$. We construct the confidence set around the MLE as follows:

$$\mathcal{G} = \{\tilde{g} \in \mathcal{F}_g \mid \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\log \frac{\hat{g}(y|x)}{\tilde{g}(y|x)}\right] \leq \beta\}$$

Lemma A.1 (MLE Guarantee, Lemma 1 in Zhan et al. [2023a])**.** *Let $\delta \in (0, 1]$ and define the event $\mathcal{E}$ that $g \in \mathcal{G}$. If*

$$\beta = \frac{c_{MLE} \log(\mathcal{N}_{\mathcal{F}_g}(1/N)/\delta)}{N},$$

where $c_{MLE} > 0$ is a universal constant, then $P(\mathcal{E}) \geq 1 - \delta/2$.

Proof. The proof follows that of Lemma 1 in Zhan et al. [2023a] and uses Cramér-Chernoff's method.

Let $\bar{\mathcal{B}}$ be a $1/N$-bracket of $\mathcal{F}_g$ with $|\bar{\mathcal{B}}| = \mathcal{N}_{\mathcal{F}_g}(1/N)$. Denote the set of all right brackets in $\bar{\mathcal{B}}$ by $\tilde{\mathcal{B}} = \{b : \exists b' s.t. [b', b] \in \bar{\mathcal{B}}\}$. For $b \in \tilde{\mathcal{B}}$, we have:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{n=1}^N \log \frac{b(y^n|x^n)}{g(y^n|x^n)}\right)\right] &= \prod_{n=1}^N \mathbb{E}\left[\exp\left(\log \frac{b(y^n|x^n)}{g(y^n|x^n)}\right)\right] \\
&= \prod_{n=1}^N \mathbb{E}\left[\frac{b(y^n|x^n)}{g(y^n|x^n)}\right] \\
&= \prod_{n=1}^N \mathbb{E}\left[\sum_y b(y|x^n)\right] \\
&\leq (1 + 1/N)^N \leq e.
\end{aligned}$$

as samples in $\mathcal{D}$ as i.i.d. We use the Tower property in the third step and the fact that b is a $1/N$-bracket for $\mathcal{F}_g$ in the fourth: there exists $g' \in \mathcal{F}_g$ such that $\|g(\cdot|x) - b(\cdot|x)\|_1 \leq 1/N$ and thus $\|b(\cdot|x)\|_1 \leq 1 + 1/N$, for all $x \in \mathcal{X}$.

Then by Markov's inequality, for any $\delta \in (0, 1]$, we have:

$$\begin{aligned}
P\left(\sum_{n=1}^N \log \frac{b(y^n|x^n)}{g(y^n|x^n)} > \log(1\delta)\right) &\leq \mathbb{E}\left[\exp\left(\sum_{n=1}^N \log \frac{b(y^n|x^n)}{g(y^n|x^n)}\right)\right] \cdot \exp(-\log(1/\delta)) \\
&\leq e\delta.
\end{aligned}$$

By union bound, we have for all $b \in \tilde{\mathcal{B}}$,

$$P\left(\sum_{n=1}^N \log \frac{b(y^n|x^n)}{g(y^n|x^n)} > c_{MLE} \log(\mathcal{N}_{\mathcal{F}_g}(1/N)/\delta)\right) \leq \delta/2,$$

where $c_{MLE} > 0$ is a universal constant.

Finally, for all $\tilde{g} \in \mathcal{F}_g$, there exists $b \in \tilde{\mathcal{B}}$ such that $g(\cdot|x) \leq \tilde{g}(\cdot|x)$ for all $x \in \mathcal{X}$. As a result, for all $\tilde{g} \in \mathcal{F}_g$, we have:

$$P\left(\sum_{n=1}^{N} \log \frac{\tilde{g}(y^n|x^n)}{g(y^n|x^n)} > c_{MLE} \log(\mathcal{N}_{\mathcal{F}_g}(1/N)/\delta)\right) \leq \delta/2.$$

□

Under this event $\mathcal{E}$, we have $g \in \mathcal{G}$ with probability $1 - \delta/2$. A confidence interval constructed via loglikelihood also incurs a bound on the total variation (TV) distance between g and $\tilde{g} \in \mathcal{G}$:

Lemma A.2 (TV-distance to MLE)**.** *Under the event $\mathcal{E}$, we have, with probability $1-\delta$, for all $\tilde{g} \in \mathcal{G}$:*

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\|g(\cdot|x) - \tilde{g}(\cdot|x)\|_1^2\right] \leq \frac{c \log(\mathcal{N}_{\mathcal{F}_g}(1/N)/\delta)}{N}, \tag{4}$$

where $c > 0$ is a universal constant.

Proof. The proof follows that of Liu et al. [2022], Proposition 14. □

This guarantees that the true reward function is within an interval around the MLE estimate with high probability.

We apply these lemmas to our MLE estimates of transition and reward functions in Algorithm 1 to obtain the following guarantees.

Let $\mathcal{E}_R$ denote the event $R \in \mathcal{R}$ and $\mathcal{E}_T$ denote the event $T \in \mathcal{T}$, $\mathcal{R}$ and $\mathcal{T}$ denote the respective confidence sets around the MLE. By Lemma A.1, events $\mathcal{E}_R$ and $\mathcal{E}_T$ have probability $1 - \delta/2$ if we choose $\beta_R = c'_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)/N_p$ and $\beta_T = c'_T \log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)/N_o$, where c'_R, c'_T are universal constants.

A.2 Model-based Pessimism and Uncertainty Penalties

Lemma A.3 (Telescoping Lemma)**.** *For any reward model $R \in \mathcal{F}_R$, and any two transition models $T, \hat{T} \in \mathcal{F}_T$:*

$$V^{\pi}_{T,R} - V^{\pi}_{\hat{T},R} \leq \mathbb{E}_{\tau\sim d^{\pi}_{\hat{T}}}\left[\sum_{s_j,a_j\in\tau} \|T(\cdot|s_j,a_j) - \hat{T}(\cdot|s_j,a_j)\|_1\right] \cdot R_{max}$$

Proof. The proof follows that of Lemma 4.1 in Yu et al. [2020] or Lemma 4 in Zhan et al. [2023a].

Let W_j be the expected return under policy π, with transition model $\hat{T}$ for the first j steps, then transition model T for the rest of the episode. We have:

$$V^{\pi}_{T,R} - V^{\pi}_{\hat{T},R} = \sum_{j=0}^{H-1} W_j - W_{j+1}.$$

Now,

$$W_j = R_j + \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\mathbb{E}_{s_{j+1}\sim T(\cdot|s_j,a_j)}[V^{\pi}_{T,R}(s_{j+1})]\right]$$
$$W_{j+1} = R_j + \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\mathbb{E}_{s_{j+1}\sim\hat{T}(s_j,a_j)}[V^{\pi}_{T,R}(s_{j+1})]\right]$$

where R_j is the expected return of the first j steps taken in $\hat{T}$. Therefore,

$$\begin{aligned} W_j - W_{j+1} &= \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\mathbb{E}_{s_{j+1}\sim T(\cdot|s_j,a_j)}[V^{\pi}_{T,R}(s_{j+1})] - \mathbb{E}_{s_{j+1}\sim\hat{T}(s_j,a_j)}[V^{\pi}_{T,R}(s_{j+1})]\right] \\ &\leq \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\|T(\cdot|s_j,a_j) - \hat{T}(\cdot|s_j,a_j)\|_1 \cdot R_{max}\right] \end{aligned}$$

under the boundedness assumption for R. Finally, we have:

$$\begin{aligned}
V^\pi_{T,R} - V^\pi_{\hat{T},R} &= \sum_{j=0}^{H-1} W_j - W_{j+1} \\
&= \sum_{j=0}^{H-1} \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\mathbb{E}_{s_{j+1}\sim T(\cdot|s_j,a_j)}[V^\pi_{T,R}(s_{j+1})] - \mathbb{E}_{s_{j+1}\sim \hat{T}(s_j,a_j)}[V^\pi_{T,R}(s_{j+1})]\right] \\
&\leq \sum_{j=0}^{H-1} \mathbb{E}_{s_j,a_j\sim\pi,\hat{T}}\left[\|T(\cdot|s_j,a_j) - \hat{T}(\cdot|s_j,a_j)\|_1 \cdot R_{max}\right] \\
&= \mathbb{E}_{\tau\sim d^\pi_{\hat{T}}}\left[\sum_{s_j,a_j\in\tau} \|T(\cdot|s_j,a_j) - \hat{T}(\cdot|s_j,a_j)\cdot R_{max}\|_1\right]
\end{aligned}$$

□

Lemma A.4 (Pessimistic Transition Model)**.** *Under event $\mathcal{E}_T$, for all $\pi\in\Pi, \tilde{R}\in\mathcal{F}_R$:*

$$V^\pi_{\hat{T},\tilde{R}-u_T} \leq V^\pi_{T,\tilde{R}}.$$

Proof.

$$\begin{aligned}
V^\pi_{T,\tilde{R}} &= V^\pi_{\hat{T},\tilde{R}} - (V^\pi_{\hat{T},\tilde{R}} - V^\pi_{T,\tilde{R}}) \\
&\geq \mathbb{E}_{\tau\sim d^\pi_{\hat{T}}}\left[\tilde{R}(\tau)\right] - \mathbb{E}_{\tau\sim d^\pi_{\hat{T}}}[u_T(\tau)] \\
&= \mathbb{E}_{\tau\sim d^\pi_{\hat{T}}}\left[\tilde{R}(\tau) - u_T(\tau)\right]
\end{aligned}$$

where we have used the telescoping lemma (Lemma A.3), and where $u_T(\tau) = \sum_{(s,a)\in\tau} u_T(s,a) \geq \sum_{(s,a)\in\tau} \|\hat{T}(\cdot|s,a) - T(\cdot|s,a)\|_1 \cdot R_{max}$ under event $\mathcal{E}_T$. □

Lemma A.5 (Pessimistic Reward Model)**.** *Under event $\mathcal{E}_R$, for all $\pi\in\Pi, \tilde{T}\in\mathcal{F}_T$:*

$$V^\pi_{\tilde{T},\hat{R}-u_R} \leq V^\pi_{\tilde{T},R}.$$

Proof.

$$\begin{aligned}
V^\pi_{\tilde{T},R} &= V^\pi_{\tilde{T},\hat{R}} - (V^\pi_{\tilde{T},\hat{R}} - V^\pi_{\tilde{T},R}) \\
&= \mathbb{E}_{\tau\sim d^\pi_{\tilde{T}}}\left[\hat{R}(\tau)\right] - \mathbb{E}_{\tau\sim d^\pi_{\tilde{T}}}\left[\hat{R}(\tau) - R(\tau)\right] \\
&\geq \mathbb{E}_{\tau\sim d^\pi_{\tilde{T}}}\left[\hat{R}(\tau) - u_R(\tau)\right]
\end{aligned}$$

where we have used the fact that $|\hat{R}(\tau) - R(\tau)| \leq \sum_{s,a\in\tau} |\hat{R}(s,a) - R(s,a)| = \sum_{(s,a)\in\tau} u_R(s,a) = u_R(\tau)$ under event $\mathcal{E}_R$. □

Combining the above two lemmas gives the following result:

Corollary A.1. *Under events $\mathcal{E}_T$ and $\mathcal{E}_R$, for all $\pi\in\Pi$:*

$$V^\pi_{\hat{T},\hat{R}-u_T-u_R} \leq V^\pi_{T,R}.$$

This justifies the overall objective considered in our pessimistic policy optimization procedure in Section 4.

A.3 Suboptimality lower bound: a counterexample

Let $\pi^*_{offline} = \text{argmax}_{\pi\in\Pi} \min_{\tilde{T}\in\mathcal{T}} V^\pi_{\tilde{T},R}$ denote the optimal offline policy, which has access to the ground-truth reward function. In this section, we ask whether its suboptimality $\epsilon_T = V^{\pi^*}_{T,R} - V^{\pi^*_{offline}}_{T,R}$ is a lower bound for the suboptimality of our learned policy $\hat{\pi}^*$ after preference elicitation.

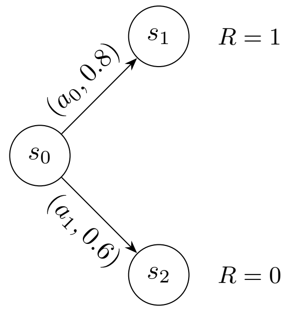


Figure 3: **Tabular MDP.** The environment starts in state s_0 and has horizon $H = 1$. Transition probabilities from state s_0 are given for the two binary actions a_0, a_1 (which send the agent to the other state with complementary probability).

Counterexample. Consider the MDP illustrated in Figure 3. Assume the following MLE estimate and uncertainty function for both the transition and reward models:

$$\begin{aligned} \hat{T}(s_1|s_0, a_0) &= 0.5; & u_T(s_0, a_0) &= 0.4 \\ \hat{T}(s_1|s_0, a_1) &= 0.5; & u_T(s_0, a_1) &= 0.1 \\ \hat{r}(s_1) = \hat{r}(s_2) &= 0.5; & u_R(s_1) = u_R(s_2) &= 0.5 \end{aligned}$$

Assuming access to the learned transition model and the *true* reward function, we pessimistically estimate the value of both actions:

$$\begin{aligned} V^{a_0}_{\hat{T}_{inf}, R} &= 0.1 \cdot 1 + 0.9 \cdot 0 = 0.1 \\ V^{a_1}_{\hat{T}_{inf}, R} &= 0.6 \cdot 0 + 0.4 \cdot 1 = 0.4 \end{aligned}$$

Thus, we have: $\pi^*_{offline}(s_0) = \text{argmax}_a V^a_{\hat{T}_{inf}, R} = a_1$. The offline policy picks the suboptimal action since the worst-case returns of this action are lower than those estimated for a_0. Evaluating this policy in the real environment, we get $\epsilon_T = V^{\pi^*}_{T,R} - V^{\pi^*_{offline}}_{T,R} = 0.6 \cdot 0 + 0.4 \cdot 1 = 0.4$.

We now estimate the optimal policy in the learned transition and reward model. Applying pessimism with respect to both models, we get an equal estimated value of 0 for both actions a_0 and a_1. If policy optimization converges to $\hat{\pi}^* = a_0$, we reach the suboptimality $V^{\pi^*}_{T,R} - V^{\hat{\pi}^*}_{T,R} = 0.8 \cdot 1 + 0.2 \cdot 0 = 0.8 > \epsilon_T$.

This example demonstrates that ϵ_T is not a lower bound for the suboptimality of $\hat{\pi}^*$, as policy $\hat{\pi}^*$ can achieve lower suboptimality than $\pi^*_{offline}$ if errors in transition and reward model estimation compensate each other.

A.4 Suboptimality of OPRL: Proof of Theorem 5.1

A.4.1 Suboptimality Decomposition

Recall that $\hat{T}_{inf}, \hat{R}_{inf} = \text{argmin}_{\hat{T} \in \mathcal{T}, \hat{R} \in \mathcal{R}} V^{\pi}_{\hat{T}, \hat{R}}$ denote the pessimistic transition and reward models, such that $\hat{\pi}^* = \text{argmax}_{\pi \in \Pi} V^{\pi}_{\hat{T}_{inf}, \hat{R}_{inf}}$. We have:

$$\begin{aligned} V^{\pi^*} - V^{\hat{\pi}^*} &= V^{\pi^*}_{T,R} - V^{\hat{\pi}^*}_{T,R} \\ &= (V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf}, \hat{R}_{inf}}) - (V^{\hat{\pi}^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf}, \hat{R}_{inf}}) \\ &\leq (V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf}, \hat{R}_{inf}}) - (V^{\hat{\pi}^*}_{T,R} - V^{\hat{\pi}^*}_{\hat{T}_{inf}, \hat{R}_{inf}}) \\ &\leq V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf}, \hat{R}_{inf}}, \end{aligned} \tag{5}$$

where we have first used the optimality of $\hat{\pi}^*$ (stating that $V^{\hat{\pi}^*}_{\hat{T}_{inf}, \hat{R}_{inf}} \geq V^{\pi}_{\hat{T}_{inf}, \hat{R}_{inf}}$, for all π) and then the pessimism principle (stating that $V^{\hat{\pi}^*}_{\hat{T}_{inf}, \hat{R}_{inf}} \leq V^{\hat{\pi}^*}_{T,R}$).

Finally, we decompose the last term above as follows:

$$V^{\pi^*} - V^{\hat{\pi}^*} \leq \underbrace{(V^{\pi^*}_{T,\hat{R}_{inf}} - V^{\pi^*}_{\hat{T}_{inf},\hat{R}_{inf}})}_{\text{transition term}} + \underbrace{(V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}})}_{\text{reward term}} \tag{6}$$

We further analyze each term in the following sections.

A.4.2 Analysis of the transition term

In this section, we now upper bound the transition term defined in Equation (6).

Lemma A.6 (Lemma 4, Zhan et al. [2023a]). *Under the event $\mathcal{E}_T$, with probability $1-\delta$, we have for all $\tilde{T} \in \mathcal{T}$, for all $\tilde{R} \in \mathcal{G}_R$, for all π:*

$$\mathbb{E}_{d^\pi_T}[\tilde{R}(\tau)] - \mathbb{E}_{d^\pi_{\tilde{T}}}[\tilde{R}(\tau)] \leq HR_{max}C_T(\mathcal{F}_T, \pi)\sqrt{\frac{c_T \log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)}{N_o}},$$

where $c_T > 0$ is a constant.

Proof. From the telescoping lemma (Lemma A.3), we have:

$$\begin{aligned} V^\pi_{T,\tilde{R}} - V^\pi_{\tilde{T},\tilde{R}} &\leq R_{max}\mathbb{E}_{\tau\sim d^\pi_T}\left[\sum_{s_j,a_j\in\tau}\|T(\cdot|s_j,a_j) - \tilde{T}(\cdot|s_j,a_j)\|_1\right] \\ &\leq HR_{max}\mathbb{E}_{(s,a)\sim d^\pi_T}\left[\|T(\cdot|s,a) - \tilde{T}(\cdot|s,a)\|_1\right] \\ &\leq HR_{max}C_T(\mathcal{F}_T,\pi)\sqrt{\mathbb{E}_{(s,a)\sim D_{offline}}[\|T(\cdot|s,a) - \tilde{T}(\cdot|s,a)\|_1^2]} \end{aligned}$$

Under event $\mathcal{E}_T$, by Lemma A.2, we have, with probability $1-\delta$, for all $\tilde{T} \in \mathcal{T}$:

$$\mathbb{E}_{(s,a)\sim D_{offline}}[\|T(\cdot|s,a) - \tilde{T}(\cdot|s,a)\|_1^2] \leq \frac{1}{N_o}c_T \log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)$$

This concludes our proof.

$\square$

A.4.3 Analysis of the reward term

Next, we upper bound the reward term defined in Equation (6).

As in Zhan et al. [2023a], we consider the following value function: $V^\pi_{T,R} = \mathbb{E}_{\tau\sim d^\pi_T}[R(\tau)] - \mathbb{E}_{\tau\sim d_{pref}}[R(\tau)]$, where d_{pref} is a fixed reference trajectory distribution. This baseline subtraction, which doesn't affect either the optimal policy or the analysis of the transition term, is needed as the approximated confidence set is based on the uncertainty in *preference* between two trajectories, not in the reward of a single one.

Definition A.1 (Preference concentrability coefficient). *The concentrability coefficient w.r.t. reward classes $\mathcal{F}_R$, a target policy π^* and a reference trajectory distribution d_{pref} is defined as:*

$$C_R(\mathcal{F}_R, \pi^*) = \frac{\mathbb{E}_{\tau_1\sim d^{\pi^*}_T, \tau_2\sim d_{pref}}[u_{P_R}(\tau_1,\tau_2)]}{\mathbb{E}_{\tau_1,\tau_2\sim\mathcal{D}_{offline}}[u_{P_R}(\tau_1,\tau_2)]}$$

Note that, for the purpose of our analysis, our definition differs from that of Zhan et al. [2023a] who instead consider the max ratio of difference in rewards term: $|R(\tau_1) - R(\tau_2) - \tilde{R}(\tau_1) + \tilde{R}(\tau_2)|$ over the entire function class $\mathcal{F}_\mathcal{R}$.

Lemma A.7. *Let trajectories for preference elicitation be sampled uniformly from the offline dataset. Under the event $\mathcal{E}_R$, with probability $1-\delta$, we have for all $\tilde{T} \in \mathcal{G}_T$, for all $\tilde{R} \in \mathcal{R}$, for all π:*

$$V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} \leq 2\kappa C_R(\mathcal{F}_T,\pi)\sqrt{\frac{c_R\log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}},$$

where $c_R > 0$ is a constant and $\kappa = \sup_{r \in [-R_{max}, R_{max}]} \frac{1}{\sigma'(r)}$ measures the degree of non-linearity of the sigmoid function.

Proof.

$$\begin{aligned}
V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} &= \mathbb{E}_{\tau \sim d^{\pi^*}_T}[R(\tau)] - \mathbb{E}_{\tau \sim d_{pref}}[R(\tau)] - \mathbb{E}_{\tau \sim d^{\pi^*}_T}[\hat{R}_{inf}(\tau)] + \mathbb{E}_{\tau \sim d_{pref}}[\hat{R}_{inf}(\tau)] \\
&= \mathbb{E}_{\tau_1 \sim d^{\pi^*}_T, \tau_2 \sim d_{pref}}[R(\tau_1) - R(\tau_2)] - (\hat{R}_{inf}(\tau_1) - \hat{R}_{inf}(\tau_2))] \\
&\leq \kappa \mathbb{E}_{\tau_1 \sim d^{\pi^*}_T, \tau_2 \sim d_{pref}}[|P_R(\tau_1 \succ \tau_2) - P_{\hat{R}_{inf}}(\tau_1 \succ \tau_2)|] \\
&\leq \kappa \mathbb{E}_{\tau_1 \sim d^{\pi^*}_T, \tau_2 \sim d_{pref}}[u_{P_R}(\tau_1, \tau_2)] \\
&= \kappa C_R(\mathcal{F}_T, \pi^*) \mathbb{E}_{\tau_1, \tau_2 \sim \mathcal{D}_{offline}}[u_{P_R}(\tau_1, \tau_2)] \qquad (7)
\end{aligned}$$

where $\kappa = \sup_{r \in [-R_{max}, R_{max}]} \frac{1}{\sigma'(r)}$ measures the degree of non-linearity of the sigmoid function. In the first inequality, we have applied the mean value theorem, under Assumption 3.2. In the second inequality, we have used the definition of uncertainty function u_{P_R} as we know $\hat{R}_{inf} \in \mathcal{R}$.

Now, under event $\mathcal{E}_R$, by Lemma A.2, we have, with probability $1 - \delta$ for all $\tilde{R} \in \mathcal{R}$:

$$\mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{pref}}[\|P_R(\tau_1 \succ \tau_2) - P_{\tilde{R}}(\tau_1 \succ \tau_2)\|_1^2] \leq \frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}, \qquad (8)$$

where $c_R > 0$ is a constant. This implies the following upper bound for the preference uncertainty function:

$$\mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{pref}}[u_{P_R}(\tau_1, \tau_2)] \leq 2\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_g}(1/N_p)/\delta)}{N_p}} \qquad (9)$$

Under uniform sampling, the distribution of preferences in $\mathcal{D}_{pref}$ is that of the offline dataset:

$$\mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{offline}}[u_{P_R}(\tau_1, \tau_2)] = \mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{pref}}[u_{P_R}(\tau_1, \tau_2)]$$

Thus,

$$V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} \leq 2\kappa C_R(\mathcal{F}_T, \pi)\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}}.$$

$\square$

Lemma A.8. *Let trajectories for preference elicitation be sampled through uncertainty sampling from the offline dataset. Under the event $\mathcal{E}_R$, with probability $1 - \delta$, we have for all $\tilde{T} \in \mathcal{G}_T$, for all $\tilde{R} \in \mathcal{R}$, for all π:*

$$V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} \leq 2\alpha\kappa C_R(\mathcal{F}_T, \pi)\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}},$$

where $c_R > 0$ is a constant and $\alpha \leq 1$.

Proof. The proof follows closely that of Lemma A.7. We introduce the preference concentrability coefficient defined for a general preference dataset:

$$C'_R(\mathcal{F}_R, \pi^*) = \frac{\mathbb{E}_{\tau_1 \sim d^{\pi^*}_T, \tau_2 \sim d_{pref}}[u_{P_R}(\tau_1, \tau_2)]}{\mathbb{E}_{\tau_1, \tau_2 \sim \mathcal{D}_{pref}}[u_{P_R}(\tau_1, \tau_2)]}$$

We start from Equation (7):

$$\begin{aligned}
V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} &\leq \kappa \mathbb{E}_{\tau_1 \sim d^{\pi^*}_T, \tau_2 \sim d_{pref}}[u_{P_R}(\tau_1, \tau_2)] \\
&= \kappa C'_R(\mathcal{F}_T, \pi^*) \mathbb{E}_{\tau_1, \tau_2 \sim \mathcal{D}_{pref}}[u_{P_R}(\tau_1, \tau_2)] \\
&\leq 2\kappa C'_R(\mathcal{F}_T, \pi^*)\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_g}(1/N_p)/\delta)}{N_p}}
\end{aligned}$$

where we have used Equation (9).

Now consider the dataset of uncertainty-sampled preferences $\mathcal{D}_{pref}$. By definition, we have:

$$\mathbb{E}_{\tau_1,\tau_2 \sim \mathcal{D}_{pref}} \left[u_{P_R}(\tau_1, \tau_2)\right] \geq \mathbb{E}_{\tau_1,\tau_2 \sim \mathcal{D}_{offline}} \left[u_{P_R}(\tau_1, \tau_2)\right]$$

Thus, we have: $C'_R(\mathcal{F}_T, \pi^*) \leq C_R(\mathcal{F}_T, \pi^*)$. In other words, we can write: $C'_R(\mathcal{F}_T, \pi^*) = \alpha C_R(\mathcal{F}_T, \pi^*)$, where $\alpha \leq 1$. This concludes our proof. □

We now conclude the proof of Theorem 5.1 under events $\mathcal{E}_R$ and $\mathcal{E}_T$.

From Lemma A.6, we upper bound the transition term:

$$V^{\pi^*}_{T,\hat{R}_{inf}} - V^{\pi^*}_{\hat{T}_{inf},\hat{R}_{inf}} \leq HR_{max} C_T(\mathcal{F}_T, \pi^*) \sqrt{\frac{c_T \log(H\mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)}{N_o}}$$

From Lemmas A.7 and A.8, we upper bound the reward term:

$$V^{\pi^*}_{T,R} - V^{\pi^*}_{T,\hat{R}_{inf}} \leq 2\alpha\kappa C_R(\mathcal{F}_T, \pi^*) \sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}},$$

where $\alpha = 1$ for uniform sampling or $\alpha \leq 1$ for uncertainty sampling.

Combining with Equation (6), we obtain Theorem 5.1.

A.5 Suboptimality of Sim-OPRL: Proof of Theorem 6.1

A.5.1 Suboptimality Decomposition

We decompose the suboptimality slightly differently to Equation (5), introducing the optimal offline policy (optimal in the pessimistic model under the *true* reward function): $\pi^*_{offline} = \text{argmax}_{\pi \in \Pi} V^{\pi}_{\hat{T}_{inf},R}$.

$$\begin{aligned}
V^{\pi^*} - V^{\hat{\pi}^*} &= V^{\pi^*}_{T,R} - V^{\hat{\pi}^*}_{T,R} \\
&= (V^{\pi^*}_{T,R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}}) - (V^{\hat{\pi}^*}_{T,R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}}) \\
&\leq (V^{\pi^*}_{T,R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}}) - (V^{\hat{\pi}^*}_{T,R} - V^{\hat{\pi}^*}_{\hat{T}_{inf},\hat{R}_{inf}}) \\
&\leq V^{\pi^*}_{T,R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}} \\
&= (V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf},R}) + (V^{\pi^*}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}}) \\
&\leq \underbrace{(V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf},R})}_{\text{transition term}} + \underbrace{(V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}})}_{\text{reward term}}
\end{aligned} \tag{10}$$

where we have followed the same analysis as in Appendix A.4.1 and used the optimality of $\pi^*_{offline}$ in the last inequality.

The analysis of the transition term is identical to the above (Appendix A.4.2). We analyze the reward term next.

A.5.2 Analysis of the reward term

Lemma A.9 (Optimal Offline Policy In Set)**.** *Let $\Pi_{offline}$ denote the following set of near-optimal pessimistic policies, under the pessimitic transition model $\hat{T}_{inf}$ and the reward confidence set $\mathcal{R}$:*

$$\Pi_{offline} = \{\pi \mid \pi = \textit{argmax}_{\pi \in \Pi} \mathbb{E}_{\tau \sim d^{\pi}_{\hat{T}_{inf}}} \left[\tilde{R}(\tau)\right] \forall \tilde{R} \in \mathcal{R}\}$$

Under event $\mathcal{E}_R$, we have $\pi^*_{offline} \in \Pi_{offline}$.

Proof. Recall the definition of $\pi^*_{offline}$: $\pi^*_{offline} = \text{argmax}_{\pi \in \Pi} V^{\pi}_{\hat{T}_{inf},R}$. Note that there is no need to consider the preference baseline term in V^{π} when building $\Pi_{offline}$ since it is independent of the policy. Under event $\mathcal{E}_R$, we have $R \in \mathcal{R}$. Thus, $\pi^*_{offline} \in \Pi_{offline}$. $\square$

Lemma A.10. *Under event $\mathcal{E}_R$, we have, with probability $1-\delta$:*

$$V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}} \leq 2\sqrt{\kappa^2 c_R / N_p \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}$$

Proof.

$$\begin{aligned}
&V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}} \\
&\qquad = (V^{\pi^*}_{\hat{T}_{inf},R} - V^{\pi^*}_{\hat{T}_{inf},\hat{R}}) + (V^{\pi^*}_{\hat{T}_{inf},\hat{R}} - V^{\pi^*}_{\hat{T}_{inf},\hat{R}_{inf}}) \\
&\qquad = \mathbb{E}_{\tau \sim d^{\pi^*_{offline}}_{\hat{T}_{inf}}}[R(\tau)] - \mathbb{E}_{\tau \sim d_{pref}}[R(\tau)] - \mathbb{E}_{\tau \sim d^{\pi^*_{offline}}_{\hat{T}_{inf}}}[\hat{R}_{inf}(\tau)] + \mathbb{E}_{\tau \sim d_{pref}}[\hat{R}_{inf}(\tau)] \\
&\qquad = \mathbb{E}_{\tau_1 \sim d^{\pi^*_{offline}}_{\hat{T}_{inf}}, \tau_2 \sim d_{pref}}[R(\tau_1) - R(\tau_2)] - \mathbb{E}_{\tau_1 \sim d^{\pi^*_{offline}}_{\hat{T}_{inf}}, \tau_2 \sim d_{pref}}[\hat{R}_{inf}(\tau_1) - \hat{R}_{inf}(\tau_2)] \\
&\qquad \leq \kappa \mathbb{E}_{\tau_1 \sim d^{\pi^*_{offline}}_{\hat{T}_{inf}}, \tau_2 \sim d_{pref}}[P_R(\tau_1 \succ \tau_2) - P_{\hat{R}_{inf}}(\tau_1 \succ \tau_2)],
\end{aligned}$$

where $\kappa = \sup_{r \in [-R_{max}, R_{max}]} \frac{1}{\sigma'(r)}$ measures the degree of non-linearity of the sigmoid function. We have applied the mean value theorem, under Assumption 3.2.

As $R_{inf} \in \mathcal{R}$, we have: $P_R(\tau_1 \succ \tau_2) - P_{\hat{R}_{inf}}(\tau_1 \succ \tau_2) \leq u_{P_R}(\tau_1, \tau_2)$.

Let d_{pref} correspond to the distribution of the preference data, which consists of rollouts from exploratory policies within the learned environment model: $d_{pref} = d^{\pi_1}_{\hat{T}_{inf}}/2 + d^{\pi_2}_{\hat{T}_{inf}}/2$. Recall that the near-optimal policy set $\Pi_{offline}$ includes policy $\pi^*_{offline}$ (Lemma A.9) and that π_1, π_2 are the two more exploratory policies within this set:

$$\mathbb{E}_{\tau_1 \sim d^{\pi^*_{offline}}_{\hat{T}}, \tau_2 \sim d_{pref}}[u_{P_R}(\tau_1, \tau_2)] \leq \max_{\pi_1, \pi_2 \in \Pi_{offline}} \mathbb{E}_{\tau_1 \sim d^{\pi_1}_{\hat{T}}, \tau_2 \sim d^{\pi_2}_{\hat{T}}}[u_{P_R}(\tau_1, \tau_2)].$$

Now, under event $\mathcal{E}_R$, by Lemma A.2, we have, with probability $1-\delta$ for all $\tilde{R} \in \mathcal{R}$:

$$\mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{pref}}[\|P_R(\tau_1 \succ \tau_2) - P_{\tilde{R}}(\tau_1 \succ \tau_2)\|_1^2] \leq \frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p},$$

where $c_R > 0$ is a constant. This implies the following upper bound for the preference uncertainty function:

$$\mathbb{E}_{(\tau_1,\tau_2) \sim \mathcal{D}_{pref}}[u_{P_R}(\tau_1, \tau_2)] \leq 2\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_g}(1/N_p)/\delta)}{N_p}}.$$

Thus, we obtain:

$$V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}} \leq 2\kappa\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_g}(1/N_p)/\delta)}{N_p}}.$$

The resulting sample complexity of $\mathcal{O}(\frac{\kappa^2 d}{\epsilon^2})$ matches that of active preference learning within a known environment [Saha et al., 2023, Chen et al., 2022].

$\square$

We now conclude the proof of Theorem 6.1 under events $\mathcal{E}_R$ and $\mathcal{E}_T$.

From Lemma A.6, we upper bound the transition term:

$$V^{\pi^*}_{T,R} - V^{\pi^*}_{\hat{T}_{inf},R} \leq H R_{max} C_T(\mathcal{F}_T, \pi^*) \sqrt{\frac{c_T \log(H \mathcal{N}_{\mathcal{F}_T}(1/N_o)/\delta)}{N_o}}.$$

From Lemma A.10, we upper bound the reward term:

$$V^{\pi^*_{offline}}_{\hat{T}_{inf},R} - V^{\pi^*_{offline}}_{\hat{T}_{inf},\hat{R}_{inf}} \leq 2\kappa\sqrt{\frac{c_R \log(\mathcal{N}_{\mathcal{F}_R}(1/N_p)/\delta)}{N_p}}.$$

Combining with Equation (10), we obtain Theorem 6.1.

B Implementation Details

We trained all models on two 64-core AMD processors or a single NVIDIA RTX2080Ti GPU. The total wall-clock time for running all experiments presented in this paper amounted to less than 72 hours.

Transition and Reward Function Training. For all baselines, transition and reward models were implemented as linear classifiers (for the Star MDP) or as two-layer perceptions with ReLU activation and hidden layer dimension 32 (Gridworld and Sepsis environments). Training was carried out for two or one epochs for the transition and reward models respectively, with the Adam optimizer [Kingma and Ba, 2014] and a learning rate of 10^{-3}.

We provide a more detailed practical algorithm for Sim-OPRL in Algorithm 3. For both our method and baselines relying on uncertainty sets (OPRL and PbOP), we estimated uncertainty sets by training models initialized with different random seeds on different bootstraps of the data (sampling 90% of the data with replacement). We consider ensembles of size $|\mathcal{T}| = |\mathcal{R}| = 5$ for both transition and reward models. Hyperparameters λ_T, λ_R control the degree of pessimism in practice and could be considered equivalent to adjusting margin parameters β_T, β_R in our conceptual algorithm proposed in Section 4. Since the exact values prescribed by our theoretical analysis cannot be estimated, the user must set these parameters themselves. Hyperparameter optimization in offline RL is a challenging problem [Levine et al., 2020]; for our experiments, we simply set $\lambda_T = 0.5$, $\lambda_R = 0.1$ (StarMDP, Gridworld) and $\lambda_T = \lambda_R = 1$ for the Sepsis environment.

Algorithm 3 Sim-OPRL: Practical Algorithm

Input: Observational trajectories dataset $\mathcal{D}_{offline}$. Hyperparameters λ_T, λ_R.
Output: $\hat{\pi}^*$

Train an ensemble $\mathcal{T}$ of transition models via bootstrapping on the observational data $\mathcal{D}_{offline}$:

$$\hat{T}(\cdot|s,a) = \frac{1}{|\mathcal{T}|}\sum_{\tilde{T}\in\mathcal{T}} \tilde{T}(\cdot|s,a); \quad u_T(s,a) = \max_{T_1,T_2\in\mathcal{T}} |T_1(\cdot|s,a) - T_2(\cdot|s,a)| \cdot R_{max}$$

$\mathcal{D}_{pref} \leftarrow \emptyset$.
for $k = 1, ...N_p$ **do**
  Estimate optimal offline policy set:

$$\Pi_{offline} = \{\pi \mid \pi = \text{argmax}_{\pi\in\Pi}\mathbb{E}_{(s,a)\sim d^\pi_{\hat{T}}}\left[\tilde{R}(s,a) - \lambda_T u_T(s,a)\right] \forall \tilde{R} \in \mathcal{R}\}$$

  Identify exploratory policies: $\pi_1, \pi_2 = \text{argmax}_{\pi_1,\pi_2\in\Pi_{offline}}\mathbb{E}_{\tau_1\sim d^{\pi_1}_{\hat{T}},\tau_2\sim d^{\pi_2}_{\hat{T}}}\left[u_{P_R}(\tau_1,\tau_2)\right]$
  Rollouts in model: $\tau_1 \sim d^{\pi_1}_{\hat{T}}, \tau_2 \sim d^{\pi_2}_{\hat{T}}$.
  Collect preference label o for (τ_1, τ_2).
  $\mathcal{D}_{pref} \leftarrow \mathcal{D}_{pref} \cup \{(\tau_1, \tau_2, o)\}$.
  Train an ensemble $\mathcal{R}$ of reward models via bootstrapping of the preference data $\mathcal{D}_{pref}$:

$$\hat{R}(s,a) = \frac{1}{|\mathcal{R}|}\sum_{\tilde{R}\in\mathcal{R}} \tilde{R}(s,a); \quad u_R(s,a) = \max_{R_1,R_2\in\mathcal{R}} |R_1(\cdot|s,a) - R_2(\cdot|s,a)|$$

end for
$\hat{\pi}^* \leftarrow \text{argmax}_{\pi\in\Pi}\mathbb{E}_{(s,a)\sim d^\pi_{\hat{T}}}[\hat{R}(s,a) - \lambda_R u_R(s,a) - \lambda_T u_T(s,a)]$

Near-Optimal Policy Set and Exploratory Policies. Both Sim-OPRL and PbOP require constructing a set of near-optimal policies within a learned model of the environment. Note that the PbOP algorithm in Chen et al. [2022] proposes to construct the near-optimal policy set by considering all policies that have a preference greater than $1/2$ over *all other policies in* Π, under a transition and preference uncertainty bonus. This is infeasible to estimate in practice; we modified the algorithm to allow for practical implementation. The motivation in building the set of plausibly optimal policies remains the same, but the theoretical guarantees may not hold.

We build $\Pi_{offline}$ by maintaining a policy model for all $\tilde{R} \in \mathcal{R}$, i.e., each element of the reward ensemble. Policy models are optimized to maximize returns under the transition model $\hat{T}$ and the reward function $\tilde{R} - \lambda_T u_T$ (Sim-OPRL) or $\tilde{R} + \lambda_T u_T$ (PbOP). Next, the most exploratory policies are identified by generating 10 rollouts of each of the candidate policies within the learned (SimOPRL) or true (PbOP) model. The trajectories (τ_1, τ_2) maximizing the preference uncertainty function $u_{P_R}(\tau_1, \tau_2)$ are used for preference feedback. In PbOP, the trajectories are then added to the trajectories buffer and the transition model is retrained for 20 (Star MDP, Gridworld) or 200 steps (Sepsis).

Preference Feedback Collection. Preference labels are provided through the ground-truth reward function associated with every environment. As stated in Section 4, for computational efficiency, we sample preferences in batches of 4 (Star MDP, Gridworld) or 100 (Sepsis) to reduce the number of model updates needed.

Policy Optimization. Policy optimization stages, both in estimating optimal policy sets in Sim-OPRL and PbOP and in outputting final policies, are carried out exactly through linear programming for the Star MDP and Gridworld using `cvxopt` [Diamond and Boyd, 2016], based on code from Lindner et al. [2021], and using Proximal Policy Optimization [Schulman et al., 2017] implemented in `stable-baselines3` [Raffin et al., 2021] for the Sepsis environment. In the latter case, after every preference collection episode, reward and policy models were trained from the checkpoint of the previous iteration, for only 20 steps to minimize computation.

Baselines and Ablations. We implement both OPRL baselines within our model-based offline preference-based algorithm described in Section 4. Uncertainty sampling is taking the pair with maximum preference uncertainty over 45 pairs for every sample, to reduce the load of computing preference uncertainty over the entire trajectory buffer.

Our ablation study for Figure 1c is conducted as follows. For Sim-OPRL without pessimism in the output policy, we output the policy that maximizes the value function under the MLE estimate of the transition and reward function, $\hat{T}$ and $\hat{R}$, after preference acquisition. For Sim-OPRL without pessimism in the simulated rollouts, we estimate the optimal policy set $\Pi_{offline}$ in the MLE estimate of the transition model instead of its pessimistic counterpart. Finally, for Sim-OPRL without optimism in the simulated rollouts, we generate rollouts from any two policies in $\Pi_{offline}$ instead of the most explorative ones.

C Environment Details

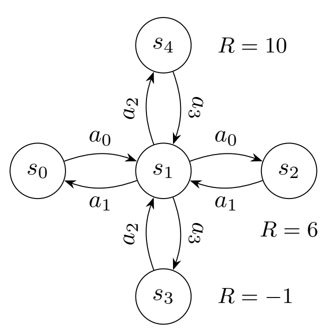


Figure 4: **Star MDP illustrated in Figure 1a**. Transition probabilities are 0.9 for all solid arrows. Omitted actions or complementary transitions keep the state unchanged.

Star MDP. We illustrate the transition dynamics underlying the Star MDP in Figure 4. Transition probabilities are 0.9 for all depicted solid arrows, and leave the state unchanged otherwise. Other actions also keep the state unchanged with probability 1. Episodes have length $H = 3$ and start from s_0. Unless specified otherwise, the offline dataset $\mathcal{D}_{offline}$ consists of 40 trajectories which only cover states (s_0, s_1, s_3) and (s_3, s_1, s_2).

Start		-1	-1
		-1	-1
	20		
			10

Figure 5: **Gridworld environment.** Rewards at every state are indicated if non-zero. Transition probabilities are 0.9. Thick lines indicate an obstacle, through which state transitions have probability zero.

Gridworld. We illustrate the gridworld environment in Figure 5. The environment consists of a 4×4 grid with states associated with different rewards, including a negative-reward region in the top-right corner, a high-reward but unreachable state, and a moderate-reward state at the bottom right corner. Each episode starts in the top-left corner. Transition probabilities for each of the four actions (`top, left, bottom, right`) are 0.9 for the intended direction, and 0.1 for the others; and action `stay` remains in the current state with probability 1. Transitions beyond the grid limits or through obstacles have probability zero, with the remainder of the probability mass for each action being distributed amongst other directions equally. The offline dataset contains 150 episodes and the behavioral policy is ϵ-optimal with noise $\epsilon = 0.1$. Episodes have length $H = 10$.

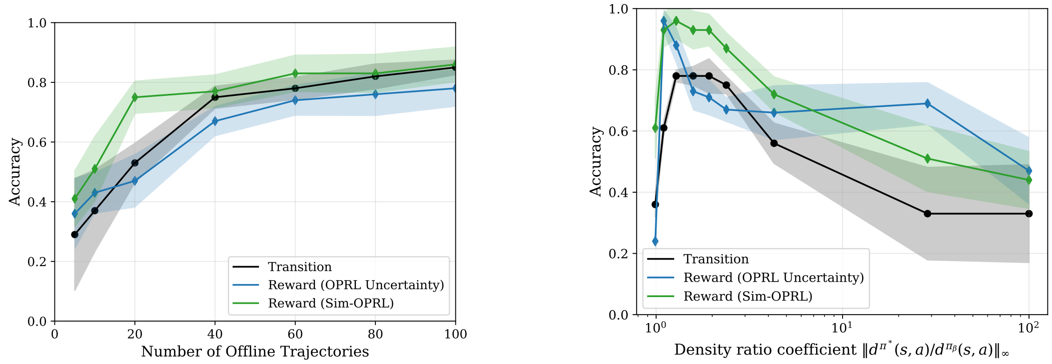
Sepsis Simulation. The sepsis simulator [Oberst and Sontag, 2019] is a commonly used environment for medically-motivated RL work [Tang and Wiens, 2021]. We use the original authors' publicly available code: `https://github.com/clinicalml/gumbel-max-scm/tree/sim-v2/sepsisSimDiabetes` (MIT license). The state space consists of five discrete observational variables (heart rate, blood pressure, oxygen concentration, glucose, diabetes status) and the action space consists of three different binary treatment options (antibiotic administration, vasopressor administration, mechanical ventilation). The probability that each treatment affects the value of each vital sign is determined by Oberst and Sontag [2019] to reflect patients' physiology. The ground truth reward function is sparse and only assigns a positive reward of $+1$ to surviving patients and a negative reward of -1 if death occurs (3 or more abnormal vitals) during their stay. The offline trajectories dataset includes 10,000 episodes following an ϵ-optimal policy with noise $\epsilon = 0.1$ and the episode length is $H = 20$.

D Additional Results

We include additional results in this section.

In Figure 6, we report the accuracy of the transition and preference model achieved for the Star MDP as we vary the size of optimality of the offline dataset. Accuracy is measured against all possible state transitions and over 100 pairs of random trajectories (random combinations of the 5 states and 4 actions in a sequence of $H = 3$). This complements our analysis in Section 7 and fig. 2. We see a steady improvement in both transition and reward model quality as we increase the amount of observational data in Figure 6a, which explains the observed dependence of N_p on N_o in Figure 2a.

In Figure 6b, we notice low model performance at both extremes of the x-axis. When the dataset is fully optimal, we find that all trajectories involve the same sequence of actions and states, so learning a transition or reward model from this data is challenging. We reach a similar conclusion at the other end of the spectrum at high density ratios, where the coverage the optimal states reduces. We reach



(a) As a function of offline dataset size.

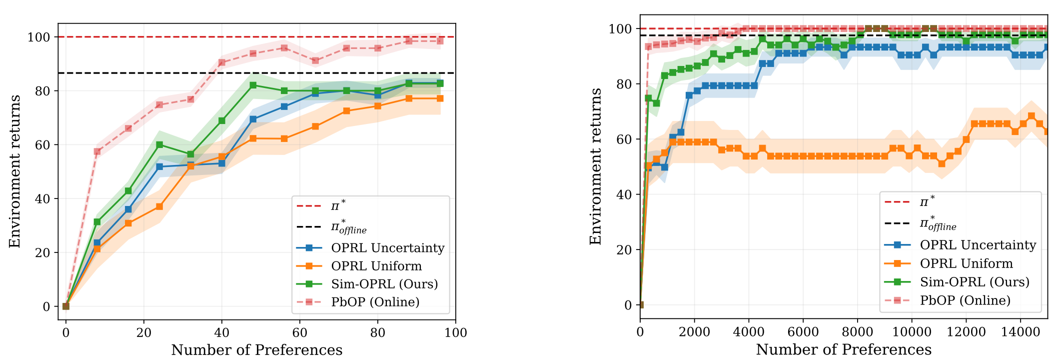
(b) As a function of optimal coverage.

Figure 6: **Transition and preference model accuracy as function of the properties of the observational data** (Star MDP). Preference elicitation is carried out until 10 preferences are queried. Mean and 95% confidence intervals over 20 experiments. Note that the transition model is the same for the two methods, as they have access to the same dataset.

highest performance for both models at intermediate values, when diversity of the observational data is high.

Still, it is important to stress that the highest accuracy of both models does not necessarily translate to the best-performing policy: good performance on the distribution induced by the optimal policy is more important, as formalized by the concentrability coefficients.

Next, we plot performance as a function of preferences sampled for our two additional environments in Figure 7. We reach similar conclusion to those drawn from the Star MDP in Section 7: within the offline preference elicitation approaches, OPRL with uniform sampling is the least efficient, OPRL with uncertainty sampling performs better, and Sim-OPRL even better. The PbOP method naturally reaches a superior policy with fewer samples as it allows environment interaction and can thus improve its estimate of the transition model in parallel to learning the preference function.



(a) Gridworld.

(b) Sepsis simulation.

Figure 7: **Empirical results on additional environments.** Mean and 95% confidence interval over 20 experiments. Environment returns are normalised between 0 and 100. Only OPRL and Sim-OPRL are fully offline.

E Broader Impact

Better preference elicitation strategies for offline reinforcement learning have the potential to facilitate and improve decision-making in real-world safety-critical domains like healthcare or economics, by reducing reliance on direct environment interaction and reducing human effort in providing feedback. Potential downsides could include the amplification of biases in the offline data, potentially leading to suboptimal or unfair policies. Thorough evaluation is therefore crucial to mitigate this before deploying models in such real-world applications. In addition, human preferences may not be fully captured by binary comparisons. As noted in our conclusion, we hope that future work will explore richer feedback mechanisms to better model complex decision-making objectives.