REPRESENTATION LEARNING FOR ONLINE AND OFFLINE RL IN LOW-RANK MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work studies the question of *Representation Learning in RL:* how can we learn a compact low-dimensional representation such that on top of the representation we can perform RL procedures such as exploration and exploitation, in a sample efficient manner. We focus on the low-rank Markov Decision Processes (MDPs) where the transition dynamics correspond to a low-rank transition matrix. Unlike prior works that assume the representation is known (e.g., linear MDPs), here we need to learn the representation for the low-rank MDP. We study both the online RL and offline RL settings. For the online setting, operating with the same computational oracles used in FLAMBE(Agarwal et al., 2020b)—-the state-of-art algorithm for learning representations in low-rank MDPs, we propose an algorithm REP-UCB—Upper Confidence Bound driven REPresentation learning for RL, which *significantly improves* the sample complexity from $\widetilde{O}(A^9d^7/(\epsilon^{10}(1-\gamma)^{22}))$ for FLAMBE to $\widetilde{O}(A^4d^4/(\epsilon^2(1-\gamma)^2))$ with d being the rank of the transition matrix (or dimension of the ground truth representation), A being the number of actions, and γ being the discount factor. Notably, REP-UCB is simpler than FLAMBE, as it directly balances the interplay between representation learning, exploration, and exploitation, while FLAMBE is an explore-then-commit style approach and has to perform reward-free exploration step-by-step forward in time. For the offline RL setting, we develop an algorithm that leverages pessimism to learn under a partial coverage condition: our algorithm is able to compete against *any policy* as long as it is covered by the offline data distribution.

1 INTRODUCTION

When applying Reinforcement Learning (RL) to large-scale problems where data is complex and high-dimensional, learning effective transformations of the data, i.e., representation learning, can often significantly improve the sample and computation efficiency of the RL procedure. Indeed, several empirical works have shown that leveraging representation learning techniques developed in supervised or unsupervised learning settings can accelerate the search for good decision-making strategies (Silver et al., 2018; Stooke et al., 2021; Srinivas et al., 2020; Yang & Nachum, 2021). However, representation learning in RL is far more subtle than it is for non-sequential and non-interactive learning tasks (e.g., supervised learning). Prior works have shown that even if one is given the magic representation that exactly linearizes the optimal policy (Du et al., 2019b) or the optimal value functions (Wang et al., 2020; Weisz et al., 2021), RL is still challenging (i.e., one may still need exponentially many samples to learn). This indicates that an effective representation that permits efficient RL needs to encode more information about the underlying Markov Decision Processes (MDPs). Despite the recent empirical success of representation learning in RL , its statistical guarantee and theoretical properties remain under-investigated.

In this work, we study the representation learning question under the low-rank MDP assumption. Concretely, a low-rank MDP assumes that the MDP transition matrix admits a low-rank factorization, i.e., there exists two *unknown* mappings $\mu(s'), \phi(s,a)$, such that $P(s'|s,a) = \mu(s')^\top \phi(s,a)$ for all s, a, s', where $P(s'|s,a)$ is the probability of transiting to the next state s' under the current state and action (s,a). The representation ϕ in a low-rank MDP not only linearizes the optimal state-action value function of the MDP (Jin et al., 2020a), but also linearizes the transition operator. A low-rankness assumption on large stochastic matrices is a common assumption and has enabled successful development of algorithms for real world applications such as movie recommendation systems

Methods	Setting	Sample Complexity	Computation
OLIVE (Jiang et al., 2017)	Low Bellman rank	$\frac{d^2A}{\epsilon^2(1-\gamma)^4}$	Inefficient
Witness rank (Sun et al., 2019)	Low Witness rank	$\frac{d^2A}{\epsilon^2(1-\gamma)^4}$	Inefficient
BLin-UCB (Du et al., 2021)	Bilinear Class	$\frac{d^2A}{\epsilon^2(1-\gamma)^7}$	Inefficient
Moffle (Modi et al., 2021)	Low-nonnegative-rank MDP	$\frac{d^6A^{13}}{\epsilon^2\eta^5(1-\gamma)^5}$	Oracle efficient
FLAMBE Agarwal et al. (2020b)	Low-rank MDP	$\frac{d^7A^9}{\epsilon^{10}(1-\gamma)^{22}}$	Oracle efficient
REP-UCB (Ours)	Low-rank MDP	$\frac{d^4A^4}{\epsilon^2(1-\gamma)^2}$	Oracle efficient

Table 1: Comparison among different provable representation learning algorithms in online RL. Algorithms such as OLIVE, Witness rank, and BLin-UCB work for settings which are more general than low-rank MDPs and have tight sample complexity. However, these algorithms are version space algorithms and thus are not computationally efficient. Moffle is an oracle-efficient algorithm (with a much stronger oracle than the one in FLAMBE and ours), but the assumptions under which Moffle operates essentially imply that the MDP's transition has low non-negative matrix rank (nnr) (see the detailed discussion in the related work section). Note that a nnr is at least as large as and could be exponentially larger than the rank (Agarwal et al., 2020b). Finally, FLAMBE operates exactly under the same setting of ours, i.e., the same function approximation and computation oracle. Our algorithm *significantly improves* the sample complexity from FLAMBE in *all parameters*. Note the horizon dependence is not exactly comparable as these prior works originally considered the finite horizon setting with nonstationary transition, and we convert their results to the discounted setting by simply replacing the finite horizon H by $\Theta(1/(1-\gamma))$.

(Koren et al., 2009). We note that a low-rank MDP strictly generalizes a linear MDP model (Yang & Wang, 2020; Jin et al., 2020a) which assumes ϕ is known a priori. The unknown representation ϕ makes learning in low-rank MDPs much more challenging than that in linear MDPs since one can no longer directly use linear function approximations. On the other hand, the fact that linear MDPs can be solved statistical and computational efficiently if ϕ is known a priori implies that if one could learn the representation of the low-rank MDP, one could then efficiently learn the optimal policy.

Indeed, prior works have shown that learning in low-rank MDPs is statistically feasible (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021) via leveraging rich function approximators. However, these algorithms are version space algorithms and are not computationally efficient. Recent work FLAMBE proposes an oracle-efficient algorithm [1] that learns in low-rank MDPs with a polynomial sample complexity, where the computation oracle is Maximum Likelihood Estimation (MLE). In this work, we follow the same setup from FLAMBE (Agarwal et al., 2020b), and propose a new algorithm — *Upper Confidence Bound driven Representation Learning, Exploration and Exploitation (*REP-UCB*)*, which can learn a near optimal policy for a low-rank MDP with a polynomial sample complexity and is oracle-efficient. Comparing to FLAMBE, our algorithm *significantly improves* the sample complexity from $O(d^7A^9/(\epsilon^{10}(1-\gamma)^{22})$ for FLAMBE to $O(d^4A^4/(\epsilon^2(1-\gamma)^2)$, where d is the rank of the transition matrix (or dimension of the true representation), A is the number of actions, ϵ is the suboptimality gap and $\gamma \in [0, 1)$ is the discount factor in the MDP. Our algorithm is also arguably much simpler than FLAMBE: FLAMBE is an explore-then-commit algorithm, has to explore in a layer-by-layer forward way, and does not permit data sharing across different time steps. In contrast, REP-UCB carefully trades exploration versus exploitation by combining the reward signal and exploration bonus (constructed using the latest learned representation), and enables data sharing across all time steps. Our sample complexity nearly matches the ones from those computationally inefficient algorithms (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021). We summarize the comparison with the prior works that study representation learning in Table 1.

In addition to the online exploration setting, we also show that our new techniques can be directly used for designing offline RL algorithms for low-rank MDPs under partial coverage. More specifically, we propose an algorithm REP-LCB—*Lower Confidence Bound driven Reprepresentation Learning for offline RL*, that given an offline dataset, can learn to compete against any policy (including history-dependent policies) as long as it is covered by the offline data where the coverage

[1] The oracle generally refers to empirical risk minimization oracle employing non-bilevel optimization. We seek to design an algorithm that runs in polynomial time, with each oracle call counting as $O(1)$. This algorithmic framework has lead to many successful algorithms in the literature of contextual bandit (Agarwal et al., 2014; Dudík et al., 2017; Foster & Rakhlin, 2020).

is measured using the relative condition number (Agarwal et al., 2021) associated with the ground truth representation. Thus, our offline RL result generalizes prior offline RL works on linear MDPs (Jin et al., 2020b; Zhang et al., 2021b) and Kernelized nonlinear regulators (KNRs) (Chang et al., 2021), which all assume representation is known a priori and use linear function approximation. Computation-wise, our approach uses one call to the MLE computation oracle, and hence is oracle-efficient. REP-LCB is *the first oracle efficient offline algorithm for low-rank MDP enjoying the aforementioned statistical guarantee.* See Section 2 for a more detailed comparison with the existing literature on representation learning in offline RL.

Our contributions. We develop new representation learning RL algorithms that enable sample efficient learning in low-rank MDPs under both online and offline settings:

1. In the online episodic learning setting, our new algorithm REP-UCB integrates representation learning, exploration, and exploitation together, and significantly improves the sample complexity of the prior state-of-art algorithm FLAMBE;
2. In the offline learning setting, we propose a natural concentrability coefficient (i.e., relative condition number under the true representation) that captures the partial coverage condition in low-rank MDP, and our algorithm REP-LCB learns to compete against any policy (including history-dependent ones) under such a partial coverage condition.

2 RELATED WORK

Online Setting We list the comparison as follows, which is summarized in Table 1.

FLAMBE (Agarwal et al., 2020b) was a state-of-the-art oracle-efficient algorithm for low-rank MDPs. In all parameters, the statistical complexity is much worse than REP-UCB . Our algorithm and FLAMBE operate under the same computation oracle. FLAMBE does not balance exploration and exploitation, and uses explore-then-committee style techniques (i.e., constructions of absorbing MDPs (Brafman & Tennenholtz, 2002)) which results in its worse sample complexity.

With a more complex oracle, Moffle (Modi et al., 2021) is a model-free algorithm for low-rank MDPs, with two additional assumptions: (1) the transition has low *non-negative rank (nnr)*, and (2) reachability in latent states. The first assumption significantly restricts the scope of low-rank MDPs as there are matrices whose nnr is exponentially larger than the rank (Agarwal et al., 2020b). The sample complexity of Moffle can scale $O(d^6|\mathcal{A}|^{13}/(\epsilon^2\eta^5(1-\gamma)^5))$, where η is the reachability probability, and $1/\eta$ could be as large as the nnr of the transition matrix.

OLIVE (Jiang et al., 2017), Witness rank (Sun et al., 2019) and Bilinear-UCB (Du et al., 2021), when specialized to low-rank MDPs, have slightly tighter dependence on d (e.g., $O(d^2/\epsilon^2)$). But these algorithms are computationally inefficient as they are version space algorithms. Dann et al. (2021) shows that with a policy class, solving a low-rank MDP can take $\Omega(2^d)$ samples. In this work, similar to Witness rank (Sun et al., 2019) and FLAMBE, we use function approximators to model the transition. Thus our positive result is not in contradiction to the result from (Dann et al., 2021).

VALOR (Dann et al., 2018), PCID (Du et al., 2019a), HOMER (Misra et al., 2020), RegRL (Foster et al., 2020), and the approach from Feng et al. (2020) are algorithms for block MDPs which is a more restricted setting than low-rank MDPs. These works require additional assumptions such as deterministic transitions (Dann et al., 2018), reachability (Misra et al., 2020; Du et al., 2019a), Bellman completion (Foster et al., 2020), and strong unsupervised learning oracles (Feng et al., 2020).

Finally, we mention works that tackle representation learning in quite different settings. Zhang et al. (2021a) consider how to choose the best representation among *correct representations* inspired by Papini et al. (2021) (i.e., the MDP is a linear MDP under any representation in the function class Φ). In contrast, we only assume the MDP is linear under *some* $\phi^\star \in \Phi$. Hao et al. (2021b) consider feature selection in sparse linear MDPs which a given exploratory distribution.

Offline Setting We discuss related works in offline RL.

Uehara & Sun (2021) obtained similar results on low-rank MDPs. Though the sample complexity in their algorithm is slightly tighter, our algorithm is oracle-efficient, while the CPPO algorithm from Uehara & Sun (2021) is a version space algorithm.

Xie et al. (2021) propose a (general) pessimistic model-free algorithm in the offline setting. We can also apply their algorithm to low-rank MDPs and show some finite-sample guarantee. However, it is

unclear whether the final bounds in their results can be characterized by the relative condition number only using the true representation, and whether they can compete with history-dependent policies. Thus, our result is still considered superior on low-rank MDPs. The detail is given in Section D.

In addition to the above two works, the pessimistic approach in offline RL has been extensively investigated. Empirically, it works well on complex control tasks (Kidambi et al., 2020; Yu et al., 2020; Kumar et al., 2020; Liu et al., 2020; Chang et al., 2021). On the theoretical side, pessimism allows us to obtain the PAC guarantee on various models when a comparator policy is covered by offline data in some forms (Jin et al., 2020b; Rashidinejad et al., 2021; Yin et al., 2021; Zanette et al., 2021b; Zhang et al., 2021b; Chang et al., 2021). However, these algorithms and their analysis rely on a *known representation*. We emphasize the differences in Section 5.

Finally, we mention works that tackle representation learning from different viewpoints and settings. Lu et al. (2021) consider multitask representation learning under a generative model assumption. Hao et al. (2021a) study the feature selection problem in sparse linear MDPs under the assumption that the offline data admits some form of full coverage condition.

3 PRELIMINARIES

We consider a Markov Decision Process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma, d_0 \rangle$ specified by a state space $\mathcal{S}$, a discrete action space $\mathcal{A}$, a transition model $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a discount factor $\gamma \in [0, 1)$, and an initial distribution $d_0 \in \Delta(\mathcal{S})$. To simplify the presentation, we assume $r(s, a)$ and d_0 are known (e.g., when d_0 is a probability mass only on s_0, agent always starts from a fixed initial state s_0)[2]. Following prior work (Jiang et al., 2017; Sun et al., 2019), we assume *trajectory reward is normalized*, i.e., for any trajectory $\{s_h, a_h\}_{h=0}^{\infty}$, we have $\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \in [0, 1]$. Since the ground truth $P^\star$ is unknown, we need to learn it by interacting with environments in an online manner or utilizing offline data at hand. We remark that the extension of our all results to the finite horizon nonstationary case is straightforward.

We use the following notation. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which maps from state to distribution over actions and some model P, we define the value function $V_P^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | s_0 = s, P, \pi\right]$ to represent the expected total discounted reward of π under P starting at s. Similarly, we define the state-action Q function $Q_P^\pi(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V_P^\pi(s')$. The expected total discounted reward of a policy π is denoted as $V_{P,r}^\pi := \mathbb{E}_{s_0 \sim d_0} V_P^\pi(s_0)$. We define the discounted state-action occupancy distribution $d_P^\pi(s, a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t d_{P,t}^\pi(s, a)$, where $d_{P,t}^\pi(s, a)$ is the probability of π visiting (s, a) at time step t under π and P. We slightly abuse the notation, and denote $d_P^\pi(s)$ as the state visitation, which is equal to $\sum_{a \in \mathcal{A}} d_P^\pi(s, a)$. When P is the ground truth transition model $P^\star$, we drop the subscript and simply use d^π to denote its state-action distribution. Unless otherwise noted, Π denotes the class of all polices $\{S \to \Delta(\mathcal{A})\}$. We denote total variation distance of P_1 and P_2 by $\|P_1 - P_2\|_1$. Finally, given a vector a, we define $\|a\|_2 = \sqrt{a^\top a}, \|a\|_B = \sqrt{a^\top B a}$. $c_0, c_1, \cdots$ are universal constants.

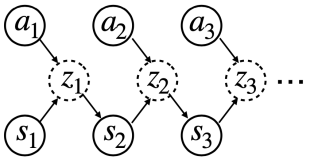

Figure 1: A latent state model captured by low-rank MDP. Here $\phi^\star(s, a)$ is a distribution over a discrete latent state space $\mathcal{Z}$. Note that this is still a Markovian model as there is no transition among latent states.

We study low-rank MDPs defined as follows (Jiang et al., 2017; Agarwal et al., 2020b). The conditions on the upper bounds of the norm of $\phi^\star, \mu^\star$ are just for normalization.

Definition 1 (Low-rank MDP)**.** *Given function classes $\Phi \subset \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\Psi \subset \mathcal{S} \to \mathbb{R}^d$, a transition model $P^\star : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ admits a low rank decomposition with dimension $d \in \mathbb{N}$ if there exists two embedding functions $\phi^\star \in \Phi, \mu^\star \in \Psi$ such that*

$$\forall s, s' \in \mathcal{S}, a \in \mathcal{A} : P^\star(s' \mid s, a) = \mu^\star(s')^\top \phi^\star(s, a)$$

where $\|\phi^*(s, a)\|_2 \leq 1$ for all (s, a) and for any function $g : \mathcal{S} \to [0, 1]$, $\| \int \mu^\star(s) g(s) \mathrm{d}(s) \|_2 \leq \sqrt{d}$. An MDP is a low rank MDP if $P^\star$ admits such a low rank decomposition.

[2]Extension to the unknown case is straightforward. Recall the major challenging of RL is due to the unknown transition model.

Low-rank MDPs capture the latent variable model shown in Figure 1 where $\phi^\star(s,a)$ is a distribution over a discrete latent state space $\mathcal{Z}$. The block-MDP model (Du et al., 2019a) is a special instance of the latent variable model with $\phi^\star(s,a)$ being a one hot encoding vector. Note the linear MDPs (Yang & Wang, 2020; Jin et al., 2020a) assume $\phi^\star$ is known.

Next, we explain two settings: the online learning setting and the offline learning setting. Then, we present our function approximation setup and computational oracles.

Episodic Online learning In online learning, our overall goal is to learn a stationary policy $\hat{\pi}$ so that it maximizes $V^{\hat{\pi}}_{P^\star,r}$, where $P^\star$ is the ground truth transition. We assume that we operate under the episodic learning setting where we can only reset to states sampled from the initial distribution d_0 (e.g., to emphasize the challenge from exploration, we can consider the special case where we can only reset to a fixed s_0). In the episodic setting, given a policy π, sampling a state s from the state visitation d^π_P is done by the following roll-in procedure: starting at $s_0 \sim d_0$, at every time step t, we terminate and return s_t with probability $1-\gamma$, and otherwise we execute $a_t \sim \pi(s_t)$ and move to $t+1$, i.e., $s_{t+1} \sim P(\cdot|s_t,a_t)$. Such a sampling procedure is widely used in the episodic learning setting for discounted infinite horizon MDPs (e.g., (Kakade & Langford, 2002; Agarwal et al., 2021; 2020a)).

Offline learning In the offline RL, we are given a static dataset in the form of quadruples:

$$\mathcal{D} = \{s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}\}_{i=1}^n \sim \rho(s,a)\delta(r = r(s,a))P^\star(s' \mid s,a).$$

For simplicity, we assume $\rho = d^{\pi_b}_{P^\star}$, where $\pi_b \in \mathcal{S} \to \Delta(\mathcal{A})$ is a fixed behavior policy. We denote $\mathbb{E}_{\mathcal{D}}[f(s,a,s')] = 1/n \sum_{(s,a,s')\in\mathcal{D}} f(s,a,s')$. To succeed in offline RL, we in general need some coverage property of ρ. One common assumption is that ρ globally covers every possible policies' state-action distribution, i.e., $\max_{\pi,s,a} d^\pi_{P^\star}(s,a)/\rho(s,a) < \infty$ (Antos et al., 2008). In this work, we relax such a global coverage assumption and work under the *partial* coverage condition where ρ may not cover distributions of all possible policies. Instead of competing against the optimal policy under the global coverage, we aim to compete against any policies covered by the offline data. In section 5, we will precisely define the partial coverage condition using the concept of the relative condition number (Agarwal et al., 2021).

Function approximation setup and computational oracles Since $\mu^\star$ and $\phi^\star$ are unknown, we use function classes to capture them. Our function approximation and computational oracles are *exactly the same as the ones used in* FLAMBE. For completeness, we state the function approximation and computational oracles below.

Assumption 2. *We have a model class $\mathcal{M} = \{(\mu,\phi) : \mu \in \Gamma, \phi \in \Phi\}$, where $\mu^\star \in \Gamma$, $\phi^\star \in \Phi$.*

Following the norm bounds on $\mu^\star, \phi^\star$ we simply assume that the same norm bounds hold for our function approximator, i.e., for any $\mu \in \Gamma, \phi \in \Phi$, $\|\phi(s,a)\|_2 \leq 1$, $\forall(s,a)$ and $\|\int \mu(s)g(s)\mathrm{d}(s)\|_2 \leq \sqrt{d}, \forall g : \mathcal{S} \to [0,1]$, and $\int \mu^\top(s')\phi(s,a)\mathrm{d}(s') = 1$, $\forall(s,a)$.

As for computational oracles, we use a supervised learning style MLE oracle.

Definition 3 (Maximum Likelihood Oracle (MLE))**.** *Consider the model class $\mathcal{M}$ and a dataset $\mathcal{D}$ in the form of (s,a,s'), the MLE oracle returns the maixmum likelihood estimator $\hat{P} := (\hat{\mu},\hat{\phi}) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}} \ln(\mu(s')^\top\phi(s,a))$.*

We also invoke a planning procedure for *known linear MDPs with potentially nonlinear rewards*, which can be done in polynomial time (after all, we know that online learning in linear MDPs can be done statistically and computationally efficient). Given a reward r and a model $P := (\mu,\phi)$ with $P(s'|s,a) = \mu(s')^\top\phi(s,a)$ (i.e., a *known linear transition* with a *known* feature ϕ), we can compute the optimal policy $\arg\max_\pi V^\pi_P$ by standard least square value iteration which uses linear regression. A planning procedure for a known linear MDP is also used in FLAMBE, see Section 5.1 in Agarwal et al. (2020b) for details of how to implement this procedure for a known linear MDP in polynomial computation complexity.

4 REPRESENTATION LEARNING IN ONLINE SETTING

We consider the online episodic learning setting where the agent can only reset based on the initial state distribution d_0. Thus, to find a near-optimal policy for a low-rank MDP efficiently, we need to carefully interleave representation learning, exploration, and exploitation.

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

1: **Input:** Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N
2: Initialize $\pi_0(a \mid s)$ to be uniform; $\mathcal{D}_0 = \emptyset$
3: **for** episode $n = 1, \cdots, N$ **do**
4: Collect a transition (s, a, s') where $s \sim d_{P^\star}^{\pi_{n-1}}$, $a \sim U(\mathcal{A})$, $s' \sim P^\star(\cdot|s, a)$
5: Add (s, a, s') to the dataset: $\mathcal{D}_n = \mathcal{D}_{n-1} + \{s, a, s'\}$
6: Learn representation via MLE: $\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_n}\left[\ln \mu^\top(s')\phi(s, a)\right]$
7: Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a\in\mathcal{D}_n} \hat{\phi}_n(s, a)\hat{\phi}_n(s, a)^\top + \lambda_n I$
8: Set the exploration bonus:

$$\hat{b}_n(s, a) := \min\left(\alpha_n \sqrt{\hat{\phi}_n(s, a)^\top \hat{\Sigma}_n^{-1} \hat{\phi}_n(s, a)}, 2\right) \tag{1}$$

9: Update policy $\pi_n = \arg\max_\pi V^\pi_{\hat{P}_n, r+\hat{b}_n}$
10: **end for**
11: **Return** $\pi_1, \cdots, \pi_N$

4.1 ALGORITHM

We present our proposed algorithm in the online setting. This is described in Algorithm 1. Every iteration, Algorithm 1 rollouts its current policy π to collect a triple (s, a, s') where $s \sim d^\pi_{P^\star}$, $a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s, a)$ where $U(\mathcal{A})$ is a uniform distribution over actions (see Section 3 in terms of how to get a sample from $d^\pi_{P^\star}$). After collecting new data and concatenating it with the existing data, we perform representation learning, i,e, learning a factorization and a representation by MLE (line 6), set the bonus based on the learned feature (Eq. 1), and update the policy via planning inside the learned model with the bonus-enhanced reward (Line 9). Note the learned transition $\hat{P}$ from MLE is linear with the learned feature $\hat{\phi}$, and planning in a known linear MDP is known as computationally efficient (Jin et al., 2020a) (see the explanation after Definition 3 as well). We emphasize that our computation MLE oracle is the same as the one used in FLAMBE (Agarwal et al., 2020b).

4.2 ANALYSIS

We show that our algorithm has the following PAC bound.

Theorem 4 (PAC Bound for REP-UCB)**.** *Fix $\delta \in (0, 1), \epsilon \in (0, 1)$. Let $\hat{\pi}$ be a random output of $\pi_1, \cdots, \pi_N$ and $\pi^* = \arg\max_\pi V^\pi_{P^\star, r}$. By setting parameters as follows:*

$$\alpha_n = O\left(\sqrt{(|\mathcal{A}| + d^2)\,\gamma \ln(|\mathcal{M}|n/\delta)}\right), \quad \lambda_n = O\left(d \ln(|\mathcal{M}|n/\delta)\right),$$

with probability at least $1 - \delta$, we have

$$V^{\pi^\star}_{P^\star, r} - V^{\hat{\pi}}_{P^\star, r} \le \epsilon,$$

where the number of collected trajectories is at most

$$N \log(N/\delta), \quad N := O\left(\frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2} \cdot \ln^2\left(1 + \frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2}\right)\right).$$

The theorem shows that REP-UCB learns in low-rank MDPs in a statistically efficient and oracle-efficient manner. To the best of our knowledge, this algorithm has the best sample complexity among all oracle efficient algorithms for low-rank MDPs.

Highlight of the analysis Below we highlight our key lemmas and proof techniques.

First, why learning in a low-rank MDP is harder than learning in models with linear structures? Unlike standard linear models such as linear MDPs (Yang & Wang, 2020; Jin et al., 2020a), KNRs (Kakade et al., 2020; Abbasi-Yadkori & Szepesvári, 2011; Mania et al., 2020; Song & Sun, 2021), and GP / kernel models (Chowdhury & Gopalan, 2019; Curi et al., 2020), we cannot get uncertainty quantification on the model in a point-wise manner. When models are linear, one can get the following style of point-wise uncertainty quantification for the learned model $\hat{P}$: $\forall s, a : \ell(\hat{P}(\cdot|s, a), P^\star(\cdot|s, a)) \le \sigma(s, a)$ where $\sigma(s, a)$ is the uncertainty measure, and ℓ is some distance metric (e.g., ℓ_1 norm). With proper scaling, the uncertainty measure $\sigma(s, a)$ is then used

for the bonus. For example, in linear MDPs, given a dataset $\mathcal{D} = \{s, a, s'\}$, we can learn a model $\hat{P}(s'|s,a) := \hat{\mu}(s')^\top \phi^\star(s,a)$, and get point-wise uncertainty quantification:

$$\forall (s,a), |\int f(s')\hat{\mu}^\top(s')\phi^\star(s,a)\mathrm{d}(s') - \int f(s')\mu^{\star\top}(s')\phi^\star(s,a)\mathrm{d}(s')| \leq c\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\phi^\star}} \quad (2)$$

for some family of functions $f : \mathcal{S} \to \mathbb{R}$ with $\Sigma_{\phi^\star} = \sum_{s,a\in\mathcal{D}} \phi^\star(s,a)\phi^\star(s,a) + \lambda I$ (Lykouris et al., 2021; Neu & Pike-Burke, 2020). To set the scaling c properly, since $\phi^\star$ is known a priori, the linear regression analysis applies here, and one can apply Cauchy-Schwarz inequality to the LHS of (2) to pull out $\phi^\star$ and get an upper bound in the form of

$$\underbrace{\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\phi^\star}}}_{(a)} \underbrace{\|\int f(s)\{\hat{\mu}(s) - \mu^\star(s)\}\mathrm{d}(s)\|_{\Sigma_{\phi^\star}}}_{(b)}$$

where c is set to be the linear regression training error measured in the term (b) above.

However, when we jointly learn μ and ϕ, since nonlinear function approximation is used[3], we cannot get point-wise uncertainty quantification via linear regression-based analysis. We stress that our bonus is not designed to capture the uncertainty quantification on the model error between $\hat{P}(\cdot|s,a) = \hat{\mu}^\top \hat{\phi}(s,a)$ and $P^\star(\cdot|s,a) = \mu^{\star\top}\phi^\star(s,a)$ in a point-wise way, which is not tractable as $\hat{P}$ and $P^\star$ does not even share the same representation. Instead, the bonus is carefully designed so that it only provides *near-optimism at the initial state distribution*. This is formalized as follows.

Lemma 5 (Almost Optimism at the Initial State Distribution)**.** *Set the parameters as in Theorem 4. With probability* $1-\delta$*,*

$$\forall n \in [1, \cdots, N], \forall \pi \in \Pi, V^\pi_{\hat{P}_n, r+\hat{b}_n} - V^\pi_{P^\star, r} \geq -c_1\sqrt{\frac{|\mathcal{A}|\ln(|\mathcal{M}|n/\delta)(1-\gamma)}{n}}.$$

We remark that the idea of optimism with respect to the initial state distribution has been used in prior works (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Zanette et al., 2020). However, these algorithms are not computationally efficient (i.e., they use version space instead of reward bonus), and their version-space based analysis is different from ours.

Proof sketch for Lemma 5 We start by using the simulation lemma (Lemma 20) *inside the learned model* which is important since our bonus $\hat{b}_n$ uses $\hat{\phi}_n$ associated with the learned model $\hat{P}_n$:

$$\begin{aligned} V^\pi_{\hat{P}_n, r+\hat{b}_n} - V^\pi_{P^\star, r} &= \mathbb{E}_{s,a\sim d^\pi_{\hat{P}_n}}\left[\hat{b}_n(s,a) + \mathbb{E}_{s'\sim\hat{P}_n(\cdot|s,a)}V^\pi_{P^\star}(s') - \mathbb{E}_{s'\sim P^\star(\cdot|s,a)}V^\pi_{P^\star}(s')\right] \\ &\geq \mathbb{E}_{s,a\sim d^\pi_{\hat{P}_n}}\left[\hat{b}_n(s,a) - \|\hat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1\right], \end{aligned}$$

from where we show that $\mathbb{E}_{s,a\sim d^\pi_{\hat{P}_n}}\|\hat{P}(\cdot|s,a) - P^\star(\cdot|s,a)\|_1$ *as a whole* nearly lower bounds $\mathbb{E}_{s,a\sim d^\pi_{\hat{P}_n}}[\hat{b}_n(s,a)]$. The detailed procedure is illustrated in Lemma 8 in Appendix A.

Second, our bonus is using representation $\hat{\phi}_n$ that is being updated every round, and our empirical covariance matrix $\hat{\Sigma}_n$ is also updated whenever we update $\hat{\phi}_n$, which means that standard elliptical potential based analysis (i.e., analysis used in linear bandits/MDPs with known features) cannot work here as our feature changes every iteration. Instead, in our analysis, we have to keep tracking a potential function that is defined using the unknown ground truth representation $\phi^\star$, i.e., the elliptical potential $\|\phi^\star(s,a)\|^2_{\Sigma^{-1}_{\rho_n,\phi^\star}}$, where $\Sigma_{\rho_n,\phi^\star} = n\mathbb{E}_{(s,a)\sim\rho_n}\phi^\star(s,a)\phi^\star(s,a)^\top + \lambda_n I$ and ρ_n is the marginal state-action distribution of $\mathcal{D}_n$ at an iteration n, i.e., $\rho_n(s,a) = \sum_{i=0}^{n-1} d^{\pi_i}_{P^\star}(s)U(a)/n$. Since this potential function uses the fixed representation $\phi^\star$, we can apply the standard elliptical potential argument to track the progress that our algorithm makes during learning. Below we illustrate the procedure of linking the bonus under $\hat{\phi}_n$ to the potential function $\|\phi^\star(s,a)\|^2_{\Sigma^{-1}_{\rho_n,\phi^\star}}$ defined with respect to the true feature $\phi^\star$.

[3]There are nonlinear models using Eluder dimension (Russo & Van Roy, 2014) as the complexity measure. However, the only known examples that admit low Eluder dimension are (generalized) linear models.

Algorithm 2 LCB-driven Representation Learning in offline RL (REP-LCB)

1: **Input:** Regularizer λ, Model classes $\mathcal{M} = \{\mu^\top \phi : \mu \in \Psi, \phi \in \Phi\}$, Dataset $\mathcal{D}$.
2: Learn a model $\hat{P}$ by MLE: $\hat{P} = \hat{\mu}^\top \hat{\phi} = \arg\max_{P \in \mathcal{M}} \mathbb{E}_{\mathcal{D}}[\ln P(s' \mid s, a)]$.
3: Set the empirical covariance matrix $\hat{\Sigma} = \sum_{(s,a)\in\mathcal{D}} \hat{\phi}(s,a)\hat{\phi}^\top(s,a) + \lambda I$.
4: Set the reward penalty:
$$\hat{b}(s,a) = \min\left(\alpha\sqrt{\hat{\phi}(s,a)^\top \hat{\Sigma}^{-1}\hat{\phi}(s,a)}, 2\right).$$
5: Solve $\hat{\pi} = \arg\max_\pi V^\pi_{\hat{P}, r-\hat{b}}$.

Linking bonus under $\hat{\phi}_n$ to the elliptical potential function under $\phi^\star$ With near optimism, using the simulation lemma (Lemma 20), we can upper bound the per-iteration regret as follows:

$$V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \leq |\mathcal{A}| \mathbb{E}_{a\sim\mathcal{U}(\mathcal{A}), s\sim d^{\pi_n}_{P^\star}}\left[\hat{b}_n(s,a) + \tfrac{1}{(1-\gamma)} f_n(s,a)\right],$$

where $f_n(s,a) := \|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$. To connect the right-hand side to the elliptical potential under the fixed feature $\phi^\star$, we show that for any function $g \in \mathcal{S}\times\mathcal{A} \to [0, B]$ for $B \in \mathbb{R}^+$,

$$\mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[g(s,a)] \leq \mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}\left[\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\right]\sqrt{n\gamma|\mathcal{A}|\mathbb{E}_{\rho_n}[g^2(s,a)] + \gamma\lambda_n d B^2} + \sqrt{(1-\gamma)|\mathcal{A}|\mathbb{E}_{\rho_n} g^2(s,a)}.$$

See Lemma 13 in Appendix A. By substituting g with $\hat{b}_n + f_n/(1-\gamma)$, the first term of the RHS of the above inequality can be upper bounded as:

$$\underbrace{2\,\mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}]}_{(\mathcal{G}_1)}\underbrace{\sqrt{n|\mathcal{A}|\mathbb{E}_{\rho_n}\left[\frac{f_n^2(s,a)}{(1-\gamma)^2} + \hat{b}_n^2(s,a)\right] + \lambda_n d}}_{(\mathcal{G}_2)}.$$

In the term $(\mathcal{G}_2)$, we expect $n\mathbb{E}_{\rho_n}[f_n^2(s,a)]$ to be $O(1)$ as $\mathbb{E}_{s,a\sim\rho_n}[f_n^2(s,a)]$ is in order of $1/n$ since it is the generalization bound of the MLE estimator $\hat{P}_n$ under the training distribution ρ_n. For $n\mathbb{E}_{\rho_n}[\hat{b}_n^2(s,a)]$, we expect it to be in the order of d as the (unnormalized) data covariance matrix $\hat{\Sigma}_n$ in the bonus $\hat{b}_n$ uses training data from ρ_n, i.e., we are measuring the expected bonus under the training distribution. In other words, the term $(\mathcal{G}_2)$ scales in order of poly(d). For the term $(\mathcal{G}_1)$, since it contains the potential function based on $\phi^\star$, the sum of the term $(\mathcal{G}_1)$ over all episodes can be controlled by the standard elliptical potential argument (see Lemma 18 and Lemma 19).

5 REPRESENTATION LEARNING IN OFFLINE SETTING

In this section, we study representation learning in the offline setting. We consider the setting where the offline data does not have full global coverage. We present our algorithm Low Confidence Bound driven Representation Learning in offline RL (REP-LCB) in Algorithm 2. Our proposed algorithm consists of three parts. The first part is MLE which learns a representation. The second part is the construction of a penalty term $\hat{b}$. Using the learned representation $\hat{\phi}$, we use a standard bonus in linear bandits as the penalty term as if $\hat{\phi}$ were the true feature. The third part is planning with the learned model $\hat{P}$ and reward $r - \hat{b}$. Again this is just planning inside a known linear MDP.

5.1 ANALYSIS

We present the PAC guarantee of REP-LCB. Before proceeding, we define a relative condition number as a mean to measure the deviation between a comparator policy π and the offline data:

$$C^\star_\pi = \sup_{x\in\mathbb{R}^d} \frac{x^\top \mathbb{E}_{d^\pi_{P^\star}}[\phi^\star(s,a)\phi^{\star\top}(s,a)]x}{x^\top \mathbb{E}_\rho[\phi^\star(s,a)\phi^{\star\top}(s,a)]x}.$$

In the special case where the MDP is just a tabular MDP, this is reduced to a density ratio $C^\star_\infty = \max_{s,a} d^\pi_{P^\star}(s,a)/\rho(s,a)$. The relative condition number $C^\star_\pi$ is always no larger than the density

ratio and could be much smaller for MDPs with large state spaces. Note that we quantify the relative condition number using the unknown true representation $\phi^\star$. With the above setup, now we are ready to state the main theorem for REP-LCB.

Theorem 6 (PAC Bound for REP-LCB)**.** *Let* $\omega = \max_{a,s}(1/\pi_b(a \mid s))$. *Denote* $\hat{\pi}$ *as the output of* REP-LCB. *There exists a set of parameters such that with probability at least* $1-\delta$*, for* any *policy* π *(including history-dependent non-Markovian policies),*

$$V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \leq c\sqrt{\frac{d^4\omega^2 C^\star_\pi \log(|\mathcal{M}|/\delta)}{(1-\gamma)^2 n}}.$$

We explain several implications. First of all, this theorem shows that we can uniformly compete with any policy including history-dependent non-Markovian policies [4] satisfying the partial coverage $C^\star_\pi < \infty$. Particularly, if the optimal policy $\pi^\star$ is covered by the offline data, i.e., $C^\star_{\pi^\star} < \infty$, then our algorithm is able to compete against it [5]. Note that assuming offline data covers $\pi^\star$ is still a weaker assumption than the global coverage such as $\sup_\pi \sup_{(s,a)} d^\pi_{P^\star}(s,a)/\rho(s,a)$ in prior offline RL works (Antos et al., 2008; Chen & Jiang, 2019) Second, our coverage condition is measured by a relative condition number defined using the unknown ground truth representation $\phi^\star$ but not depending on other features. Prior works that use relative condition numbers as measures of coverage are restricted to the settings where the ground truth representation $\phi^\star$ is known (Jin et al., 2020b; Zhang et al., 2021b; Chang et al., 2021; Zanette et al., 2021b).

To sum up, our algorithm is the first oracle efficient algorithm which does not need to know $\phi^\star$, and requires partial coverage only in terms of $\phi^\star$. Note while Uehara & Sun (2021) has a similar guarantee on low-rank MDPs, their algorithm is not oracle-efficient as it is a version space algorithm.

Highlight of the analysis Prior offline RL works that use reward penalties (Rashidinejad et al., 2021; Zhang et al., 2021b; Chang et al., 2021) all assume the representation $\phi^\star$ is known a priori, which allows them to use linear regression analysis to derive point-wise model uncertainty quantification which in turn serves as a penalty. In contrast, our function class is nonlinear. As in the online setting, what we can show is near pessimism in the initial state distribution d_0.

Lemma 7 (Almost Pessimism at the Initial State Distribution)**.** *There exists a set of parameters that with probability* $1-\delta$*, we have*

$$\forall \pi \in \Pi : V^{\pi}_{\hat{P},r-b} - V^{\pi}_{P^\star,r} \leq c_1\sqrt{\frac{\omega \log(|\mathcal{M}|/\delta)(1-\gamma)}{n}}.$$

We leave the detailed proof to Lemma 14 in the Appendix B, which is similar to the proof of Lemma 5. Though the pessimism in the initial state distribution was recently also used in Xie et al. (2021); Zanette et al. (2021b), the derivation is totally different.

Similar to the online setting, to obtain a result only depending on the relative condition number using the ground truth representation $\phi^\star$ but not the learned feature $\hat{\phi}$, we again need to translate the penalty defined with $\hat{\phi}$ to the potential function $\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho,\phi^\star}}$. Similarly, the techniques that we developed for the online setting can be leveraged here to achieve the above-mentioned goal.

6 CONCLUSION

We study online/offline RL on low-rank MDPs, where the ground truth feature is not known a priori. For online RL, our new algorithm REP-UCB significantly improves the same complexity of the piror state-of-the-art algorithm FLAMBE in all parameters while using the same computational oracles. REP-UCB has the best sample complexity among existing oracle efficient algorithms for low-rank MDPs by a margin. For offline RL, our new algorithm REP-LCB is the first oracle efficient algorithm that has a PAC guarantee under a partial coverage condition measured by the relative condition number defined with the true feature representation.

[4] Given $\pi = \{\pi_i\}_{i=0}^\infty$ where π_i depends on $s_0, a_0, \ldots s_i$, $V^\pi_{P^\star,r}$ and $d^\pi_{P^\star}(s,a)$ are still well-defined.

[5] We also require $\omega < \infty$, which is a mild assumption since it does not involve $P^\star$. Indeed, it is much weaker than the global coverage type assumption $1/\rho(s,a) < \infty, \forall (s,a)$.

REFERENCES

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26. JMLR Workshop and Conference Proceedings, 2011.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *NeurIPS*, 2020a.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020b.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Jonathan D Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data without great coverage. 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1042–1051, 2019.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3197–3205. PMLR, 2019.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *arXiv preprint arXiv:2006.08684*, 2020.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *arXiv preprint arXiv:1803.00606*, 2018.

Christoph Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Agnostic reinforcement learning with low-rank mdps and rich observations. *arXiv preprint arXiv:2106.11519*, 2021.

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, 2019a.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019b.

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. *ICML*, 2021.

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. In *2017 ieee 58th annual symposium on foundations of computer science (focs)*, pp. 528–539. IEEE, 2017.

Fei Feng, Ruosong Wang, Wotao Yin, Simon S Du, and Lin F Yang. Provably efficient exploration for reinforcement learning using unsupervised learning. *arXiv preprint arXiv:2003.06898*, 2020.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pp. 4063–4073. PMLR, 2021a.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 316–324. PMLR, 2021b.

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*, 2017.

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, 2020a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020b.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15312–15325, 2020.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1264–1274, 2020.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

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.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. *arXiv preprint arXiv:2104.03781*, 2021.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Yuda Song and Wen Sun. Pc-mlp: Model-based reinforcement learning with policy cover guided exploration. In *International Conference on Machine Learning*, pp. 9801–9811. PMLR, 2021.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pp. 9870–9879. PMLR, 2021.

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 and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10746–10756, 2020.

Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making. *arXiv preprint arXiv:2102.05815*, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021.

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

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, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021a.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021b.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935*, 2021a.

Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*, 2021b.

A PROOF OF THE THEORETICAL PROPERTY OF REP-UCB

Hereafter, we assume $c_0, c_1, \cdots$, are some universal constants. Hereafter, the notation

$$f(1/(1-\gamma), |\mathcal{A}|, \ln(1/\delta), \ln(|\mathcal{M}|), d, n) \lesssim g(1/(1-\gamma), |\mathcal{A}|, \ln(1/\delta), \ln(|\mathcal{M}|), d, n)$$

means there exists a some large constant $c_1 > 0$ satisfying

$$f(1/(1-\gamma), |\mathcal{A}|, \ln(1/\delta), \ln(|\mathcal{M}|), d, n) \leq c_1 g(1/(1-\gamma), |\mathcal{A}|, \ln(1/\delta), \ln(|\mathcal{M}|), d, n)$$

for any $0 \leq \gamma < 1, |\mathcal{A}|, \ln(1/\delta), \ln(|\mathcal{M}|), d, n$.

We first prove a lemma to show the optimism. Hereafter, we use the following notation

$$d^{\bar{\pi}}(s,a) \coloneqq d^{\pi}_{P^\star}(s)u(a).$$

where $u(a)$ is a uniform distribution over actions.

First, we prove the optimism at the initial distribution. This is proved by using a simulation lemma inside the learned model.

Lemma 8 (Almost Optimism at the Initial Distribution)**.** *Consider an episode $n (1 \leq n \leq N)$ and set*

$$\alpha_n = O(\sqrt{(|\mathcal{A}| + d^2)\,\gamma \ln(|\mathcal{M}|n/\delta)}), \quad \lambda_n = O\left(d \ln(|\mathcal{M}|n/\delta)\right), \zeta_n = O\left(\frac{\ln(|\mathcal{M}|n/\delta)}{n}\right).$$

With probability $1-\delta$, we have

$$\forall n \in [1, \cdots, N], \forall \pi \in \Pi, V^{\pi}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi}_{P^\star, r} \geq -\sqrt{(1-\gamma)|\mathcal{A}|\zeta_n}.$$

Proof. We define

$$\rho_n = \frac{1}{n}\sum_{j=0}^{n} d^{\bar{\pi}_j}(s,a),\ \Sigma_{\rho_n,\phi} = n\mathbb{E}_{(s,a)\sim\rho_n}[\phi\phi^\top] + \lambda_n I.$$

In this proof, letting $f_n(s,a) = \|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$, we condition on the event

$$\forall n, \quad \mathbb{E}_{(s,a)\sim\rho_n}[f_n^2(s,a)] \leq \zeta_n, \quad \forall n, \forall \phi, \|\phi(s,a)\|_{\hat{\Sigma}_{n,\phi}} = \Theta(\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi}}),$$

where

$$\hat{\Sigma}_{n,\phi} = n\mathbb{E}_{(s,a)\sim\mathcal{D}_n}[\phi\phi^\top] + \lambda_n I.$$

From Lemma 11 and Lemma 17, this event happens with probability $1-\delta$. Then, for any policy π, from simulation lemma 20,

$$\begin{aligned}
&V^{\pi}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi}_{P^\star, r} \\
&= \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left[\hat{b}_n(s,a) + \gamma\mathbb{E}_{s'\sim\hat{P}_n(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \gamma\mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right] \\
&\geq \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left[\min\left(\alpha_n\|\hat{\phi}_n(s,a)\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}, 2\right) + \gamma\mathbb{E}_{s'\sim\hat{P}_n(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \gamma\mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right].
\end{aligned} \tag{3}$$

Here, since $\|V^{\pi}_{P^\star,r}\|_\infty \leq 1$, we have:

$$\left|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left\{\mathbb{E}_{s'\sim\hat{P}_n(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right\}\right| \leq \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left\{f_n(s,a)\right\}.$$

This is further bounded by Lemma 12:

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left\{f_n(s,a)\right\} \lesssim \sqrt{\alpha'_n}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}}\|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)}. \tag{4}$$

where

$$\alpha'_n = \gamma\{n|\mathcal{A}|\zeta_n + \lambda_n d + n\zeta_n\} \lesssim \gamma\left(|\mathcal{A}| + d^2\right)\ln(|\mathcal{M}|n/\delta).$$

Note we here use $f_n(s,a) \le 2$ and $\mathbb{E}_{\rho_n}[f_n(s,a)^2] \le \zeta_n$.

Combining all things together,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{\mathbb{E}_{s'\sim \hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right\}\right| \le 2\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left\{f_n(s,a)\right\} \\
&\lesssim \sqrt{\alpha'_n}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}}\|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} + \sqrt{(1-\gamma)|\mathcal{A}|\zeta_n} \\
&\le \alpha_n \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}}\|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} + \sqrt{(1-\gamma)|\mathcal{A}|\zeta_n}, \quad \text{where } \alpha_n := \sqrt{\alpha'_n}. \qquad (5)
\end{aligned}$$

Going back to (3), we have

$$\begin{aligned}
&V^{\pi}_{\hat{P}_n,r+\hat{b}_n} - V^{\pi}_{P^\star,r} \\
&\ge \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left[\min\left(\alpha_n\|\hat{\phi}_n(s,a)\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}, 2\right) + \gamma\mathbb{E}_{s'\sim\hat{P}_n(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \gamma\mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right] \\
&\ge \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\left[\min\left(\alpha_n\|\hat{\phi}_n(s,a)\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}, 2\right) - \min\left(\alpha_n\|\hat{\phi}_n(s,a)\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} + \sqrt{(1-\gamma)|\mathcal{A}|\zeta_n}, 2\right)\right] \\
&\ge -\sqrt{(1-\gamma)|\mathcal{A}|\zeta_n}.
\end{aligned}$$

This concludes the proof. From the second line to the third line, we use $\|V^{\pi}_{P^\star,r}\|_\infty \le O(1)$ and (10). □

Next, we obtain the upper bound of $\sum_{n=0}^{N} V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r}$. Recall $\pi^\star$ is the optimal policy. Though this form is the same as a standard regret form, since we need to use several samples to get one triple per iteration, we cannot get the regret guarantee. However, it suffices for the PAC guarantee.

Lemma 9. *With probability $1-\delta$, we have*

$$\sum_{n=1}^{N} V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \lesssim \sqrt{N\ln\left(1+\frac{N}{d^2\ln(|\mathcal{M}|/\delta)}\right)\ln(N|\mathcal{M}|/\delta)}\frac{|\mathcal{A}|^2 d^2}{(1-\gamma)}.$$

Proof. In this proof, letting $f_n(s,a) = \|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$, we condition on the event

$$\forall n, \quad \mathbb{E}_{(s,a)\sim\rho_n}[f_n^2(s,a)] \le \zeta_n, \quad \forall n, \forall \phi, \|\phi(s,a)\|_{\hat{\Sigma}_n,\phi} = \Theta(\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi}}). \qquad (6)$$

From Lemma 11 and Lemma 17, this event happens with probability $1-\delta$.

For any fixed episode n and any policy π, we have

$$\begin{aligned}
&V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \\
&\le V^{\pi_n}_{\hat{P}_n,r+\hat{b}_n} - V^{\pi_n}_{P^\star,r} + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)} \qquad \text{(Lemma 8)} \\
&= \mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[\hat{b}_n(s,a) + \gamma\mathbb{E}_{\hat{P}_n(s'|s,a)}[V^{\pi_n}_{\hat{P}_n,r+\hat{b}_n}(s')] - \gamma\mathbb{E}_{P^\star(s'|s,a)}[V^{\pi_n}_{\hat{P}_n,r+\hat{b}_n}(s')]] + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)}.
\end{aligned}$$

We use simulation lemma, Lemma 20, in the last display.

Then, noting $\|\hat{b}_n\|_\infty \le 2$, we have $\|V^{\pi_n}_{\hat{P}_n,r+\hat{b}_n}\|_\infty \le 2/(1-\gamma)$. Combining this fact with the above expansion, we have

$$\begin{aligned}
&V^{\pi}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \\
&\le \mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[\hat{b}_n(s,a)] + \left(\frac{1}{1-\gamma}\right)\mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[f_n(s,a)] + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)} \\
&\le \underbrace{|\mathcal{A}|\mathbb{E}_{(s,a)\sim d^{\bar{\pi}_n}_{P^\star}}\left[\hat{b}_n(s,a)\right]}_{\text{(a)}} + \underbrace{\left(\frac{1}{1-\gamma}\right)|\mathcal{A}|\mathbb{E}_{(s,a)\sim d^{\bar{\pi}_n}_{P^\star}}[f_n(s,a)]}_{\text{(b)}} + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)}
\end{aligned}$$

(Importance sampling)

First, we calculate the first term (a). Following Lemma 13 and noting the bonus $\hat{b}_n$ is $O(1)$, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_{P^\star}^{\bar{\pi}_n}}\left[\hat{b}_n(s,a)\right] \\
&\lesssim \mathbb{E}_{(s,a)\sim d_{P^\star}^{\bar{\pi}_n}}\left[\min\left(\alpha_n\|\phi^\star(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}, 2\right)\right] && \text{(From (6))} \\
&\lesssim \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{n\gamma|\mathcal{A}|\alpha_n^2\mathbb{E}_{(s,a)\sim\rho_n}\left[\|\hat{\phi}_n(s,a)\|^2_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}\right]+d\gamma\lambda_n} \\
&+\sqrt{|\mathcal{A}|\alpha_n^2\mathbb{E}_{(s,a)\sim\rho_n}\left[\|\hat{\phi}_n(s,a)\|^2_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}\right](1-\gamma)}.
\end{aligned}$$

In addition, we have

$$n\mathbb{E}_{(s,a)\sim\rho_n}\left[\|\hat{\phi}_n(s,a)\|^2_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}\right] = n\,\mathrm{Tr}(\mathbb{E}_{\rho_n}[\hat{\phi}_n\hat{\phi}_n^\top]\{n\mathbb{E}_{\rho_n}[\hat{\phi}_n\hat{\phi}_n^\top]+\lambda_n I\}^{-1}) \le d.$$

Then,

$$\mathbb{E}_{(s,a)\sim d_{P^\star}^{\bar{\pi}_n}}\left[\hat{b}_n(s,a)\right] \le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{\gamma d|\mathcal{A}|\alpha_n^2+\gamma d\lambda_n}+\sqrt{d|\mathcal{A}|\alpha_n^2(1-\gamma)/n}.$$

Second, we further calculate the second term (b). Following Lemma 13 and noting $f_n^2(s,a)$ is upper-bounded by 4, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_{P^\star}^{\bar{\pi}_n}}[f_n(s,a)] \\
&\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{\left\{n|\mathcal{A}|\gamma\mathbb{E}_{(s,a)\sim\rho_n}\left[f_n^2(s,a)\right]\right\}+4\gamma\lambda_n d}+\sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[f_n^2(s,a)(1-\gamma)\right]} \\
&\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{n|\mathcal{A}|\gamma\zeta_n+4\gamma\lambda_n d}+\sqrt{|\mathcal{A}|\zeta_n(1-\gamma)} \\
&\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\alpha_n+\sqrt{|\mathcal{A}|\zeta_n(1-\gamma)}.
\end{aligned}$$

In the last line, recall $\sqrt{\gamma}\sqrt{n|\mathcal{A}|\zeta_n+\lambda_n d+n\zeta_n}\lesssim\alpha_n$.

Then, by combining the above calculation of the term (a) and term (b),

$$\begin{aligned}
V^{\pi}_{P^\star,r}-V^{\pi_n}_{P^\star,r} \lesssim |\mathcal{A}|&\left(\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{d|\mathcal{A}|\alpha_n^2+d\lambda_n}+\sqrt{\frac{d|\mathcal{A}|\alpha_n^2(1-\gamma)}{n}}\right) \\
&+\frac{|\mathcal{A}|}{(1-\gamma)}\left(\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\alpha_n+\sqrt{|\mathcal{A}|\zeta_n(1-\gamma)}\right).
\end{aligned}$$

Hereafter, we take the dominating term out. First, recall

$$\alpha_n \lesssim \sqrt{\{|\mathcal{A}|+d^2\}\ln(N|\mathcal{M}|/\delta))} \lesssim \sqrt{|\mathcal{A}|d^2\ln(N|\mathcal{M}|/\delta)}.$$

Second, we also use

$$\begin{aligned}
\sum_{n=1}^N \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}} &\le \sqrt{N\sum_{n=1}^N \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}[\phi^\star(\tilde{s},\tilde{a})^\top\Sigma^{-1}_{\rho_n,\phi^\star}\phi^\star(\tilde{s},\tilde{a})]} && \text{(CS inequality)} \\
&\lesssim \sqrt{N\left(\ln\det(\sum_{n=1}^N \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\bar{\pi}_n}}[\phi^\star(\tilde{s},\tilde{a})\phi^\star(\tilde{s},\tilde{a})^\top])-\ln\det(\lambda_1 I)\right)} && \text{(Lemma 18 and } \lambda_1\le\cdots\le\lambda_N) \\
&\le \sqrt{dN\ln\left(1+\frac{N}{d\lambda_1}\right)}. && \text{(Potential function bound, Lemma 19 noting } \|\phi^\star(s,a)\|_2\le 1 \text{ for any } (s,a).)
\end{aligned}$$

Finally,

$$\begin{aligned}
\sum_{n=1}^{N} V^{\pi}_{P^\star,r} - V^{\pi_n}_{P^\star,r} &\lesssim |\mathcal{A}| \left(\sqrt{dN \ln\left(1+\frac{N}{d\lambda_1}\right)} \sqrt{d|\mathcal{A}|\alpha_N^2 + d\lambda_N} + \sum_{n=1}^{N} \sqrt{\frac{d|\mathcal{A}|\alpha_n^2(1-\gamma)}{n}} \right) \\
&\quad + \frac{|\mathcal{A}|}{(1-\gamma)} \left(\sqrt{dN \ln\left(1+\frac{N}{d\lambda_1}\right)} \alpha_N + \sum_{n=1}^{N} \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)} \right) \\
&\lesssim |\mathcal{A}| \sqrt{dN \ln\left(1+\frac{N}{d\lambda_1}\right)} \sqrt{d|\mathcal{A}|\alpha_N^2} + \frac{|\mathcal{A}|}{(1-\gamma)} \sqrt{dN \ln\left(1+\frac{N}{d\lambda_1}\right)} \alpha_N \\
&\qquad \text{(Some algebra. We take the dominating term out.)} \\
&\lesssim \sqrt{dN \ln\left(1+\frac{N}{d\lambda_1}\right)} \frac{|\mathcal{A}|^2 d^{3/2} \ln^{1/2}(N|\mathcal{M}|/\delta)}{(1-\gamma)}.
\end{aligned}$$

This concludes the proof. □

Using Lemma 9, we can immediately obtain the PAC guarantee.

Theorem 10 (PAC guarantee of REP-UCB)**.** *By using a number of interactions with the environment at most*

$$N \log(N/\delta), \quad N := O\left(\frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2} \ln^2\left(1 + \frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2}\right)\right).$$

with probability $1-\delta$, we can ensure $V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \le \epsilon$.

Proof. From Lemma 9 and Lemma 22, when N is

$$O\left(\frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2} \ln^2\left(1 + \frac{d^4|\mathcal{A}|^4 \ln(|\mathcal{M}|/\delta)}{(1-\gamma)^2\epsilon^2}\right)\right),$$

with probability $1-\delta$, we can ensure

$$\frac{1}{N}\sum_{n=1}^{N} V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \le \epsilon.$$

With probability $1-\delta$, we need $\ln(1/\delta)$ interactions with the environment to get one triple (s,a,s') from $d^{\pi}_{P^\star}(s,a)$. Thus, the total sample complexity is $O(N\ln(N/\delta))$.

□

Next, we provide an important lemma to ensure the concentration of the bonus term. The version for fixed ϕ is proved in Zanette et al. (2021a, Lemma 39). Here, we take a union bound over the whole feature $\phi \in \Phi$. Recall

$$\rho_n = \frac{1}{n}\sum_{i=0}^{n-1} d^{\bar{\pi}_i}_{P^\star}(s,a), \quad (s^{(i)}, a^{(i)}) \sim d^{\bar{\pi}_i}_{P^\star}(s,a).$$

The extension to the offline case is again straightforward.

Lemma 11 (Concentration of the bonus term)**.** *Set $\lambda_n = \Theta(d\ln(n|\Phi|/\delta))$ for any n. Define*

$$\Sigma_{\rho_n,\phi} = \mathbb{E}_{(s,a)\sim\rho_n}[\phi(s,a)\phi^\top(s,a)] + \lambda_n I, \quad \hat{\Sigma}_{n,\phi} = \sum_{i=0}^{n-1} \phi(s^{(i)}, a^{(i)})\phi^\top(s^{(i)}, a^{(i)}).$$

With probability $1-\delta$, we have

$$\forall n \in \mathbb{N}^+, \forall \phi \in \Phi, c_1\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi}} \le \|\phi(s,a)\|_{\hat{\Sigma}^{-1}_{n,\phi}} \le c_2\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n,\phi}}.$$

Next, we show our key lemma leveraging low-rank MDP structures. For any $g \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, This lemma shows that $\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\{g(s,a)\}$ is upper-bounded using $\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}} \|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}}$ as long as we have the convergence guarantee for $\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]$.

Lemma 12 (One-step back inequality for the learned model)**.** *Take any $g \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ s.t. $\|g\|_\infty \le B$. We condition on the event where the MLE guarantee* (14)*:*

$$\mathbb{E}_{(s,a)\sim\rho_n}[\|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2] \lesssim \zeta_n,$$

holds. Then, for any policy π, we have

$$|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\{g(s,a)\}| \le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}} \|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} \sqrt{\gamma}\sqrt{\left\{n|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]\right\} + \lambda_n d B^2 + nB^2\zeta_n}$$
$$+ \sqrt{(1-\gamma)|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]}.$$

Recall $\Sigma_{\rho_n,\hat{\phi}_n} = \mathbb{E}_{(s,a)\sim\rho_n}[\hat{\phi}_n(s,a)\hat{\phi}_n^\top(s,a)] + \lambda_n I$.

Proof. First, we have an equality:

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\{g(s,a)\} = \gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}, s\sim\hat{P}_n(\tilde{s},\tilde{a}), a\sim\pi(s)}\{g(s,a)\} + (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\pi(s_0)}\{g(s,a)\}. \tag{7}$$

The second term in (7) is upper-bounded by

$$(1-\gamma)\sqrt{\max_{(s,a)} \frac{d_0(s)\pi(a \mid s)}{\rho_n(s,a)} \mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]}$$
$$\le (1-\gamma)\sqrt{\max_{(s,a)} \frac{d_0(s)\pi(a \mid s)}{(1-\gamma)d_0(s)u(a)} \mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]} \le \sqrt{(1-\gamma)|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]}.$$

Next we consider the first term in (7). By CS inequality, we have

$$\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}, s\sim\hat{P}_n(\tilde{s},\tilde{a}), a\sim\pi(s)}\{g(s,a)\} = \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}} \hat{\phi}_n(\tilde{s},\tilde{a})^\top \int \hat{\mu}_n(s)\pi(a \mid s)g(s,a)d(s,a)$$
$$\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}} \|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}_n}} \left\|\int \hat{\mu}_n(s)\pi(a \mid s)g(s,a)d(s,a)\right\|_{\Sigma_{\rho_n,\hat{\phi}_n}}.$$

Then,

$$\|\int \hat{\mu}_n(s)\pi(a \mid s)g(s,a)d(s,a)\|^2_{\Sigma_{\rho_n,\hat{\phi}_n}}$$
$$\le \left\{\int \hat{\mu}_n(s)\pi(a \mid s)g(s,a)d(s,a)\right\}^\top \left\{n\mathbb{E}_{(s,a)\sim\rho_n}[\hat{\phi}_n\hat{\phi}_n^\top] + \lambda_n I\right\} \left\{\int \hat{\mu}_n(s)\pi(a \mid s)g(s,a)d(s,a)\right\}$$
$$\le n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n}\left[\int \hat{\mu}_n(s)^\top\hat{\phi}_n(\tilde{s},\tilde{a})\pi(a \mid s)g(s,a)d(s,a)\right]\right\}^2 + B^2\lambda_n d$$

(Use the assumption $\|g\|_\infty \le B$ and $\|\hat{\mu}_n(s)\mathrm{d}(s)\|_2 \le \sqrt{d}$.)

$$= n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n, s\sim\hat{P}_n(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[g(s,a)\right]\right\}^2 + B^2\lambda_n d$$
$$\le n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n, s\sim\hat{P}_n(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[g^2(s,a)\right]\right\} + B^2\lambda_n d. \qquad \text{(Jensen)}$$

Finally, the first term is upper bounded by

$$n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n, s\sim\hat{P}_n(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[g^2(s,a)\right]\right\} + \lambda_n d B^2$$
$$\le n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n, s\sim P^\star(\tilde{s},\tilde{a}), a\sim\pi(s)}\left[g^2(s,a)\right]\right\} + \lambda_n d B^2 + n\zeta_n B^2 \qquad \text{(MLE guarantee)}$$
$$\le \frac{n}{\min\{u(a)\}}\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n, s\sim P^\star(\tilde{s},\tilde{a}), a\sim u(s)}\left[g^2(s,a)\right]\right\} + \lambda_n d B^2 + n\zeta_n B^2$$
$$\le n|\mathcal{A}|\left\{\frac{1}{(\gamma)}\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]\right\} + \lambda_n d B^2 + n\zeta_n B^2.$$

Then, the final statement is immediately concluded. □

Here, we show a similar lemma as Lemma 12. The difference is we aim for calculating $\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}\{g(s,a)\}$ instead of $\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\{g(s,a)\}$. For any $g \in \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, this lemma shows that $\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}\{g(s,a)\}$ is upper-bounded using $\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\hat{\phi}^\star}}$ as long as we have the convergence guarantee for $\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]$. Note comparing to Lemma 12, this is not a probabilistic statement.

Lemma 13 (One-step back inequality for the true model). *Take any $g\in\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ such that $\|g\|_\infty\le B$. Then,*

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}\{g(s,a)\}\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{\gamma}\sqrt{n|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]+\lambda_n dB^2} + \sqrt{(1-\gamma)|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]}.$$

Recall $\Sigma_{\rho_n,\phi^\star}=n\mathbb{E}_{(s,a)\sim\rho_n}[\phi^\star\{\phi^\star\}^\top]+\lambda_n I$.

Proof. First, we have

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}\{g(s,a)\}=\gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star},s\sim P^\star(\tilde{s},\tilde{a}),a\sim\pi(s)}\{g(s,a)\}+(1-\gamma)\mathbb{E}_{s\sim d_0,a\sim\pi(s_0)}\{g(s,a)\}. \tag{8}$$

The second term in (8) is upper-bounded by

$$(1-\gamma)\sqrt{\max_{(s,a)}\frac{d_0(s)\pi(a\mid s)}{\rho_n(s,a)}\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]}\le\sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right](1-\gamma)}.$$

By CS inequality, the first term in (8) is further bounded as follows:

$$\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star},s\sim P^\star(\tilde{s},\tilde{a}),a\sim\pi(s)}\{g(s,a)\}=\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}}\phi^\star(\tilde{s},\tilde{a})^\top\int\mu^\star(s)\pi(a\mid s)g(s,a)d(s,a)$$
$$\le\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\left\|\int\mu^\star(s)\pi(a\mid s)g(s,a)d(s,a)\right\|_{\Sigma_{\rho_n,\phi^\star}}.$$

Here, we have

$$\begin{aligned}
&\|\int\mu^\star(s)\pi(a\mid s)g(s,a)d(s,a)\|^2_{\Sigma_{\rho_n,\phi^\star}}\\
&\le\left\{\int\mu^\star(s)\pi(a\mid s)g(s,a)d(s,a)\right\}^\top\left\{n\mathbb{E}_{(s,a)\sim\rho_n}[\phi^\star(s,a)\{\phi^\star(s,a)\}^\top]+\lambda_n I\right\}\left\{\int\mu^\star(s)\pi(a\mid s)g(s,a)d(s,a)\right\}\\
&\le n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n}\left[\int\mu^\star(s)^\top\phi^\star(\tilde{s},\tilde{a})\pi(a\mid s)g(s,a)d(s,a)\right]\right\}^2+\lambda_n dB^2\\
&=n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n,s\sim P^\star(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[g(s,a)\right]\right\}^2+\lambda_n dB^2\\
&\le n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n,s\sim P^\star(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[g^2(s,a)\right]\right\}+\lambda_n dB^2. \qquad \text{(Jensen)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n,s\sim P^\star(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[g^2(s,a)\right]\right\}+\lambda_n dB\\
&\le\frac{n}{\min\{u(a)\}}\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_n,s\sim P^\star(\tilde{s},\tilde{a}),a\sim u(s)}\left[g^2(s,a)\right]\right\}+\lambda_n dB^2\\
&\le n|\mathcal{A}|\left\{\frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]\right\}+\lambda_n dB^2.
\end{aligned}$$

Finally, we have

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}\{g(s,a)\}\le\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n,\phi^\star}}\sqrt{\left\{\frac{n|\mathcal{A}|}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]\right\}+\lambda_n dB^2} + (1-\gamma)\sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_n}\left[g^2(s,a)\right]/(1-\gamma)}.$$

□

B PROOF OF THE THEORETICAL PROPERTY OF REP-LCB

Below we first prove that $V^{\pi}_{\hat{P},r-b}$ is an almost pessimistic estimator of $V^{\pi}_{P^\star,r}$.

Lemma 14 (Almost Pessimism at the Initial Distribution). *Let* $\omega = \max_{a,s} 1/\pi_b(a \mid s)$. *Set*

$$\alpha = c_1\sqrt{(\omega + d^2)\,\gamma \ln(|\mathcal{M}|/\delta)}, \quad \lambda = O(d\ln|\mathcal{M}|/\delta), \zeta = O\left(\frac{\ln(|\mathcal{M}|/\delta)}{n}\right).$$

With probability $1-\delta$*, for any policy* π*, we have*

$$V^{\pi}_{\hat{P},r-\hat{b}} - V^{\pi}_{P^\star,r} \le \sqrt{\frac{\omega \ln(|\mathcal{M}|/\delta)(1-\gamma)}{n}}.$$

Proof. We define

$$\Sigma_{\rho,\phi} = n\mathbb{E}_{(s,a)\sim\rho}[\phi\phi^\top] + \lambda I, \quad \hat{\Sigma}_\phi = n\mathbb{E}_{\mathcal{D}}[\phi\phi^\top] + \lambda I.$$

where $\lambda = O(d\ln|\mathcal{M}|/\delta)$. In this proof, letting $f(s,a) = \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$, we condition on the events:

$$\mathbb{E}_{(s,a)\sim\rho}[f^2(s,a)] \le \zeta, \quad \forall \phi \in \Phi : \|\phi(s,a)\|_{\hat{\Sigma}^{-1}_\phi} = \Theta(\|\phi(s,a)\|_{\Sigma^{-1}_{\rho,\phi}}). \tag{9}$$

where $\zeta = O(\ln(|\mathcal{M}|/\delta)/n)$. From the offline version of Lemma 17 and Lemma 11 [6], this event happens with probability $1-\delta$.

Then, from simulation lemma (Lemma 20),

$$\begin{aligned}
&V^{\pi}_{\hat{P},r-\hat{b}} - V^{\pi}_{P^\star,r}\\
&= \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left[-\hat{b}(s,a) + \gamma\mathbb{E}_{s'\sim\hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \gamma\mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right]\\
&\le \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left[-\min\left(\alpha\|\hat{\phi}(s,a)\|_{\Sigma^{-1}_{\rho,\hat{\phi}}}, 2\right) + \gamma\mathbb{E}_{s'\sim\hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \gamma\mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right]. \quad \text{(From (9))}
\end{aligned}$$

Here, we have

$$\left|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{\mathbb{E}_{s'\sim\hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right\}\right| \le \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{f(s,a)\right\},$$

noting $\|V^{\pi}_{P^\star,r}\|_\infty = O(1)$. This is further bounded by Lemma 16:

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{f(s,a)\right\} \lesssim \sqrt{\alpha'}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}}\|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}} + \sqrt{\omega\zeta(1-\gamma)}. \tag{10}$$

where

$$\alpha' = n\gamma\omega\zeta + \gamma^2\lambda d + \gamma^2 n\zeta \lesssim \left(\omega + d^2\right)\gamma\ln(|\mathcal{M}|/\delta).$$

Here, we use $f(s,a) \le 2$ in Lemma 16 and $\mathbb{E}_{(s,a)\sim\rho}[f^2(s,a)] \le \zeta$.

Thus,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{\mathbb{E}_{s'\sim\hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right\}\right| \le \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left\{f(s,a)\right\}\\
&\le \sqrt{\alpha'}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}}\|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}} + \sqrt{\omega\zeta(1-\gamma)}\\
&= \alpha\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}}\|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}} + \sqrt{\omega\zeta(1-\gamma)}, \quad \alpha = \sqrt{\alpha'}.
\end{aligned}$$

Going back to the simulation lemma 20, we have

$$\begin{aligned}
&V^{\pi}_{\hat{P},r-\hat{b}} - V^{\pi}_{P^\star,r}\\
&\le \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left[-\min\left(\alpha\|\hat{\phi}(s,a)\|_{\Sigma^{-1}_{\rho,\hat{\phi}}}, 2\right) + \mathbb{E}_{s'\sim\hat{P}(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right] - \mathbb{E}_{s'\sim P^\star(s,a)}\left[V^{\pi}_{P^\star,r}(s')\right]\right]\\
&\le \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}}\left[-\min\left(\alpha\|\hat{\phi}(s,a)\|_{\Sigma^{-1}_{\rho,\hat{\phi}}}, 2\right) + \min\left(\alpha\|\hat{\phi}(s,a)\|_{\Sigma^{-1}_{\rho,\hat{\phi}}} + \sqrt{\omega\zeta(1-\gamma)}, 2\right)\right]\\
&\le \sqrt{\omega\zeta(1-\gamma)}.
\end{aligned}$$

This concludes the proof. ∎

[6] We can remove $\ln n$ since n is fixed in the offline setting.

With the above lemma, now we can proceed to prove the main theorem.

Theorem 15 (PAC guarantee of REP-LCB)**.** *Set the parameters as in Lemma 14. With probability $1-\delta$, for any comparator policy π including history-dependent non-Markovian policies, we have*

$$V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \lesssim \frac{\omega d^2}{(1-\gamma)}\sqrt{\frac{C^\star_\pi \ln(|\mathcal{M}|/\delta)}{n}}, \quad C^\star_\pi := \sup_{x\in\mathbb{R}} \frac{x^\top \mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\phi^\star(s,a)\{\phi^\star(s,a)\}^\top]x}{x^\top \mathbb{E}_{(s,a)\sim\rho}[\phi^\star(s,a)\{\phi^\star(s,a)\}^\top]x}.$$

Proof. In this proof, letting $f(s,a) = \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$ we condition on the events:

$$\mathbb{E}_{(s,a)\sim\rho}[f^2(s,a)] \le \zeta, \quad \forall \phi \in \Phi : \|\phi(s,a)\|_{\hat{\Sigma}^{-1}_\phi} = \Theta(\|\phi(s,a)\|_{\Sigma^{-1}_{\rho,\phi}}). \tag{11}$$

From Lemma 11 and Lemma 17, this event happens with probability $1-\delta$.

For any policy π, we have

$$\begin{aligned}
&V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \\
&\le V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{\hat{P},r-\hat{b}} + \sqrt{\omega\zeta(1-\gamma)} && \text{(Lemma 14)} \\
&\le V^{\pi}_{P^\star,r} - V^{\pi}_{\hat{P},r-\hat{b}} + \sqrt{\omega\zeta(1-\gamma)} \\
&\lesssim \underbrace{\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\hat{b}(s,a)]}_{\text{(a)}} + \underbrace{\left(\frac{1}{1-\gamma}\right)\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[f(s,a)]}_{\text{(b)}} + \sqrt{\omega\zeta(1-\gamma)}.
\end{aligned}$$

Recall $f(s,a) = \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$.

From the second line to the third line, note though $\hat{\pi}$ is the argmax over Markoovian polices, $\hat{\pi}$ is also the argmax over all history-dependent polices. In the last line, we use a simulation lemma 20, which is tailored to a time-inhomogeneous policy. We here use $\|V^{\pi}_{\hat{P},r-\hat{b}}\|_\infty \le 2/((1-\gamma))$. noting $\|\hat{b}\|_\infty = O(1)$.

We further calculate the first term (a). Considering the offline version of Lemma 13 and noting $\|\hat{b}\|_\infty \le 2$, we have

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\hat{b}(s,a)] \lesssim \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \sqrt{n\omega\left\{\gamma\mathbb{E}_{(s,a)\sim\rho}\left[\hat{b}^2(s,a)\right]\right\} + \gamma\lambda d} \\
+ \sqrt{\omega(1-\gamma)}\{\mathbb{E}_\rho[\hat{b}^2(s,a)]\}^{1/2}.
\end{aligned}$$

From (11), we have

$$\begin{aligned}
&n\mathbb{E}_{(s,a)\sim\rho}\left[\hat{b}^2(s,a)\right] \le n\mathbb{E}_{(s,a)\sim\rho}\left[\min\left(\alpha^2\|\hat{\phi}(s,a)\|^2_{\Sigma^{-1}_{\rho,\hat{\phi}}}, 4\right)\right] \le n\mathbb{E}_{(s,a)\sim\rho}\left[\alpha^2\|\hat{\phi}(s,a)\|^2_{\Sigma^{-1}_{\rho,\hat{\phi}}}\right] \\
&\le \mathrm{Tr}[n\mathbb{E}[\hat{\phi}\hat{\phi}^\top]\{n\mathbb{E}[\hat{\phi}\hat{\phi}^\top] + \lambda I\}^{-1}] \le \mathrm{Tr}[n(\mathbb{E}[\hat{\phi}\hat{\phi}^\top] + \lambda I)\{n\mathbb{E}[\hat{\phi}\hat{\phi}^\top] + \lambda I\}^{-1}] \le d.
\end{aligned} \tag{12}$$

Thus,

$$\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[\hat{b}(s,a)] \le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \sqrt{\omega d\alpha^2\gamma + \gamma\lambda d} + \sqrt{\frac{\omega d\alpha^2(1-\gamma)}{n}}.$$

Second, we further calculate the second term (b). Considering the offline version of Lemma 13 and noting $f^2(s,a)$ is upper-bounded by 4,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{P^\star}}[f(s,a)] \\
&= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \sqrt{n\omega\left\{\gamma\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]\right\} + 4\gamma\lambda d} + \sqrt{\omega\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right](1-\gamma)} \\
&\lesssim \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \sqrt{\omega\left\{n\gamma\zeta\right\} + \gamma\lambda d} + \sqrt{\omega\zeta(1-\gamma)} \\
&\lesssim \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \alpha + \sqrt{\omega\zeta(1-\gamma)}.
\end{aligned}$$

In the final line, recall $\sqrt{\omega\{n\gamma\zeta\} + \gamma\lambda d + \gamma n\zeta} \leq \alpha$.

Finally, by combining the calculation of the first term (a) and the second term (b), we have

$$V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \lesssim \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \sqrt{d\alpha^2\omega\gamma + \gamma\lambda d} + \sqrt{\frac{\omega\alpha^2 d(1-\gamma)}{n}}$$
$$+ \frac{\alpha}{(1-\gamma)} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} + \sqrt{\frac{\omega\zeta}{(1-\gamma)}}$$

Now, we use the fact $\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}}$ is upper-bounded as

$$\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\phi^\star}} \leq \sqrt{\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star}} \|\phi^\star(\tilde{s},\tilde{a})\|^2_{\Sigma^{-1}_{\rho,\phi^\star}}} \leq \sqrt{C^\star \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho} \|\phi^\star(\tilde{s},\tilde{a})\|^2_{\Sigma^{-1}_{\rho,\phi^\star}}}$$
(Refer to Lemma 21)
$$\leq \sqrt{C^\star d/n}. \qquad \text{(From (12))}$$

Finally, we have

$$\begin{aligned}
&V^{\pi}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \\
&\lesssim \sqrt{\frac{C^\star d}{n}} \sqrt{d\alpha^2\omega\gamma + \gamma\lambda d} + \frac{\alpha}{(1-\gamma)} \sqrt{\frac{C^\star d}{n}} + \sqrt{\frac{\omega\alpha^2 d(1-\gamma)}{n}} + \sqrt{\frac{\omega\zeta}{(1-\gamma)}} \\
&\lesssim \sqrt{\frac{C^\star d}{n}} \sqrt{d\alpha^2\omega\gamma} + \frac{\alpha}{(1-\gamma)} \sqrt{\frac{C^\star d}{n}} \qquad \text{(Take out two dominating terms)} \\
&\lesssim \frac{\omega d^2}{(1-\gamma)} \sqrt{\frac{C^\star \ln(|\mathcal{M}|/\delta)}{n}}.
\end{aligned}$$

□

The lemma below is a key technical lemma for our proof. It shows that one can relate the expected value of any function $f(s,a)$ with respect to $d^{\pi}_{\hat{P}}$ (i.e., inside the learned model $\hat{P}$) to the potential function with respect to $d^{\pi}_{\hat{P}}$, i.e., $\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}} \|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}}$. Pairing $\hat{\phi}$ and $\hat{P}$ is important since $\hat{P}$ is the low-rank transition model defined using $\hat{\phi}$. As we have seen in the above analysis, when using the lemma below, we instantiate $f(s,a) := \|\hat{P}(\cdot|s,a) - P^\star(\cdot|s,a)\|_1$.

Lemma 16 (One-step back inequality for the learned model)**.** *Take any $f \subset \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ s.t. $\|f\|_\infty \leq B$. We condition on the event where the MLE guarantee holds:*

$$\mathbb{E}_{(s,a)\sim\rho} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|^2_1 \lesssim \zeta.$$

Then, letting $\omega = \max_{s,a}(1/\pi_b(a \mid s))$, for any policy π, we have

$$|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}} \{f(s,a)\}| \leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}} \|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}} \sqrt{\left\{n\omega\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]\right\} + \gamma^2\lambda d B^2 + n\gamma^2\zeta B^2}$$
$$+ \sqrt{\omega\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right](1-\gamma)}.$$

Proof. First, we have an equality:

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}}} \{f(s,a)\} = \gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}, s\sim\hat{P}(\tilde{s},\tilde{a}), a\sim\pi(s)} \{f(s,a)\} + (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\pi(s_0)} \{f(s,a)\}. \tag{13}$$

The second term in (13) is upper-bounded by

$$\mathbb{E}_{s\sim d_0, a\sim\pi(s_0)} \{f(s,a)\} \leq \mathbb{E}_{s\sim d_0, a\sim\pi(s_0)} \left\{f^2(s,a)\right\}^{1/2} = \sqrt{\omega\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]/(1-\gamma)}.$$

Next we consider the first term in (13). By CS inequality, we have

$$\left|\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}},s\sim\hat{P}(\tilde{s},\tilde{a}),a\sim\pi(s)}\left\{f(s,a)\right\}\right| = \left|\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}}\hat{\phi}(\tilde{s},\tilde{a})^{\top}\int\hat{\mu}(s)\pi(a\mid s)f(s,a)d(s,a)\right|$$
$$\leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}}}\|\hat{\phi}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho,\hat{\phi}}}\|\int\hat{\mu}(s)\pi(a\mid s)f(s,a)d(s,a)\|_{\Sigma_{\rho,\hat{\phi}}}.$$

Then,

$$\begin{aligned}
&\|\int\hat{\mu}(s)\pi(a\mid s)f(s,a)d(s,a)\|^2_{\Sigma_{\rho,\hat{\phi}}}\\
&\leq\left\{\int\hat{\mu}(s)\pi(a\mid s)f(s,a)d(s,a)\right\}^{\top}\left\{n\mathbb{E}_{(s,a)\sim\rho}[\hat{\phi}\hat{\phi}^{\top}]+\lambda I\right\}\left\{\int\hat{\mu}(s)\pi(a\mid s)f(s,a)d(s,a)\right\}\\
&\leq n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho}\left[\int\hat{\mu}(s)^{\top}\hat{\phi}(\tilde{s},\tilde{a})\pi(a\mid s)f(s,a)d(s,a)\right]\right\}^2+B^2\lambda d\\
&\qquad\qquad\text{(Use the assumption } \|f(s,a)\|_\infty\leq B \text{ and } \|\textstyle\int\hat{\mu}(s)\mathrm{d}(s)\|_2\leq\sqrt{d})\\
&=n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho,s\sim\hat{P}(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[f(s,a)\right]\right\}^2+B^2\lambda d\\
&\leq n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho,s\sim\hat{P}(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[f^2(s,a)\right]\right\}+B^2\lambda d. \qquad\text{(Jensen)}
\end{aligned}$$

Finally, the first term in (13) is upper-bounded by

$$\begin{aligned}
&n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho,s\sim\hat{P}(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[f^2(s,a)\right]\right\}+\lambda dB^2\\
&\leq n\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho,s\sim P^{\star}(\tilde{s},\tilde{a}),a\sim\pi(s)}\left[f^2(s,a)\right]\right\}+\lambda dB^2+nB^2\zeta\\
&\qquad\qquad\text{(MLE guarantee and } \|\mathbb{E}_{a\sim\pi(\cdot)}[f^2(\cdot,a)]\|_\infty\leq B^2.)\\
&\leq n\omega\left\{\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho,s\sim P^{\star}(\tilde{s},\tilde{a}),a\sim\pi_b(s)}\left[f^2(s,a)\right]\right\}+\lambda dB^2+nB^2\zeta \qquad\text{(Importance sampling)}\\
&\leq n\omega\left\{\frac{1}{\gamma}\mathbb{E}_{(s,a)\sim\rho}\left[f^2(s,a)\right]\right\}+\lambda dB^2+nB^2\zeta. \qquad\text{(Definition of } \rho)
\end{aligned}$$

The final statement is immediately concluded.

□

C AUXILIARY LEMMAS

First, we present the MLE guarantee. Regarding the proof, refer to Agarwal et al. (2020b, Theorem 21). Note $\hat{P}_n$ and $\bar{\pi}_n$ are the quantities appearing in the proposed online algorithm. We can also immediately obtain the statement to the offline case.

Lemma 17 (MLE guarantee)**.** *For a fixed episode n, with probability $1-\delta$,*

$$\sum_{j=0}^{n-1}\mathbb{E}_{(s,a)\sim d^{\bar{\pi}_j}}[\|\hat{P}_n(\cdot\mid s,a)-P^{\star}(\cdot\mid s,a)\|_1^2]\lesssim n\zeta,\quad \zeta\coloneqq\frac{\ln(|\mathcal{M}|/\delta)}{n}.$$

As a straightforward corollary, with probability $1-\delta$,

$$\forall n\in\mathbb{N}^{+},\sum_{j=0}^{n-1}\mathbb{E}_{(s,a)\sim d^{\bar{\pi}_j}}[\|\hat{P}_n(\cdot\mid s,a)-P^{\star}(\cdot\mid s,a)\|_1^2]\lesssim n\zeta_n,\quad \zeta_n\coloneqq\frac{\ln(|\mathcal{M}|n/\delta)}{n}. \tag{14}$$

The following is a standard inequality to prove regret bounds for linear models. Refer to Agarwal et al. (2020a, Lemma G.2.)

Lemma 18. *Consider the following process. For $n=1,\cdots,N$, $M_n=M_{n-1}+G_n$ with $M_0=\lambda_0 I$ and G_n being a positive semidefinite matrix with eigenvalues upper-bounded by 1. We have that:*

$$2\ln\det(M_N)-2\ln\det(\lambda_0 I)\geq\sum_{n=1}^{N}\mathrm{Tr}(G_n M_{n-1}^{-1}).$$

Lemma 19 (Potential function lemma)**.** *Suppose* $\mathrm{Tr}(G_n) \leq B^2$.

$$2\ln\det(M_N) - 2\ln\det(\lambda_0 I) \leq d\ln\left(1 + \frac{NB^2}{d\lambda_0}\right).$$

Proof. Let $\sigma_1, \cdots, \sigma_d$ be the set of singular values of M_N recalling M_N is a positive semidefinite matrix. Then, by the AM-GM inequality,

$$\ln\det(M_N)/\det(\lambda_0 I) = \ln\prod_{i=1}^{d}(\sigma_i/\lambda_0) \leq \ln d\left(\frac{1}{d}\sum_{i=1}^{d}(\sigma_i/\lambda_0))\right)$$

Since we have $\sum_i \sigma_i = \mathrm{Tr}(M_N) \leq d\lambda_0 + NB^2$, the statement is concluded. □

Lemma 20 (Simulation lemma)**.** *Given two MDPs* $(P', r+b)$ *and* (P, r)*, for any policy* π*, we have:*

$$V^{\pi}_{P',r+b} - V^{\pi}_{P,r} = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P'}}[b(s,a) + \gamma\mathbb{E}_{P'(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)] - \gamma\mathbb{E}_{P(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)]]$$

and

$$V^{\pi}_{P',r+b} - V^{\pi}_{P,r} = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P}}[b(s,a) + \gamma\mathbb{E}_{P'(s'|s,a)}[Q^{\pi}_{P,r+b}(s',\pi)] - \gamma\mathbb{E}_{P(s'|s,a)}[Q^{\pi}_{P',r+b}(s',\pi)]].$$

Proof. We use

$$V^{\pi}_{P} - f(s_0,\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi}_{P}}[r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}[f(s',\pi)] - f(s,a)]]$$

Then,

$$\begin{aligned}
V^{\pi}_{P',r+b} - V^{\pi}_{P,r} &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P'}}[r(s,a) + b(s,a) + \gamma\mathbb{E}_{P'(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)] - Q^{\pi}_{P,r}(s,a)]] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P'}}[b(s,a) + \gamma\mathbb{E}_{P'(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)] - \gamma\mathbb{E}_{P(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)]].
\end{aligned}$$

Similarly,

$$\begin{aligned}
V^{\pi}_{P,r} - V^{\pi}_{P',r+b} &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P}}[r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}[Q^{\pi}_{P',r+b}(s',\pi)] - Q^{\pi}_{P',r+b}(s,a)]] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi}_{P}}[-b(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}[Q^{\pi}_{P',r+b}(s',\pi)] - \gamma\mathbb{E}_{P'(s'|s,a)}[Q^{\pi}_{P,r}(s',\pi)]].
\end{aligned}$$

□

The following lemma is used to deal with the distribution shift in the offline setting. For the proof, refer to Chang et al. (2021).

Lemma 21 (Distribution shift lemma)**.** *Consider any policy* π *and state-action distribution* ρ*, and any representation* $\phi^\star$*, we have:*

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}[\phi^\star(s,a)\{\phi^\star(s,a)\}^\top] \leq C^\star\mathbb{E}_{\rho}[\phi^\star(s,a)\{\phi^\star(s,a)\}^\top], \quad C^\star := \sup_{x\in\mathbb{R}}\frac{x^\top\mathbb{E}_{(s,a)\sim d^{\pi}_{P^\star}}[\phi^\star\{\phi^\star\}^\top]x}{x^\top\mathbb{E}_{(s,a)\sim\rho}[\phi^\star\{\phi^\star\}^\top]x}.$$

This is some auxiliary lemma to convert the finite sample error bound into the sample complexity.

Lemma 22 (Conversion of finite sample error bounds into sample complexities)**.** *By taking*

$$N = 1/\epsilon'^2 \times \ln^2(1 + 1/\epsilon'^2), \epsilon' = \frac{\epsilon}{a_1\ln^{1/2}(e+a_2)\ln^{1/2}(e+a_3)}.$$

It satisfies

$$a_1\sqrt{1/N}\ln^{1/2}(1+a_2N)\ln^{1/2}(1+a_3N) < c\epsilon.$$

where c *is a constant independent of* a_1, a_2, a_3.

Proof. We first have
$a_1\sqrt{1/N}\ln^{1/2}(1+a_2N)\ln^{1/2}(1+a_3N) \le a_1\max(\ln^{1/2}(1+a_2)\ln^{1/2}(1+a_3),1)\sqrt{1/N}\ln(1+N)$.
Here, we use

$$\ln^{1/2}(1+a_2N) \le \{\ln(1+a_2)+\ln(1+N)\}^{1/2} \le \sqrt{\max(1,\ln(1+a_2))\ln(1+N)}.$$

Then, we prove when $N = 1/\epsilon^2 \times \ln^2(1+1/\epsilon^2)$.

$$\sqrt{1/N}\ln(1+N) < \epsilon.$$

This is proved by

$$\begin{aligned}
\sqrt{1/N}\ln(1+N) &\le \epsilon \times \frac{\ln(1+1/\epsilon^2\times\ln^2(1+1/\epsilon^2))}{\ln(1+1/\epsilon^2)} \\
&\le \epsilon \times \frac{\ln(1+1/\epsilon^2)+\ln(1+\ln^2(1+1/\epsilon^2))}{\ln(1+1/\epsilon^2)} \\
&\le \epsilon + \epsilon \times \frac{\ln(1+\ln^2(1+1/\epsilon^2))}{\ln(1+1/\epsilon^2)} \\
&\le \epsilon + \epsilon \times \frac{0.5\{1+\ln^2(1+1/\epsilon^2))\}^{1/2}-1}{\ln(1+1/\epsilon^2)} \\
&\lesssim \epsilon.
\end{aligned}$$

From the third line to the fourth line, we use $\ln(x) \le 0.5(x^{1/2}-1)$ for $x > 0$. Then, the final statement is concluded. □

D More comparison to Xie et al. (2021)

We briefly explain the guarantee when we use Algorithm 1 (Xie et al., 2021). For a given reward r, we first define a new feature class Φ_r^+.

Definition 23 (Augmented feature)**.** *Let* $\phi = [\phi_1, \cdots, \phi_d]$.

$$\Phi_r^+ = \{\phi_r^+ ; \phi \in \Phi\}, \quad \phi_r^+ = [\phi_1, \cdots, \phi_d, r].$$

Then, we set

$$\mathcal{F} = \{a^\top \phi_r^+ \mid \|a\|_2 \le c\sqrt{d}+1, \phi_r^+ \in \Phi_r^+\}.$$

where c is some suitable constant. Given the hypothesis class $\mathcal{F}$ for the Q-function, we can run Algorithm 1 in (Xie et al., 2021). We denote the output policy as $\hat{\pi}$.

We check two assumptions to ensure the algorithm works. The first assumption is realizability. This is satisfied since for any policy $\pi \in \Pi$ (Π is the class of all Markovian polices), we have $Q^\pi_{P^\star,r} \in \mathcal{F}$. The second assumption is completeness. This is also satisfied since $\mathcal{T}^\pi_{P^\star,r}\mathcal{F} \subset \mathcal{F}$ for any policy $\pi \in \Pi$ where $\mathcal{T}^\pi_{P^\star,r}$ is a Bellman-operator s.t.

$$\mathcal{T}^\pi_{P^\star,r} : \{\mathcal{S}\times\mathcal{A} \to \mathbb{R}\} \ni f \mapsto r(s,a) + \gamma\mathbb{E}_{s'\sim P^\star(s,a)}[f(s',\pi)] \in \{\mathcal{S}\times\mathcal{A}\to\mathbb{R}\},$$

where we denote $f(s,\pi) = \mathbb{E}_{a\sim\pi(s)}f(s,a)$. Then, by invoking their Corollary 5, we have

Theorem 24 (PAC bound based on Xie et al. (2021))**.** *With probability* $1-\delta$*,*

$$\forall \pi \in \Pi : V^\pi_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \le c\frac{\sqrt{C^\dagger_{\pi,r}}}{(1-\gamma)^2}\left(\frac{(d+1)\log(1/\delta)\log|\mathcal{A}|}{n}\right)^{1/5}.$$

where

$$C^\dagger_{\pi,r} = \sup_{\phi_r^+\in\Phi_r^+}\sup_{a\in\mathbb{R}^{d+1}} \frac{a^\top\mathbb{E}_{d^\pi_{P^\star}}[\phi_r^+\{\phi_r^+\}^\top]a}{a^\top\mathbb{E}_\rho[\phi_r^+\{\phi_r^+\}^\top]a}.$$

We compare the above result with our result in Theorem 6. First, since $C^\dagger_r$ includes r and all possible features in Φ, this partial coverage condition is stronger than ours (recall our partial coverage condition is only related to the true representation $\phi^\star$), and we always have $C^\star_\pi \le C^\dagger_{\pi,r}$. Secondly, the dependence on n is much worse. Third, it is unclear whether the learned policy can compete against any history-dependent policy. Recall in Theorem 6, we show that our algorithm can compete with any history-dependent policies.