Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach

Xuezhou Zhang [1] **Yuda Song** [2] **Masatoshi Uehara** [3] **Mengdi Wang** [1] **Alekh Agarwal** [4] **Wen Sun** [3]

Abstract

We present BRIEE (Block-structured Representation learning with Interleaved Explore Exploit), an algorithm for efficient reinforcement learning in Markov Decision Processes with block structured dynamics (i.e., Block MDPs), where rich observations are generated from a set of unknown latent states. BRIEE interleaves latent states discovery, exploration, and exploitation together, and can provably learn a near-optimal policy with sample complexity scaling polynomially in the number of latent states, actions, and the time horizon, with no dependence on the size of the potentially infinite observation space. Empirically, we show that BRIEE is more sample efficient than the state-of-art Block MDP algorithm HOMER and other empirical RL baselines on challenging rich-observation combination lock problems which require deep exploration.

1. Introduction

Representation learning in Reinforcement Learning (RL) has gained increasing attention in recent years from both theoretical and empirical research communities (Schwarzer et al., 2020; Laskin et al., 2020) due to its potential in enabling sample-efficient non-linear function approximation, the benefits in multitask settings (Zhang et al., 2020; Yang et al., 2022; Sodhani et al., 2021), and the potential to leverage advances on representation learning in related areas such as computer vision and natural language processing. Despite this interest, there remains a gap between the theoretical and empirical literature, where the theoretically sound methods are seldom evaluated or even implemented and often rely on strong assumptions, while the empirical techniques are not backed with any theoretical guarantees even under stylistic assumptions. This leaves open the key challenge of designing representation learning methods that are both theoretically sound and empirically effective.

In this work, we tackle this challenge for a special class of problems called *Block MDPs*, where the high dimensional and rich observations of the agent are generated from certain latent states and there exists some fixed, but unknown mapping from observations to the latent states (each observation is generated only by one latent state). Prior works (Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Zhang et al., 2020; Sodhani et al., 2021) have motivated the Block MDP model through scenarios such as navigation tasks and image based robotics tasks where the observations can often be reasonably mapped to the latent physical location and states. We develop a new algorithm BRIEE, which finds a provably good policy for any Block MDP. It performs model-free representation learning with a form of adversarial training to learn the features, interleaved with deep exploration and exploitation. Unlike prior theoretical works, our new approach does not require *uniform reachability*, i.e., every latent state is reachable with sufficient probability, which is a strong assumption that cannot be guaranteed or verified. We also demonstrate the empirical effectiveness of our algorithm in Block MDPs that are challenging to explore. Importantly, our technique is model-free which means there is no need to model the observation generation process that can be complex for high dimensional sensory data.

Contributions Our key contributions are three folds:

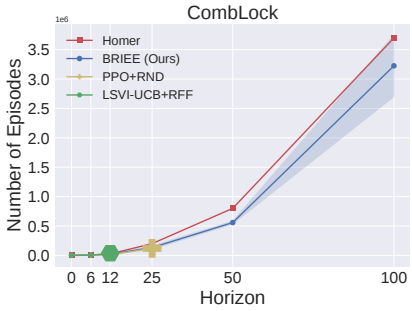
1. We design a new algorithm BRIEE that solves any Block MDP with polynomial sample complexity, with no explicit dependence on the number of states which could be infinite;
2. BRIEE does not require reachability assumption, and can directly optimize a given reward function;
3. BRIEE's computation oracles can be easily implemented using standard gradient based optimization. Our experiments show that BRIEE is more sample efficient than HOMER (Misra et al., 2020), can be extended to richer MDPs where the block structure does not hold, and can leverage dense reward structure to achieve

[1]Princeton University [2]Carnegie Mellon University [3]Cornell University [4]Google Research. Correspondence to: Xuezhou Zhang <zhangxz1123@gmail.com>, Wen Sun <ws455@cornell.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

	Sample Complexity	Reward
OLIVE (Jiang et al., 2017)	$\frac{\lvert\mathcal{Z}\rvert^3 H^3 \lvert\mathcal{A}\rvert^2 \log\lvert\Phi\rvert}{\epsilon^2}$	Yes
REP-UCB (Uehara et al., 2021)	$\frac{\lvert\mathcal{Z}\rvert^4 H^5 \lvert\mathcal{A}\rvert^2 \ln(\lvert\Phi\rvert\lvert\Upsilon\rvert)}{\epsilon^2}$	Yes
MOFFLE (Modi et al., 2021)	$\frac{\lvert\mathcal{Z}\rvert^7 H^8 \lvert\mathcal{A}\rvert^{13} \ln\lvert\Phi\rvert}{\min(\epsilon^2\eta_{\min}, \eta_{\min}^5)}$	No
HOMER (Misra et al., 2019)	$\frac{\lvert\mathcal{Z}\rvert^6 H \lvert\mathcal{A}\rvert(\lvert\mathcal{Z}\rvert^2\lvert\mathcal{A}\rvert^3 + \ln\lvert\Phi\rvert)}{\min(\eta_{\min}^3, \epsilon^2)}$	No
BRIEE (this paper)	$\frac{\lvert\mathcal{Z}\rvert^8 H^9 \lvert\mathcal{A}\rvert^{14} \ln\lvert\Phi\rvert}{\epsilon^4}$	Yes

(a)



(b)

Figure 1. **(a)** Sample complexities of different approaches to learning an ϵ-optimal policy for Block MDPs on a given reward function. Some approaches are reward-free and we only count the sample complexity of exploration in these cases. For MOFFLE and HOMER, $\eta_{\min}$ is the minimum probability of reaching any latent state and can be arbitrarily small or even zero in the worst case. Υ in REP-UCB refers to the function class for modeling the emission distribution since their approach is model-based, which is an additional inductive bias. We omit all log factors other than those depending on function class sizes. Our bound for OLIVE is a factor of $|\mathcal{Z}||\mathcal{A}|$ larger than that of (Jiang et al., 2017) as we include the uniform convergence over the linear function class $\{w^\top\phi : \|w\|_2 \leq 1, \phi \in \Phi\}$, where $\phi \in \mathbb{R}^{|\mathcal{Z}||\mathcal{A}|}$, to capture the $Q^\star$ function in a Block MDP. **(b)** Empirical evaluation of BRIEE against baselines in a challenging block MDP (see Section 5 for details), showing the number of episodes required to find a near-optimal solution for varying horizon lengths. Large marker indicates that the method fails to solve the problem (within the sample budget) for larger horizon. HOMER and ours can solve the problem up to $H = 100$, but our method consistently outperforms HOMER.

improved sample efficiency.

We summarize our theoretical results in Figure 1(a). Note that our strong empirical performance beyond Block MDPs (e.g., low-rank MDPs) suggests that it might be possible to extend the theoretical analysis to more general settings, and we leave this as an important direction for future work.

1.1. Related Work

We survey some of the relevant literature here.

RL with function approximation. There has been considerable progress on sample-efficient RL with function approximation in recent years. While some of it focuses on the linear case (e.g. (Jin et al., 2020; Yang & Wang, 2019)) which does not involve representation learning, other works have developed information-theoretically efficient methods for non-linear function approximation (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Jin et al., 2021), some of which subsume our setup in this paper. Particularly relevant is OLIVE (Jiang et al., 2017) which can solve Block MDPs in a model-free manner and has a better sample complexity than BRIEE. However, it is known to be computationally intractable (Dann et al., 2018) even for tabular MDPs.

Low-rank MDPs Low-rank MDP is *strictly more general* than linear MDPs which assume representation is known a priori. There are several related papers come from the recent literature on provable representation learning for low-rank MDPs (Agarwal et al., 2020b; Modi et al., 2021; Uehara et al., 2021; Ren et al., 2021). Low-rank MDPs generalize Block MDPs, so these algorithms are applicable in our setting. Of these, however, only Modi et al. (2021) handles the model-free case, while the other approaches are model-based and pay a significant sample complexity overhead in modeling the generative process of the observations. The model-free approach of Modi et al. (2021) is the closest to our work and we build on some of their algorithmic and analysis ideas. However, their work makes a significantly stronger assumption that each latent state must be reachable with at least constant probability and do not interleave exploration and exploitation. They also do not provide any empirical validation of their approach. Taking a slightly different approach, Sekhari et al. (2021) studies low-rank MDP from an agnostic policy-based perspective and only requires a policy class that not necessarily contains the optimal policy (thus the name agnostic). However, they show that exponential sample complexity in this setting is not avoidable, which indeed justifies the need for using function approximation to capture representations. Zhang et al. (2021) and Papini et al. (2021) also refer to their settings as *representation learning*, their goal is to choose the most efficient representation among a set of correct representations (i.e., every representation still linearizes the transition), which is stronger than the typical notion of feature learning from an arbitrary function class.

Block MDPs There are two prior results on sample-efficient and practical learning in Block MDPs (Du et al., 2019; Misra et al., 2019). Du et al. (2019) requires the start state to be deterministic, makes the reachability assumption on latent states, and their sample complexity has an undesir-

able polynomial dependency on the failure probability $1/\delta$. Misra et al. (2019) removes the deterministic start state assumption but still requires reachability. Both approaches are tailor-made for Block MDPs. These approaches also have to learn in a layer-by-layer forward fashion, which is not ideal in practice (i.e., they cannot learn stationary policies for episodic infinite horizon discounted setting, while our approach can be extended straightforwardly). In contrast, BRIEE learns in all layers simultaneously. (Zhang et al., 2020) extends the Block MDP to multi-task learning and study how the error from a *given* state abstraction affects the multi-task performance, but do not theoretically study how to learn such an abstraction. Feng et al. (2020) assume a high level oracle that can decode latent states. Foster et al. (2021) focus on instance-dependent bounds, but their bounds scale with a *value function disagreement coefficient* and inverse value gap, both of which can be arbitrarily large in general Block MDPs (e.g., disagreement coefficient is a stronger notation than the usual classic notation of uniform convergence which is what we use here). Finally, Duan et al. (2019) and Ni et al. (2021) study state abstraction learning from logged data, without identifying the optimal policy.

Approaches from the empirical literature There are exploration techniques with non-linear function approximation from the deep reinforcement learning literature (e.g. (Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2018; Machado et al., 2020; Sekar et al., 2020). Of these, we include the RND approach of Burda et al. (2018) in our empirical evaluation. The use of adversarial discriminators for feature learning is somewhat related to the insights in Bellemare et al. (2019), but unlike our approach, they use random adversaries in the empirical evaluation, and do not focus on strategic exploration and data collection.

2. Preliminaries

We consider a finite horizon episodic Markov Decision Process $\langle \mathcal{S}, \mathcal{A}, \{r_h\}_{h=0}^{H-1}, H, \{P_h\}_{h=0}^{H-1}, d_0 \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, P_h, r_h are the transition and reward at time step $h \in [H]$, H being the episode length; $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. For normalization, we assume the trajectory cumulative reward is bounded in $[0, 1]$.

An MDP is called a **low-rank MDP** (Rendle et al., 2010; Yao et al., 2014; Jiang et al., 2017) if the transition matrix at any time step h is low-rank, i.e., there exist two mappings $\mu_h^\star : \mathcal{S} \mapsto \mathbb{R}^d$, and $\phi_h^\star : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, such that for any s, a, s', we have $P_h^\star(s'|s,a) = \mu_h^\star(s')^\top \phi_h^\star(s,a)$. Denote d the rank of $P_h^\star$. Note that for low-rank MDP, *neither* $\mu_h^\star$ *nor* $\phi_h^\star$ *are known*, which is fundamentally different from the linear MDP model (Jin et al., 2020; Yang & Wang, 2019) where $\phi_h^\star$ is known. Learning in low-rank MDPs requires either directly learning a near-optimal policy through general function approximation, or doing representation learning first (again through nonlinear function approximation), followed by linear techniques. Either way, low-rank MDPs provide an expressive framework for analyzing non-linear function approximation in RL.

In this work, we mainly focus on analyzing a special case of low-rank MDPs, called **Block MDPs** (Du et al., 2019; Misra et al., 2020). We denote $\mathcal{Z}$ as a latent state space where $|\mathcal{Z}|$ is small. Denote $\mathcal{Z} \times \mathcal{A}$ as a joint space whose size is $|\mathcal{Z}||\mathcal{A}|$. In a Block MDP, each state s in generated from a unique latent state z as described below (hence the name block), which means that the latent state is decodable by just looking at the state. Denote the (unknown) ground truth mapping from s to the corresponding z as $\psi_h^\star : \mathcal{S} \mapsto \mathcal{Z}$ for all h. A Block MDP is formally defined as follows.

Definition 2.1 (Block MDP)**.** Consider any $h \in [H]$. A Block MDP has an emission distribution $o_h(\cdot|z) \in \Delta(\mathcal{S})$ and a latent state space transition $T_h(z'|z,a)$, such that for any $s \in \mathcal{S}$, $o_h(s|z) > 0$ for a unique $z \in \mathcal{Z}$ denoted as $\psi_h^\star(s)$. Together with the ground truth decoder $\psi_h^\star$, it defines the transitions $P_h^\star(s'|s,a) = \sum_{z' \in \mathcal{Z}} o_h(s'|z') T_h(z'|\psi_h^\star(s), a)$.

The Block MDP structure allows us to model the setting where the states s are high dimensional rich observations (e.g., images) and the state space $\mathcal{S}$ is exponentially large or even infinite. In the rest of the paper, we use words *state* and *observation* interchangeably for s with the impression that s is a high dimensional object from an extremely large space $\mathcal{S}$. Block MDPs are generalized by the low-rank MDP model. Denote the ground truth feature vector $\phi_h^\star(s,a)$ at step h as a $|\mathcal{Z}||\mathcal{A}|$-dimensional vector $e_{(\psi^\star(s),a)}$ where e_i is the i_{th} basis vector, so that it is non-zero only in the coordinate corresponding to $(\psi_h^\star(s), a) \in \mathcal{Z} \times \mathcal{A}$. Correspondingly, for any $s \in \mathcal{S}$, $\mu_h^\star(s)$ is a $|\mathcal{Z}||\mathcal{A}|$ dimensional vector such that the $(z,a)_{th}$ entry is $\sum_{z' \in \mathcal{Z}} o_h(s|z') T_h(z'|z,a)$. Then $P_h^\star(s'|s,a) = \mu_h^\star(s')^\top \phi_h^\star(s,a)$, so that the Block MDP is a low-rank MDP with rank $d = |\mathcal{Z}||\mathcal{A}|$. We assume that the reward function $r_h(s,a)$ is known.

Function approximation Our representation learning approach to learn Block MDPs requires a feature class $\{\Phi_h\}_{h=0}^{H-1}$. Since the features $\phi_h^\star$ are one-hot in a Block MDP as described above, it is natural to use the same structure in the class Φ_h as well, since it yields statistical and algorithmic advantages as we will explain in the sequel. So any $\phi_h \in \Phi_h$ is parameterized by a candidate decoder $\psi_h : \mathcal{S} \to \mathcal{Z}$ that aims to approximate $\psi_h^\star$, with $\phi_h(s,a) = e_{(\psi_h(s),a)}$. In algorithm and analysis, we will mostly work with the state-action representation class Φ_h directly, but discuss the benefits of the specific Block MDP structure when important. This is because we intend to

make our algorithm as general as possible and indeed as we will see, our algorithm can be directly applied to low-rank MDP, although our analysis only focuses on Block MDPs.

We aim to learn a near optimal policy with sample complexity scaling polynomially with respect to $|\mathcal{Z}|, |\mathcal{A}|, H$, and the statistical complexity of Φ_h rather than the size of the state space $|\mathcal{S}|$ which could be infinite here. In this work, we will focus our analysis on the setting of finite Φ_h and thus the statistical complexity Φ_h is simply $\log(|\Phi_h|)$. Extending to continuous Φ_h is straightforward by using statistical complexities such as covering number, since our analysis only uses the standard uniform convergence property on Φ_h.

Model-free vs model-based Like HOMER (Misra et al., 2020), we only model $\phi^\star$ via function approximation (hence model-free), while FLAMBE and Rep-UCB (Agarwal et al., 2020b; Uehara et al., 2021) additionally model omission distributions $o(\cdot|z)$ (i.e., model-based), which could be complex when observations are high dimensional.

Notation We denote $\pi = \{\pi_0, \ldots, \pi_{H-1}\}$ as the non-stationary Markovian policy, where each π_h maps from a state s to a distribution over actions $\Delta(\mathcal{A})$, and $V_h^\pi(s)$ as the value function of π at time step h, i.e., $V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r_h | \pi, s_h = s\right]$. We denote $Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h^\star(s,a)} V_{h+1}^\pi(s')$. We denote $V_{P,r}^\pi \in \mathbb{R}^+$ as the expected total reward of π under non-stationary transitions $P := \{P_h\}_h$ and rewards $r := \{r_h\}_h$.

We define $d_h^\pi(s,a)$ as the probability of π visiting a state-action pair (s,a) at time step h. We abuse notation a bit and denote $d_h^\pi(s)$ as the marginalized state distribution, i.e., $d_h^\pi(s) = \sum_a d_h^\pi(s,a)$. Given d_h^π, we denote $s \sim d_h^\pi$ as sampling a state at time step h from d_h^π, which can be done by executing π for $h-1$ steps starting from $h=0$. We denote $U(\mathcal{A})$ as a uniform distribution over action space $\mathcal{A}$. For a vector x and a PSD matrix Σ, we denote $\|x\|_\Sigma^2 = x^\top \Sigma x$. For $n \in \mathbb{N}^+$, we use $[n] = \{0, 1, \ldots, n-1\}$. Lastly, we denote $|\Phi| := \max_{h \in [H]} |\Phi_h|$, and $a \wedge b = \min(a,b)$.

3. Our Algorithm

In this section, we present our algorithm BRIEE: *Block-structured Representation learning with Interleaved Explore Exploit*. We first give an overview of our algorithm and then describe how to perform representation learning.

Algorithm Overview Algorithm 1 operates in an episodic setting. In episode n, we use the latest policy $\hat{\pi}^{n-1}$ to collect new data for every time step h. Note that in our data collection scheme, for each time step h, we maintain two replay buffers $\mathcal{D}_h$ and $\mathcal{D}'_h$ of transitions (s,a,s') which draw the state s from slightly different distributions (line 4). With

Algorithm 1 Block-structured Representation learning with Interleaved Explore Exploit (BRIEE)

1: **Input:** Representation classes $\{\Phi_h\}_{h=0}^{H-1}$, discriminator classes $\{\mathcal{F}_h\}_{h=0}^{H-1}$, parameters $N, T_n, \alpha_n, \lambda_n$
2: Initialize policy $\hat{\pi}^0 = \{\pi_0, \ldots, \pi_{H-1}\}$ arbitrarily and replay buffers $\mathcal{D}_h = \emptyset, \mathcal{D}'_h = \emptyset$ for all h
3: **for** $n = 1 \to N$ **do**
4: Data collection from $\hat{\pi}^{n-1}$: $\forall h \in [H]$,
$s \sim d_h^{\hat{\pi}^{n-1}}, a \sim U(\mathcal{A}), s' \sim P_h^\star(s,a)$;
$\tilde{s} \sim d_{h-1}^{\hat{\pi}^{n-1}}, \tilde{a} \sim U(\mathcal{A}), \tilde{s}' \sim P_{h-1}^\star(\tilde{s}, \tilde{a})$,
$\tilde{a}' \sim U(\mathcal{A}), \tilde{s}'' \sim P_h^\star(\tilde{s}', \tilde{a}')$
$\mathcal{D}_h = \mathcal{D}_h \cup \{s, a, s'\}$ and $\mathcal{D}'_h = \mathcal{D}'_h \cup \{\tilde{s}', \tilde{a}', \tilde{s}''\}$.
5: Learn representations for all $h \in [H]$:
$\hat{\phi}_h^n = \text{REPLEARN}(\mathcal{D}_h \cup \mathcal{D}'_h, \Phi_h, \mathcal{F}_h, \lambda_n, T_n, \ell_n)$
6: Define exploration bonus for all $h \in [H]$:
$\hat{b}_h^n(s,a) := \min\left\{\alpha_n \sqrt{\hat{\phi}_h^n(s,a)^\top \Sigma_h^{-1} \hat{\phi}_h^n(s,a)}, 2\right\}$,
with $\Sigma_h := \sum_{s,a,s' \sim \mathcal{D}_h} \hat{\phi}_h^n(s,a) \hat{\phi}_h^n(s,a)^\top + \lambda_n I$.
7: Set $\hat{\pi}^n$ as the policy returned by:
$\text{LSVI}(\{r_h + \hat{b}_h^n\}_{h=0}^{H-1}, \{\hat{\phi}_h^n\}_{h=0}^{H-1}, \{\mathcal{D}_h \cup \mathcal{D}'_h\}_{h=0}^{H-1}, \lambda_n)$.
8: **end for**
9: **Return** $\hat{\pi}^0, \ldots, \hat{\pi}^N$

$\mathcal{D}_h$ and $\mathcal{D}'_h$, we update the representation $\hat{\phi}_h$ for time step h by calling our REPLEARN oracle which is described in Algorithm 2. We then formulate the linear-bandit and linear MDP style bonus $\hat{b}_h^n(s,a)$ using the latest representation $\hat{\phi}_h^n$. Note that the bonus is constructed using only the replay buffer $\mathcal{D}_h$. When the features $\hat{\phi}_h^n$ are one-hot for a Block MDP, the first term inside the minimum in the bonus definition (line 6) simplifies to $\alpha_n / \sqrt{\lambda_n + N_h^n(\hat{\psi}_h^n(s), a)}$, where $\hat{\psi}_h^n(s)$ is the estimated latent state for s corresponding to the index of the non-zero entry in $\hat{\phi}_h^n(s,a)$ and $N_h^n(\hat{\psi}_h^n(s), a)$ is the number of times we observe a transition $(\bar{s}, \bar{a}, \bar{s}')$ in $\mathcal{D}_h$ with $\hat{\psi}_h^n(\bar{s}) = \hat{\psi}_h^n(s)$ and $\bar{a} = a$. With bonus $\hat{b}_h^n$, the representation $\hat{\phi}_h^n$, and the dataset $\mathcal{D}_h + \mathcal{D}'_h$, we use the standard Least Square Value Iteration (LSVI) (Algorithm 3) to update our policy to π^n using the combined reward $r_h + \hat{b}_h^n$.

Our algorithm is conceptually simple: it resembles the UCB style LSVI algorithm designed for linear MDPs where the ground truth features $\phi^\star$ are known. However, since $\phi^\star$ is unknown, we additionally update the representation $\hat{\phi}_h^n$ in every episode. Note that if the features $\hat{\phi}_h^n$ are one-hot, we can alternatively use the counts $N(\hat{\psi}_h^n(s), a)$ to estimate a tabular transition model over the inferred latent states and do tabular value iteration when the rewards only depend on the latent states. We choose to use the more general LSVI approach as it keeps our algorithm more general and we will comment more on this aspect at the end of this section.

Algorithm 2 Representation Learning Oracle (REPLEARN)

1: **Input:** Dataset $\mathcal{D} = \{s, a, s'\}$, representation class Φ, discriminator class $\mathcal{F}$, regularization λ, iterations T, termination threshold ℓ.
2: Initialize $\phi^0 \in \Phi$ arbitrarily
3: Denote least squares loss:
$\mathcal{L}_{\lambda,\mathcal{D}}(\phi, w, f) := \mathbb{E}_{\mathcal{D}}\left(w^\top \phi(s,a) - f(s')\right)^2 + \lambda \|w\|_2^2$.
4: **for** $t = 0 \to T-1$ **do**
5: Discriminator selection:
 # find a discriminator that cannot be linearly predicted by the current features
$f^t = \text{argmax}_f \max_{\tilde{\phi} \in \Phi} \Big[\min_w \left[\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f)\right] - \min_{\tilde{w}} \left[\mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}, \tilde{w}, f)\right] \Big]$
6: **If** f^t achieves an objective value at most ℓ: **Break** and **Return** ϕ^t.
7: Feature selection via Least Square minimization:
 # find a feature map that can linearly predict all discriminators' values at next states
$\phi^{t+1} = \text{argmin}_{\phi \in \Phi} \min_{\{w_i\}_{i=0}^t} \sum_{i=0}^t \mathcal{L}_{\lambda,\mathcal{D}}(\phi, w_i, f^i)$.
8: **end for**

Representation Learning Now we explain our representation learning algorithm (Algorithm 2). This representation learning oracle follows the algorithm from MOFFLE (Modi et al., 2021). For completeness, we explain the intuition of the representation learning oracle here. Given a dataset $\mathcal{D} = \{(s, a, s')\}$, Algorithm 2 aims to learn a representation via adversarial training using the following *ideal objective*:

$$\min_{\phi \in \Phi_h} \max_{f \in \mathcal{F}_h} \left[\min_w \mathbb{E}_{s,a \in \mathcal{D}} \left(w^\top \phi(s,a) - \mathbb{E}_{s' \sim P^\star(s,a)} f(s')\right)^2 \right]$$

where $\mathcal{F}_h \subset [\mathcal{S} \to \mathbb{R}]$ are the discriminators. In Section 4, we instantiate $\mathcal{F}_h$ as a class of linear functions on top of the representations in Φ_{h+1}. To understand the intuition here, first note that regardless of f, $\mathbb{E}_{s' \sim P_h^\star(s,a)} f(s')$ is always a linear function with respect to the ground truth features $\phi_h^\star$ (see e.g. Proposition 2.3 in Jin et al. (2020)). Hence, $\phi_h^\star$ is always a minimizer of the objective above for any class $\mathcal{F}_h$, and by using a sufficiently rich class $\mathcal{F}_h$, we hope that any other approximate optimum is also a good approximation to $\phi_h^\star$ under the same distribution.

However, the conditional expectation inside the squared loss in our ideal objective precludes easy optimization, or even direct unbiased estimation from samples, related to the "double sampling" issue in Bellman Residual objectives from the policy evaluation literature. Following a standard approach from offline RL (Antos et al., 2008) also used in MOFFLE, we instead rewrite the ideal objective with an additional term to remove the residual variance:

$$\min_{\phi \in \Phi_h} \max_{f \in \mathcal{F}_h} \Big[\min_w \mathbb{E}_{\mathcal{D}} (w^\top \phi(s,a) - f(s'))^2 - \min_{\tilde{w}, \tilde{\phi} \in \Phi_h} \mathbb{E}_{\mathcal{D}} (\tilde{w}^\top \tilde{\phi} - f(s'))^2 \Big]. \quad (1)$$

Algorithm 2 optimizes Eq. (1) through alternating updates over f and ϕ. At each iteration, it first picks features that can linearly capture expectations of all the discriminators found so far by solving a least squares problem (line 7). If ϕ is a parametric function, we can *solve the least square regression problem via gradient descent* on the parameters of ϕ and w_i. Given the latest representation ϕ^t, we simply search for a discriminator which cannot be linearly predicted by ϕ^t for any choice of weights (line 5). If no such discriminator can be found (line 6), then the current features ϕ^t are near optimal and the algorithm terminates and returns ϕ^t. For a Block MDP, our analysis shows this process will terminate in polynomial number of rounds. Note that in line 5, since we use ridge linear regression for w and $\tilde{w}$, both w and $\tilde{w}$ have closed form solutions given ϕ^t and $\tilde{\phi}$. Thus, if f and $\tilde{\phi}$ are parameterized functions, we can compute the gradient of the objective function, which allows us to *directly optimize f and $\tilde{\phi}$ jointly via gradient ascent*.

Extensions and Computation Note that our algorithm is stated in a general way that does not use any Block MDP structures. This means that the algorithm can be applied to any low-rank MDP directly. While our theoretical results only hold for Block MDPs, our experimental results indicate that the algorithm can work for more general low-rank MDPs. Note that all prior Block MDP algorithms cannot be directly applied to low-rank MDPs, and use very different approaches than general low-rank MDP learning algorithms.

Another benefit of the algorithm being not tailored to the Block MDP structure is that when we apply our algorithm to linear MDPs where $\phi_h^\star$ are known a priori, we can simply set $\Phi_h = \{\phi_h^\star\}$ as a singleton, rendering it as efficient as LSVI-UCB in this setting.

Compared to more general approaches such as OLIVE, our main benefits are algorithmic and computational. Note that OLIVE is provably intractable for even a tabular or linear MDP (Dann et al., 2018), due to complicated version space constraints. Empirically, the constraints in OLIVE are unsuitable for gradient-based approaches due to the presence of indicator functions involving the parameters.

4. Analysis

We start by specifying the discriminator class $\mathcal{F}_h$ constructed using the representations from Φ_{h+1}. Define two sets of discriminators $\mathcal{F}_h^{(1)}$ and $\mathcal{F}_h^{(2)}$ as

$$\Big\{ f(s) : \mathbb{E}_{a \sim U(\mathcal{A})}[\phi(s,a)^\top \theta - \phi'(s,a)^\top \theta'] \mid \phi, \phi' \in \Phi_{h+1} \ \max(\|\theta\|_\infty, \|\theta'\|_\infty) \le 1 \Big\}, \text{ and}$$

$$\Big\{ f(s) : \max_a \left(\frac{r_{h+1}(s,a) + w^\top \phi(s,a) \wedge 2}{2H+1} + \phi(s,a)^\top w' \right) \mid \phi \in \Phi_{h+1};\ \|w\|_\infty \le c, \|w'\|_\infty \le 1 \Big\},$$

Algorithm 3 Least Square Value Iteration LSVI

Input: Reward $\{r_h(s,a)\}_{h=0}^{H-1}$, features $\{\phi_h\}_{h=0}^{H-1}$, datasets $\{\mathcal{D}_h\}_{h=0}^{H-1}$, regularization λ
Initialize $V_H(s) = 0, \forall s$
for $h = H-1 \to 0$ **do**
  $\Sigma_h = \sum_{s,a,s' \in \mathcal{D}_h} \phi_h(s,a)\phi_h(s,a)^\top + \lambda I$
  $w_h = \Sigma_h^{-1} \sum_{s,a,s' \in \mathcal{D}_h} \phi_h(s,a) V_{h+1}(s')$.
  Set $Q_h(s,a) = w_h^\top \phi_h(s,a) + r_h(s,a)$, and $V_h(s) = \max_a Q_h(s,a)$
  Set $\pi_h(s) = \operatorname{argmax}_a Q_h(s,a)$
end for
Return $\pi := \{\pi_0, \cdots, \pi_{H-1}\}$

respectively, where $c \in \mathbb{R}^+$ is some positive constant that we will specify in the main theorem. We let our discriminator class be $\mathcal{F}_h = \mathcal{F}_h^{(1)} \cup \mathcal{F}_h^{(2)}$. Note that $\mathcal{F}_h$ contains linear functions of the features in Φ_{h+1} which makes bounding the statistical complexity (e.g., covering number) of $\mathcal{F}_h$ straightforward. Using this definition of $\mathcal{F}_h$ in Algorithm 1, we have the following guarantee when the environment is a Block MDP:

Theorem 4.1 (PAC bound of BRIEE)**.** *Consider a Block MDP (Definition 2.1) and assume that $\phi_h^\star \in \Phi_h$ for all $h \in [H]$. Fix $\delta, \epsilon \in (0,1)$, and let $\hat{\pi}$ be a uniform mixture of $\hat{\pi}^0, ..., \hat{\pi}^{N-1}$. By setting the parameters as*

$$\alpha_n = \tilde{\Theta}\Big((nd^5)^{\frac{1}{4}} |\mathcal{A}| \ln \frac{|\Phi| n}{\delta}\Big),\ \lambda_n = \Theta\Big(d \ln \frac{|\Phi| n}{\delta}\Big),$$

$$T_n = \left\lceil \sqrt{\frac{n}{d \ln(\frac{|\Phi|}{\delta})}} \right\rceil, \ell_n = \Theta\left(d^2 \sqrt{\frac{\ln(\frac{|\Phi|}{\delta})}{n}}\right),\ c = \frac{\alpha_N}{\sqrt{\lambda_N}}$$

with probability at least $1-\delta$, we have $V_{P^\star}^{\pi^\star} - V_{P^\star}^{\hat{\pi}} \le \epsilon$, after at most

$$H \cdot N \le O\left(\frac{H^9 |\mathcal{A}|^{14} |\mathcal{Z}|^8 \ln(H|\mathcal{A}||\mathcal{Z}||\Phi|/\delta\epsilon)}{\epsilon^4}\right)$$

episodes of interaction with the environment.

Theorem 4.1 certifies a polynomial sample complexity of BRIEE for all Block MDPs. Furthermore, the bound on T means that our overall computational complexity is favorable as long as one iteration of REPLEARN can be efficiently executed. Compared to other efficient methods for solving Block MDPs, our result is fully general and places no additional assumptions on the reachability or stochasticity in the underlying dynamics. Compared with statistically superior methods like OLIVE, our algorithm is amenable to practical implementation as we demonstrate through our experiments. Finally, compared with prior works which are primarily model-based, our model-free guarantees do not require modeling the emission process of observations in a Block MDP. Compared with the prior model-free representation learning guarantee of Modi et al. (2021), we do not require reachability of latent states.

Our next result shows that our BRIEE framework is flexible enough that a variation of the algorithm can solve the reward-free exploration problem. To perform reward-free exploration, we simply run BRIEE by setting reward $r_h(s,a) = 0, \forall h, s, a$. After the exploration phase, once we are given a new reward function that is linear in the ground truth feature $\phi^\star$, we can find a near optimal policy via LSVI. The detailed procedure is described in Algorithm 4 in appendix.

Theorem 4.2 (Reward-free Exploration)**.** *Consider a Block MDP and assume that $\phi_h^\star \in \Phi_h$ for all $h \in [H]$. Fix $\delta, \epsilon \in (0,1)$, by setting the parameters as in Theorem 4.1, given N episodes in the exploration phase, with probability at least $1-\delta$, for all reward function $r(s,a)$ that is linear in $\phi^\star$, Algorithm 4 finds a policy $\hat{\pi}$ in the planning phase such that*

$$V_{P^\star, r}^{\pi^\star} - V_{P^\star, r}^{\hat{\pi}} \le O\left(H^{5/2} |\mathcal{A}|^{1/2} d \log(|\Phi|/\delta)^{1/4} N^{-1/4}\right).$$

4.1. The Representation Learning Oracle's Guarantee

The following lemma gives a bound on the reconstruction error with the learned features at any iteration of BRIEE, under the data distribution used for representation learning.

Lemma 4.3 (REPLEARN guarantee)**.** *Consider parameters defined in Theorem 4.1 and time step h. Denote ρ as the joint distribution for (s,a,s') in the dataset $\mathcal{D}$ of size n. Let $\hat{\phi} = \text{REPLEARN}(\mathcal{D}, \Phi_h, \mathcal{F}_h, \lambda_n, T_n, \ell_n)$. Denote $\widehat{w}_f = \operatorname{argmin}_w \sum_{s,a,s' \in \mathcal{D}} (w^\top \hat{\phi}(s,a) - f(s'))^2 + \lambda_n \|w\|_2^2$ for any $f : \mathcal{S} \to \mathbb{R}$. Then, with probability at least $1-\delta$:*

$$\max_{f \in \mathcal{F}_h} \mathbb{E}_{s,a \sim \rho}\left[\left(\hat{w}_f^\top \hat{\phi}(s,a) - \mathbb{E}_{s' \sim P_h^\star(s,a)}[f(s')]\right)^2\right] \le \zeta_n = O\left(d^2 \sqrt{\log(n|\Phi|/\delta)/n}\right).$$

Note that the guarantee above holds under the data distribution ρ, which might not be sufficiently exploratory at intermediate iterations of the algorithm. However, it shows that our representation is always good for the entire set of discriminators on the available data distribution, and any deficiencies of the representation can only be addressed by improving the data coverage in subsequent iterations. As mentioned earlier, the number of iterations T of REPLEARN before this guarantee is achieved is bounded by the setting of T in Theorem 4.1.

Indeed the main challenge in the proof of Lemma 4.3 is to establish that REPLEARN terminates in a small number of iterations T. While this need not be true for a general MDP, the low-rank property which implies that

$\mathbb{E}[f(s')|s,a] = w_f^{*\top}\phi^\star(s,a)$ allows us to achieve this. We do so by following a similar elliptical potential argument proposed in MOFFLE. The main difference here is that we need to incorporate ridge regularization to make it consistent with the ridge regression used in LSVI. Note that ridge linear regression is important in implementation as well since it has closed-form solution which enables direct gradient ascent optimization on discriminators. While the key ideas follow from MOFFLE, there are technical differences in our analysis, as well as room to additionally leverage the Block MDP structure, which results in an improved bound. To contrast, MOFFLE requires $\tilde{O}(d^{7/3}n^{-1/3})$ samples (c.f. Lemma 5 in Modi et al. (2021)).

4.2. Proof Sketch

We now give a high-level proof sketch for Theorem 4.1. Even though our algorithm is model-free, it would be helpful to leverage an equivalent non-parametric model-based interpretation which has also been discussed before in Parr et al. (2008); Jin et al. (2020); Lykouris et al. (2021). We describe this model-based perspective for completeness here, before discussing how we utilize it.

A Model-based Perspective: Let us focus on an iteration n. For notational simplicity, we drop the superscript n here. Denote the learned features as $\hat{\phi}_h$ for $h \in [H]$. Define the non-parametric model $\hat{P}_h$ as follows:

$$\hat{P}_h(s'|s,a) = \hat{\phi}_h(s,a)^\top \Sigma_h^{-1} \sum_{\tilde{s},\tilde{a},\tilde{s}' \in \mathcal{D}_h \cup \mathcal{D}'_h} \hat{\phi}_h(\tilde{s},\tilde{a})\mathbf{1}(s' = \tilde{s}') = \frac{N_{\mathcal{D}_h \cup \mathcal{D}'_h}(\hat{\phi}(s,a), s')}{N_{\mathcal{D}_h \cup \mathcal{D}'_h}(\hat{\phi}(s,a)) + \lambda},$$

where $\Sigma_h := \sum_{s,a \in \mathcal{D}_h \cup \mathcal{D}'_h} \hat{\phi}_h(s,a)\hat{\phi}_h(s,a)^\top + \lambda I$, $\mathbf{1}(s = s')$ is the indicator function, and $N_{\mathcal{D}}(x)$ is the number of occurrences of x in the dataset $\mathcal{D}$. The intuition behind this model is that we learn the model via multivariate linear regression from representation $\hat{\phi}_h(s,a)$ to the vector $e_{s'}$. In particular, the conditional expectation $\mathbb{E}_{\hat{P}_h}[f(s')|s,a]$ is $w^\top \hat{\phi}_h(s,a)$, where w minimizes $\sum_{(s,a,s') \in \mathcal{D}_h \cup \mathcal{D}'_h} (w^\top \hat{\phi}_h(s,a) - f(s'))^2$, underscoring the connection between this model definition and our LSVI planner. The second equality is specific to the block structure of our features and does not hold in general low-rank MDPs. Interestingly, this is perhaps the only ingredient of our analysis which does not generalize beyond Block MDPs. Notice that $\hat{P}_h$ is not a normalized conditional distribution, in the sense that $\sum_{s' \in \mathcal{S}} \hat{P}_h(s'|s,a) \neq 1$ for $\lambda > 0$. Nevertheless, it is still non-negative and we can still define the occupancy measures and value functions inductively as follows:

$$d^\pi_{\hat{P};0}(s,a) = d_0(s)\pi(a|s); \quad V^\pi_{(\hat{P},r);H}(s) = 0;$$

$$d^\pi_{\hat{P};h+1}(s,a) = \sum_{\tilde{s},\tilde{a}} d^\pi_{\hat{P};h}(\tilde{s},\tilde{a})\hat{P}_h(s|\tilde{s},\tilde{a})\pi(a|s);$$

$$V^\pi_{(\hat{P},r);h}(s) = \mathbb{E}_{a\sim\pi(s)}\left[r(s,a) + \hat{P}_h(\cdot|s,a)^\top V^\pi_{(\hat{P},r);h+1}\right].$$

where $P(\cdot|s,a)^\top f$ is in short of $\sum_{s' \in \mathcal{S}} P(s'|s,a)f(s')$. As we will see, these constructions are sufficient for us to carry out a model-based analysis on our model-free algorithm.

Using our constructed representation $\{\hat{\phi}^n_h\}_{h\in[H]}$ and bonus $\{\hat{b}^n_h\}_{h\in[H]}$, we can prove the following near-*optimism* claim. Note that the optimism only holds at the initial state distribution, in contrast to the stronger versions that hold in a point-wise manner. Note that from here, our analysis significantly departs from the prior Block MDP works' analysis and the analysis of MOFFLE which rely on a reachability assumption. This part of our proof leverages some ideas in the analysis of a recent model-based representation learning algorithm REP-UCB (Uehara et al., 2021).

For $\hat{\pi} := \text{LSVI}\left(\{r_h + \hat{b}_h\}, \{\hat{\phi}_h\}, \{\mathcal{D}_h + \mathcal{D}'_h\}, \lambda_n\right)$, it is not hard to see that $\hat{\pi}$ is indeed the optimal policy for the MDP model with transitions $\{\hat{P}_h\}$ and rewards $\{r_h + \hat{b}_h\}$, i.e., $\hat{\pi} = \text{argmax}_{\pi'} V^{\pi'}_{\hat{P},r+\hat{b}}$ from our construction of $\hat{P}_h$. The following lemma shows that $V^\pi_{\hat{P},r+\hat{b}}$ almost upper bounds $V^{\pi^\star}_{P^\star,r}$, i.e., we achieve almost optimism.

Lemma 4.4 (Optimism)**.** *Using the settings of Theorem 4.1, with probability $1-\delta$, we have for all iterations $n \in [N]$, $V^{\hat{\pi}^n}_{\hat{P}^n,r+\hat{b}^n} - V^{\pi^\star}_{P^\star,r} \geq -\sigma_n$, where $\sigma_n = O(|\mathcal{A}|^{1/2}d\,(\log(n|\Phi|/\delta)/n)^{1/4})$.*

Optimism allows us to upper bound policy regret as follows. For $\hat{\pi}^n$, conditioned on the optimism event and via the standard simulation lemma (which also works on the unnormalized transitions $\hat{P}^n$), we have:

$$V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}^n}_{P^\star,r} \leq V^{\hat{\pi}^n}_{\hat{P}^n,r+\hat{b}^n} - V^{\hat{\pi}^n}_{P^\star,r} + \sigma_n \tag{2}$$
$$= \sum_{h=0}^{H-1} \mathbb{E}_{d^{\hat{\pi}^n}_h}[\hat{b}^n_h(s,a) + \underbrace{(\hat{P}^n_h - P^\star_h)(s,a)^\top \hat{V}^{\hat{\pi}^n}_{h+1}}_{\xi^n_h(s,a)}] + \sigma_n$$

where we denote $\hat{V}^{\hat{\pi}^n}_{h+1}(\cdot)$ as the expected total reward function of $\hat{\pi}^n$ starting from time step $h+1$, under model $\hat{P}^n$ and reward $r + \hat{b}^n$.

The rest of the proof is to directly control the expectation of $\hat{b}^n_h + \xi^n_h$ under $d^{\hat{\pi}^n}_h$, which is done in Lemma A.2. Lemma A.2 uses a key property of the low-rank MDP which is that for any f, $\mathbb{E}_{s,a\sim d^\pi_h} f(s,a)$ can be written in a bilinear form:

$$\left\langle \mathbb{E}_{\tilde{s},\tilde{a}\sim d^\pi_{h-1}} \phi^\star_{h-1}(\tilde{s},\tilde{a}), \int_s \mu^\star_{h-1}(s)\mathbb{E}_{a\sim\pi(s)} f(s,a)ds \right\rangle.$$

This "one step back" trick (i.e., moving from h to $h-1$) leverages the bilinear structure in $P^\star_{h-1}$, and was first

used in Agarwal et al. (2020b). Denote $\gamma_h^n(s,a) = \sum_{i=0}^{n-1} d_h^{\pi^i}(s,a)/n$ as the mixture state-action distribution, and $\Sigma_{\gamma_h^n,\phi^\star} = n\mathbb{E}_{s,a\sim\gamma_h^n}\phi_h^\star(s,a)(\phi_h^\star(s,a))^\top + \lambda I$ as the regularized covariance matrix under the ground truth representation $\phi^\star$. We can further upper bound the above bilinear form via Cauchy-Schwart under norm induced by $\Sigma_{\gamma_h^n,\phi^\star}$:

$$\left\langle \mathbb{E}_{\tilde{s},\tilde{a}\sim d_{h-1}^\pi}\phi_{h-1}^\star(\tilde{s},\tilde{a}), \int_s \mu_{h-1}^\star(s)\mathbb{E}_{a\sim\pi(s)}f(s,a)ds \right\rangle$$
$$\leq \mathbb{E}_{\tilde{s},\tilde{a}\sim d_{h-1}^\pi} \left\|\phi_{h-1}^\star(\tilde{s},\tilde{a})\right\|_{\Sigma^{-1}_{\gamma^n_{h-1},\phi^\star}}$$
$$\times \left\| \int_s \mu_{h-1}^\star(s)\mathbb{E}_{a\sim\pi(s)}f(s,a)ds \right\|_{\Sigma^n_{\gamma^n_{h-1},\phi^\star}}.$$

The first term on the RHS of the above inequality is related to the classic elliptical potential function defined with the true $\phi^\star$. The second term on the RHS can be further converted to the expectation of f^2 under the training data distribution (i.e., the distribution over $\mathcal{D}_h$), which is controllable when $f := \hat{b}_h^n + \xi_h^n$. This step of distribution change from d^π to the training data distribution is the key to directly upper bounding regret. The detailed proof for Theorem 4.1 can be found in Appendix A.

5. Experiments

We test BRIEE on MDPs motivated by the rich observation combination lock benchmark created by (Misra et al., 2019), which contains latent states but the observed states are high-dimensional continuous observations. We ask:

1. Is BRIEE more sample efficient than HOMER on their own benchmark?
2. Can BRIEE solve an MDP where the block structure does not hold, e.g., low-rank MDP?
3. If the MDP happens to have dense rewards that make it easy for policy optimization, can BRIEE leverage that?

In short, our experiments provide affirmative answers to all three questions above. Note that prior algorithms such as HOMER and PCID cannot be applied to low-rank MDPs, and are not able to leverage dense reward structures, since they require a reward-free exploration phase. In what follows we describe the experiment setup.

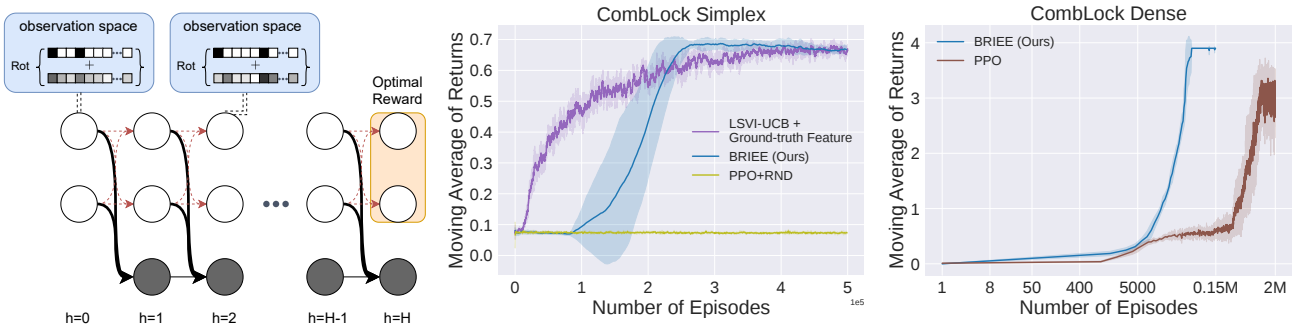
Reproducibility Our code can be find at `https://github.com/yudasong/briee`. We include experiment details in Appendix E.

The Environment We evaluate our algorithm on the *diabolical combination lock* (comblock) problem (Fig. 2a), which has horizon H and 10 actions. At each step h, there are three latent states $z_{i;h}$ for $i \in \{0,1,2\}$. We call $z_{i;h}$ for $i \in \{0,1\}$ good states and $z_{2;h}$ bad states. For each $z_{i;h}$ with $i \in \{0,1\}$, we randomly pick an action $a^*_{i;h}$ from the 10 actions. While at $z_{i;h}$ for $i \in \{0,1\}$, taking action $a^*_{i;h}$ transits the agent to $z_{0;h+1}$ and $z_{1;h+1}$ with equal probability. Taking other actions transits the agent to $z_{2;h+1}$ deterministically. At $z_{2;h}$, regardless what action the agent takes, it will transit to $z_{2;h+1}$. For reward function, we give reward 1 at state $z_{i;H}$ for $i \in \{0,1\}$, i.e., good states at H have reward 1. With probability 0.5, the agent will also receive an anti-shaped reward 0.1 from transiting from a good state to a bad state. We have reward zero for any other states and transitions. The observation s has dimension $2^{\lceil \log(H+4)\rceil}$, created by concatenating the one-hot vectors of latent state z and the one-hot vectors of horizon h, followed by adding noise sampled from $\mathcal{N}(0, 0.1)$ on each dimension, appending 0 if necessary, and multiplying with a Hadamard matrix. The initial state distribution is uniform over $z_{i;0}, i \in \{0,1\}$. Note that the optimal policy picks the special action $a^*_{i;h}$ at every h. Once the agent hits a bad state, it will stay in bad states for the entire episode, missing the large reward signal at the end. This is an extremely challenging exploration problem, since a uniform random policy will only have 10^{-H} probability of hitting the goals.

BRIEE Implementation Here we provide the details for implementing Algorithm 2. For features ϕ_h we have: $\phi_h(s,a) = \text{softmax}(A_h s/\tau) \otimes \mathbf{1}_a$, where $A_h \in \mathbb{R}^{3\times \dim_s}$, τ is the temperature, and $\mathbf{1}_a \in \{0,1\}^{|\mathcal{A}|}$ is the one-hot indicator vector. This design of decoder follows from HOMER, for the purpose of a fair comparison. For discriminator we use two-layer neural network with tanh activation. In each step of line 5, we perform gradient ascent on $\tilde{\phi}$ and f jointly. Similarly for line 7, we perform gradient descent on A_h.

Baselines In the following experiments, in addition to HOMER, we compare with empirical deep RL baselines Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Random Network Distillation (PPO+RND) (Burda et al., 2018). We also include LSVI-UCB (Jin et al., 2020) with ground-truth features (i.e., it is an aspirational baseline with access to the latent state information), and LSVI-UCB with Random Fourier Features, i.e., RFF directly on top of (s,a) (Rahimi & Recht, 2007) (LSVI-UCB+RFF) as baselines.

Comparison with HOMER In this experiment we focus on the comblock environment. We test BRIEE and baselines for different horizon values. We record the number of episodes that each algorithm needs to identify the optimal policy (i.e., a policy that achieves the optimal total reward 1). The results are shown in Fig. 1a. We compare with HOMER, PPO+RND and LSVI-UCB+RFF. We reuse the results of HOMER and PPO-RND from (Misra et al., 2019). For BRIEE, we run 5 random seeds and plot the confidence



(a) Rich-obs Comblock w/ sparse reward. (b) Comblock with simplex features. (c) Comblock dense reward.

Figure 2. **(a):** Visualization of comblock. The blue area denotes the (infinitely many) possible observations from one latent state. The top vector denotes the one-hot vector and the bottom vector denotes the noise. Rot$\{\cdot\}$ denotes multiplying with the Hadamard matrix. The dark arrows denote transiting from good states (white) to bad states (gray). Once the agent transits to a bad state, it stays in bad states for the entire episode thus missing the goal. Note that in order to reach the goal (orange), one has to stay in good states which can only be achieved by picking the right action (red) at every time step (thus the name *combination lock*). **(b):** Moving average of the evaluation returns for comblock with simplex feature and $H = 30$. The converged return of LSVI-UCB with ground truth feature suggests the optimal value. **(c):** Moving average of evaluation returns for comblock with dense reward and $H = 30$. The x-axis is scaled in natural log.

interval within 1 std. Note that LSVI-UCB + RFF can only solve H up to 12, indicating that simply lifting linear MDPs to RBF kernel space is not enough to capture the underlying nonlinearity in our problem. PPO+RND can solve up to $H = 25$. Both HOMER and BRIEE solve up to $H = 100$ with BRIEE being more sample efficient.

Visualization of decoders We visualize the outputs of decoders in Fig. 3 in Appendix E.2. In short, our decoders successfully decode latent states (up to a permutation).

Comblock with Simplex Feature Here we extend the above environment beyond Block MDP. Instead of decoding s to a unique latent state, we modify the ground truth decoder to make it stochastic. Given s, a, the ground truth decoder maps s to a distribution over latent state space, then a latent state is sampled from this distribution, and then together with a, it transits to the next latent state, followed by generating the next s' from the emission distribution. This is not a block MDP anymore, and indeed it is a low-rank MDP (i.e., $\phi^\star_h(s, a)$ is not one hot, and it is from $\Delta(\mathcal{Z} \times \mathcal{A})$). Note that HOMER provably fails in this example. We show the results for $H = 30$ in Fig. 2b. While PPO+RND completely fails, BRIEE matches the return from LSVI-UCB with the features $\phi^\star$ (that are unknown to BRIEE and PPO+RND).

Dense Reward Comblock We also test if BRIEE can leverage dense reward signals to further speed up learning. We modify the reward as follows. Instead of getting an anti-shaped reward from transition to a bad state, we get a positive reward every time one transits to a good state. Thus a greedy algorithm that collects one step immediate reward should be able to reach the final goals. In Fig. 2c we show the mean evaluation returns in the $H = 30$ dense-reward comblock environment. Compared with PPO – a greedy policy gradient algorithm, BRIEE can consistently reach the optimal value (3.9) and uses orders of magnitude fewer samples. Note that compared to the results in Fig. 1b where reward is sparse and anti-shaped, BRIEE is able to solve the problem two times faster, indicating that it indeed can leverage the dense reward structure for further speedup. Note that HOMER cannot leverage such informative reward signals and will have to perform reward-free exploration regardless, which is a huge waste of samples.

6. Conclusion

In this paper, we present a new algorithm BRIEE that provably solves block MDPs. Unlike prior block MDP algorithms, BRIEE does not require any reachability assumption and can directly optimize the given reward function. Unlike FLAMBE and REP-UCB, BRIEE is model-free which means that it is more suitable for tasks where states are high dimensional objects. Experimentally, on the benchmarks motivated by HOMER, we show our approach is more sample efficient than HOMER and other empirical RL baselines. We also demonstrate the efficacy of our approach on a low-rank MDP where the block structure does not hold.

Acknowledgement

Xuezhou Zhang and Mengdi Wang acknowledge support by NSF grants IIS-2107304, CMMI-1653435, AFOSR grant and ONR grant 1006977. Masatoshi Uehara is partly supported by MASASON Foundation.

References

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 2020a.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 2020b.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29:1471–1479, 2016.

Bellemare, M., Dabney, W., Dadashi, R., Ali Taiga, A., Castro, P. S., Le Roux, N., Schuurmans, D., Lattimore, T., and Lyle, C. A geometric perspective on optimal representations for reinforcement learning. *Advances in neural information processing systems*, 32:4358–4369, 2019.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1429–1439, 2018.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *International Conference on Machine Learning*, 2021.

Duan, Y., Ke, T., and Wang, M. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32:4486–4495, 2019.

Feng, F., Wang, R., Yin, W., Du, S. S., and Yang, L. Provably efficient exploration for reinforcement learning using unsupervised learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *Conference on learning theory*, 2021.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 2021.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Machado, M. C., Bellemare, M. G., and Bowling, M. Count-based exploration with the successor representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5125–5133, 2020.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. *arXiv preprint arXiv:1911.05815*, 2019.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Ni, C., Zhang, A., Duan, Y., and Wang, M. Learning good state and action representations via tensor decomposition. *arXiv preprint arXiv:2105.01136*, 2021.

Papini, M., Tirinzoni, A., Pacchiano, A., Restelli, M., Lazaric, A., and Pirotta, M. Reinforcement learning in

linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34, 2021.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 752–759, 2008.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1177–1184, 2007.

Ren, T., Zhang, T., Szepesvári, C., and Dai, B. A free lunch from the noise: Provable and practical exploration for representation learning. *arXiv preprint arXiv:2111.11485*, 2021.

Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 811–820, 2010.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A., and Bachman, P. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2020.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *ICML*, 2020.

Sekhari, A., Dann, C., Mohri, M., Mansour, Y., and Sridharan, K. Agnostic reinforcement learning with low-rank mdps and rich observations. *Advances in Neural Information Processing Systems*, 34, 2021.

Sodhani, S., Meier, F., Pineau, J., and Zhang, A. Block contextual mdps for continual learning. *arXiv preprint arXiv:2110.06972*, 2021.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. 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.

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

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Provable benefits of representation learning in linear bandits. *ICLR*, 2022.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Yao, H., Szepesvári, C., Pires, B. A., and Zhang, X. Pseudo-mdps and factored linear action models. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2014.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *COLT*, 2021.

Zhang, A., Sodhani, S., Khetarpal, K., and Pineau, J. Learning robust state abstractions for hidden-parameter block mdps. In *International Conference on Learning Representations*, 2020.

Zhang, W., He, J., Zhou, D., Zhang, A., and Gu, Q. Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935*, 2021.

A. Sample Complexity Analysis

Recall that we define the non-parametric model $\hat{P}_h^n$ as follows:

$$\hat{P}_h^n(s'|s,a) = \hat{\phi}_h^n(s,a)^\top \Sigma_h^{-1} \sum_{\tilde{s},\tilde{a},\tilde{s}' \in \mathcal{D}_h \cup \mathcal{D}'_h} \hat{\phi}_h^n(\tilde{s},\tilde{a}) \mathbf{1}(s' = \tilde{s}'), \quad \Sigma_h := \Big(\sum_{s,a \in \mathcal{D}'_h + \mathcal{D}'_h} \hat{\phi}_h^n(s,a) \hat{\phi}_h^n(s,a)^\top + \lambda_n I \Big)^{-1}.$$

and we define

$$\hat{\mu}_h^n = \Sigma_h^{-1} \sum_{\tilde{s},\tilde{a},\tilde{s}' \in \mathcal{D}_h \cup \mathcal{D}'_h} \hat{\phi}_h^n(\tilde{s},\tilde{a}) \mathbf{1}(s' = \tilde{s}').$$

Throughout the appendix, we also abuse the notation $\mathbb{E}_{(s,a)\sim p(s,a)}[f(s,a)] = \int f(s,a)p(s,a)d(s,a)$ for any non-negative function $p(s,a) \geq 0$.

Recall that in the case of Block MDP where $\hat{\phi}_h^n$ are one-hot vectors, and thus $\hat{P}_h^n(\cdot|s,a)$ can be further simplified as

$$\hat{P}_h^n(s'|s,a) = \frac{N_{\mathcal{D}_h + \mathcal{D}'_h}(\hat{\phi}_h^n(s,a), s')}{N_{\mathcal{D}_h + \mathcal{D}'_h}(\hat{\phi}_h^n(s,a)) + \lambda_n}$$

where $N_{\mathcal{D}}(\hat{\phi}_h^n(s,a), s')$ denotes the number of triples $(\tilde{s},\tilde{a},\tilde{s}') \in \mathcal{D}$ such that $\hat{\phi}_h^n(\tilde{s},\tilde{a}) = \hat{\phi}_h^n(s,a)$ and $\tilde{s}' = s'$, and $N_{\mathcal{D}}(\hat{\phi}_h^n(s,a)) = \sum_{s' \in \mathcal{S}} N_{\mathcal{D}}(\hat{\phi}_h^n(s,a), s')$. Therefore, we can clearly see that $\hat{P}_h^n(s'|s,a) > 0$ and $\sum_{s' \in \mathcal{S}} \hat{P}_h^n(s'|s,a) < 1$ for $\lambda_n > 0$.

For $\hat{\pi}^n := \text{LSVI}\Big(\{r_h + \hat{b}_h^n\}_{h=0}^{H-1}, \{\hat{\phi}_h^n\}_{h=0}^{H-1}, \{\mathcal{D}_h + \mathcal{D}'_h\}_{h=0}^{H-1}, \lambda_n\Big)$, $\hat{\pi}^n$ is indeed the optimal policy for the MDP model with transitions $\{\hat{P}_h^n\}$ and rewards $\{r_h + \hat{b}_h^n\}$ (i.e., the output of Value Iteration in $\{\hat{P}_h^n, r_h + \hat{b}_h^n\}_{h=0}^{H-1}$). In this section, we will take the model-based perspective and analyze based on this fitted model $\hat{P}^n$.

We define a few mixture distributions that will be used extensively in the analysis. For any n, h, define $\rho_h^n \in \Delta(\mathcal{S} \times \mathcal{A})$ as follows:

$$\rho_h^n(s,a) = \frac{1}{n} \sum_{i=0}^{n-1} d_h^{\hat{\pi}_i}(s) U(a), \forall s,a \in \mathcal{S} \times \mathcal{A}.$$

For any $n, h \geq 1$ define β_h^n as follows:

$$\beta_h^n(s,a) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_{\tilde{s} \sim d_{h-1}^{\hat{\pi}_i}, \tilde{a} \sim U(\mathcal{A})} P^\star(s|\tilde{s},\tilde{a}) U(a).$$

For any n, h, we also define $\gamma_h^n \in \Delta(\mathcal{S} \times \mathcal{A})$ as follows:

$$\gamma_h^n(s,a) = \frac{1}{n} \sum_{i=0}^{n-1} d_h^{\hat{\pi}_i}(s,a).$$

For an iteration n, a distribution ρ and a feature ϕ, we denote the expected feature covariance as

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

which in the case of Block MDP is a diagonal matrix. Notice that the dataset $\mathcal{D}_h$ of size n is sampled from ρ_h^n and the dataset $\mathcal{D}'_h$ of size n is sampled from β_h^n. Below we focus on a particular iteration n, and drop the n superscript.

For the remainder of this section, we assume that we have learned a representation $\hat{\phi}_h$ such that the following generalization bound holds:

$$\begin{aligned} &\max_{f \in \mathcal{F}_h} \mathbb{E}_{(s,a)\sim\rho_h}[(\hat{P}_h(\cdot \mid s,a)^\top f - P_h^\star(\cdot \mid s,a)^\top f)^2] \leq \zeta_n, \forall h \in [H]. \\ &\max_{f \in \mathcal{F}_h} \mathbb{E}_{(s,a)\sim\beta_h}[(\hat{P}_h(\cdot \mid s,a)^\top f - P_h^\star(\cdot \mid s,a)^\top f)^2] \leq \zeta_n, \forall h \in [H]\backslash\{0\}. \end{aligned} \tag{3}$$

where again the discriminator class is set to $\mathcal{F} = \mathcal{F}_h^{(1)} \bigcup \mathcal{F}_h^{(2)}$.

$$\mathcal{F}_h^{(1)} = \left\{ f(s) := \mathbb{E}_{a\sim U(\mathcal{A})} \left[\phi(s,a)^\top \theta - \phi'(s,a)^\top \theta'\right] \;\middle|\; \phi, \phi' \in \Phi_{h+1} \;\; \max(\|\theta\|_\infty, \|\theta'\|_\infty) \le 1 \right\}, \tag{4}$$

$$\mathcal{F}_h^{(2)} = \left\{ f(s) := \max_a \left(\frac{r_{h+1}(s,a) + \min\left\{w^\top \phi(s,a), 2\right\}}{2H+1} + {w'}^\top \phi(s,a) \right) \middle| \phi \in \Phi_{h+1};\; \|w\|_\infty \le c, \|w'\|_\infty \le 1 \right\}$$

In the analysis below, we actually need $\mathcal{F}_h$ to capture the follow two forms of function:

$$f^{(1)}(s) = \mathbb{E}_{a\sim U(\mathcal{A})} \left[\mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] \right] \tag{5}$$

$$f^{(2)}(s) = \frac{1}{2H+1} V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+1}(s) = \frac{1}{2H+1} \left(\max_a r_{h+1}(s,a) + \hat{b}_{h+1}(s,a) + \hat{\phi}_{h+1}^\top(s,a) \int \hat{\mu}_{h+1}(s') V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+2}(s') \mathrm{d}s' \right) \tag{6}$$

$f^{(1)}(s)$ is naturally captured by $\mathcal{F}^{(1)}(s)$ since $V^{\pi^\star}_{P^\star,r,h+1}(s'_h)$ is bounded in $[0,1]$, and by Lemma D.1, the expectation of any bounded function under $\hat{P}_h$ (resp. P^*_h) is a linear function of $\hat{\phi}_h$ (resp. ϕ^*_h). For $f^{(2)}(s)$, the last term is captured by ${w'}^\top \phi(s,a)$ in $\mathcal{F}^{(2)}$, because $V^{\pi^n}_{\hat{P},r+b,h+2}(s)$ is bounded in $[0, 2H]$. For the bonus term, recall that due to the Block MDP structure, the bonus takes the form of

$$\hat{b}^n_h(s,a) = \alpha_n \sqrt{\hat{\phi}^n_h(s,a)^\top \Sigma_h^{-1} \hat{\phi}^n_h(s,a)} = \hat{\phi}^n_h(s,a) \cdot \sqrt{\frac{\alpha_n^2}{N(\hat{\phi}^n_h(s,a)) + \lambda_n}}$$

which is in fact linear in $\hat{\phi}^n_h(s,a)$ and can be captured by the 2nd term in $\mathcal{F}_h^{(2)}$, with $c = \frac{\alpha_N}{\sqrt{\lambda_N}}$. Note that $\frac{\alpha_N}{\sqrt{\lambda_N}} \ge \frac{\alpha_n}{\sqrt{\lambda_n}}$ for all $n \in [N]$.

To begin with, we establish two forms of one-step-back tricks that are central to our analysis. They are of close resemblance to the one-step-back lemmas in REP-UCB (Uehara et al., 2021).

Lemma A.1 (One-step-back inequality in the learned model)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, s.t. $\|g_h\|_\infty \le B$ and $\mathbb{E}_{a\sim U(\mathcal{A})} g_{h+1}(s,a) \in \mathcal{F}_h$ for all $h \in [H]$. We condition on the event where the* REPLEARN *guarantee (3) holds, where $\mathcal{F}_h$ are defined as in Eq. (4). Then, we have for any policy π,*

$$\sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d^\pi_{\hat{P},h}}[g_h(s,a)] \le \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{P},h}} \|\hat{\phi}_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} \cdot \sqrt{n|\mathcal{A}|^2 \mathbb{E}_{(s,a)\sim\beta_{h+1}}\left[g^2_{h+1}(s,a)\right] + B^2\lambda_n d + n|\mathcal{A}|^2\zeta_n} + \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}[g_0^2(s,a)]}.$$

Proof. For step $h = 0$, we have

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{\hat{P},0}}[g_0(s,a)] =& \mathbb{E}_{s\sim d_0, a\sim\pi_0(s)}[g_0(s,a)] \\
\le& \sqrt{\max_{(s,a)} \frac{d_0(s)\pi_0(a \mid s)}{\rho_0(s,a)} \mathbb{E}_{(s,a)\sim\rho_0}\left[g_0^2(s,a)\right]} && \text{(Jensen)} \\
\le& \sqrt{\max_{(s,a)} \frac{d_0(s)\pi_0(a \mid s)}{d_0(s)u(a)} \mathbb{E}_{(s,a)\sim\rho_0}\left[g_0^2(s,a)\right]} && \text{(behavior policy has uniform action)} \\
\le& \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}\left[g_0^2(s,a)\right]}.
\end{aligned}$$

For step $h = 1, ..., H-1$, we observe the following one-step-back decomposition:

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{\hat{P},h}}[g_h(s,a)] =& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{P},h-1}, s\sim\hat{P}_{h-1}(\tilde{s},\tilde{a}), a\sim\pi_{h-1}(s)}[g_h(s,a)] \\
=& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{P},h-1}} \hat{\phi}_{h-1}(\tilde{s},\tilde{a})^\top \int \sum_a \hat{\mu}_{h-1}(s)\pi_{h-1}(a \mid s) g_h(s,a) \mathrm{d}s
\end{aligned}$$

$$\leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P},h-1}} \|\hat{\phi}_{h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_{h-1},\hat{\phi}_{h-1}}} \left\| \int \sum_a \hat{\mu}_{h-1}(s)\pi_{h-1}(a \mid s) g_h(s,a) d(s) \right\|_{\Sigma_{\rho_{h-1},\hat{\phi}_{h-1}}}.$$

For any h,

$$\begin{aligned}
&\left\| \int \sum_a \hat{\mu}_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\|^2_{\Sigma_{\rho_h,\hat{\phi}_h}} \\
=&\left\{ \int \sum_a \hat{\mu}_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\}^{\top} \left\{ n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h}[\hat{\phi}_h \hat{\phi}_h^{\top}] + \lambda_n I \right\} \left\{ \int \sum_a \hat{\mu}_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\} \\
\leq& n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h} \left\{ \left[\int \sum_a \hat{\mu}_h(s)^{\top} \hat{\phi}_h(\tilde{s},\tilde{a}) \pi_h(a \mid s) g_{h+1}(s,a) d(s) \right]^2 \right\} + B^2 \lambda_n d \\
&\qquad \text{(Use the assumption } \|\textstyle\sum_a \pi_h(a \mid s) g_{h+1}(s,a)\|_\infty \leq B \text{ and } \textstyle\int \|\hat{\mu}_h(s) h(s) d(s)\|_2 \leq \sqrt{d} \text{ for any } h : \mathcal{S} \to [0,1].) \\
=& n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h} \left[\left\{ \mathbb{E}_{s\sim\hat{P}_h(\tilde{s},\tilde{a}), a\sim\pi_h(s)} \left[g_{h+1}(s,a) \right] \right\}^2 \right] + B^2 \lambda_n d \\
=& n|\mathcal{A}|^2 \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h} \left[\left\{ |\mathcal{A}| \mathbb{E}_{s\sim\hat{P}_h(\tilde{s},\tilde{a}), a\sim U(\mathcal{A})} \left[g_{h+1}(s,a) \right] \right\}^2 \right] + B^2 \lambda_n d && \text{(Importance Sampling)} \\
\leq& n|\mathcal{A}|^2 \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h} \left[\left\{ \mathbb{E}_{s\sim P^{\star}_h(\tilde{s},\tilde{a}), a\sim U(\mathcal{A})} \left[g_{h+1}(s,a) \right] \right\}^2 \right] + B^2 \lambda_n d + n|\mathcal{A}|^2 \zeta_n && \text{(REPLEARN: } \mathbb{E}_{a\sim U(\mathcal{A})} g_{h+1}(s,a) \in \mathcal{F}_h) \\
\leq& n|\mathcal{A}|^2 \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_h, s\sim P^{\star}_h(\tilde{s},\tilde{a}), a\sim U(\mathcal{A})} \left[g^2_{h+1}(s,a) \right] + B^2 \lambda_n d + n|\mathcal{A}|^2 \zeta_n. && \text{(Jensen)} \\
=& n|\mathcal{A}|^2 \mathbb{E}_{(s,a)\sim\beta_{h+1}} \left[g^2_{h+1}(s,a) \right] + B^2 \lambda_n d + n|\mathcal{A}|^2 \zeta_n
\end{aligned}$$

Summing the decomposition for all steps h gives the desired result. □

Lemma A.2 (One-step back inequality for the true model)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, s.t. $\|g_h\|_\infty \leq B$ for all $h \in [H]$. Then, for any policy π,*

$$\sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d^{\pi}_{P^{\star},h}} [g_h(s,a)] \leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^{\star},h}} \|\phi^{\star}_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^{\star}_h}} \cdot \sqrt{n|\mathcal{A}| \mathbb{E}_{(s,a)\sim\rho_{h+1}} \left[g^2_{h+1}(s,a) \right] + B^2 \lambda_n d} + \sqrt{|\mathcal{A}| \mathbb{E}_{(s,a)\sim\rho_0} [g^2_0(s,a)]}.$$

Proof. For step $h = 0$, we similarly have

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi}_{P^{\star},0}} [g_0(s,a)] =& \mathbb{E}_{s\sim d_0, a\sim\pi_0(s)} [g_0(s,a)] \\
\leq& \sqrt{\max_{(s,a)} \frac{d_0(s)\pi_0(a \mid s)}{\rho_0(s,a)} \mathbb{E}_{(s,a)\sim\rho_0} \left[g^2_0(s,a) \right]} && \text{(Jensen)} \\
\leq& \sqrt{\max_{(s,a)} \frac{d_0(s)\pi_0(a \mid s)}{d_0(s)u(a)} \mathbb{E}_{(s,a)\sim\rho_0} \left[g^2_0(s,a) \right]} && \text{(behavior policy has uniform action)} \\
\leq& \sqrt{|\mathcal{A}| \mathbb{E}_{(s,a)\sim\rho_0} \left[g^2_0(s,a) \right]}.
\end{aligned}$$

For step $h = 1, ..., H-1$, we observe the following one-step-back decomposition:

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi}_{P^{\star},h}} [g_h(s,a)] =& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^{\star},h-1}, s\sim P^{\star}_{h-1}(\tilde{s},\tilde{a}), a\sim\pi_{h-1}(s)} [g_h(s,a)] \\
=& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^{\star},h-1}} \phi^{\star}_{h-1}(\tilde{s},\tilde{a})^{\top} \int \sum_a \mu^{\star}_{h-1}(s)\pi_{h-1}(a \mid s) g_h(s,a) \mathrm{d}s
\end{aligned}$$

$$\leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{P^\star,h-1}} \|\phi^\star_{h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-1},\phi^\star_{h-1}}} \left\| \int \sum_a \mu^\star_{h-1}(s)\pi_{h-1}(a \mid s) g_h(s,a) d(s) \right\|_{\Sigma_{\gamma_{h-1},\phi^\star_{h-1}}}.$$

For any h,

$$\begin{aligned}
&\left\| \int \sum_a \mu^\star_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\|_{\Sigma_{\gamma_h,\phi^\star_h}} \\
\leq &\left\{ \int \sum_a \mu^\star_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\}^\top \left\{ n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h}[\phi^\star_h \phi^{\star\top}_h] + \lambda_n I \right\} \left\{ \int \sum_a \mu^\star_h(s)\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right\} \\
\leq &n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h} \left\{ \left[\int \sum_a \mu^\star_h(s)^\top \phi^\star_h(\tilde{s},\tilde{a})\pi_h(a \mid s) g_{h+1}(s,a) d(s) \right]^2 \right\} + B^2\lambda_n d \\
&\qquad \text{(Use the assumption } \|\textstyle\sum_a \pi_h(a \mid s) g(s_i,a_i)\|_\infty \leq B \text{ and } \textstyle\int \|\mu^\star_h(s)h(s)d(s)\|_2 \leq \sqrt{d} \text{ for any } h : \mathcal{S} \to [0,1].) \\
= &n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h} \left[\left\{ \mathbb{E}_{s\sim P^\star_h(\tilde{s},\tilde{a}),a\sim\pi_h(s)} \left[g_{h+1}(s,a) \right] \right\}^2 \right] + B^2\lambda_n d \\
\leq &n\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h, s\sim P^\star_h(\tilde{s},\tilde{a}), a\sim\pi_h(s)} \left[g^2_{h+1}(s,a) \right] + B^2\lambda_n d. && \text{(Jensen)} \\
\leq &n|\mathcal{A}|\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_h, s\sim P^\star_h(\tilde{s},\tilde{a}), a\sim U(\mathcal{A})} \left[g^2_{h+1}(s,a) \right] + B^2\lambda_n d && \text{(Importance Sampling)} \\
\leq &n|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_{h+1}} \left[g^2_{h+1}(s,a) \right] + B^2\lambda_n d
\end{aligned}$$

Then, the final statement is immediately concluded. □

Notice that compared to Lemma A.1, Lemma A.2 post no structural assumption on g_h other than boundedness, and does not rely on the REPLEARN guarantee. Next, we prove the almost optimism Lemma presented in Lemma 4.4, restated below.

Lemma A.3 (Almost Optimism at the Initial Distribution)**.** *Consider an episode $n (1 \leq n \leq N)$ and set*

$$\alpha_n = \sqrt{n|\mathcal{A}|^2\zeta_n + 4\lambda_n d + n\zeta_n}/c, \quad \lambda_n = O\left(d \ln(|\Phi|n/\delta)\right).$$

where c is an absolute constant. Conditioning on the event that the REPLEARN *guarantee* (3) *holds, then with probability $1-\delta$, we have for all $n \in [1, \cdots, N]$,*

$$V^{\pi^\star}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi^\star}_{P^\star, r} \geq -\sqrt{|\mathcal{A}|\zeta_n}.$$

Proof. Then, from simulation lemma (Lemma D.6), we have

$$\begin{aligned}
&V^{\pi^\star}_{\hat{P}, r+\hat{b}} - V^{\pi^\star}_{P^\star, r} \\
= &\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \left[\hat{b}_h(s_h,a_h) + \mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] \right] \\
\geq &\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \left[\min\left(c\alpha_n \|\hat{\phi}_h(s,a)\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}, 2 \right) + \mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] \right]
\end{aligned} \tag{7}$$

where in the last step, we apply Lemma D.5 to replace the empirical covariance by the population covariance and c is an absolute constant. Denote

$$f_h(s,a) = \mathbb{E}_{\hat{P}_h(s'_h|s,a)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s,a)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)]$$

Notice that we have $\|f_h(s,a)\|_\infty \leq 2$,and since $V^{\pi^\star}_{P^\star,r,h+1}(s') = \max_a r_{h+1}(s',a) + \mathbb{P}^\star_{h+1} V^\star_{P^\star,r,h+2}$ and $\mathbb{P}^\star_{h+1} V^\star_{P^\star,r,h+2}$ is linear in $\phi^\star_{h+1} \in \Phi_{h+1}$, we know $V^{\pi^\star}_{P^\star,r,h+1} \in \mathcal{F}_h$. Then, by the REPLEARN guarantee, we have

$$\mathbb{E}_{(s,a)\sim\rho_h}\left[f^2_h(s,a)\right] \leq \zeta_n, \mathbb{E}_{(s,a)\sim\beta_h}\left[f^2_h(s,a)\right] \leq \zeta_n$$

Also, since $\mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)]$ is linear in $\hat{\phi}_h$ and $\mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)]$ is linear in $\phi^\star_h$, we have $\mathbb{E}_{a\sim U(\mathcal{A})} f_h(s,a) \in \mathcal{F}_h$ as well.

Then, substituting $g_h(s,a) = f_h(s,a)$ in Lemma A.1, we have:

$$
\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{P},h}}[g_h(s,a)] \\
\le &\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi^\star}_{\hat{P},h}} \|\hat{\phi}_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} \cdot \sqrt{n|\mathcal{A}|^2 \mathbb{E}_{(s,a)\sim\beta_h}\left[g^2_{h+1}(s,a)\right] + 4\lambda_n d + n\zeta_n} + \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}[g^2_0(s,a)]} \\
\le &\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi^\star}_{\hat{P},h}} \|\hat{\phi}_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} \cdot \sqrt{n|\mathcal{A}|^2 \zeta_n + 4\lambda_n d + n\zeta_n} + \sqrt{|\mathcal{A}|\zeta_n} \\
\le &\sum_{h=0}^{H-2} c\alpha_n \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi^\star}_{\hat{P},h}} \|\hat{\phi}_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} + \sqrt{|\mathcal{A}|\zeta_n}
\end{aligned}
$$

where in the last step we denote

$$\alpha_n = \sqrt{n|\mathcal{A}|^2\zeta_n + 4\lambda_n d + n\zeta_n}/c$$

Going back to (7), we have

$$
\begin{aligned}
&V^{\pi^\star}_{\hat{P},r+b} - V^{\pi^\star}_{P^\star,r} \\
= &\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \left[\min\left(c\alpha_n \|\hat{\phi}_h(s,a)\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}, 2\right) + \mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\pi^\star}_{P^\star,r,h+1}(s'_h)]\right] \\
= &\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \left[\min\left(c\alpha_n \|\hat{\phi}_h(s,a)\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}}, 2\right)\right] - \sum_{h=0}^{H-2} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}_{\hat{P},h}} \left[\min\left(c\alpha_n \|\hat{\phi}_h(s,a)\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} + \sqrt{|\mathcal{A}|\zeta_n}, 2\right)\right] \\
\ge &-\sqrt{|\mathcal{A}|\zeta_n}
\end{aligned}
$$

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

With the above preparations, we are now ready to prove our main theorem.

Theorem A.4 (Pseudo-Regret of BRIEE)**.** *With probability* $1-\delta$*, we have*

$$\sum_{n=0}^{N-1} V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}^n}_{P^\star,r} \le O\left(H^{5/2}|\mathcal{A}|^{1/2} d \log(|\Phi|/\delta)^{1/4} N^{3/4}\right)$$

Proof. Similar to Lemma A.3, we condition on the event that the REPLEARN guarantee (3) holds, which by Theorem B.4 happens with probability $1-\delta$.

For any fixed episode n we have

$$
\begin{aligned}
&V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}^n}_{P^\star,r} \\
\le &V^{\pi^\star}_{\hat{P},r+\hat{b}} - V^{\hat{\pi}^n}_{P^\star,r} + \sqrt{|\mathcal{A}|\zeta_n} && \text{(Lemma A.3)} \\
\le &V^{\hat{\pi}^n}_{\hat{P},r+\hat{b}} - V^{\hat{\pi}^n}_{P^\star,r} + \sqrt{|\mathcal{A}|\zeta_n} && (\hat{\pi}^n = \operatorname{argmax}_\pi V^{\pi}_{\hat{P}^n,r+\hat{b}^n}) \\
= &\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} \left[\hat{b}_h(s_h,a_h) + \mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+1}(s'_h)]\right] + \sqrt{|\mathcal{A}|\zeta_n}
\end{aligned}
$$

We used the 2nd form of Simulation Lemma (Lemma D.6) in the last display. Denote

$$f_h(s,a) = \frac{1}{2H+1}\left\{\mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\hat{\pi}^n}_{\hat{P},r+\hat{b},h+1}(s'_h)]\right\}$$

Then, noting $\|\hat{b}\|_\infty \leq 2$, we have $\|V^{\hat{\pi}^n}_{\hat{P},r+b,h+1}\|_\infty \leq (2H+1)$, and $\frac{1}{2H+1}V^{\hat{\pi}^n}_{\hat{P},r+b,h+1} \in \mathcal{F}_h$. Combining this fact with the above expansion, we have

$$V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}^n}_{P^\star,r} = \underbrace{\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} \left[\hat{b}_h(s_h,a_h)\right]}_{\text{(a)}} + \underbrace{(2H+1)\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} [f_h(s_h,a_h)]}_{\text{(b)}} + \sqrt{|\mathcal{A}|\zeta_n} \tag{8}$$

First, we calculate the first term (a) in Eq. (8). Following Lemma A.2 and noting again $\|\hat{b}_h\|_\infty \leq 2$, we have

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} \left[\hat{b}_h(s_h,a_h)\right] \\
&\leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}} \sqrt{n|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_{h+1}}\left[(\hat{b}_{h+1}(s,a))^2\right] + 4\lambda_n d} + \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}[(\hat{b}_0(s,a))^2]}. \\
&\leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}} \sqrt{n|\mathcal{A}|\alpha_n^2\mathbb{E}_{(s,a)\sim\rho_{h+1}}\left[\|\hat{\phi}_{h+1}\|^2_{\Sigma^{-1}_{\rho_{h+1},\hat{\phi}_{h+1}}}\right] + 4\lambda_n d} + \sqrt{|\mathcal{A}|\alpha_0^2\mathbb{E}_{(s,a)\sim\rho_0}\left[\|\hat{\phi}_0\|^2_{\Sigma^{-1}_{\rho_0,\hat{\phi}_0}}\right]}.
\end{aligned}$$

Note that we use the fact that $B = 2$ when applying Lemma A.2. In addition, we have that for any $h \in [H]$,

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

Then,

$$\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\hat{\pi}^n}_{P^\star,h}}[b_h(s,a)] \leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\phi^\star_h}} \sqrt{|\mathcal{A}|\alpha_n^2 d + 4\lambda_n d} + \sqrt{|\mathcal{A}|\alpha_1^2 d/n}.$$

Second, we calculate the term (b) in Eq. (8). Following Lemma A.2 and noting $\|f_h(s,a)\|_\infty \leq 1$, we have

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} [f_h(s_h,a_h)] \\
&\leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}} \sqrt{n|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_{h+1}}\left[f^2_{h+1}(s,a)\right] + 4\lambda_n d} + \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}[f_0^2(s,a)]}. \\
&\leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}} \sqrt{n|\mathcal{A}|\zeta_n + 4\lambda_n d} + \sqrt{|\mathcal{A}|\zeta_n}.
\end{aligned}$$

where in the second inequality, we use $\mathbb{E}_{s,a\sim\rho_h}[f_h^2(s,a)] \leq \zeta_n$. Then, by combining the above calculation of the term (a) and term (b) in Eq. (8), we have:

$$\begin{aligned}
&V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}^n}_{P^\star,r} \\
&= \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} [b_h(s_h,a_h)] + (2H+1)\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}^n}_{P^\star,h}} [f_h(s_h,a_h)] + \sqrt{|\mathcal{A}|\zeta_n} \\
&\leq \sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}} \|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}} \sqrt{|\mathcal{A}|\alpha_n^2 d + 4\lambda_n d} + \sqrt{|\mathcal{A}|\alpha_1^2 d/n} +
\end{aligned}$$

$$(2H+1)\sum_{h=0}^{H-2}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}}\|\phi^\star_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_h,\phi^\star_h}}\sqrt{n|\mathcal{A}|\zeta_n+4\lambda_n d}+(2H+1)\sqrt{|\mathcal{A}|\zeta_n}+\sqrt{|\mathcal{A}|\zeta_n}.$$

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

$$\alpha_n=O(\sqrt{n|\mathcal{A}|^2\zeta_n+4\lambda_n d+n\zeta_n})$$

Second, recall that $\gamma^n_h(s,a)=\frac{1}{n}\sum_{i=0}^{n-1}d^{\pi_i}_h(s,a)$, and thus

$$\sum_{n=0}^{N-1}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}}\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma^n_h,\phi^\star_h}}\le\sqrt{N\sum_{n=1}^{N}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}}[\phi^\star_h(\tilde{s},\tilde{a})^\top\Sigma^{-1}_{\gamma^n_h,\phi^\star_h}\phi^\star_h(\tilde{s},\tilde{a})]}\qquad\text{(CS inequality)}$$

$$\le\sqrt{N\left(\ln\det(\sum_{n=1}^{N}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}^n}_{P^\star,h}}[\phi^\star_h(\tilde{s},\tilde{a})\phi^\star_h(\tilde{s},\tilde{a})^\top])-\ln\det(\lambda_1 I)\right)}\qquad\text{(Lemma D.2 and }\lambda_1\le\cdots\le\lambda_N\text{)}$$

$$\le\sqrt{dN\ln\left(1+\frac{N}{d\lambda_1}\right)}.\qquad\text{(Potential function bound, Lemma D.3 noting }\|\phi^\star_h(s,a)\|_2\le 1\text{ for any }(s,a).\text{)}$$

Finally, The REPLEARN guarntee gives

$$\zeta_n=O\left(d^2\sqrt{\frac{\log(dn|\Phi|/\delta)}{k}}\right)$$

Combining all of the above, we have

$$\sum_{n=1}^{N}V^{\pi^\star}_{P^\star,r}-V^{\hat{\pi}^n}_{P^\star,r}\le O\left(H^2|\mathcal{A}|^{3/2}d^2n^{3/4}\log(dn|\Phi|/\delta)^{1/4}\right)$$

This concludes the proof and gives us a sample complexity of $O\left(\frac{H^8|\mathcal{A}|^6d^8\log(d|\Phi|/\delta\epsilon)}{\epsilon^4}\right)$. □

B. Representation Learning Analysis

In this section we prove Lemma 4.3. Below we omit the superscript n and subscript h when clear from the context. Denote

$$\mathcal{L}_{\lambda,\mathcal{D}}(\phi,w,f)=\frac{1}{|\mathcal{D}|}\sum_{(s,a,s')\in\mathcal{D}}\left(\phi(s,a)^\top w-f(s')\right)^2+\frac{\lambda}{|\mathcal{D}|}\|w\|_2^2 \tag{9}$$

$$\mathcal{L}_{\mathcal{D}}(\phi,w,f)=\frac{1}{|\mathcal{D}|}\sum_{(s,a,s')\in\mathcal{D}}\left(\phi(s,a)^\top w-f(s')\right)^2 \tag{10}$$

$$\mathcal{L}_{\rho}(\phi,w,f)=\mathbb{E}_{(s,a)\sim\rho,s'\sim P^\star(\cdot|s,a)}\left(\phi(s,a)^\top w-f(s')\right)^2 \tag{11}$$

The following lemma quantifies the complexity of our discriminator class $\mathcal{F}_h$ using its covering number.

Lemma B.1 (Covering Number of $\mathcal{F}_h$)**.** *The γ-covering number of Eq. $\mathcal{F}_h$ defined in (4) is at most* $2|\Phi_{h+1}|^2\cdot\left(\frac{2\alpha_N}{\sqrt{\lambda_N}\gamma}\right)^{2d}$.

Proof. Recall that the discriminator class is defined as follows:

$$\mathcal{F}^{(1)}_h=\left\{f(s):=\mathbb{E}_{a\sim U(\mathcal{A})}\left[\phi(s,a)^\top\theta-\phi'(s,a)^\top\theta'\right]\;\middle|\;\phi,\phi'\in\Phi_{h+1}\;\max(\|\theta\|_\infty,\|\theta'\|_\infty)\le 1\right\},$$

$$\mathcal{F}^{(2)}_h=\left\{f(s):\max_a\left(\frac{r_{h+1}(s,a)+\min\left\{w^\top\phi(s,a),2\right\}}{2H+1}+{w'}^\top\phi(s,a)\right)\middle|\phi,\phi'\in\Phi_{h+1};\;\|w\|_\infty\le c,\|w'\|_\infty\le 1\right\}$$

We cover $\mathcal{F}_h^{(1)}$ and $\mathcal{F}_h^{(2)}$ separately. For $\mathcal{F}_h^{(1)}$, let Θ be an ℓ_∞-cover of the set $\{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq 1\}$ at scale γ. Then, we know $|\Theta| \leq \left(\frac{2}{\gamma}\right)^d$. Define the γ-covering set of $\mathcal{F}_h^{(1)}$ as

$$\tilde{\mathcal{F}}_h^{(1)} = \left\{ s \mapsto \left(\mathbb{E}_{a \sim U(\mathcal{A})} \left[\phi(s,a)^\top \theta - \phi'(s,a)^\top \theta' \right] \right) \mid \phi, \phi' \in \Phi_{h+1}, \theta, \theta' \in \Theta \right\}$$

Then, we have that for any $f \in \mathcal{F}_h^{(1)}$, there exists a $\tilde{f} \in \tilde{\mathcal{F}}_h^{(1)}$, s.t. $\|f - \tilde{f}\|_\infty \leq \gamma$, where we use the fact that $\phi(s,a)$ are one-hot vectors, and we have $|\tilde{\mathcal{F}}_h^{(1)}| \leq |\Phi_{h+1}|^2 \cdot \left(\frac{2}{\gamma}\right)^{2d}$.

For $\mathcal{F}_h^{(2)}$, similarly let $\mathcal{W}$ be an ℓ_∞-cover of the set $\{w \in \mathbb{R}^d : \|w\|_\infty \leq c\}$ at scale γ. Let $\mathcal{W}'$ be an ℓ_∞-cover of the set $\{w' \in \mathbb{R}^d : \|w'\|_\infty \leq 1\}$ at scale γ. Then, we know that $|\mathcal{W}'| \leq \left(\frac{2}{\gamma}\right)^d$ and $|\mathcal{W}| \leq \left(\frac{2c}{\gamma}\right)^d, c = \alpha_N / \sqrt{\lambda_N}$. Define the γ-covering set of $\mathcal{F}_h^{(2)}$ as

$$\tilde{\mathcal{F}}_h^{(2)} = \left\{ s \mapsto \max_a \left(\frac{r_{h+1}(s,a) + \min\left\{ w^\top \phi(s,a), 2 \right\}}{2H+1} + {w'}^\top \phi(s,a) \right) \;\middle|\; \phi \in \Phi_{h+1}, w \in \mathcal{W}, w' \in \mathcal{W}' \right\}$$

Then, we have that for any $g \in \mathcal{G}'$, there exists a $\tilde{g}' \in \tilde{\mathcal{G}}'$, s.t. $\|g - \tilde{g}\|_\infty \leq \gamma$, and $|\tilde{\mathcal{G}}'| \leq |\Phi_{h+1}| \cdot \left(\frac{2}{\gamma}\right)^d \left(\frac{2\alpha_N}{\sqrt{\lambda_N}\gamma}\right)^d$. So the γ-covering number of $\mathcal{F}_h$ is

$$\begin{aligned} |\tilde{\mathcal{G}}| + |\tilde{\mathcal{G}}'| &= |\Phi_{h+1}|^2 \cdot \left(\frac{2}{\gamma}\right)^{2d} + |\Phi_{h+1}| \cdot \left(\frac{2}{\gamma}\right)^d \left(\frac{2\alpha_N}{\sqrt{\lambda_N}\gamma}\right)^d \\ &\leq 2|\Phi_{h+1}|^2 \cdot \left(\frac{2\alpha_N}{\sqrt{\lambda_N}\gamma}\right)^{2d} \end{aligned}$$

where the last step is due to $\alpha_N / \sqrt{\lambda_N} \geq 1$. □

Lemma B.2 (Uniform Convergence for Square Loss)**.** *Let there be a dataset $\mathcal{D} \coloneqq \{(s_i, a_i, s_i')\}_{i=1}^k$ collected in k episodes. Denote that the data generating distribution in iteration i by d_i, and $\rho = \frac{1}{k}\sum_{i=0}^k d_i$. Note that d_i can depend on the randomness in episodes $0, ..., i-1$. For a finite feature class Φ and a discriminator class $\mathcal{F} : \mathcal{S} \to [0, L]$ with γ-covering number $\|\mathcal{F}\|_\gamma$, we will show that, with probability at least $1-\delta$:*

$$\begin{aligned} &\left| \left[\mathcal{L}_\rho(\phi, w, f) - \mathcal{L}_\rho(\phi^\star, \theta_f^\star, f) \right] - \left[\mathcal{L}_\mathcal{D}(\phi, w, f) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_f^\star, f) \right] \right| \\ \leq &\frac{1}{2}\left[\mathcal{L}_\rho(\phi, w, f) - \mathcal{L}_\rho(\phi^\star, \theta_f^\star, f) \right] + \frac{28L^2 \log(\frac{2(4k)^d \cdot |\Phi| \cdot \|\mathcal{F}\|_{L/2k}}{\delta})}{k} \end{aligned}$$

for all $\phi \in \Phi$, $\|w\|_\infty \leq L$ and $f \in \mathcal{F}$, where recall that $\phi^\star$ is the true feature and $\theta_f^\star$ is defined as $\mathbb{E}_{s' \sim P^\star(\cdot|s,a)}[f(s')|s,a] = \langle \phi^\star(s,a), \theta_f^\star \rangle$.

Proof. Note that in REPLEARN, everything is happening at a fixed time step and we drop the time step indexing for brevity. To start, we focus on a given $f \in \mathcal{F}$. We first give a high probability bound on the following deviation term:

$$\left| \mathcal{L}_\rho(\phi, w, f) - \mathcal{L}_\rho(\phi^\star, \theta_f^\star, f) - \left(\mathcal{L}_\mathcal{D}(\phi, w, f) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_f^\star, f) \right) \right|.$$

Denote $g(s_i, a_i) = \langle \phi(s_i, a_i), w \rangle$ and $g^\star(s_i, a_i) = \left\langle \phi^\star(s_i, a_i), \theta_f^\star \right\rangle$.

At episode i, let $\mathcal{F}_{i-1}$ be the σ-field generated by all the random variables over the first $k-1$ episodes, for the random variable $Y_i \coloneqq \left(g(s_i, a_i) - f(s_i')\right)^2 - \left(g^\star(s_i, a_i) - f(s_i')\right)^2$, we have:

$$\begin{aligned} \mathbb{E}[Y_i | \mathcal{F}_{i-1}] &= \mathbb{E}\left[\left(g(s_i, a_i) - f(s_i')\right)^2 - \left(g^\star(s_i, a_i) - f(s_i')\right)^2 \right] \\ &= \mathbb{E}\left[\left(g(s_i, a_i) + g^\star(s_i, a_i) - 2f(s_i')\right) \left(g(s_i, a_i) - g^\star(s_i, a_i)\right) \right] \\ &= \mathbb{E}\left[\left(g(s_i, a_i) - g^\star(s_i, a_i)\right)^2 \right]. \end{aligned}$$

Here the conditional expectation is taken according to the distribution $[d_i|\mathcal{F}_{i-1}]$. The last equality is due to the fact that

$$\mathbb{E}\left[\left(g^\star(s_i,a_i)-f(s_i')\right)\left(g(s_i,a_i)-g^\star(s_i,a_i)\right)\right]=\mathbb{E}_{s_i,a_i}\mathbb{E}_{s_i'|s_i,a_i}\left[\left(g^\star(s_i,a_i)-f(s_i')\right)\left(g(s_i,a_i)-g^\star(s_i,a_i)\right)\right]=0$$

Next, for the conditional variance of the random variable, we have:

$$\begin{aligned}\mathbb{V}[Y_i|\mathcal{F}_{i-1}]&\le\mathbb{E}\left[Y_i^2|\mathcal{F}_{i-1}\right]=\mathbb{E}\left[\left(g(s_i,a_i)+g^\star(s_i,a_i)-2f(s_i')\right)^2\left(g(s_i,a_i)-g^\star(s_i,a_i)\right)^2\right]\\&\le 16L^2\mathbb{E}\left[\left(g(s_i,a_i)-g^\star(s_i,a_i)\right)^2\right]\\&\le 16L^2\mathbb{E}[Y_i|\mathcal{F}_{i-1}].\end{aligned}$$

Noticing $Y\in[-4L^2,4L^2]$.

From here on, we use $\mathbb{E}[Y]$ to denote the conditional expectation and $\mathbb{V}[Y]$ to denote the conditional variance for all Y_i, since they are all the same. Now, applying Azuma-Bernstein's inequality on $Y_1+...+Y_k$ with respect to filtration $\{\mathcal{F}_k\}_{k\ge0}$, with probability at least $1-\delta'$, we can bound the deviation term above as:

$$\begin{aligned}&\left|\mathcal{L}_\rho(\phi,w,f)-\mathcal{L}_\rho(\phi^\star,\theta_f^\star,f)-\left(\mathcal{L}_\mathcal{D}(\phi,w,f)-\mathcal{L}_\mathcal{D}(\phi^\star,\theta_f^\star,f)\right)\right|\\&\le\sqrt{\frac{2\mathbb{V}[Y]\log\frac{2}{\delta'}}{k}}+\frac{16L^2\log\frac{2}{\delta'}}{3k}\\&\le\sqrt{\frac{32L^2\mathbb{E}[Y]\log\frac{2}{\delta'}}{k}}+\frac{16L^2\log\frac{2}{\delta'}}{3k}\end{aligned}$$

where in the last inequality is obtained by choosing $\gamma=\frac{L}{2k}$.

Further, consider a finite point-wise cover of the function class $\mathcal{G}:=\{g(s,a)=\langle\phi(s,a),w\rangle:\phi\in\Phi,\|w\|_\infty\le L\}$. Note that, with a ℓ_∞-cover $\overline{\mathcal{W}}$ of $\mathcal{W}=\{\|w\|_\infty\le L\}$ at scale γ, we have for all (s,a) and $\phi\in\Phi$, there exists $\bar w\in\overline{\mathcal{W}}$, $|\langle\phi(s,a),w-\bar w\rangle|\le\gamma$, and we have $|\mathcal{W}|=\left(\frac{2L}{\gamma}\right)^d$. Let $\tilde{\mathcal{F}}$ be a γ-covering set of $\mathcal{F}$.

Then, applying a union bound over elements in $\Phi\times\overline{\mathcal{W}}\times\tilde{\mathcal{F}}$, with probability $1-|\Phi|\cdot|\overline{\mathcal{W}}|\cdot|\tilde{\mathcal{F}}|\delta'$, for all $w\in\mathcal{W}$, $f\in\mathcal{F}$, we have:

$$\begin{aligned}&\left|\mathcal{L}_\rho(\phi,w,f)-\mathcal{L}_\rho(\phi^\star,\theta_f^*,f)-\left(\mathcal{L}_\mathcal{D}(\phi,w,f)-\mathcal{L}_\mathcal{D}(\phi^\star,\theta_f^*,f)\right)\right|\\\le&\left|\mathcal{L}_\rho(\phi,\bar w,f)-\mathcal{L}_\rho(\phi^\star,\theta_f^\star,f)-\left(\mathcal{L}_\mathcal{D}(\phi,\bar w,f)-\mathcal{L}_\mathcal{D}(\phi^\star,\theta_f^\star,f)\right)\right|+4L\gamma\\\le&\sqrt{\frac{32L^2\mathbb{E}[Y]\log\frac{2}{\delta'}}{k}}+\frac{16L^2\log\frac{2}{\delta'}}{3k}+4L\gamma\\\le&\frac{1}{2}\mathbb{E}[Y_{\bar w}]+\frac{16L^2\log\frac{2}{\delta'}}{k}+\frac{16L^2\log\frac{2}{\delta'}}{3k}+4L\gamma\\\le&\frac{1}{2}\mathbb{E}[Y_w]+2L\gamma+\frac{16L^2\log\frac{2}{\delta'}}{k}+\frac{16L^2\log\frac{2}{\delta'}}{3k}+4L\gamma\\\le&\frac{1}{2}\left(\mathcal{L}_\rho(\phi,w,f)-\mathcal{L}_\rho(\phi^\star,\theta_f^\star,f)\right)+\frac{22L^2\log\frac{2}{\delta'}}{k}+6L\gamma\\\le&\frac{1}{2}\left(\mathcal{L}_\rho(\phi,w,f)-\mathcal{L}_\rho(\phi^\star,\theta_f^\star,f)\right)+\frac{28L^2\log\frac{2}{\delta'}}{k}&\text{(setting }\gamma=L/k)\end{aligned}$$

where we add subscript to Y to distinguish $Y_{\bar w}:=\left(\langle\phi(s,a),\bar w\rangle-f(s')\right)^2-\left(g^*(s,a)-f(s')\right)^2$ from $Y_w:=\left(\langle\phi(s,a),w\rangle-f(s_i')\right)^2-\left(g^*(s_i,a_i)-f(s_i')\right)^2$.

Finally, setting $\delta=\delta'/\left(|\Phi||\overline{\mathcal{W}}||\tilde{\mathcal{F}}|\right)$, we get $\log\frac{2}{\delta'}\le\log\frac{2(4k)^d|\Phi||\tilde{\mathcal{F}}|}{\delta}$. This completes the proof. □

We will see now how the above lemma can be adapted to the regularized objective.

Below, we use $\hat w_f^t$ to denote $\operatorname{argmin}_w\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t,w,f)$, $\hat w_i^t$ a shorthand for $\hat w_{f^i}^t$, and $\theta_i^\star=\operatorname{argmin}_\theta\mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star,\theta,f^i)$.

Lemma B.3 (Deviation Bounds for Algorithm 2)**.** *Let* $\tilde{\epsilon} = \frac{56L^2 \log(\frac{2(4k)^d \cdot |\Phi| \cdot \|\mathcal{F}\|_{L/2k}}{\delta})}{k}$*. If Algorithm 2 is called with a dataset* $\mathcal{D}$ *of size* k *and terminal loss cutoff* $\ell = \frac{3}{2}\epsilon_1 + \tilde{\epsilon} + \frac{2\lambda L^2 d}{k}$*, then with probability at least* $1-\delta$*, for any* $f \in \mathcal{F} \subset (S \to [0, L])$ *and* $t \leq T$*, we have*

$$\sum_{i \leq t} \mathbb{E}_\rho \left[\left(\phi^t(s,a)^\top \hat{w}_i^t - \phi^\star(s,a)^\top \theta_i^\star \right)^2 \right] \leq t \left(\tilde{\epsilon} + \frac{2\lambda L^2 d}{k} \right)$$
$$\mathbb{E}_\rho \left[\left(\phi^t(s,a)^\top w - \phi^\star(s,a)^\top \theta_{t+1}^\star \right)^2 \right] \geq \epsilon_1, \text{ for all } w \in \mathbb{R}^d.$$

Furthermore, at termination, the learned feature ϕ^t *satisfies:*

$$\max_{f \in \mathcal{F}} \mathbb{E}_\rho \left[(\phi^{t\top} \hat{w}_f^t - \phi^{\star\top} \theta_f^\star)^2 \right] \leq 3\epsilon_1 + 3\tilde{\epsilon} + \frac{4\lambda L^2 d}{k}.$$

Proof. We begin by using the result in Lemma B.2 such that, with probability at least $1-\delta$, for all $\|w\|_\infty \leq L$, $\phi \in \Phi$ and $f \in \mathcal{F}$, we have

$$\left| \left[\mathcal{L}_\rho(\phi, w, f) - \mathcal{L}_\rho(\phi^\star, \theta_f^\star, f) \right] - \left[\mathcal{L}_\mathcal{D}(\phi, w, f) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_f^\star, f) \right] \right| \leq \frac{1}{2} \left[\mathcal{L}_\rho(\phi, w, f) - \mathcal{L}_\rho(\phi^\star, \theta_f^\star, f) \right] + \tilde{\epsilon}/2.$$

Thus, for the feature selection step in iteration t, with probability at least $1-\delta$ we have:

$$\begin{aligned}
&\sum_{i \leq t} \mathbb{E}_\rho \left[\left(\phi^{t\top} \hat{w}_i^t - \phi^{\star\top} \theta_i^\star \right)^2 \right] \\
&= \sum_{i \leq t} \left(\mathcal{L}_\rho(\phi^t, \hat{w}_i^t, f^i) - \mathcal{L}_\rho(\phi^\star, \theta_i^\star, f^i) \right) && \text{(since } \mathbb{E}_{s' \sim P^\star(s,a)} f^i = (\theta_i^\star)^\top \phi^*(s,a)\text{)} \\
&\leq \sum_{i \leq t} 2 \left(\mathcal{L}_\mathcal{D}(\phi^t, \hat{w}_i^t, f^i) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_i^\star, f^i) \right) + t\tilde{\epsilon} && \text{(Lemma B.2, and } \|\hat{w}_i^t\|_\infty \leq L \text{ by Lemma D.1)} \\
&\leq \sum_{i \leq t} 2 \left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, \hat{w}_i^t, f^i) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta_i^\star, f^i) + \frac{\lambda}{k} \|\theta_i^\star\|_2^2 \right) + t\tilde{\epsilon} \\
&\leq t \left(\tilde{\epsilon} + \frac{2\lambda L^2 d}{k} \right), && \text{(by the optimality of } \phi^t, \hat{w}_i^t \text{ under } \mathcal{L}_{\lambda,\mathcal{D}}(\cdot,\cdot,f^i)\text{, see Algorithm 2 line 7)}
\end{aligned}$$

which means the first inequality in the lemma statement holds. Here, we use $\|\theta_i^*\|_2^2 \leq L^2 d$, which is easily derived using the block MDP assumption.

For the discriminator selected at iteration t, let $\bar{w} \coloneqq \operatorname{argmin}_w \mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f^{t+1})$. Using the same sample size for the adversarial test function at each non-terminal iteration with loss cutoff ℓ, for any vector $w \in \mathbb{R}^d$, $\|w\|_\infty \leq L$ we get:

$$\begin{aligned}
&\mathbb{E}_\rho \left[\left(\phi^{t\top} w - \phi^{\star\top} \theta_{t+1}^\star \right)^2 \right] \\
&= \mathcal{L}_\rho(\phi^t, w, f^{t+1}) - \mathcal{L}_\rho(\phi^\star, \theta_{t+1}^\star, f^{t+1}) \\
&\geq \frac{2}{3} \left(\mathcal{L}_\mathcal{D}(\phi^t, w, f^{t+1}) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_{t+1}^\star, f^{t+1}) \right) - \frac{\tilde{\epsilon}}{3} && \text{(Lemma B.2)} \\
&\geq \frac{2}{3} \left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f^{t+1}) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta_{t+1}^\star, f^{t+1}) - \frac{\lambda L^2 d}{k} \right) - \frac{\tilde{\epsilon}}{3} \\
&\geq \frac{2}{3} \left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, \bar{w}, f^{t+1}) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta_{t+1}^\star, f^{t+1}) - \frac{\lambda L^2 d}{k} \right) - \frac{\tilde{\epsilon}}{3} && (\bar{w} \coloneqq \operatorname{argmin}_w \mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f^{t+1})) \\
&\geq \frac{2\ell}{3} + \frac{2}{3} \left(\min_{\tilde{\phi} \in \Phi_h, \tilde{w}} \mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}, \tilde{w}, f^{t+1}) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta_{t+1}^\star, f^{t+1}) \right) - \frac{2\lambda L^2 d}{3k} - \frac{\tilde{\epsilon}}{3} \\
&\geq \frac{2\ell}{3} + \frac{2}{3} \left(\min_{\tilde{\phi} \in \Phi_h, \tilde{w}} \mathcal{L}_\mathcal{D}(\tilde{\phi}, \tilde{w}, f^{t+1}) - \mathcal{L}_\mathcal{D}(\phi^\star, \theta_{t+1}^\star, f^{t+1}) - \frac{\lambda L^2 d}{k} \right) - \frac{2\lambda L^2 d}{3k} - \frac{\tilde{\epsilon}}{3}
\end{aligned}$$

$$\geq \frac{2\ell}{3} + \frac{1}{3}\left(\mathcal{L}(\tilde{\phi}_{t+1}, \tilde{w}_{t+1}, f^{t+1}) - \mathcal{L}(\phi^\star, \theta^\star_{t+1}, f^{t+1})\right) - \frac{4\lambda L^2 d}{3k} - \frac{2\tilde{\epsilon}}{3} \quad (\|\tilde{w}_{t+1}\|_\infty \leq L, \|\theta^*_{t+1}\|_\infty \leq L, \text{Lemma B.2})$$
$$\geq \frac{2\ell}{3} - \frac{4\lambda L^2 d}{3k} - \frac{2\tilde{\epsilon}}{3}.$$

where $(\tilde{\phi}_{t+1}, \tilde{w}_{t+1})$ denote $\operatorname{argmin}_{\tilde{\phi}\in\Phi_h, \tilde{w}} \mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}, \tilde{w}, f^{t+1})$. In the first inequality, we invoke Lemma B.2 to move to empirical losses. In the fourth inequality, we add and subtract the bias correction term along with the fact that the termination condition is not satisfied for f^{t+1}. In the next step, we again use Lemma B.2 for the bias correction term for f^{t+1}.

Thus, if we set the cutoff ℓ for test loss to $3\epsilon_1/2 + \tilde{\epsilon} + \frac{2\lambda L^2 d}{k}$, for a non-terminal iteration t, for any $w \in \mathbb{R}^d$ with $\|w\|_\infty \leq L$, we have:

$$\mathbb{E}_\rho\left[\left(\phi^{t\top} w - \phi^{\star\top}\theta^\star_{t+1}\right)^2\right] \geq \epsilon_1. \tag{12}$$

Now, since we know $\phi^{\star\top}\theta^\star_{t+1} = \mathbb{E}_{s'\sim P^\star(s,a)} f^{t+1} \leq L$, and we know the Bayes optimal solution

$$w_{\phi^t} = \operatorname*{argmin}_{w\in\mathbb{R}^d} \mathbb{E}_\rho\left[\left(\phi^{t\top} w - \phi^{\star\top}\theta^\star_{t+1}\right)^2\right]$$

satisfies $\|w_{\phi^t}\|_\infty \leq L$ by Lemma D.1. Therefore, Eq. (12) also applies to w_{ϕ^t}, and since w_{ϕ^t} is the minimizer, we have in fact for all $w \in \mathbb{R}^d$,

$$\mathbb{E}_\rho\left[\left(\phi^{t\top} w - \phi^{\star\top}\theta^\star_{t+1}\right)^2\right] \geq \epsilon_1.$$

which implies the second inequality in the lemma statement holds.

At the same time, for the last iteration t, for all $f \in \mathcal{F}$, define $\hat{w}_f = \operatorname{argmin}_w \mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f)$, then the feature ϕ^t satisfies:

$$\begin{aligned}
&\mathcal{L}_\rho(\phi^t, \hat{w}_f, f) - \mathcal{L}_\rho(\phi^\star, \theta^\star_f, f) \\
&\leq 2\left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, \hat{w}_f, f) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta^\star_f, f) + \frac{\lambda L^2 d}{k}\right) + \tilde{\epsilon} && (\text{Lemma B.2, and } \|\theta^*_f\|_\infty \leq L \text{ by Lemma D.1}) \\
&\leq 2\left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, \hat{w}_f, f) - \mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f) + \mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f) - \mathcal{L}_{\lambda,\mathcal{D}}(\phi^\star, \theta^\star_f, f)\right) + \frac{2\lambda L^2 d}{k} + \tilde{\epsilon} \\
&\leq 2\left(\mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, \hat{w}_f, f) - \mathcal{L}_{\lambda,\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f)\right) + \frac{2\lambda L^2 d}{k} + \tilde{\epsilon} && (\text{optimality of } \tilde{\phi}_f, \tilde{w}_f \text{ on } \mathcal{L}_{\lambda,\mathcal{D}}(\cdot,\cdot,f).) \\
&\leq 2\ell + \frac{2\lambda L^2 d}{k} + \tilde{\epsilon} \\
&= 3\epsilon_1 + 3\tilde{\epsilon} + \frac{4\lambda L^2 d}{k}.
\end{aligned}$$

This gives us the third inequality in the lemma, thus completes the proof. □

Theorem B.4 (Sample and Iteration Complexity of Algorithm 2)**.** *Let ϵ_1 be set to $16\sqrt{2}Ld^{3/2}\tilde{\epsilon}^{1/2}$ and the termination threshold be set to $\ell = \frac{3}{2}\epsilon_1 + \tilde{\epsilon} + \frac{2\lambda L^2 d}{k}$ as in Lemma B.3, then Algorithm 2 terminates in at most $T = \sqrt{\frac{L^2 d}{2\tilde{\epsilon}}}$ iterations, and returns a ϕ^t such that*

$$\max_{f\in\mathcal{F}} \mathbb{E}_\rho\left[(\phi^{t\top}\hat{w}_f - \phi^{\star\top}\theta^\star_f)^2\right] \leq 75Ld^{3/2}\tilde{\epsilon}^{1/2}.$$

For $\mathcal{F}_h$ defined in Lemma B.1, we have

$$\max_{f\in\mathcal{F}_h} \mathbb{E}_\rho\left[(\phi^{t\top}\hat{w}_f - \phi^{\star\top}\theta^\star_f)^2\right] \leq \zeta_n := O\left(d^2\sqrt{\frac{\log(dk|\Phi|/\delta)}{k}}\right)$$

Proof. We first incur Lemma B.3 and get that when $\ell = \frac{3}{2}\epsilon_1 + \tilde{\epsilon} + \frac{2\lambda L^2 d}{k}$,

$$\sum_{i\leq t} \mathbb{E}_\rho\left[\left(\phi^t(s,a)^\top \hat{w}^t_i - \phi^\star(s,a)^\top\theta^\star_i\right)^2\right] \leq t\left(\tilde{\epsilon} + \frac{2\lambda L^2 d}{k}\right) \leq 2t\tilde{\epsilon} \tag{13}$$

$$\mathbb{E}_\rho\left[\left(\phi^t(s,a)^\top w - \phi^\star(s,a)^\top \theta^\star_{t+1}\right)^2\right] \geq \epsilon_1 \tag{14}$$

for all w and $t \leq T$.

At round t, for functions $f^1, \ldots, f^t \in \mathcal{F}$ in Algorithm 2, let $\theta^\star_i = \theta^\star_{f^i}$ as before and further let $\Lambda_t = \sum_{i=1}^t \theta^\star_i \theta^{\star\top}_i + \lambda' I_{d\times d}$.

Further, let $W_t = [w_{t,1} \mid w_{t,2} \mid \ldots \mid w_{t,t}] \in \mathbb{R}^{d\times t}$ be the matrix with columns W^i_t as the linear parameter $\hat{w}^t_i = \operatorname{argmin}_w \mathcal{L}_{\lambda,\mathcal{D}}(\phi^t, w, f^i)$. Similarly, let $A_t = [\theta^\star_1 \mid \theta^\star_2 \mid \ldots \mid \theta^\star_t]$.

Using the linear parameter $\theta^\star_{t+1}$ of the adversarial test function f^{t+1}, define $\hat{w}_t = W_t A_t^\top \Lambda_t^{-1} \theta^\star_{t+1}$. For this $\hat{w}_t$, we can bound its norm as:

$$\|W_t A_t^\top \Lambda_t^{-1} \theta^\star_{t+1}\|_2 \leq \|W_t\|_2 \|A_t^\top \Lambda_t^{-1}\|_2 \|\theta^\star_{t+1}\|_2 \leq L^2 d \sqrt{\frac{t}{4\lambda'}}.$$

Here, $\|W_t\|_2 \leq L\sqrt{dt}$, $\|\theta^\star_{t+1}\|_2 \leq L\sqrt{d}$, and $\|A_t^\top \Lambda_t^{-1}\|_{\text{op}}$ can be shown to be less than $\sqrt{1/4\lambda'}$ [1]. From Eq. (14), we have

$$\begin{aligned}
\epsilon_1 \leq \mathbb{E}\left[\left(\phi^{t\top}\hat{w}_t - \phi^{\star\top}\theta^\star_{t+1}\right)^2\right] &= \mathbb{E}\left[\left(\hat{\phi}_t^\top W_t A_t^\top \Lambda_t^{-1}\theta^\star_{t+1} - \phi^{\star\top}\Lambda_t\Lambda_t^{-1}\theta^\star_{t+1}\right)^2\right] \\
&\leq \|\Lambda_t^{-1}\theta^\star_{t+1}\|_2^2 \cdot \mathbb{E}\left[\|\phi^{t\top} W_t A_t^\top - \phi^{\star\top}\Lambda_t\|_2^2\right] \\
&\leq 2\|\Lambda_t^{-1}\theta^\star_{t+1}\|_2^2 \cdot \mathbb{E}\left[\|\phi^{t\top} W_t A_t^\top - \phi^{t\top} A_t A_t^\top\|_2^2 + \lambda'^2\|\phi^{\star\top}\|_2^2\right] \\
&\leq 2\|\Lambda_t^{-1}\theta^\star_{t+1}\|_2^2 \cdot \left(\sigma_1^2(A_t)\mathbb{E}\left[\|\phi^{t\top} W_t - \phi^{\star\top} A_t\|_2^2\right] + \lambda'^2\right) \\
&\leq 2\|\Lambda_t^{-1}\theta^\star_{t+1}\|_2^2 \cdot \left(2L^2 dt^2\tilde{\epsilon} + \lambda'^2\right).
\end{aligned}$$

The second inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$. The last inequality applies the upper bound $\sigma_1(A_t) \leq L\sqrt{dt}$ and the guarantee from Eq. (13). Using the fact that $t \leq T$, this implies that

$$\|\Lambda_t^{-1}\theta^\star_{t+1}\|_2 \geq \sqrt{\frac{\epsilon_1}{2(2L^2 dT^2\tilde{\epsilon} + \lambda'^2)}}.$$

We now use the generalized elliptic potential lemma to upper bound the total value of $\|\Lambda_t^{-1}\theta^\star_{t+1}\|_2$. From Lemma D.4, if we set $\lambda' = L^2 d$ and we do not terminate in T rounds, then

$$T\sqrt{\frac{\epsilon_1}{2(2L^2 dT^2\tilde{\epsilon} + \lambda'^2)}} \leq \sum_{t=1}^T \|\Lambda_t^{-1}\theta^\star_{t+1}\|_2 \leq 2\sqrt{\frac{Td}{\lambda'}}.$$

From this chain of inequalities, we can deduce

$$T\epsilon_1 \leq 8(d/\lambda')\left(2L^2 dT^2\tilde{\epsilon} + \lambda'^2\right),$$

therefore

$$T \leq \frac{8d\lambda'}{\epsilon_1 - 16L^2 d^2 T\tilde{\epsilon}/\lambda'}.$$

Now, if we set $\epsilon_1 = 32L^2 d^2 T\tilde{\epsilon}/\lambda'$ in the above inequality, we can deduce that

$$T \leq \frac{\lambda'^2}{2L^2 dT\tilde{\epsilon}} \implies T \leq \sqrt{\frac{L^2 d}{2\tilde{\epsilon}}}.$$

and the proper value for ϵ_1 is $16\sqrt{2}Ld^{3/2}\tilde{\epsilon}^{1/2}$. The sample complexity then readily follows from Lemma B.3.

Combining with Lemma B.1, and noting that $\|f\|_\infty \leq 2$ for all $f \in \mathcal{F}_h$, we have that

$$\max_{f\in\mathcal{F}_h} \mathbb{E}_\rho\left[(\phi^{t\top}\hat{w}_f - \phi^{\star\top}\theta^\star_f)^2\right] \leq \zeta_n \coloneqq O\left(d^2\sqrt{\frac{\log\left(dk|\Phi|/\delta\right)}{k}}\right)$$

which gives us Lemma 4.3. □

[1] Applying SVD decomposition and the property of matrix norm, $\|A_t^\top \Lambda_t^{-1}\|_{\text{op}}$ can be upper bounded by $\max_{i\leq d} \frac{\sqrt{\lambda_i}}{\lambda_i + \lambda'} \leq \frac{1}{\sqrt{4\lambda'}}$ where λ_i are the eigenvalues of $A_t A_t^\top$ and the final inequality holds by AM-GM.

Algorithm 4 Reward-free BRIEE

Exploration Phase:

Input: Representation classes $\{\Phi_h\}_{h=0}^{H-1}$, discriminator classes $\{\mathcal{F}_h\}_{h=0}^{H-1}$, parameters $N, T_n, \alpha_n, \lambda_n$
Initialize policy $\hat{\pi}^0 = \{\pi_0, \ldots, \pi_{H-1}\}$ arbitrarily and replay buffers $\mathcal{D}_h = \emptyset, \mathcal{D}'_h = \emptyset$ for all h
for $n = 1 \to N$ **do**
  Data collection from $\hat{\pi}^{n-1}$: $\forall h \in [H]$,
    $s \sim d_h^{\hat{\pi}^{n-1}}, a \sim U(\mathcal{A}), s' \sim P_h^\star(s,a)$;
    $\tilde{s} \sim d_{h-1}^{\hat{\pi}^{n-1}}, \tilde{a} \sim U(\mathcal{A}), \tilde{s}' \sim P_{h-1}^\star(\tilde{s},\tilde{a}), \tilde{a}' \sim U(\mathcal{A}), \tilde{s}'' \sim P_h^\star(\tilde{s}',\tilde{a}')$
    $\mathcal{D}_h = \mathcal{D}_h \cup \{s, a, s'\}$ and $\mathcal{D}'_h = \mathcal{D}'_h \cup \{\tilde{s}', \tilde{a}', \tilde{s}''\}$.
  Learn representations for all $h \in [H]$:
  $\hat{\phi}_h^n = \text{REPLEARN}\left(\mathcal{D}_h \cup \mathcal{D}'_h, \Phi_h, \mathcal{F}_h, \lambda_n, T_n, \ell_n\right)$
  Define exploration bonus for all $h \in [H]$:
  $\hat{b}_h^n(s,a) := \min\left\{\alpha_n \sqrt{\hat{\phi}_h^n(s,a)^\top \Sigma_h^{-1} \hat{\phi}_h^n(s,a)}, 2\right\}$, with $\Sigma_h := \sum_{s,a,s' \sim \mathcal{D}_h} \hat{\phi}_h^n(s,a)\hat{\phi}_h^n(s,a)^\top + \lambda_n I$.
  $\left(\hat{\pi}^n, \hat{V}_0^n(s_0)\right) \leftarrow \text{LSVI}(\{\hat{b}_h^n\}_{h=0}^{H-1}, \{\hat{\phi}_h^n\}_{h=0}^{H-1}, \{\mathcal{D}_h \cup \mathcal{D}'_h\}_{h=0}^{H-1}, \lambda_n)$.
end for
$\tilde{n} = \operatorname{argmin}_n \hat{V}_0^n(s_0)$.

Planning Phase:

Input: Reward function $r_h(s,a)$.
$\hat{\pi} \leftarrow \text{LSVI}(\{r_h + \hat{b}_h^{\tilde{n}}\}_{h=0}^{H-1}, \{\hat{\phi}_h^{\tilde{n}}\}_{h=0}^{H-1}, \{\mathcal{D}_h^{\tilde{n}} \cup \mathcal{D}_h'^{\tilde{n}}\}_{h=0}^{H-1}, \lambda_{\tilde{n}})$.

C. Reward-free Exploration

Exploration Stage Run BRIEE with reward function set to zero and discriminator class set to

$$\left\{f(s)=\mathbb{E}_{a\sim U(\mathcal{A})}[\phi(s,a)^\top\theta - \phi'(s,a)^\top\theta'] \mid \phi,\phi' \in \Phi_{h+1} \ \max(\|\theta\|_\infty, \|\theta'\|_\infty)\le 1\right\} \bigcup$$

$$\left\{f(s)=\frac{\phi'(s,\pi'(s))^\top\theta + \min\{w^\top\phi(s,\pi(s)), 2\}}{2H+1} + w'^\top\phi(s,\pi(s)) \mid \phi,\phi' \in \Phi_{h+1}; \|w\|_\infty, \|w'\|_\infty, \|\theta\|_\infty \le c; \pi,\pi' \in \tilde{\Pi}\right\}$$

where

$$\tilde{\Pi} = \{f(\psi(s)) \mid f : \mathcal{Z} \to \mathcal{A}, \psi \in \Psi\}$$

Then main difference to the discriminator class in the reward-driven setting is that we replace the known reward function $r_{h+1}(s,a)$ in $\mathcal{F}^{(2)}$ with a linear term $\phi'(s,a)^\top\theta$ which covers all possible reward functions that depends on the latent states and actions, i.e., $r_{h+1}(z,a)$. We also replace $\max_a$ by a pre-defined policy class that covers all possible policies $\hat{\pi}$ that can be returned by Algorithm 4. $\tilde{\Pi}$ has cardinality $|\Phi|\cdot|\mathcal{A}|^{|\mathcal{Z}|}$. So the log-covering number of $\mathcal{F}_h$ does not change more than constant factors and Lemma 4.3 remains to hold.

We collect all historical version of the datasets as $\{\mathcal{D}_h^n\}$ and $\{\mathcal{D}_h^{n\prime}\}$, bonuses as $\{\hat{b}_h^n\}$ and learned features as $\{\phi_h^n\}$, for all $h \in [H]$ and $n \in [N]$.

Planning Stage The key observation here is that $V_{\hat{P}^n,\hat{b}^n}^{\pi^n}$ can actually observed as the returned value $\hat{V}_0^n(s_0)$ of LSVI for each iteration n (assuming a unique starting state). Therefore, we can choose $\tilde{n} = \operatorname{argmin}_n V_{\hat{P}^n,\hat{b}^n}^{\pi^n}$ and let the exploration stage output $\hat{P}^{\tilde{n}}, \hat{b}^{\tilde{n}}$.

Then, during planning stage, simply return $\hat{\pi} = \operatorname{argmax}_\pi V_{\hat{P}^{\tilde{n}}, r+\hat{b}^{\tilde{n}}}^{\pi}$, which is done by LSVI.

Lemma C.1. *For any $n \in [N]$,*

$$NV_{\hat{P}^n,\hat{b}^{\tilde{n}}}^{\hat{\pi}^{\tilde{n}}} \le \sum_{n=0}^{N-1} V_{\hat{P}^n,\hat{b}^n}^{\hat{\pi}^n} \le O\left(H^{5/2}|\mathcal{A}|^{1/2} d \log(|\Phi|/\delta)^{1/4} N^{3/4}\right).$$

Proof. This follows the exact same proof as Theorem A.4 (starting from the 3rd equation), plus noting that during exploration, since the reward is 0, $V^{\pi^n}_{P^\star,r} = 0$. □

Lemma C.2. *We have*

$$V^{\hat{\pi}}_{\hat{P}^n, r+\hat{b}^n} - V^{\hat{\pi}}_{P^\star, r} \le C \cdot V^{\hat{\pi}}_{\hat{P}^n, \hat{b}^n} + \sqrt{|\mathcal{A}|\zeta_n}$$

for an absolute constant C.

Proof. Below we drop the superscript n. First we apply simulation lemma (Lemma D.6)

$$\begin{aligned}
&V^{\hat{\pi}}_{\hat{P}, r+\hat{b}} - V^{\hat{\pi}}_{P^\star, r} \\
=&\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}}_{\hat{P},h}} \left[\hat{b}_h(s_h,a_h) + \mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\hat{\pi}}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\hat{\pi}}_{P^\star,r,h+1}(s'_h)]\right] \\
=&V^{\hat{\pi}}_{\hat{P},\hat{b}} + \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}}_{\hat{P},h}} \left[\underbrace{\mathbb{E}_{\hat{P}_h(s'_h|s_h,a_h)}[V^{\hat{\pi}}_{P^\star,r,h+1}(s'_h)] - \mathbb{E}_{P^\star_h(s'_h|s_h,a_h)}[V^{\hat{\pi}}_{P^\star,r,h+1}(s'_h)]}_{g_h(s,a)}\right]
\end{aligned}$$

We know $g_h \in \mathcal{F}_h$, $\|g_h\|_\infty \le 2$. Therefore, by the REPLEARN guarantee, we have

$$\mathbb{E}_{(s,a)\sim\rho_h}\left[g_h^2(s,a)\right] \le \zeta_n, \mathbb{E}_{(s,a)\sim\beta_h}\left[g_h^2(s,a)\right] \le \zeta_n$$

Then, we apply one-step-back trick to $\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\hat{\pi}}_{\hat{P},h}} g_h(s,a)$,

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d^{\hat{\pi}}_{\hat{P},h}}[g_h(s,a)] \\
\le&\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}}_{\hat{P},h}} \|\phi_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\phi_h}} \cdot \sqrt{n|\mathcal{A}|^2 \mathbb{E}_{(s,a)\sim\beta_h}\left[g^2_{h+1}(s,a)\right] + 4\lambda_n d + n\zeta_n} + \sqrt{|\mathcal{A}|\mathbb{E}_{(s,a)\sim\rho_0}[g_0^2(s,a)]} \\
\le&\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}}_{\hat{P},h}} \|\phi_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\phi_h}} \cdot \sqrt{n|\mathcal{A}|^2\zeta_n + 4\lambda_n d + n\zeta_n} + \sqrt{|\mathcal{A}|\zeta_n} \\
\le&\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}}_{\hat{P},h}} \alpha_n \cdot \|\phi_h(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_h,\hat{\phi}_h}} + \sqrt{|\mathcal{A}|\zeta_n} \\
\le&\sum_{h=0}^{H-2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\hat{\pi}}_{\hat{P},h}} c \cdot \hat{b}(\tilde{s},\tilde{a}) + \sqrt{|\mathcal{A}|\zeta_n} && \text{(by Lemma D.5)}
\end{aligned}$$

□

Theorem C.3 (Reward-free PAC bound)**.** *Let $\pi^\star$ be the optimal policy under $P^\star$ and r. Then,*

$$V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \le O\left(H^{5/2}|\mathcal{A}|^{1/2} d \log(|\Phi|/\delta)^{1/4} N^{-1/4}\right)$$

Proof. We have

$$\begin{aligned}
V^{\pi^\star}_{P^\star,r} - V^{\hat{\pi}}_{P^\star,r} \le& \sqrt{|\mathcal{A}|\zeta_n} + V^{\pi^\star}_{\hat{P}^{\bar{n}}, r+\hat{b}^{\bar{n}}} - V^{\hat{\pi}}_{P^\star,r} && \text{(Apply Lemma A.3 to } \pi^*) \\
\le& \sqrt{|\mathcal{A}|\zeta_n} + V^{\hat{\pi}}_{\hat{P}^{\bar{n}}, r+\hat{b}^{\bar{n}}} - V^{\hat{\pi}}_{P^\star,r} && (\pi^\star \in \tilde{\Pi} \text{ and } \hat{\pi} = \operatorname{argmax}_\pi V^{\pi}_{\hat{P}^{\bar{n}}, r+\hat{b}^{\bar{n}}}) \\
\le& 2\sqrt{|\mathcal{A}|\zeta_n} + C V^{\hat{\pi}}_{\hat{P}^{\bar{n}}, \hat{b}^{\bar{n}}} && \text{(by Lemma C.2)}
\end{aligned}$$

$$\leq 2\sqrt{|\mathcal{A}|\zeta_n} + CV^{\hat{\pi}^n}_{\hat{P}^{\tilde{n}},\hat{b}^{\tilde{n}}} \qquad (\hat{\pi}^n = \operatorname{argmax}_\pi V^\pi_{\hat{P}^n,\hat{b}^n})$$

$$\leq O\left(H^{5/2}|\mathcal{A}|^{1/2} d \log(|\Phi|/\delta)^{1/4} N^{-1/4}\right) \qquad \text{(by Lemma C.1)}$$

□

D. Auxiliary Lemmas

Lemma D.1 (bounded LSVI solution for Block MDP)**.** *For any $\phi \in \Phi$, $f : \mathcal{S} \to [0, L]$, dataset $\mathcal{D} = \{(s_i, a_i, s'_i)\}_{i=1}^n$, the ridge regression solution*

$$\hat{w}_f = \operatorname*{argmin}_w \sum_{i=1}^n (\phi(s,a)^\top w - f(s'))^2 + \lambda \|w\|_2^2 \tag{15}$$

satisfies $\|\hat{w}_f\|_\infty \leq L$.

Proof. In Block MDP, $\hat{w}_f$ takes the following closed-form

$$\hat{w}_f(z,a) = \frac{1}{N_{\mathcal{D}}(\phi(s,a)) + \lambda} \sum_{i:\phi(s_i,a_i)=(z,a)} f(s'_i) \leq \frac{N_{\mathcal{D}}(\phi(s,a))}{N_{\mathcal{D}}(\phi(s,a)) + \lambda} L \leq L$$

as needed. □

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

Lemma D.2. *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:*

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

Lemma D.3 (Potential Function Lemma)**.** *Suppose $Tr(G_n) \leq B^2$.*

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

Lemma D.4 (Generalized Elliptic Potential Lemma, Lemma 24 of (Modi et al., 2021))**.** *For any sequence of vectors $\theta_1^\star, \theta_2^\star \ldots, \theta_T^\star \in \mathbb{R}^{d\times T}$ where $\|\theta_i^\star\| \leq L\sqrt{d}$, for $\lambda \geq L^2 d$, we have:*

$$\sum_{t=1}^T \|\Sigma_t^{-1} \theta^\star_{t+1}\|_2 \leq 2\sqrt{\frac{dT}{\lambda}}.$$

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. (2021, Lemma 39). Here, we take a union bound over the whole feature $\phi \in \Phi$.

Lemma D.5 (Concentration of the Bonus)**.** *Set $\lambda_n = \Theta(d \ln(n|\Phi|/\delta))$ for any n. Let $\mathcal{D} = \{s_i, a_i\}_{i=0}^{n-1}$ be a stochastic sequence of data where $(s_i, a_i) \sim \rho_i$ where ρ_i can depend on the history of time steps $1, ..., i-1$. Let $\rho = \frac{1}{n}\sum_{i=0}^{n-1} \rho$ and define*

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

Then, with probability $1-\delta$, we have

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

Lemma D.6 (Simulation Lemma). *Given two MDPs $(P', r+b)$ and (P, r), for any policy π, we have:*

$$V^{\pi}_{P',r+b} - V^{\pi}_{P,r} = \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d^{\pi}_{P',h}} \left[b_h(s_h,a_h) + \mathbb{E}_{P'_h(s'_h|s_h,a_h)}[V^{\pi}_{P,r,h+1}(s'_h)] - \mathbb{E}_{P_h(s'_h|s_h,a_h)}[V^{\pi}_{P,r,h+1}(s'_h)] \right]$$

and

$$V^{\pi}_{P',r+b} - V^{\pi}_{P,r} = \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d^{\pi}_{P,h}} \left[b_h(s_h,a_h) + \mathbb{E}_{P'_h(s'_h|s_h,a_h)}[V^{\pi}_{P,r+b,h+1}(s'_h)] - \mathbb{E}_{P_h(s'_h|s_h,a_h)}[V^{\pi}_{P,r+b,h+1}(s'_h)] \right]$$

We note that since both the occupancy measure and Bellman updates under $\hat{P}$ are defined in the exact same way as if $\hat{P}$ is a proper probability matrix, the classic simulation lemma also applies to $\hat{P}$.

E. Experiment Details

E.1. Setup Details and Hyperparameters

CombLock environment Details For the design of the diabolical combination lock environment in our comparison with HOMER, we follow the design as in HOMER, but we present the environment hyperparameters in Table. 1 for completeness. We also provide a more detailed explanation of how we record result in Fig. 1b: after each policy update, we perform 20 i.i.d. rollouts using the latest policy and record the mean returns in these 20 evaluation runs. If one algorithm can get a mean return of 1 for 5 consecutive policy updates, we count the algorithm solving the environment and record the number of episodes it uses.

	Value
Horizon	6,12,25,50,100
Switch probability	0.5
Anti reward	0.1
Anti reward probability	0.5
Final reward	1
Number of actions	10
Observation noise std	0.1
Random seeds	1,12,123,1234,12345

Table 1. Hyperparameters for the rich observation comblock environment with sparse and anti-shaped reward.

BRIEE Implementation for CombLock Environment For the dense reward diabolical combination lock environment, we first use a random policy to collect 10000 episodes of samples before our first iteration of feature learning. We perform this sample warm-up procedure for $H = 50$ and $H = 100$ experiments only. We maintain a separate buffer for each timestep h, and in practice we mix the samples in $\mathcal{D}$ and $\mathcal{D}'$ together. For each buffer, we limit the size of the buffer to 10000 and update the buffer with first-in-first-out procedure. Between each update we rollout $50 \times H$ episodes to collect data. For the optimization method we use SGD with a momentum factor of 0.99. Finally due to a different latent state distribution, we use softmax with temperature 0.1 for ϕ_0 and a temperature of 1 for all the other timesteps. We provide the full list of hyperparameters in Table. 2.

Baseline Implementation for CombLock Environment In this section we provide the hyperparameteres we use for PPO-RND in Table. 3 and LSVI-UCB in Table. 4.[2] For the RFF feature, we choose the bandwidth with median trick.

Simplex Feature Experiment Details We first present detailed description of dynamics of the new MDP. Given a observation-action pair s, a, the simplex feature is given by $z(s) = \text{softmax}(I_3 R^{-1} s/\tau_{\text{env}})$, where $R^{-1} \in \mathbb{R}^{\dim_s \times \dim_s}$ is the inverse of the Hadamard matrix, and $I_3 \in \mathbb{R}^{3\times \dim_s}$, a matrix with a 3×3 identity matrix at its first three columns and

[2] We use the public code for PPO-RND, which is available at here.

	Value Considered	Final Value
Decoder ϕ learning rate	{1e-2, 5e-3, 1e-3}	1e-2
Discriminator f learning rate	{1e-2, 5e-3, 1e-3}	1e-2
Discriminator f hidden layer size	{256,512}	256
RepLearn Iteration T	{10,20,30}	30
Decoder ϕ number of gradient steps	{32,64,128,256}	64
Discriminator f number of gradient steps	{128,256}	128
Decoder ϕ batch size	{128,256,512}	512
Discriminator f batch size	{128,256,512}	512
RepLearn regularization coefficient λ	{1,0.1,0.01}	0.01
Decoder ϕ softmax temperature	{1,0.5,0.1}	1
Decoder ϕ_0 softmax temperature	{1,0.1}	0.1
LSVI bonus coefficient β	$\{10, \frac{H}{5}\}$	$\frac{H}{5}$
LSVI regularization coefficient λ	{1}	1
Buffer size	{1e5}	1e5
Update frequency	{50,100}	50
Warm up samples ($H = 50, 100$)	{10000}	10000

Table 2. Hyperparameters for BRIEE in sparse reward comblock experiment.

	Value
Learning rate	1e-3
Hidden layer size	64
τ_{GAE}	0.95
Gradient clipping	5.0
Entropy bonus	0.01
Clip ratio	0.2
Minibatch size	160
Optimization epoch	5
Intrinsic reward normalization	False
Intrinsic reward coefficient	1e3
Extrinsic reward coefficient	1.0

Table 3. Hyperparameters for PPO-RND in comblock sparse reward experiment.

zero everywhere else. Thus the ground truth feature is given by $\phi^* = z(s) \otimes a$. With a observation-action pair s, a, the environment first sample a latent state according to the probability simplex $z(s)$, and then transit according to the action a and the transition rules of the original comblock environment. For our results in Fig. 2b, we use $\tau_{\text{env}} = 0.2$. The moving average is cross 50 evaluations (and 20 i.i.d. rollouts per evaluation). We use the same set of hyperparameters for BRIEE as in Table. 2 and for LSVI as in Table. 4.

Dense Reward Environment Details In the dense reward environment, we remove the anti-shaped reward and the agent receives rewards while staying in the good states. We keep the final reward if the correct actions are taken in the last layer. We provide the hyperparameters of the environment in Table. 5. We provide the hyperparameters of BRIEE for the dense reward environment in Table.6 (note only the LSVI bonus coefficients are different). We provide the hyperparameters of PPO for the dense reward environment in Table.7.

E.2. Visualization of the decoder

In Figure 3, we visualize the decoder on the combination lock example (with the Block MDP structure). We run BRIEE on the Block MDP combination lock until it solves the problem (i.e., achieve the optimal total reward). Denote the learned decoders as $\hat{\psi}_h$ for all $h \in [H]$. Note that $\hat{\psi}_h$ maps from state (i.e., observation) s to a 3-dimensional vector in the simplex (since we use softmax, i.e., $\hat{\psi}(s) = \text{softmax}(A_h s/\tau)$, as defined in the implementation section). Ideally, we hope that $\hat{\psi}_h$ can output pretty deterministic distribution over 3 latent states (i.e., $\hat{\psi}_h(s)$ is close to a one-hot encoding vector), and can

	Value Considered	Final Value
LSVI bonus coefficient β	$\{10, \frac{H}{5}\}$	$\frac{H}{5}$
LSVI regularization coefficient λ	$\{1\}$	1
Buffer size	{1e5,5e5,1e6}	1e6
Update frequency	{50,100,250}	250
Kernel bandwidth	Median Trick	5
Feature dimension $H = 6$	{200}	200
Feature dimension $H = 12$	{200,500}	500
Feature dimension $H = 25$	{500,800,1000}	1000

Table 4. Hyperparameters for for LSVI-UCB with RFF feature in comblock sparse reward experiment.

	Value
Horizon	30
Switch probability	0.5
Step reward	0.1
Final reward	1
Reward probability	1
Number of actions	10
Observation noise std	0.1
Random seeds	1,12,123,1234,12345

Table 5. Hyperparameters for the rich observation comblock environment with dense reward.

decode the latent state (up to permutation).

We test the decoders as follows. For each state $z_{i;h}$ for $i \in \{0, 1, 2\}$ and $h \in [H]$, where in this section we use $H = 25$, we sample 50 observations for each state (following the emission distribution described in the environment section), and we take the average of the 50 decoded states (decoded by $\hat{\psi}_h$).

In Fig. 3, we demonstrate the decoded states. The h-th column in the i-th image denotes the average of the 50 decoded states from observations generated from $z_{i,h}$, for $i \in \{0, 1, 2\}$ and $h \in [25]$ (i.e., the image number denotes the ground truth state, the x-axis denotes the timestep, and the y-axis in each image denotes the averaged value of the decoded states on each dimension).

Interestingly, we notice that our decoder $\hat{\psi}_h$ fails to decode the two good latent states (i.e., state 0 and state 1) confidently at $h = 10$ and $h = 20$. However, this is not a failure case. The reason is that at $h = 10$ and $h = 20$, the two good latent states share the same optimal action (i.e., the action that transits the agent from a good state to the next two good states). Namely, the two good states at $h = 10$ (and $h = 20$) share the same transition. Hence, there is no need for the decoder to distinguish these two states. Note that our decoders still successfully differentiate the bad state and the two good states at $h = 10$ and 20. This phenomenon is also observed in HOMER. Also note that for $h = 0$, the decoder is only required to distinguish state 0 from state 1, because the initial distribution is uniform over state 0 and state 1 only, and assigns 0 mass to state 2 (i.e., we never reach state 2 in $h = 0$).

	Value Considered	Final Value
Decoder ϕ learning rate	{1e-2, 5e-3, 1e-3}	1e-2
Discriminator f learning rate	{1e-2, 5e-3, 1e-3}	1e-2
Discriminator f hidden layer size	{256,512}	256
RepLearn Iteration T	{10,20,30}	30
Decoder ϕ number of gradient steps	{32,64,128,256}	64
Discriminator f number of gradient steps	{128,256}	128
Decoder ϕ batch size	{128,256,512}	512
Discriminator f batch size	{128,256,512}	512
RepLearn regularization coefficient λ	{1,0.1,0.01}	0.01
Decoder ϕ softmax temperature	{1,0.5,0.1}	1
Decoder ϕ_0 softmax temperature	{1,0.1}	0.1
LSVI bonus coefficient β	$\{\frac{H}{5}, \frac{H}{50}\}$	$\frac{H}{50}$
LSVI regularization coefficient λ	{1}	1
Buffer size	{1e5}	1e5
Update frequency	{50,100}	50
Warm up samples	{0}	0

Table 6. Hyperparameters for BRIEE in sparse reward comblock experiment.

	Value Considered	Value
Learning rate	{1e-3,5e-4,1e-4}	1e-3
Hidden layer size	{64}	64
τ_{GAE}	{0.95}	0.95
Gradient clipping	{5.0}	5.0
Entropy bonus	{0.01,0.001}	0.01
Clip ratio	{0.2}	0.2
Minibatch size	{160}	160
Optimization epoch	{5}	5

Table 7. Hyperparameters for PPO in comblock sparse reward experiment.

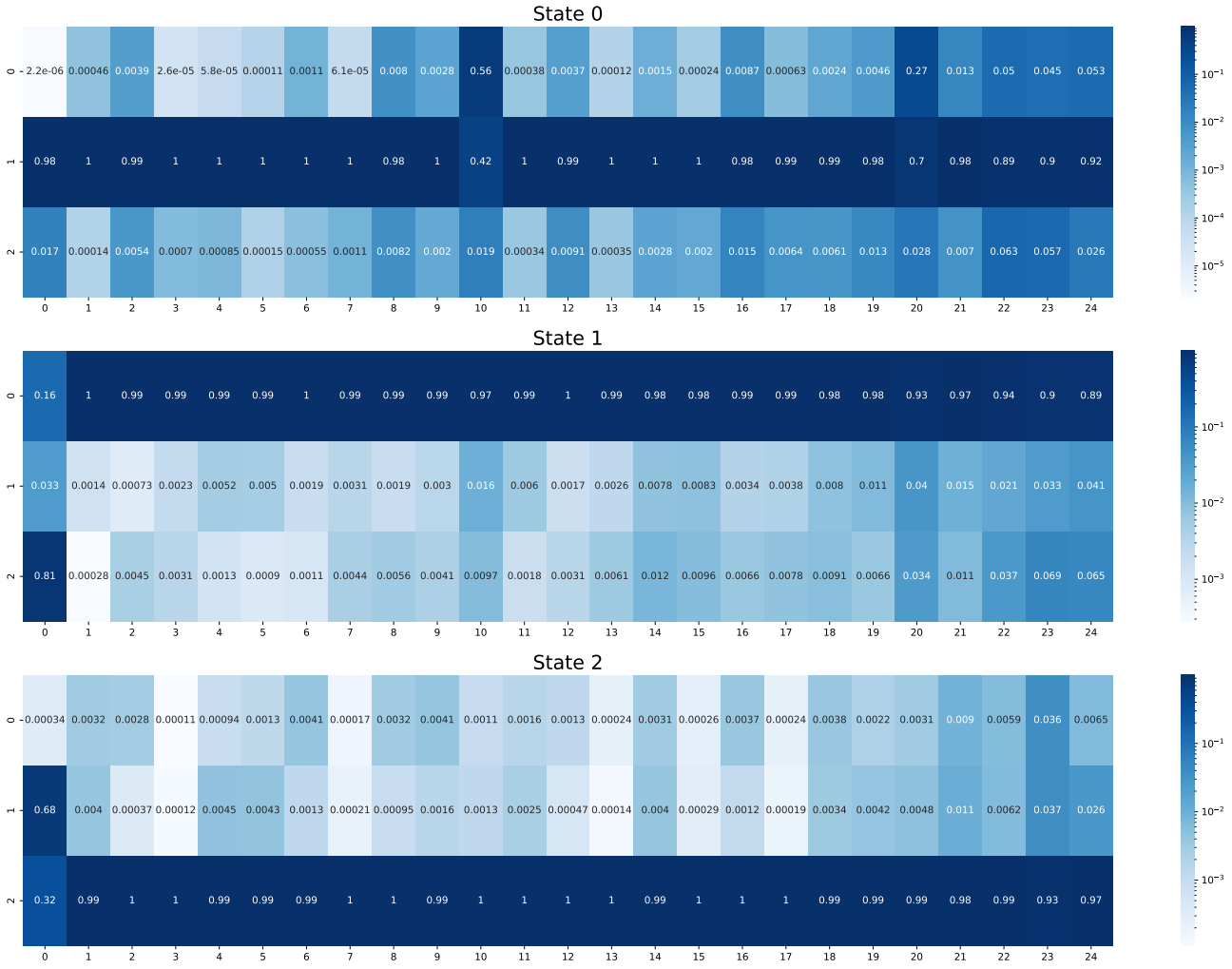


Figure 3. Visualization of the decoder. The h-th column in the i-th image denotes the averaged decoded states from the 50 observations generated by latent state $z_{i,h}$, for $i \in \{0, 1, 2\}$ and $h \in [25]$. The heatmap color transits in log scale. Note that we have near one-hot encoding in most of the levels, except $h = 10$ and $h = 20$. This is because we have the same optimal actions for the good states in $h = 10$ and $h = 20$ thus we still recover the ground-truth features (see text for more details). Also note that the optimal features are recovered up to permutation (e.g., in this case we decode state 0 as state 1 and state 1 as state 0, but this is still the optimal decoding).