IMPROVED SAMPLE COMPLEXITY FOR REWARD-FREE REINFORCEMENT LEARNING UNDER LOW-RANK MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

In reward-free reinforcement learning (RL), an agent explores the environment first without any reward information in order to achieve certain learning goals afterwards for any given reward. While reward-free RL has been well studied under the tabular setting with minimax optimal sample complexity being achieved, theoretical study of reward-free RL with complicated function approximation is still limited. In this paper we focus on reward-free RL under low-rank MDP models, which capture the representation learning in RL. We propose a new model-based algorithm, coined RAFFLE, and show that it can both find an ϵ-optimal policy and achieve an ϵ-accurate system identification via reward-free exploration, with a sample complexity of $\tilde{O}(\frac{H^3 d^2 K(d^2+K)}{\epsilon^2})$, where d, H and K respectively denote the representation dimension, episode horizon, and action space cardinality. This significantly improves the sample complexity of $\tilde{O}(\frac{H^{22} K^9 d^7}{\epsilon^{10}})$ in Agarwal et al. (2020) for the same learning goals. We further provide a sample complexity lower bound of $\tilde{\Omega}(\frac{HdK}{\epsilon^2})$ that holds for any reward-free algorithm under low-rank MDPs, which matches our upper bound in the dependence on ϵ, as well as on K in the large d regime. Comparing this lower bound for low-rank MDPs with the upper bound for linear MDPs in Wang et al. (2020), it implies that reward-free RL under low-rank MDPs is strictly harder than linear MDPs. Finally, we complete our study by reusing RAFFLE to learn representation. We estimate the representation individually with only access to the learned transition kernels from RAFFLE and without interacting with true environment, and then theoretically characterize the closeness between the learned and the ground truth representation. The learned representation can be further used for few shot RL as in supervised learning (Du et al., 2021b).

1 INTRODUCTION

Reward-free reinforcement learning, recently formalized by Jin et al. (2020b), arises as a powerful framework to accommodate diverse demands in sequential learning applications. Under the reward-free RL framework, an agent first explores the environment without reward information during the exploration phase, with the objective to achieve certain learning goals later on for any given reward during the planning phase. The learning goal can be to find an ϵ-optimal policy, to achieve an ϵ-accurate system identification, etc. The reward-free RL paradigm may find broad application in many real-world engineering problems. For instance, reward-free exploration can be efficient when various reward functions are taken into consideration over a single environment, such as safe RL (Miryoosefi & Jin, 2021; Huang et al., 2022), multi-objective RL (Wu et al., 2021), multi-task RL (Agarwal et al., 2022; Cheng et al., 2022), etc. Studies of reward-free RL on the theoretical side have been largely focused on characterizing the sample complexity to achieve a learning goal under various MDP models. Specifically, reward-free tabular RL has been studied in Jin et al. (2020a); Ménard et al. (2021); Kaufmann et al. (2021); Zhang et al. (2020). For reward-free RL with function approximation, Wang et al. (2020) studied *linear MDPs* introduced by Jin et al. (2020b), where both the transition and the reward are linear functions of a given feature extractor, Zhang et al. (2021b) studied *linear mixture MDPs* introduced by Ayoub et al. (2020), and Zanette et al. (2020b) considered a classes of MDPs with *low inherent Bellman error* introduced by Zanette et al. (2020a).

Table 1: Comparison of reward-free RL algorithms under low-rank MDPs.

METHODS	SETTING	COMPLEXITY
FLAMBE (AGARWAL ET AL., 2020)	LOW-RANK MDP	$\tilde{O}(\frac{H^{22}K^9d^7}{\epsilon^{10}})$
MOFFLE (MODI ET AL., 2021)	LOW-NONNEGATIVE-RANK MDP	$\tilde{O}(\frac{H^5d_{LV}^3K^5}{\epsilon^2\eta})$
RFOLIVE [1](CHEN ET AL., 2022)	LOW BELLMAN ELUDER DIMENSION	$\tilde{O}(\frac{H^8d^3Kd_{\mathcal{R}}}{\epsilon^2})$
RAFFLE (OURS)	LOW-RANK MDP	$\tilde{O}(\frac{H^3d^2K(d^2+K)}{\epsilon^2})$

[1] The result of RFOLIVE (Chen et al., 2022) is not directly comparable with our result: (i) Similar to its reward-known version OLIVE (Jiang et al., 2017), RFOLIVE is not computationally efficient as noted in Agarwal et al. (2020); Uehara et al. (2022b), (ii) RFOLIVE assumes parametric growth of the covering number w.r.t $d_{\mathcal{R}}$ for the reward class, whereas we do not make any assumption on the reward class.

In this paper, we focus on reward-free RL under low-rank MDPs, where the transition kernel admits a decomposition into two embedding functions that map to low dimensional spaces. Compared with linear MDPs, the feature functions (i.e., the representation) under low-rank MDPs are unknown, hence the design further requires representation learning and becomes more challenging. Reward-free RL under low-rank MDPs was first studied by Agarwal et al. (2020), and the authors introduced a provably efficient algorithm FLAMBE, which achieves the learning goal of system identification with a sample complexity of $\tilde{O}(\frac{H^{22}K^9d^7}{\epsilon^{10}})$, where d, H and K respectively denote the representation dimension, episode horizon, and action space cardinality. Later on, Modi et al. (2021) proposed a model-free algorithm MOFFLE for reward-free RL under low-nonnegative-rank MDPs (where feature functions are non-negative), for which the sample complexity for finding an ϵ-optimal policy scales as $\tilde{O}(\frac{H^5K^5d_{LV}^3}{\epsilon^2\eta})$ (which is rescaled under the condition of $\sum_{h=1}^H r_h \leq 1$ for fair comparison). Here, d_{LV} denotes the non-negative rank of the transition kernel, which may be exponentially larger than d as shown in Agarwal et al. (2020), and η denotes the positive reachability probability to all states, where $1/\eta$ can be as large as $\sqrt{d_{LV}}$ as shown in Uehara et al. (2022b). Recently, a reward-free algorithm called RFOLIVE has been proposed under non-linear MDPs with low Bellman Eluder dimension by Chen et al. (2022), which can be specialized to low-rank MDPs. However, RFOLIVE is computationally more costly and considers a special reward function class, making their complexity result not directly comparable to other studies on reward-free low-rank MDPs.

This paper further investigates reward-free RL under low-rank MDPs to address the following three important open questions:

- Can we design a reward-free algorithm for low-rank MDPs that improves the sample complexity of FLAMBE in Agarwal et al. (2020) and MOFFLE in Modi et al. (2021), and achieves the goal of system identification as in Agarwal et al. (2020) and finds near-optimal policy as in Modi et al. (2021)?
- What is the fundamental limit on the required sample complexity to learn a low-rank MDP? While this question can be answered by a lower bound on the sample complexity for low-rank MDPs, none of the previous studies on low-rank MDPs establish such a lower bound.
- Is it possible to reuse the learned representation (i.e., features) of the transition kernels in other RL environment? Intuitively, the features can be utilized if learned accurately. However, previous studies on low-rank MDPs did not provide any estimation accuracy guarantee on the learned representation, but only on the transition kernels.

1.1 MAIN CONTRIBUTIONS

We summarize our main contributions in this work below.

- We propose a new model-based reward-free RL algorithm under low-rank MDPs. Our algorithm features carefully designed exploration-driven reward functions, exploration policies, and exploration termination criterion, and hence enjoys provably much better sample efficiency than the existing reward-free RL algorithms under low-rank MDPs.
- We show that our algorithm can both find an ϵ-optimal policy and achieve an ϵ-accurate system identification via reward-free exploration, with a sample complexity of $\tilde{O}(\frac{H^3d^2K(d^2+K)}{\epsilon^2})$,

which significantly improves that of $\tilde{O}(\frac{H^{22}K^9d^7}{\epsilon^{10}})$ in Agarwal et al. (2020) to achieve the same goal. Our result also improves the sample complexity of $\tilde{O}(\frac{H^5K^5d_{LV}^3}{\epsilon^2\eta})$ in Modi et al. (2021) in three aspects: order on K is reduced; d can be exponentially smaller than d_{LV} as shown in Agarwal et al. (2020); and no introduction of η, where $1/\eta$ can be as large as $\sqrt{d_{LV}}$. Further, our result on reward-free RL naturally achieves the goal of reward-known RL, which improves that of $\tilde{O}\left(\frac{H^5d^4K^2}{\epsilon^2}\right)$ in (Uehara et al., 2022b) by a factor of $O(H^2)$.

- We provide a lower bound $\tilde{\Omega}(\frac{HdK}{\epsilon^2})$ on the sample complexity that holds for any reward-free algorithm under low-rank MDPs, which matches our upper bound in terms of the dependence on ϵ as well as on K in the large d regime. Comparing this lower bound for low-rank MDPs with the upper bound for linear MDPs in Wang et al. (2020) implies that reward-free RL under low-rank MDPs is strictly harder than linear MDPs.
- We further design an algorithm to exploit the learned transition kernel from our reward-free algorithm to further learn a provably near-accurate representation of the transition kernel without requiring further interaction with the environment. To the best of our knowledge, this is the first work that provides such guarantee on representation learning in reward-free RL.

2 Related Work

Reward-free RL. While various studies (Oudeyer et al., 2007; Bellemare et al., 2016; Burda et al., 2018; Colas et al., 2018; Nair et al., 2018; Eysenbach et al., 2018; Co-Reyes et al., 2018; Hazan et al., 2019; Du et al., 2019; Pong et al., 2019; Misra et al., 2020) proposed exploration algorithms for good coverage on the state space without using explicit reward signals, theoretically speaking, the paradigm of reward-free RL was first formalized by Jin et al. (2020a), where they provided both upper and lower bounds on the sample complexity. For tabular case, several follow-up studies (Kaufmann et al., 2021; Ménard et al., 2021) further improved the sample complexity, and Zhang et al. (2020) established the minimax optimality guarantee. Reward-free RL was also studied with function approximation. Wang et al. (2020) studied *linear MDPs*, and Zhang et al. (2021b) studied *linear mixture MDPs*. Further, Zanette et al. (2020b) considered a class of MDPs with *low inherent Bellman error* introduced by Zanette et al. (2020a). Chen et al. (2022) proposed a reward-free algorithm called RFOLIVE under non-linear MDPs with low Bellman Eluder dimension. In addition, Miryoosefi & Jin (2021) proposed a reward-free approach to solving constrained RL problems with any given reward-free RL oracle under both the tabular and linear MDPs.

Reward-free RL under low-rank MDPs. As discussed in Section 1, reward-free RL under low-rank MDPs have been studied recently (Agarwal et al., 2020; Modi et al., 2021), and our result significant improves the sample complexity therein. When finite latent state space is assumed, low-rank MDPs specialize to block MDPs, under which algorithms termed as PCID (Du et al., 2019) and HOMER (Misra et al., 2020) achieve sample complexities of $\tilde{O}(\frac{d^4H^2K^4}{\min(\eta^4\gamma_n^2,\epsilon^2)})$ and $\tilde{O}(\frac{d^8H^4K^4}{\min(\eta^3,\epsilon^2)})$, respectively. Our result on the general low-rank MDPs can be utilized to further improve those results for block MDPs.

Reward-known RL under low-rank MDPs. For reward-known RL, Uehara et al. (2022b) proposed a computationally efficient algorithm REP-UCB under low-rank MDPs. Our design of the reward-free algorithm for low-rank MDPs is inspired by their algorithm with several new ingredients as discussed in Section 4, and improves their sample complexity on the dependence of H. Meanwhile, algorithms have been proposed for MDP models with low Bellman rank (Jiang et al., 2017), low witness rank (Sun et al., 2019), bilinear classes (Du et al., 2021a) and low Bellman eluder dimension (Jin et al., 2021), which can specialize to low-rank MDPs. However, those algorithms are computationally more costly as remarked in Uehara et al. (2022b), although their sample complexity may have sharper dependence on d, K or H. Specializing to block MDPs, Zhang et al. (2022) proposed an algorithm called BRIEE, which empirically achieves the state-of-art sample complexity for block MDP models. Besides, Zhang et al. (2021a) proposed an algorithm coined ReLEX for a slightly different low-rank model, and obtained a problem dependent regret upper bound.

Low-rank MDPs in extended RL settings. After this work was initially submitted for potential publication early this year, many concurrent studies have been developed on various extended RL settings with low-rank MDP models. Wang et al. (2022); Uehara et al. (2022a) studied partially

observable Markov decision process (POMDP) with latent low-rank structure. Zhan et al. (2022) studied predictive state representations model, and applied their results to POMDP with latent low-rank structure. Cheng et al. (2022); Agarwal et al. (2022) studied benefits of multitask representation learning under low-rank MDPs. Huang et al. (2022) proposed a general safe RL framework and instantiated it to low-rank MDPs. Ren et al. (2022) studied reward-known RL under low-rank MDPs and proposed a spectral method to replace the computation oracle. This work focuses on the most fundamental setting in the study of low-rank MDPs, and our novel algorithm design and theoretical analysis can be adopted to further improve the sample complexity in many of these studies.

3 Preliminaries and Problem Formulation

Notation. For any $H \in \mathbb{N}$, we denote $[H] := \{1, \ldots, H\}$. For any vector x and symmetric matrix A, we denote $\|x\|_2$ as the ℓ_2 norm of x and $\|x\|_A := \sqrt{x^\top A x}$. For any matrix A, we denote $\|A\|_F$ as its Frobenius norm and let $\sigma_i(A)$ be its i-th largest singular value. For two probability measures $P, Q \in \Omega$, we use $\|P - Q\|_{TV}$ to denote their total variation distance.

3.1 Episodic MDPs

We consider an episodic Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ can be an arbitrarily large state space; $\mathcal{A}$ is a finite action space with cardinality K; H is the number of steps in each episode; $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the time-dependent transition kernel, where $P_h(s_{h+1}|s_h, a_h)$ denotes the transition probability from the state-action pair (s_h, a_h) at step h to state s_{h+1} in the next step; $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the deterministic reward function at step h; We further normalize the summation of reward function as $\sum_h^H r_h \leq 1$. A policy π is a set of mappings $\{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h \in [H]}$, where $\Delta(\mathcal{A})$ is the set of all probability distributions over the action space $\mathcal{A}$. Further, $a \sim \mathcal{U}(\mathcal{A})$ indicates the uniform selection of an action a from $\mathcal{A}$.

In each episode of the MDP, we assume that a fixed initial state s_1 is drawn. Then, at each step $h \in [H]$. the agent observes state $s_h \in \mathcal{S}$, takes an action $a_h \in \mathcal{A}$ under a policy π_h, and receives a reward $r_h(s_h, a_h)$ (in the reward-known setting), and then the system transits to the next state s_{h+1} with probability $P_h(s_{h+1}|s_h, a_h)$. The episode ends after H steps.

As standard in the literature, we use $s_h \sim (P, \pi)$ to denote a state sampled by executing the policy π under the transition kernel P for $h - 1$ steps. If the previous state-action pair (s_{h-1}, a_{h-1}) is given, we use $s_h \sim P$ to denote that s_h follows the distribution $P_h(\cdot|s_{h-1}, a_{h-1})$. We use the notation $\mathbb{E}_{(s_h, a_h) \sim (P, \pi)}[\cdot]$ to denote the expectation over states $s_h \sim (P, \pi)$ and actions $a_h \sim \pi$.

For a given policy π and an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$, we denote the value function starting from state s_h at step h as $V^\pi_{h,P,r}(s_h) := \mathbb{E}_{(s_{h'}, a_{h'}) \sim (P, \pi)} \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) | s_h \right]$. We use $V^\pi_{P,r}$ to denote $V^\pi_{1,P,r}(s_1)$ for simplicity. Similarly, we denote the action-value function starting from state action pair (s_h, a_h) at step h as $Q^\pi_{h,P,r}(s_h, a_h) := r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim P} \left[V^\pi_{h+1,P,r}(s_{h+1}) | s_h, a_h \right]$.

We use $P^\star$ to denote the transition kernel of the true environment and for simplicity, denote $\mathbb{E}_{(s_h, a_h) \sim (P^\star, \pi)}[\cdot]$ as $\mathbb{E}^\star_\pi[\cdot]$. Given a reward function r, there always exists an optimal policy $\pi^\star$ that yields the optimal value $V^\star_{P^\star, r} = \sup_\pi V^\pi_{P^\star, r}$.

3.2 Low-rank MDPs

This paper focuses on the low-rank MDP (Jiang et al., 2017; Agarwal et al., 2020) defined as follows.

Definition 1. (Low-rank MDP). A transition probability $P^\star_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ admits a low-rank decomposition with dimension $d \in \mathbb{N}$ if there exists two embedding functions $\phi^\star_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu^\star_h : \mathcal{S} \to \mathbb{R}^d$ such that

$$P^\star_h(s'|s, a) = \langle \phi^\star_h(s, a), \mu^\star_h(s') \rangle, \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}.$$

For normalization, we assume $\|\phi^\star_h(s, a)\|_2 \leq 1$ for all (s, a), and for any function $g : \mathcal{S} \to [0, 1]$, $\| \int \mu^\star_h(s) g(s) ds \|_2 \leq \sqrt{d}$. An MDP $\mathcal{M}$ is a low-rank MDP with dimension d if for each

$h \in [H]$, P_h admits a low rank decomposition with dimension d. We use $\phi^\star = \{\phi_h^\star\}_{h\in[H]}$ and $\mu^\star = \{\mu_h^\star\}_{h\in[H]}$ to denote the embeddings for $P^\star$.

We remark that when ϕ_h is revealed to the agent, low-rank MDPs specialize to linear MDPs (Wang et al., 2020; Jin et al., 2020b). Essentially, low-rank MDPs do not assume that the features $\{\phi_h\}_h$ are known a priori. The lack of knowledge on features in fact invokes a nonlinear structure, which makes the model strictly harder than linear MDPs or tabular models. Since it is impossible to learn a model in polynomial time if there is no assumption on features ϕ_h and μ_h, we adopt the following conventional assumption from the recent studies on low-rank MDPs.

Assumption 1. (Realizability). A learning agent can access to a model class $\{(\Phi, \Psi)\}$ that contains the true model, i.e., the embeddings $\phi^\star \in \Phi, \mu^\star \in \Psi$.

While we assume the cardinality of the function classes to be finite for simplicity, extensions to infinite classes with bounded statistical complexity (such as bounded covering number) are not difficult (Sun et al., 2019; Agarwal et al., 2020).

We further assume that for a given low-rank MDP with any reward function, there exists an algorithm that returns the optimal policy in polynomial time. Such a computational oracle is commonly adopted in the literature (Agarwal et al., 2020; Uehara et al., 2022b).

3.3 Reward-free RL and Learning Objectives

Reward-free RL typically has two phases: exploration and planning. In the *exploration* phase, an agent explores the state space via interaction with the true environment and can collect samples over multiple episodes, but without access to the reward information. In the *planning* phase, the agent is no longer allowed to interact with the environment, and for any given reward function, is required to achieve certain learning goals (to be elaborated below) based on the outcome of exploration phase.

The planning phase may require the agent to achieve different learning goals. In this paper, we focus on two of such goals. The most popular goal in reward-free RL is to find a near-optimal policy that achieves the best value function under the true environment with ϵ-accuracy, as defined below.

Definition 2. (ϵ-optimal policy). Fix $\epsilon > 0$. For any given reward function r, a learned policy π is ϵ-optimal if it satisfies

$$V^{\pi^\star}_{P^\star,r} - V^{\pi}_{P^\star,r} \leq \epsilon. \tag{1}$$

For model-based learning, Agarwal et al. (2020) proposed *system identification* as another useful learning goal, defined as follows.

Definition 3 (ϵ-accurate system identification)**.** Fix $\epsilon > 0$. Given a model class (Φ, Ψ), a learned model $(\hat{\phi}, \hat{\mu})$ is said to achieve ϵ-accurate system identification if it uniformly approximates the true model $P^\star$, i.e., $\forall \pi, h \in [H]$, $\mathbb{E}^\star_\pi \left[\left\| \left\langle \hat{\phi}_h(s_h, a_h), \hat{\mu}_h(\cdot) \right\rangle - P^\star_h(\cdot|s_h, a_h) \right\|_{TV} \right] \leq \epsilon$.

4 The RAFFLE Algorithm

In this section, we describe our proposed algorithm RAFFLE for reward-free RL under low-rank MDPs. The algorithm is presented in Algorithm 1.

Exploration and MLE model estimation. In each iteration n during the exploration phase, for each $h \in [H]$, the agent executes the exploration policy π_{n-1} (defined in the previous iteration) up to step $h-1$, after which it takes two uniformly selected actions, and stop after step $h+1$. Differently from FLAMBE that collects a large number of samples for each episode, our algorithm uses each exploration policy to collect only one sample trajectory at each episode, indexed by (n, h). Hence, the sample complexity of RAFFLE is much smaller than that of FLAMBE. In fact, such an efficient sampling together with our new termination idea introduced later benefit sample complexity.

Then, the agent estimates the low-rank components $\hat{\phi}_h^{(n)}$ and $\hat{\mu}_h^{(n)}$ via the MLE oracle with given model class (Φ, Ψ) and a dataset $\mathcal{D}_h^n$ as follows

$$\mathrm{MLE}(\mathcal{D}_h^n) := \arg\max_{\phi\in\Phi,\mu\in\Psi} \textstyle\sum_{(s,a,s')\in\mathcal{D}_h^n} \log \langle \phi_h(s,a), \mu_h(s') \rangle .$$

Algorithm 1 RAFFLE (**R**ew**A**rd-**F**ree **F**eature **LE**arning)

Input: $\hat{\alpha}_n$, ζ_n, $\epsilon > 0$, $\delta \in (0,1)$, regularizer λ_n, model classes $\{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$.
Initialize $\pi_0(\cdot|s)$ to be uniform; set $\mathcal{D}_h^0 = \emptyset$.
Phase I: Exploration Phase
for $n = 1, \ldots$ **do**
  for $h = 1, \ldots, H$ **do**
    Use π_{n-1}: roll into s_{h-1}, uniformly choose a_{h-1}, a_h, enter into s_h, s_{h+1}.
    Collect data $s_1^{(n,h)}, a_1^{(n,h)}, \ldots, s_h^{(n,h)}, a_h^{(n,h)}, s_{h+1}^{(n,h)}$.
    Add the triple $(s_h^{(n,h)}, a_h^{(n,h)}, s_{h+1}^{(n,h)})$ to the dataset $\mathcal{D}_h^n = \mathcal{D}_h^{n-1} \cup \{(s_h^{(n,h)}, a_h^{(n,h)}, s_{h+1}^{(n,h)})\}$.
    Learn $(\hat{\phi}_h^{(n)}, \hat{\mu}_h^{(n)}) = \text{MLE}(\mathcal{D}_h^n)$.
    Update transition dynamics $\hat{P}^{(n)}$ as $\hat{P}_h^{(n)}(s'|s,a) = \langle \hat{\phi}_h^{(n)}(s,a), \hat{\mu}_h^{(n)}(s') \rangle$.
  end for
  Update empirical covariance matrix $\hat{U}_h^{(n)}$ as in Equation (2).
  Define exploration-driven reward function $\hat{b}_h^{(n)}$ as in Equation (3).
  Define an estimated value function $\hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}}$ based on $\hat{P}^{(n)}$ and $\hat{b}^{(n)}$ as Equation (4).
  Find exploration policy $\pi_n = \arg\max_{\pi} \hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}}$.
  if $2\hat{V}^{\pi_n}_{\hat{P}^{(n)}, \hat{b}^{(n)}} + 2\sqrt{K\zeta_n} \leq \epsilon$ **then**
    Terminate **Phase I: Exploration Phase** and set $\hat{P}^{\epsilon} = \hat{P}^{(n)}, \hat{b}^{\epsilon} = \hat{b}^{(n)}, \pi_{\epsilon} = \pi_n, n_{\epsilon} = n$.
  end if
end for
Phase II: Planning Phase
Receive reward function $r = \{r_h\}_{h=1}^H$, and compute policy $\bar{\pi} = \arg\max_{\pi} V^{\pi}_{\hat{P}^{\epsilon}, r}$.
Output: policy $\bar{\pi}$, learned transition dynamics $\hat{P}^{\epsilon}$.

The agent then updates the estimated transition kernel with $\hat{\phi}_h^{(n)}$ and $\hat{\mu}_h^{(n)}$.

Design of exploration reward. The agent updates the empirical covariance matrix $\hat{U}_h^{(n)}$ as

$$\hat{U}_h^{(n)} = \textstyle\sum_{\tau=1}^{n} \hat{\phi}_h^{(n)}(s_h^{(\tau,h+1)}, a_h^{(\tau,h+1)})(\hat{\phi}_h^{(n)}(s_h^{(\tau,h+1)}, a_h^{(\tau,h+1)}))^{\top} + \lambda_n I, \tag{2}$$

where $\{s_h^{(\tau,h+1)}, a_h^{(\tau,h+1)}, s_{h+1}^{(\tau,h+1)}\}$ is collected at iteration τ, episode $(h+1)$, and step h.

Next, the agent uses both $\hat{\phi}_h^{(n)}$ and $\hat{U}_h^{(n)}$ to construct an *exploration-driven reward* function as

$$\hat{b}_h^{(n)}(s,a) = \min\{\hat{\alpha}_n \|\hat{\phi}_h^{(n)}(s,a)\|_{(\hat{U}_h^{(n)})^{-1}}, 1\} \tag{3}$$

where $\hat{\alpha}_n$ is a pre-determined parameter. We note that although individual $\hat{b}_h^{(n)}(s,a)$ for each step may not represent point-wise uncertainty as indicated in Uehara et al. (2022b), we find its total cumulative version $\hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}}$ can serve as a *trajectory-wise* uncertainty measure to select exploration policy. To see this, it can be shown that for any π and h, $\mathbb{E}_{\pi}^{\star}\left[\left\|\left\langle \hat{\phi}_h^{(n)}(s_h,a_h), \hat{\mu}_h^{(n)}(\cdot)\right\rangle - P_h^{\star}(\cdot|s_h,a_h)\right\|_{TV}\right] \leq c'\hat{V}^{\pi_n}_{\hat{P}^{(n)}, \hat{b}^{(n)}} + \sqrt{\frac{c_n}{n}}$, where c' is a constant and $c_n = O(\log n)$. As iteration number n grows, the second term diminishes to zero, which indicates that $\hat{V}^{\pi_n}_{\hat{P}^{(n)}, \hat{b}^{(n)}}$ (under the reward of $\hat{b}^{(n)}$) serves as a good upper bound on the estimation error for the true transition kernel. Hence, exploration guided by maximizing $\hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}}$ will collect more trajectories over which the learned transition kernels are not estimated well. This will help to reduce the model estimation error in the future.

Design of exploration policy. The agent defines an *truncated value function* iteratively using the estimated transition kernel and the exploration-driven reward as follows

$$\begin{aligned}
\hat{Q}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h,a_h) &= \min\left\{1, \hat{b}_h^{(n)}(s_h,a_h) + \hat{P}_h^{(n)}\hat{V}^{\pi}_{h+1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h,a_h)\right\}, \\
\hat{V}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h) &= \mathop{\mathbb{E}}_{\pi}\left[\hat{Q}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h,a_h)\right].
\end{aligned} \tag{4}$$

The truncation technique here is important for the improvement of the sample complexity on the dependence of H. The agent finally finds an optimal policy maximizing $\hat{V}^{\pi}_{\hat{P}^{(n)},\hat{b}^{(n)}}$, and uses this policy as the exploration policy for the next iteration.

Novel termination criterion. RAFFLE does not require a pre-determined maximum number of iterations as its input. Instead, it will **terminate** and output the current estimated model if the optimal value function $\hat{V}^{\pi}_{\hat{P}^{(n)},\hat{b}^{(n)}}$ plus a minor term is below a threshold. Such a termination criterion essentially guarantees that the value functions under the estimated and true models are close to each other under any reward and policy, hence the exploration can be terminated in finite step. Such a termination criterion enables our algorithm to identify an accurate model and later find a near-optimal policy with fewer sample collections than FLAMBE in Agarwal et al. (2020) as we discuss in the exploration phase. Additionally, our termination criterion provides strong performance guarantees on the output policy and estimator from the last iteration. On the contrary, Uehara et al. (2022b) can only provide guarantees on a random mixture of the policies obtained over all iterations.

Planning phase. Given any reward function r, the agent finds a near-optimal policy by planning with the specially selected learned transition dynamics $\hat{P}^{\epsilon}$ and the given reward r. Note that such planning with a known low-rank MDP is computationally efficient by assumption.

5 Main Results

5.1 Upper Bounds on Sample Complexity

In this section, we first show that the policy returned by RAFFLE is an ϵ-optimal policy with respect to any given reward r in the planning phase. The detailed proof can be found in Appendix A.

Theorem 1 (ϵ-optimal policy)**.** *Assume $\mathcal{M}$ is a low-rank MDP with dimension d, and Assumption 1 holds. Given any $\delta, \epsilon \in (0,1)$, and any reward function r, let $\bar{\pi}$ and $\hat{P}^{\epsilon}$ be the output of RAFFLE and $\pi^{\star} := \arg\max_{\pi} V^{\pi}_{P^{\star},r}$ be the optimal policy under the true model $P^{\star}$. Set $\hat{\alpha}_n = \tilde{O}(\sqrt{(K+d^2)})$ and $\lambda_n = \tilde{O}(d)$. Then, with probability at least $1-\delta$, we have $V^{\pi^{\star}}_{P^{\star},r} - V^{\bar{\pi}}_{P^{\star},r} \leq \epsilon$, and the total number of trajectories collected by RAFFLE is upper bounded by*

$$\tilde{O}\left(\frac{H^3 d^2 K(d^2+K)\log^2(|\Phi||\Psi|/\delta)}{\epsilon^2}\right).$$

Compared with MOFFLE (Modi et al., 2021), which also finds an ϵ-optimal policy in reward-free RL, our result improves their sample complexity of $\tilde{O}\left(\frac{H^5 K^5 d_{LV}^3}{\epsilon^2 \eta}\right)$ in three aspects. First, the order on K is reduced. Second, the dimension d_{LV} of the underlying function class in MOFFLE can be exponentially larger than d as shown in Agarwal et al. (2020). Finally, MOFFLE requires reachability assumption, leading to a factor $1/\eta$ in the sample complexity, which can be as large as $\sqrt{d_{LV}}$. Further, Theorem 1 naturally achieves the goal of reward-known RL with the same sample complexity, which improves that of $\tilde{O}\left(\frac{H^5 d^4 K^2}{\epsilon^2}\right)$ in (Uehara et al., 2022b) by a factor of $O(H^2)$.

Proceeding to the learning objective of system identification, the learned transition kernel output by Algorithm 1 achieves the goal of ϵ-accurate system identification with the same sample complexity as follows. The detailed proof can be found in Appendix B.

Theorem 2 (ϵ-accurate system identification)**.** *Under the same condition of Theorem 1 and let $\hat{P}^{\epsilon} = \{\hat{\phi}^{\epsilon}_h, \hat{\mu}^{\epsilon}_h\}$ be the output of RAFFLE. Then, with probability at least $1-\delta$, $\hat{P}^{\epsilon}$ achieves ϵ-accurate system identification, i.e. for any π and h: $\mathbb{E}^{\star}_{\pi}\left[\left\|\left\langle \hat{\phi}^{\epsilon}_h(s_h,a_h), \hat{\mu}^{\epsilon}_h(\cdot)\right\rangle - P^{\star}_h(\cdot|s_h,a_h)\right\|_{TV}\right] \leq \epsilon$, and the number of trajectories collected by RAFFLE is upper bounded by*

$$\tilde{O}\left(\frac{H^3 d^2 K(d^2+K)\log^2(|\Phi||\Psi|/\delta)}{\epsilon^2}\right).$$

Theorem 2 significantly improves the sample complexity of $\tilde{O}(\frac{H^{22}K^9 d^7}{\epsilon^{10}})$ in Agarwal et al. (2020) on the dependence of all involved parameters for achieving ϵ-accurate system identification.

5.2 LOWER BOUND ON SAMPLE COMPLEXITY

In this subsection, we provide a lower bound on the sample complexity that all reward-free RL algorithms must satisfy under low-rank MDPs. The detailed proof can be found in Appendix C.

Theorem 3 (Lower bound)**.** *For any reward-free algorithm* `Alg`*, which can output an ϵ-optimal policy satisfying Equation* (1)*, if $H > \max(24\epsilon, 4)$ and $\delta < 1/16$, then there exists a low-rank MDP model $\mathcal{M}$ such that the number of trajectories sampled by* `Alg` *is at least $\Omega\left(\frac{HdK}{\epsilon^2}\right)$.*

Theorem 3 has two important implications. First, the lower bound in Theorem 3 matches the upper bound in Theorem 1 in terms of the dependence on ϵ as well as on K in the large d regime.

Second, Theorem 3 also shows that reward-free RL under *low-rank* MDPs is strictly harder than *linear* MDPs by at least a factor of $\Omega(K)$, because the lower bound in Theorem 3 for low-rank MDPs has an additional term K compared to the upper bound $\tilde{O}\left(\frac{d^3H^4}{\epsilon^2}\right)$ provided in Wang et al. (2020) for linear MDPs. This can be explained intuitively as follows. In linear models, all the representations $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ are known. Then it requires at most $O(d)$ actions with linearly independent features to realize all transitions $\langle \phi, \mu \rangle$. However, learning low-rank MDPs requires the agent to further select $O(K)$ actions to access to unknown features, leading to a dependence on K.

5.3 REPRESENTATION LEARNING VIA REWARD-FREE RL

In low-rank MDPs, the representation ϕ, if learned accurately, can be *reused* in other RL environments. Although RAFFLE provides an estimation of ϕ during its execution, such an estimation does not come with an accuracy guarantee. Note that Theorem 2 on the system identification does not provide the guarantee on the representation $\hat{\phi}$, but only on the entire transition kernel $\hat{P}$. None of previous studies of reward-free RL under low-rank MDPs (Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2022b) established the guarantee on $\hat{\phi}$.

In this section, we design an algorithm, which exploits the learned transition kernel from RAFFLE and learns an accurate representation without additional interaction with the environment.

We first introduce the following divergence-based metric to quantify the distance between two representations, which has been used in supervised learning (Du et al., 2021b).

Definition 4. (Divergence between two representations). Given a distribution q over $\mathcal{S} \times \mathcal{A}$ and two representations $\phi, \phi' \in \Phi$, define the covariance between ϕ and ϕ' w.r.t q as $\Sigma_{(s,a)\sim q}(\phi, \phi') = \mathbb{E}\left[\phi(s,a)\phi'(s,a)^\top\right]$. Then, the divergence between ϕ and ϕ' with respect to q is defined as

$$D_q(\phi, \phi') = \Sigma_{(s,a)\sim q}(\phi', \phi') - \Sigma_{(s,a)\sim q}(\phi', \phi)\left(\Sigma_{(s,a)\sim q}(\phi, \phi)\right)^\dagger \Sigma_{(s,a)\sim q}(\phi, \phi').$$

It can be verified that $D_q(\phi, \phi') \succeq 0$ (i.e., positive semidefinite) and $D_q(\phi, \phi) = 0$ for any $\phi, \phi' \in \Phi$.

Algorithm 2 for representation learning via RAFFLE. The formal version of Algorithm 2 is presented in Appendix D. We explain the main idea of Algorithm 2 as follows. First, for each $h \in [H], t \in [T]$, where T is the number of rewards, N_f pairs of state-action (s_h, a_h) are generated based on a distribution $\{q_h\}$. Note that the agent does not interact with the true environment during such data generation. Then, for any h, if we set the reward r at step h to be zero, $Q^\pi_{P^\star,h,r}(s_h, a_h)$ can have a linear structure in terms of the true representation $\phi^\star_h(s_h, a_h)$. Namely, there exists a w_h decided by r, $P^\star$ and π such that $Q^\pi_{P^\star,h,r}(s_h, a_h) = \langle \phi^\star_h(s_h, a_h), w_h \rangle$. Then, with the estimated transition kernel $\hat{P}$ from RAFFLE and the well-designed rewards $r^{h,t}$, $Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h, a_h)$ can be computed efficiently and serve as a target to learn the representation via the following regression problem:

$$\arg\min_{\phi_h \in \Phi, w^t_h \in \mathbb{R}^d} \sum_{t\in[T]} \sum_{(s_h,a_h)\in\mathcal{D}^t_h} (Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h, a_h) - \langle \phi_h(s_h, a_h), w^t_h \rangle)^2. \tag{5}$$

The main difference between our algorithm and that in Lu et al. (2021) for representation learning is that, our data generation is based on $\hat{P}$ from RAFFLE, which carries a natural estimation error but requires no interaction with true environment, whereas their algorithm assumes a generative model to collect data from the ground truth transition kernel.

In order to guarantee the learned representation by Algorithm 2 is close enough to the ground truth, we need to employ the following two somewhat necessary assumptions.

First, for the distributions $\{q_h\}_{h=1}^H$ in Algorithm 2 of the state-action pair, it is desirable to have $\hat{P}(\cdot|s,a)$ approximates true $P^\star(\cdot|s,a)$ well over distribution $\{q_h\}_{h=1}^H$, so that $Q^\pi_{\hat{P},h,r}(s_h,a_h)$ can approximate the ground truth well. Intuitively, if some state-action pairs can hardly be visited under any policy, the output of Algorithm 1 $\hat{P}(\cdot|s,a)$ cannot approximate true $P^\star(\cdot|s,a)$ well over these state-action pairs. Hence, $\hat{P}(\cdot|s,a)$ naturally only performs well on the input distributions with small density on these state-action pairs. We formalize such an assumption as follows.

Assumption 2. For the distributions $\{q_h\}_{h=1}^H$ in Algorithm 2, there exists a policy π^0 and a constant $C_{\min}$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $q_h(s,a) \leq C_{\min}\mathbb{P}_h^{\pi^0}(s,a)$, where $\mathbb{P}_h^{\pi^0}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the density over $\mathcal{S} \times \mathcal{A}$ induced by using policy π^0 to roll into state-action pair (s,a) at time h.

Assumption 2 is easy to satisfy if each state-action pair under the true MDP can be explored well by some policies. In fact, under standard reachability type assumptions such as in Modi et al. (2021), Assumption 2 holds for any distributions with finite bound (see Appendix D.1 for details).

Next, we assume that the rewards chosen for generating target Q-functions are sufficiently diverse so that the target $Q^\pi_{P^\star,h,r}(s_h,a_h)$ spans over the entire representation space to guarantee accurate representation learning in Equation (5). Such an assumption is commonly adopted in multi-task representation learning literature (Du et al., 2021b; Yang et al., 2021; Lu et al., 2021). To formally state the assumption, for any $h \in [H]$, let $\{r^{h,t}\}_{t\in[T]}$ be a set of T rewards (where $T \geq d$) satisfying $r_h^{h,t} = 0$. As a result, for each t, given π^t, there exists $w_h^{t\star} \in \mathbb{R}^d$ such that $Q^\pi_{P^\star,h,r^{h,t}}(s_h,a_h) = \langle \phi_h^\star(s_h,a_h), w_h^{t\star}\rangle$. Let $W_h^\star = [w_h^{1\star},\ldots,w_h^{T\star}] \in \mathbb{R}^{d\times T}$.

Assumption 3 (Diverse rewards)**.** The smallest singular value $\sigma_d(W_h^\star)$ of $W_h^\star$ defined above satisfies $\sigma_d^2(W_h^\star) \geq \Omega(\frac{T}{d})$, i.e., there exists a constant $C_D > 0$ such that $\sigma_d^2(W_h^\star) \geq \frac{C_D T}{d}$.

We next characterize the accuracy of the output representation of Algorithm 2 in terms of its closeness to the ground truth representation in the following theorem and defer the proof to Appendix D.

Theorem 4 (Guarantee for representation learning)**.** *Under Assumptions 1 - 3, for any $\delta,\epsilon \in (0,1)$, any $h \in [H]$, and sufficiently large N_f, let $\hat{P}$ be the output transition kernel of RAFFLE that satisfies Theorem 2. Then, with probability at least $1-\delta$, the output of Algorithm 2 $\tilde{\phi}$ satisfies*

$$\left\|D_{q_h}(\phi_h^\star,\tilde{\phi}_h)^{1/2}\right\|_F^2 = O\left(\frac{\epsilon d C_{\min}}{C_D} + \frac{d}{C_D}\sqrt{\frac{\log\frac{2}{\delta}}{TN_f}}\right). \tag{6}$$

We explain the bound in Equation (6) as follows. The first term arises from the system identification error ϵ by using the output of RAFFLE, and can be made as small as possible by choosing appropriate ϵ. The second term is related to the randomness caused by sampling state-action pairs from input distribution $\{q_h\}_{h\in[H]}$, and vanishes as N_f becomes large. Note that N_f is the number of simulated samples in Algorithm 2, which does not require interaction with the true environment. Hence, it can be made sufficiently large to guarantee a small error.

Theorem 4 shows that Algorithm 2 can learn a near-accurate representation, based on which the learned representation can be further used in other RL environments that share a common representation, similar to how representation learning has been exploited in supervised learning (Du et al., 2021b).

6 Conclusions

In this paper, we investigate the reward-free reinforcement learning, in which the underlying model admits a low-rank structure. Without further assumptions, we propose an algorithm called RAFFLE which significantly improves the state-of-the-art sample complexity for accurate model estimation and near-optimal policy identification. We further design an algorithm to exploit the model learned by RAFFLE for accurate representation learning without further interaction with the environment. Although ϵ-accurate system identification can easily induce $H\epsilon$-optimal policy, the relationship between these two learning goals under general reward-free exploration setting remain under-explored, and is an interesting topic for future investigation.

REFERENCES

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e894d787e2fd6c133af47140aa156f00-Paper.pdf.

Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. *arXiv preprint arXiv:2205.14571*, 2022.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29:1471–1479, 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear RL. *CoRR*, abs/2206.10770, 2022. doi: 10.48550/arXiv.2206.10770. URL https://doi.org/10.48550/arXiv.2206.10770.

Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning. *CoRR*, abs/2206.05900, 2022. doi: 10.48550/arXiv.2206.05900. URL https://doi.org/10.48550/arXiv.2206.05900.

John Co-Reyes, YuXuan Liu, Abhishek Gupta, Benjamin Eysenbach, Pieter Abbeel, and Sergey Levine. Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings. In *International Conference on Machine Learning*, pp. 1009–1018. PMLR, 2018.

Cédric Colas, Pierre Fournier, Olivier Sigaud, and Pierre-Yves Oudeyer. CURIOUS: intrinsically motivated multi-task, multi-goal reinforcement learning. *CoRR*, abs/1810.06284, 2018. URL http://arxiv.org/abs/1810.06284.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021a.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019.

Ruiquan Huang, Jing Yang, and Yingbin Liang. Safe exploration incurs nearly no additional sample complexity for reward-free rl. *arXiv preprint arXiv:2206.14057*, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 13–18 Jul 2020a. URL `https://proceedings.mlr.press/v119/jin20d.html`.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms, 2021.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

Rui Lu, Gao Huang, and Simon S Du. On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*, 2021.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning, 2021.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*, 2018.

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022a.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank mdps. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=J4iSIR9fhY0.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Embed to control partially observed systems: Representation learning with provable sample efficiency. *CoRR*, abs/2205.13476, 2022. doi: 10.48550/arXiv.2205.13476. URL https://doi.org/10.48550/arXiv.2205.13476.

Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

Jingfeng Wu, Vladimir Braverman, and Lin Yang. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13112–13124, 2021.

Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon Shaolei Du. Impact of representation learning in linear bandits. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=edJ_HipawCa.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020b.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation learning in low-rank markov decision processes. *CoRR*, abs/2106.11935, 2021a. URL https://arxiv.org/abs/2106.11935.

Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34, 2021b.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Wen Sun, and Alekh Agarwal. Efficient reinforcement learning in block mdps: A model-free representation learning approach. *arXiv preprint arXiv:2202.00063*, 2022.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020.

Appendices

A Proof of Theorem 1

We first provide a proof outline to highlight our key ideas in the analysis of Theorem 1, and then provide the detailed proof. To simplify the notation, we denote the total variation distance $\left\|\hat{P}_h^{(n)}(s_h, a_h) - P_h^\star(s_h, a_h)\right\|_{TV}$ by $f_h^{(n)}(s_h, a_h)$.

Proof Outline for Theorem 1. **Step 1** provides an upper bound on the difference of value functions under estimated model $\hat{P}^{(n)}$ and the true model $P^\star$ for any given policy π and reward r, as given in the following proposition (see proposition 4 in appendix A.2).

Proposition 1 (Informal)**.** *There exist constants $c_n = O(\log n)$ and $c' = O(1)$ such that for any policy π and reward r, with high probability, we have*

$$\left|V^\pi_{P^\star,r} - V^\pi_{\hat{P}^{(n)},r}\right| \leq \hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{\frac{c_n}{n}}.$$

This proposition is inspired by Uehara et al. (2022b), while it generalizes to non-stationary setting and arbitrary reward scenario from infinite stationary MDP and fixed reward. The main proof idea is to first notice that for any reward, we have

$$\left|V^\pi_{P^\star,r} - V^\pi_{\hat{P}^{(n)},r}\right| \leq \hat{V}^\pi_{\hat{P}^{(n)},f^{(n)}}. \tag{7}$$

Then, we show that $\hat{V}^\pi_{\hat{P}^{(n)},f^{(n)}} \leq \hat{V}^\pi_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{\frac{c_n}{n}} \leq \hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{\frac{c_n}{n}}$ due to the optimality of the exploration policy π_n.

Step 2 shows sublinearity of the summation of $\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}}$, which is an upper bound of value function difference under the true model with the estimated model (see proposition 5 in appendix A.3).

Proposition 2 (Informal)**.** *Under the same setting of Proposition 1, with high probability, the summation of value functions $V^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}}$ under exploration policies $\{\pi_n\}_{n\in[N]}$ with exploration-driven reward functions $\hat{b}^{(n)}$ is sublinear, as given by*

$$\sum_{n\in[N]} \hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} \leq \tilde{O}\left(Hd\sqrt{K(K+d^2)N}\right).$$

Step 3 combines Step 2 and Step 1. We are able to conclude that RAFFLE will terminate with polynomial sample complexity such that, the value function difference of $P^\star$ and the returned environment $\hat{P}^\epsilon$ is at most ϵ.

Proposition 3. *With high probability, RAFFLE terminates after at most $\tilde{O}\left(H^2d^2K(d^2+K)/\epsilon^2\right)$ iterations, and the output model $\hat{P}^\epsilon$ satisfies*

$$\left|V^\pi_{P^\star,r} - V^\pi_{\hat{P}^{(\epsilon)},r}\right| \leq \hat{V}^{\pi_{n_\epsilon}}_{\hat{P}^{(\epsilon)},\hat{b}^{(n_\epsilon)}} + \sqrt{\frac{c_{n_\epsilon}}{n_\epsilon}} \leq \epsilon/2,$$

where n_ϵ is the iteration number where RAFFLE terminates, i.e. $\hat{P}^\epsilon = \hat{P}^{(n_\epsilon)}$.

Finally, with some algebraic operations, Proposition 3 concludes the proof of Theorem 1. □

A.1 Supporting Lemmas

We first present the high probability event.

Lemma 1. *We define Π_n to be a uniform mixture of previous $n-1$ exploration policies:*

$$\Pi_n = \mathcal{U}(\pi_0, \pi_1, ..., \pi_{n-1}).$$

Denote the total variation of $\hat{P}^{(n)}$ and $P^\star$, and the expected matrix of $\hat{U}_h^{(n)}$ as follows.

$$f_h^{(n)}(s_h, a_h) = \left\| P_h^\star(\cdot|s_h, a_h) - \hat{P}_h^{(n)}(\cdot|s_h, a_h) \right\|_{TV}, \tag{8}$$

$$U_{h,\phi}^{(n)} = n \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi_n) \\ a_h \sim \mathcal{U}(\mathcal{A})}} \left[\phi(s_h, a_h)(\phi(s_h, a_h))^\top \right] + \lambda_n I, \tag{9}$$

$$W_{h,\phi}^{(n)} = n \mathop{\mathbb{E}}_{(s_h, a_h) \sim (P^\star, \Pi_n)} \left[\phi(s_h, a_h)(\phi(s_h, a_h))^\top \right] + \lambda_n I. \tag{10}$$

where $\lambda_n = \beta_3 d \log(2nH|\Phi|/\delta))$ and $\beta_3 = O(1)$ is a constant coefficient. Suppose Algorithm 1 runs N iterations and let events $\mathcal{E}_0$ and $\mathcal{E}_1$ be defined as follows.

$$\mathcal{E}_0 = \left\{ \forall n \in [N], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}, \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi_n) \\ a_h \sim \mathcal{U}}} \left[f_h^{(n)}(s_h, a_h)^2 \right] \leq \zeta_n \right\},$$

$$\mathcal{E}_1 = \Bigg\{ \forall n \in [N], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A},$$
$$\frac{1}{5} \left\| \hat{\phi}_{h-1}^{(n)}(s, a) \right\|_{(U_{h-1,\hat{\phi}}^{(n)})^{-1}} \leq \left\| \hat{\phi}_{h-1}^{(n)}(s, a) \right\|_{(\hat{U}_{h-1}^{(n)})^{-1}} \leq 3 \left\| \hat{\phi}_{h-1}^{(n)}(s, a) \right\|_{(U_{h-1,\hat{\phi}}^{(n)})^{-1}} \Bigg\}.$$

where $\zeta_n = \log(2|\Phi||\Psi|nH/\delta)/n$. We further denote $\zeta = N\zeta_N = \log(2|\Phi||\Psi|NH/\delta)$ for simplicity. Denote $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1$ as the intersection of the two events. Then, $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$.

Proof. By Corollary 2 in Appendix E, we have $\mathbb{P}[\mathcal{E}_0] \geq 1 - \delta/2$. Further, by Lemma 39 in Zanette et al. (2021) for the version of fixed ϕ and Lemma 11 in Uehara et al. (2022b), we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - \delta/2$. Therefore, $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$. □

Based on Lemma 1, we can bound the exlporation-driven reward in RAFFLE as follows.

Corollary 1. *Given that the event $\mathcal{E}$ occurs, the following inequality holds for any $n \in [N], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}$:*

$$\min \left\{ \frac{\hat{\alpha}_n}{5} \left\| \hat{\phi}_h^{(n)}(s_h, a_h) \right\|_{(U_{h,\hat{\phi}}^{(n)})^{-1}}, 1 \right\} \leq \hat{b}_h^{(n)}(s_h, a_h) \leq 3\hat{\alpha}_n \left\| \hat{\phi}_h^{(n)}(s_h, a_h) \right\|_{(U_{h,\hat{\phi}}^{(n)})^{-1}},$$

where $\hat{\alpha}_n = 5\sqrt{2\beta_3 n \zeta_n (K + d^2)}$.

Proof. Recall $\hat{b}_h^{(n)}(s_h, a_h) = \min \left\{ \hat{\alpha}_n \left\| \hat{\phi}_h^{(n)}(s, a) \right\|_{(\hat{U}_h^{(n)})^{-1}}, 1 \right\}$. Applying Lemma 1, we can immediately obtain the result. □

The following lemma extends the Lemmas 12 and 13 under infinite discount MDPs in Uehara et al. (2022b) to episodic MDPs. We provide the proof for completeness.

Lemma 2. *Let $P_{h-1} = \langle \phi_{h-1}, \mu_{h-1} \rangle$ be a generic MDP model, and Π be an arbitrary and possibly mixture policy. Define an expected Gram matrix as follows*

$$M_{h-1,\phi} = \lambda_n I + n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[\phi_{h-1}(s_{h-1}, a_{h-1}) \left(\phi_{h-1}(s_{h-1}, a_{h-1}) \right)^\top \right].$$

Further, let $f_{h-1}(s_{h-1}, a_{h-1})$ be the total variation between $P^\star_{h-1}$ and P_{h-1} at time step $h-1$. Suppose $g \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is bounded by $B \in (0, \infty)$, i.e., $\|g\|_\infty \leq B$. Then, $\forall h \geq 2, \forall$ policy π_h,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{s_h \sim P_{h-1} \\ a_h \sim \pi_h}} [g(s_h, a_h) | s_{h-1}, a_{h-1}] \\
&\qquad \leq \|\phi_{h-1}(s_{h-1}, a_{h-1})\|_{(M_{h-1,\phi})^{-1}} \times \\
&\qquad \sqrt{nK \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi) \\ a_h \sim \mathcal{U}}} [g^2(s_h, a_h)] + \lambda_n d B^2 + n B^2 \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} [f_{h-1}(s_{h-1}, a_{h-1})^2]}.
\end{aligned}$$

Proof. We first derive the following bound:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{s_h \sim P_{h-1} \\ a_h \sim \pi_h}} [g(s_h, a_h) | s_{h-1}, a_{h-1}] \\
&= \int_{s_h} \sum_{a_h} g(s_h, a_h) \pi(a_h | s_h) \langle \phi_{h-1}(s_{h-1}, a_{h-1}), \mu_{h-1}(s_h) \rangle ds_h \\
&\leq \|\phi_{h-1}(s_{h-1}, a_{h-1})\|_{(M_{h-1,\phi})^{-1}} \left\| \int \sum_{a_h} g(s_h, a_h) \pi(a_h | s_h) \mu_{h-1}(s_h) ds_h \right\|_{M_{h-1,\phi}},
\end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality. We further expand the second term in the RHS of the above inequality as follows.

$$\begin{aligned}
&\left\| \int \sum_{a_h} g(s_h, a_h) \pi(a_h | s_h) \mu_{h-1}(s_h) ds_h \right\|^2_{M_{h-1,\phi}} \\
&\overset{(i)}{\leq} n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[\left(\int_{s_h} \sum_{a_h} g(s_h, a_h) \pi_h(a_h | s_h) \mu(s_h)^\top \phi(s_{h-1}, a_{h-1}) ds_h \right)^2 \right] + \lambda_n d B^2 \\
&= n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[\left(\mathop{\mathbb{E}}_{\substack{s_h \sim P_{h-1} \\ a_h \sim \pi_h}} \left[g(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \right)^2 \right] + \lambda_n d B^2 \\
&\overset{(ii)}{\leq} n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[\mathop{\mathbb{E}}_{\substack{s_h \sim P_{h-1} \\ a_h \sim \pi_h}} \left[g(s_h, a_h)^2 \middle| s_{h-1}, a_{h-1} \right] \right] + \lambda_n d B^2 \\
&\overset{(iii)}{\leq} n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[\mathop{\mathbb{E}}_{\substack{s_h \sim P^\star_{h-1} \\ a_h \sim \pi_h}} \left[g(s_h, a_h)^2 \middle| s_{h-1}, a_{h-1} \right] \right] + \lambda_n d B^2 \\
&\quad + n B^2 \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[f_{h-1}(s_{h-1}, a_{h-1})^2 \right] \\
&\overset{(iv)}{\leq} nK \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi) \\ a_h \sim \mathcal{U}}} \left[g(s_h, a_h)^2 \right] + \lambda_n d B^2 + n B^2 \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi) \\ a_{h-1} \sim \Pi}} \left[f_{h-1}(s_{h-1}, a_{h-1})^2 \right],
\end{aligned}$$

where (i) follows from the assumption that $\|g\|_\infty \leq B$, (ii) follows from Jensen's inequality, (iii) is due to that $f_{h-1}(s_{h-1}, a_{h-1})$ is the total variation between $P^\star_{h-1}$ and P_{h-1} at time step $h-1$, and (iv) is due to importance sampling. This finishes the proof. □

Based on Lemma 2, we summarize three useful inequalities which bridges the total variation $f_h^{(n)}$ and the exploration-driven reward $\hat{b}_h^{(n)}$.

Lemma 3. *Define*

$$W^{(n)}_{h,\phi} = n \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi_n) \\ a_h \sim \Pi_n}} \left[\phi(s_h, a_h) (\phi(s_h, a_h))^\top \right] + \lambda_n I, \tag{11}$$

where $\lambda_n = \beta_3 d \log(2nH|\Phi|/\delta)$. Given that the event $\mathcal{E}$ occurs, the following inequalities hold. For any n, when $h \geq 2$,

$$\mathop{\mathbb{E}}_{\substack{s_h \sim \hat{P}^{(n)}_{h-1} \\ a_h \sim \pi}} \left[f^{(n)}_h(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \leq \alpha_n \left\| \hat{\phi}^{(n)}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}}, \tag{12}$$

$$\mathop{\mathbb{E}}_{\substack{s_h \sim P^{*}_{h-1} \\ a_h \sim \pi}} \left[f^{(n)}_h(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \leq \alpha_n \left\| \phi^{*}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\phi^\star})^{-1}}, \tag{13}$$

$$\mathop{\mathbb{E}}_{\substack{s_h \sim P^{*}_{h-1} \\ a_h \sim \pi}} \left[\hat{b}^{(n)}_h(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \leq \gamma_n \left\| \phi^{*}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(W^{(n)}_{h-1,\phi^\star})^{-1}}, \tag{14}$$

where

$$\alpha_n = \sqrt{2\beta_3 n \zeta_n (K + d^2)}, \quad \gamma_n = \sqrt{45 \beta_3 n \zeta_n K d (K + d^2)}.$$

Specially, when $h = 1$,

$$\mathop{\mathbb{E}}_{a_1 \sim \pi} \left[f^{(n)}_1(s_1, a_1) \right] \leq \sqrt{K \zeta_n}, \qquad \mathop{\mathbb{E}}_{a_1 \sim \pi} \left[\hat{b}(s_1, a_1) \right] \leq 15 \alpha_n \sqrt{\frac{dK}{n}}. \tag{15}$$

Proof. We start by developing Equation (12) as follows. Given that the event $\mathcal{E}$ occurs, for $h \geq 2$ we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{s_h \sim \hat{P}^{(n)}_{h-1} \\ a_h \sim \pi}} \left[f^{(n)}_h(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \\
&\quad \overset{(i)}{\leq} \left\| \hat{\phi}^{(n)}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}} \times \\
&\quad \sqrt{nK \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi_n) \\ a_{h-1}, a_h \sim \mathcal{U} \\ s_h \sim P^\star{}_h(\cdot | s_{h-1}, a_{h-1})}} [f^{(n)}_h(s_h, a_h)^2] + \lambda_n d + n \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi_n) \\ a_{h-1} \sim \mathcal{U}}} \left[f^{(n)}_{h-1}(s_{h-1}, a_{h-1})^2 \right]} \\
&\quad \overset{(ii)}{\leq} \left\| \hat{\phi}^{(n)}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}} \times \\
&\quad \sqrt{nK \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi_n) \\ a_{h-1}, a_h \sim \mathcal{U} \\ s_h \sim P^\star{}_h(\cdot | s_{h-1}, a_{h-1})}} [f^{(n)}_h(s_h, a_h)^2] + \lambda_n d + nK \mathop{\mathbb{E}}_{\substack{s_{h-2} \sim (P^\star, \Pi_n) \\ a_{h-2}, a_{h-1} \sim \mathcal{U} \\ s_{h-1} \sim P^\star{}_{h-1}(\cdot | s_{h-2}, a_{h-2})}} \left[f^{(n)}_{h-1}(s_{h-1}, a_{h-1})^2 \right]} \\
&\quad \overset{(iii)}{\leq} \left\| \hat{\phi}^{(n)}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}} \sqrt{2n\zeta_n K + \beta_3 n \zeta_n d^2} \\
&\quad \leq \alpha_n \left\| \hat{\phi}^{(n)}_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}},
\end{aligned}$$

where (i) follows from Lemma 2 and the fact that $f^{(n)}_h(s_h, a_h) \leq 1$, (ii) follows from importance sampling at time step $h - 2$, and (iii) follows from Lemma 1.

Equation (13) follows from the arguments similar to the above.

To obtain Equation (14), we first apply Lemma 2 and obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{s_h \sim P^\star_{h-1} \\ a_h \sim \pi_n}} \left[\hat{b}^{(n)}_h(s_h, a_h) \middle| s_{h-1}, a_{h-1} \right] \\
&\leq \left\| \phi^\star_{h-1}(s_{h-1}, a_{h-1}) \right\|_{(W^{(n)}_{h-1,\phi^\star})^{-1}} \sqrt{nK \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \Pi_n) \\ a_h \sim \mathcal{U}}} [\{\hat{b}^{(n)}_h(s_h, a_h)\}^2] + \lambda_n d},
\end{aligned}$$

where we use the fact that $\hat{b}_h^{(n)}(s_h, a_h) \leq 1$. We further bound the term $n\,\mathbb{E}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}[(\hat{b}_h^{(n)}(s_h,a_h))^2]$ as follows:

$$
\begin{aligned}
& n \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}\left[\left(\hat{b}_h^{(n)}(s_h,a_h)\right)^2\right] \\
&\leq n \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}\left[\hat{\alpha}_n^2 \left\|\hat{\phi}_h^{(n)}(s_h,a_h)\right\|^2_{(\hat{U}^{(n)}_{h,\hat{\phi}})^{-1}}\right] \\
&\overset{(i)}{\leq} n \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}\left[9\hat{\alpha}_n^2 \left\|\hat{\phi}_h^{(n)}(s_h,a_h)\right\|^2_{(U^{(n)}_{h,\hat{\phi}})^{-1}}\right] \\
&= 9\hat{\alpha}_n^2 \mathrm{tr}\left\{ n \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}\left[\hat{\phi}_h^{(n)}(s_h,a_h)\hat{\phi}_h^{(n)}(s_h,a_h)^\top \left(n \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\Pi_n)\\ a_h\sim\mathcal{U}}}\left[\hat{\phi}_h(s_h,a_h)\hat{\phi}_h^{(n)}(s_h,a_h)^\top\right] + \lambda_n I\right)^{-1}\right]\right\} \\
&\leq 9\hat{\alpha}_n^2 \mathrm{tr}(I) = 9\hat{\alpha}_n^2 d,
\end{aligned}
$$

where (i) follows from Lemma 1, and we use $\mathrm{tr}(A)$ to denote the trace of any matrix A.

Hence,

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{s_h\sim P^*_{h-1}\\ a_h\sim\pi}}\left[\hat{b}_h^{(n)}(s_h,a_h)\middle| s_{h-1},a_{h-1}\right] &\leq \left\|\phi^*_{h-1}(s_{h-1},a_{h-1})\right\|_{(W^{(n)}_{h-1,\phi^\star})^{-1}} \sqrt{9K\hat{\alpha}_n^2 d + \lambda_n d} \\
&\leq \gamma_n \left\|\phi^*_{h-1}(s_{h-1},a_{h-1})\right\|_{W^{(n)}_{h-1,\phi^\star})^{-1}},
\end{aligned}
$$

where the last inequality follows from that $\hat{\alpha}_n = 5\alpha_n$ and the definition of γ_n In addition, for $h = 1$, we have

$$
\mathop{\mathbb{E}}_{a_1\sim\pi_n}\left[f_1^{(n)}(s_1,a_1)\right] \overset{(i)}{\leq} \sqrt{K \mathop{\mathbb{E}}_{a_1\sim\mathcal{U}}\left[f_1^{(n)}(s_1,a_1)^2\right]} \leq \sqrt{K\zeta_n},
$$

and

$$
\begin{aligned}
\mathop{\mathbb{E}}_{a_1\sim\pi_n}\left[\hat{b}(s_1,a_1)\right] &\overset{(ii)}{\leq} \hat{\alpha}_n \sqrt{K \mathop{\mathbb{E}}_{a_1\sim\mathcal{U}}\left[\|\hat{\phi}_1(s_1,a_1)\|^2_{(\hat{U}^{(n)}_{1,\hat{\phi}})^{-1}}\right]} \\
&\leq 3\hat{\alpha}_n \sqrt{K \mathop{\mathbb{E}}_{a_1\sim\mathcal{U}}\left[\|\hat{\phi}_1(s_1,a_1)\|^2_{(U^{(n)}_{1,\hat{\phi}})^{-1}}\right]} \\
&\leq 3\sqrt{\frac{25K\alpha_n^2 d}{n}} = 15\alpha_n\sqrt{\frac{dK}{n}},
\end{aligned}
$$

□

A.2 Proof of Proposition 1

Equipped with Lemma 3, the following proposition provides an upper bound on the difference of value functions under the estimated model $\hat{P}^{(n)}$ and the true model $P^\star$ for any given policy π and reward r.

Proposition 4 (Restatement of Proposition 1)**.** *For all $n \in [N]$, policy π and reward r, given that the event $\mathcal{E}$ occurs, we have*

$$
\left|V^\pi_{P^\star,r} - V^\pi_{\hat{P}^{(n)},r}\right| \leq \hat{V}^\pi_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{K\zeta_n}.
$$

Proof. **Step 1.** We first show that $\left|V^\pi_{P^\star,r} - V^\pi_{\hat{P}^{(n)},r}\right| \leq \hat{V}^\pi_{\hat{P}^{(n)},f^{(n)}}$.

Recall the definition of estimated value functions $\hat{V}_{h,\hat{P}^{(n)},r}(s_h)$ and $\hat{Q}_{h,\hat{P}^{(n)},r}(s_h,a_h)$:

$$\hat{Q}^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h) = \min\left\{1, r_h(s_h,a_h) + \hat{P}^{(n)}_h \hat{V}^{\pi}_{h+1,\hat{P}^{(n)},r}(s_h,a_h)\right\},$$
$$\hat{V}^{\pi}_{h,\hat{P}^{(n)},r}(s_h) = \mathop{\mathbb{E}}_{\pi}\left[\hat{Q}^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h)\right].$$

We develop the proof by induction. For the base case $h = H+1$, we have $\left|V^{\pi}_{H+1,\hat{P}^{(n)},r}(s_{H+1}) - V^{\pi}_{H+1,P^\star,r}(s_{H+1})\right| = 0 = \hat{V}^{\pi}_{H+1,\hat{P}^{(n)},f^{(n)}}(s_{H+1})$.

Assume that $\left|V^{\pi}_{h+1,\hat{P}^{(n)},r}(s_{h+1}) - V^{\pi}_{h+1,P^\star,r}(s_{h+1})\right| \le \hat{V}^{\pi}_{h+1,\hat{P}^{(n)},f^{(n)}}(s_{h+1})$ holds for any s_{h+1}.

Then, from Bellman equation, we have,

$$\begin{aligned}
&\left|Q^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h) - Q^{\pi}_{h,P^\star,r}(s_h,a_h)\right| \\
&\quad= \left|\hat{P}^{(n)}_h V^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h) - P^\star_h V^{\pi}_{h+1,P^\star,r}(s_h,a_h)\right| \\
&\quad= \left|\hat{P}^{(n)}_h \left(V^{\pi}_{h+1,\hat{P}^{(n)},r} - V^{\pi}_{h+1,P^\star,r}\right)(s_h,a_h) + \left(\hat{P}^{(n)}_h - P^\star_h\right) V^{\pi}_{h,P^\star,r}(s_h,a_h)\right| \\
&\quad\overset{(i)}{\le} \min\left\{1, f^{(n)}_h(s_h,a_h) + \hat{P}^{(n)}_h \left|V^{\pi}_{h+1,\hat{P}^{(n)},r} - V^{\pi}_{h+1,P^\star,r}\right|(s_h,a_h)\right\} \\
&\quad\overset{(ii)}{\le} \min\left\{1, f^{(n)}_h(s_h,a_h) + \hat{P}^{(n)}_h \hat{V}^{\pi}_{h+1,\hat{P}^{(n)},f^{(n)}}(s_h,a_h)\right\} \\
&\quad= \hat{Q}^{\pi}_{h,\hat{P}^{(n)},f^{(n)}}(s_h,a_h),
\end{aligned} \tag{16}$$

where (i) follows from the (action) value function is at most 1, and (ii) follows from the induction hypothesis.

Then, by the definition of $\hat{V}^{\pi}_{h,\hat{P}^{(n)},r}(s_h)$, we have

$$\begin{aligned}
&\left|V^{\pi}_{h,\hat{P}^{(n)},r}(s_h) - V^{\pi}_{h,P^\star,r}(s_h)\right| \\
&\quad= \left|\mathop{\mathbb{E}}_{\pi}\left[Q^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h)\right] - \mathop{\mathbb{E}}_{\pi}\left[Q^{\pi}_{h,P^\star,r}(s_h,a_h)\right]\right| \\
&\quad\le \mathop{\mathbb{E}}_{\pi}\left[\left|Q^{\pi}_{h,\hat{P}^{(n)},r}(s_h,a_h) - Q^{\pi}_{h,P^\star,r}(s_h,a_h)\right|\right] \\
&\quad\overset{(i)}{\le} \mathop{\mathbb{E}}_{\pi}\left[\hat{Q}^{\pi}_{h,\hat{P}^{(n)},f^{(n)}}(s_h,a_h)\right] \\
&\quad= \hat{V}^{\pi}_{h,\hat{P}^{(n)},f^{(n)}}(s_h),
\end{aligned}$$

where (i) follows from Equation (16).

Therefore, by induction, we have

$$\left|V^{\pi}_{P^\star,r} - V^{\pi}_{\hat{P}^{(n)},r}\right| \le \hat{V}^{\pi}_{\hat{P}^{(n)},f^{(n)}}.$$

Step 2. Then, we show that $\hat{V}^{\pi}_{\hat{P}^{(n)},f^{(n)}} \le \hat{V}^{\pi}_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{K\zeta_n}$.

By Equation (12) and the fact that the total variation distance is upper bounded by 1, with probability at least $1-\delta/2$, we have

$$\mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[f^{(n)}_h(s_h,a_h)\middle| s_{h-1}\right] \le \mathop{\mathbb{E}}_{a_{h-1}\sim\pi}\left[\min\left(\alpha_n \left\|\hat{\phi}^{(n)}_{h-1}(s_{h-1},a_{h-1})\right\|_{(U^{(n)}_{h-1,\hat{\phi}})^{-1}}, 1\right)\right], \forall h \ge 2. \tag{17}$$

Similarly, when $h = 1$,

$$\mathop{\mathbb{E}}_{a_1\sim\pi}\left[f_1^{(n)}(s_1,a_1)\right] \le \sqrt{K \mathop{\mathbb{E}}_{a\sim\mathcal{U}}\left[\left(f_1^{(n)}(s_1,a_1)\right)^2\right]} \le \sqrt{K\zeta_n}. \tag{18}$$

Based on Corollary 1, Equation (17) and $\alpha_n = 5\hat{\alpha}_n$, we have

$$\mathop{\mathbb{E}}_{\pi}\left[\hat{b}_h^{(n)}(s_h,a_h)\middle| s_h\right] \ge \mathop{\mathbb{E}}_{\pi}\left[\min\left(\alpha_n \left\|\hat{\phi}_h^{(n)}(s_h,a_h)\right\|_{(U_{h,\hat{\phi}}^{(n)})^{-1}}, 1\right)\right] \ge \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[f_{h+1}^{(n)}(s_{h+1},a_{h+1})\middle| s_h\right]. \tag{19}$$

For the base case $h = H$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{H,\hat{P}^{(n)},f^{(n)}}(s_H)\middle| s_{H-1}\right] &= \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[f_H^{(n)}(s_H,a_H)\middle| s_{H-1}\right] \\
&\le \mathop{\mathbb{E}}_{\pi}\left[b_{H-1}^{(n)}(s_{H-1},a_{H-1})|s_{H-1}\right] \\
&\le \min\left\{1, \mathop{\mathbb{E}}_{\pi}\left[\hat{Q}^{\pi}_{H-1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{H-1},a_{H-1})\middle| s_{H-1}\right]\right\} \\
&= \hat{V}^{\pi}_{H-1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{H-1}).
\end{aligned}$$

Assume that $\mathbb{E}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{h+1,\hat{P}^{(n)},f^{(n)}}(s_{h+1})\middle| s_h\right] \le \hat{V}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h)$ holds for step $h+1$. Then, by Jensen's inequality, we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{h,\hat{P}^{(n)},f^{(n)}}(s_h)\middle| s_{h-1}\right] \\
&\quad\le \min\left\{1, \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[f_h^{(n)}(s_h,a_h) + \hat{P}_h^{(n)}\hat{V}^{\pi}_{h+1,\hat{P}^{(n)},f^{(n)}}(s_h,a_h)\middle| s_{h-1}\right]\right\} \\
&\quad\overset{(i)}{\le} \min\left\{1, \mathop{\mathbb{E}}_{\pi}\left[\hat{b}_{h-1}^{(n)}(s_{h-1},a_{h-1})\right] + \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{h+1,\hat{P}^{(n)},f^{(n)}}(s_{h+1})\middle| s_h\right]\middle| s_{h-1}\right]\right\} \\
&\quad\overset{(ii)}{\le} \min\left\{1, \mathop{\mathbb{E}}_{\pi}\left[b_{h-1}^{(n)}(s_{h-1},a_{h-1})\right] + \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h)\middle| s_{h-1}\right]\right\} \\
&\quad= \min\left\{1, \mathop{\mathbb{E}}_{\pi}\left[\hat{Q}^{\pi}_{h-1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{h-1},a_{h-1})\right]\right\} \\
&\quad= \hat{V}^{\pi}_{h-1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{h-1}),
\end{aligned}$$

where (i) follows from Equation (19), and (ii) is due to the induction hypothesis.

By induction, we conclude that

$$\begin{aligned}
\hat{V}^{\pi}_{\hat{P}^{(n)},f^{(n)}} &= \mathop{\mathbb{E}}_{\pi}\left[f_1^{(s)}(s_1,a_1)\right] + \mathop{\mathbb{E}}_{\hat{P}^{(n)},\pi}\left[\hat{V}^{\pi}_{2,\hat{P}^{(n)},f^{(n)}}(s_2)\middle| s_1\right] \\
&\le \sqrt{K\zeta_n} + \hat{V}^{\pi}_{\hat{P}^{(n)},\hat{b}^{(n)}}.
\end{aligned}$$

Combining Step 1 and Step 2, we conclude that

$$\left|V^{\pi}_{P^\star,r} - V^{\pi}_{\hat{P}^{(n)},r}\right| \le \sqrt{K\zeta_n} + \hat{V}^{\pi}_{\hat{P}^{(n)},\hat{b}^{(n)}}.$$

□

A.3 PROOF OF PROPOSITION 2

The following lemma is key to ensure that RAFFLE terminates in finite episodes.

Proposition 5 (Restatement of Proposition 2)**.** *Given that the event $\mathcal{E}$ occurs, $\zeta = \log(2|\Phi||\Psi|NH/\delta)$ the summation of the truncated value functions $\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}}$ under exploration policies $\{\pi_n\}_{n\in[N]}$ is sublinear, i.e., the following bound holds:*

$$\sum_{n\in[N]} \hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{K\zeta_n} \leq 32\zeta H d\sqrt{\beta_3 K(d^2+K)N}.$$

Proof. Note that $\hat{V}^{\pi}_{h,\hat{P}^{(n)},\hat{b}^{(n)}} \leq 1$ holds for any policy π and $h \in [H]$. We first have

$$\begin{aligned}
\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} - V^{\pi_n}_{P^\star,\hat{b}^{(n)}} &\leq \mathop{\mathbb{E}}_{\pi_n}\left[\hat{P}^{(n)}_1 \hat{V}^{\pi_n}_{2,\hat{P}^{(n)},\hat{b}^{(n)}}(s_1,a_1) - P^\star_1 V^{\pi_n}_{2,P^\star,\hat{b}^{(n)}}(s_1,a_1)\right] \\
&= \mathop{\mathbb{E}}_{\pi_n}\left[\left(\hat{P}^{(n)}_1 - P^\star_1\right)\hat{V}^{\pi_n}_{2,\hat{P}^{(n)},\hat{b}^{(n)}}(s_1,a_1) + P^\star_1\left(\hat{V}^{\pi_n}_{2,\hat{P}^{(n)},\hat{b}^{(n)}} - V^{\pi_n}_{2,P^\star,\hat{b}^{(n)}}\right)(s_1,a_1)\right] \\
&\leq \mathop{\mathbb{E}}_{\pi_n}\left[f^{(n)}_1(s_1,a_1) + P^\star_1\left(\hat{V}^{\pi_n}_{2,\hat{P}^{(n)},\hat{b}^{(n)}} - V^{\pi_n}_{2,P^\star,\hat{b}^{(n)}}\right)\right] \\
&\leq \dots \\
&\leq \mathbb{E}_{(s_h,a_h)\sim(P^\star,\pi_n)}\left[\sum_{h=1}^{H} f^{(n)}(s_h,a_h)\right] = V^{\pi_n}_{P^\star,f^{(n)}},
\end{aligned}$$

which implies $\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} \leq V^{\pi_n}_{P^\star,\hat{b}^{(n)}} + V^{\pi_n}_{P^\star,f^{(n)}}$.

Applying the Equation (14) and Equation (15), we obtain the following bound on the value function $V^{\pi_n}_{P^\star,\hat{b}^{(n)}}$:

$$\begin{aligned}
V^{\pi_n}_{P^\star,\hat{b}^{(n)}} &= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\hat{b}_n(s_h,a_h)\right] \\
&\leq \sum_{h=2}^{H} \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(P^\star,\pi_n)\\ a_{h-1}\sim\pi_n}}\left[\gamma_n \left\|\phi^\star_{h-1}(s_{h-1},a_{h-1})\right\|_{(W^{(n)}_{h-1,\phi^\star})^{-1}}\right] + 15\alpha_n\sqrt{\frac{dK}{n}} \\
&\leq \sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\gamma_n \left\|\phi^\star_{h}(s_h,a_h)\right\|_{(W^{(n)}_{h,\phi^\star})^{-1}}\right] + 15\alpha_n\sqrt{\frac{dK}{n}}.
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
V^{\pi_n}_{P^\star,f^{(n)}} &= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[f^{(n)}_h(s_h,a_h)\right] \\
&\leq \sum_{h=2}^{H} \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(P^\star,\pi_n)\\ a_{h-1}\sim\pi_n}}\left[\alpha_n \left\|\phi^\star_{h-1}(s_{h-1},a_{h-1})\right\|_{(U^{(n)}_{h-1,\phi^\star})^{-1}}\right] + \sqrt{K\zeta_n} \\
&\leq \sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\alpha_n \left\|\phi^\star_{h}(s_h,a_h)\right\|_{(U^{(n)}_{h,\phi^\star})^{-1}}\right] + \sqrt{K\zeta_n}.
\end{aligned}$$

Then, taking the summation of $V^{\pi_n}_{P^\star,\hat{b}^{(n)}+f^{(n)}}$ over $n \in [N]$, we have

$$
\begin{aligned}
&\sum_{n\in[N]} V^{\pi_n}_{P^\star,f^{(n)}+\hat{b}^{(n)}} + \sqrt{K\zeta_n} \\
&\quad\le \sum_{n\in[N]} 15\alpha_n\sqrt{\frac{dK}{n}} + 2\sum_{n\in[N]}\sqrt{K\zeta_n} + \sum_{n\in[N]}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\gamma_n \left\|\phi^\star_h(s_h,a_h)\right\|_{(W^{(n)}_{h,\phi^\star})^{-1}}\right] \\
&\qquad + \sum_{n\in[N]}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\alpha_n \left\|\phi^\star_h(s_h,a_h)\right\|_{(U^{(n)}_{h,\phi^\star})^{-1}}\right] \\
&\quad\overset{(i)}{\le} 17\alpha_N\sqrt{dKN} + \gamma_N\sum_{h=1}^{H}\sqrt{N\sum_{n\in[N]} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\pi_n}}\left[\left\|\phi^\star_h(s_h,a_h)\right\|^2_{(W^{(n)}_{h,\phi^\star})^{-1}}\right]} \\
&\qquad + \alpha_N\sum_{h=1}^{H}\sqrt{KN\sum_{n\in[N]} \mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi_n)\\ a_h\sim\mathcal{U}}}\left[\left\|\phi^\star_h(s_h,a_h)\right\|^2_{(U^{(n)}_{h,\phi^\star})^{-1}}\right]} \\
&\quad\overset{(ii)}{\le} 17\sqrt{\zeta}\sqrt{2\beta_3 dK(K+d^2)N} + H\sqrt{45\beta_3\zeta dK(K+d^2)}\sqrt{dN\zeta} \\
&\qquad + H\sqrt{\beta_3\zeta(K+d^2)}\sqrt{dKN\zeta} \\
&\quad\le 32\zeta Hd\sqrt{\beta_3 K(d^2+K)N},
\end{aligned}
$$

where (i) follows from Cauchy-Schwarz inequality and importance sampling, and (ii) follows from Lemma 10. Hence, the statement of Proposition 5 is verified. □

A.4 Proof of Proposition 3

Based on Proposition 5, we argue that with enough number of iterations, RAFFLE can find $\hat{P}^\epsilon$ satisfying the condition in line 15 of Algorithm 1.

Proposition 6 (Restatement of Proposition 3)**.** *Fix any $\delta \in (0,1), \epsilon > 0$. Suppose the algorithm runs for $N = \frac{2^{14}\beta_3 H^2 d^2 K(d^2+K)\log^2(2|\Phi||\Psi|H^3 d^2 K(d^2+K)/(\delta\epsilon^2))}{\epsilon^2}$ iterations, with probability at least $1-\delta$, RAFFLE can find an $n_\epsilon \le N$ in the exploration phase such that $2\hat{V}^{\pi_{n_\epsilon}}_{\hat{P}^{(n_\epsilon)},\hat{b}^{(n_\epsilon)}} + 2\sqrt{K\zeta_{n_\epsilon}} \le \epsilon$. In other words, Algorithm 1 can output $\hat{P}^\epsilon = \hat{P}^{(n_\epsilon)}$ satisfying the condition in line 15. In addition*

$$
\left|V^{\pi}_{P^\star,r} - V^{\pi}_{\hat{P}^{(n_\epsilon)},r}\right| \le \epsilon/2.
$$

Proof. We show that the algorithm terminates by contradiction. If it does not stop, applying Proposition 5, we have

$$
\begin{aligned}
\epsilon N/2 &< \sum_{n\in[N]}\left(\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}} + \sqrt{K\zeta_n}\right) \\
&\le 32\zeta Hd\sqrt{\beta_3 K(d^2+K)N}.
\end{aligned}
$$

Therefore,

$$
N < \frac{2^{12}\beta_3 H^2 d^2 K(d^2+K)\zeta^2}{\epsilon^2}
$$

Recall $\zeta = N\zeta_N = \log(2|\Phi||\Psi|NH/\delta)$. Using the fact that $n \le c\log^2(\alpha_n n) \Rightarrow n \le 4c\log^2(\alpha_n c), \forall c \ge e^2, n \ge 1, \alpha_n \in \mathbb{R}^+$, it can be concluded that

$$
N < \frac{2^{14}\beta_3 H^2 d^2 K(d^2+K)\log^2(2|\Phi||\Psi|H^3 d^2 K(d^2+K)/(\delta\epsilon^2))}{\epsilon^2},
$$

which is a contradiction.

Therefore, there exists an $n_\epsilon = O(\frac{H^2 d^2 K(d^2+K)\log^2(|\Phi||\Psi|H^3 d^2 K(d^2+K)/(\delta\epsilon^2))}{\epsilon^2})$ such that $\hat{P}^\epsilon = \hat{P}^{(n_\epsilon)}$ satisfies

$$2\hat{V}^{\pi_{n_\epsilon}}_{\hat{P}^{(n_\epsilon)},\hat{b}^{(n_\epsilon)}} + 2\sqrt{K\zeta_{n_\epsilon}} \leq \epsilon.$$

Combining Proposition 4, we finish the proof.

□

Proof of Theorem 1. Recall that $\hat{P}^\epsilon$ is the output of RAFFLE in the n_ϵ-iteration. Then, by Proposition 6

$$\begin{aligned}
&V^\star_{P^\star,r} - V^{\bar{\pi}}_{P^\star,r} \\
&\quad \leq V^{\pi^\star}_{\hat{P}^\epsilon,r} - V^{\bar{\pi}}_{P^\star,r} + \epsilon/2 \\
&\quad \overset{(i)}{\leq} V^{\bar{\pi}}_{\hat{P}^\epsilon,r} - V^{\bar{\pi}}_{P^\star,r} + \epsilon/2 \\
&\quad \leq \epsilon/2 + \epsilon/2 \\
&\quad = \epsilon,
\end{aligned}$$

where (i) follows from the definition of $\bar{\pi}$. The number of trajectories $n_\epsilon H$ is at at most

$$O\left(\frac{H^3 d^2 K(d^2+K)\log^2(|\Phi||\Psi|H^3 d^2 K(d^2+K)/(\delta\epsilon^2))}{\epsilon^2}\right)$$

□

B Proof of Theorem 2

In this section, we adopt the same notations as in Appendix A. The following lemma provides an upper bound for the estimation error of any learned model from the true model.

Lemma 4. *Fix $\delta \in (0,1)$, for any $h \in [H], n \in \mathbb{N}^+$, any policy π, with probability at least $1-\delta/2$,*

$$\mathop{\mathbb{E}}_{\substack{s_h\sim(P^\star,\pi)\\ s_h\sim\pi}}\left[f^{(n)}_h(s_h,a_h)\right] \leq 2\sqrt{K\zeta_n} + 2\hat{V}^{\pi_n}_{\hat{P}^{(n)},\hat{b}^{(n)}}.$$

Proof. Recall that $f^{(n)}_h(s,a) = \left\|\hat{P}^{(n)}_h(\cdot|s,a) - P^\star_h(\cdot|s,a)\right\|_{TV}$. Fix any policy π, for any $h \geq 2$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{s_h\sim(\hat{P}^{(n)},\pi)\\ a_h\sim\pi}}\left[\hat{Q}^\pi_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_h,a_h)\right] &= \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(\hat{P}^{(n)},\pi)\\ a_{h-1}\sim\pi}}\left[\hat{P}^{(n)}_h \hat{V}^\pi_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{h-1},a_{h-1})\right] \\
&\leq \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(\hat{P}^{(n)},\pi)\\ a_{h-1}\sim\pi}}\left[\min\left\{1, \hat{b}^{(n)}_{h-1}(s_{h-1},a_{h-1}) + \hat{P}^{(n)}_{h-1}\hat{V}^\pi_{h,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{h-1},a_{h-1})\right\}\right] \\
&= \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(\hat{P}^{(n)},\pi)\\ a_{h-1}\sim\pi}}\left[\hat{Q}^\pi_{h-1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_{h-1},a_{h-1})\right] \\
&\leq \cdots \\
&\leq \mathop{\mathbb{E}}_{a_1\sim\pi}\left[\hat{Q}^\pi_{1,\hat{P}^{(n)},\hat{b}^{(n)}}(s_1,a_1)\right] \\
&= \hat{V}^\pi_{\hat{P}^{(n)},\hat{b}^{(n)}}. \qquad (20)
\end{aligned}$$

Hence, for $h \geq 2$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{s_h \sim (\hat{P}^{(n)}, \pi) \\ a_h \sim \pi}} \left[f_h^{(n)}(s_h, a_h) \right] &\overset{(i)}{\leq} \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (\hat{P}^{(n)}, \pi) \\ a_{h-1} \sim \pi}} \left[\hat{b}_{h-1}^{(n)}(s_{h-1}, a_{h-1}) \right] \\
&\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (\hat{P}^{(n)}, \pi) \\ a_{h-1} \sim \pi}} \left[\hat{Q}^{\pi}_{h-1, \hat{P}^{(n)}, \hat{b}^{(n)}}(s_{h-1}, a_{h-1}) \right] \\
&\overset{(iii)}{\leq} \hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}}, \qquad (21)
\end{aligned}$$

where (i) follows from Equation (19), (ii) follows from the definition of $\hat{Q}^{\pi}_{h-1, \hat{P}^{(n)}, \hat{b}^{(n)}}(s_{h-1}, a_{h-1})$ and (iii) follows from Equation (20).

$$\begin{aligned}
\mathbb{E}_{\substack{\sim (P^\star, \pi) \\ s_h \sim \pi}} \left[f_h^{(n)}(s_h, a_h) \right] &\leq \mathop{\mathbb{E}}_{\substack{s_h \sim (\hat{P}^{(n)}, \pi) \\ a_h \sim \pi}} \left[f_h^{(n)}(s_h, a_h) \right] + \left| \mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \pi) \\ a_h \sim \pi}} \left[f_h^{(n)}(s_h, a_h) \right] - \mathop{\mathbb{E}}_{\substack{s_h \sim (\hat{P}^{(n)}, \pi) \\ a_h \sim \pi}} \left[f_h^{(n)}(s_h, a_h) \right] \right| \\
&\overset{(i)}{\leq} (\hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}} + \sqrt{K\zeta_n}) + \left(\sqrt{K\zeta_n} + \hat{V}^{\pi}_{\hat{P}^{(n)}, \hat{b}^{(n)}} \right) \\
&\overset{(ii)}{\leq} 2\sqrt{K\zeta_n} + 2\hat{V}^{\pi_n}_{\hat{P}^{(n)}, \hat{b}^{(n)}},
\end{aligned}$$

where the first term in (i) is due to Equation (21) and Equation (15), the second term in (i) is due to Proposition 4 and (ii) follows from the definition of π_n. □

Proof of Theorem 2. By Proposition 6, let $n_\epsilon = O\left(\frac{H^2 d^2 K (d^2 + K) \log^2(|\Phi||\Psi| N H^3 d^2 K (K + d^2)/(\delta \epsilon^2))}{\epsilon^2} \right)$, with no more than $n_\epsilon H$ trajectories, RAFFLE can learn a model $\hat{P}^\epsilon$, bonus $\hat{b}^\epsilon$ and policy π_ϵ at the n_ϵ-th iteration satisfying $2V^{\pi_\epsilon}_{\hat{P}^\epsilon, \hat{b}^\epsilon} + 2\sqrt{K\zeta_{n_\epsilon}} \leq \epsilon$. Then, following Lemma 4, we have

$$\mathop{\mathbb{E}}_{\substack{s_h \sim (P^\star, \pi) \\ s_h \sim \pi}} \left[\|\hat{P}^\epsilon_h(\cdot|s_h, a_h) - P^\star_h(\cdot|s_h, a_h)\|_{TV} \right] \leq 2\sqrt{K\zeta_{n_\epsilon}} + 2V^{\pi_\epsilon}_{\hat{P}^\epsilon, \hat{b}^\epsilon} \leq \epsilon.$$

□

C Proof of Theorem 3 (Lower Bound on Sample Complexity)

We first summarize the outline of our proof here. The idea of the proof is as follows. We first construct a low-rank representation for the tabular case with low dimension. The state-of-the-art lower bound for tabular MDPs is given by $\Omega\big(\frac{H|\mathcal{S}||\mathcal{A}|}{\epsilon^2}\big)$, where $\mathcal{S}$ and $|\mathcal{A}|$ are the state and action spaces, respectively. By embedding $\mathcal{S} \times \mathcal{A}$ into a $d = \Omega\left(|\mathcal{S}|\right)$-dimensional space, we are able to show that the lower bound for low-rank MDPs is given by $\Omega\big(\frac{HdK}{\epsilon^2}\big)$.

Proof of Theorem 3. Section 3.1 in Domingues et al. (2020) provides a lower bound $\Omega\left(\frac{H|\mathcal{S}||\mathcal{A}|\log(1/\delta)}{\epsilon^2}\right)$ for tabular MDP by constructing a set of MDP instances. Specifically, There are $H|\mathcal{S}||\mathcal{A}|$ MDPs which are hard to distinguish in KL divergence, while the corresponding optimal policies are completely different. By leveraging tools from information theory, it shows that at least $\Omega\left(H|\mathcal{S}||\mathcal{A}|/\epsilon^2\right)$ sample trajectories are required to identify an ϵ-optimal policy.

Below, we restate the MDP structure as in Figure 1 and then embed the states and actions into $\mathbb{R}^{\Omega(|\mathcal{S}|)}$, where $\mathcal{S}$ is the state space. We normalize the reward functions to satisfy $\sum_h r_h \leq 1$.. Note that for any transition probability $P(s_{h+1}|s_h, a_h) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, we can always find embeddings $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu \to \mathbb{R}^d$, where $d = |\mathcal{S}||\mathcal{A}|$. If we index coordinates by (s, a), then $\phi(s, a)$ is a standard unit vector $e_{(s,a)}$ and the (s, a)-th coordinate of $\mu(s_{h+1})$ is the transition probability $P(s_{h+1}|s, a)$. We call such construction the standard embedding with dimension $d = |\mathcal{S}||\mathcal{A}|$. Another fact is that, we can always embedding a low dimensional space into a higher dimensional space.

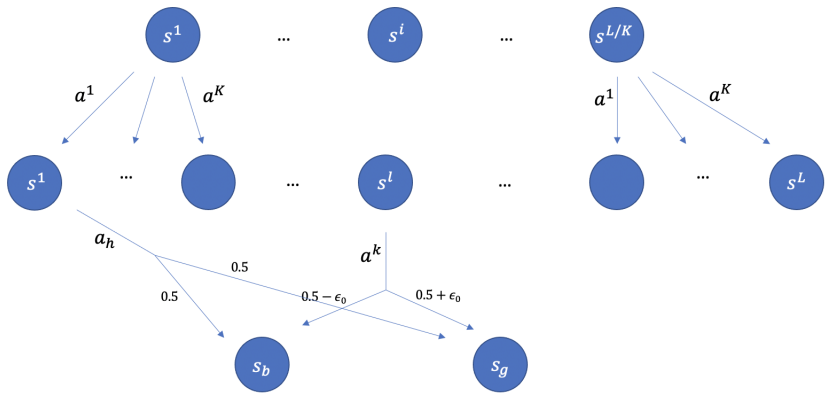

Figure 1: Tree-structure of the MDP.

Now, we are ready to construct the representations for instances in Domingues et al. (2020). Since the MDP has a K-ary tree, as illustrated in Figure 1, at the level of leaves, we have the most states denoted by $s^1, \ldots, s^L$. Let h be the step where the leaves lie. Note that at $h+1$, we only have two states s_g and s_b representing a good state and a bad state. Suppose choosing a^k at state s^ℓ will increase the probability visiting s_g by ϵ_0. Then, the transition probabilities at step h are defined as follows.

$$P_h(s_g|s_h, a_h) = \frac{1}{2} + \epsilon_0 \mathbf{1}_{\{s_h = s^\ell, a_h = a^k\}}, \tag{22}$$

$$P_h(s_b|s_h, a_h) = \frac{1}{2} - \epsilon_0 \mathbf{1}_{\{s_h = s^\ell, a_h = a^k\}}. \tag{23}$$

We can define 2-dimensional features over the transition probabilities as follow.

$$\phi_h(s_h, a_h) = \left(\frac{1}{2}, \frac{1}{2}\right)^\top + \mathbf{1}_{\{s_h = s^\ell, a_h = a^k\}}(\epsilon_0, -\epsilon_0)^\top, \tag{24}$$

$$\mu_h(s_g) = (1, 0), \tag{25}$$

$$\mu_h(s_b) = (0, 1). \tag{26}$$

We finally construct a standard embedding with dimension L on the $(h-1)$-th step, since the number of state is L/K and the number of actions is K. If an algorithm `Alg` can identify an ϵ-optimal policy on those MDPs, according to Section 3.1 in Domingues et al. (2020), `Alg` should collect at least $\Omega\left(HL|\mathcal{A}|\log(1/\delta)/\epsilon^2\right)$ sample trajectories. With the fact that $d = L$, a lower bound $\Omega\left(\frac{HdK}{\epsilon^2}\right)$ for low-rank MDP is hence established. □

D Algorithm for Representation Learning in Section 5.3 and Proof of Theorem 4

We first present the full algorithm in Section 5.3 below as Algorithm 2.

D.1 Discussion on Reachability Assumption: Assumption 2

As discussed in Section 5.3, intuitively, if some state action pairs can be hardly visited by any policy, the output of Algorithm 1 $\hat{P}(\cdot|s,a)$ can not approximate true $P^\star(\cdot|s,a)$ well over these state action pairs, so the input distribution naturally requires restrictions. On the other hand, it is easy for input distributions $\{q_h\}_{h=1}^H$ to satisfy Assumption 2 if each state action pair under true MDPs can be explored well by some policies. In fact if we assume following reachability assumption, Assumption 2 holds with for any $\{q_h\}_{h=1}^H$ with finite bound C_B. Such reachability type assumption is common in relevant literature (Modi et al., 2021; Agarwal et al., 2020).

Algorithm 2 Representation Learning via RAFFLE

Input: Sample size N_f, state action pair distribution $\{q_h\}_{h=1}^H$, special designed reward function $\{r^{h,t}\}_{h\in[H],t\in[T]}$ and policy $\{\pi^t\}_{t\in[T]}$ and estimated transition kernel $\hat{P}$ from the output of Algorithm 1, model class: Φ.
Initialize $\mathcal{D}_h^{0,0} = \emptyset$.
for $n = 1, \ldots, N_f$ **do**
  for $h = 1, \ldots, H$ **do**
    for $t = 1, \ldots, T$ **do**
      Choose $(s_h^{n,t}, a_h^{n,t}) \sim q_h$ and add $(s_h^{n,t}, a_h^{n,t})$ to the dataset $\mathcal{D}_h^{n,t} = \mathcal{D}_h^{n-1,t} \cup (s_h^{n,t}, a_h^{n,t})$.
    end for
  end for
end for
for $h = 1, \ldots, H$ **do**
  Learn $(\tilde{\phi}_h, \tilde{w}_h^1, \ldots, \tilde{w}_h^T)$ by

$$\underset{\phi_h\in\Phi, w_h^t\in\mathbb{R}^d}{\arg\min} \sum_{t\in[T]} \sum_{(s_h,a_h)\in\mathcal{D}_h^{N_f,t}} (Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h,a_h) - \langle \phi_h(s_h,a_h), w_h^t\rangle)^2.$$

end for
Output: $\tilde{\phi} = \{\tilde{\phi}_h\}_{h\in[H]}$.

Assumption 4 (Reachability)**.** For the true transition kernel $P^\star$, there exists a policy π^0 such that $\min_{s\in\mathcal{S}} \mathbb{P}_h^{\pi^0}(s) \geq \eta_{\min}$, where $\mathbb{P}_h^{\pi^0}(\cdot) : \mathcal{S} \to \mathbb{R}$ is the density function over $\mathcal{S}$ using policy π^0 to roll into state s at timestep h.

Under Assumption 4, for any bounded input distributions $\{q_h\}_{h=1}^H$ with bound C_B, Assumption 4 holds with $C_{\min} = \frac{C_B}{\eta_{\min}}$.

D.2 Supporting Lemmas

We first show that $Q^\pi_{\hat{P},h,r}$ can approximate $Q^\pi_{P^\star,h,r}$ well over distribution $\{q_h\}_{h\in[H]}$ by following two lemmas.

Lemma 5. *Given any $\delta \in (0,1)$. Let $\hat{P}$ be the output of Algorithm 1, for any policy π and rewards r, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{(s_h,a_h)\sim(\hat{P},\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h) - Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right] \leq \epsilon$$
$$\mathbb{E}_{(s_h,a_h)\sim(P^\star,\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h) - Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right] \leq \epsilon.$$

Proof. Define $\hat{f}_h(s_h,a_h) = \left\|\hat{P}_h(\cdot|s_h,a_h) - P^\star(\cdot|s_h,a_h)\right\|_{TV}$ and $\hat{f}$ is a collection of all $\hat{f}_h$, i.e. $\left\{\hat{f}_h\right\}_{h\in[H]}$.

$$\begin{aligned}
&\mathbb{E}_{(s_h,a_h)\sim(\hat{P},\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h) - Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right] \\
&\overset{(i)}{\leq} \mathbb{E}_{(s_h,a_h)\sim(\hat{P},\pi)}\left[\hat{Q}^\pi_{h,\hat{P},\hat{f}}(s_h,a_h)\right] \\
&\overset{(ii)}{\leq} \hat{V}^\pi_{\hat{P},\hat{f}} \\
&\overset{(iii)}{\leq} \epsilon/2,
\end{aligned} \tag{27}$$

where (i) follows from Equation (16), (ii) follows from the definition of $\hat{V}$ and $\hat{Q}$, and (iii) follows from the proof of Theorem 1.

Then for the second inequality,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(s_h,a_h)\sim(P^\star,\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\\
&\leq \mathop{\mathbb{E}}_{(s_h,a_h)\sim(\hat{P},\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\\
&+\left|\mathbb{E}_{(s_h,a_h)\sim(P^\star,\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]-\mathbb{E}_{(s_h,a_h)\sim(\hat{P},\pi)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\right|\\
&\overset{(i)}{\leq} \epsilon/2+V^\pi_{\hat{P},\hat{f}}\\
&\overset{(ii)}{\leq} \epsilon,
\end{aligned}
$$

where the first term in (i) is due to Equation (27) and Equation (15), the second term in (i) is due to Step 1 in Proposition 4 and (ii) follows from the proof of Theorem 1. □

Lemma 6. *Given any $\delta,\epsilon\in(0,1)$ and the output of Algorithm 1, under Assumption 2, for any policy π and rewards r, let $\{q_h\}_{h=1}^H$ be any distribution satisfying Assumption 2. Then with probability at least $1-\delta$, for each $h\in[H]$, $\mathbb{E}_{(s_h,a_h)\sim q_h}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\leq\epsilon C_{\min}$.*

Proof.

$$
\begin{aligned}
&\mathbb{E}_{(s_h,a_h)\sim q_h}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\\
&=\int q_h(s_h,a_h)\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|ds_hda_h\\
&\overset{(i)}{=}\int C_{\min}\mathbb{P}^{\pi^0}_h(s_h)\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|ds_hda_h\\
&=C_{\min}\mathop{\mathbb{E}}_{(s_h,a_h)\sim(P^\star,\pi^0)}\left[\left|Q^\pi_{P^\star,h,r}(s_h,a_h)-Q^\pi_{\hat{P},h,r}(s_h,a_h)\right|\right]\\
&\overset{(ii)}{\leq}\epsilon C_{\min},
\end{aligned}
$$

where (i) follows Assumption 4 and (ii) follows Lemma 5. □

The lemma above shows that for any reward r, $Q^\pi_{\hat{P},h,r}(s_h,a_h)$ can be the target of $Q^\pi_{P^\star,h,r}(s_h,a_h)$ when (s_h,a_h) are chosen from given distribution q_h.

We then show that for any h, when reward r is set to be zero at step h, $Q^\pi_{P^\star,h,r}$ has a linear structure w.r.t $\phi^\star_h$.

Lemma 7. *For any $h\in[H]$, policy π and given $(s_h,a_h)\in\mathcal{S}\times\mathcal{A}$, given any r such that r is set to be zero at step h, i.e. $r_h=0$, then $Q^\pi_{P^\star,h,r}(s_h,a_h)$ is linear with respect to $\phi^\star(s_h,a_h)$, i.e. there exist a $w^\star_h$ such that $Q^\pi_{P^\star,h,r}(s_h,a_h)=\langle\phi^\star_h(s_h,a_h),w^\star_h\rangle$.*

Proof.

$$
\begin{aligned}
Q^\pi_{P^\star,h,r}(s_h,a_h)&=r_h(s_h,a_h)+\mathbb{E}_{s_{h+1}\sim P^\star(\cdot|s_h,a_h)}\left[V^\pi_{P^\star,h+1,r}(s_{h+1})\middle|s_h,a_h\right]\\
&=\int P^\star(s_{h+1}|s_h,a_h)V^\pi_{P^\star,h+1,r}(s_{h+1})ds_{h+1}\\
&=\left\langle\phi^\star_h(s_h,a_h),\int\mu^\star_h(s_{h+1})V^\pi_{P^\star,h+1,r}(s_{h+1})ds_{h+1}\right\rangle\\
&=\left\langle\phi^\star_h(s_h,a_h),w^\star_h\right\rangle,
\end{aligned}
$$

where $w^\star_h=\int\mu^\star_h(s_{h+1})V^\pi_{P^\star,h+1,r}(s_{h+1})ds_{h+1}$.

□

D.3 PROOF OF THEOREM 4

Proof of Theorem 4. We allow ϕ and $Q^{\pi^t}_{\hat{P},h,r^{h,t}}$ to apply to all the samples in a dataset matrix simultaneously, i.e. $\phi_h(\mathcal{D}_h^{N_f,t}) = (\phi_h(s_h^{1,t}),\ldots,\phi_h(s_h^{N_f,t}))^\top \in \mathbb{R}^{N_f\times d}$ and $Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t}) = (Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{1,t}),\ldots,Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{N_f,t}))^\top \in \mathbb{R}^{N_f}$. After we estimating $\tilde{\phi}_h$ and then taking it as known, from the property of linear regression in Equation (5), for any reward function $r^{h,t}$, any policy π^t we got

$$\widetilde{w}_h^t = (\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top\tilde{\phi}_h(\mathcal{D}_h^{N_f,t}))^\dagger\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top Q^{\pi}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})$$
$$\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})\widetilde{w}_h^t = P_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t}),$$

where $P_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})} = \tilde{\phi}_h(\mathcal{D}_h^{N_f,t})(\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top\tilde{\phi}_h(\mathcal{D}_h^{N_f,t}))^\dagger\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top$, represents the projection operator to the column spaces of $\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})$. Then

$$\begin{aligned}
&\sum_{t\in[T]}\left\|P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right\|^2\\
&=\sum_{t\in[T]}\left\|\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})\widetilde{w}_h^t - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right\|^2\\
&\overset{(i)}{\le}\sum_{t\in[T]}\left\|\phi_h^\star(\mathcal{D}_h^{N_f,t})w_h^{t\,\star} - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right\|^2\\
&=\sum_{t\in[T]}\sum_{n=1}^{N_f}\left(Q^{\pi^t}_{P^\star,h,r^{h,t}}(s_h^{n,t},a_h^{n,t}) - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{n,t},a_h^{n,t})\right)^2\\
&\overset{(ii)}{\le} N_f\sum_{t\in[T]}\mathbb{E}_{(s_h,a_h)\sim q_h}\left[\left(Q^{\pi^t}_{P^\star,h,r^{h,t}}(s_h^{n,t},a_h^{n,t}) - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{n,t},a_h^{n,t})\right)^2\right] + \sqrt{\frac{TN_f\log\frac{2}{\delta}}{2}}\\
&\overset{(iii)}{\le} N_f\sum_{t\in[T]}\mathbb{E}_{(s_h,a_h)\sim q_h}\left[\left|Q^{\pi^t}_{P^\star,h,r^{h,t}}(s_h^{n,t},a_h^{n,t}) - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{n,t},a_h^{n,t})\right|\right] + \sqrt{\frac{TN_f\log\frac{2}{\delta}}{2}}\\
&\overset{(iv)}{\le}\epsilon C_{\min}N_fT + \sqrt{\frac{TN_f\log\frac{2}{\delta}}{2}},
\end{aligned}\tag{28}$$

where (i) follows from minimality of $\{\tilde{\phi}_h^t\}_{t\in[T]}$ and $\{\widetilde{w}_h^t\}_{t\in[T]}$, (ii) follows Hoeffding's inequality, (iii) follows that $\left|Q^{\pi^t}_{P^\star,h,r^{h,t}}(s_h^{n,t},a_h^{n,t}) - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(s_h^{n,t},a_h^{n,t})\right| \le 1$ and (iv) follows Lemma 6.

As a result:

$$\begin{aligned}
&\sum_{t\in[T]}\left\|P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}\phi_h^\star(\mathcal{D}_h^{N_f,t})w_h^{t\,\star}\right\|^2\\
&=\sum_{t\in[T]}\left\|P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}Q^{\pi^t}_{P^\star,h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right\|^2\\
&\le\sum_{t\in[T]}\left\{\left\|P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right\|^2 + \left\|P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}\left(Q^{\pi^t}_{P^\star,h,r^{h,t}}(\mathcal{D}_h^{N_f,t}) - Q^{\pi}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\right)\right\|^2\right\}\\
&\overset{(i)}{\le}\sum_{t\in[T]}\epsilon C_{\min}N_fT + \sqrt{\frac{TN_f\log\frac{2}{\delta}}{2}} + \sum_{t\in[T]}\sigma_1^2(P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})})\|Q^{\pi^t}_{P^\star,h,r^{h,t}}(\mathcal{D}_h^{N_f,t})) - Q^{\pi^t}_{\hat{P},h,r^{h,t}}(\mathcal{D}_h^{N_f,t})\|^2\\
&\overset{(ii)}{\le} 2\epsilon C_{\min}N_fT + \sqrt{2TN_f\log\frac{2}{\delta}},
\end{aligned}$$

where (i) follows from $\|Av\|_2 \leq \sigma_1(A)\|v\|_2$ and (ii) follows from that $\sigma_1(P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})}) \leq 1$ and the process to derive Equation (28).

We finally use the technique in Du et al. (2021b) to derive the super population guarantee.

$$
\begin{aligned}
&2\epsilon C_{\min} N_f T + \sqrt{2TN_f \log \frac{2}{\delta}} \\
&\geq \sum_{t\in[T]} \left\| P^{\perp}_{\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})} \phi_h^\star(\mathcal{D}_h^{N_f,t}) w_h^{t\star} \right\|_F^2 \\
&= \left\| \left(I - \tilde{\phi}_h(\mathcal{D}_h^{N_f,t}) \left(\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top \tilde{\phi}_h(\mathcal{D}_h^{N_f,t}) \right)^\dagger \tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top \right) \phi_h^\star(\mathcal{D}_h^{N_f,t}) w_h^{t\star} \right\|_F^2 \\
&= \sum_{t\in[T]} (w_h^{t\star})^\top \phi_h^\star(\mathcal{D}_h^{N_f,t})^\top \left(I - \tilde{\phi}_h(\mathcal{D}_h^{N_f,t})(\tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top \tilde{\phi}_h(\mathcal{D}_h^{N_f,t}))^\dagger \tilde{\phi}_h(\mathcal{D}_h^{N_f,t})^\top \right) \phi_h^\star(\mathcal{D}_h^{N_f,t}) w_h^{t\star} \\
&= \sum_{t\in[T]} N_f (w_h^{t\star})^\top D_{\mathcal{D}_h^{N_f,t}}(\phi_h^\star, \tilde{\phi}_h) w_h^{t\star} \\
&\overset{(i)}{\geq} 0.9 \sum_{t\in[T]} N_f (w_h^{t\star})^\top D_{q_h}(\phi_h^\star, \tilde{\phi}_h) w_h^{t\star} \\
&= 0.9 N_f \sum_{t\in[T]} \left\| D_{q_h}(\phi_h^\star, \tilde{\phi}_h)^{1/2} w_h^{t\star} \right\|^2 \\
&\geq 0.9 N_f \left\| D_{q_h}(\phi_h^\star, \tilde{\phi}_h)^{1/2} W_h^\star \right\|_F^2 \\
&\overset{(ii)}{\geq} 0.9 N_f \left\| D_{q_h}(\phi_h^\star, \tilde{\phi}_h)^{1/2} \right\|_F^2 \sigma_d^2(W_h^\star) \\
&\geq 0.9 N_f \left\| D_{q_h}(\phi_h^\star, \tilde{\phi}_h)^{1/2} \right\|_F^2 \frac{C_D T}{d},
\end{aligned}
$$

where (i) follows from lemma B.1 in Du et al. (2021b) and N_f can be large enough, and (ii) follows from that $\|AB\|_F^2 \geq \|A\|_F^2 \sigma_{\min}^2(B)$, where $\sigma_{\min}(B)$ denote the smallest singular value of matrix B. Then

$$
\left\| D_{q_h}(\phi_h^\star, \tilde{\phi}_h)^{1/2} \right\|_F^2 \leq \frac{2\epsilon d C_{\min}}{0.9 C_D} + \frac{d}{0.9 C_D} \sqrt{\frac{2 \log \frac{2}{\delta}}{T N_f}}.
$$

□

E AUXILIARY LEMMAS

Recall $f_h^{(n)}(s,a) = \|\hat{P}_h^{(n)}(\cdot|s,a) - P_h^\star(\cdot|s,a)\|_{TV}$ represents the estimation error in the n-th iteration at step h, given state s and action a, in terms of the total variation distance. By using Theorem 21 in Agarwal et al. (2020), we are able to guarantee that under all exploration policies, the estimation error can be bounded with high probability.

Lemma 8. (MLE guarantee). *Given $\delta \in (0,1)$, we have the following inequality holds for any $n, h \geq 2$ with probability at least $1 - \delta/2$:*

$$
\sum_{\tau=0}^{n-1} \mathop{\mathbb{E}}_{\substack{s_{h-1}\sim(P^\star,\pi_\tau) \\ (a_{h-1},a_h)\sim\mathcal{U}(\mathcal{A}) \\ s_h\sim P^\star(\cdot|s_{h-1},a_{h-1})}} \left[f_h^n(s_h,a_h)^2 \right] \leq n\zeta_n, \quad \textit{where } \zeta_n := \frac{\log\left(2|\Phi||\Psi|nH/\delta\right)}{n}.
$$

In addition, for $h = 1$,

$$
\sum_{\tau=0}^{n-1} \mathop{\mathbb{E}}_{a_1\sim\mathcal{U}(\mathcal{A})} \left[f_1^n(s_1,a_1)^2 \right] \leq n\zeta_n.
$$

Divide both sides of the result of lemma Lemma 8 by n, and define $\Pi_n = \mathcal{U}(\pi_1, \ldots, \pi_{n-1})$, we have the following corollary which will be intensively used in the analysis.

Corollary 2. *Given $\delta \in (0,1)$, the following inequality holds for any $n, h \geq 2$ with probability at least $1 - \delta/2$:*

$$\mathop{\mathbb{E}}_{\substack{s_{h-1} \sim (P^\star, \Pi_n) \\ a_{h-1}, a_h \sim \mathcal{U}(\mathcal{A}) \\ s_h \sim P^\star(\cdot | s_{h-1}, a_{h-1})}} \left[f_h^n(s_h, a_h)^2 \right] \leq \zeta_n.$$

In addition, for $h = 1$,

$$\mathop{\mathbb{E}}_{a_1 \sim \mathcal{U}(\mathcal{A})} \left[f_1^n(s_1, a_1)^2 \right] \leq \zeta_n.$$

The following lemma (Dann et al., 2017) will be useful to measure the difference between two value functions under two MDPs and reward functions. We define $P_h V_{h+1}(s_h, a_h) = \mathbb{E}_{s \sim P_h(\cdot | s_h, a_h)} \left[V(s) \right]$ for shorthand.

Lemma 9. (Simulation lemma). *Suppose P_1 and P_2 are two MDPs and r_1, r_2 are the corresponding reward functions. Given a policy π, we have,*

$$\begin{aligned}
&V^\pi_{h,P_1,r_1}(s_h) - V^\pi_{h,P_2,r_2}(s_h) \\
&\quad = \sum_{h'=h}^{H} \mathop{\mathbb{E}}_{\substack{s_{h'} \sim (P_2, \pi) \\ a_{h'} \sim \pi}} \left[r_1(s_{h'}, a_{h'}) - r_2(s_{h'}, a_{h'}) + (P_{1,h'} - P_{2,h'}) V^\pi_{h'+1,P_1,r}(s_{h'}, a_{h'}) | s_h \right] \\
&\quad = \sum_{h'=h}^{H} \mathop{\mathbb{E}}_{\substack{s_{h'} \sim (P_1, \pi) \\ a_{h'} \sim \pi}} \left[r_1(s_{h'}, a_{h'}) - r_2(s_{h'}, a_{h'}) + (P_{1,h'} - P_{2,h'}) V^\pi_{h'+1,P_2,r}(s_{h'}, a_{h'}) | s_h \right].
\end{aligned}$$

The following lemma is a standard inequality in regret analysis for linear models in reinforcement learning (see Lemma G.2 in Agarwal et al. (2020) and Lemma 10 in Uehara et al. (2021)).

Lemma 10. (Elliptical potential lemma). *Consider a sequence of $d \times d$ positive semidefinite matrices $X_1, \ldots, X_N$ with $\mathrm{tr}(X_n) \leq 1$ for all $n \in [N]$. Define $M_0 = \lambda_0 I$ and $M_n = M_{n-1} + X_n$. Then*

$$\sum_{n=1}^{N} \mathrm{tr}\left(X_n M_{n-1}^{-1} \right) \leq 2 \log \det(M_N) - 2 \log \det(M_0) \leq 2d \log \left(1 + \frac{N}{d\lambda_0} \right).$$

If we choose any subset of the set $\{X_n M_{n-1}^{-1}\}_{n=1}^N$, we can still get a sublinear summation as follows.