Journal of Machine Learning Research 25 (2024) 1-76 Submitted 6/22; Revised 4/23; Published 1/24

Model-Free Representation Learning and Exploration in Low-Rank MDPs

Aditya Modi * ADMODI@UMICH.EDU
Microsoft
Mountain View, CA 94043, USA

Jinglin Chen * JINGLINC@ILLINOIS.EDU
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research
New York, NY 10011, USA

Nan Jiang NANJIANG@ILLINOIS.EDU
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA

Alekh Agarwal ALEKHAGARWAL@GOOGLE.COM
Google Research

Editor: Tor Lattimore

Abstract

The low-rank MDP has emerged as an important model for studying representation learning and exploration in reinforcement learning. With a known representation, several model-free exploration strategies exist. In contrast, all algorithms for the unknown representation setting are model-based, thereby requiring the ability to model the full dynamics. In this work, we present the first model-free representation learning algorithms for low-rank MDPs. The key algorithmic contribution is a new minimax representation learning objective, for which we provide variants with differing tradeoffs in their statistical and computational properties. We interleave this representation learning step with an exploration strategy to cover the state space in a reward-free manner. The resulting algorithms are provably sample efficient and can accommodate general function approximation to scale to complex environments.

Keywords: Reinforcement learning, representation learning, low-rank MDPs, sample complexity analysis, reward-free exploration

1. Introduction

A key driver of recent empirical successes in machine learning is the use of rich function classes for discovering transformations of complex data, a sub-task referred to as *representation learning*. For example, when working with images or text, it is standard to train extremely

∗. Equal contribution.

large neural networks in a self-supervised fashion on large datasets, and then fine-tune the network on supervised tasks of interest. The representation learned in the first stage is essential for sample-efficient generalization on the supervised tasks. Can we endow Reinforcement Learning (RL) agents with a similar capability to discover representations that provably enable sample efficient learning in downstream tasks?

In the empirical RL literature, representation learning often occurs implicitly simply through the use of deep neural networks, for example in DQN (Mnih et al., 2015). Recent work has also considered more explicit representation learning via auxiliary losses like inverse dynamics (Pathak et al., 2017), the use of explicit latent state space models (Hafner et al., 2019; Sekar et al., 2020), and via bisimulation metrics (Gelada et al., 2019; Zhang et al., 2020). Crucially, these explicit representations are again often trained in a way that they can be reused across a variety of related tasks, such as domains sharing the same (latent state) dynamics but differing in reward functions.

While these works demonstrate the value of representation learning in RL, theoretical understanding of such approaches is limited. Indeed obtaining sample complexity guarantees is quite subtle as recent lower bounds demonstrate that various representations are not useful or not learnable (Modi et al., 2020; Du et al., 2019b; Van Roy and Dong, 2019; Lattimore and Szepesvari, 2020; Hao et al., 2021). Despite these lower bounds, some prior theoretical works do provide sample complexity guarantees for non-linear function approximation (Jiang et al., 2017; Sun et al., 2019a; Osband and Roy, 2014; Wang et al., 2020b; Yang et al., 2020), but these approaches do not obviously enable generalization to related tasks. More direct representation learning approaches were recently studied in Du et al. (2019a); Misra et al. (2020); Agarwal et al. (2020b), who develop algorithms that provably enable sample efficient learning in any downstream task that shares the same dynamics.

Our work builds on the most general of the direct representation learning approaches, namely the Flambe algorithm of Agarwal et al. (2020b), that finds features under which the transition dynamics are nearly linear. The main limitation of Flambe is the assumption that the dynamics can be described in a parametric fashion. In contrast, we take a model-free approach to this problem, thereby accommodating much richer dynamics.

Concretely, we study the low-rank MDP, in which the transition operator $T : (x, a) \to \Delta(\mathcal{X})$ admits a low-rank factorization as $T(x' \mid x, a) = \langle \phi^*(x, a), \mu^*(x') \rangle$ for feature maps ϕ^*, μ^*. For model-free representation learning, we assume access to a function class Φ containing the underlying feature map ϕ^*. This is a much weaker inductive bias than prior work in the "known features" setting where ϕ^* is known in advance (Jin et al., 2020b; Yang and Wang, 2020; Agarwal et al., 2020a) and the model-based setting (Agarwal et al., 2020b) that assumes realizability for both μ^* and ϕ^*.

While our model-free setting captures richer MDP models, addressing the intertwined goals of representation learning and exploration is much more challenging. In particular, the forward and inverse dynamics prediction problems used in prior works are no longer admissible under our weak assumptions. Instead, we address these challenges with a new representation learning procedure based on the following insight: for any function $f : \mathcal{X} \to \mathbb{R}$, the Bellman backup of f is a linear function in the feature map ϕ^*. This leads to a natural minimax objective, where we search for a representation $\hat{\phi}$ that can linearly approximate the Bellman backup of all functions in some "discriminator" class $\mathcal{F}$. Importantly, the discriminator class $\mathcal{F}$ is induced directly by the class Φ, so no additional realizability assumptions are required.

We also provide an incremental approach for expanding the discriminator set, which leads to a more computationally practical variant of our algorithm. The two algorithms reduce to minimax optimization problems over non-linear function classes. While such problems can be solved empirically with modern deep learning libraries, they do not come with rigorous computational guarantees. To this end, we further show that when Φ is efficiently enumerable, our optimization problems can be reduced to eigenvector computations, which leads to provable computational efficiency.[1]

Paper outline and summary of contributions The organization of the rest of this paper and our main contributions and are summarized below:

- In Section 2, we formally describe the problem setting of this paper. The related work and comparison with existing literature is discussed in Section 3.
- In Section 4, we present our main algorithm MOFFLE which interleaves the exploration and representation learning components. We further describe how the learned representation can be used for planning in downstream tasks by using a standard offline planning algorithm, namely, FQI.
- In Section 5, we present our novel representation learning objective for low-rank MDPs, a min-max-min optimization problem defined using the feature class Φ. A sample complexity result is then presented for MOFFLE under a min-max-min computational oracle assumption.
- To address the computational tractability of our representation learning objective, we propose a computationally friendly iterative greedy approach in Section 6. We state a formal guarantee on the iteration complexity of this approach and show that the resulting instance of MOFFLE is provably sample efficient.
- In Section 7, we show that for the special case of enumerable feature classes, MOFFLE can be used for sample efficient reward-free exploration and that the main representation learning objective can be reduced to a computationally tractable eigenvector computation problem.
- In Section 8, we give a proof outline of our main results along with the complete proofs for each instantiation of MOFFLE and, finally, conclude in Section 9.
- The following supporting results are delegated to the appendix thereafter: (i) details and guarantees for an elliptical planning algorithm (Appendix B), (ii) deviation bounds for our main results (Appendix C), (iii) sample complexity results for FQI planning and FQE methods (Appendix D and Appendix E), and (iv) auxiliary results e.g., deviation bounds for regression with squared loss (Appendix F).
- **Key contribution** We propose the first model-free representation learning algorithms for low-rank MDPs. On the algorithmic side, we propose a minimax representation learning objective and variants which can learn a representation $\phi \in \Phi$ which is sufficiently accurate for downstream planning. In addition to this, we also analyze different approaches to solve for this objective which shows differing tradeoffs in their statistical and computational

1. For the enumerable case, our algorithm collects an exploratory dataset, which is sufficient for downstream planning but requires the entire feature class Φ. See Section 7 for more details.

properties. The representation learning step interleaves with an exploration strategy to cover the state space in a reward-free manner and provides the first provably sample-efficient algorithm for model-free and reward-free exploration for low-rank MDPs.

2. Problem Setting

We consider an episodic MDP $\mathcal{M}$ with a state space $\mathcal{X}$, a finite action space $\mathcal{A} = \{1, \ldots, K\}$ and horizon H. In each episode, an agent generates a trajectory $\tau = (x_0, a_0, x_1, \ldots, x_{H-1}, a_{H-1}, x_H)$, where (i) x_0 is a starting state drawn from some initial distribution, (ii) $x_{h+1} \sim T_h(\cdot \mid x_h, a_h)$, and (iii) the actions are chosen by the agent according to some non-stationary policy $a_h \sim \pi(\cdot \mid x_h)$. Here, T_h denotes the (possibly non-stationary) transition dynamics $T_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ for each timestep. For notation, π_h denotes an h-step policy that chooses actions $a_0, \ldots, a_h$. We also use $\mathbb{E}_\pi[\cdot]$ and $\mathbb{P}_\pi[\cdot]$ to denote the expectations over states and actions and probability of an event respectively, when using policy π in $\mathcal{M}$. Further, we use $[H]$ to denote $\{0, 1, \ldots, H-1\}$.

We consider learning in a low-rank MDP defined as:

Definition 1 *An operator $T : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ admits a low-rank decomposition of dimension d if there exists functions $\phi^* : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu^* : \mathcal{X} \to \mathbb{R}^d$ such that: $\forall x, x' \in \mathcal{X}, a \in \mathcal{A}$: $T(x' \mid x, a) = \langle \phi^*(x, a), \mu^*(x') \rangle$,[2] and additionally $\|\phi^*(x, a)\|_2 \leq 1$ and for all $g : \mathcal{X} \to [0, 1]$, $\left\| \int g(x) \mu^*(x) dx \right\|_2 \leq \sqrt{d}$. We assume that $\mathcal{M}$ is low-rank with embedding dimension d, i.e., for each $h \in [H]$, the transition operator T_h admits a rank-d decomposition.*

We denote the embedding for T_h by ϕ_h^* and μ_h^*. In addition to the low-rank representation, we also consider a latent variable representation of $\mathcal{M}$, as defined in Agarwal et al. (2020b), as follows:

Definition 2 *The latent variable representation of a transition operator $T : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ is a latent space $\mathcal{Z}$ along with functions $\psi : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{Z})$ and $\nu : \mathcal{Z} \to \Delta(\mathcal{X})$, such that $T(\cdot \mid x, a) = \int \nu(\cdot \mid z) \psi(z \mid x, a) dz$. The latent variable dimension of T, denoted d_{LV} is the cardinality of smallest latent space $\mathcal{Z}$ for which T admits a latent variable representation. In other words, this representation gives a non-negative factorization of T.*

When state space $\mathcal{X}$ is finite, all transition operators $T_h(\cdot \mid x, a)$ admit a trivial latent variable representation. More generally, the latent variable representation enables us to augment the trajectory τ as: $\tau = \{x_0, a_0, z_1, x_1, \ldots, z_{H-1}, x_{H-1}, a_{H-1}, z_H, x_H\}$, where $z_{h+1} \sim \psi_h(\cdot \mid x_h, a_h)$ and $x_{h+1} \sim \nu_h(\cdot \mid z_{h+1})$. In general we neither assume access to nor do we learn this representation, and it is solely used to reason about the following reachability assumption:

Assumption 1 (Reachability) *There exists a constant $\eta_{\min} > 0$, such that $\forall h \in [H], z \in \mathcal{Z}_{h+1}$: $\max_\pi \mathbb{P}_\pi [z_{h+1} = z] \geq \eta_{\min}$.*

Assumption 1 posits that in MDP $\mathcal{M}$, for each factor (latent variable) at any level h, there exists a policy which reaches it with a non-trivial probability. This generalizes the reachability of latent states assumption from prior block MDP results (Du et al., 2019a; Misra et al., 2020).

2. We consider the exact low-rank decomposition for simplicity. Our results also extend to approximately case $T(x' \mid x, a) \approx \langle \phi^*(x, a), \mu^*(x') \rangle$, and more discussions can be found in Section 8.1.

Note that, exploring all latent states is still non-trivial, as a policy which chooses actions uniformly at random may hit these latent states with an exponentially small probability. More discussions are deferred to Section 3 and Section 4.

Representation learning in low-rank MDPs We consider MDPs where the state space $\mathcal{X}$ is large and the agent must employ function approximation to enable efficient learning. Given the low-rank MDP assumption, we grant the agent access to a class of representation functions mapping a state-action pair (x, a) to a d-dimensional embedding. Specifically, the feature class is $\Phi = \bigcup_{h\in[H]} \Phi_h$, where each mapping $\phi_h \in \Phi_h$ is a function $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. The feature class can now be used to learn ϕ^* and exploit the low-rank decomposition for efficient learning[3]. We assume that our feature class Φ is rich enough:

Assumption 2 (Realizability) *For each $h \in [H]$, we have $\phi_h^* \in \Phi_h$. Further, we assume that $\forall \phi_h \in \Phi_h, \forall (x, a) \in \mathcal{X} \times \mathcal{A}$, $\|\phi_h(x, a)\|_2 \leq 1$.*

Learning goal We focus on the problem of representation learning (Agarwal et al., 2020b) in low-rank MDPs where the agent tries to learn good enough features and collect a suitable dataset that enables offline optimization of any given reward in downstream tasks instead of optimizing a fixed and explicit reward signal. We consider a model-free setting and we provide this *reward-free* learning guarantee for any reward function $R = R_{0:H-1}$ ($R_{0:H-1} \coloneqq \{R_0, \dots R_{H-1}\}$ with $R_h : \mathcal{X}\times\mathcal{A} \to [0, 1], \forall h \in [H]$) in a bounded reward class $\mathcal{R}$.[4] Specifically, for such a bounded reward function R, the learned features $\{\bar{\phi}_h\}_{h\in[H]}$ and the collected data should allow the agent to compute a near-optimal policy π_R, such that $v_R^{\pi_R} \geq v_R^* - \varepsilon$, where $v_R^\pi := \mathbb{E}_\pi \left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right]$ is the expected return of policy π under reward function R, and $v_R^* := \max_\pi v_R^\pi$ is the optimal expected return for R. We desire (w.p. $\geq 1 - \delta$) sample complexity bounds which are

$$\text{poly}(d, H, K, 1/\eta_{\min}, 1/\varepsilon, \log(|\Phi|), \log(|\mathcal{R}|), \log(1/\delta)).$$

For the simplicity of presentation, we consider finite Φ and $\mathcal{R}$ classes. The results can be straightforwardly extended to infinite function classes by applying standard tools in statistical learning theory (Natarajan, 1989; Pollard, 2012; Devroye et al., 2013).

3. Related Literature

Much recent attention has been devoted to linear function approximation (c.f., Jin et al., 2020b; Yang and Wang, 2020). These results provide important building blocks for our work. In particular, the low-rank MDP model we study is from Jin et al. (2020b) who assume that the feature map ϕ^* is known in advance. However, as we are focused on nonlinear function approximation, it is more apt to compare to related nonlinear approaches, which can be categorized in terms of their dependence on the size of the function class:

Polynomial in $|\Phi|$ approaches Many approaches, while not designed explicitly for our setting, can yield sample complexity scaling polynomially with $|\Phi|$ in our setup. Note, however, that polynomial-in-$|\Phi|$ scaling can be straightforwardly obtained by concatenating

3. Sometimes we drop h in the subscript for brevity.
4. The subscript $i{:}j$ for any $i \leq j$ is also used similarly in other variables.

all of candidate feature maps and running the algorithm of Jin et al. (2020b). Further, this is the only obvious way to apply Eluder dimension results here (Osband and Roy, 2014; Wang et al., 2020b; Ayoub et al., 2020), and it also pertains to work on model selection (Pacchiano et al., 2020; Lee et al., 2021). Indeed, the key observation that enables a logarithmic-in-$|\Phi|$ sample complexity is that all value function are in fact represented as *sparse* linear functions of this concatenated feature map.

However, exploiting sparsity in RL (and in contextual bandits) is quite subtle. In both settings, it is not possible to obtain results scaling logarithmically in both the ambient dimension *and* the number of actions (Lattimore and Szepesvári, 2020; Hao et al., 2021). That said, it is possible to obtain results scaling polynomially with the number of actions and logarithmically with the ambient dimension, as we do here.

Logarithmic in $|\Phi|$ approaches For logarithmic-in-$|\Phi|$ approaches, the assumptions and results vary considerably. Several results focus on the block MDP setting (Du et al., 2019a; Misra et al., 2020; Foster et al., 2020), where the dynamics are governed by a discrete latent state space, which is decodable from the observations. This setting is a special case of our low-rank MDP setting. Additionally, these works make stronger function approximation assumptions than we do. As such our work can be seen as generalizing and relaxing assumptions, when compared with existing block MDP results.

Most closely related to our work are Olive (Jiang et al., 2017), Witness Rank (Sun et al., 2019b), Flambe (Agarwal et al., 2020b), and BLin-Ucb (Du et al., 2021) algorithms. Olive is a model-free RL algorithm that can be instantiated to produce a logarithmic-in-$|\Phi|$ sample complexity guarantee in the reward-aware (single reward) low-rank MDP setting (it also applies more generally). However, it is not computationally efficient even in tabular settings (Dann et al., 2018). Like Olive, Witness Rank and BLin-Ucb are also statistically efficient in more general settings but computationally intractable in similar ways. Witness Rank is a model-based algorithm and can handle our setting given a stronger function approximation assumption. BLin-Ucb works with a general hypothesis class, which can be either model-free or model-based and generalize the previous two approaches. All these three algorithms are restricted to the reward-aware setting, but do not require the reachability assumption as we do.

Flambe is computationally efficient with Maximum Likelihood Estimation (MLE) and sampling oracles, but it is model-based, so the function approximation assumptions are stronger than ours. Thus the key challenge, as well as the main advancement, is our weaker model-free function approximation assumption which does not allow modeling μ^* (and hence the MDP dynamics) whatsoever. On the other hand, Flambe does not require the reachability assumption when two computational oracles are available. However, in the absence of the sampling oracle for the estimated model, it does require the reachability assumption to establish theoretical guarantees. We leave fully eliminating the reachability assumption with a computationally efficient model-free approach as an important open problem.

Our proposed algorithms address the reward-free learning goal with differing tradeoffs in their statistical and computational properties. Moffle is computationally efficient with the min-max-min oracle (Eq. (5)) or the oracles for squared loss minimization and a saddle-point formulation (Algorithm 3). For the special case of enumerable feature class, our optimization

problems further reduce to a fully computationally tractable eigenvector computation problem (Eq. (12)); note that even in this special case, the computation of Olive, Witness rank, and BLin-Ucb is still inefficient as they need to enumerate over infinite function classes (see Appendix A for more details). In addition, follow-up empirical evaluations (Zhang et al., 2022) have also confirmed the practical feasibility of the optimization oracle (Algorithm 3) required by our algorithm.

In Table 1, we present a more detailed comparison between our algorithms and the closely related ones in the low-rank MDP setting. For comparisons among algorithms that tackle the more restricted block MDPs setting, we refer the reader to Zhang et al. (2022). The details of how we instantiate Olive, Witness rank, and BLin-Ucb in our setting can be found in Appendix A.

Algorithm	R-F?	Realizability	Sample Complexity	Computation
Olive (Jiang et al., 2017)	No	$\phi^* \in \Phi$	$\frac{d^3 K H^5 \log(\lvert\Phi\rvert/\delta)}{\varepsilon^2}$	Enumeration over the value class
Witness rank (Sun et al., 2019b)	No	$\phi^* \in \Phi$ $\mu^* \in \Upsilon$	$\frac{d^3 K H^5 \log(\lvert\Phi\rvert\lvert\Upsilon\rvert/\delta)}{\varepsilon^2}$	Enumeration over the model class
BLin-Ucb (Du et al., 2021)	No	$\phi^* \in \Phi$	$\frac{d^3 K H^7 \log(\lvert\Phi\rvert/\delta)}{\varepsilon^2}$	Enumeration over the hypothesis class
Flambe (Agarwal et al., 2020b)	Yes	$\phi^* \in \Phi$ $\mu^* \in \Upsilon$	$\frac{d^7 K^9 H^{22} \log(\lvert\Phi\rvert\lvert\Upsilon\rvert/\delta)}{\varepsilon^{10}}$	MLE oracle + sampling oracle
Rep-Ucb (Uehara et al., 2021)	No	$\phi^* \in \Phi$ $\mu^* \in \Upsilon$	$\frac{d^4 K^2 H^5 \log(\lvert\Phi\rvert\lvert\Upsilon\rvert/\delta)}{\varepsilon^2}$	MLE oracle + sampling oracle
RFOlive (Chen et al., 2022)	Yes	$\phi^* \in \Phi$ $\mu^* \in \Upsilon$	$\frac{d^3 K H^8 \log(\lvert\Phi\rvert\lvert\mathcal{R}\rvert/\delta)}{\varepsilon^2}$	Enumeration over the value class
Moffle (Ours)	Yes	$\phi^* \in \Phi$	$\frac{d^{11} K^{14} H^7 \log(\lvert\Phi\rvert\lvert\mathcal{R}\rvert/\delta)}{\min\{\varepsilon^2 \eta_{\min}, \eta_{\min}^5\}}$	Min-max-min oracle (Eq. (5))
Moffle (Ours)	Yes	$\phi^* \in \Phi$	$\frac{d^{19} K^{32} H^{19} \log(\lvert\Phi\rvert\lvert\mathcal{R}\rvert/\delta)}{\min\{\varepsilon^6 \eta_{\min}^3, \eta_{\min}^{11}\}}$	Squared loss minimization + saddle-point formulation (Algorithm 3)
Explore (Ours) + FQI	Yes	$\phi^* \in \Phi$	$\frac{d^{25} K^{50} H^7 \log^5(\lvert\Phi\rvert\lvert\mathcal{R}\rvert/\delta)}{\min\{\varepsilon^2 \eta_{\min}, \eta_{\min}^{17}\}}$	Enumeration over Φ + eigenvector computation (Eq. (12))

Table 1: Comparisons among algorithms for low-rank MDPs (with unknown features). R-F column refers to whether the algorithm can handle reward-free learning. Υ is the additional candidate feature class used in Sun et al. (2019b), Agarwal et al. (2020b) and Uehara et al. (2021) to capture the model-based relizability. For the sample complexity, we only show the orders and hide polylog terms (i.e., using $\tilde{O}(\cdot)$ notation). Since the sample complexity bounds for our proposed algorithms are too long, we only show their simplified versions here. We convert the $1/(1-\gamma)$ horizon dependence in Rep-Ucb (Uehara et al., 2021) to H. See the text for more discussions on realizability and computation.

Related algorithmic approaches Central to our approach is the idea of embedding plausible futures into a "discriminator" class and using this class to guide the learning process. Bellemare et al. (2019) also propose a min-max representation learning objective using a class of *adversarial value functions*, but their work only empirically demonstrates its usefulness as an auxiliary task during learning and does not study exploration. Similar ideas of using a

discriminator class have been deployed in model-based RL (Farahmand et al., 2017; Sun et al., 2019a; Modi et al., 2020; Ayoub et al., 2020), but the application to model-free representation learning and exploration is novel to our knowledge.

Subsequent works After the initial version of our work was released, there have been several follow-up papers that also investigate representation learning. Similar to Flambe, Uehara et al. (2021) develop a model-based representation-learning algorithm in low-rank MDPs, which is computationally efficient with MLE and sampling oracles. Their algorithm Rep-Ucb requires stronger function approximation assumptions (model realizability) and does not come with reward-free learning guarantees. On the other hand, Rep-Ucb does not require the reachability assumption and has a sharper rate. Ren et al. (2022) also proposes a model-based representation-learning algorithm, which exploits the noise assumption in the stochastic control model. Their setting has a low-rank structure but is different from our low-rank MDP setting.

Another recent work (Zhang et al., 2022) builds on our representation learning oracle and analysis. They present various experimental results, which can be regarded as empirical evidence and support that the representation learning oracle proposed in our work is empirically tractable and effective. As for theoretical guarantees, their results are restricted to the reward-dependent block MDP setting—which is more restrictive than low-rank MDPs—but they do not need the reachability assumption.

On the statistical side, Chen et al. (2022) achieves much better rates than ours, extend our results to the more general settings, and illustrate that the reachability assumption is not necessary for sample efficient algorithms. However, their algorithm suffers the same computational bottleneck like Olive. We believe it would be an interesting future avenue to devise an algorithm that is both computationally friendly like ours and at the same time achieve sharper rates like Chen et al. (2022). In a very recent work (Mhammedi et al., 2023), the reachability assumption has also been removed while using the same min-max-min representation learning objective as ours.

Orthogonal to our work, Zhang et al. (2021a) and Papini et al. (2021) assume a candidate set of "correct" representations is given and propose algorithms to select the "good" one (in a certain technical sense) from this candidate set. In contrast, our function-approximation assumption is much weaker: we only require *one* "correct" representation (in their terminology) to lie in the candidate set.

Finally, the recent work (Huang et al., 2021) studies deployment-efficient RL in linear MDPs, where the goal is to minimize the number of policy changes ("deployment complexity") for real-world deployment considerations. Our algorithm Moffle only requires H deployments, thus matching the optimal $\tilde{\Omega}(H)$ deployment complexity up to polylog terms and in a strictly more general setting.[5]

4. Main Algorithmic Framework

In this section, we describe the overall algorithmic framework that we propose for representation learning for low-rank MDPs. A key component in this general framework, which

5. Our earlier version of Moffle employs the online "elliptical planner", which needs $\tilde{O}(Hd^3K^4/\eta_{\min}^2)$ deployments overall. Inspired by the techniques in Huang et al. (2021), we integrate the more advanced offline "elliptical planner" to Moffle and achieve the optimal deployment complexity.

specifies how to learn a good representation once exploratory data (at previous levels) has been acquired, is left unspecified in this section and instantiated with two different choices in the subsequent sections. We also present sample complexity guarantees for each choice in the corresponding sections.

At the core of our model-free approach is the following well-known property of a low-rank MDP due to Jin et al. (2020b). We provide a proof in Appendix F.1 for completeness.

Lemma 3 (Jin et al. (2020b)) *For a low-rank MDP $\mathcal{M}$ with embedding dimension d, for any $f : \mathcal{X} \to [0,1]$, we have: $\mathbb{E}\left[f(x_{h+1}) \mid x_h, a_h\right] = \langle \phi_h^*(x_h, a_h), \theta_f^* \rangle$, where $\theta_f^* \in \mathbb{R}^d$ and $\|\theta_f^*\|_2 \leq \sqrt{d}$.*

We turn this property into an algorithm by finding a feature map in the candidate class Φ which can certify this condition for a sufficiently rich class of functions $\mathcal{F}$. The key insight in our algorithm is that this property depends solely on ϕ^*, so we do not require additional modeling assumptions.

Before turning to the algorithm description, we clarify a useful notation for h step policies. An h-step policy[6] ρ_h chooses actions $a_0, \ldots, a_h$, consequently inducing a distribution over (x_h, a_h, x_{h+1}). We routinely append several random actions to such a policy, and we use ρ_h^{+i} to denote the policy that chooses $a_{0:h}$ according to ρ_h and then takes actions uniformly at random for i steps, inducing a distribution over $(x_{h+i}, a_{h+i}, x_{h+i+1})$. As an edge case, for $i \geq j \geq 0$, ρ_{-j}^{+i} takes actions $a_0, \ldots, a_{i-j}$ uniformly. The mnemonic is that the last action taken by ρ_j^{+i} is a_{i+j}.

Our algorithm, Model-Free Feature Learning and Exploration (MOFFLE) shown in Algorithm 1, takes as input a feature set Φ, a reward class $\mathcal{R}$, the reachability coefficient $\eta_{\min}$ (Assumption 1), the sub-optimality parameter ε, and the high-probability parameter δ. It outputs feature maps $\bar{\phi}_{0:H-1}$ and a dataset $\mathcal{D} = (\mathcal{D}_{0:H-1})$ such that Fitted Q-Iteration (FQI), using linear functions of the returned features, can be run with the returned dataset to obtain a ε-optimal policy for any reward function in $\mathcal{R}$. The algorithm runs in the following two stages.

Algorithm 1 MOFFLE $(\mathcal{R}, \Phi, \eta_{\min}, \varepsilon, \delta)$: Model-Free Feature Learning and Exploration

Set $\beta \leftarrow \tilde{O}\left(\frac{\eta_{\min}^2}{dK^4B^2}\right)$, $\kappa \leftarrow \frac{64dK^4\log(1+8/\beta)}{\eta_{\min}}$, and $\varepsilon_{\text{apx}} \leftarrow \frac{\varepsilon^2}{16H^4\kappa K}$.
Compute the exploratory policy cover: $\left\{\rho_{h-3}^{+3}\right\}_{h\in[H]} \leftarrow$ EXPLORE $(\Phi, \eta_{\min}, \delta)$.
for $h \in [H]$ **do**
  Collect dataset $\mathcal{D}_h^{\bar{\phi}}$ of size $n_{\bar{\phi}}$ using ρ_{h-3}^{+3}.
  Learn representation $\bar{\phi}_h$ by solving Eq. (5) (or calling Algorithm 3) with feature class Φ_h, discriminator class $\mathcal{V} = \mathcal{G}_{h+1}$, dataset $\mathcal{D}_h^{\bar{\phi}}$ and tolerance ε_{apx}.
  Collect dataset $\mathcal{D}_h$ of size n_{plan} using ρ_{h-3}^{+3}.
  Set $\mathcal{D} \leftarrow \mathcal{D} \bigcup \{\mathcal{D}_h\}$.
end for
return $\mathcal{D}, \bar{\phi}_{0:H-1}$.

6. We use π_h to denote a standard (single) policy and ρ_h to denote an (exploratory) mixture policy, i.e., uniformly sampling from a set of policies.

Exploration In line 2 of Moffle, we use the Explore sub-routine (Algorithm 2) to compute exploratory policies ρ_{h-3}^{+3} for each timestep $h \in [H]$. It takes as input a feature set Φ, the reachability coefficient $\eta_{\min}$, and the high-probability parameter δ. Intuitively, Algorithm 2 returns a set of exploratory policies ρ_{h-3}^{+3} for each timestep $h \in [H]$ such that ρ_{h-3}^{+3} hits each latent state in the set $\mathcal{Z}_h$ with a large enough probability.

Algorithm 2 Explore $(\Phi, \eta_{\min}, \delta)$

1: Set $\beta \leftarrow \tilde{O}\left(\frac{\eta_{\min}^2}{dK^4B^2}\right)$, and $\varepsilon_{\text{reg}} \leftarrow \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^9\log^2(1+8/\beta)}\right)$.
2: **for** $h = 0, \ldots, H-1$ **do**
3: Set exploratory policy for step h to ρ_{h-3}^{+3}.
4: Collect dataset $\mathcal{D}_h^{\hat{\phi}}$ of size $n_{\hat{\phi}}$ using ρ_{h-3}^{+3}.
5: Learn representation $\hat{\phi}_h$ for timestep h by solving Eq. (5) (or calling Algorithm 3) with feature class Φ_h, discriminator class $\mathcal{F}_{h+1}$, dataset $\mathcal{D}_h^{\hat{\phi}}$ and tolerance ε_{reg}.
6: Collect dataset $\mathcal{D}_h^{\text{ell}}$ of size n_{ell} using ρ_{h-3}^{+3}.
7: Call offline elliptical planner (Algorithm 4) with features $\hat{\phi}$, dataset $\mathcal{D}_{0:h}^{\text{ell}}$ and β to obtain policy ρ_h.
8: **end for**
9: **return** Exploratory policies $\left\{\rho_{h-3}^{+3}\right\}_{h\in[H]}$.

Algorithm 2 uses a step-wise forward exploration scheme similar to Flambe (Agarwal et al., 2020b). The algorithm proceeds in stages. For each level h, we first collect a dataset $\mathcal{D}_h^{\hat{\phi}}$ of size $n_{\hat{\phi}}$ (the superscript $\hat{\phi}$ implies that the dataset will be used to learn $\hat{\phi}$) with the exploratory policy ρ_{h-3}^{+3} constructed in the previous level (line 4) and then use such dataset to learn feature $\hat{\phi}_h$ (line 5) by calling a feature learning sub-routine (discussed in the sequel). The feature $\hat{\phi}_h$ is computed to approximate the property in Lemma 3 for the discriminator function class $\mathcal{F}_{h+1} \subseteq (\mathcal{X} \to [0,1])$ defined as

$$\mathcal{F}_{h+1} := \left\{\text{clip}_{[0,1]}(\mathbb{E}_{\text{unif}(\mathcal{A})}\langle\phi_{h+1}(x_{h+1}, a), \theta\rangle) : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le B\right\}, \text{ where } B \ge \sqrt{d}. \quad (1)$$

Using the policy ρ_{h-3}^{+3}, we also collect the exploratory dataset $\mathcal{D}_h^{\text{ell}}$ of size n_{ell} ("ell" stands for elliptical) for step h (line 6). With the collected datasets[7] $\mathcal{D}_{0:h}^{\text{ell}}$ and features $\hat{\phi}_h$ we call an *offline* "elliptical" planning subroutine (Algorithm 4) to compute the policy ρ_h. This planning algorithm is inspired by techniques used in reward-free exploration (Wang et al., 2020a; Zanette et al., 2020b; Huang et al., 2021). Most elliptical planning algorithms in the literature require online interactions with the environment, where in each round the agent sets the reward appropriately to collect data from a previously unexplored direction. In contrast, our offline elliptical planner only uses the offline data, which is built on Huang et al. (2021) but handles the more computationally friendly continuous function class and requires more involved analysis . The detailed description with the pseudocode is deferred to Appendix B.

7. For a dataset $\mathcal{D}_h$, subscript h denotes that it is a collection of tuples (x_h, a_h, x_{h+1}). $\mathcal{D}_{0:h}^{\text{ell}}$ denotes $\{\mathcal{D}_0^{\text{ell}}, \ldots, \mathcal{D}_h^{\text{ell}}\}$.

In Algorithm 2, parameters $\beta, \varepsilon_{\text{reg}}$ are set according to Theorem 8, which is deferred to Section 8.1. The missing values $B, n_{\hat{\phi}}$, and n_{ell} are specifically assigned for different instantiations, and we will present them in detail when later stating the formal theoretical guarantees. The proposed algorithm requires reachability parameter $\eta_{\min}$ as an input, which is usually unknown in the practical case. Notice that $\eta_{\min}$ is used to set the size of the dataset and other parameters. In the experiment, we can set the sample size in an ad hoc way and fine-tune all parameters by checking the performance.

Representation learning In MOFFLE, we subsequently learn a feature $\bar{\phi}_h$ for each level—again by invoking the representation learning subroutine—that allows us to use FQI to plan for any reward $R \in \mathcal{R}$ afterwards. Similarly to EXPLORE sub-routine (Algorithm 2), we collect a dataset $\mathcal{D}_h^{\bar{\phi}}$ of size $n_{\bar{\phi}}$ (the superscript $\bar{\phi}$ implies that the dataset will be used to learn $\bar{\phi}$) in line 4, which is then used to learn features $\bar{\phi}_h$ in line 5. Additionally, we use the exploratory mixture policy ρ_{h-3}^{+3} to collect a dataset $\mathcal{D}_h$ of size n_{plan} for the downstream planning (line 6). Here for learning feature $\bar{\phi}_h$, we use a discriminator function class $\mathcal{G}_{h+1} \subseteq (\mathcal{X} \to [0, H])$ defined as

$$\mathcal{G}_{h+1} := \Big\{ \text{clip}_{[0,H]}\Big(\max_a (R_{h+1}(x_{h+1}, a) + \langle \phi_{h+1}(x_{h+1}, a), \theta \rangle) \Big) : R \in \mathcal{R}, \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le B \Big\} \text{ where } B \ge H\sqrt{d}. \tag{2}$$

Note that the class $\mathcal{G}_{h+1}$, while still derived from Φ, is quite different from the class $\mathcal{F}_{h+1}$ (Eq. (1)) used to learn features inside the exploration module. Recall that in $\mathcal{F}_{h+1}$, we clip the functions to $[0, 1]$, set $R_{h+1}(x_{h+1}, a) = 0$, and take expectation with respect to $a \sim \text{unif}(\mathcal{A})$ instead of a maximum. Finally, MOFFLE returns the computed features $\bar{\phi}_{0:H-1}$ and the exploratory dataset $\mathcal{D}_{0:H-1}$. The intuition that we use a different function class $\mathcal{G}_{h+1}$ and learn feature $\bar{\phi}_h$ is that the previous function class $\mathcal{F}_{h+1}$ and learned $\hat{\phi}_h$ only capture the probability of hitting latent states for a given policy whereas here we want to ensure that we capture the Bellman backup of all plausible reward-appended value functions.

In Algorithm 1, β and κ are set according to Theorem 8, and ε_{apx} is set according to Theorem 9. The missing values $B, n_{\bar{\phi}}$, and n_{plan} are again specifically chosen for different instantiations, and we discuss them in detail later.

Careful readers may have noticed that we use the same exploratory mixture policy ρ_{h-3}^{+3} to collect different datasets in several places. In the practical implementation, we can equivalently collect a large enough dataset $\mathcal{D}_h$ once and then use it to unify current $\mathcal{D}_h^{\hat{\phi}}, \mathcal{D}_h^{\bar{\phi}}, \mathcal{D}_h^{\text{ell}}$, and $\mathcal{D}_h$. Therefore, in our subsequent discussions, sometimes we will simply use $\mathcal{D}_h$ to refer to such a dataset and do not differentiate them.

Planning in downstream tasks For downstream planning with any reward $R \in \mathcal{R}$, we use FQI (Ernst et al., 2005; Antos et al., 2007, 2008; Munos and Szepesvári, 2008; Szepesvári, 2010; Chen and Jiang, 2019) with the following Q-function class defined using the features $\bar{\phi}_{0:H-1}$:

$$\mathcal{Q}(\bar{\phi}, R) := \bigcup_{h \in [H]} \mathcal{Q}_h(\bar{\phi}_h, R_h), \tag{3}$$

$$\mathcal{Q}_h(\bar{\phi}_h, R_h) := \Big\{ \text{clip}_{[0,H]}(R_h(x_h, a_h) + \langle \bar{\phi}_h(x_h, a_h), w \rangle) : \|w\|_2 \le B \Big\}.$$

Note that features $\bar{\phi}$ are computed to approximate the backup of candidate functions of this form (class $\mathcal{G}_{h+1}$), and thus, satisfy the conditions stated in Chen and Jiang (2019) for using FQI.

5. Min-Max-Min Representation Learning

In this section, we describe our novel representation learning objective for finding $\hat{\phi}$ and $\bar{\phi}$. The key insight is that the low-rank property of the MDP $\mathcal{M}$ can be used to learn a feature map $\hat{\phi}$ which can approximate the Bellman backup of all linear functions under feature maps $\phi \in \Phi$, and that approximating the backups of these functions enables subsequent near-optimal planning.

We present the algorithm with an abstract discriminator class $\mathcal{V} \subseteq (\mathcal{X} \to [0, L])$ that is instantiated either with $\mathcal{F}_{h+1}$ (with $L = 1$) or $\mathcal{G}_{h+1}$ (with $L = H$) defined in Eq. (1) and Eq. (2), respectively. In order to describe our objective, it is helpful to introduce the shorthand

$$\text{b\_err}\left(\pi_h, \phi_h, v; B\right) = \min_{\|w\|_2 \le B} \mathbb{E}_{\pi_h}\left[\left(\langle \phi_h(x_h, a_h), w\rangle - \mathbb{E}\left[v(x_{h+1}) \mid x_h, a_h\right]\right)^2\right] \tag{4}$$

for any policy π_h, feature ϕ_h, function v, and constant B, which we set so that $B \ge L\sqrt{d}$. This is the error in approximating the conditional expectation of $v(x_{h+1})$ using linear functions in the features $\phi_h(x_h, a_h)$. For approximating backups of all functions $v \in \mathcal{V}$, we seek feature $\hat{\phi}_h$ which minimizes $\max_{v \in \mathcal{V}} \text{b\_err}\left(\rho^{+3}_{h-3}, \hat{\phi}_h, v; B\right)$ up to an error of ε_{tol}.

Unfortunately, the quantity $\text{b\_err}\left(\cdot\right)$ contains a conditional expectation inside the square loss, so we cannot estimate it from samples $(x_h, a_h, x_{h+1}) \sim \rho^{+3}_{h-3}$. This is an instance of the well-known double sampling issue (Baird III, 1995; Antos et al., 2008). Instead, we introduce the loss function

$$\mathcal{L}_{\rho^{+3}_{h-3}}(\phi_h, w, v) = \mathbb{E}_{\rho^{+3}_{h-3}}\left[\left(\langle \phi_h(x_h, a_h), w\rangle - v(x_{h+1})\right)^2\right],$$

which is amenable to estimation from samples. However this loss function contains an undesirable conditional variance term, since via the bias-variance decomposition we have

$$\mathcal{L}_{\rho^{+3}_{h-3}}(\phi_h, w, v) = \text{b\_err}\left(\rho^{+3}_{h-3}, \phi_h, v; B\right) + \mathbb{E}_{\rho^{+3}_{h-3}}\left[\mathbb{V}\left[v(x_{h+1}) \mid x_h, a_h\right]\right].$$

The excess variance term can lead the agent to erroneously select a bad feature $\hat{\phi}_h$. However, via Lemma 3, we can rewrite the conditional variance as $\mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_v, v)$ for some $\|\theta^*_v\|_2 \le L\sqrt{d}$. Therefore, we can instead optimize the following *variance-corrected* objective which includes a correction term:

$$\operatorname*{argmin}_{\phi_h \in \Phi_h} \max_{v \in \mathcal{V}} \Big\{ \min_{\|w\|_2 \le B} \mathcal{L}_{\mathcal{D}_h}(\phi_h, w, v) - \min_{\tilde{\phi}_h \in \Phi_h, \|\tilde{w}\|_2 \le L\sqrt{d}} \mathcal{L}_{\mathcal{D}_h}(\tilde{\phi}_h, \tilde{w}, v) \Big\}. \tag{5}$$

Here we set the constant $B \ge L\sqrt{d}$, and $\mathcal{L}_{\mathcal{D}_h}(\cdot)$ is the empirical estimate of $\mathcal{L}_{\rho^{+3}_{h-3}}(\cdot)$ using dataset $\mathcal{D}_h = \left\{\left(x^{(i)}_h, a^{(i)}_h, x^{(i)}_{h+1}\right)\right\}_{i=1}^n$, which is defined as

$$\mathcal{L}_{\mathcal{D}_h}(\phi_h, w, v) := \sum_{i=1}^n \left(\left\langle \phi_h\left(x^{(i)}_h, a^{(i)}_h\right), w\right\rangle - v\left(x^{(i)}_{h+1}\right)\right)^2.$$

We now state our first result, which is an information-theoretic guarantee and assumes that an oracle solver for the objective in Eq. (5) is available when we run Moffle. A complete proof will be given in Section 8.4.

Theorem 4 *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Definition 1 and Assumption 1, Assumption 2 hold. If an oracle solution to Eq. (5) is available, then by setting*

$$B = \sqrt{d}, \quad n_{\hat{\phi}} = \tilde{O}\left(\frac{d^4 K^9 \log(|\Phi|/\delta)}{\eta_{\min}^3}\right), \quad n_{\text{ell}} = \tilde{O}\left(\frac{H^5 d^{11} K^{14} \log(|\Phi|/\delta)}{\eta_{\min}^5}\right),$$
$$n_{\bar{\phi}} = \tilde{O}\left(\frac{H^6 d^3 K^5 \log\left(|\Phi||\mathcal{R}|/\delta\right)}{\varepsilon^2 \eta_{\min}}\right), \quad n_{\text{plan}} = \tilde{O}\left(\frac{H^6 d^2 K^5 \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2 \eta_{\min}}\right),$$

with probability at least $1-\delta$, Moffle *returns an exploratory dataset $\mathcal{D}$ s.t. for any $R \in \mathcal{R}$, running* FQI *with value function class $\mathcal{Q}(\bar{\phi}, R)$ defined in Eq. (3) returns an ε-optimal policy for MDP $\mathcal{M}$. The total number of episodes used by the algorithm is*

$$\tilde{O}\left(\frac{H^6 d^{11} K^{14} \log(|\Phi|/\delta)}{\eta_{\min}^5} + \frac{H^7 d^3 K^5 \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2 \eta_{\min}}\right).$$

When compared with the most related work Flambe, our sample complexity bound has worse dependence on K, d and better dependence on ε, H. We additionally pay $1/\eta_{\min}$ and $\log(|\mathcal{R}|)$ since our algorithm is reward-free model-free and requires the reachability assumption. On the other side, Flambe has a $\log(|\Upsilon|)$ dependence, where Υ is another function class that captures the second component of the low-rank decomposition. We refer the reader to Table 1 for more detailed comparisons.

6. Iterative Greedy Representation Learning

The min-max-min objective (Eq. (5)) in the previous section is not provably computationally tractable for non-enumerable and non-linear function classes. However, recent empirical work (Lin et al., 2020) has considered a heuristic approach for solving similar min-max-min objectives by alternating between updating the outer min and inner max-min components. In this section, we show that a similar iterative approach that alternates between a *squared loss minimization problem* and a *max-min objective* in each iteration can be used to provably solve our representation learning problem.

This iterative procedure is displayed in Algorithm 3. Given the discriminator class $\mathcal{V}$ (instantiated with $\mathcal{F}_{h+1}$ as defined in Eq. (1) for learning $\hat{\phi}$ or $\mathcal{G}_{h+1}$ as defined in Eq. (2) for learning $\bar{\phi}$), the algorithm grows finite subsets $\mathcal{V}^1, \mathcal{V}^2, \ldots \subseteq \mathcal{V}$ in an incremental and greedy fashion with $\mathcal{V}^1 = \{v_1\}$ initialized arbitrarily. In the t^{th} iteration, we have the discriminator class $\mathcal{V}^t$ and we estimate a feature $\hat{\phi}_{t,h}$ which has a low total squared loss with respect to all functions in $\mathcal{V}^t$ (line 6). Importantly, the total *squared loss* (sum) avoids the double sampling issue that arises with the worst case loss over class $\mathcal{V}^t$ (max), so no correction term is required. More specifically, it is easy to see from Eq. (5) and Eq. (6) that once all v_i are fixed, the conditional variance terms and their sum is also fixed. Thus it can be dropped when we minimize over ϕ_h and W.

Next, we try to certify that $\hat{\phi}_{t,h}$ is a good representation by searching for a *witness* function $v_{t+1} \in \mathcal{V}$ for which $\hat{\phi}_{t,h}$ has large excess square loss (line 7). The optimization

problem in Eq. (7) does require a correction term to address double sampling, but since $\hat{\phi}_{t,h}$ is fixed, it can be written as a simpler *max-min* program when compared to the previous oracle approach. If the objective value l (line 8) is smaller than the threshold ε_{tol} (instantiated with ε_{reg} for learning $\hat{\phi}$ or ε_{apx} for learning $\bar{\phi}$), then our certification successfully verifies that $\hat{\phi}_{t,h}$ can approximate the Bellman backup of all functions in $\mathcal{V}$, so we terminate and output $\hat{\phi}_{t,h}$. On the other hand, if the objective is large, we add the witness v_{t+1} to our growing discriminator class and advance to the next iteration.

Algorithm 3 Feature Selection via Greedy Improvement

1: **input:** Feature class Φ_h, discriminator class $\mathcal{V}$, dataset $\mathcal{D}$ and tolerance ε_{tol}.
2: Set $\mathcal{V}^0 \leftarrow \emptyset$ and choose $v_1 \in \mathcal{V}$ arbitrarily.
3: Set $\varepsilon_0 \leftarrow \varepsilon_{\text{tol}}/52d^2$, $t \leftarrow 1$ and $l \leftarrow \infty$.
4: **repeat**
5: Set $\mathcal{V}^t \leftarrow \mathcal{V}^{t-1} \bigcup \{v_t\}$.
6: **(Fit feature)** Compute $\hat{\phi}_{t,h}$ as:

$$\hat{\phi}_{t,h}, W_t = \operatorname*{argmin}_{\phi_h \in \Phi_h, W \in \mathbb{R}^{d \times t}, \|W\|_{2,\infty} \le L\sqrt{d}} \sum_{i=1}^{t} \mathcal{L}_{\mathcal{D}_h}(\phi_h, W^i, v_i). \tag{6}$$

7: **(Find witness)** Find test witness function:

$$v_{t+1} = \operatorname*{argmax}_{v \in \mathcal{V}} \max_{\tilde{\phi}_h \in \Phi_h, \|\tilde{w}\|_2 \le L\sqrt{d}} \left(\min_{\|w\|_2 \le \frac{L\sqrt{dt}}{2}} \mathcal{L}_{\mathcal{D}_h}(\hat{\phi}_{t,h}, w, v) - \mathcal{L}_{\mathcal{D}_h}(\tilde{\phi}_h, \tilde{w}, v) \right). \tag{7}$$

8: Set test loss l to the objective value in Eq. (7).
9: **until** $l < 24d^2\varepsilon_0 + \varepsilon_0^2$.
10: **return** Feature $\hat{\phi}_{T,h}$ from last iteration T.

One technical point worth noting is that in Eq. (7) we relax the norm constraint on w to allow it to grow with $\sqrt{t}$ (Eq. (36) in the proof of Lemma 14). This is required by our iteration complexity analysis which we summarize in the following lemma:

Lemma 5 *(Informal) Fix $\delta \in (0,1)$. If the dataset $\mathcal{D}$ is sufficiently large, then with probability at least $1-\delta$, Algorithm 3 terminates after $T = \frac{52L^2d^2}{\varepsilon_{\text{tol}}}$ iterations and returns a feature $\hat{\phi}_h$ such that:*

$$\max_{v \in \mathcal{V}} \text{b\_err}\left(\rho_{h-3}^{+3}, \hat{\phi}_h, v; \sqrt{\frac{13L^4d^3}{\varepsilon_{\text{tol}}}}\right) \le \varepsilon_{\text{tol}}. \tag{8}$$

The size of $\mathcal{D}$ scales polynomially with the relevant parameters, e.g., for $\mathcal{V} = \mathcal{F}_{h+1}$, we set $n = \tilde{O}\left(\frac{d^7 \log(|\Phi|/\delta)}{\varepsilon_{\text{tol}}^3}\right)$.

A formal statement (Lemma 14) along with its complete proof is provided in Section 8.5. Eq. (8) in Lemma 5 shows that the learned feature $\hat{\phi}_h$ does have a small Bellman backup error (as defined in Eq. (4)) for all discriminator functions in the class $\mathcal{F}_{h+1}$. Notice that the norm bound in Eq. (8) scales with the accuracy parameter ε_{tol} and degrades the overall

sample complexity when compared with using the oracle approach (Eq. (5)). However, it can be used in Moffle leading to a more computationally viable algorithm. In the following we state sample complexity result for Moffle when Algorithm 3 is used as the feature learning sub-routine. The detailed analysis for the result can be found in Section 8.5.

Theorem 6 *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Definition 1 and Assumption 1, Assumption 2 hold. If Eq. (5) is solved via the iterative greedy approach (Algorithm 3), then by setting*

$$B = \tilde{O}\left(\sqrt{\frac{d^5K^9}{\eta_{\min}^3}}\right),\ n_{\hat{\phi}} = \tilde{O}\left(\frac{d^{13}K^{27}\log(|\Phi|/\delta)}{\eta_{\min}^9}\right),\ n_{\text{ell}} = \tilde{O}\left(\frac{H^5d^{19}K^{32}\log(|\Phi|/\delta)}{\eta_{\min}^{11}}\right),$$
$$n_{\bar{\phi}} = \tilde{O}\left(\frac{H^{18}d^{10}K^{15}\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^6\eta_{\min}^3}\right), \quad n_{\text{plan}} = \tilde{O}\left(\frac{H^6d^2K^5\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2\eta_{\min}}\right),$$

with probability at least $1-\delta$, Moffle *returns an exploratory dataset $\mathcal{D}$ s.t. for any $R \in \mathcal{R}$, running* FQI *with value function class $\mathcal{Q}(\bar{\phi}, R)$ defined in Eq. (3) returns an ε-optimal policy for MDP $\mathcal{M}$. The total number of episodes used by the algorithm is*

$$\tilde{O}\left(\frac{H^6d^{19}K^{32}\log(|\Phi|/\delta)}{\eta_{\min}^{11}} + \frac{H^{19}d^{10}K^{15}\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^6\eta_{\min}^3}\right).$$

Applying Algorithm 3 leads to a more computationally viable algorithm by breaking the feature learning objective in Eq. (5) into familiar computational primitives: *squared loss minimization* (Eq. (6)) and a *saddle point problem* (Eq. (7)). Apart from the squared loss minimization which is considered tractable, saddle point problems have also become common in recent off-policy RL methods like Dai et al. (2018); Zhang et al. (2019) where scalable optimization heuristics are presented as well. For the practical implementation of Algorithm 3, we can choose parameterized Φ and $\mathcal{V}$ classes, which allow us to use gradient descent based methods for learning features and finding witness functions (discriminators) in a scalable manner. We want to highlight that the subsequent work of Zhang et al. (2022), which builds on the same representation learning oracle as ours (with minor differences in the discriminator class), provides empirical evidence that our iterative greedy representation learning oracle (Algorithm 3) is indeed implementable and more computationally viable.

7. Provably Computationally-Tractable Reward-Free RL with an Enumerable Feature Class

A critical component of Moffle (Algorithm 1) is the stage of collecting exploratory data. The problem of reward-free exploration only asks the agent to collect a dataset with good coverage over the state space and does not require the agent to learn a representation. This problem has been studied in recent literature for tabular (Jin et al., 2020a; Kaufmann et al., 2021; Ménard et al., 2021; Zhang et al., 2021c), block MDPs (Misra et al., 2020), and linear-MDP/low inherent Bellman error/linear-mixture setting with known features (Wang et al., 2020a; Zanette et al., 2020b; Zhang et al., 2021b; Huang et al., 2021; Wagenmaker et al., 2022).[8] In this section, we describe a special case where the min-max-min objective (Eq. (5)) from Moffle results in a provably computationally-tractable reward-free exploration scheme.

8. We only provide an incomplete list here.

In particular, we show that when Φ is efficiently enumerable, we can use Algorithm 2 to compute an exploratory policy cover in a computationally tractable manner. We can learn $\hat{\phi}$ using a slightly different min-max-min objective

$$\operatorname*{argmin}_{\phi \in \Phi_h} \max_{f \in \mathcal{F}_{h+1}, \tilde{\phi} \in \Phi_h, \|\tilde{w}\|_2 \le B} \left\{ \min_{\|w\|_2 \le B} \mathcal{L}_{\mathcal{D}_h}(\phi, w, f) - \mathcal{L}_{\mathcal{D}_h}(\tilde{\phi}, \tilde{w}, f) \right\}, \tag{9}$$

where with a slight abuse of notation $\mathcal{F}_{h+1}$ (different from that in Eq. (1)) now is the discriminator class that contains all *unclipped* functions f in form of

$$\mathcal{F}_{h+1} := \left\{ \mathbb{E}_{\mathrm{unif}(\mathcal{A})} \left[\langle \phi_{h+1}(x_{h+1}, a), \theta \rangle \right] : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le \sqrt{d} \right\}. \tag{10}$$

Consider the min-max-min objective Eq. (9) and fix $\phi, \tilde{\phi} \in \Phi_h$. We show that it can be can be reduced to

$$\max_{f \in \mathcal{F}_{h+1}} f(\mathcal{D}_h)^\top \left(A(\phi)^\top A(\phi) - A(\tilde{\phi})^\top A(\tilde{\phi}) \right) f(\mathcal{D}_h) \tag{11}$$

where $A(\phi) = I_{n\times n} - X \left(\frac{1}{n} X^\top X + \lambda I_{d\times d} \right)^{-1} \left(\frac{1}{n} X^\top \right)$, $A(\tilde{\phi}) = I_{n\times n} - \tilde{X} \left(\frac{1}{n} \tilde{X}^\top \tilde{X} + \lambda I_{d\times d} \right)^{-1} \left(\frac{1}{n} \tilde{X}^\top \right)$ for a parameter λ. Additionally, $X, \tilde{X} \in \mathbb{R}^{n\times d}$ are the sample covariate matrices for features $\phi, \tilde{\phi}$ respectively, and we overload the notation and use $f(\mathcal{D}_h) \in \mathbb{R}^n$ to denote the value of any $f \in \mathcal{F}_{h+1}$ on the n samples. The objective in Eq. (11) is obtained by using a ridge regression solution for $w, \tilde{w}$ in Eq. (9) and the details are deferred to Section 8.6.

Finally, for any fixed feature ϕ' in the definition of $f = X'\theta \in \mathcal{F}_{h+1}$, we can rewrite Eq. (11) as

$$\max_{\|\theta\|_2 \le \sqrt{d}} \theta^\top X'^\top \left(A(\phi)^\top A(\phi) - A(\tilde{\phi})^\top A(\tilde{\phi}) \right) X'\theta$$

where $X' \in \mathbb{R}^{n\times d}$ is again a sample matrix defined using $\phi' \in \Phi_{h+1}$.

Thus, for a fixed tuple of $(\phi, \tilde{\phi}, \phi')$, the maximization problem reduces to a tractable eigenvector computation problem. As a result, we can efficiently solve the min-max-min objective in Eq. (9) by enumerating over each candidate feature in $(\phi, \tilde{\phi}, \phi')$ to solve

$$\operatorname*{argmin}_{\phi \in \Phi_h} \max_{\tilde{\phi} \in \Phi_h, \phi' \in \Phi_{h+1}, \|\theta\|_2 \le \sqrt{d}} \theta^\top X'^\top \left(A(\phi)^\top A(\phi) - A(\tilde{\phi})^\top A(\tilde{\phi}) \right) X'\theta. \tag{12}$$

While the analysis is more technical, Eq. (12) still allows us to plan using $\hat{\phi}_h$ in Explore (Algorithm 2) to guarantee that the policies ρ_{h-3}^{+3} are exploratory. With the exploratory data, we can subsequently call FQI with the following Q-function class defined using the entire feature class for the downstream planning:

$$\mathcal{Q}(R) := \bigcup_{h \in [H]} \mathcal{Q}_h(R), \tag{13}$$

$$\mathcal{Q}_h(R) := \left\{ \mathrm{clip}_{[0,H]} \left(R_h(x,a) + \langle \phi_h(x,a), w \rangle \right) : \|w\|_2 \le B, \phi_h \in \Phi_h \right\}, \text{ where } B \ge H\sqrt{d}.$$

We summarize the overall result as follows and its proof can be found in Section 8.6.

Theorem 7 *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Definition 1 and Assumption 1, Assumption 2 hold. In* Explore *(Algorithm 2), if $\hat{\phi}_h$ is learned using the eigenvector formulation Eq.* (12)*, then by setting*

$$B = \tilde{\Theta}\left(\frac{d^4K^9\log\left(|\Phi|/\delta\right)}{\eta_{\min}^3}\right), \quad n_{\hat{\phi}} = \tilde{O}\left(\frac{d^{12}K^{27}\log^3(|\Phi|/\delta)}{\eta_{\min}^9}\right),$$
$$n_{\mathrm{ell}} = \tilde{O}\left(\frac{H^5d^{25}K^{50}\log^5(|\Phi|/\delta)}{\eta_{\min}^{17}}\right), \quad n_{\mathrm{plan}} = \tilde{O}\left(\frac{H^6d^3K^5\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2\eta_{\min}}\right),$$

Moffle *returns an exploratory dataset $\mathcal{D}$ such that for any $R \in \mathcal{R}$, running* FQI *with the collected dataset and the value function class $\mathcal{Q}(R)$ defined in Eq.* (13) *returns an ε-optimal policy with probability at least $1-\delta$. The total number of episodes used by the algorithm is*

$$\tilde{O}\left(\frac{H^6d^{25}K^{50}\log^5(|\Phi|/\delta)}{\eta_{\min}^{17}} + \frac{H^7d^3K^5\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2\eta_{\min}}\right).$$

From the result, we can see that although Explore with Eq. (12) enumerates over the candidate feature class Φ, the sample complexity is still logarithmic in $|\Phi|$. We believe there may be still room for further improving the bound, and we leave it to future work.

8. Proofs

In this section, we present detailed proofs. We start with the overall proof outline for Moffle in Section 8.1, and then provide the proofs for the exploration and representation learning components in Section 8.2 and Section 8.3 respectively. The concrete results for min-max-min oracle representation learning (Eq. (5)), iterative greedy representation learning (Algorithm 3), and enumerable case (Eq. (12)) are shown in Section 8.4, Section 8.5, and Section 8.6 respectively. We defer the statements and proofs for the offline elliptical planner (Algorithm 4), FQI (Algorithm 5), FQE (Algorithm 6), and auxiliary results (e.g., concentration arguments and deviation bounds for regression with squared loss) to the appendix.

8.1 Proof Outline

We provide some intuition behind the design choices in Moffle and give a sketch of the proof of the main results. We divide the proof sketch into four stages: (i) establishing the exploratory nature of the policies ρ^{+3}, (ii) representation learning guarantees for features $\bar{\phi}$ used for the downstream planning, (iii) concentration arguments for learned features $\hat{\phi}$ and $\bar{\phi}$, and (iv) final planning in downstream tasks.

Computing exploratory policies To understand the intuition behind Explore (Algorithm 2), it is helpful to consider how we can discover a policy cover over the latent state space $\mathcal{Z}_{h+1}$. If we knew the mapping to latent states, we could create the reward functions $\mathbf{1}[z_{h+1} = z]$ for all $z \in \mathcal{Z}_{h+1}$ and compute policies to optimize such rewards, but here we do not have access to this mapping. Additionally, we do not have access to the true features ϕ^* to enable tractable planning even for the known rewards. Explore tackles both of these challenges. Note that in this section, we will establish the coverage over $\mathcal{Z}_{h+1}$ through

learning feature $\hat{\phi}_{h-2}$ and calling the offline "elliptical planner" (Algorithm 4) to build an exploratory mixture policy ρ_{h-2}. This is for the simplicity of presentation. To connect it to the computation at level h in Explore, we need to add all subscripts by 2, i.e., changing $\mathcal{Z}_{h+1}$ to $\mathcal{Z}_{h+3}$, $\hat{\phi}_{h-2}$ to $\hat{\phi}_h$, ρ_{h-2} to ρ_h, etc.

For the first challenge, we note that by Definition 2 of the latent variables and Lemma 3, there always exists $f(x_h, a_h) = \mathbb{P}[z_{h+1} = z \mid x_h, z_h]$ such that

$$\begin{aligned}&\mathbb{P}[z_{h+1} = z \mid x_{h-1}, a_{h-1}] = \mathbb{E}[\mathbf{1}[z_{h+1} = z] \mid x_{h-1}, a_{h-1}] \\ &= \mathbb{E}[f(x_h, a_h) \mid x_{h-1}, a_{h-1}] = \langle \phi^*_{h-1}(x_{h-1}, a_{h-1}), \theta^*_f \rangle.\end{aligned}$$

This essentially says the desired indicator function for reaching latent states at level $h+1$ can be written as some function f at level h. Although such f cannot be directly captured by the given the candidate feature class Φ, after one Bellman backup, it can be represented by a linear function of the true feature ϕ^*_{h-1} (at level $h-1$). This inspires us to construct the discriminator class $\mathcal{F}_{h-1}$ (the clipped version is defined in Eq. (1) and the unclipped version is defined in Eq. (10)), which includes $\left\langle \phi^*_{h-1}(x_{h-1}, a_{h-1}), \theta^*_f \right\rangle$. This overcomes the first challenge.

Now we discuss how to tackle the second challenge: finding good features ($\hat{\phi}_{h-2}$) to enable planning for the known rewards. Given the discriminator class, in the sequel, we specify the objective of our feature learning step. In Explore (Algorithm 2), we learn $\hat{\phi}_{h-2}$ so that for any appropriately bounded θ, there is a w such that

$$\mathbb{E}[\langle \phi^*_{h-1}(x_{h-1}, a_{h-1}), \theta \rangle \mid x_{h-2}, a_{h-2}] \approx \langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w \rangle, \tag{14}$$

More specifically, in the feature learning step (line 5 of Algorithm 2), we learn $\hat{\phi}_{h-2}$ such that, for some fixed scalar B and the discriminator class $\mathcal{F}_{h-1}$ discussed above, it satisfies

$$\max_{f \in \mathcal{F}_{h-1}} \mathrm{b\_err}\left(\rho^{+3}_{h-5}, \hat{\phi}_{h-2}, f; B\right) \leq \varepsilon_{\mathrm{reg}}, \tag{15}$$

where (recall that) $\mathrm{b\_err}(\cdot)$ is defined in Eq. (4).

Intuitively, it is guaranteed that we can find such a good $\hat{\phi}_{h-2} \in \Phi_{h-2}$ because again from Lemma 3, we know that true feature ϕ^*_{h-2} satisfies Eq. (14). We defer the detailed discussion on how to find $\hat{\phi}_{h-2}$ that satisfies Eq. (15) to the later part as we focus on computing exploratory policies here.

For building an (exploratory) mixture policy ρ_{h-2} that effectively covers all directions spanned by $\hat{\phi}_{h-2}$, we employ the *offline* "elliptical planner" for reward-free exploration (Huang et al., 2021) and optimize reward functions that are quadratic in the learned features $\hat{\phi}_{h-2}$. To do so, in Algorithm 4, we repeatedly (i) update the elliptical reward, (ii) invoke FQI subroutine (Algorithm 5) to obtain a greedy policy w.r.t. the elliptical reward, and (iii) call FQE (Fitted Q-Evaluation, Le et al. (2019)) subroutine (Algorithm 6) with the policy obtained from FQI to estimate the covariance matrix (to update the elliptical reward in the next iteration) and estimate the expected return (to check the stopping criterion). For both FQI and FQE, we use a function class comprising of all reward-appended linear functions of $\phi \in \Phi$.

Instead of using the offline "elliptical planner", we can also employ an online version, where we substitute the FQE component with Monte Carlo rollout. However, this requires

collecting additional samples and leads to worse sample complexity bound due to lack of data reuse. One appealing guarantee we can obtain by using our offline "elliptical planner" is the optimal deployment complexity formulated in Huang et al. (2021) and it can be easily verified. On the technical side, our "elliptical planner" builds on Huang et al. (2021), but is established without discretization over the value function class and the reward functions. This advance makes the algorithm more computationally friendly. However, we need to apply the more involved concentration analysis (uniform Bernstein's inequality) for the infinite function class to achieve sharp rates. We adapt the tools and analysis from Dong et al. (2020) and show a key concentration result in Corollary 43, which is then applied in the squared loss deviation result in Appendix F.2. We provide a complete description and analysis for the "elliptical planner" in Appendix B, and its induced guarantee for proving the distribution shift argument in Lemma 11. The related FQI (Algorithm 5) and FQE (Algorithm 6) analyses used in the "elliptical planner" are presented in Appendix D.4 and Appendix E respectively.

With the help of Lemma 11 and based on our earlier intuition for covering $\mathcal{Z}_{h+1}$ by translating the indicator reward at level $h+1$ to feature $\hat{\phi}_{h-2}$, we show that the policy $\rho_{h-2}^{+2} = \rho_{h-2} \circ \mathrm{unif}(\mathcal{A}) \circ \mathrm{unif}(\mathcal{A})$ is exploratory and hits all latent states $z \in \mathcal{Z}_{h+1}$,

$$\max_{\pi} \mathbb{P}_{\pi}\left[z_{h+1} = z\right] \leq \kappa \mathbb{P}_{\rho_{h-2}^{+2}}\left[z_{h+1} = z\right], \tag{16}$$

where $\kappa > 0$ is a constant specified in Theorem 8.

The formal proof is established in an inductive way, i.e., we assume Eq. (16) holds for all $h' \leq h$ and then show it also holds for $h+1$. The main reason is that we need exploratory policies/datasets at the prior levels to be fed into the offline "elliptical planner". For the induction base ($h = 0$), it is easy to verify that the null policy ρ_{-2}^{+2} satisfies the exploration guarantee in Eq. (16).

One thing that eluded the objective Eq. (15) for learning $\hat{\phi}$ is that we need exploratory policies. More specifically, we can see that Eq. (15) is defined with an exploratory policy ρ_{h-5}. Therefore, the goals of representation learning and exploration are intertwined. However, this is indeed not a concern since we need $\hat{\phi}_{h-2}$ when exploring/covering $\mathcal{Z}_{h+1}$ (or building ρ_{h-2}) while learning $\hat{\phi}_{h-2}$ only requires ρ_{h-5} (exploratory policies at the prior step). By our inductive proof, we can observe that ρ_{h-5} has already been established at this stage.

Also notice that we plan in the previously learned features $\hat{\phi}_{h-2}$ to obtain a cover over $\mathcal{Z}_{h+1}$. This way, planning trails feature learning like Flambe, but with an additional step of lag due to differences between model-free and model-based reasoning.

Based on the exploratory property in Eq. (16), taking another action a_{h+1} uniformly at random further returns an exploratory policy ρ_{h-2}^{+3} for state-action pairs (x_{h+1}, a_{h+1}). In summary, we provide the following result for the policies returned by Algorithm 2 with details in Section 8.2:

Theorem 8 *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Definition 1 and Assumption 1, Assumption 2 hold. If the features $\hat{\phi}_h$ learned in line 5 of* Explore *(Algorithm 2) satisfy the condition in Eq. (15) for $B \geq \sqrt{d}$, and $\varepsilon_{\mathrm{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2 K^9 \log^2(1+8/\beta)}\right)$, then with probability at least $1-\delta$, the sub-routine* Explore *collects an exploratory mixture policy ρ_{h-3}^{+3} for each level h such that*

$$\forall \pi, \forall f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+, \text{ we have } \mathbb{E}_{\pi}[f(x_h, a_h)] \leq \kappa K \mathbb{E}_{\rho_{h-3}^{+3}}[f(x_h, a_h)], \tag{17}$$

where $\kappa = \frac{64dK^4 \log(1+8/\beta)}{\eta_{\min}}$. *The total number of episodes used in line 7 by* Explore *is*

$$\tilde{O}\left(\frac{H^5 d^9 K^{14} B^4 \log(|\Phi|/\delta)}{\eta_{\min}^5}\right),$$

with β *chosen to satisfy* $\beta \log(1+8/\beta) \leq \frac{\eta_{\min}^2}{128dK^4B^2}$ *and a sufficient one is* $\beta = \tilde{O}\left(\frac{\eta_{\min}^2}{dK^4B^2}\right)$.

This result specifies the required size n_{ell} of dataset $\mathcal{D}^{\text{ell}}$ in Algorithm 4. We need to choose different values of B to guarantee approximation error bound for $\hat{\phi}$ (Eq. (15)) holds for different instantiations, and we discuss the details in the deviation bounds for learned features part.

The precise dependence on parameters d, H, K, and $\eta_{\min}$ is likely improvable. The exponent on K arises from multiple importance sampling steps over the uniform action choice and can be improved when the features $\phi(x,a) \in \Delta(d)$ for all x, a, and $\phi \in \Phi$ (Section 8.2.2). Improving these dependencies further is an interesting avenue for future progress.

Representation learning for downstream tasks For showing planning guarantees using FQI, we need to ensure that the following requirements stated in Chen and Jiang (2019) are satisfied: (i) (*concentrability*) we have adequate coverage over the state space, (ii) (*realizability*) we can express Q^* (more specifically it is Q^*_R, i.e., Q^* under the reward function R that we consider) with our function class, and (iii) (*completeness*) our class is closed under Bellman backups. Condition (i) is implied by Theorem 8. For (ii) and (iii), we learn features $\bar{\phi}_{0:H-1} \in \Phi$ such that $\bar{\phi}_h$ satisfies

$$\max_{g \in \mathcal{G}_{h+1}} \text{b\_err}\left(\rho_{h-3}^{+3}, \bar{\phi}_h, g; B\right) \leq \varepsilon_{\text{apx}}. \tag{18}$$

Here $\mathcal{G}_{h+1}$ as defined in Eq. (2) is the discriminator class containing all reward-appended linear candidate Q-value functions, which in turn includes the true Q^* value function for all $R \in \mathcal{R}$. The main conceptual difference over Eq. (15) is that the discriminator class $\mathcal{G}_{h+1}$ now incorporates reward information, which enables downstream planning. The objective for learning $\bar{\phi}$ (Eq. (18)) again includes the exploratory policies ρ, but they have been fully established in the exploration phase. Using Eq. (18) and the low-rank MDP properties, we show that this function class satisfies approximate realizability and approximate completeness, so we can invoke results for FQI and obtain the following representation learning guarantee. The details can be found in Section 8.3.

Theorem 9 *Fix* $\delta \in (0,1)$ *and consider an MDP* $\mathcal{M}$ *that satisfies Definition 1 and Assumption 1, Assumption 2 hold. If the features* $\bar{\phi}_{0:H-1}$ *learned by* Moffle *satisfy the condition in Eq.* (18) *for all* h *with* $\varepsilon_{\text{apx}} = \tilde{O}\left(\frac{\varepsilon^2 \eta_{\min}}{dH^4K^5}\right)$, *then for any reward function* $R \in \mathcal{R}$, *running* FQI *with the value function class* $\mathcal{Q}(\bar{\phi}, R)$ *in Eq.* (3) *and an exploratory dataset* $\mathcal{D}$, *returns a policy* $\hat{\pi}$, *which satisfies* $v_R^{\hat{\pi}} \geq v_R^* - \varepsilon$ *with probability at least* $1-\delta$. *The total number of episodes collected by* Moffle *in line 6 is:*

$$\tilde{O}\left(\frac{H^7 d^2 K^5 \log(|\Phi||\mathcal{R}|B/\delta)}{\varepsilon^2 \eta_{\min}}\right).$$

Deviation bounds for learned features One thing we skipped in the earlier discussion is how to establish guarantees for learning $\hat{\phi}$ that satisfies Eq. (15) and $\bar{\phi}$ that satisfies Eq. (18). The key technical component used in all proofs is the Bernstein's version of uniform concentration result (Corollary 43). With this careful concentration argument, in Lemma 16, we show that the min-max-min oracle subroutine (Eq. (5)) can be used to achieve these goals with appropriate choices of $n_{\hat{\phi}}$ and $n_{\bar{\phi}}$. The corresponding guarantees of the parameters for iterative greedy representation learning (Algorithm 3) is presented in Lemma 14, where the analysis is more complicated and we additionally use a potential argument to give the bound on its iteration complexity.

For feature learning in the enumerable case (Section 7), we only provide the guarantee to learn $\hat{\phi}$ that satisfies Eq. (15) with large enough $n_{\hat{\phi}}$ in Lemma 18, as we only present the reward-free exploration result for this version. We invoke the Bernstein's type concentration result on the ridge regression estimator. Therefore, we need to treat the scale of the regularizer and the bias term carefully, which leads to a worse rate in the sample complexity bound.

The detailed choices of $n_{\hat{\phi}}$ and $n_{\bar{\phi}}$ are calculated from the deviation bounds and the thresholds ε_{reg} (Theorem 8) and ε_{apx} (Theorem 9). Recall that previously we skipped the choices of B for setting n_{ell} in Theorem 9. The proper way to specify them are also discussed in the deviation results (Lemma 16, Lemma 14, and Lemma 18).

We also remark that our results naturally extend to approximate low-rank MDPs ($T_h(x_{h+1} \mid x_h, a_h) \approx \langle \phi_h^*(x_h, a_h), \mu_h^*(x_{h+1}) \rangle$) by following the standard misspecification analyses in Jin et al. (2020b); Zanette et al. (2020a). In the approximate case, feature ϕ_h^* satisfies approximate Bellman completeness when backing up functions at level $h+1$. The corresponding approximation errors will enter ε_{reg} in Eq. (15) and ε_{apx} Eq. (18), inducing an additive polynomially term in the performance difference bound.

Planning in downstream tasks We combine the representation learning guarantees with the sample complexity analysis for FQI to set the size n_{plan} of dataset $\mathcal{D}$ for planning in downstream tasks. The specific values are set according to Theorem 9. For the enumerable feature instance, we integrate the reward-free exploration guarantee and the FQI result for planning for a reward class with the full representation class (Corollary 20), which also gives us the choice of n_{plan}.

8.2 Proofs for Exploration and Sample Complexity Results for Algorithm 2

In this section, we present proofs for the exploration and sample complexity results for Explore (Algorithm 2). We provide the result for the low-rank setting in Section 8.2.1 and an improved result for the simplex feature setting in Section 8.2.2.

8.2.1 Proof of Theorem 8

Proof of Theorem 8 We will now prove the result assuming that the following condition from Eq. (15) is satisfied by $\hat{\phi}_h$ for all $h \in [H]$ with probability at least $1 - \delta/4$

$$\max_{f \in \mathcal{F}_{h+1}} \min_{\|w\|_2 \le B} \mathbb{E}_{\rho_{h-3}^{+3}} \left[\left(\left\langle \hat{\phi}_h(x_h, a_h), w \right\rangle - \mathbb{E}\left[f(x_{h+1}) \mid x_h, a_h \right] \right)^2 \right] \le \varepsilon_{\text{reg}}. \tag{19}$$

Now, let us turn to the inductive argument to show that the constructed policies ρ_{h-3}^{+3} are exploratory for every h. We will establish the following inductive statement for each

timestep h:

$$\forall z \in \mathcal{Z}_{h+1} : \max_{\pi} \mathbb{P}_{\pi}\left[z_{h+1} = z\right] \leq \kappa \mathbb{P}_{\rho_{h-2}^{+2}}\left[z_{h+1} = z\right]. \tag{20}$$

Assume that the exploration statement Eq. (20) is true for all timesteps $h' \leq h$. We first show an error guarantee similar to Eq. (17) under distribution shift:

Lemma 10 *If the inductive assumption in Eq.* (20) *is true for all $h' \leq h$, then for all $v : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ we have*

$$\forall \pi : \mathbb{E}_{\pi}[v(x_h, a_h)] \leq \kappa K \mathbb{E}_{\rho_{h-3}^{+3}}[v(x_h, a_h)]. \tag{21}$$

Proof Consider any timestep h and non-negative function v. Using the inductive assumption, we have

$$\begin{aligned}
\mathbb{E}_{\pi}\left[v(x_h, a_h)\right] &= \sum_{z \in \mathcal{Z}_h} \mathbb{P}_{\pi}[z_h = z] \cdot \int \mathbb{E}_{\pi_h}[v(x_h, a_h)] \nu^*(x_h \mid z) d(x_h) \\
&\leq \kappa \sum_{z \in \mathcal{Z}_h} \mathbb{P}_{\rho_{h-3}^{+2}}[z_h = z] \cdot \int \mathbb{E}_{\pi_h}[v(x_h, a_h)] \nu^*(x_h \mid z) d(x_h) \\
&= \kappa \mathbb{E}_{\rho_{h-3}^{+2}}[\mathbb{E}_{\pi_h}[v(x_h, a_h)]] \\
&\leq \kappa K \mathbb{E}_{\rho_{h-3}^{+3}}[v(x_h, a_h)].
\end{aligned}$$

Therefore, the result holds for any policy π, timestep $h' \leq h$, and non-negative function v. ■

Choosing $v(x_h, a_h) = \left(\left\langle \hat{\phi}_h(x_h, a_h), w \right\rangle - \mathbb{E}\left[f(x_{h+1}) \mid x_h, a_h\right]\right)^2$ and using the feature learning guarantee in Eq. (19) along with Eq. (21), we have

$$\forall \pi, \forall f \in \mathcal{F}_{h+1} : \min_{\|w\|_2 \leq B} \mathbb{E}_{\pi}\left[\left(\left\langle \hat{\phi}_h(x_h, a_h), w \right\rangle - \mathbb{E}\left[f(x_{h+1}) \mid x_h, a_h\right]\right)^2\right] \leq \kappa K \varepsilon_{\mathrm{reg}}. \tag{22}$$

We now outline our key argument to establish exploration: Fix a latent variable $z \in \mathcal{Z}_{h+1}$ and let $\pi \coloneqq \pi_h$ be the policy which maximizes $\mathbb{P}_{\pi}[z_{h+1} = z]$. Thus, with

$$f(x_h, a_h) = \mathbb{P}_{\pi}[z_{h+1} = z \mid x_h, a_h]$$

we have

$$\begin{aligned}
\mathbb{E}_{\pi}\left[f(x_h, a_h)\right] &\leq K^2 \mathbb{E}_{\pi_{h-2} \circ \mathrm{unif}(\mathcal{A}) \circ \mathrm{unif}(\mathcal{A})}\left[f(x_h, a_h)\right] \\
&= K^2 \mathbb{E}_{\pi_{h-2}}\left[\mathbb{E}_{\mathrm{unif}(\mathcal{A})}\left[g(x_{h-1}, a_{h-1}) \mid x_{h-2}, a_{h-2}\right]\right] \\
&\leq K^2 \mathbb{E}_{\pi_{h-2}}\left[\left|\langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w_g \rangle\right|\right] + \sqrt{\kappa K^5 \varepsilon_{\mathrm{reg}}}.
\end{aligned} \tag{23}$$

The first inequality follows by using importance weighting on timesteps $h-1$ and h, where we choose actions uniformly at random among $\mathcal{A}$. In the next step, we define

$$g(x_{h-1}, a_{h-1}) = \mathbb{E}_{\mathrm{unif}(\mathcal{A})}[f(x_h, a_h) \mid x_{h-1}, a_{h-1}] = \left\langle \phi^*_{h-1}(x_{h-1}, a_{h-1}), \theta^*_f \right\rangle$$

with $\|\theta_f^*\|_2 \le \sqrt{d}$. For the last inequality, we first use the result from Lemma 10 that $\hat{\phi}_{h-2}$ has a small squared loss for the regression target specified by $g(\cdot)$ with a vector w_g defined as

$$w_g := \operatorname*{argmin}_{\|w\|_2 \le B} \mathbb{E}_{\rho_{h-5}^{+3}} \left[\left(\left\langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w \right\rangle - \mathbb{E}_{\text{unif}(\mathcal{A})} \left[g(x_{h-1}, a_{h-1}) \mid x_{h-2}, a_{h-2} \right] \right)^2 \right].$$

We further use Eq. (22) to translate the error from ρ_{h-5}^{+3} to π_{h-2} and apply the weighted RMS-AM inequality in the same step to bound the mean absolute error using the squared error bound.

Lemma 11 *If the offline elliptical planner (Algorithm 4) is called with a sample of size*

$$\tilde{O}\left(\frac{H^4 d^6 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2} \right),$$

then with probability at least $1-\delta$, for all $h \in [H]$, we have

$$\mathbb{E}_{\pi_{h-2}} \left[\left| \langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w_g \rangle \right| \right] \le \frac{\alpha}{2} \mathbb{E}_{\rho_{h-2}} \left[\left(\langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w_g \rangle \right)^2 \right] + \frac{T\beta}{2\alpha} + \frac{\alpha \|w_g\|_2^2}{2T} + \frac{\alpha\beta \|w_g\|_2^2}{2}.$$

Proof Applying Cauchy-Schwarz inequality followed by AM-GM, for any matrix $\widehat{\Sigma}$, we have

$$\begin{aligned} \mathbb{E}_{\pi_{h-2}} \left[\left| \langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w_g \rangle \right| \right] &\le \mathbb{E}_{\pi_{h-2}} \left[\left\| \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}) \right\|_{\widehat{\Sigma}^{-1}} \cdot \|w_g\|_{\widehat{\Sigma}} \right] \\ &\le \frac{1}{2\alpha} \mathbb{E}_{\pi_{h-2}} \left[\left\| \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}) \right\|_{\widehat{\Sigma}^{-1}}^2 \right] + \frac{\alpha}{2} \|w_g\|_{\widehat{\Sigma}}^2 . \end{aligned}$$

Here, we choose $\widehat{\Sigma}$ to be the (normalized) matrix returned by the elliptic planner in Algorithm 4. As can be seen in the algorithm pseudocode, $\widehat{\Sigma}$ is obtained by summing up a (normalized) identity matrix and the empirical estimates of the population covariance matrix $\Sigma_{\pi_\tau} = \mathbb{E}_{\pi_\tau} \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}) \hat{\phi}_{h-2}(x_{h-2}, a_{h-2})^\top$, where $\{\pi_\tau\}_{1 \le \tau \le T}$ are the T policies computed by the planner. Noting that ρ_{h-2} is a mixture of these T policies, we consider the following empirical and population quantities:

$$\Sigma_{\rho_{h-2}} = \frac{1}{T} \sum_{t=1}^{T} \Sigma_{\pi_t}, \qquad \Sigma = \Sigma_{\rho_{h-2}} + \frac{1}{T} I_{d \times d}, \qquad \widehat{\Sigma} = \frac{1}{T} \Gamma_T = \frac{1}{T} \sum_{i=1}^{T} \widehat{\Sigma}_{\pi_i} + \frac{1}{T} I_{d \times d}.$$

Now, we use the termination conditions satisfied by the elliptic planner (shown in Lemma 15) in the following steps:

$$\begin{aligned} &\mathbb{E}_{\pi_{h-2}} \left[\left| \langle \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}), w_g \rangle \right| \right] \\ &\le \frac{1}{2\alpha} \mathbb{E}_{\pi_{h-2}} \left[\left\| \hat{\phi}_{h-2}(x_{h-2}, a_{h-2}) \right\|_{\widehat{\Sigma}^{-1}}^2 \right] + \frac{\alpha}{2} \|w_g\|_{\widehat{\Sigma}}^2 \qquad (24) \\ &\le \frac{T\beta}{2\alpha} + \frac{\alpha}{2} \|w_g\|_{\widehat{\Sigma}}^2 \le \frac{T\beta}{2\alpha} + \frac{\alpha}{2} \|w_g\|_{\Sigma}^2 + \frac{\alpha}{2} \beta \|w_g\|_2^2 \qquad (25) \end{aligned}$$

$$= \frac{T\beta}{2\alpha} + \frac{\alpha}{2}\mathbb{E}_{\rho_{h-2}}\left[\left(\langle\hat{\phi}_{h-2}(x_{h-2},a_{h-2}),w_g\rangle\right)^2\right] + \frac{\alpha\|w_g\|_2^2}{2T} + \frac{\alpha\beta\|w_g\|_2^2}{2}.$$

For the second inequality, note that $\frac{1}{T}\left\|\hat{\phi}_{h-2}(x_{h-2},a_{h-2})\right\|^2_{\widehat{\Sigma}^{-1}}$ is the reward function optimized by the offline elliptical planner in the last iteration. Let v_T^π denote the expected return of the policy π for this reward function and MDP $\mathcal{M}$. From the termination condition and the results for the offline elliptical planner in Lemma 15, we get

$$\max_\pi v_T^\pi \le v_T^{\pi_T} + \beta/8 \le \hat{v}_T^{\pi_T} + \beta/4 \le \beta.$$

Therefore, the first term on the RHS in Eq. (24) can be bounded by $T\beta/(2\alpha)$. In Eq. (25), we use the estimation guarantee for $\Sigma = \Gamma_T/T$ for the FQI planner shown in Lemma 15. Then, in the last equality step, we expand the norm of w_g using the definition of Σ to arrive at the desired result.

Putting everything together, we now compute the number of samples used during elliptical planning for the required error tolerance. Lemma 15 states that for a sample of size n, the computed policy is sub-optimal by a value difference of order upto $\tilde{O}\left(\sqrt{\frac{H^4d^6\kappa K^2\log(|\Phi|/\delta)}{n}}\right)$. Setting the failure probability of elliptical planning to be $\delta/(4H)$ for each level $h \in [H]$, and setting the planning error to $\beta/8$, we conclude that the total number of episodes used by Algorithm 4 for each timestep h is $\tilde{O}\left(\frac{H^4d^6\kappa K^2\log(|\Phi|/\delta)}{\beta^2}\right)$. ■

Using Lemma 11 in Eq. (23), we get

$$\begin{aligned}
\mathbb{E}_\pi\left[f(x_h,a_h)\right] &\le \frac{\alpha K^2}{2}\mathbb{E}_{\rho_{h-2}}\left[\left(\langle\hat{\phi}_{h-2}(x_{h-2},a_{h-2}),w_g\rangle\right)^2\right] + \frac{\beta K^2 T}{2\alpha} + \frac{\alpha K^2\|w_g\|_2^2}{2T} \\
&\quad + \frac{\alpha\beta K^2\|w_g\|_2^2}{2} + \sqrt{\kappa K^5\varepsilon_{\text{reg}}} \\
&\le \alpha K^2\mathbb{E}_{\rho_{h-2}}\left[\left(\langle\phi^*_{h-2}(x_{h-2},a_{h-2}),\theta^*_g\rangle\right)^2\right] + \alpha\kappa K^3\varepsilon_{\text{reg}} + \frac{K^2T\beta}{2\alpha} + \frac{\alpha K^2\|w_g\|_2^2}{2T} \\
&\quad + \frac{\alpha\beta K^2\|w_g\|_2^2}{2} + \sqrt{\kappa K^5\varepsilon_{\text{reg}}}. \qquad (26)
\end{aligned}$$

The second inequality uses the approximation guarantee for features $\hat{\phi}_{h-2}$ in Eq. (22) (derived from Eq. (19)), the definition of w_g, and the inequality $(a+b)^2 \le 2a^2 + 2b^2$. Finally, we note that the inner product inside the expectation is always bounded between $[0,1]$ which allows use to use the fact that $f(x)^2 \le f(x)$ for $f : \mathcal{X} \to [0,1]$. Substituting the upper bound for $\|w_g\|_2$, we get

$$\begin{aligned}
&\mathbb{E}_\pi\left[f(x_h,a_h)\right] \\
&\le \alpha K^2\mathbb{E}_{\rho_{h-2}}\left[\langle\phi^*_{h-2}(x_{h-2},a_{h-2}),\theta^*_g\rangle\right] + \alpha\kappa K^3\varepsilon_{\text{reg}} + \sqrt{\kappa K^5\varepsilon_{\text{reg}}} \\
&\quad + \frac{\beta K^2T}{2\alpha} + \frac{\alpha\beta K^2B^2}{2} + \frac{\alpha K^2B^2}{2T} \\
&= \alpha K^2\mathbb{P}_{\rho^{+2}_{h-2}}\left[z_{h+1} = z\right] + \alpha\kappa K^3\varepsilon_{\text{reg}} + \sqrt{\kappa K^5\varepsilon_{\text{reg}}} + \frac{\beta K^2T}{2\alpha} + \frac{\alpha\beta K^2B^2}{2} + \frac{\alpha K^2B^2}{2T}. \qquad (27)
\end{aligned}$$

Eq. (27) follows by the definition of the function $g(\cdot)$.

We now set $\kappa \geq 2\alpha K^2$ in Eq. (27). Therefore, if we set the parameters $\alpha, \beta, \varepsilon_{\text{reg}}$ such that

$$\max\left\{\alpha\kappa K^3\varepsilon_{\text{reg}} + \sqrt{\kappa K^5 \varepsilon_{\text{reg}}}, \frac{\beta K^2 T}{2\alpha}, \frac{\alpha\beta K^2 B^2}{2}, \frac{\alpha K^2 B^2}{2T}\right\} \leq \eta_{\min}/8, \tag{28}$$

Eq. (27) can be re-written as

$$\max_{\pi} \mathbb{P}_\pi\left[z_{h+1} = z\right] \leq \frac{\kappa}{2}\mathbb{P}_{\rho_{h-2}^{+2}}\left[z_{h+1} = z\right] + \frac{\eta_{\min}}{2} \leq \kappa \mathbb{P}_{\rho_{h-2}^{+2}}\left[z_{h+1} = z\right]$$

where in the last step, we use Assumption 1. Hence, we prove the exploration guarantee in Theorem 8 by induction.

To find the feasible values for the constants in Eq. (28), we first note that $T \leq 8d\log\left(1 + 8/\beta\right)/\beta$ (Lemma 15). We start by setting $\frac{\beta K^2 T}{2\alpha} = \eta_{\min}/8$ which gives $\alpha/T = \frac{4\beta K^2}{\eta_{\min}}$. Using the upper bound on T, we get $\alpha \leq \frac{32dK^2\log(1+8/\beta)}{\eta_{\min}}$. Next, we set the term $\alpha\kappa K^3\varepsilon_{\text{reg}} + \sqrt{\kappa K^5\varepsilon_{\text{reg}}} \leq \eta_{\min}/8$. Using the value of $\kappa = 2\alpha K^2$ we get

$$2\alpha^2 K^5 \varepsilon_{\text{reg}} + \sqrt{2\alpha K^7 \varepsilon_{\text{reg}}} \leq \eta_{\min}/8,$$

which is satisfied by $\varepsilon_{\text{reg}} = \Theta\left(\frac{\eta_{\min}^3}{d^2K^9\log^2(1+8/\beta)}\right)$.

Lastly, we will consider the term $\frac{\alpha\beta K^2 B^2}{2}$ and by setting it less than $\eta_{\min}/8$, we get

$$\beta\log\left(1 + 8/\beta\right) \leq \frac{\eta_{\min}^2}{128dB^2K^4}.$$

One can verify that under this condition we also have $\frac{\alpha K^2 B^2}{2T} \leq \eta_{\min}/8$, and setting $\beta = \tilde{O}\left(\frac{\eta_{\min}^2}{dB^2K^4}\right)$ satisfies the feasibility constraint for β. Here, we assume that B only has a polylog dependence on β and show later that this is true for all our feature selection methods. Notably, the only cases when B depends on β in our results is when $B = O\left(\frac{1}{\varepsilon_{\text{reg}}^c}\right)$ for a constant $c = \{1/2, 1\}$ which has a $\log^2(1 + 8/\beta)$ term.

Substituting the value of κ and β in Lemma 11 with an additional factor of H to account for all h gives us the final sample complexity bound in Theorem 8. The change of measure guarantee (Eq. (17)) follows from the result in Lemma 10.

8.2.2 Improved Sample Complexity Bound for Simplex Features

We can obtain more refined results when the agent instead has access to a latent variable feature class $\{\Psi_h\}_{h\in[H]}$ with $\psi_h : \mathcal{X}\times\mathcal{A} \to \Delta(d_{\text{LV}})$. We call this the *simplex features* setting (Agarwal et al., 2020b) and show the improved results in this section. For notation simplicity, we still use Φ_h and ϕ_h to represent the features. In order to achieve this improved result, we make two modifications to Explore: (i) We use a smaller discriminator function class $\mathcal{F}_{h+1} \coloneqq \{f(x_{h+1}, a_{h+1}) = \mathbb{E}_{\text{unif}(\mathcal{A})}[\phi_{h+1}(x_{h+1}, a_{h+1})[i]] : \phi_{h+1} \in \Phi_{h+1}, i \in [d_{\text{LV}}]\}$ and (ii) in Explore, instead of calling the planner with learned features $\hat{\phi}_{h-2}$ and taking three uniform actions, we plan for the features $\hat{\phi}_{h-1}$ and add two uniform actions to collect data for feature

learning in timestep h. The key idea here is that instead of estimating the expectation of any bounded function f, we only need to focus on the expectation of coordinates of ϕ^* as included in class $\mathcal{F}_{h+1}$. Further, since $\phi^*_{h+1}[i]$ is already a linear function of the feature ϕ^*_{h+1}, we take only one action at random at timestep h.

Theorem 12 (Exploration with simplex features) *Fix $\delta \in (0,1)$. Consider an MDP $\mathcal{M}$ which admits a low-rank factorization with dimension d in Definition 1 and satisfies Assumption 1. If Assumption 2 holds, the features $\hat{\phi}_h$ learned in line 5 in Algorithm 2 satisfy the condition in Eq. (15) for $B \geq \sqrt{d}$, and $\varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^5\log^2(1+8/\beta)}\right)$, then with probability at least $1-\delta$, the sub-routine* Explore *collects an exploratory mixture policy ρ_{h-3}^{+3} for each level h such that*

$$\forall \pi : \mathbb{E}_\pi[f(x_h, a_h)] \leq \kappa K \mathbb{E}_{\rho_{h-3}^{+3}}[f(x_h, a_h)] \tag{29}$$

for any $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ and $\kappa = \frac{64dK^2\log(1+8/\beta)}{\eta_{\min}}$. The total number of episodes used in line 7 by Algorithm 2 is

$$\tilde{O}\left(\frac{H^5 d^9 K^8 B^4 \log(|\Phi|/\delta)}{\eta_{\min}^5}\right).$$

β is chosen such that $\beta \log(1+8/\beta) \leq \frac{\eta_{\min}^2}{128dK^2B^2}$ and a sufficient one is $\beta = \tilde{O}\left(\frac{\eta_{\min}^2}{dK^4B^2}\right)$.

Proof For simplex features, the key observation is that for any latent state $z \in \mathcal{Z}_{h+1}$, the function $f(x_h) = \mathbb{E}_{\text{unif}(\mathcal{A})}[\mathbb{P}[z_{h+1} = z | x_h, a_h]]$ is already a member of the discriminator function class $\mathcal{F}_h := \{f(x_h) = \mathbb{E}_{\text{unif}(\mathcal{A})}[\phi_h(x_h, a_h)[i]] : \phi_h \in \Phi_h, i \in [d_{\text{LV}}]\}$. Thus, when we rewrite the term $\mathbb{E}_\pi[f(x_h, a_h)]$ as a linear function, we only need to backtrack one timestep to use the feature selection guarantee

$$\begin{aligned}\mathbb{E}_\pi[f(x_h, a_h)] &\leq K\mathbb{E}_{\pi_{h-1}\circ\,\text{unif}(\mathcal{A})}[f(x_h, a_h)] = K\mathbb{E}_{\pi_{h-1}}[g(x_{h-1}, a_{h-1})] \\ &\leq K\mathbb{E}_{\pi_{h-1}}\left[\left|\langle\hat{\phi}_{h-1}(x,a), w_g\rangle\right|\right] + \sqrt{\kappa K^3 \varepsilon_{\text{reg}}},\end{aligned} \tag{30}$$

where we define $g(x_{h-1}, a_{h-1}) = \mathbb{E}_{\text{unif}(\mathcal{A})}[f(x_h, a_h)|x_{h-1}, a_{h-1}]$. Therefore, the new value of κ becomes $2\alpha K$ and by shaving off this K factor in the chain of inequalities, we get the following constraint set for the parameters:

$$\max\left\{\alpha\kappa K^2\varepsilon_{\text{reg}} + \sqrt{\kappa K^3\varepsilon_{\text{reg}}}, \frac{\beta KT}{2\alpha}, \frac{\alpha\beta KB^2}{2}, \frac{\alpha KB^2}{2T}\right\} \leq \eta_{\min}/8. \tag{31}$$

Thus, the values of these parameters for the simplex features case are as follows:

$$\frac{\alpha}{T} = \frac{4\beta K}{\eta_{\min}}, \quad \alpha \leq \frac{32dK\log(1+8/\beta)}{\eta_{\min}}, \quad \varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^5\log^2(1+8/\beta)}\right).$$

Hence, the updated constraint for β is

$$\beta\log(1+8/\beta) \leq \frac{\eta_{\min}^2}{64dB^2K^2}.$$

Other than the values for these parameters, the algorithm remains the same. Therefore, substituting the new values of κ and β in the expression $\tilde{O}\left(\frac{H^4 d^6 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2}\right)$ as before, we get the improved sample complexity result. ■

8.3 Proofs for Representation Learning Guarantees for Downstream Tasks

We show that after obtaining the exploratory policies ρ_{h-3}^{+3} for all $h \in [H]$ using MOFFLE, we can collect a dataset $\mathcal{D}$ to learn a feature $\bar{\phi}_h \in \Phi_h$ for all levels and use FQI to plan for any reward function $R \in \mathcal{R}$. Specifically, with min-max-min oracle or iterative greedy representation learning, we compute a feature $\bar{\phi}_h \in \Phi_h$ such that

$$\max_{g \in \mathcal{G}_{h+1}} \min_{\|w\|_2 \le B} \mathbb{E}_{\rho_{h-3}^{+3}} \left[\left(\left\langle \bar{\phi}_h(x_h, a_h), w \right\rangle - \mathbb{E}\left[g(x_{h+1}) \mid x_h, a_h \right] \right)^2 \right] \le \varepsilon_{\text{apx}}, \tag{32}$$

where $\mathcal{G}_{h+1} \subseteq (\mathcal{X} \to [0, H])$ is defined in Eq. (2). For ease of discussion, we also present it here: $\mathcal{G}_{h+1} := \Big\{ \text{clip}_{[0,H]}\Big(\max_a (R_{h+1}(x_{h+1}, a) + \langle \phi_{h+1}(x_{h+1}, a), \theta \rangle) \Big) : R \in \mathcal{R}, \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le B \Big\}$, where $B \ge H\sqrt{d}$. Also recall that $\mathcal{Q}(\bar{\phi}, R)$ as defined in Eq. (3) is: $\mathcal{Q}(\bar{\phi}, R) := \bigcup_{h \in [H]} \mathcal{Q}_h(\bar{\phi}_h, R_h)$, $\mathcal{Q}_h(\bar{\phi}_h, R_h) := \Big\{ \text{clip}_{[0,H]}(R_h(x_h, a_h) + \langle \bar{\phi}_h(x_h, a_h), w \rangle) : \|w\|_2 \le B \Big\}$.

The learned feature serves two purposes as discussed before:

- (*realizability*) The optimal value function for any timestep $h+1$ and reward $R_{h+1} \in \mathcal{R}$, is defined as $V_{h+1}^*(x') = \max_a \left(R_{h+1}(x', a) + \mathbb{E}[Q_{h+2}^*(\cdot) \mid x', a] \right) = \max_a \left(R_{h+1}(x', a) + \langle \phi_{h+1}^*, \theta_{h+1}^* \rangle \right)$. Thus, we have realizability as $V_{h+1}^* \in \mathcal{G}_{h+1}$, which in turn implies that $\exists Q_h \in \mathcal{Q}_h(\bar{\phi}_h, R_h)$, s.t. $Q_h \approx R_h + \mathbb{E}[V_{h+1}^*(\cdot)]$.
- (*completeness*) For completeness, note that $\mathcal{G}_{h+1}$ contains the Bellman backup of all possible $Q_{h+1}(\cdot)$ value functions we may encounter while running FQI with $\mathcal{Q}(\bar{\phi}, R)$. Therefore, for any such Q_{h+1}, we have that $\exists Q_h \in \mathcal{Q}_h(\bar{\phi}_h, R_h)$, s.t. $Q_h \approx \mathcal{T} Q_{h+1}$.

Proof of Theorem 9 We run FQI with the learned representation $\bar{\phi}_h$ using the value function class $\mathcal{Q}_h(\bar{\phi}_h, R_h)$ defined for each $h \in [H]$. Lemma 22 shows that when Eq. (32) is satisfied with an error ε_{apx}, running FQI using a total of $n_h = \tilde{O}\left(\frac{H^6 d \kappa K \log(|\mathcal{R}|B/\delta')}{\beta^2}\right)$ episodes collected from each exploratory policy $\{\rho_{h-3}^{+3}\}$ returns a policy $\hat{\pi}$ which satisfies

$$\mathbb{E}_{\hat{\pi}} \left[\sum_{h=0}^{H-1} R_h(x_h, a_h) \right] \ge \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{h=0}^{H-1} R_h(x_h, a_h) \right] - \beta - H^2 \sqrt{\kappa K \varepsilon_{\text{apx}}}$$

with probability at least $1 - \delta'$.

Then union bounding over all possible $\bar{\phi}$, and setting $\delta = \delta'/|\Phi|$, $\beta = \varepsilon/2$, $\varepsilon_{\text{apx}} = \frac{\varepsilon^2}{16 H^4 \kappa K}$, we get the final planning result with a value error of ε and probability at least $1 - \delta$. Substituting $\kappa = \tilde{O}\left(\frac{32 d K^4}{\eta_{\min}}\right)$, we get $n_h = \tilde{O}\left(\frac{H^6 d^2 K^5 \log(|\Phi||\mathcal{R}|B/\delta)}{\varepsilon^2 \eta_{\min}}\right)$. The final sample complexity is $\tilde{O}\left(\frac{H^7 d^2 K^5 \log(|\Phi||\mathcal{R}|B/\delta)}{\varepsilon^2 \eta_{\min}}\right)$, where we sum up the collected episodes across all levels. ■

8.4 Proofs for Oracle Representation Learning

In this section, we present the sample complexity result and the proof for MOFFLE when a computational oracle FLO is available. Since we need to set $B \geq L\sqrt{d}$ in the min-max-min objective (Eq. (5)), we assume FLO solves Eq. (5) with $B = L\sqrt{d}$. The computational oracle is defined as follows:

Definition 13 (Optimization oracle, FLO) *Given a feature class Φ_h and an abstract discriminator class $\mathcal{V} \subseteq (\mathcal{X} \to [0, L])$, we define the Feature Learning Oracle (FLO) as a subroutine that takes a dataset $\mathcal{D}$ of tuples (x_h, a_h, x_{h+1}) and returns a solution to the following objective:*

$$\hat{\phi}_h = \operatorname*{argmin}_{\phi \in \Phi_h} \max_{v \in \mathcal{V}} \left\{ \min_{\|w\|_2 \leq L\sqrt{d}} \mathcal{L}_{\mathcal{D}}(\phi, w, v) - \min_{\tilde{\phi} \in \Phi_h, \|\tilde{w}\|_2 \leq L\sqrt{d}} \mathcal{L}_{\mathcal{D}}(\tilde{\phi}, \tilde{w}, v) \right\}. \tag{33}$$

With this definition of FLO, we will use the sample complexity result in Lemma 16, shown for the min-max-min objective (Eq. (5)) against a general discriminator function class $\mathcal{V}$ consisting of the set of functions

$$v(x_{h+1}) = \text{clip}_{[0,L]}(\mathbb{E}_{a_{h+1} \sim \pi_{h+1}(x_{h+1})}[R(x_{h+1}, a_{h+1}) + \langle \phi_{h+1}(x_{h+1}, a_{h+1}), \theta \rangle])$$

where $\phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \leq L\sqrt{d}, R \in \mathcal{R}$ for a prespecified policy π_{h+1} over x_{h+1}. Note that, $\mathcal{F}_h$ in the main text uses a singleton reward class $R(x_{h+1}, a_{h+1}) = 0$ with $L = 1$ and $\pi_{h+1} = \text{unif}(\mathcal{A})$. Similarly, $\mathcal{G}_h$ uses $L = H$ with π_{h+1} as the greedy arg-max policy.

Sample Complexity of MOFFLE with Min-Max-Min Oracle We now give a proof for the final sample complexity result for MOFFLE as instantiated with the oracle FLO.

Proof of Theorem 4 Let us start with any fixed $h \in [H]$ and calculate the required number of samples per level.

Firstly, we consider learning $\hat{\phi}_h$ that satisfies Eq. (15). We use the discriminator class $\mathcal{V} = \mathcal{F}$ as defined in Eq. (1) and set $B = \sqrt{d}$. Then applying Lemma 16 with $L = 1$, we know that condition Eq. (15) holds with probability at least $1 - \delta/(4H)$, if

$$n \geq \frac{16 c_3 d^2 \log(2n\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/(\delta/4H))}{\varepsilon_{\text{reg}}},$$

where c_3 is the constant in Lemma 34.

Setting $\varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2 K^9 \log^2(1+8/\beta)}\right)$ and noting $\beta = \tilde{O}\left(\frac{\eta_{\min}^2}{dK^4 B^2}\right)$, we get

$$n_{\hat{\phi}} = \tilde{O}\left(\frac{d^2 \log(|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{\varepsilon_{\text{reg}}}\right) = \tilde{O}\left(\frac{d^4 K^9 \log(|\Phi|/\delta)}{\eta_{\min}^3}\right).$$

Substituting the value $B = \sqrt{d}$ in Theorem 8, we know that we can get an exploratory dataset with probability at least $1 - \delta/(4H)$ and the corresponding sample complexity for the elliptic planner is

$$n_{\text{ell}} = \tilde{O}\left(\frac{H^5 d^9 K^{14} B^4 \log(|\Phi|/\delta)}{\eta_{\min}^5}\right) = \tilde{O}\left(\frac{H^5 d^{11} K^{14} \log(|\Phi|/\delta)}{\eta_{\min}^5}\right)$$

Then we consider learning $\bar{\phi}_h$ that satisfies Eq. (18). We use the discriminator class $\mathcal{V} = \mathcal{G}$ as defined in Eq. (2) and set $B = \sqrt{d}$. Noticing that $\kappa = \frac{64dK^4 \log(1+8/\beta)}{\eta_{\min}}$ and β is a polynomial term, we know that $\frac{\varepsilon^2}{16H^4\kappa K} = \tilde{O}\left(\frac{\varepsilon^2\eta_{\min}}{dH^4K^5}\right)$. Setting $\varepsilon_{\text{apx}} = \frac{\varepsilon^2}{16H^4\kappa K}$ and applying Lemma 16 with $L = H$, we have that condition Eq. (18) is satisfied with probability at least $1 - \delta/(4H)$ if

$$n_{\bar{\phi}} = \tilde{O}\left(\frac{H^6 d^3 K^5 \log\left(|\Phi||\mathcal{R}|/\delta\right)}{\varepsilon^2 \eta_{\min}}\right).$$

Notice that Eq. (18) holds and we collect an exploratory dataset by applying Theorem 8. Then Theorem 9 implies the required sample complexity for offline FQI planning with $\bar{\phi}_{0:H-1}$ to learn an ε-optimal policy with probability at least $1 - \delta/(4H)$ is

$$n_{\text{plan}} = \tilde{O}\left(\frac{H^6 d^2 K^5 \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2 \eta_{\min}}\right).$$

Union bounding over $h \in [H]$, the final sample complexity is $H(n_{\hat{\phi}} + n_{\text{ell}} + n_{\bar{\phi}} + n_{\text{plan}})$, and the result holds with probability $1 - \delta$. Reorganizing terms completes the proof. ■

8.5 Proofs for Iterative Greedy Representation Learning Method

We start by showing the main iteration complexity result and a feature selection guarantee for Algorithm 3 below.

Lemma 14 (Iteration complexity for Algorithm 3) *Fix $\delta \in (0,1)$. If the iterative greedy feature selection algorithm (Algorithm 3) is run with a sample $\mathcal{D}$ of size $n = \tilde{O}\left(\frac{L^6 d^7 \log(|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{\varepsilon_{\text{tol}}^3}\right)$, then with $B = \sqrt{\frac{13L^4d^3}{\varepsilon_{\text{tol}}}}$, it terminates after $T = \frac{52L^2d^2}{\varepsilon_{\text{tol}}}$ iterations and returns a feature $\hat{\phi}_h$ such that for $\mathcal{V} \subseteq (\mathcal{X} \to [0, L])$, $\mathcal{V} \coloneqq \{v(x_{h+1}) = \text{clip}_{[0,L]}(\mathbb{E}_{a_{h+1}\sim\pi_{h+1}(x_{h+1})}[R_{h+1}(x_{h+1}, a_{h+1}) + \langle\phi_{h+1}(x_{h+1}, a_{h+1}), \theta\rangle]) : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \leq L\sqrt{d}, R \in \mathcal{R}\}$, where policy π is the greedy policy or the uniform policy, we have*

$$\max_{v\in\mathcal{V}} \text{b\_err}\left(\rho_{h-3}^{+3}, \hat{\phi}_h, v; B\right) \leq \varepsilon_{\text{tol}}.$$

Proof For ease of notation, we will not use the subscript ρ_{h-3}^{+3} in the expectations below ($\mathcal{L}(\cdot) \coloneqq \mathcal{L}_{\rho_{h-3}^{+3}}(\cdot)$). Similarly, we will use ϕ_t to denote feature $\phi_{t,h}(x_h, a_h)$ of iteration t and (x', a') for (x_{h+1}, a_{h+1}) unless required by context. Further, for any iteration t, 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_t^i as the linear parameter $w_{t,i} = \text{argmin}_{\|w\|_2 \leq L\sqrt{d}} \mathcal{L}_{\mathcal{D}}(\hat{\phi}_{t,h}, w, v_i)$. Similarly, let $A_t = [\theta_1^* \mid \theta_2^* \mid \ldots \mid \theta_t^*]$.

In the proof, we assume that the total number of iterations T does not exceed $\frac{52L^2d^2}{\varepsilon_{\text{tol}}}$ and set parameters accordingly. We later verify that this assumption holds. Further, let $\tilde{\varepsilon} = \frac{\varepsilon_{\text{tol}}^2}{2704L^2d^3}$ and $\varepsilon_0 = T_{\max} \cdot \tilde{\varepsilon} = \frac{\varepsilon_{\text{tol}}}{52d}$.

To begin, based on the deviation bound in Lemma 17, we note that if the sample $\mathcal{D}$ in Algorithm 3 is of size $n = \tilde{O}\left(\frac{L^6 d^7 \log(|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{\varepsilon_{\text{tol}}^3}\right)$ and the termination loss cutoff is set to $3\varepsilon_1/2 + \tilde{\varepsilon}$ such that, with probability at least $1-\delta$, for all non-terminal iterations t we have

$$\sum_{v_i \in \mathcal{V}^t} \mathbb{E}\left[\left(\hat{\phi}_t^\top W_t^i - \phi^{*\top} A_t^i\right)^2\right] \leq t\tilde{\varepsilon} \leq \varepsilon_0, \tag{34}$$

$$\mathbb{E}\left[\left(\hat{\phi}_t^\top w - \phi^{*\top} \theta_{t+1}^*\right)^2\right] \geq \varepsilon_1 \tag{35}$$

where $\tilde{\varepsilon}$ is an error term dependent on the size of $\mathcal{D}$ and w is any vector with $\|w\|_2 \leq B_t \leq B_T \leq B$. Further, when the algorithm does terminate, we get the loss upper bound to be $3\varepsilon_1 + 4\tilde{\varepsilon}$.

Using Eq. (34) and Eq. (35), we will now show that the maximum iterations in Algorithm 3 is bounded. At round t, for functions $v_1, \ldots, v_t \in \mathcal{V}$ in Algorithm 3, let $\theta_i^* = \theta_{v_i}^*$ as before and further let $\Sigma_t = A_t A_t^\top + \lambda I_{d\times d}$. Using the linear parameter θ_{t+1}^* of the adversarial test function v_{t+1}, define $\hat{w}_t = W_t A_t^\top \Sigma_t^{-1} \theta_{t+1}^*$. For this $\hat{w}_t$, we can bound its norm as

$$\|W_t A_t^\top \Sigma_t^{-1} \theta_{t+1}^*\|_2 \leq \|W_t\|_2 \|A_t^\top \Sigma_t^{-1}\|_2 \|\theta_{t+1}^*\|_2 \leq L^2 d\sqrt{\frac{t}{4\lambda}}. \tag{36}$$

Here $\|W_t\|_2 \leq L\sqrt{dt}$ and $\|\theta_{t+1}^*\|_2 \leq L\sqrt{d}$. Applying SVD decomposition and the property of matrix norm, $\|A_t^\top \Sigma_t^{-1}\|_2$ 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 $A_t A_t^\top$. Then noticing AM-GM inequality, we get $\|A_t^\top \Sigma_t^{-1}\|_2 \leq \sqrt{1/4\lambda}$.

Setting $B_t = L^2 d\sqrt{\frac{t}{4\lambda}}$, from Eq. (35), we have

$$\begin{aligned}
\varepsilon_1 &\leq \mathbb{E}\left[\left(\hat{\phi}_t^\top \hat{w}_t - \phi^{*\top}\theta_{t+1}^*\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\phi}_t^\top W_t A_t^\top \Sigma_t^{-1}\theta_{t+1}^* - \phi^{*\top}\Sigma_t\Sigma_t^{-1}\theta_{t+1}^*\right)^2\right] \\
&\leq \|\Sigma_t^{-1}\theta_{t+1}^*\|_2^2 \cdot \mathbb{E}\left[\|\hat{\phi}_t^\top W_t A_t^\top - \phi^{*\top}\Sigma_t\|_2^2\right] \\
&\leq 2\|\Sigma_t^{-1}\theta_{t+1}^*\|_2^2 \cdot \mathbb{E}\left[\|\hat{\phi}_t^\top W_t A_t^\top - \phi_t^\top A_t A_t^\top\|_2^2 + \lambda^2\|\phi^{*\top}\|_2^2\right] \\
&\leq 2\|\Sigma_t^{-1}\theta_{t+1}^*\|_2^2 \cdot \left(\sigma_1^2(A_t)\mathbb{E}\left[\|\hat{\phi}_t^\top W_t - \phi^{*\top}A_t\|_2^2\right] + \lambda^2\right) \leq 2\|\Sigma_t^{-1}\theta_{t+1}^*\|_2^2 \cdot \left(L^2 dt\varepsilon_0 + \lambda^2\right).
\end{aligned}$$

The second inequality uses Cauchy-Schwarz. The last inequality applies the upper bound $\sigma_1(A_t) \leq L\sqrt{dt}$ and the guarantee from Eq. (34). Using the fact that $t \leq T$, this implies that

$$\|\Sigma_t^{-1}\theta_{t+1}^*\|_2 \geq \sqrt{\frac{\varepsilon_1}{2(L^2 dT\varepsilon_0 + \lambda^2)}}.$$

We now use the generalized elliptic potential lemma from Carpentier et al. (2020) to upper bound the total value of $\|\Sigma_t^{-1}\theta_{t+1}^*\|_2$. From Lemma 35 in Appendix F.3, if $\lambda \geq L^2 d$ and we do not terminate in T rounds, then

$$T\sqrt{\frac{\varepsilon_1}{2(L^2 dT\varepsilon_0 + \lambda^2)}} \leq \sum_{t=1}^{T} \|\Sigma_t^{-1}\theta_{t+1}^*\|_2 \leq 2\sqrt{\frac{Td}{\lambda}}.$$

From this chain of inequalities, we can deduce $T\varepsilon_1 \leq 8(d/\lambda)\left(L^2 dT\varepsilon_0 + \lambda^2\right)$, therefore $T \leq \frac{8d\lambda}{\varepsilon_1 - 8L^2d^2\varepsilon_0/\lambda}$. Now, if we set $\varepsilon_1 = 16L^2d^2\varepsilon_0/\lambda$ in the above inequality, we can deduce

$$T \leq \frac{\lambda^2}{L^2 d\varepsilon_0}.$$

Putting everything together, for input parameter ε_{tol}, the termination threshold for the loss l is set such that $\frac{48L^2d^2\varepsilon_0}{\lambda} + \frac{4L^2d\varepsilon_0^2}{\lambda^2} \leq \varepsilon_{\text{tol}}$ which is satisfied for $\varepsilon_0 = \frac{\lambda\varepsilon_{\text{tol}}}{52L^2d^2}$. In addition, with $\lambda = L^2 d$, we set the constants for Algorithm 3 as follows:

$$T \leq \frac{52L^2d^2}{\varepsilon_{\text{tol}}}, \qquad \varepsilon_0 = \frac{\varepsilon_{\text{tol}}}{52d}, \qquad B_t \coloneqq \sqrt{\frac{L^2dt}{4}}, \qquad B \coloneqq \sqrt{\frac{13L^4d^3}{\varepsilon_{\text{tol}}}}.$$

Further, for Lemma 17, we set $\tilde{\varepsilon}$ to $\varepsilon_0/T = O\left(\frac{\varepsilon_{\text{tol}}^2}{L^2d^3}\right)$. Note that from Lemma 17, the loss upper bound is $3\varepsilon_1 + 4\tilde{\varepsilon}$ when the algorithm terminates. By our choice of the parameters, we can verify that $3\varepsilon_1 + 4\tilde{\varepsilon} \leq \varepsilon_{\text{tol}}$ and T does not exceed $\frac{52L^2d^2}{\varepsilon_{\text{tol}}}$, which completes the proof. ■

Sample Complexity of Moffle with Iterative Greedy Representation Learning With the feature selection guarantee in Lemma 14, we can now finish the proof for the final sample complexity result for the greedy iterative algorithm.

Proof of Theorem 6 Let us start with any fixed $h \in [H]$ and calculate the required number of samples per level.

Firstly, we consider learning $\hat{\phi}_h$ that satisfies Eq. (15). We use the discriminator class $\mathcal{V} = \mathcal{F}$ as defined in Eq. (1) and set $B = \sqrt{\frac{13d^3}{\varepsilon_{\text{reg}}}} = \tilde{O}\left(\sqrt{\frac{d^5K^9}{\eta_{\min}^3}}\right)$ (since $\varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^9\log^2(1+8/\beta)}\right)$ and β is a polynomial term). Applying Lemma 14, we know that for an approximation error of ε_{tol}, we need to set the sample size to $n = \tilde{O}\left(\frac{L^6d^7\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon_{\text{tol}}^3}\right)$.

Setting the values of the parameter $\varepsilon_{\text{tol}} = \varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^9\log^2(1+8/\beta)}\right)$ (according to Theorem 8) and $L = 1$ in Lemma 14, we get the number of episodes for learning $\hat{\phi}_h$ that satisfies Eq. (15) with probability at least $1 - \delta/(4H)$ is

$$n_{\hat{\phi}} = \tilde{O}\left(\frac{L^6d^7\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon_{\text{reg}}^3}\right) = \tilde{O}\left(\frac{d^{13}K^{27}\log(|\Phi|/\delta)}{\eta_{\min}^9}\right).$$

Substituting the value $B = \sqrt{\frac{13d^3}{\varepsilon_{\text{reg}}}} = \tilde{O}\left(\sqrt{\frac{d^5K^9}{\eta_{\min}^3}}\right)$ in Theorem 8, we know that we can get an exploratory dataset with probability at least $1 - \delta/(4H)$ and the corresponding sample complexity for the elliptic planner is

$$n_{\text{ell}} = \tilde{O}\left(\frac{H^5d^9K^{14}B^4\log(|\Phi|/\delta)}{\eta_{\min}^5}\right) = \tilde{O}\left(\frac{H^5d^{19}K^{32}\log(|\Phi|/\delta)}{\eta_{\min}^{11}}\right).$$

Next, we consider learning $\bar{\phi}_h$ that satisfies Eq. (18). Noticing that $\kappa = \frac{64dK^4\log(1+8/\beta)}{\eta_{\min}}$ and β is a polynomial term, we have that $\frac{\varepsilon^2}{16H^4\kappa K} = \tilde{O}\left(\frac{\varepsilon^2\eta_{\min}}{dH^4K^5}\right)$. Setting $\varepsilon_{\text{tol}} = \varepsilon_{\text{apx}} = \frac{\varepsilon^2}{16H^4\kappa K}$

and applying Lemma 14 with $L = H$, we know that if

$$n_{\bar{\phi}} = \tilde{O}\left(\frac{L^6 d^7 \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon_{\text{apx}}^3}\right) = \tilde{O}\left(\frac{H^{18} d^{10} K^{15} \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^6 \eta_{\min}^3}\right),$$

then condition Eq. (18) is satisfied with probability at least $1 - \delta/(4H)$.

Notice that Eq. (18) holds and we collect an exploratory dataset by applying Theorem 8. Then Theorem 9 implies the required sample complexity for offline FQI planning with $\bar{\phi}_{0:H-1}$ to learn an ε-optimal with probability at least $1 - \delta/(4H)$ is

$$n_{\text{plan}} = \tilde{O}\left(\frac{H^6 d^2 K^5 \log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2 \eta_{\min}}\right).$$

Union bounding over $h \in [H]$, the final sample complexity is $H(n_{\hat{\phi}} + n_{\text{ell}} + n_{\bar{\phi}} + n_{\text{plan}})$, and the result holds with probability $1 - \delta$. Reorganizing terms completes the proof. ■

8.6 Proofs for Enumerable Representation Class

We first derive the ridge regression based reduction of the min-max-min objective to eigenvector computation problems. Recall that for the enumerable feature class, we solve the following modified objective (Eq. (9)) in Algorithm 2

$$\operatorname*{argmin}_{\phi \in \Phi_h} \max_{f \in \mathcal{F}_{h+1}, \tilde{\phi} \in \Phi_h, \|\tilde{w}\|_2 \le B} \left\{ \min_{\|w\|_2 \le B} \mathcal{L}_{\mathcal{D}_h}(\phi, w, f) - \mathcal{L}_{\mathcal{D}_h}(\tilde{\phi}, \tilde{w}, f) \right\}$$

where $\mathcal{F}_{h+1}$ is now the discriminator class that contains all *unclipped* functions f in form of

$$f(x_{h+1}) = \mathbb{E}_{\text{unif}(\mathcal{A})}\left[\langle \phi_{h+1}(x_{h+1}, a), \theta \rangle\right], \text{ for } \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le \sqrt{d}.$$

Consider the min-max-min objective and fix $\phi, \tilde{\phi} \in \Phi_h$. Rewriting the objective for a sample of size n, we get the following updated objective:

$$\max_{f \in \mathcal{F}_{h+1}} \min_{\|w\|_2 \le \sqrt{d}} \|Xw - f(\mathcal{D}_h)\|_2^2 - \min_{\|\tilde{w}\|_2 \le \sqrt{d}} \|\tilde{X}\tilde{w} - f(\mathcal{D}_h)\|_2^2$$

where $X, \tilde{X} \in \mathbb{R}^{n \times d}$ are the covariate matrices for features ϕ and $\tilde{\phi}$ respectively.

We overload the notation and use $f(\mathcal{D}_h) \in \mathbb{R}^n$ to denote the value of any $f \in \mathcal{F}_{h+1}$ on the n samples. Now, instead of solving the constrained least squares problem, we use a ridge regression solution with regularization parameter λ. Thus, for any target f in the min-max objective, for feature ϕ, we get

$$w_f = \left(\tfrac{1}{n} X^\top X + \lambda I_{d \times d}\right)^{-1} \left(\tfrac{1}{n} X^\top f(\mathcal{D}_h)\right)$$

$$\|Xw - f(\mathcal{D}_h)\|_2^2 = \left\| X \left(\tfrac{1}{n} X^\top X + \lambda I_{d \times d}\right)^{-1} \left(\tfrac{1}{n} X^\top f(\mathcal{D}_h)\right) - f(\mathcal{D}_h) \right\|_2^2 = \|A(\phi) f(\mathcal{D}_h)\|_2^2$$

where $A(\phi) = I_{n \times n} - X \left(\tfrac{1}{n} X^\top X + \lambda I_{d \times d}\right)^{-1} \left(\tfrac{1}{n} X^\top\right)$.

Similarly, for the feature $\tilde{\phi}$, we have

$$\|\tilde{X}\tilde{w} - f(\mathcal{D}_h)\|_2^2 = \|A(\tilde{\phi})f(\mathcal{D}_h)\|_2^2,$$

where $A(\tilde{\phi}) = I_{n\times n} - \tilde{X}\left(\frac{1}{n}\tilde{X}^\top\tilde{X} + \lambda I_{d\times d}\right)^{-1}\left(\frac{1}{n}\tilde{X}^\top\right)$.

In addition, any regression target f can be rewritten as $f = X'\theta$ for a feature $\phi' \in \Phi_{h+1}$ and $\|\theta\|_2 \le \sqrt{d}$. Thus, for a fixed ϕ', ϕ and $\tilde{\phi}$, the maximization problem for $\mathcal{F}_{h+1}$ is the same as

$$\max_{\|\theta\|_2\le\sqrt{d}} \theta^\top X'^\top \left(A(\phi)^\top A(\phi) - A(\tilde{\phi})^\top A(\tilde{\phi})\right) X'\theta. \tag{37}$$

where $X' \in \mathbb{R}^{n\times d}$ is again the sample matrix defined using $\phi' \in \Phi_{h+1}$.

For each tuple of $(\phi, \tilde{\phi}, \phi')$, the maximization problem reduces to an eigenvector computation. As a result, we can efficiently solve the min-max-min objective in Eq. (9) by enumerating over each candidate feature in $(\phi, \tilde{\phi}, \phi')$ to solve

$$\operatorname*{argmin}_{\phi\in\Phi_h} \max_{\tilde{\phi}\in\Phi_h, \phi'\in\Phi_{h+1}, \|\theta\|_2\le\sqrt{d}} \theta^\top X'^\top \left(A(\phi)^\top A(\phi) - A(\tilde{\phi})^\top A(\tilde{\phi})\right) X'\theta. \tag{38}$$

Sample Complexity of MOFFLE for the Enumerable Feature Class We now prove the sample complexity result.

Proof of Theorem 7 Let us start with any fixed $h \in [H]$ and calculate the required number of samples per level.

Firstly, we consider learning $\hat{\phi}_h$ that satisfies Eq. (15). We use the discriminator class $\mathcal{V} = \mathcal{F}$ as defined in Eq. (10) and the error threshold ε_{reg}. Setting the values of the parameter $\varepsilon_{\text{reg}} = \tilde{\Theta}\left(\frac{\eta_{\min}^3}{d^2K^9\log^2(1+8/\beta)}\right)$ (according to Theorem 8) in Lemma 18 and noting β is a polynomial term, we get the number of episodes for learning $\hat{\phi}_h$ that satisfies Eq. (15) with probability at least $1-\delta/(3H)$ is

$$n_{\hat{\phi}} = \tilde{O}\left(\frac{d^6\log^3(|\Phi_h||\Phi_{h+1}|/\delta)}{\varepsilon_{\text{reg}}^3}\right) = \tilde{O}\left(\frac{d^{12}K^{27}\log^3(|\Phi|/\delta)}{\eta_{\min}^9}\right).$$

Now, substituting the value $B = 1/\lambda = \tilde{\Theta}\left(n_{\hat{\phi}}^{1/3}\right) = \tilde{\Theta}\left(\frac{d^4K^9\log(|\Phi|/\delta)}{\eta_{\min}^3}\right)$ in Theorem 8, we know that we can get an exploratory dataset with probability at least $1-\delta/(3H)$ and the corresponding sample complexity for the elliptic planner is

$$n_{\text{ell}} = \tilde{O}\left(\frac{H^5d^9K^{14}B^4\log(|\Phi|/\delta)}{\eta_{\min}^5}\right) = \tilde{O}\left(\frac{H^5d^{25}K^{50}\log^5(|\Phi|/\delta)}{\eta_{\min}^{17}}\right).$$

Finally, using Corollary 20 from Appendix D.2 and noticing $\kappa = \frac{64dK^4\log(1+8/\beta)}{\eta_{\min}}$, the number of episodes collected for running FQI with $\mathcal{Q}(R)$ to learn an ε-optimal policy with probability at least $1-\delta/(3H)$ can be bounded by

$$n_{\text{plan}} = \tilde{O}\left(\frac{H^6d^2\kappa K\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2}\right) = \tilde{O}\left(\frac{H^6d^3K^5\log(|\Phi||\mathcal{R}|/\delta)}{\varepsilon^2\eta_{\min}}\right).$$

Union bounding over $h \in [H]$, the final sample complexity is $H(n_{\hat{\phi}} + n_{\text{ell}} + n_{\text{plan}})$, and the result holds with probability at least $1-\delta$. Reorganizing terms completes the proof. ■

9. Conclusion

In this paper, we present MOFFLE, a new model-free algorithm, for representation learning and exploration in low-rank MDPs. We develop several representation learning schemes that vary in their computational and statistical properties, each yielding a different instantiation of the overall algorithm. Importantly MOFFLE can leverage a general function class Φ for representation learning, which provides it with the expressiveness and flexibility to scale to rich observation environments in a provably sample-efficient manner.

Acknowledgments

Part of this work was done while AM was at University of Michigan and was supported in part by a grant from the Open Philanthropy Project to the Center for Human-Compatible AI, and in part by NSF grant CAREER IIS-1452099. JC would like to thank Kefan Dong for helpful discussions related to Bernstein's version of uniform deviation bounds. NJ acknowledges funding support from ARL Cooperative Agreement W911NF-17-2-0196, NSF IIS-2112471, NSF CAREER IIS-2141781, and Adobe Data Science Research Award.

Contents

Appendix A. Additional Discussions Among the Closely Related Works

In this section, we provide more details about comparisons with Olive, Witness rank, and BLin-Ucb. They are statistically efficient for more general settings beyond low-rank MDPs. Strictly speaking, their realizability assumptions and sample complexity terms are different from what we present in Table 1. Olive requires the realizability of the value function class $Q^* \in \mathcal{F}^{\text{CLASS}}$ and has $\log(|\mathcal{F}^{\text{CLASS}}|)$ dependence. Witness rank requires the realizability of the model class $\mathcal{M}^* \in \mathcal{M}^{\text{CLASS}}$ and an induced value function $\mathcal{F}^{\text{CLASS}}$ class from $\mathcal{M}^{\text{CLASS}}$, and consequently pays $\log(|\mathcal{F}^{\text{CLASS}}||\mathcal{M}^{\text{CLASS}}|)$. BLin-Ucb makes a more complicated realizability assumption on a hypothesis function class $\mathcal{H}^{\text{CLASS}}$, whose complexity $\log(|\mathcal{H}^{\text{CLASS}}|)$ shows up on the bound (please refer to Du et al. (2021) for more details). Here we add the superscript class to function classes to differentiate them from the notations in other parts of the paper.

For the purpose of comparison, we instantiate their sample complexity bounds in our setting. We design function classes $\mathcal{H}^{\text{CLASS}} = \mathcal{F}^{\text{CLASS}} = \mathcal{F}_0^{\text{CLASS}} \times \ldots \times \mathcal{F}_{H-1}^{\text{CLASS}}$, where $\mathcal{F}_h^{\text{CLASS}} = \left\{ f_h(x_h, a_h) = R_h(x_h, a_h) + \langle \phi_h(x_h, a_h), \theta_h \rangle : \phi_h \in \Phi_h, \|\theta_h\|_2 \leq \sqrt{d} \right\}$. For Witness rank, we additionally construct $\mathcal{M}^{\text{CLASS}} = \{\langle \phi, \mu \rangle : \phi \in \Phi, \mu \in \Upsilon\}$. The sample complexities are then obtained by calculating the complexity of these function classes and multiplying an H^2 factor to translate the results from the bounded total reward setting ($0 \leq \sum_{h=0}^{H-1} r_h \leq 1$) in Jiang and Agarwal (2018) to our uniformly bounded reward setting ($r_h \in [0, 1], \forall h \in [H]$).

Appendix B. The Analysis of Elliptical Planner

In this section, we show the iteration and sample complexities and the estimation guarantee for *offline* "elliptical planner" (Algorithm 4). The algorithm and analysis follows a similar approach as Algorithm 2 in Agarwal et al. (2020b), while the major difference here is that we call FQI for the policy optimization step because we do not have the model. In addition, we cannot directly estimate the covariance matrix by sampling data from the estimated model as in Agarwal et al. (2020b).

Inspired by Huang et al. (2021), we perform Fitted Q-Evaluation (FQE) with the exploratory data in the prior levels to substitute the Monte Carlo estimation counterpart used in the online "elliptical planner". Different from Huang et al. (2021), we no longer run our algorithm on the discretized value functions and rewards. In contrast, we directly use the original elliptical reward and perform FQI and FQE on the original continuous function class. This makes the algorithm more computationally handy.

The detailed algorithm is shown in Algorithm 4. The algorithm proceeds in iterations. In each round, we first use the current covariance matrix Γ_{t-1} to set the elliptical reward line 4. Then in line 5 we call FQI (Algorithm 5) to get policy π_t that explores the uncovered direction set by the elliptical reward. Next, we call FQE (Algorithm 6) to estimate the covariance matrix $\widehat{\Sigma}_{\pi_t}$ (more specifically, each (i, j)-th coordinate, $i, j \in \{1, \ldots, d\}$ in the covariance matrix respectively) for all policy π_t (line 6-12). The covariance matrix Γ_t is updated in line 13. FQE is also used to the expected return (line 14) to check the stopping condition.

On the technical side, because the number of arbitrary policies (the policy set of π_t) and the value function class are exponentially large, we need to invoke covering argument on the infinite function class. We apply the more involved concentration analysis (uniform

Algorithm 4 Elliptical Planner with FQI and FQE

input: Features $\hat{\phi}$, exploratory dataset $\mathcal{D} \coloneqq \mathcal{D}_{0:\tilde{H}}$ with size n at each level $h \in [\tilde{H}]$, and threshold $\beta > 0$.
Initialize $\Gamma_0 = I_{d\times d}$.
for $t = 1, 2, \ldots,$ **do**
  Define the elliptical reward $R^{\mathrm{FQI},t}$ as $R^{\mathrm{FQI},t}_{\tilde{H}} = \left\|\hat{\phi}_{\tilde{H}}\right\|^2_{\Gamma_{t-1}^{-1}}$ and $R^{\mathrm{FQI},t}_h = \mathbf{0}, \forall h \in [\tilde{H}]$.
  Using Algorithm 5, compute

$$\pi_t = \text{FQI-ELLIPTICAL}\left(\mathcal{D}, R^{\mathrm{FQI},t}\right).$$

  Estimate feature covariance matrix $\widehat{\Sigma}_{\pi_t}$ as
  for $i = 1, \ldots, d$ **do**
    for $j = 1, \ldots, d$ **do**
      Define reward function $R^{\mathrm{FQE},ij}$ as $R^{\mathrm{FQE},ij}_{\tilde{H}}(\cdot,\cdot) = \frac{1+\hat{\phi}_{\tilde{H}}(\cdot,\cdot)[i]\hat{\phi}_{\tilde{H}}(\cdot,\cdot)[j]}{2}$ and $R^{\mathrm{FQE},ij}_h = \mathbf{0}, \forall h \in [\tilde{H}]$.
      Estimate the (i,j)-th coordinate of $\widehat{\Sigma}_{\pi_t}$ using FQE (Algorithm 6)

$$\widehat{\Sigma}_{\pi_t}[i,j] := 2\left[\mathrm{FQE}\left(\mathcal{D}, R^{\mathrm{FQE},ij}, \pi_t\right)\right] - 1.$$

    end for
  end for
  Update $\Gamma_t \leftarrow \Gamma_{t-1} + \widehat{\Sigma}_{\pi_t}$.
  Estimate the expected return of π_t under the elliptical reward R^t as

$$\hat{v}_t^{\pi_t} := \mathrm{FQE}\left(\mathcal{D}, R^{\mathrm{FQI},t}, \pi_t\right).$$

  If the estimated objective $\hat{v}_t^{\pi_t} \le \frac{3\beta}{4}$, halt and output $\rho \coloneqq \mathrm{unif}(\{\pi_\tau\}_{1\le\tau\le t})$.
end for

Bernstein's inequality) for the infinite function class to achieve sharp rates. By adapting the tools and analysis from Dong et al. (2020), we show a key concentration result in Corollary 43, which is then applied in the squared loss deviation result in Appendix F.2. As discussed in Section 8.1, FQE procedure is not the only solution for the "elliptical planner". Instead of running FQE in the offline "elliptical planner", in each iteration we can also collect new data according to policy π_t and use Monte Carlo evaluation to estimate the covariance matrix, which yields the online "elliptical planner". However, it leads to a worse rate due to additional collection and inefficient usage of prior data. Another advantage of using FQE is that it matches the optimal $\tilde{\Omega}(H)$ deployment complexity as discussed in Huang et al. (2021).

Now we state and prove the theoretical guarantee in Lemma 15.

Lemma 15 (Estimation and iteration guarantees for Algorithm 4) *If Algorithm 4 is run with a dataset of size* $n \ge \tilde{O}(\frac{H^4 d^6 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2})$ *for a fix* $\beta > 0, \delta \in (0,1)$*, then upon*

termination, it outputs a matrix Γ_T *and a policy* ρ *that with probability at least* $1-\delta$

$$\forall \pi : \mathbb{E}_\pi \left[\hat{\phi}_{\tilde{H}-1}(x_{\tilde{H}-1}, a_{\tilde{H}-1})^\top (\Gamma_T)^{-1} \hat{\phi}_{\tilde{H}-1}(x_{\tilde{H}-1}, a_{\tilde{H}-1}) \right] \leq O(\beta), \tag{39}$$

$$\left\| \frac{\Gamma_T}{T} - \left(\mathbb{E}_\rho \left[\hat{\phi}_{\tilde{H}-1}(x_{\tilde{H}-1}, a_{\tilde{H}-1}) \hat{\phi}_{\tilde{H}-1}(x_{\tilde{H}-1}, a_{\tilde{H}-1})^\top \right] + \frac{I_{d\times d}}{T} \right) \right\|_{\text{op}} \leq O(\beta/d). \tag{40}$$

Further, the iteration complexity is also bounded $T \leq \frac{8d}{\beta} \log\left(1 + \frac{8}{\beta}\right)$.

Proof As notation, we use v_t^π to denote the expected return of any policy π under the elliptical reward $R^{\text{FQI},t}$ ($R^{\text{FQI},t}$ is defined as $R^{\text{FQI},t}_{\tilde{H}} = \|\hat{\phi}_{\tilde{H}}\|^2_{\Gamma^{-1}_{t-1}}$ and $R^{\text{FQI},t}_h = \mathbf{0}, \forall h \in [\tilde{H}]$).

We start with showing

$$\max_{t\in[T]} \max \left\{ d \cdot \left\| \widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t} \right\|_{\text{op}}, \left| \hat{v}_t^{\pi_t} - v_t^{\pi_t} \right|, \max_\pi v_t^\pi - v_t^{\pi_t} \right\} \leq \beta/8. \tag{41}$$

For the second term in Eq. (41), from Corollary 30 we know that if $n \geq \tilde{O}(\frac{H^4 d^5 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2})$, then with probability at least $1-\delta/3$, we have for $|\hat{v}_t^{\pi_t} - v_t^{\pi_t}| \leq \beta/8$ any $t \in [T]$.

For the third term in Eq. (41), Lemma 24 tells us that if $n \geq \tilde{O}(\frac{H^4 d^3 \kappa K \log(|\Phi|/\delta)}{\beta^2})$, then with probability at least $1-\delta/3$, we have $\max_\pi v_t^\pi - v_t^{\pi_t}$ for any $t \in [T]$.

Then we consider the more complicated first term in Eq. (41). We will show that for any policy π_t, such quantity can be upper bound by some FQE errors.

For any fixed policy π_t, noticing the spectral norm result from Lemma 5.4 of Vershynin (2010), we get

$$\left\| \widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t} \right\|_{\text{op}} \leq (1-2\gamma')^{-1} \sup_{v \in \mathcal{N}_{\gamma'}} |\langle (\widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t}) v, v \rangle|,$$

where we use $\mathcal{N}_{\gamma'}$ to denote the ℓ_2-cover of the unit ball $\{v \in \mathbb{R}^d : \|v\|_2 \leq 1\}$ at scale γ' and its size is $(\frac{2}{\gamma'})^d$.

For given π_t and v, we consider the reward function R^v with $R^v_{\tilde{H}}(x,a) = (v^\top \hat{\phi}_{\tilde{H}}(x,a))^2 = v^\top \hat{\phi}_{\tilde{H}}(x,a) \hat{\phi}_{\tilde{H}}(x,a)^\top v$ ($= \sum_{i=1}^d \sum_{j=1}^d v[i] \hat{\phi}_{\tilde{H}}(x,a)[i] \hat{\phi}_{\tilde{H}}(x,a)[j] v[j]$) and $R^v_h = \mathbf{0}, \forall h \in [\tilde{H}]$. In addition, we use $v^{\pi_t}_{R^v}$ to denote the expected return of policy π_t under reward function R^v and use $\hat{v}^{\pi_t}_{R^v}$ to denote $\text{FQE}(\mathcal{D}, R^v, \pi_t)$.

One thing to notice is that here we do not really run FQE with reward R^v for all $v \in \mathcal{N}_{\gamma'}$ to get the estimate $\hat{v}^{\pi_t}_{R^v}$ in Algorithm 4. Instead, it only appears in the analysis. Once we calculate $\widehat{\Sigma}_{\pi_t}$ by estimating the covariance matrix (line 6-12 in Algorithm 4), we can immediately get $v^\top \widehat{\Sigma}_{\pi_t} v$, which it is equivalent to the real output of FQE (i.e., $\hat{v}^{\pi_t}_{R^v}$). More specifically, from the equivalence between FQE and a model-based plug-in formulation (Duan et al., 2020, Theorem 1) and noticing the definition of $R^v_{\tilde{H}}, R^{\text{FQE},ij}$, we have the following crucial equation

$$\hat{v}^{\pi_t}_{R^v} = \sum_{i=1}^d \sum_{j=1}^d v[i] (2\hat{v}^{\pi_t}_{R^{\text{FQE},ij}} - 1) v[j] = \sum_{i=1}^d \sum_{j=1}^d v[i] \widehat{\Sigma}_\pi[i,j] v[j] = v^\top \widehat{\Sigma}_{\pi_t} v.$$

This further implies that $v_{R^v}^{\pi_t} = v^\top \Sigma_{\pi_t} v$ and $|\langle(\widehat{\Sigma}_\pi - \Sigma_\pi)v, v\rangle| = |\hat{v}_{R^v}^{\pi_t} - v_{R^v}^{\pi_t}|$ is indeed the FQE error. Therefore, bounding the operator norm $\|\widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t}\|_{\text{op}}$ can be reduced to bounding the FQE error for a class of rewards.

Applying Lemma 27 with $\mathcal{R} = \{R^v : v \in \mathcal{N}_{\gamma'}\}$ and $\gamma' = \frac{\beta}{32d}$, we get that if $n \geq \tilde{O}(\frac{H^4 d^6 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2})$, then (assuming $\frac{\beta}{32d} \leq \frac{1}{4}$) with probability at least $1 - \delta/3$, we have for any $t \in [T]$

$$\left\|\widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t}\right\|_{\text{op}} \leq (1-2\gamma')^{-1} \sup_{v \in \mathcal{N}_{\gamma'}} |\langle(\widehat{\Sigma}_{\pi_t} - \Sigma_{\pi_t})v, v\rangle| = (1-2\gamma')^{-1} \sup_{v \in \mathcal{N}_{\gamma'}} |\hat{v}_{R^v}^{\pi_t} - v_{R^v}^{\pi_t}| \leq \frac{\beta}{8d}.$$

In summary, the required number of samples for Eq. (41) is

$$n \geq \tilde{O}\left(\frac{H^4 d^6 \kappa K^2 \log(|\Phi|/\delta)}{\beta^2}\right).$$

Now, if we terminate in iteration T, we know that $\hat{v}_T^{\pi_T} \leq 3\beta/4$. This implies

$$\max_\pi v_T^\pi \leq v_T^{\pi_T} + \beta/8 \leq \hat{v}_T^{\pi_T} + \beta/4 \leq \beta. \tag{42}$$

Therefore Eq. (39) holds. From Eq. (41), it is also easy to see that Eq. (40) holds.

Then, we turn to the iteration complexity. Similarly to Eq. (42), we have

$$\begin{aligned} T\left(3\beta/4 - \beta/4\right) &\leq \sum_{t=1}^T (\hat{v}_t^{\pi_t} - \beta/4) \leq \sum_{t=1}^T (v_t^{\pi_t} - \beta/8) = \sum_{t=1}^T \left(\mathbb{E}_{\pi_t}\left[\hat{\phi}_{\tilde{H}}^\top \Gamma_{t-1}^{-1} \hat{\phi}_{\tilde{H}}\right] - \beta/8\right) \\ = \sum_{t=1}^T \left(\text{tr}(\Sigma_{\pi_t}\Gamma_{t-1}^{-1}) - \beta/8\right) &\leq \sum_{t=1}^T \text{tr}(\widehat{\Sigma}_{\pi_t}\Gamma_{t-1}^{-1}) \leq 2d \log\left(1 + \frac{T}{d}\right). \end{aligned}$$

In the last step, we apply elliptical potential lemma (e.g., Lemma 26 of Agarwal et al. (2020b)).

Reorganizing the equation yields $T \leq \frac{4d}{\beta}\log\left(1 + \frac{T}{d}\right)$. Further, if $T \leq \frac{8d}{\beta}\log\left(1 + \frac{8}{\beta}\right)$, then we have

$$\begin{aligned} T &\leq \frac{4d}{\beta}\log\left(1 + T/d\right) \leq \frac{4d}{\beta}\log\left(1 + \frac{8\log\left(1 + 8/\beta\right)}{\beta}\right) \\ &\leq \frac{4d}{\beta}\log\left(1 + (8/\beta)^2\right) \leq \frac{8d}{\beta}\log\left(1 + 8/\beta\right). \end{aligned}$$

Therefore, we obtain an upper bound on T by this set and guess approach. ■

Appendix C. Supporting Sample Complexity Results for Section 8

In this section, we state and prove the deviation bounds used for the proofs in Section 8.

C.1 Deviation Bound for FLO

We start by stating the sample complexity result for the min-max-min oracle FLO, defined in Eq. (33) in Section 8.4.

Lemma 16 *If the feature learning objective Eq. (5) is solved by the* FLO *for a sample of size n, then for $\mathcal{V} \subseteq (\mathcal{X} \to [0, L])$, $\mathcal{V} := \{v(x_{h+1}) = \text{clip}_{[0,L]}\ (\mathbb{E}_{a_{h+1}\sim\pi_{h+1}(x_{h+1})}\ [R(x_{h+1}, a_{h+1}) + \langle\phi_{h+1}(x_{h+1}, a_{h+1}), \theta\rangle]) : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \le L\sqrt{d}, R \in \mathcal{R}\}$, where π is the greedy policy or the uniform policy, with probability at least $1-\delta$, we have*

$$\max_{v\in\mathcal{V}} \text{b\_err}\left(\rho_{h-3}^{+3}, \hat{\phi}_h, v; L\sqrt{d}\right) \le \frac{16c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n},$$

where c_3 is the constant in Lemma 34.

Proof Firstly, note the term $\text{b\_err}\left(\rho_{h-3}^{+3}, \hat{\phi}_h, v; L\sqrt{d}\right)$ is a shorthand for

$$\min_{\|w\|_2\le L\sqrt{d}} \mathcal{L}_{\rho_{h-3}^{+3}}(\hat{\phi}_h, w, v) - \mathcal{L}_{\rho_{h-3}^{+3}}(\phi_h^*, \theta_v^*, v)$$

where $\theta_v^* = \text{argmin}_{\|\theta\|_2\le L\sqrt{d}} \mathcal{L}_{\rho_{h-3}^{+3}}(\phi_h^*, \theta, v)$.

Now, using the result in Lemma 34 from Appendix F.5 and denoting $\mathcal{L}_{\rho_{h-3}^{+3}}(\cdot)$ as $\mathcal{L}(\cdot)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&|\mathcal{L}(\phi, w, v) - \mathcal{L}(\phi^*, \theta_v^*, v) - (\mathcal{L}_\mathcal{D}(\phi, w, v) - \mathcal{L}_\mathcal{D}(\phi^*, \theta_v^*, v))| \\
\le\ &\frac{1}{2}\left(\mathcal{L}(\phi, w, v) - \mathcal{L}(\phi^*, \theta_v^*, v)\right) + \frac{4c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n}
\end{aligned}$$

for all $\|w\|_2 \le L\sqrt{d}$ ($B = L\sqrt{d}$), $\phi \in \Phi_h$ and $v \in \mathcal{V}$.

By the definition of FLO, for any $v \in \mathcal{V}$, we know that there exists θ_v^* such that (ϕ^*, θ_v^*) is the population minimizer $\text{argmin}_{\tilde{\phi}\in\Phi_h, \|\tilde{w}\|_2\le L\sqrt{d}} \mathcal{L}(\tilde{\phi}, \tilde{w}, v)$. Using this and the concentration result, for all $v \in \mathcal{V}$, with $(\tilde{\phi}_v, \tilde{w}_v)$ as the solution of the innermost min in Eq. (33), we have

$$\begin{aligned}
&\mathcal{L}_\mathcal{D}(\phi^*, \theta_v^*, v) - \mathcal{L}_\mathcal{D}(\tilde{\phi}_v, \tilde{w}_v, v) \\
\le\ &\frac{3}{2}\left(\mathcal{L}(\phi^*, \theta_v^*, v) - \mathcal{L}(\tilde{\phi}_v, \tilde{w}_v, v)\right) + \frac{4c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n} \\
\le\ &\frac{4c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n}.
\end{aligned}$$

Now, for the oracle solution $\hat{\phi}$, for all $v \in \mathcal{V}$, we have

$$\begin{aligned}
&\mathcal{L}_\mathcal{D}(\hat{\phi}, \hat{w}_v, v) - \mathcal{L}_\mathcal{D}(\tilde{\phi}_v, \tilde{w}_v, v) \\
=\ &\mathcal{L}_\mathcal{D}(\hat{\phi}, \hat{w}_v, v) - \mathcal{L}_\mathcal{D}(\phi^*, \theta_v^*, v) + \mathcal{L}_\mathcal{D}(\phi^*, \theta_v^*, v) - \mathcal{L}_\mathcal{D}(\tilde{\phi}_v, \tilde{w}_v, v) \\
\ge\ &\frac{1}{2}\left(\mathcal{L}(\hat{\phi}, \hat{w}_v, v) - \mathcal{L}(\phi^*, \theta_v^*, v)\right) - \frac{4c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n}.
\end{aligned}$$

Combining the two chains of inequalities, we get

$$\mathcal{L}(\hat{\phi}, \hat{w}_v, v) - \mathcal{L}(\phi^*, \theta_v^*, v)$$

$$
\begin{aligned}
&\leq 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}, \hat{w}_v, v) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_v, \tilde{w}_v, v)\right) + \frac{8c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n} \\
&\leq 2\max_{g\in\mathcal{V}}\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}, \hat{w}_g, g) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_g, \tilde{w}_g, g)\right) + \frac{8c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n} \\
&\leq 2\max_{g\in\mathcal{V}}\left(\mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_g, g) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_g, \tilde{w}_g, g)\right) + \frac{8c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n} \\
&\leq \frac{16c_3L^2d^2\log(2nL\sqrt{d}|\Phi_h||\Phi_{h+1}||\mathcal{R}|/\delta)}{n}.
\end{aligned}
$$

Hence, we have proved the desired result. ■

Here, we explicitly give a result for the discriminator function classes used by Moffle. However, a similar result can be easily derived for a general discriminator class $\mathcal{V}$ with the dependence $\log(N)$ where N is either the cardinality or an appropriate complexity measure of $\mathcal{V}$.

C.2 Deviation Bounds for Greedy Selection

Below, we state the deviation bounds used in the proof of Lemma 14 in Section 8.5.

Lemma 17 *Let*

$$\tilde{\varepsilon} = \frac{2c_3d(B+L\sqrt{d})^2\log(n(B+L\sqrt{d})|\Phi||\Phi'||\mathcal{R}|/\delta)}{n},$$

where c_3 is the constant in Lemma 34. If Algorithm 3 is called with a dataset $\mathcal{D}$ of size n and termination loss cutoff $3\varepsilon_1/2 + \tilde{\varepsilon}$, then with probability at least $1-\delta$, for all $v \in \mathcal{V} \subseteq (\mathcal{X} \to [0, L])$, $\mathcal{V} \coloneqq \{v(x_{h+1}) = \mathrm{clip}_{[0,L]}(\mathbb{E}_{a_{h+1}\sim\pi_{h+1}(x_{h+1})}[R(x_{h+1}, a_{h+1}) + \langle\phi_{h+1}(x_{h+1}, a_{h+1}), \theta\rangle]) : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \leq L\sqrt{d}, R \in \mathcal{R}\}$, $\|w\|_2 \leq B_t$, and $t \leq T$, we have

$$\sum_{i\leq t}\mathbb{E}_{\rho^{+3}_{h-3}}\left[\left(\hat{\phi}_{t,h}(x_h, a_h)^\top w_{t,i} - \phi^*_h(x_h, a_h)^\top\theta^*_i\right)^2\right] \leq t\tilde{\varepsilon}$$

$$\mathbb{E}_{\rho^{+3}_{h-3}}\left[\left(\hat{\phi}_{t,h}(x_h, a_h)^\top w - \phi^*_h(x_h, a_h)^\top\theta^*_{t+1}\right)^2\right] \geq \varepsilon_1$$

where $w_{t,i} = \mathrm{argmin}_{\|\tilde{w}\|_2\leq B_t}\mathcal{L}_{\mathcal{D}}(\hat{\phi}_{t,h}, \tilde{w}, v_i)$, $\theta^*_i = \mathrm{argmin}_{\|\tilde{w}\|_2\leq L\sqrt{d}}\mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \tilde{w}, v_i)$, $B \geq B_t = \frac{L\sqrt{dt}}{2}$ in Algorithm 3, and policy π is the greedy policy or the uniform policy.

Further, at termination, the learned feature $\hat{\phi}_{T,h}$ satisfies

$$\max_{v\in\mathcal{V}}\mathrm{b\_err}\left(\rho^{+3}_{h-3}, \hat{\phi}_{T,h}, v; B\right) \leq 3\varepsilon_1 + 4\tilde{\varepsilon}.$$

Proof We again denote $\mathcal{L}_{\rho^{+3}_{h-3}}(\cdot)$ as $\mathcal{L}(\cdot)$ and set $\tilde{\varepsilon} = \frac{2c_3d(B+L\sqrt{d})^2\log(n(B+L\sqrt{d})|\Phi||\Phi'||\mathcal{R}|/\delta)}{n}$. Further, we remove the subscript h for simplicity unless not clear by context. We begin by using the result in Lemma 34 such that, with probability at least $1-\delta$, for all $\|w\|_2 \leq B$ ($B \geq L\sqrt{d}$), $\phi \in \Phi_h$ and $v \in \mathcal{V}$, we have

$$|\mathcal{L}(\phi, w, v) - \mathcal{L}(\phi^*, \theta^*_v, v) - (\mathcal{L}_{\mathcal{D}}(\phi, w, v) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_v, v))|$$

$$\leq \frac{1}{2}\left(\mathcal{L}(\phi, w, v) - \mathcal{L}(\phi^*, \theta_v^*, v)\right) + \tilde{\varepsilon}/2.$$

Thus, for the feature fitting step in line 6 of Algorithm 3 in iteration t, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\sum_{v_i \in \mathcal{V}^t} \mathbb{E}\left[\left(\hat{\phi}_t^\top w_{t,i} - \phi^{*\top}\theta_i^*\right)^2\right] = \sum_{v_i \in \mathcal{V}^t}\left(\mathcal{L}(\hat{\phi}_t, w_{t,i}, v_i) - \mathcal{L}(\phi^*, \theta_i^*, v_i)\right)\\
\leq &\sum_{v_i \in \mathcal{V}^t} 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_t, w_{t,i}, v_i) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_i^*, v_i)\right) + |\mathcal{V}^t|\tilde{\varepsilon} \leq \ t\tilde{\varepsilon},
\end{aligned}$$

which means the first inequality in the lemma statement holds.

For the adversarial test function at iteration t with $B_t \leq B$, let $\bar{w} \coloneqq \operatorname{argmin}_{\|w\|_2 \leq B_t} \mathcal{L}_{\mathcal{D}}(\hat{\phi}_t, w, v_{t+1})$. Using the same sample size for the adversarial test function at each non-terminal iteration with loss cutoff c_{cutoff}, for any vector $w \in \mathbb{R}^d$ with $\|w\|_2 \leq B_t$ we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\phi}_t^\top w - \phi^{*\top}\theta_{t+1}^*\right)^2\right] = \mathcal{L}(\hat{\phi}_t, w, v_{t+1}) - \mathcal{L}(\phi^*, \theta_{t+1}^*, v_{t+1})\\
\geq &\frac{2}{3}\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_t, w, v_{t+1}) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_{t+1}^*, v_{t+1})\right) - \frac{\tilde{\varepsilon}}{3}\\
\geq &\frac{2}{3}\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_t, \bar{w}, v_{t+1}) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_{t+1}^*, v_{t+1})\right) - \frac{\tilde{\varepsilon}}{3}\\
\geq &\frac{2c_{\text{cutoff}}}{3} + \frac{2}{3}\left(\min_{\tilde{\phi}\in\Phi_h, \|\tilde{w}\|_2 \leq L\sqrt{d}} \mathcal{L}_{\mathcal{D}}(\tilde{\phi}, \tilde{w}, v_{t+1}) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_{t+1}^*, v_{t+1})\right) - \frac{\tilde{\varepsilon}}{3}\\
\geq &\frac{2c_{\text{cutoff}}}{3} + \frac{1}{3}\left(\mathcal{L}(\tilde{\phi}_{t+1}, \tilde{w}_{t+1}, v_{t+1}) - \mathcal{L}(\phi^*, \theta_{t+1}^*, v_{t+1})\right) - \frac{2\tilde{\varepsilon}}{3} \geq \frac{2c_{\text{cutoff}}}{3} - \frac{2\tilde{\varepsilon}}{3}.
\end{aligned}$$

In the first inequality, we invoke Lemma 34 to move to empirical losses. In the third inequality, we add and subtract the ERM loss over (ϕ, w) pairs along with the fact that the termination condition is not satisfied for v_{t+1}. In the next step, we again use Lemma 34 for the ERM pair $(\tilde{\phi}_{t+1}, \tilde{w}_{t+1})$ for v_{t+1}.

Thus, if we set the cutoff c_{cutoff} for test loss to $3\varepsilon_1/2 + \tilde{\varepsilon}$, for a non-terminal iteration t, for any $w \in \mathbb{R}^d$ with $\|w\|_2 \leq B_t$, we have

$$\mathbb{E}\left[\left(\hat{\phi}_t^\top w - \phi^{*\top}\theta_{t+1}^*\right)^2\right] \geq \varepsilon_1, \tag{43}$$

which implies the second inequality in the lemma statement holds.

At the same time, for the last iteration, for all $v \in \mathcal{V}$, the feature $\hat{\phi}_T$ satisfies

$$\begin{aligned}
&\min_{\|w\|_2 \leq B} \mathbb{E}\left[\left(\hat{\phi}_T^\top w - \phi^{*\top}\theta_v^*\right)^2\right] \leq 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_T, \hat{w}_v, v) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_v^*, v)\right) + \tilde{\varepsilon}\\
\leq &\ 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_T, \hat{w}_v, v) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_v, \tilde{w}_v, v) + \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_v, \tilde{w}_v, v) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta_v^*, v)\right) + \tilde{\varepsilon}\\
\leq &\ 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}_T, \hat{w}_v, v) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_v, \tilde{w}_v, v)\right) + \tilde{\varepsilon} \leq 3\varepsilon_1 + 4\tilde{\varepsilon}.
\end{aligned}$$

This gives us the third inequality in the lemma, thus completing the proof. ■

C.3 Deviation Bound for Enumerable Feature Setting

We now show that using the ridge estimator for an enumerable feature class, discriminator class $\mathcal{F}_{h+1}$ and an appropriately set value of λ still allows us to establish a feature approximation result similar to FLO (Eq. (5)) and iterative greedy feature selection (Algorithm 3).

Lemma 18 *For the feature $\hat{\phi}_h$ learned via the ridge estimator Eq. (38), with $B = \frac{1}{\lambda}$ and $\lambda = \tilde{\Theta}\left(\frac{1}{n^{1/3}}\right)$ for any function $f \in \mathcal{F}_{h+1}$, with probability at least $1-\delta$, we have*

$$\max_{f \in \mathcal{F}_{h+1}} \text{b\_err}\left(\rho^{+3}_{h-3}, \hat{\phi}_h, f; B\right) \leq \tilde{O}\left(\frac{d^2 \log(2n|\Phi_h||\Phi_{h+1}|/\delta)}{n^{1/3}}\right)$$

where the discriminator function class is $\mathcal{F}_{h+1} := \{f(x_{h+1}) = \mathbb{E}_{\text{unif}(\mathcal{A})}\left[\langle \phi_{h+1}(x_{h+1}, a), \theta\rangle\right] : \phi_{h+1} \in \Phi_{h+1}, \|\theta\|_2 \leq \sqrt{d}\}$.

Proof Firstly, from the definition of $\mathcal{F}_{h+1}$, we note that note the term $\text{b\_err}\left(\rho^{+3}_{h-3}, \hat{\phi}_h, f; 1/\lambda\right)$ can be written as

$$\min_{\|\hat{w}_f\|_2 \leq 1/\lambda} \mathcal{L}_{\rho^{+3}_{h-3}}(\hat{\phi}_h, \hat{w}_f, f) - \mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_f, f)$$

where $\theta^*_f = \text{argmin}_{\|\theta\|_2 \leq \sqrt{d}} \mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta, f)$.

We again denote $\mathcal{L}_{\rho^{+3}_{h-3}}(\cdot)$ as $\mathcal{L}(\cdot)$, $\mathcal{F}_{h+1}$ as $\mathcal{F}$. Firstly, as the discriminator function class $\mathcal{F}$ is defined without clipping, we now have: $\mathbb{E}[f(x')|x, a] = \langle \phi^*(x, a), \theta^*_f\rangle$ with $\|\theta^*_f\|_2 \leq d$. Also, the scale of the ridge estimator $w_f = \left(\frac{1}{n}X^\top X + \lambda I_{d\times d}\right)^{-1}\left(\frac{1}{n}X^\top f\right)$ now scales as $\frac{1}{\lambda}$. Now, applying Lemma 34, for all $\phi \in \Phi_h$, $\|w\|_2 \leq 1/\lambda$ and $f \in \mathcal{F}$, we have

$$\begin{aligned}
&\left|\mathcal{L}(\phi, w, f) - \mathcal{L}(\phi^*, \theta^*_f, f) - \left(\mathcal{L}_{\mathcal{D}}(\phi, w, f) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f)\right)\right| \\
\leq\ & \frac{1}{2}\left(\mathcal{L}(\phi, w, f) - \mathcal{L}(\phi^*, \theta^*_f, f)\right) + \frac{c_3 d(1/\lambda + \sqrt{d})^2 \log(n(1/\lambda + \sqrt{d})|\Phi_h||\Phi_{h+1}|/\delta)}{n},
\end{aligned}$$

where c_3 is the constant in Lemma 34. Now, let w^*_f denote the population ridge regression estimator for target $f \in \mathcal{F}$ for features ϕ^*. Assume $\lambda \leq 1/d$ Then an upper bound on the second term in the RHS is $\gamma := \frac{4c_3 d \log(2n(1/\lambda)|\Phi_h||\Phi_{h+1}|/\delta)}{\lambda^2 n}$. For the selected feature $\hat{\phi}$, we have

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f) \\
=\ & \mathcal{L}_{\mathcal{D}}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f) + \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f) \\
\geq\ & \frac{1}{2}\left(\mathcal{L}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}(\phi^*, \theta^*_f, f)\right) + \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f) - \gamma \\
\geq\ & \frac{1}{2}\left(\mathcal{L}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}(\phi^*, \theta^*_f, f)\right) + \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f) - \mathcal{L}_{\mathcal{D}}(\phi^*, w^*_f, f) - \gamma.
\end{aligned}$$

Thus, with the feature selection output, we have

$$\begin{aligned}
&\mathcal{L}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}(\phi^*, \theta^*_f, f) \\
\leq\ & 2\left(\mathcal{L}_{\mathcal{D}}(\hat{\phi}, \hat{w}_f, f) - \mathcal{L}_{\mathcal{D}}(\tilde{\phi}_f, \tilde{w}_f, f)\right) + 2\left(\mathcal{L}_{\mathcal{D}}(\phi^*, w^*_f, f) - \mathcal{L}_{\mathcal{D}}(\phi^*, \theta^*_f, f)\right) + 2\gamma
\end{aligned}$$

$$
\begin{aligned}
&\le 2\max_{g\in\mathcal{F}}\Big(\mathcal{L}_{\mathcal{D}}(\hat{\phi},\hat{w}_g,g)-\mathcal{L}_{\mathcal{D}}(\tilde{\phi}_g,\tilde{w}_g,g)\Big)+3\left(\mathcal{L}(\phi^*,w_f^*,f)-\mathcal{L}(\phi^*,\theta_f^*,f)\right)+4\gamma\\
&\le 2\max_{g\in\mathcal{F}}\Big(\mathcal{L}_{\mathcal{D}}(\phi^*,w_g^*,g)-\mathcal{L}_{\mathcal{D}}(\tilde{\phi}_g,\tilde{w}_g,g)\Big)+3\left(\mathcal{L}(\phi^*,w_f^*,f)-\mathcal{L}(\phi^*,\theta_f^*,f)\right)+4\gamma\\
&\le 2\max_{g\in\mathcal{F}}\left(\mathcal{L}_{\mathcal{D}}(\phi^*,w_g^*,g)-\mathcal{L}_{\mathcal{D}}(\phi^*,\theta_g^*,g)\right)+2\max_{g\in\mathcal{F}}\Big(\mathcal{L}_{\mathcal{D}}(\phi^*,\theta_g^*,g)-\mathcal{L}_{\mathcal{D}}(\tilde{\phi}_g,\tilde{w}_g,g)\Big)\\
&\quad+3\left(\mathcal{L}(\phi^*,w_f^*,f)-\mathcal{L}(\phi^*,\theta_f^*,f)\right)+4\gamma\\
&\le 2\max_{g\in\mathcal{F}}\left(\mathcal{L}_{\mathcal{D}}(\phi^*,w_g^*,g)-\mathcal{L}_{\mathcal{D}}(\phi^*,\theta_g^*,g)\right)+3\left(\mathcal{L}(\phi^*,w_f^*,f)-\mathcal{L}(\phi^*,\theta_f^*,f)\right)+6\gamma\\
&\le 6\max_{g\in\mathcal{F}}\left(\mathcal{L}(\phi^*,w_g^*,g)-\mathcal{L}(\phi^*,\theta_g^*,g)\right)+8\gamma.
\end{aligned}
$$

The third inequality uses the fact that $\hat{\phi}$ is the solution of the ridge regression based feature selection objective. Further, in all steps, we repeatedly apply the deviation bound from Lemma 34 to move from $\mathcal{L}_{\mathcal{D}}(\cdot)$ to $\mathcal{L}(\cdot)$.

Now, for ridge regression estimate w_g^*, we can bound the bias term on the RHS as follows:

$$
\begin{aligned}
&\mathcal{L}(\phi^*,w_g^*,g)-\mathcal{L}(\phi^*,\theta_g^*,g)=\mathbb{E}\left[\left(\left\langle\phi^*,w_g^*\right\rangle-\left\langle\phi^*,\theta_g^*\right\rangle\right)^2\right]\\
&=\|w_g^*-\theta_g^*\|_{\Sigma^*}^2=\sum_{i=1}^d\lambda_i\left\langle v_i,w_g^*-\theta_g^*\right\rangle^2=\sum_{i=1}^d\lambda_i\left(\frac{\lambda_i}{\lambda+\lambda_i}\left\langle v_i,\theta_g^*\right\rangle-\left\langle v_i,\theta_g^*\right\rangle\right)^2\\
&=\sum_{i=1}^d\frac{\lambda_i\lambda^2\left\langle v_i,\theta_g^*\right\rangle^2}{(\lambda_i+\lambda)^2}\le\frac{\lambda}{4}\|\theta_g^*\|_2^2\le\frac{\lambda d^2}{4},
\end{aligned}
$$

where (λ_i,v_i) denote the i-th eigenvalue-eigenvector pair of the population covariance matrix Σ^* for feature ϕ^*. In the derivation above, we use the fact that $w_g^*=(\Sigma+\lambda I)^{-1}\mathbb{E}[\phi^*g]=\frac{\lambda_i}{\lambda+\lambda_i}\left\langle v_i,\theta_g^*\right\rangle$.

Therefore, the final deviation bound for $\hat{\phi}$ is

$$
\mathcal{L}(\hat{\phi},\hat{w}_f,f)-\mathcal{L}(\phi^*,\theta_f^*,f)\le\frac{3\lambda d^2}{2}+\frac{32c_3d\log(2n(1/\lambda)|\Phi_h||\Phi_{h+1}|/\delta)}{\lambda^2n}.
$$

Setting $\lambda=\tilde{O}\left(\frac{1}{n^{1/3}}\right)$ gives us the final result. ■

Appendix D. FQI Planning Results

In this section, we provide various FQI (Fitted Q-iteration) planning results. In Appendix D.1, we provide the general framework of FQI algorithms. In Appendix D.2, we show the sample complexity of FQI-FULL-CLASS that handles the offline planning for a class of rewards. In Appendix D.3, we discuss the sample complexity result of planning with the learned feature $\bar{\phi}$. In Appendix D.4, we provide the sample complexity guarantee for planning for the elliptical reward class. We want to mention that we abuse some notations in this section. For example, $\mathcal{F}_h,\mathcal{G}_h,\mathcal{V}_h$ may have different meanings from the main text. However, they should be clear within the context.

For simplicity, we use the horizon H in Algorithm 5 and all statements. When FQI is called with a smaller horizon $\tilde{H} \leq H$, we can just replace all H by $\tilde{H}$. The analyses still go through and the sample complexity will not exceed the one instantiated with H.

D.1 FQI Planning Algorithm

In this part, we present the general framework of FQI planner in Algorithm 5, which subsumes three different algorithms: FQI-FULL-CLASS, FQI-REPRESENTATION, and FQI-ELLIPTICAL. FQI-FULL-CLASS and FQI-REPRESENTATION will be used to plan for a finite deterministic reward class, while FQI-ELLIPTICAL is specialized in planning for the elliptical reward class.

Algorithm 5 FQI: Fitted Q-Iteration

input: (1) exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ sampled from ρ_{h-3}^{+3} with size n at each level $h \in [H]$, (2) reward function $R = R_{0:H-1}$ with $R_h : \mathcal{X} \times \mathcal{A} \to [0,1], \forall h \in [H]$, (3) function class: (i) for FQI-FULL-CLASS, $\mathcal{F}_h(R_h) \coloneqq \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \leq H\sqrt{d}, \phi_h \in \Phi_h\}, h \in [H]$; (ii) for FQI-REPRESENTATION, $\mathcal{F}_h(R_h) \coloneqq \{R_h + \mathrm{clip}_{[0,H]}\left(\langle \bar{\phi}_h, w_h \rangle\right) : \|w_h\|_2 \leq B\}, h \in [H]$; (iii) for FQI-ELLIPTICAL, $\mathcal{F}_h(R_h) \coloneqq \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \leq \sqrt{d}, \phi_h \in \Phi_h\}, h \in [H]$.
Set $\hat{V}_H(x) = 0$.
for $h = H-1, \ldots, 0$ **do**
  Pick n samples $\left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^n$ from the exploratory dataset $\mathcal{D}_h$.
  Solve least squares problem:

$$\hat{f}_{h,R_h} \leftarrow \operatorname*{argmin}_{f_h \in \mathcal{F}_h(R_h)} \mathcal{L}_{\mathcal{D}_h,R_h}(f_h, \hat{V}_{h+1}), \tag{44}$$

  where $\mathcal{L}_{\mathcal{D}_h,R_h}(f_h, \hat{V}_{h+1}) := \sum_{i=1}^n \left(f_h\left(x_h^{(i)}, a_h^{(i)}\right) - R_h\left(x_h^{(i)}, a_h^{(i)}\right) - \hat{V}_{h+1}\left(x_{h+1}^{(i)}\right)\right)^2$.
  if FQI-FULL-CLASS or FQI-REPRESENTATION **then**
    Define $\hat{\pi}_h(x) = \mathrm{argmax}_a \hat{f}_{h,R_h}(x,a)$ and $\hat{V}_h(x) = \mathrm{clip}_{[0,H]}\left(\max_a \hat{f}_{h,R_h}(x,a)\right)$.
  else if FQI-ELLIPTICAL **then**
    Define $\hat{\pi}_h(x) = \mathrm{argmax}_a \hat{f}_{h,R_h}(x,a)$ and $\hat{V}_h(x) = \mathrm{clip}_{[0,1]}\left(\max_a \hat{f}_{h,R_h}(x,a)\right)$.
  end if
end for
return $\hat{\pi} = (\hat{\pi}_0, \ldots, \hat{\pi}_{H-1})$.

This leads to the different bounds of parameters in the Q-value function classes and different clipping thresholds of the state-value functions. In addition, for FQI-FULL-CLASS and FQI-ELLIPTICAL, we use all features in Φ to construct the Q-value function classes, while in FQI-REPRESENTATION we only use the the learned representation $\bar{\phi}$. The details can be found below. When calling Algorithm 5 and there is no confusion, we sometimes drop the input (3) function class for simplicity.

Unlike the regression problem in the main text, the objective here includes an additional reward function component. Therefore, we define a new loss function $\mathcal{L}_{D_h,R_h}$, and will use $\mathcal{L}_{\rho_{h-3}^{+3},R_h}$ to denote its population version. Notice that the function class $\mathcal{F}_h(R_h)$ in

Algorithm 5 also depends on the reward function R. If we pull out the reward term from $\hat{f}_{h,R_h}$, we can obtain an equivalent solution of the least squares problem Eq. (44) as below

$$\hat{f}_{h,R_h} = R_h + \operatorname*{argmin}_{f_h \in \mathcal{F}_h(\mathbf{0})} \mathcal{L}_{\mathcal{D}_h}(f_h, \hat{V}_{h+1}),$$

where

$$\mathcal{L}_{\mathcal{D}_h}(f_h, \hat{V}_{h+1}) := \sum_{i=1}^{n} \left(f_h\left(x_h^{(i)}, a_h^{(i)}\right) - \hat{V}_{h+1}\left(x_{h+1}^{(i)}\right)\right)^2.$$

Intuitively, the reward function R_h only makes the current least squares solution offset the original (reward-independent) least squares solution by R_h.

D.2 Planning for a Reward Class with the Full Representation Class

In this part, we first establish the sample complexity of planning for a prespecified deterministic reward function R in Lemma 19. We will choose FQI-FULL-CLASS as the planner, where the Q-value function class consists of linear function of all features in the feature class with reward appended. Specifically, we have $\mathcal{F}_h(R_h) := \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \le H\sqrt{d}, \phi_h \in \Phi_h\}, h \in [H]$. Equipped with this lemma, we also provide the sample complexity of planning for a finite deterministic reward class $\mathcal{R}$ in Corollary 20. The analysis is similar to that of Chen and Jiang (2019); Agarwal et al. (2020b).

Lemma 19 (Planning for a prespecified reward with the entire representation class) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$, which is collected from ρ_{h-3}^{+3} and satisfies Eq. (17) for all $h \in [H]$. For a prespecified deterministic reward function $R = R_{0:H-1}, R_h : \mathcal{X} \times \mathcal{A} \to [0,1], \forall h \in [H]$ and $\delta \in (0,1)$, if we set*

$$n \ge \frac{32 c_3 H^6 d^2 \kappa K}{\beta^2} \log\left(\frac{16 c_3 H^6 d^2 \kappa K}{\beta^2}\right) + \frac{32 c_3 H^6 d^2 \kappa K}{\beta^2} \log\left(\tfrac{4|\Phi| H^3 d}{\delta}\right),$$

where c_3 is the constant in Lemma 34, then with probability at least $1-\delta$, the policy $\hat{\pi}$ returned by FQI-FULL-CLASS *satisfies*

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] \ge \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \beta.$$

Proof We first bound the difference in cumulative rewards between $\hat{\pi} := \hat{\pi}_{0:H-1}$ and the optimal policy π^* for the given reward function. Recall that $\hat{\pi}_0$ is greedy w.r.t. $\hat{f}_{0,R_0}$, which implies that $\hat{f}_{0,R_0}(x_0, \hat{\pi}_0(x_0)) \ge \hat{f}_0(x_0, \pi^*(x_0))$ for all x_0. Hence, we have

$$\begin{aligned}
& v_R^* - v_R^{\hat{\pi}} \\
&= \mathbb{E}_{\pi^*}\left[R_0(x_0,a_0) + V_1^*(x_1)\right] - \mathbb{E}_{\hat{\pi}}\left[R_0(x_0,a_0) + V_1^{\hat{\pi}}(x_1)\right] \\
&\le \mathbb{E}_{\pi^*}\left[R_0(x_0,a_0) + V_1^*(x_1) - \hat{f}_0(x_0,a_0)\right] - \mathbb{E}_{\hat{\pi}}\left[R_0(x_0,a_0) + V_1^{\hat{\pi}}(x_1) - \hat{f}_{0,R_0}(x_0,a_0)\right] \\
&= \mathbb{E}_{\pi^*}\left[R_0(x_0,a_0) + V_1^*(x_1) - \hat{f}_{0,R_0}(x_0,a_0)\right] - \mathbb{E}_{\hat{\pi}}\left[R_0(x_0,a_0) + V_1^*(x_1) - \hat{f}_{0,R_0}(x_0,a_0)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{\hat{\pi}}\left[V_1^*(x_1) - V_1^{\hat{\pi}}(x_1)\right] \\
&= \mathbb{E}_{\pi^*}\left[Q_0^*(x_0,a_0) - \hat{f}_{0,R_0}(x_0,a_0)\right] - \mathbb{E}_{\hat{\pi}}\left[Q_0^*(x_0,a_0) - \hat{f}_{0,R_0}(x_0,a_0)\right] \\
&\quad + \mathbb{E}_{\hat{\pi}}\left[V_1^*(x_1) - V_1^{\hat{\pi}}(x_1)\right].
\end{aligned}
$$

Continuing unrolling to $h = H-1$, we get

$$
\begin{aligned}
&v_R^* - v_R^{\hat{\pi}} \\
&\leq \sum_{h=0}^{H-1} \mathbb{E}_{\hat{\pi}_{0:h-1}\circ\pi^*}\left[Q_h^*(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right] - \sum_{h=0}^{H-1} \mathbb{E}_{\hat{\pi}_{0:h}}\left[Q_h^*(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right].
\end{aligned}
$$

Now we bound each of these terms. The two terms only differ in the policies that generate the data and can be handled similarly. Therefore, in the following, we focus on just one of them. For any function $V_{h+1} : \mathcal{X} \to \mathbb{R}$, we introduce Bellman backup operator $(\mathcal{T}_h V_{h+1})(x_h,a_h) := R_h(x_h,a_h) + \mathbb{E}\left[V_{h+1}(x_{h+1}) \mid x_h, a_h\right]$. Let's call the roll-in policy π and drop the dependence on h. This gives us

$$
\begin{aligned}
&\left|\mathbb{E}_\pi\left[Q^*(x,a) - \hat{f}_R(x,a)\right]\right| = \left|\mathbb{E}_\pi\left[R(x,a) + \mathbb{E}\left[V^*(x') \mid x,a\right] - \hat{f}_R(x,a)\right]\right| \\
&\leq \mathbb{E}_\pi\left[\left|R(x,a) + \mathbb{E}\left[V^*(x') \mid x,a\right] - \hat{f}_R(x,a)\right|\right] \\
&\leq \mathbb{E}_\pi\left[\left|\mathbb{E}\left[V^*(x') \mid x,a\right] - \mathbb{E}\left[\hat{V}(x') \mid x,a\right]\right| + \left|R(x,a) + \mathbb{E}\left[\hat{V}(x') \mid x,a\right] - \hat{f}_R(x,a)\right|\right] \\
&\leq \mathbb{E}_\pi\left[\left|V^*(x') - \hat{V}(x')\right| + \left|(\mathcal{T}\hat{V})(x,a) - \hat{f}_R(x,a)\right|\right],
\end{aligned}
$$

where the last inequality is due to Jensen's inequality.

From the definition of $\hat{V}(x')$, we have

$$
\begin{aligned}
\mathbb{E}_\pi\left[\left|V^*(x') - \hat{V}(x')\right|\right] &\leq \mathbb{E}_\pi\left[\left|\max_a Q^*(x',a) - \max_{a'} \hat{f}_R(x',a')\right|\right] \\
&\leq \mathbb{E}_{\pi\circ\tilde{\pi}}\left[\left|Q^*(x',a') - \hat{f}_R(x',a')\right|\right].
\end{aligned}
$$

In the last inequality, we define $\tilde{\pi}$ to be the greedy one between two actions, that is we set $\tilde{\pi}(x') = \mathrm{argmax}_{a'} \max\{Q^*(x',a'), \hat{f}(x',a')\}$. This expression has the same form as the initial one, while at the next timestep. Keep unrolling yields

$$
\begin{aligned}
&\left|\mathbb{E}_\pi\left[Q_h^*(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right]\right| \\
&\leq \sum_{\tau=h}^{H-1} \max_{\pi_\tau} \mathbb{E}_{\pi_\tau}\left[\left|(\mathcal{T}_\tau \hat{V}_{\tau+1})(x_\tau,a_\tau) - \hat{f}_{\tau,R_\tau}(x_\tau,a_\tau)\right|\right] \\
&\leq \sum_{\tau=h}^{H-1} \max_{\pi_\tau} \sqrt{\mathbb{E}_{\pi_\tau}\left[\left[(\mathcal{T}_\tau \hat{V}_{\tau+1})(x_\tau,a_\tau) - \hat{f}_{\tau,R_\tau}(x_\tau,a_\tau)\right]^2\right]} \\
&\leq \sum_{\tau=h}^{H-1} \sqrt{\kappa K \mathbb{E}_{\rho_{\tau-3}^{+3}}\left[\left[(\mathcal{T}_\tau \hat{V}_{\tau+1})(x_\tau,a_\tau) - \hat{f}_{\tau,R_\tau}(x_\tau,a_\tau)\right]^2\right]},
\end{aligned}
$$

where the last inequality is due to condition Eq. (17).

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

$$\begin{aligned}
&\mathbb{E}_{\rho^{+3}_{\tau-3}}\left[\left((\mathcal{T}_\tau \hat{V}_{\tau+1})(x_\tau,a_\tau)-\hat{f}_{\tau,R_\tau}(x_\tau,a_\tau)\right)^2\right] \\
&=\mathbb{E}_{\rho^{+3}_{\tau-3}}\Big[\left(R_\tau(x_\tau,a_\tau)+\hat{V}_{\tau+1}(x_{\tau+1})-\hat{f}_{\tau,R_\tau}(x_\tau,a_\tau)\right)^2 \\
&\qquad -\left(R_\tau(x_\tau,a_\tau)+\hat{V}_{\tau+1}(x_{\tau+1})-(\mathcal{T}_\tau\hat{V}_{\tau+1})(x_\tau,a_\tau)\right)^2\Big] \\
&=\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_\tau,R_\tau}(\hat{f}_{\tau,R_\tau},\hat{V}_{\tau+1})\right]-\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_\tau,R_\tau}(\mathcal{T}_\tau\hat{V}_{\tau+1},\hat{V}_{\tau+1})\right] \\
&\le \frac{16c_3H^2d^2\log(4nH^3d|\Phi|/\delta)}{n}. \qquad \text{(Step (*), Lemma 21)}
\end{aligned}$$

Plugging this back into the overall value performance difference, the bound is

$$v^*_R-v^{\hat{\pi}}_R\le H^2\sqrt{\kappa K}\sqrt{\frac{16c_3H^2d^2\log(4nH^3d|\Phi|/\delta)}{n}}.$$

Setting RHS to be less than β and reorganize, we get

$$n\ge\frac{16c_3H^6d^2\kappa K\log(4nH^3d|\Phi|/\delta)}{\beta^2}.$$

A sufficient condition for the inequality above is

$$n\ge\frac{32c_3H^6d^2\kappa K}{\beta^2}\log\left(\frac{16c_3H^6d^2\kappa K}{\beta^2}\right)+\frac{32c_3H^6d^2\kappa K}{\beta^2}\log\left(\tfrac{4|\Phi|H^3d}{\delta}\right),$$

which completes the proof. ■

Corollary 20 (Planning for a reward class with a full representation class) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$, which is collected from ρ^{+3}_{h-3} and satisfies Eq. (17) for all $h\in[H]$, and we are given a finite deterministic reward class $\mathcal{R}=\mathcal{R}_0\times\ldots\times\mathcal{R}_{H-1},\mathcal{R}_h\subseteq(\mathcal{X}\times\mathcal{A}\to[0,1]),\forall h\in[H]$. For $\delta\in(0,1)$ and any reward function $R\in\mathcal{R}$, if we set*

$$n\ge\frac{32c_3H^6d^2\kappa K}{\beta^2}\log\left(\frac{16c_3H^6d^2\kappa K}{\beta^2}\right)+\frac{32c_3H^6d^2\kappa K}{\beta^2}\log\left(\tfrac{4|\Phi|\mathcal{R}||H^3d}{\delta}\right),$$

where c_3 is the constant in Lemma 34, then with probability at least $1-\delta$, the policy $\hat{\pi}$ returned by FQI-full-class *satisfies*

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{h=0}^{H-1}R_h(x_h,a_h)\right]\ge\max_\pi\mathbb{E}_\pi\left[\sum_{h=0}^{H-1}R_h(x_h,a_h)\right]-\beta.$$

Proof For any fixed reward $R \in \mathcal{R}$, applying Lemma 19 yields that with probability $1-\delta'$,

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] \geq \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \beta,$$

if we set n according to the lemma statement.

Then union bounding over $R \in \mathcal{R}$ and setting $\delta = \delta'/|\mathcal{R}|$ gives us the desired result. ■

Lemma 21 (Deviation bound for Lemma 19) *Given a deterministic reward function $R = R_{0:H-1}, R_h : \mathcal{X} \times \mathcal{A} \to [0,1], \forall h \in [H]$ and a dataset $\{\mathcal{D}\}_{0:H-1}$ collected from ρ^{+3}_{h-3}, where $\mathcal{D}_h$ is $\left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^{n}$. With probability at least $1-\delta$, $\forall\, h \in [H], V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, we have*

$$\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\hat{f}_{h,R_h}, V_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})\right]\right| \leq \frac{16 c_3 H^2 d^2 \log(4nH^3 d|\Phi|/\delta)}{n},$$

Here c_3 is the constant in Lemma 34, $\mathcal{V}_{h+1}(R_{h+1}) := \{\mathrm{clip}_{[0,H]}\left(\max_a f_{h+1,R_{h+1}}(x_{h+1}, a)\right) : f_{h+1,R_{h+1}} \in \mathcal{F}_{h+1}(R_{h+1})\}$ for $h \in [H-1]$ and $\mathcal{V}_H = \{\mathbf{0}\}$ is the state-value function class, and $\mathcal{F}_h(R_h) := \{R_h + \langle \phi_h, w_h\rangle : \|w_h\|_2 \leq H\sqrt{d}, \phi_h \in \Phi_h\}$ for $h \in [H]$ is the reward dependent Q-value function class.

Proof Recall the definition, we have $\hat{f}_{h,R_h} = R_h + \hat{f}_h$, $\hat{f}_h = \mathrm{argmin}_{f_h \in \mathcal{F}_h(\mathbf{0})} \mathcal{L}_{\mathcal{D}_h}(f_h, V_{h+1})$, $\mathcal{F}_h(\mathbf{0}) := \{\langle \phi_h, w_h\rangle : \|w_h\|_2 \leq H\sqrt{d}, \phi_h \in \Phi_h\}$, and $\mathcal{L}_{\mathcal{D}_h}(f_h, V_{h+1}) := \sum_{i=1}^{n}\left(f_h\left(x_h^{(i)}, a_h^{(i)}\right) - V_{h+1}\left(x_{h+1}^{(i)}\right)\right)^2$. Therefore we have $\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h}, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})$ and $\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1}, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\mathcal{T}_h V_{h+1} - R_h, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\langle \phi_h^*, \theta^*_{V_{h+1}}\rangle, V_{h+1})$. Then it suffices to show the bound between $\mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})$ and $\mathcal{L}_{\mathcal{D}_h}(\langle \phi_h^*, \theta^*_{V_{h+1}}\rangle, V_{h+1})$.

Firstly, we fix $h \in [H]$. Noticing the structure of $\mathcal{F}_h(\mathbf{0})$, for any $f_h \in \mathcal{F}_h(\mathbf{0})$, we can associate it with some $\phi_h \in \Phi_h$ and w_h that satisfies $\|w_h\|_2 \leq H\sqrt{d}$. Therefore, we can equivalently write $\mathcal{L}_{D_h}(f_h, V_{h+1})$ as $\mathcal{L}_{D_h}(\phi_h, w_h, V_{h+1})$. Also noticing the structure of $\mathcal{V}_{h+1}(R_{h+1})$, we can directly apply Lemma 34 with ρ^{+3}_{h-3}, Φ_h, Φ_{h+1}, $B = H\sqrt{d}$, $L = H$, $\mathcal{R} = \{R\}$, and π' be the greedy policy.

This implies that for all $\|w_h\|_2 \leq H\sqrt{d}$, $\phi_h \in \Phi_h$, and $V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, with probability at least $1 - \delta/H$, we have

$$\begin{aligned}
&\mathcal{L}_{\rho^{+3}_{h-3}}(\phi_h, w_h, V_{h+1}) - \mathcal{L}_{\rho^{+3}_{h-3}}(\phi_h^*, \theta^*_{V_{h+1}}, V_{h+1}) \\
&\leq 2\left(\mathcal{L}_{\mathcal{D}_h}(\phi_h, w_h, V_{h+1}) - \mathcal{L}_{\mathcal{D}_h}(\phi_h^*, \theta^*_{V_{h+1}}, V_{h+1})\right) + \frac{16 c_3 H^2 d^2 \log(4nH^2 d|\Phi|H/\delta)}{n}. \qquad (45)
\end{aligned}$$

Notice that here we use the property that $\phi_h^* \in \Phi_h$ and $\|\theta^*_{V_{h+1}}\|_2 \leq H\sqrt{d}$ from Lemma 3.

From the definition, we have $\hat{f}_h = \mathrm{argmin}_{f_h \in \mathcal{F}_h(\mathbf{0})} \mathcal{L}_{\mathcal{D}_h}(f_h, V_{h+1})$. Noticing the structure of $\mathcal{F}_h(\mathbf{0})$, we can write $\hat{f}_h = \langle \hat{\phi}_h, \hat{w}_h\rangle$, where $\hat{\phi}_h, \hat{w}_h = \mathrm{argmin}_{\phi_h \in \Phi_h, \|w_h\|_2 \leq H\sqrt{d}} \mathcal{L}_{\mathcal{D}_h}(\phi_h, w_h, V_{h+1})$ (here we abuse the notation of $\hat{\phi}_h$, which is reserved for the learned feature).

This implies $\mathcal{L}_{\mathcal{D}_h}(\hat{\phi}_h, \hat{w}_h, V_{h+1}) - \mathcal{L}_{\mathcal{D}_h}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1}) \le 0$. Therefore, with probability at least $1 - \delta/H$

$$\left|\mathcal{L}_{\rho^{+3}_{h-3}}(\hat{\phi}_h, \hat{w}_h, V_{h+1}) - \mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})\right| \le \frac{16c_3H^2d^2\log(4nH^3d|\Phi|/\delta)}{n}.$$

Finally, by definition $\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})\right] = \mathcal{L}_{\rho^{+3}_{h-3}}(\hat{\phi}_h, \hat{w}_h, V_{h+1})$ and $\mathbb{E}[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})] = \mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})$, union bounding over $h \in [H]$, we complete the proof. ■

D.3 Planning for a Reward Class with the Learned Representation Function

In this part, we will show that the learned feature $\bar{\phi}$ enables the downstream policy optimization for a finite deterministic reward class $\mathcal{R}$. The sample complexity is shown in Lemma 22. We will choose FQI-REPRESENTATION as the planner, where the Q-value function class only consists of linear function of learned feature $\bar{\phi}$ with reward appended. Specifically, we have $\mathcal{F}_h(R_h) \coloneqq \{R_h + \mathrm{clip}_{[0,H]}\left(\langle\bar{\phi}_h, w_h\rangle\right) : \|w_h\|_2 \le B\}, h \in [H]$. In addition to constructing the function class with learned feature itself, we also perform clipping in $\mathcal{F}_h(R_h)$. This clipping variant helps us avoid the poly(B) dependence in the sample complexity. Notice that clipped Q-value function classes also work for FQI-FULL-CLASS and FQI-ELLIPTICAL, and would save d factor. We only introduce this variant here because B is much larger than $H\sqrt{d}$ or d.

Lemma 22 (Planning for a reward class with a learned representation) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ (collected from ρ^{+3}_{h-3} and satisfies Eq. (17) for all $h \in [H]$), a learned feature $\bar{\phi}_h$ that satisfies the condition in Eq. (18), and a finite deterministic reward class $\mathcal{R} = \mathcal{R}_0 \times \ldots \times \mathcal{R}_{H-1}, \mathcal{R}_h \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1]), \forall h \in [H]$. For $\delta \in (0,1)$ and any reward function $R \in \mathcal{R}$, if we set*

$$n \ge \frac{2c_4H^6d\kappa K}{\beta^2}\log\left(\frac{c_4H^6d\kappa K}{\beta^2}\right) + \frac{2c_4H^6d\kappa K}{\beta^2}\log\left(\tfrac{|\mathcal{R}|BH}{\delta}\right),$$

where c_4 is the constant in Lemma 23, then with probability at least $1-\delta$, the policy $\hat{\pi}$ returned by FQI-Representation *satisfies*

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] \ge \max_{\pi}\mathbb{E}_{\pi}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \beta - H^2\sqrt{\kappa K\varepsilon_{\mathrm{apx}}}.$$

Proof Following similar steps in the proof of Lemma 19 and replacing Lemma 21 with Lemma 23 in Step(*), with probability at least $1-\delta$, we have

$$\begin{aligned} v^*_R - v^{\hat{\pi}}_R \le& H^2\sqrt{\kappa K}\sqrt{\frac{c_4dH^2\log(nB|\mathcal{R}|H/\delta)}{n} + \varepsilon_{\mathrm{apx}}} \\ \le& H^2\sqrt{\kappa K}\sqrt{\frac{c_4dH^2\log(nB|\mathcal{R}|H/\delta)}{n}} + H^2\sqrt{\kappa K\varepsilon_{\mathrm{apx}}}. \end{aligned}$$

Setting RHS to be less than $\beta + H^2\sqrt{\kappa K\varepsilon_{\mathrm{apx}}}$ and reorganize, we get the condition

$$n \ge \frac{c_4H^6d\kappa K\log(n|\mathcal{R}|BH/\delta)}{\beta^2}.$$

A sufficient condition for the inequality above is

$$n \geq \frac{2c_4 H^6 d\kappa K}{\beta^2} \log\left(\frac{c_4 H^6 d\kappa K}{\beta^2}\right) + \frac{2c_4 H^6 d\kappa K}{\beta^2} \log\left(\tfrac{|\mathcal{R}|BH}{\delta}\right),$$

which completes the proof. ■

Lemma 23 (Deviation bound for Lemma 22) *Assume that we have an exploratory dataset $\mathcal{D}_h \coloneqq \left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^n$ collected from ρ_{h-3}^{+3}, $h \in [H]$, a learned feature $\bar{\phi}_h$ that satisfies the condition in Eq. (18), and a finite deterministic reward class $\mathcal{R} = \mathcal{R}_0 \times \ldots \times \mathcal{R}_{H-1}, \mathcal{R}_h \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1]), \forall h \in [H]$. Then, with probability at least $1-\delta$, $\forall R \in \mathcal{R}, h \in [H], V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, we have*

$$\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h}, V_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})\right]\right| \leq \frac{c_4 d H^2 \log(nB|\mathcal{R}|H/\delta)}{n} + \varepsilon_{\text{apx}},$$

Here c_4 is some universal constant, $\mathcal{V}_{h+1}(R_{h+1}) := \{\text{clip}_{[0,H]}\left(\max_a f_{h+1,R_{h+1}}(x_{h+1}, a)\right) : f_{h+1,R_{h+1}} \in \mathcal{F}_{h+1}(R_{h+1})\}$ for $h \in [H-1]$ and $\mathcal{V}_H = \{\mathbf{0}\}$ are the state-value function class, $\mathcal{F}_h(R_h) \coloneqq \{R_h + \text{clip}_{[0,H]}\left(\langle\bar{\phi}_h, w_h\rangle\right) : \|w_h\|_2 \leq B\}$ for $h \in [H]$ is the reward dependent Q-value function class.

Proof We start with any fixed $h \in [H]$. Recall the definition $\hat{f}_{h,R_h} = R_h + \hat{f}_h$, where $\hat{f}_h = \text{argmin}_{f_h \in \mathcal{F}_h(\mathbf{0})} \mathcal{L}_{\mathcal{D}_h}(f_h, V_{h+1})$, $\mathcal{L}_{\mathcal{D}_h}(f_h, V_{h+1}) \coloneqq \sum_{i=1}^n \left(f_h\left(x_h^{(i)}, a_h^{(i)}\right) - V_{h+1}\left(x_{h+1}^{(i)}\right)\right)^2$, and $\mathcal{F}_h(\mathbf{0}) \coloneqq \{\text{clip}_{[0,H]}\left(\langle\bar{\phi}_h, w_h\rangle\right) : \|w_h\|_2 \leq B\}$. Therefore we get $\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h}, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})$ and $\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1}, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\mathcal{T}_h V_{h+1} - R_h, V_{h+1}) = \mathcal{L}_{\mathcal{D}_h}(\langle\phi_h^*, \theta_{V_{h+1}}^*\rangle, V_{h+1})$. Then it suffices to show the bound between $\mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})$ and $\mathcal{L}_{\mathcal{D}_h}(\langle\phi_h^*, \theta_{V_{h+1}}^*\rangle, V_{h+1})$.

Firstly, we show that condition Eq. (18) implies the following approximation guarantee for $\bar{\phi}_h$. For all $h, V_{h+1} \in \{\text{clip}_{[0,H]}\left(\max_a(R_{h+1}(x_{h+1}, a) + \text{clip}_{[0,H]}\left(\langle\phi_{h+1}(x_{h+1}, a), \theta_{h+1}\rangle\right))\right) : \phi_{h+1} \in \Phi_{h+1}, \|\theta_{h+1}\|_2 \leq B, R \in \mathcal{R}\}$, let $\bar{w}_{V_{h+1}} = \text{argmin}_{\|w_h\|_2 \leq B} \mathcal{L}_{\mathcal{D}_h}(\bar{\phi}_h,\ w_h, V_{h+1})$ with $B \geq H\sqrt{d}$, then we have

$$\mathbb{E}_{\rho_{h-3}^{+3}}\left[\left(\left\langle\bar{\phi}_h(x_h, a_h), \bar{w}_{V_{h+1}}\right\rangle - \mathbb{E}\left[V_{h+1}(x_{h+1})|x_h, a_h\right]\right)^2\right] \leq \varepsilon_{\text{apx}}.$$

This is because the order of taking max and clipping doesn't matter:

$$\begin{aligned}
&\text{clip}_{[0,H]}\left(\max_a\left(R_{h+1}(x_{h+1}, a) + \text{clip}_{[0,H]}\left(\langle\phi_{h+1}(x_{h+1}, a), \theta_{h+1}\rangle\right)\right)\right) \\
&= \text{clip}_{[0,H]}\left(\max_a\left(R_{h+1}(x_{h+1}, a) + \langle\phi_{h+1}(x_{h+1}, a), \theta_{h+1}\rangle\right)\right).
\end{aligned}$$

Since we have clipping, we now define $\mathcal{L}^{\text{c}}_{\rho_{h-3}^{+3}}(\phi_h, w_h, V_{h+1}) \coloneqq \mathbb{E}_{\rho_{h-3}^{+3}}[(\text{clip}_{[0,H]}\left(\langle\phi_h, w_h\rangle\right) - V_{h+1})^2]$ and $\mathcal{L}^{\text{c}}_{\mathcal{D}_h}(\cdot)$ as its empirical version. Then we can follow the similar steps in Lemma 34 and get the concentration result. For any $R \in \mathcal{R}, V_{h+1} \in \mathcal{V}_{h+1}(R)$ and $\|w_h\|_2 \leq B$, we have that with probability at least $1-\delta'$

$$\left|\mathcal{L}^{\text{c}}_{\rho_{h-3}^{+3}}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\text{c}}_{\rho_{h-3}^{+3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right.$$

$$
\begin{aligned}
&\quad - \left(\mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right)\Big| \\
&\leq \frac{1}{2}\left(\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right) + \frac{c_4' dH^2 \log(nB|\mathcal{R}|/\delta')}{n},
\end{aligned}
$$

where c_4' is some universal constant. Note that the slight difference is that here we will set the feature class to be $\{\bar{\phi}_h\}$ and $\{\bar{\phi}_{h+1}\}$, change the term related to ϕ^*, θ^*_V to $\bar{\phi}, \bar{w}_V$, change the norm constraints of the value function classes, and add the clipping on the corresponding function classes there. It is easy to show this concentration result, and now the range of the hypothesis functions is $16H^2$ and we can get rid of the union bound over the feature classes. Then we have

$$
\begin{aligned}
&\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1}) \\
\leq\ &\left(\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right) \\
&\quad + \left|\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})\right| \\
\leq\ &\left(\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right) + \varepsilon_{\mathrm{apx}} \\
\leq\ &2\left(\mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, w_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1})\right) + \frac{2c_4' dH^2 \log(nB|\mathcal{R}|/\delta')}{n} + \varepsilon_{\mathrm{apx}}
\end{aligned}
$$

The second inequality above is due to

$$
\begin{aligned}
&\left|\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \bar{w}_{V_{h+1}}, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})\right| \\
&= \mathbb{E}_{\rho^{+3}_{h-3}}\left[\left(\mathrm{clip}_{[0,H]}\left(\left\langle\bar{\phi}_h(x_h, a_h), \bar{w}_{V_{h+1}}\right\rangle\right) - \mathbb{E}\left[V_{h+1}(x_{h+1})|x_h, a_h\right]\right)^2\right] \\
&\leq \mathbb{E}_{\rho^{+3}_{h-3}}\left[\left(\left\langle\bar{\phi}_h(x_h, a_h), \bar{w}_{V_{h+1}}\right\rangle - \mathbb{E}\left[V_{h+1}(x_{h+1})|x_h, a_h\right]\right)^2\right] \leq \varepsilon_{\mathrm{apx}}.
\end{aligned}
$$

From the definition, we have $\hat{f}_h = \mathrm{argmin}_{f_h \in \mathcal{F}_h(\mathbf{0})} \mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(f_h, V_{h+1})$. Noticing the structure of $\mathcal{F}_h(\mathbf{0})$, we can write $\hat{f}_h = \mathrm{clip}_{[0,H]}\left(\langle\bar{\phi}_h, \hat{w}_h\rangle\right)$, where $\hat{w}_h = \mathrm{argmin}_{\|w_h\|_2 \leq B} \mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, w_h, V_{h+1})$. This implies $\mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, \hat{w}_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\mathcal{D}_h}(\bar{\phi}_h, \tilde{w}_{V_{h+1}}, V_{h+1}) \leq 0$.

Union bounding over $h \in [H]$, and setting $\delta = \delta'/H$, we have that with probability at least $1 - \delta$,

$$
\left|\mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \hat{w}_h, V_{h+1}) - \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})\right| \leq \frac{2c_4' dH^2 \log(nB|\mathcal{R}|H/\delta)}{n} + \varepsilon_{\mathrm{apx}}.
$$

Finally, noticing the property that $\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h}(\hat{f}_h, V_{h+1})\right] = \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\bar{\phi}_h, \hat{w}_h, V_{h+1})$ and $\mathbb{E}[\mathcal{L}_{\mathcal{D}_h, R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})] = \mathcal{L}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1}) = \mathcal{L}^{\mathrm{c}}_{\rho^{+3}_{h-3}}(\phi^*_h, \theta^*_{V_{h+1}}, V_{h+1})$, we complete the proof. ■

D.4 Planning for Elliptical Reward Functions

In this part, we show the sample complexity of FQI-ELLIPTICAL, which is specialized in planning for the elliptical reward class defined in Lemma 24. The Q-value function class consists of linear function of all features in the feature class with reward appended. Specifically, we have $\mathcal{F}_h(R_h) \coloneqq \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \leq \sqrt{d}, \phi_h \in \Phi_h\}, h \in [H]$. We still use the full representation class, but compared with FQI-FULL-CLASS, we use a different bound on the norm of the parameters.

Lemma 24 (Planning for elliptical reward functions) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ (collected from ρ^{+3}_{h-3} and satisfies Eq. (17) for all $h \in [H]$). For $\delta \in (0,1)$ and any deterministic elliptical reward function $R \in \mathcal{R}$, where $\mathcal{R} \coloneqq \{R_{0:H-1} : R_{0:H-2} = \mathbf{0}, R_{H-1} \in \{\phi_{H-1}^\top \Gamma^{-1} \phi_{H-1} : \phi_{H-1} \in \Phi_{H-1}, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$, if we set*

$$n \geq \frac{136 c_3 H^4 d^3 \kappa K}{\beta^2} \log\left(\frac{68 c_3 H^4 d^3 \kappa K}{\beta^2}\right) + \frac{136 c_3 H^4 d^3 \kappa K}{\beta^2} \log\left(\tfrac{2|\Phi|H}{\delta}\right),$$

where c_3 is the constant in Lemma 34, then with probability at least $1-\delta$, the policy $\hat{\pi}$ returned by FQI-ELLIPTICAL *satisfies*

$$\mathbb{E}_{\hat{\pi}}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] \geq \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \beta.$$

Remark 25 *Notice that the elliptical reward function only has a non-zero value at timestep $H-1$.*

Proof The proof mostly follows the steps in Lemma 19. Since we consider the deterministic elliptical reward function class, we apply Lemma 26 instead of Lemma 21 in Step (*). Then following a similar calculation gives us the result immediately. ■

Lemma 26 (Deviation bound for Lemma 24) *Consider the deterministic elliptical reward function classes $\mathcal{R} \coloneqq \{R_{0:H-1} : R_{0:H-2} = \mathbf{0}, R_{H-1} \in \mathcal{R}_{H-1} \coloneqq \{\phi_{H-1}^\top \Gamma^{-1} \phi_{H-1} : \phi_{H-1} \in \Phi_{H-1}, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$, an exploratory dataset $\mathcal{D}_h \coloneqq \left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^n$ collected from ρ^{+3}_{h-3}, $h \in [H]$. With probability at least $1-\delta$, $\forall\, R \in \mathcal{R}, h \in [H], V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, we have*

$$\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\hat{f}_{h,R_h}, V_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})\right]\right| \leq \frac{68 c_3 d^3 \log(2n|\Phi|H/\delta)}{n},$$

where c_3 is the constant in Lemma 34, $\mathcal{V}_{h+1}(R_{h+1}) \coloneqq \{\mathrm{clip}_{[0,1]}\left(\max_a f_{h+1,R_{h+1}}(x_{h+1}, a)\right) : f_{h+1,R_{h+1}} \in \mathcal{F}_{h+1}(R_{h+1})\}$ for $h \in [H]$ and $\mathcal{V}_H = \{\mathbf{0}\}$ is the state-value function class, and $\mathcal{F}_h(R_h) \coloneqq \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \leq \sqrt{d}, \phi_h \in \Phi_h\}$ for $h \in [H]$ is the reward dependent Q-value function class.

Proof Firstly, from Lemma 36, we know that there exists a γ-cover $\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}$ for the reward class $\mathcal{R}$. For any fixed $\tilde{R} \in \mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}$, we can follow the similar steps in Lemma 21 to get a concentration result. The differences are that the norm of w_h is now bounded by $\sqrt{d}$ instead of $H\sqrt{d}$, and we clip to $[0,1]$. Therefore, for this fixed $\tilde{R}$, with probability at least $1-\delta'$, we have that $\forall h \in [H], \tilde{V}_{h+1} \in \mathcal{V}_{h+1}(\tilde{R}_{h+1})$,

$$\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\hat{f}_{h,\tilde{R}_h},\tilde{V}_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\mathcal{T}_h\tilde{V}_{h+1},\tilde{V}_{h+1})\right]\right| \leq \frac{16c_3 d\log(4nH|\Phi|/\delta)}{n}.$$

Union bounding over all $\tilde{R} \in \mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}$ and setting $\delta = \delta'/|\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}|$, with probability at least $1-\delta$, we have

$$\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\hat{f}_{h,\tilde{R}_h},\tilde{V}_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\mathcal{T}_h\tilde{V}_{h+1},\tilde{V}_{h+1})\right]\right| \leq \frac{16c_3 d\log(4nH|\Phi||\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}|/\delta)}{n}.$$

Notice that $\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}$ is a γ-cover of $\mathcal{R}$, for any $R \in \mathcal{R}$, there exists $\tilde{R} \in \mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}$, such that $\|R_h - \tilde{R}_h\|_\infty \leq \gamma$. Therefore for any $f_{h,R_h} \in \mathcal{F}_h(R_h)$ and $V_h \in \mathcal{V}_h(R_h)$, there exists some $f_{h,\tilde{R}_h} \in \mathcal{F}(\tilde{R}_h)$ and $V_h \in \mathcal{V}(\tilde{R}_h)$ that satisfy $\|f_{h,R_h} - f_{h,\tilde{R}_h}\|_\infty \leq \gamma$ and $\|\tilde{V}_h - V_h\|_\infty \leq \gamma$. Hence, for any $R \in \mathcal{R}, h \in [H], V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h},V_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_hV_{h+1},V_{h+1})\right]\right| \\
\leq &\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\hat{f}_{h,\tilde{R}_h},\tilde{V}_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,\tilde{R}_h}(\mathcal{T}_h\tilde{V}_{h+1},\tilde{V}_{h+1})\right]\right| + 36\sqrt{d}\gamma \\
\leq &\frac{16c_3 d\log(4nH|\Phi||\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}|/\delta)}{n} + 36\sqrt{d}\gamma \leq \frac{68c_3 d^3\log(2n|\Phi|H/\delta)}{n}.
\end{aligned}$$

The last inequality is obtained by choosing $\gamma = \frac{\sqrt{d}}{n}$ and noticing $|\mathcal{C}_{\mathcal{R}_{H-1}^{\text{ELL}},\gamma}| = |\Phi_{H-1}|(2\sqrt{d}/\gamma)^{d^2}$ $\leq |\Phi|(2n)^{d^2}$. This completes the proof. ■

Appendix E. FQE Result

In this section, we provide FQE (Le et al., 2019) analysis. In Appendix E.1, we discuss the FQE algorithm. For simplicity, we use the horizon H in Algorithm 6 and all statements. When FQE is called with a smaller horizon $\tilde{H} \leq H$, we can just replace all H by $\tilde{H}$. The analyses still go through and the sample complexity will not exceed the one instantiated with H. We want to mention that we abuse some notations in this section. For example, $\mathcal{F}_h, \mathcal{V}_h$ may have different meanings from the main text. However, they should be clear within the context.

E.1 FQE Algorithm

In this part, we present FQE algorithm (Algorithm 6), which is similar to Algorithm 5. The difference is that we now approximate the Bellman equation instead of the Bellman optimality equation. More specifically, $\hat{V}$ is defined by evaluating $\hat{f}$ with π rather than maximizing over all actions in $\hat{f}$. In addition, we return the estimated expected return instead of the greedy policy. The details can be found below. When calling Algorithm 6 and there is no confusion, we sometimes drop the input function class (see line 1 in Algorithm 6) for simplicity.

Algorithm 6 FQE: Fitted Q-Evaluation

input: (1) exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ sampled from ρ_{h-3}^{+3} with size n at each level $h \in [H]$, (2) reward function $R = R_{0:H-1}$ with $R_h : \mathcal{X} \times \mathcal{A} \to [0,1], \forall h \in [H]$, (3) function class: $\mathcal{F}_h(R_h) \coloneqq \{R_h + \langle \phi_h, w_h \rangle : \|w_h\|_2 \leq \sqrt{d}, \phi_h \in \Phi_h\}, h \in [H]$, (4) evaluated policy π.
Set $\hat{V}_H(x) = 0$.
for $h = H-1, \ldots, 0$ **do**
  Pick n samples $\left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^n$ from the exploratory dataset $\mathcal{D}_h$.
  Solve least squares problem:

$$\hat{f}_{h,R_h} \leftarrow \operatorname*{argmin}_{f_h \in \mathcal{F}_h(R_h)} \mathcal{L}_{\mathcal{D}_h,R_h}(f_h, \hat{V}_{h+1}),$$

  where $\mathcal{L}_{\mathcal{D}_h,R_h}(f_h, \hat{V}_{h+1}) := \sum_{i=1}^n \left(f_h\left(x_h^{(i)}, a_h^{(i)}\right) - R_h\left(x_h^{(i)}, a_h^{(i)}\right) - \hat{V}_{h+1}\left(x_{h+1}^{(i)}\right)\right)^2$.
  Define $\hat{V}_h(x) = \mathrm{clip}_{[0,1]}\left(\hat{f}_{h,R_h}(x, \pi)\right)$.
end for
return $\hat{v}^\pi = \hat{V}_0(x_0)$.

E.2 FQE Analysis

Lemma 27 (FQE for a reward class) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ (collected from ρ_{h-3}^{+3} and satisfies Eq. (17) for all $h \in [H]$), and we are given a finite deterministic reward class $\mathcal{R} = \mathcal{R}_0 \times \ldots \times \mathcal{R}_{H-1}$ with $\mathcal{R}_{H-1} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ and $\mathcal{R}_h = \{\mathbf{0}\}, \forall h \in [H-1]$. For $\delta \in (0,1)$, any policy π encountered when running Algorithm 4 (see the statement of Lemma 28 for details), and any reward function $R \in \mathcal{R}$, if we set*

$$n \geq \frac{2c_5 H^4 d^3 \kappa K^2 \log^2(Kd)}{\beta^2}\left(\log\left(\frac{c_5 H^4 d^3 \kappa K^2 \log^2(Kd)}{\beta^2}\right) + \log\left(\tfrac{2|\Phi||\mathcal{R}|H}{\delta}\right)\right),$$

where c_5 is the constant in Lemma 28, then with probability at least $1-\delta$, the value $\hat{v}^\pi$ returned by FQE *(Algorithm 6) satisfies*

$$\left|\mathbb{E}_\pi\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \hat{v}^\pi\right| = |v_R^\pi - \hat{v}^\pi| \leq \beta.$$

Proof This proof shares some similarity as the proof of Lemma 19, so we will only show the different steps. For any fixed $R \in \mathcal{R}$, we have

$$\begin{aligned}
|v_R^\pi - \hat{v}^\pi| &= \left|V_0^\pi(x_0) - \hat{f}_{0,R_0}(x_0)\right| = \left|\mathbb{E}_\pi\left[Q_0^\pi(x_0,a_0) - \hat{f}_{0,R_0}(x_0,a_0)\right]\right| \\
&= \left|\mathbb{E}_\pi\left[R_0(x_0,a_0) + \mathbb{E}\left[V_1^\pi(x_1) \mid x_0, a_0\right] - \hat{f}_{0,R_0}(x_0,a_0)\right]\right| \\
&\leq \mathbb{E}_\pi\left[\left|R_0(x_0,a_0) + \mathbb{E}\left[V_1^\pi(x_1) \mid x_0, a_0\right] - \hat{f}_{0,R_0}(x_0,a_0)\right|\right] \\
&\leq \mathbb{E}_\pi\left[\left|\mathbb{E}\left[V_1^\pi(x_1) \mid x_0, a_0\right] - \mathbb{E}\left[\hat{V}_1(x_1) \mid x_0, a_0\right]\right|\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_\pi\left[\left|R_0(x_0,a_0) + \mathbb{E}\left[\hat{V}_1(x_1) \mid x_0, a_0\right] - \hat{f}_{0,R_0}(x_0,a_0)\right|\right] \\
&\le \mathbb{E}_\pi\left[\left|V_1^\pi(x_1) - \hat{V}_1(x_1)\right| + \left|(\mathcal{T}_0\hat{V}_1)(x_0,a_0) - \hat{f}_{0,R_0}(x_0,a_0)\right|\right],
\end{aligned}
$$

where the last inequality is due to Jensen's inequality.

Continuing unrolling to $h = H-1$, we get

$$
\begin{aligned}
|v_R^\pi - \hat{v}^\pi| &\le \sum_{h=0}^{H-1} \mathbb{E}_\pi\left|(\mathcal{T}_h\hat{V}_h)(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right| \\
&\le \sum_{h=0}^{H-1} \sqrt{\kappa K \mathbb{E}_{\rho_{h-3}^{+3}}\left[\left[(\mathcal{T}_h\hat{V}_{h+1})(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right]^2\right]},
\end{aligned}
$$

where the last inequality is due to condition Eq. (17).

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

$$
\begin{aligned}
&\mathbb{E}_{\rho_{h-3}^{+3}}\left[\left((\mathcal{T}_h\hat{V}_{h+1})(x_h,a_h) - \hat{f}_{h,R_h}(x_h,a_h)\right)^2\right] \\
&= \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h}, \hat{V}_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h\hat{V}_{h+1}, \hat{V}_{h+1})\right] \\
&\le \frac{c_5 K d^3 \log^2(Kd)\log\left(n|\Phi||\mathcal{R}|H/\delta\right)}{n}. \qquad \text{(Lemma 28)}
\end{aligned}
$$

Plugging this back into the overall value performance difference, the bound is

$$
|v_R^\pi - \hat{v}^\pi| \le H^2\sqrt{\kappa K}\sqrt{\frac{c_5 K d^3 \log^2(Kd)\log\left(n|\Phi||\mathcal{R}|H/\delta\right)}{n}}.
$$

Setting RHS to be less than β and reorganize, we get

$$
n \ge \frac{c_5 H^4 \kappa K^2 d^3 \log^2(Kd)\log\left(n|\Phi||\mathcal{R}|H/\delta\right)}{\beta^2}.
$$

A sufficient condition for the inequality above is

$$
n \ge \frac{2c_5 H^4 d^3 \kappa K^2 \log^2(Kd)}{\beta^2}\left(\log\left(\frac{c_5 H^4 d^3 \kappa K^2 \log^2(Kd)}{\beta^2}\right) + \log\left(\tfrac{2|\Phi||\mathcal{R}|H}{\delta}\right)\right).
$$

Notice that all the analyses above and Lemma 28 hold for any $R \in \mathcal{R}$, we complete the proof. ■

Lemma 28 (Deviation bound for Lemma 27) *Assume that we have an exploratory dataset $\mathcal{D}_h \coloneqq \left\{\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)\right\}_{i=1}^n$ collected from ρ_{h-3}^{+3}, $h \in [H]$, and a finite deterministic reward class $\mathcal{R} = \mathcal{R}_0 \times \ldots \times \mathcal{R}_{H-1}$ with $\mathcal{R}_{H-1} \subseteq (\mathcal{X}\times\mathcal{A} \to [0,1])$ and $\mathcal{R}_h = \{\mathbf{0}\}, \forall h \in [H-1]$. Then, with probability at least $1-\delta$, $\forall R \in \mathcal{R}, \pi \in \Pi, h \in [H], V_{h+1} \in \mathcal{V}_{h+1}(R_{h+1})$, we have*

$$
\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h}, V_{h+1})\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1}, V_{h+1})\right]\right| \le \frac{c_5 K d^3 \log^2(Kd)\log\left(n|\Phi||\mathcal{R}|H/\delta\right)}{n}.
$$

Here c_5 is some universal constant. The policy class is defined as $\Pi = \Pi_0 \times \dots \Pi_{H-1}$, where $\Pi_h = \{\text{argmax}_a(\langle \phi_h(x_h, a), w_h \rangle : \phi_h \in \Phi_h, \|w_h\|_2 \leq \sqrt{d}\}$ for $h \in [H-1]$ and $\Pi_{H-1} = \{\text{argmax}_a(R_{H-1}(x_{H-1}, a)) : R \in \mathcal{R}^{\text{ELL}}_{H-1}\}$ ($\mathcal{R}^{\text{ELL}}_{H-1}$ is defined in Lemma 36). Additionally, $\mathcal{V}_{h+1}(R_{h+1}) := \{\text{clip}_{[0,1]}\left(f_{h+1,R_{h+1}}(x_{h+1}, \pi_{h+1}(x_{h+1})\right) : f_{h+1,R_{h+1}} \in \mathcal{F}_{h+1}(R_{h+1}),\ \pi_{h+1} \in \Pi_{h+1}\}$ for $h \in [H-1]$, $\mathcal{V}_H = \{\mathbf{0}\}$ are the state-value function classes and $\mathcal{F}_h(R_h) := \{R_h + \langle \phi_h, w_h \rangle : \phi_h \in \Phi_h, \|w_h\|_2 \leq \sqrt{d}\}$ for $h \in [H]$ is the reward dependent Q-value function class.

Remark 29 *From Algorithm 5, we know that the policy class defined in the lemma statement includes all possible policies that can be encountered when running Algorithm 5.*

Proof For the last level $h = H-1$, we know that $\hat{f}_{H-1} = \mathbf{0}$ and $\hat{f}_{H-1,R_{H-1}} = R_{H-1} + \hat{f}_h = R_{H-1}$ because $\hat{V}_H = \mathbf{0}$ and the reward R_{H-1} is known. Thus, for $h = H-1$, we have

$$\left| \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\hat{f}_{h,R_h}, V_{h+1}) \right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h, R_h}(\mathcal{T}_h V_{h+1}, V_{h+1}) \right] \right| = 0.$$

Then we consider the cases that $h \leq H-3$ and $h = H-2$. Notice that any $\pi \in \Pi$ is a greedy policy, thus we have

$$\text{clip}_{[0,1]}\left(f_{h+1,R_{h+1}}(x_{h+1}, \pi_{h+1}(x_{h+1}))\right) = \mathbb{E}_{a_{h+1} \sim \pi_{h+1}}\left[\text{clip}_{[0,1]}\left(f_{h+1,R_{h+1}}(x_{h+1}, a_{h+1})\right)\right],$$

which implies that the $\mathcal{V}$ class has the same format as that in Lemma 32 and Lemma 33.

Now we can invoke these lemmas by setting $L = 1$ and $B = \sqrt{d}$ and get concentration results. The remaining steps follow similar ones starting at Eq. (45) in the proof of Lemma 21. ■

Corollary 30 (FQE for the elliptical reward class) *Assume that we have the exploratory dataset $\{\mathcal{D}\}_{0:H-1}$ (collected from ρ^{+3}_{h-3} and satisfies Eq. (17) for all $h \in [H]$) and consider the elliptical reward class $\mathcal{R} := \{R_{0:H-1} : R_{0:H-2} = \mathbf{0}, R_{H-1} \in \{\phi^\top_{H-1}\Gamma^{-1}\phi_{H-1} : \phi_{H-1} \in \Phi_{H-1}, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$. For $\delta \in (0,1)$, any policy π encountered when running Algorithm 4 (see the statement of Lemma 28 for details), and any reward function $R \in \mathcal{R}$, if we set*

$$n \geq \frac{4c_6 H^4 d^5 \kappa K^2 \log^2(Kd)}{\beta^2}\left(\log\left(\frac{2c_6 H^4 d^5 \kappa K^2 \log^2(Kd)}{\beta^2}\right) + \log\left(\tfrac{2|\Phi|H}{\delta}\right)\right).$$

then with probability at least $1-\delta$, the value $\hat{v}^\pi$ returned by FQE *(Algorithm 6) satisfies*

$$\left| \mathbb{E}_\pi\left[\sum_{h=0}^{H-1} R_h(x_h, a_h)\right] - \hat{v}^\pi \right| = |v^\pi_R - \hat{v}^\pi| \leq \beta.$$

Proof Similar as the proof of Lemma 26, we first instantiate the result Lemma 28 with the finite reward class $\mathcal{C}_{\mathcal{R}^{\text{ELL}}_{H-1},\gamma}$, where $\gamma = \frac{\sqrt{d}}{n}$. Then we bound the difference between any state function and its closest function in the cover. This will give us a version of Lemma 28 with

the elliptical class $\mathcal{R} \coloneqq \{R_{0:H-1} : R_{0:H-2} = \mathbf{0}, R_{H-1} \in \{\phi_{H-1}^\top \Gamma^{-1} \phi_{H-1} : \phi_{H-1} \in \Phi_{H-1}, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$ and

$$\left|\mathbb{E}\Big[\mathcal{L}_{\mathcal{D}_h,R_h}(\hat{f}_{h,R_h},V_{h+1})\Big] - \mathbb{E}[\mathcal{L}_{\mathcal{D}_h,R_h}(\mathcal{T}_h V_{h+1},V_{h+1})]\right| \leq \frac{2c_6 K d^5 \log^2(Kd)\log\left(n|\Phi|H/\delta\right)}{n}. \quad (46)$$

The final result can be obtained by following the proof of Lemma 27, while we use Eq. (46) instead of the result in Lemma 28. ■

Appendix F. Auxiliary Results

In this section, we provide detailed proofs for auxiliary lemmas.

F.1 Proof of Lemma 3

In this part, we provide the proof of Lemma 3 for completeness. This result is widely used throughout the paper.

Lemma 31 (Restatement of Lemma 3) *For a low-rank MDP $\mathcal{M}$ with embedding dimension d, for any function $f : \mathcal{X} \to [0,1]$, we have*

$$\mathbb{E}\left[f(x_{h+1})|x_h,a_h\right] = \left\langle \phi_h^*(x_h,a_h), \theta_f^* \right\rangle$$

where $\theta_f^* \in \mathbb{R}^d$ and we have $\|\theta_f^*\|_2 \leq \sqrt{d}$. A similar linear representation is true for $\mathbb{E}_{a\sim\pi_{h+1}}[f(x_{h+1},a)|x_h,a_h]$ where $f : \mathcal{X}\times\mathcal{A} \to [0,1]$ and a policy $\pi_{h+1} : \mathcal{X} \to \Delta(\mathcal{A})$.

Proof For state-value function f, we have

$$\begin{aligned}
\mathbb{E}\left[f(x_{h+1})|x_h,a_h\right] &= \int f(x_{h+1}) T_h(x_{h+1}|x_h,a_h) d(x_{h+1}) \\
&= \int f(x_{h+1}) \left\langle \phi_h^*(x_h,a_h), \mu_h^*(x_{h+1}) \right\rangle d(x_{h+1}) \\
&= \left\langle \phi_h^*(x_h,a_h), \int f(x_{h+1}) \mu_h^*(x_{h+1}) d(x_{h+1}) \right\rangle = \left\langle \phi_h^*(x_h,a_h), \theta_f^* \right\rangle,
\end{aligned}$$

where $\theta_f^* \coloneqq \int f(x_{h+1})\mu_h^*(x_{h+1})d(x_{h+1})$ is a function of f. Additionally, we obtain $\|\theta_f^*\|_2 \leq \sqrt{d}$ from Definition 1.

For Q-value function f, we similarly have

$$\mathbb{E}_{a\sim\pi_{h+1}}\left[f(x_{h+1},a)|x_h,a_h\right] = \left\langle \phi_h^*(x_h,a_h), \theta_f^* \right\rangle,$$

where $\theta_f^* \coloneqq \iint f(x_{h+1},a_{h+1})\pi(a_{h+1}|x_{h+1})\mu_h^*(x_{h+1})d(x_{h+1})d(a_{h+1})$ and $\|\theta_f^*\|_2 \leq \sqrt{d}$. ■

F.2 Deviation Bounds for Regression with Squared Loss

In this section, we derive a generalization error bound for squared loss for a class $\mathcal{F}$ which subsumes the discriminator classes $\mathcal{F}_h$ and $\mathcal{G}_h$ in the main text. In this section, we prove the bounds for a prespecified $h \in [H]$ and drop h and $h+1$ subscripts for simplicity. When we apply Lemma 32 in other parts of the paper, usually $\left(x^{(i)}, a^{(i)}, x'^{(i)}\right)$ tuples stands for $\left(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}\right)$ tuples, function classes Φ, Φ' refers to Φ_h, Φ_{h+1}, and $\mathcal{R}$ refers to $\mathcal{R}_h$. We abuse the notations of $\mathcal{F}, \mathcal{G}, \Phi, \mathcal{R}, h, w, \Theta, \varepsilon$ and they have different meanings from the main text.

Lemma 32 *For a dataset $\mathcal{D} \coloneqq \left\{\left(x^{(i)}, a^{(i)}, x'^{(i)}\right)\right\}_{i=1}^n \sim \rho$, finite feature classes Φ and Φ', and a finite reward function class $\mathcal{R} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$, we can show that, with probability at least $1-\delta$,*

$$\begin{aligned}
&|\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V) - (\mathcal{L}_\mathcal{D}(\phi, w, V) - \mathcal{L}_\mathcal{D}(\phi^*, \theta_V^*, V))| \\
\leq &\frac{1}{2}\left(\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V)\right) + \frac{c_1 K d \log^2(Kd)(B + L\sqrt{d})^2 \log\left(n|\Phi||\Phi'||\mathcal{R}|/\delta\right)}{n}
\end{aligned}$$

for all $\phi \in \Phi$, $\|w\|_2 \leq B, V \in \mathcal{V} \coloneqq \{\mathbb{E}_{a' \sim \pi'(x')}[f(x', a')] : f \in \mathcal{F}, \pi' \in \Pi'\}$, where $\mathcal{F} \coloneqq \{f(x', a') = \mathrm{clip}_{[0,L]}(R(x',\ a') + \langle \phi'(x', a'), \theta\rangle) : \phi' \in \Phi', \|\theta\|_2 \leq B, R \in \mathcal{R}\}$, $\Pi' = \{\mathrm{argmax}_{a'}(\langle \phi'(x', a'), \theta\rangle), \phi' \in \Phi', \|\theta\|_2 \leq B\}$, and c_1 is some universal constant.

Proof Firstly recall that from Lemma 3, for any $V \in \mathcal{V}$, we have $\mathbb{E}[V(x')|x, a] = \langle \phi^*(x, a), \theta_V^*\rangle$ with $\|\theta_V^*\|_2 \leq L\sqrt{d}$. From the definition, we can decompose the policy class Π' and the value function class $\mathcal{F}$ according to corresponding features and rewards. We have that $\Pi' = \bigcup_{\phi'_{\Pi'} \in \Phi'} \Pi'(\phi'_{\Pi'})$ and $\mathcal{F} = \bigcup_{\phi'_\mathcal{F} \in \Phi', R \in \mathcal{R}} \mathcal{F}(\phi'_\mathcal{F}, R)$, where $\Pi'(\phi'_{\Pi'}) \coloneqq \{\mathrm{argmax}_{a'}(\langle \phi'_{\Pi'}(x', a'), \theta\rangle) : \|\theta\|_2 \leq B\}$ and $\mathcal{F}(\phi'_\mathcal{F}, R) \coloneqq \{f(x', a') = \mathrm{clip}_{[0,L]}(R(x',\ a') + \langle \phi'_\mathcal{F}(x', a'), \theta\rangle) : \|\theta\|_2 \leq B\}$.

Then, we start with fixed $\phi, \phi'_{\Pi'}, \phi'_\mathcal{F}, R$ and show the concentration result related to $\phi, \mathcal{F}(\phi'_\mathcal{F}, R)$ and $\Pi'(\phi'_{\Pi'})$. We also define function classes $\mathcal{G}_1(\phi) = \{\langle \phi, w_1\rangle : \|w_1\|_2 \leq B\}$, and $\mathcal{G}_2 = \{\langle \phi^*, w_2\rangle : \|w_2\|_2 \leq L\sqrt{d}\}$. From the fact of pseudo dimension of linear function class and Natarajan dimension (Daniely et al., 2011, Theorem 21) and notice that the usage of clipping in $\mathcal{F}(\phi'_\mathcal{F}, R)$ will not increase the pseudo dimension, we know that

$$\mathrm{Pdim}(\mathcal{F}(\phi'_\mathcal{F}, R)) \leq d+1, \quad \mathrm{Pdim}(\mathcal{G}_1) \leq d, \quad \mathrm{Pdim}(\mathcal{G}_2) \leq d, \quad \mathrm{Ndim}(\Pi'(\phi'_{\Pi'})) \leq c_1 d \log(d),$$

where $c_1 > 0$ is some universal constant.

Consider the hypothesis class $\mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_\mathcal{F}, R)$

$$\begin{aligned}
\mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_\mathcal{F}, R) \coloneqq \Big\{&(x, a, x') \to (g_1(x, a) - f(x', \pi'))^2 - (g_2(x, a) - f(x', \pi'))^2 : \\
&f \in \mathcal{F}(\phi'_\mathcal{F}, R), g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2, \pi' \in \Pi'(\phi'_{\Pi'})\Big\}. \qquad (47)
\end{aligned}$$

For any $h \in \mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_\mathcal{F}, R)$, we can write it as

$$h(x, a, x') = (g_1(x, a))^2 - (g_2(x, a))^2 - \left(2\sum_{a' \in \mathcal{A}} f(x', a')\mathbf{1}[a' = \pi'(x')]\right)(g_1(x, a) - g_2(x, a)),$$

where $\mathbf{1}[\cdot]$ is the indicator function.

Here $\mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_{\mathcal{F}}, R)$ is a composited class of $\mathcal{F}(\phi'_{\mathcal{F}}, R), \mathcal{G}_1, \mathcal{G}_2, \Pi'(\phi'_{\Pi'})$ and the compositions belong to Lemma 50. Hence we obtain that

$$\mathrm{Pdim}(\mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_{\mathcal{F}}, R)) \le c_2 K d \log^2(Kd) =: d',$$

where c_2 is some universal constant. We briefly discuss how to use the compositions to bound the pseudo dimension of the most complex term $\sum_{a' \in \mathcal{A}} f(x', a') \mathbf{1}[a' = \pi'(x')]$, since other compositions are easy to see. We first need to augment a' to the domain (i.e., use domain (x', a'), whether keeping (x, a) does not matter because it do not depend on (x, a)). Then notice that for any fixed $\tilde{a}' \in \mathcal{A}$, $f(x', \tilde{a}')\mathbf{1}[\tilde{a}' = \pi'(x')]$ is a map from (x', a') to $\mathbb{R}$, we can apply part 3 of Lemma 50. Finally, by part 2 of Lemma 50, we take the summation over $\mathcal{A}$ we get the final bound. The term a' does not show up in the domain of $\mathcal{H}$ as its dependence disappears after we take the summation over $\mathcal{A}$.

For any $h \in \mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_{\mathcal{F}}, R)$, its range is bounded as $|h(\cdot)| \le 4(B + L\sqrt{d})^2$. By Corollary 43, with probability at least $1 - \delta'$, we have the concentration result

$$\begin{aligned}
&\left| \mathbb{E}[h(x, a, x')] - \frac{1}{n} \sum_{i=1}^{n} h\left(x^{(i)}, a^{(i)}, x'^{(i)}\right) \right| \\
&\le \sqrt{\frac{768 d' \mathbb{V}[h(x, a, x')] \log(n/\delta')}{n}} + \frac{6144 d' (B + L\sqrt{d})^2 \log(n/\delta')}{n} \quad \forall h \in \mathcal{H}(\phi, \phi'_{\Pi'}, \phi'_{\mathcal{F}}, R).
\end{aligned}$$

Union bounding over $\phi'_{\Pi'} \in \Phi', \phi'_{\mathcal{F}} \in \Phi', R \in \mathcal{R}$, we get that for any fixed ϕ and any $\|w_1\|_2 \le B, V \in \mathcal{V}$, with probability at least $1 - |\Phi'|^2 |\mathcal{R}| \delta'$ we have

$$\begin{aligned}
&|\mathcal{L}_\rho(\phi, w_1, V) - \mathcal{L}_\rho(\phi^*, \theta^*_V, V) - (\mathcal{L}_\mathcal{D}(\phi, w_1, V) - \mathcal{L}_\mathcal{D}(\phi^*, \theta^*_V, V))| \\
&\le \sqrt{\frac{768 d' \mathbb{V}[(\langle \phi(x, a), w_1 \rangle - V(x'))^2 - (\langle \phi^*(x, a), \theta^*_V \rangle - V(x'))^2] \log(n/\delta')}{n}} \\
&\quad + \frac{6144 d' (B + L\sqrt{d})^2 \log(n/\delta')}{n}. \qquad (48)
\end{aligned}$$

For the variance term, we can bound it as the following

$$\begin{aligned}
&\mathbb{V}[(\langle \phi(x, a), w_1 \rangle - V(x'))^2 - (\langle \phi^*(x, a), \theta^*_V \rangle - V(x'))^2] \\
&\le \mathbb{E}\left[\left((\langle \phi(x, a), w_1 \rangle - V(x'))^2 - (\langle \phi^*(x, a), \theta^*_V \rangle - V(x'))^2 \right)^2 \right] \\
&\le 4(B + L\sqrt{d})^2 \mathbb{E}\left[(\langle \phi(x, a), w_1 \rangle - \langle \phi^*(x, a), \theta^*_V \rangle)^2 \right] \\
&= 4(B + L\sqrt{d})^2 \left(\mathcal{L}_\rho(\phi, w_1, V) - \mathcal{L}_\rho(\phi^*, \theta^*_V, V) \right).
\end{aligned}$$

Plugging this into Eq. (48) and invoking AM-GM inequality gives us that for any fixed ϕ and any $\|w_1\|_2 \le B, V \in \mathcal{V}$, with probability at least $1 - |\Phi'|^2 |\mathcal{R}| \delta'$

$$\begin{aligned}
&|\mathcal{L}_\rho(\phi, w_1, V) - \mathcal{L}_\rho(\phi^*, \theta^*_V, V) - (\mathcal{L}_\mathcal{D}(\phi, w_1, V) - \mathcal{L}_\mathcal{D}(\phi^*, \theta^*_V, V))| \\
&\le \frac{1}{2} \left(\mathcal{L}_\rho(\phi, w_1, V) - \mathcal{L}_\rho(\phi^*, \theta^*_V, V) \right) + \frac{(768 + 6144) d' (B + L\sqrt{d})^2 \log(n/\delta')}{n}.
\end{aligned}$$

The final bound is obtained by union bounding over $\phi \in \Phi$, setting $\delta' = \delta/(|\Phi||\Phi'|^2|\mathcal{R}|)$, and noticing the definition of d'. ■

Lemma 33 *For a dataset $\mathcal{D} \coloneqq \left\{\left(x^{(i)}, a^{(i)}, x'^{(i)}\right)\right\}_{i=1}^n \sim \rho$, finite feature classes Φ and Φ', and a finite reward function class $\mathcal{R} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$, we can show that, with probability at least $1-\delta$,*

$$\begin{aligned}
&|\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V) - (\mathcal{L}_\mathcal{D}(\phi, w, V) - \mathcal{L}_\mathcal{D}(\phi^*, \theta_V^*, V))| \\
\leq &\frac{1}{2}\left(\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V)\right) + \frac{c_2 K d^2 \log^2(Kd)(B + L\sqrt{d})^2 \log\left(n|\Phi||\Phi'||\mathcal{R}|/\delta\right)}{n}
\end{aligned}$$

for all $\phi \in \Phi$, $\|w\|_2 \leq B, V \in \mathcal{V} \coloneqq \{\mathbb{E}_{a' \sim \pi'(x')}[f(x', a')] : f \in \mathcal{F}, \pi' \in \Pi'\}$, where $\mathcal{F} \coloneqq \{f(x', a') = \mathrm{clip}_{[0,L]}(R(x',\ a') + \langle \phi'(x', a'), \theta\rangle) : \phi' \in \Phi', \|\theta\|_2 \leq B, R \in \mathcal{R}\}$, $\Pi' = \{\mathrm{argmax}_{a'}(f^{\text{ELL}}(x', a')) : f^{\text{ELL}} \in \{\phi'^\top \Gamma^{-1} \phi' : \phi' \in \Phi', \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$, and c_2 is some universal constant.

Proof The proof is almost the same as the proof of Lemma 32. The only difference is that we have a different policy class Π'. It can be decomposed as $\Pi' = \bigcup_{\phi'_{\Pi'} \in \Phi'} \Pi'(\phi'_{\Pi'})$, where $\Pi'(\phi'_{\Pi'}) = \{\mathrm{argmax}_{a'}(f^{\text{ELL}}(x', a')) : f^{\text{ELL}} \in \{\phi'^\top_{\Pi'} \Gamma^{-1} \phi'_{\Pi'} : \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$. We can show that $\Pi'(\phi'_{\Pi'})$ can be represented as the greedy policy of a linear class with dimension d^2

$$\Pi'(\phi'_{\Pi'}) \subseteq \{\operatorname*{argmax}_{a'} \langle \varphi'_{\phi'_{\Pi'}}(x', a'), w\rangle : w \in \mathbb{R}^{d^2}\},$$

where $\varphi'_{\phi'_{\Pi'}}$ is a d^2 dimension function and for any $i, j \in [d]$, $\varphi'_{\phi'_{\Pi'}}(x', a')[i + jd] = \phi'_{\Pi'}(x', a')[i]$ $\phi'_{\Pi'}(x', a')[j]$.

Therefore, we have $\mathrm{Ndim}(\Pi'(\phi'_{\Pi'})) \leq c_3 d^2 \log(d^2)$. Following the steps in Lemma 32, it is easy to see that we only need to pay such additional d factor in the final result. ■

Lemma 34 *For a dataset $\mathcal{D} \coloneqq \left\{\left(x^{(i)}, a^{(i)}, x'^{(i)}\right)\right\}_{i=1}^n \sim \rho$, finite feature classes Φ and Φ' and finite reward function class $\mathcal{R} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$, we can show that, with probability at least $1-\delta$*

$$\begin{aligned}
&|\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V) - (\mathcal{L}_\mathcal{D}(\phi, w, V) - \mathcal{L}_\mathcal{D}(\phi^*, \theta_V^*, V))| \\
\leq &\frac{1}{2}\left(\mathcal{L}_\rho(\phi, w, V) - \mathcal{L}_\rho(\phi^*, \theta_V^*, V)\right) + \frac{c_3 d (B + L\sqrt{d})^2 \log(n(B + L\sqrt{d})|\Phi||\Phi'||\mathcal{R}|/\delta)}{n}
\end{aligned}$$

for all $\phi \in \Phi$, $\|w\|_2 \leq B, V \in \mathcal{V} \coloneqq \{\mathrm{clip}_{[0,L]}(\mathbb{E}_{a' \sim \pi'_f(x')}[f(x', a')])\} = \mathbb{E}_{a' \sim \pi'_f(x')}[f(x', a')]$, where $\mathcal{F} \coloneqq \{f(x', a') = \mathrm{clip}_{[0,L]}(R(x',\ a') + \langle \phi'(x', a'), \theta\rangle) : \phi' \in \Phi', \|\theta\|_2 \leq B, R \in \mathcal{R}\}$ and π'_f is a policy determined by f (e.g.., its induced greedy policy or the uniform policy), and c_3 is some universal constant.

Proof This proof mostly follows from Lemma 32. The main difference is that we do not establish the bound of the covering number through the pseudo dimension, while here we directly show it. We first similarly define the hypothesis class $\mathcal{H}$ as

$$\mathcal{H} \coloneqq \left\{(x, a, x') \to (g_1(x, a) - f(x', \pi'_f))^2 - (g_2(x, a) - f(x', \pi'_f))^2 : f \in \mathcal{F}, g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\right\},$$

where $\mathcal{G}_1 = \{\langle \phi, w_1 \rangle : \|w_1\|_2 \leq B, \phi \in \Phi\}$, $\mathcal{G}_2 = \{\langle \phi^*, w_2 \rangle : \|w_2\|_2 \leq L\sqrt{d}\}$, and $\mathcal{F} \coloneqq \{\text{clip}_{[0,L]}(\mathbb{E}_{a' \sim \pi'_f(x')}[f(x', a')]) : f(x', a') = R(x',\, a') + \langle \phi'(x', a'), \theta \rangle, \phi' \in \Phi', \|\theta\|_2 \leq B, R \in \mathcal{R}\}$.

From the standard covering result, we know that there exists an ℓ_2 cover $\overline{\mathcal{W}}_1$ of $\mathcal{W}_1 = \{w_1 : \|w_1\|_2 \leq B\}$ with scale γ of size $\left(\frac{2B}{\gamma}\right)^d$. Similarly we construct an ℓ_2 cover of $\overline{\mathcal{W}}_2 = \{w_2 : \|w_2\|_2 \leq L\sqrt{d}\}$ and an ℓ_2 cover $\overline{\Theta}$ of $\Theta = \{\theta : \|\theta\|_2 \leq B\}$. They both have scale γ and but with different sizes $\left(\frac{2L\sqrt{d}}{\gamma}\right)^d$ and $\left(\frac{2B}{\gamma}\right)^d$ respectively.

Then we define $\overline{\mathcal{G}}_1 = \{\langle \phi, w_1 \rangle : w_1 \in \overline{\mathcal{W}}_1, \phi \in \Phi\}$, $\overline{\mathcal{G}}_2 = \{\langle \phi^*, w_2 \rangle : w_2 \in \overline{\mathcal{W}}_2\}$, and $\overline{\mathcal{F}} \coloneqq \{\text{clip}_{[0,L]}(\mathbb{E}_{a' \sim \pi'_f(x')}[f(x', a']) : f(x', a') = R(x',\, a') + \langle \phi'(x', a'), \theta \rangle, \phi' \in \Phi', \theta \in \overline{\Theta}, R \in \mathcal{R}\}$. Now we define a function class

$$\overline{\mathcal{H}} := \Big\{(x, a, x') \to (g_1(x, a) - f(x', \pi'_f))^2 - (g_2(x, a) - f(x', \pi'_f))^2 : f \in \overline{\mathcal{F}}, g_1 \in \overline{\mathcal{G}}_1, g_2 \in \overline{\mathcal{G}}_2\Big\}.$$

It is easy to see that for any $h \in \mathcal{H}$, there exists $h' \in \overline{\mathcal{H}}$ such that $\|h - h'\|_\infty \leq 16\gamma(B + L\sqrt{d})$ and $|\overline{\mathcal{H}}| = |\mathcal{R}||\Phi||\Phi'|\left(\frac{2B}{\gamma}\right)^{2d}\left(\frac{2L\sqrt{d}}{\gamma}\right)^d$. From the definition of ℓ_1 cover (Definition 37), we can see that $\overline{\mathcal{H}}$ is a $16\gamma(B + L\sqrt{d})$ resolution ℓ_1 cover of $\mathcal{H}$. By setting $\gamma = \frac{\varepsilon}{16(B+L\sqrt{d})}$, we get that for any ε, n,

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, n) \leq |\mathcal{R}||\Phi||\Phi'|\,(2B/\gamma)^{2d}\left(2L\sqrt{d}/\gamma\right)^d \leq |\mathcal{R}||\Phi||\Phi'|\left(32(B + L\sqrt{d})^2/\varepsilon\right)^{3d}. \quad (49)$$

In Corollary 43, instead of bounding the covering number from the pseudo dimension, we now directly substitute Eq. (49) into Eq. (60). Following the remaining steps in Corollary 43, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$ we have

$$\left|\mathbb{E}[h(x, a, x')] - \frac{1}{n}\sum_{i=1}^{n} h\Big(x^{(i)}, a^{(i)}, x'^{(i)}\Big)\right| \leq \sqrt{\frac{c'_3 d \mathbb{V}[h(z)] \log\left(n(B + L\sqrt{d})|R||\Phi||\Phi'|/\delta\right)}{n}} + \frac{2c'_3 d(B + L\sqrt{d})^2 \log\left(n(B + L\sqrt{d})|R||\Phi||\Phi'|/\delta\right)}{n},$$

where c'_3 is some universal constant.

Following remaining steps in Lemma 32 (without the union bound) completes the proof. ■

F.3 Generalized Elliptic Potential Lemma

The following lemma is adapted from Proposition 1 of Carpentier et al. (2020).

Lemma 35 (Generalized elliptic potential lemma) *For any sequence of vectors $\theta^*_1, \theta^*_2, \ldots, \theta^*_T \in \mathbb{R}^{d \times T}$ where $\|\theta^*_i\| \leq L\sqrt{d}$, and any $\lambda \geq L^2 d$, we have*

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

Proof Proposition 1 from Carpentier et al. (2020) shows that for any bounded sequence of vectors $\theta_1^*, \theta_2^*, \ldots, \theta_T^* \in \mathbb{R}^{d\times T}$, we have

$$\sum_{t=1}^{T} \|\Sigma_t^{-1}\theta_t^*\|_2 \leq \sqrt{\frac{dT}{\lambda}}.$$

Now, we have

$$\Sigma_t = \Sigma_{t-1} + \theta_t^*\theta_t^{*\top} \preceq \Sigma_{t-1} + \lambda I_{d\times d} \preceq 2\Sigma_{t-1},$$

where we use the fact that $\|\theta_t^*\theta_t^{*\top}\|_2 \leq L^2 d \leq \lambda$. Using this relation, we can show the property that for any vector $x \in \mathbb{R}^d$, $4x^\top\Sigma_t^{-2}x \geq x^\top\Sigma_{t-1}^{-2}x$.

First noticing the above p.s.d. dominance inequality, we have $I_{d\times d}/2 \preceq \Sigma_t^{-1/2}\Sigma_{t-1}\Sigma_t^{-1/2}$. Therefore, all the eigenvalues of $\Sigma_t^{-1/2}\Sigma_{t-1}\Sigma_t^{-1/2}$ (thus $\Sigma_t^{-1}\Sigma_{t-1}$) are no less than $1/2$. Applying SVD decomposition, we can get all eigenvalues of matrix $\Sigma_{t-1}\Sigma_t^{-2}\Sigma_{t-1} = (\Sigma_t^{-1}\Sigma_{t-1})^\top(\Sigma_t^{-1}\Sigma_{t-1})$ are no less than $1/4$. Then consider any vector $y \in \mathbb{R}^d$, we have $4y^\top\Sigma_{t-1}\Sigma_t^{-2}\Sigma_{t-1}y \geq y^\top y$. Let $x = \Sigma_{y-1}^{-1}y$, we get this property.

Applying the above result, we finally have

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

which completes the proof. ■

F.4 Covering Lemma for the Elliptical Reward Class

In this part, we provide the statistical complexity of the elliptical reward class. The result is used when we analyze the elliptical planner.

Lemma 36 (Covering lemma for the elliptical reward class) *For any $h \in [H]$ and the elliptical reward class $\mathcal{R}_h \coloneqq \{\phi_h^\top\Gamma^{-1}\phi_h : \phi_h \in \Phi_h, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}$, there exists a γ-cover (in ℓ_∞ norm) $\mathcal{C}_{\mathcal{R}_h,\gamma}$ of size $|\Phi_h|(2\sqrt{d}/\gamma)^{d^2}$.*

Moreover, for any $h \in [H]$, there exists a γ-cover (in ℓ_∞ norm) $\mathcal{C}_{\mathcal{R}_h^{\text{ELL}},\gamma}$ of the reward class $\mathcal{R}_h^{\text{ELL}} \coloneqq \{R_{0:h} : R_{0:h-1} = \mathbf{0}, R_h \in \mathcal{R}_h \coloneqq \{\phi_h^\top\Gamma^{-1}\phi_h : \phi_h \in \Phi_h, \Gamma \in \mathbb{R}^{d\times d}, \lambda_{\min}(\Gamma) \geq 1\}\}$ and $|\mathcal{C}_{\mathcal{R}_h^{\text{ELL}},\gamma}| = |\mathcal{C}_{\mathcal{R}_h,\gamma}|$.

Proof Firstly, for any $\Gamma \in \mathbb{R}^{d\times d}$ with $\lambda_{\min}(\Gamma) \geq 1$, applying matrix norm inequality yields $\|\Gamma\|_{\text{F}} \geq \sqrt{d}$. Further, we have

$$\|\Gamma^{-1}\|_{\text{F}} \leq \frac{\|I_{d\times d}\|_{\text{F}}}{\|\Gamma\|_{\text{F}}} \leq \sqrt{d}.$$

Next, consider the matrix class $\bar{A} \coloneqq \{A \in \mathbb{R}^{d\times d} : \|A\|_{\text{F}} \leq \sqrt{d}\}$. From the definition of the Frobenius norm, for any $A \in \bar{A}$ and any (i,j)-th element, we have $|A_{ij}| \leq \sqrt{d}$. Applying the standard covering argument for each of the d^2 elements, there exists a γ-cover of $\bar{A}$,

whose size is upper bounded by $(2\sqrt{d}/\gamma)^{d^2}$. Denote this γ-cover as $\bar{A}_\gamma$. For any $\Gamma \in \mathbb{R}^{d\times d}$ with $\lambda_{\min}(\Gamma) \geq 1$, we can pick some $A \in \bar{A}_\gamma$ so that $\|\Gamma^{-1} - A_\Gamma\|_{\mathrm{F}} \leq \gamma$. Then for any $h \in [H], \phi_h \in \Phi_h, (x,a) \in \mathcal{X}\times\mathcal{A}$, we have

$$|\phi_h(x,a)^\top \Gamma^{-1}\phi_h(x,a) - \phi_h(x,a)^\top A_\Gamma \phi_h(x,a)| \leq \sup_{v:\|v\|_2\leq 1} |v^\top(\Gamma^{-1} - A_\Gamma)v| \leq \|\Gamma^{-1} - A_\Gamma\|_{\mathrm{F}} \leq \gamma.$$

This implies that $\|\phi_h^\top \Gamma^{-1}\phi_h - \phi_h^\top A_\Gamma \phi_h\|_\infty \leq \gamma$ and thus $\mathcal{C}_{\mathcal{R}_h,\gamma} \coloneqq \{\phi_h^\top A \phi_h : \phi_h \in \Phi_h, A \in \bar{A}_\gamma\}$ is a γ-cover of $\mathcal{R}_h$.

Finally, for any $h \in [H]$, from the definition of $\mathcal{R}_h^{\mathrm{ELL}}$ we know that at level $0,\ldots,h-1$ the reward is $\mathbf{0}$ for any reward $R \in \mathcal{R}_h^{\mathrm{ELL}}$. Therefore, $\mathcal{C}_{\mathcal{R}_h^{\mathrm{ELL}},\gamma}$ is directly a γ-cover of reward class $\mathcal{R}_h^{\mathrm{ELL}}$ and $|\mathcal{C}_{\mathcal{R}_h^{\mathrm{ELL}},\gamma}| = |\mathcal{C}_{\mathcal{R}_h,\gamma}|$. This completes the proof. ■

F.5 Probabilistic Tools

In this part, we abuse some notations (e.g., $\mathcal{Z}, \mathcal{G}, \Pi, n, \varepsilon, f, g, h, z$) and they have the different meaning from other parts of the paper.

Definition 37 (ℓ_1 covering number) *Given a hypothesis class $\mathcal{H} \subseteq (\mathcal{Z} \to \mathbb{R})$, $\varepsilon > 0$, and $Z^n = (z_1,\ldots,z_n) \in \mathcal{Z}^n$. We define the ℓ_1 covering number $\mathcal{N}_1(\varepsilon, \mathcal{H}, Z^n)$ as the minimal cardinality of a set $\mathcal{C} \subseteq \mathcal{H}$, such that for any $h \in \mathcal{H}$, there exists $h' \in \mathcal{H}$ such that $\frac{1}{n}\sum_{i=1}^n |h(z_i) - h'(z_i)| \leq \varepsilon$. We also define $\mathcal{N}_1(\varepsilon, \mathcal{H}, n) \coloneqq \max_{Z^n \in \mathcal{Z}^n} \mathcal{N}_1(\varepsilon, \mathcal{H}, Z^n)$.*

Lemma 38 (Uniform deviation bound using covering number (Hoeffding's version), Theorem 29.1 of Devroye et al. (2013)) *Let $\mathcal{H} \subseteq (\mathcal{Z} \to [0,b])$ be a hypothesis class and $Z^n = (z_1,\ldots,z_n) \in \mathcal{Z}^n$, where z_i are i.i.d. samples drawn from some distribution $\mathbb{P}(z)$ supported on $\mathcal{Z}$. Then for any n and $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}[h(z)]\right| > \varepsilon\right] \leq 8\mathcal{N}_1(\varepsilon/8, \mathcal{H}, n)\exp\left(-\frac{n\varepsilon^2}{128b^2}\right).$$

Lemma 39 (An extension of the classical Bernstein's inequality, Lemma F.2 of Dong et al. (2020); Lemma 3.1 of Massart (1986)) *For any $N \geq n \geq 1$, let w be a uniformly random permutation over $1,\ldots,N$. For any $\xi \in \mathbb{R}^N$, we define*

$$\hat{S}_N = \sum_{i=1}^N \xi_i, \quad \hat{S}_{w,n} = \sum_{i=1}^n \xi_{w(i)}, \quad \hat{\sigma}_N^2 = \left(\frac{1}{N}\sum_{i=1}^N \xi_i^2\right) - \left(\frac{1}{N}\sum_{i=1}^N \xi_i\right)^2,$$

and $\hat{U}_N = \max_{1\leq i\leq N}\xi_i - \min_{1\leq i\leq N}\xi_i$. Then for any $\varepsilon > 0$, we have

$$\mathbb{P}\left[\left|\frac{\hat{S}_{w,n}}{n} - \frac{\hat{S}_N}{N}\right| > \varepsilon\right] \leq 2\exp\left(-\frac{n\varepsilon^2}{2\hat{\sigma}_N^2 + \varepsilon\hat{U}_N}\right).$$

Lemma 40 (Uniform deviation bound using covering number (Bernstein's version), adapted from Lemma F.3 of Dong et al. (2020)) *Let $\mathcal{H} \subseteq (\mathcal{Z} \to [0,b])$ be a*

hypothesis class and $Z^n = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, where z_i are i.i.d. samples drawn from some distribution $\mathbb{P}(z)$ supported on $\mathcal{Z}$. Then for any $h \in \mathcal{H}$, we have

$$\mathbb{P}\left[\left|\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^{n} h(z_i)\right| > \varepsilon\right] \leq \inf_{N \geq 2n + 8b^2/\varepsilon^2} \left(32\mathcal{N}_1(\varepsilon/32, \mathcal{H}, N)\exp\left(-\frac{N\varepsilon^2}{2048b^2}\right) + 4\mathcal{N}_1\left(\frac{\varepsilon n}{16N}, \mathcal{H}, N\right)\exp\left(-\frac{n\varepsilon^2}{128\mathbb{V}[h(z)] + 256\varepsilon b}\right)\right).$$

Remark 41 *The major difference Lemma F.3 of Dong et al. (2020) is that we have $\mathbb{V}[h(z)]$ instead of its uniform upper bound on RHS.*

Proof The proof mostly follows from Lemma F.3 of Dong et al. (2020). Firstly, we define similar notations for the empirical sums. Let $n' = N - n$ and $Z^N = (z_1, \cdots, z_N)$ be N i.i.d. random samples. For any $h \in \mathcal{H}$, we define the following:

$$\hat{\mathbb{E}}_n[h(z)] := \frac{1}{n}\sum_{i=1}^{n} h(z_i), \quad \hat{\mathbb{E}}_{n'}[h(z)] := \frac{1}{n'}\sum_{i=n+1}^{N} h(z_i), \quad \hat{\mathbb{E}}_N[h(z)] := \frac{1}{N}\sum_{i=1}^{N} h(z_i).$$

In addition, let w be a random permutation over $1, \ldots, N$, which is independent of the choice of $(z_1, \cdots, z_N)$. The empirical sums of the permutation are defined as

$$\hat{\mathbb{E}}_{w,n}[h(z)] := \frac{1}{n}\sum_{i=1}^{n} h(z_{w(i)}), \quad \hat{\mathbb{E}}_{w,n'}[h(z)] := \frac{1}{n'}\sum_{i=n+1}^{N} h(z_{w(i)}).$$

Following the first two steps of the proof in Dong et al. (2020), for $N \geq 2n + 8b^2/\varepsilon^2$ and any $h \in \mathcal{H}$, we can easily get that

$$\begin{aligned}\mathbb{P}\left[\left|\hat{\mathbb{E}}_n[h(z)] - \mathbb{E}[h(z)]\right| > \varepsilon\right] &\leq 2\mathbb{P}\left[\left|\hat{\mathbb{E}}_n[h(z)] - \hat{\mathbb{E}}_{n'}[h(z)]\right| > \varepsilon/2\right] \\ &\leq 2\mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h(z)] - \hat{\mathbb{E}}_N[h(z)]\right| > \varepsilon/4\right]. \end{aligned} \tag{50}$$

The main difference in our proof is that we move $\sup_{h\in\mathcal{H}}$ from inside $\mathbb{P}[\cdot]$ to the outside (and change it to the argument "for any $h \in \mathcal{H}$"), and it can be easily verified.

For any fixed $Z^N = (z_1, \cdots, z_N) \in \mathcal{Z}^N$, we use $\mathcal{C} = \{h'_1, \cdots, h'_{|\mathcal{C}|}\}$ to denote the minimal $(n\varepsilon/16N)$-cover over $\mathcal{H}|_{Z^N}$. Then we have $|\mathcal{C}| \leq \mathcal{N}_1(n\varepsilon/16N, \mathcal{H}, N)$, and there exists a mapping $\alpha : \mathcal{H} \to \{1, \ldots, |\mathcal{C}|\}$ such that,

$$\frac{1}{N}\sum_{i=1}^{N}\left|h(z_i) - h'_{\alpha(h)}(z_i)\right| \leq n\varepsilon/16N, \quad \forall h \in \mathcal{H}. \tag{51}$$

Following the analysis in Dong et al. (2020), we have that for any $h \in \mathcal{H}$,

$$\left|\hat{\mathbb{E}}_{w,n}[h(z)] - \hat{\mathbb{E}}_N[h(z)]\right| \leq \left|\hat{\mathbb{E}}_{w,n}[h'_{\alpha(h)}(z)] - \hat{\mathbb{E}}_N[h'_{\alpha(h)}(z)]\right| + \varepsilon/8.$$

This implies that for fixed Z^N and any $h \in \mathcal{H}$, we have

$$\mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h(z)] - \hat{\mathbb{E}}_N[h(z)]\right| > \varepsilon/4 \,|\, Z^N\right] \leq \mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h'_{\alpha(h)}(z)] - \hat{\mathbb{E}}_N[h'_{\alpha(h)}(z)]\right| > \varepsilon/8 \,|\, Z^N\right]. \tag{52}$$

Now we define the empirical variance as

$$\hat{\mathbb{V}}_N[h(z)] := \frac{1}{N}\sum_{i=1}^N h(z_i)^2 - \left(\frac{1}{N}\sum_{i=1}^N h(z_i)\right)^2.$$

Applying Lemma 39 and union bounding over $\mathcal{C}$ yields that for any $h_i' \in \mathcal{C}, i \in \{1, 2, \ldots, |\mathcal{C}|\}$,

$$\mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h_i'(z)] - \hat{\mathbb{E}}_N[h_i'(z)]\right| > \varepsilon/8 \mid Z^N\right] \le 2|\mathcal{C}| \exp\left(-\frac{n\varepsilon^2/64}{2\hat{\mathbb{V}}_N[h_i'(z)] + \varepsilon b/8}\right). \tag{53}$$

Note that for h and $h'_{\alpha(h)}$, we have

$$\begin{aligned}
&\left|\hat{\mathbb{V}}_N[h(z)] - \hat{\mathbb{V}}_N[h'_{\alpha(h)}(z)]\right| \\
&= \frac{1}{N}\sum_{i=1}^N \left(\left(h(z_i) - h'_{\alpha(h)}(z_i)\right)\left(h(z_i) + h'_{\alpha(h)}(z_i)\right)\right) \\
&\quad - \left(\frac{1}{N}\sum_{i=1}^N h(z_i) - \frac{1}{N}\sum_{i=1}^N h'_{\alpha(h)}(z_i)\right)\left(\frac{1}{N}\sum_{i=1}^N h(z_i) + \frac{1}{N}\sum_{i=1}^N h'_{\alpha(h)}(z_i)\right) \\
&\le 2b\frac{1}{N}\sum_{i=1}^N \left|h(z_i) - h'_{\alpha(h)}(z_i)\right| + 2b\frac{1}{N}\sum_{i=1}^N \left|h(z_i) - h'_{\alpha(h)}(z_i)\right|.
\end{aligned} \tag{54}$$

Combining Eqs. (51) to (54), for any fixed Z^N and any $h \in \mathcal{H}$, we have

$$\mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h(z)] - \hat{\mathbb{E}}_N[h(z)]\right| > \varepsilon/4 \mid Z^N\right] \le 2|\mathcal{C}| \exp\left(-\frac{n\varepsilon^2/64}{2\hat{\mathbb{V}}_N[h(z)] + \varepsilon b + n\varepsilon b/(2N)}\right). \tag{55}$$

For the empirical variance and population variance, we note that

$$\begin{aligned}
&\left|\hat{\mathbb{V}}_N[h(z)] - \mathbb{V}[h(z)]\right| \\
&= \left|\left(\frac{1}{N}\sum_{i=1}^N h(z_i)^2 - \mathbb{E}[h(z)^2]\right) - \left(\left(\frac{1}{N}\sum_{i=1}^N h(z_i)\right)^2 - (\mathbb{E}[h(z)])^2\right)\right| \\
&\le \left|\frac{1}{N}\sum_{i=1}^N h(z_i)^2 - \mathbb{E}[h(z)^2]\right| + \left|\left(\frac{1}{N}\sum_{i=1}^N h(z_i) - \mathbb{E}[h(z)]\right)\left(\frac{1}{N}\sum_{i=1}^N h(z_i) + \mathbb{E}[h(z)]\right)\right| \\
&\le \left|\frac{1}{N}\sum_{i=1}^N h(z_i)^2 - \mathbb{E}[h(z)^2]\right| + 2b\left|\frac{1}{N}\sum_{i=1}^N h(z_i) - \mathbb{E}[h(z)]\right|.
\end{aligned} \tag{56}$$

Consequently,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\hat{\mathbb{V}}_N[h(z)] - \mathbb{V}[h(z)]\right| > \varepsilon b\right]$$

$$\leq \mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^{N}h(z_i)^2-\mathbb{E}[h(z)^2]\right|>\varepsilon b/2\right]+\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^{N}h(z_i)-\mathbb{E}[h(z)]\right|>\varepsilon/4\right]$$
$$\leq 8\mathcal{N}_1\left(\varepsilon b/16,\mathcal{H}^2,N\right)\exp\left(-\frac{N\varepsilon^2}{512b^2}\right)+8\mathcal{N}_1\left(\varepsilon/32,\mathcal{H},N\right)\exp\left(-\frac{N\varepsilon^2}{2048b^2}\right)$$
$$\leq 16\mathcal{N}_1\left(\varepsilon/32,\mathcal{H},N\right)\exp\left(-\frac{N\varepsilon^2}{2048b^2}\right). \tag{57}$$

Here, we define $\mathcal{H}^2=\{h^2:h\in\mathcal{H}\}$. The first inequality is due to Eq. (56). The second inequality is due to Lemma 38. Notice that for any $h_1,h_2\in\mathcal{H},z\in\mathcal{Z}$, we have $|h_1(z)^2-h_2(z)^2|=|(h_1(z)-h_2(z))(h_1(z)+h_2(z))|\leq 2b\,|h_1(z)-h_2(z)|$. This implies that $\mathcal{N}_1\left(\varepsilon b/16,\mathcal{H}^2,N\right)\leq\mathcal{N}_1\left(\varepsilon/32,\mathcal{H},N\right)$. Then it is easy to see the third inequality holds.

Combining Eqs. (55) and (57), for $N\geq 2n+8b^2/\varepsilon^2$, any fixed Z^N and any $h\in\mathcal{H}$, we get that

$$\mathbb{P}\left[\left|\hat{\mathbb{E}}_{w,n}[h(z)]-\hat{\mathbb{E}}_N[h(z)]\right|>\varepsilon/4\right]$$
$$\leq 2|\mathcal{C}|\exp\left(-\frac{n\varepsilon^2/64}{2\mathbb{V}[h(z)]+4\varepsilon b}\right)+16\mathcal{N}_1\left(\varepsilon/32,\mathcal{H},N\right)\exp\left(-\frac{N\varepsilon^2}{2048b^2}\right). \tag{58}$$

Finally, combining Eqs. (50) and (58) and noticing the range of N and the upper bound of $|\mathcal{C}|$ completes the proof. ■

Corollary 42 (Uniform deviation bound using covering number (Bernstein's version, tail bound), adapted from Lemma F.4 of Dong et al. (2020)) *For $b\geq 1$, let $\mathcal{H}\subseteq(\mathcal{Z}\to[0,b])$ be a hypothesis class and $Z^n=(z_1,\ldots,z_n)\in\mathcal{Z}^n$, where z_i are i.i.d. samples drawn from some distribution $\mathbb{P}(z)$ supported on $\mathcal{Z}$. Then for any $h\in\mathcal{H}$, we have*

$$\mathbb{P}\left[\left|\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_i)\right|>\varepsilon\right]\leq 36\mathcal{N}_1\left(\frac{\varepsilon^3}{160b^2},\mathcal{H},\frac{10nb^2}{\varepsilon^2}\right)\exp\left(-\frac{n\varepsilon^2}{128\mathbb{V}[h(z)]+256\varepsilon b}\right).$$

Proof In Lemma 40, we can set $N=10nb^2/\varepsilon^2\geq 2n+8b^2/\varepsilon^2$, which indicates $\frac{N\varepsilon^2}{2048b^2}\geq\frac{n\varepsilon^2}{256\varepsilon b}\geq\frac{n\varepsilon^2}{128\mathbb{V}[h(z)]+256\varepsilon b}$ and $\varepsilon/32\geq\varepsilon n/(16N)$. Then noticing the monotonicity of covering number $\mathcal{N}_1(\cdot,\mathcal{H},N)$ and $\exp(\cdot)$, we complete the proof. ■

Corollary 43 (Uniform deviation bound using covering number (Bernstein's version, confidence interval bound)) *For $b\geq 1$, let $\mathcal{H}\subseteq(\mathcal{Z}\to[-b,b])$ be a hypothesis class with $\mathrm{Pdim}(\mathcal{H})\leq d_{\mathcal{H}}$ and $Z^n=(z_1,\ldots,z_n)$ be i.i.d. samples drawn from some distribution $\mathbb{P}(z)$ supported on $\mathcal{Z}$. Then with probability at least $1-\delta$, we have that for any $h\in\mathcal{H}$,*

$$\left|\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_i)\right|\leq\sqrt{\frac{384d_{\mathcal{H}}\mathbb{V}[h(z)]\log(n/\delta)}{n}}+\frac{768d_{\mathcal{H}}b\log(n/\delta)}{n}.$$

Remark 44 *The slight difference from the standard Bernstein's inequality is that we have $\log(n)$ term on the numerator. It is mostly due to the ε dependence in the covering number.*

Proof Let $\mathcal{H}' = \{(h(\cdot)+b : h \in \mathcal{H}\} \subseteq (\mathcal{Z} \to [0, 2b])$. We know that shifting only changes the range of the function, but does not change the variance of the function. Therefore, applying Corollary 42 gives us that for any $h' \in \mathcal{H}'$,

$$\mathbb{P}\left[\left|\mathbb{E}[h'(z)] - \frac{1}{n}\sum_{i=1}^{n} h'(z_i)\right| > \varepsilon\right] \leq 36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}', \frac{40nb^2}{\varepsilon^2}\right)\exp\left(-\frac{n\varepsilon^2}{128\mathbb{V}[h(z)] + 512\varepsilon b}\right).$$

From Definition 37, we know that $\mathcal{H}$ and $\mathcal{H}'$ have the same covering number, i.e., $\mathcal{N}_1(\varepsilon', \mathcal{H}', m) = \mathcal{N}_1(\varepsilon', \mathcal{H}, m)$ for any $\varepsilon' \in \mathbb{R}_+$ and $m \in \mathbb{N}_+$. In addition, we have for any $h' \in \mathcal{H}'$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} h'(z_i) - \mathbb{E}[h'(z)]\right| = \left|\frac{1}{n}\sum_{i=1}^{n} h(z_i) + b - \mathbb{E}[h(z) + b]\right| = \left|\frac{1}{n}\sum_{i=1}^{n} h(z_i) - \mathbb{E}[h(z)]\right|.$$

This implies that for any $h \in \mathcal{H}$,

$$\mathbb{P}\left[\left|\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^{n} h(z_i)\right| > \varepsilon\right] \leq 36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{40nb^2}{\varepsilon^2}\right)\exp\left(\frac{-n\varepsilon^2}{128\mathbb{V}[h(z)] + 512\varepsilon b}\right). \quad (59)$$

Setting RHS of Eq. (59) to be δ, we get

$$n = \frac{(128\mathbb{V}[h(z)] + 512\varepsilon b)\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{40nb^2}{\varepsilon^2}\right)/\delta\right)}{\varepsilon^2}. \quad (60)$$

This implies the following inequality

$$\varepsilon \leq \sqrt{\frac{128\mathbb{V}[h(z)]\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{40nb^2}{\varepsilon^2}\right)/\delta\right)}{n}} + \frac{512b\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{40nb^2}{\varepsilon^2}\right)/\delta\right)}{n},$$

which can be verified by substituting n in Eq. (60).

Applying Corollary 46 and simplifying the expression yields

$$\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{72nb^2}{\varepsilon^2}\right)/\delta\right) \leq 3d_{\mathcal{H}}\log\left(\frac{54b}{\varepsilon\delta}\right).$$

From Eq. (60), we also have $n \geq \frac{512b}{\varepsilon}$. Consequently, we have

$$\varepsilon \leq \sqrt{\frac{384d_{\mathcal{H}}\mathbb{V}[h(z)]\log(n/\delta)}{n}} + \frac{768d_{\mathcal{H}}b\log(n/\delta)}{n}.$$

Plugging this into Eq. (59) completes the proof. ■

Lemma 45 (Bounding covering number by pseudo dimension (Haussler, 1995)) *Given a hypothesis class* $\mathcal{H} \subseteq (\mathcal{Z} \to [0, 1])$ *with* $\mathrm{Pdim}(\mathcal{H}) \leq d_{\mathcal{H}}$*, we have for any* $Z^n \in \mathcal{Z}^n$*,*

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, Z^n) \leq e\left(d_{\mathcal{H}} + 1\right)\left(\frac{2e}{\varepsilon}\right)^{d_{\mathcal{H}}}.$$

Corollary 46 (Bounding covering number by pseudo dimension) *Given a hypothesis class* $\mathcal{H} \subseteq (\mathcal{Z} \to [a,b])$ *with* $\mathrm{Pdim}(\mathcal{H}) \leq d_{\mathcal{H}}$*, for any* $Z^n \in \mathcal{Z}^n$*, we have*

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, Z^n) \leq e\,(d_{\mathcal{H}}+1)\left(\frac{2e(b-a)}{\varepsilon}\right)^{d_{\mathcal{H}}} \quad \text{and} \quad \mathcal{N}_1(\varepsilon, \mathcal{H}, n) \leq \left(\frac{4e^2(b-a)}{\varepsilon}\right)^{d_{\mathcal{H}}}.$$

Proof Let $\mathcal{H}' = \{(h(\cdot)-a)/(b-a) : h \in \mathcal{H}\} \subseteq (\mathcal{Z} \to [0,1])$. From the definition of pseudo dimension, it is easy to see $d_{\mathcal{H}'} = d_{\mathcal{H}}$. Noticing Definition 37 and applying Lemma 45, we get

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, Z^n) = \mathcal{N}_1(\varepsilon/(b-a), \mathcal{H}', Z^n) = (d_{\mathcal{H}}+1)\left(\frac{2e(b-a)}{\varepsilon}\right)^{d_{\mathcal{H}}}.$$

Noticing Definition 37 and following simple algebra, we can show the second part. ■

Definition 47 (VC-dimension) *For hypothesis class* $\mathcal{H} \subseteq (\mathcal{X} \to \{0,1\})$*, we define its VC-dimension VC-dim*$(\mathcal{H})$ *as the maximal cardinality of a set* $X = \{x_1, \ldots, x_{|X|}\} \subseteq \mathcal{X}$ *that satisfies* $|\mathcal{H}_X| = 2^{|X|}$ *(or* X *is* shattered *by* $\mathcal{H}$*), where* $\mathcal{H}_X$ *is the restriction of* $\mathcal{H}$ *to* X*, i.e.,* $\{(h(x_1), \ldots, h(x_{|X|})) : h \in \mathcal{H}\}$.

Definition 48 (Pseudo dimension (Haussler, 2018)) *For hypothesis class* $\mathcal{H} \subseteq (\mathcal{X} \to \mathbb{R})$*, we define its pseudo dimension* $\mathrm{Pdim}(\mathcal{H})$ *as* $\mathrm{Pdim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H}^+)$*, where* $\mathcal{H}^+ = \{(x,\xi) \mapsto \mathbf{1}[h(x) > \xi] : h \in \mathcal{H}\} \subseteq (\mathcal{X} \times \mathbb{R} \to \{0,1\})$.

Lemma 49 (Sauer's lemma) *For the hypothesis class* $\mathcal{H} \subseteq (\mathcal{X} \to \{0,1\})$ *with* $\mathrm{VCdim}(\mathcal{H}) = d_{\mathrm{VC}} < \infty$ *and any* $X = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$*, we have*

$$|\mathcal{H}_X| \leq (n+1)^{d_{\mathrm{VC}}}, \tag{61}$$

where $\mathcal{H}_X \coloneqq \{(h(x_1), h(x_2), \ldots, h(x_n)) : h \in \mathcal{H}\}$ *is the restriction of* $\mathcal{H}$ *to* X.

Lemma 50 *Let* $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{A}$ *with* $|\mathcal{A}| = K$*. Let* $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$ *be a policy class with Natarajan dimension* $\mathrm{Ndim}(\Pi) = d_\Pi \geq 6$*,* $\mathcal{F} \subseteq (\mathcal{Z} \to [0,L])$ *with pseudo dimension* $\mathrm{Pdim}(\mathcal{F}) = d_{\mathcal{F}} \geq 6$*, and* $\mathcal{G}_1, \mathcal{G}_2 \subseteq (\mathcal{X} \to [0,L])$ *with pseudo dimension* $\mathrm{Pdim}(\mathcal{G}_1) = d_{\mathcal{G}_1} \geq 6$ *and* $\mathrm{Pdim}(\mathcal{G}_2) = d_{\mathcal{G}_2} \geq 6$*. Then we have the following:*

1. *The hypothesis class* $\mathcal{H}_1 = \{x \to g_1(x)g_2(x) : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$ *has pseudo dimension* $\mathrm{Pdim}(\mathcal{H}_1) \leq 32(d_{\mathcal{G}_1}\log(d_{\mathcal{G}_1}) + d_{\mathcal{G}_2}\log(d_{\mathcal{G}_2}))$.
2. *The hypothesis class* $\mathcal{H}_2 = \{x \to g_1(x) + g_2(x) : g_1, \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$ *has pseudo dimension* $\mathrm{Pdim}(\mathcal{H}_2) \leq 32(d_{\mathcal{G}_1}\log(d_{\mathcal{G}_1}) + d_{\mathcal{G}_2}\log(d_{\mathcal{G}_2}))$.
3. *The hypothesis class* $\mathcal{H}_3 = \{(x,a) \to f(x,a)\mathbf{1}[a = \pi(x)] : f \in \mathcal{F}, \pi \in \Pi\}$ *has pseudo dimension* $\mathrm{Pdim}(\mathcal{H}_3) \leq 6(d_\Pi + d_{\mathcal{F}})\log(2eK(d_\Pi + d_{\mathcal{F}}))$.

Proof Firstly, w.l.o.g. we assume that $L = 1$ since in the pseudo dimension we can just scale all ξ in Definition 48 by $1/L$.

Part 1. Let $\mathcal{H}_1^+ \coloneqq \{(x,\zeta) \to \mathbf{1}[g_1(x)g_2(x) > \zeta] : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\} \subseteq (\mathcal{X} \times \mathbb{R} \to \{0,1\})$. From the fact that $\text{Pdim}(\mathcal{H}) = \text{VCdim}(\mathcal{H}_1^+)$, it suffices to prove that $\log|\mathcal{H}_{1,X}^+| \le n$ for any $X = ((x_1,\zeta_1),(x_2,\zeta_2),\dots,(x_n,\zeta_n)) \in (\mathcal{X}\times\mathbb{R})^n$, where $n = 32(d_{\mathcal{G}_1}\log(d_{\mathcal{G}_1}) + d_{\mathcal{G}_2}\log(d_{\mathcal{G}_2}))$ and $\mathcal{H}_{1,X}^+$ refers to the restriction of $\mathcal{H}_1^+$ to X.

Similarly we define $\mathcal{G}_1^+ \coloneqq \{(x,\xi) \to \mathbf{1}[g_1(x) > \xi] : g_1 \in \mathcal{G}_1\} \subseteq (\mathcal{X}\times\mathbb{R} \to \{0,1\})$ and $\mathcal{G}_2^+ \coloneqq \{(x,\xi') \to \mathbf{1}[g_2(x) > \xi'] : g_2 \in \mathcal{G}_2\} \subseteq (\mathcal{X}\times\mathbb{R} \to \{0,1\})$. For $X' = ((x_1,\xi_1),(x_2,\xi_2),\dots,(x_n,\xi_n)) \in (\mathcal{X}\times\mathbb{R})^n, X'' = ((x_1,\xi_1'),(x_2,\xi_2'),\dots,(x_n,\xi_n')) \in (\mathcal{X}\times\mathbb{R})^n$, we use $\mathcal{G}_{1,X'}^+$ to denote the restriction of $\mathcal{G}_1^+$ to X' and use $\mathcal{G}_{2,X''}^+$ to denote the restriction of $\mathcal{G}_2^+$ to X''. Since $\mathbf{1}[g_1(x)g_2(x) > \zeta]$ can be decomposed as $\mathbf{1}[g_1(x) > \xi]\mathbf{1}[g_2(x) > \zeta/\xi]$ for some $\xi \in \mathbb{R}$, by setting $\xi' = \zeta/\xi$ we can see that $\left|\mathcal{H}_{1,X}^+\right| \le \left|\mathcal{G}_{1,X'}^+\right|\left|\mathcal{G}_{2,X''}^+\right|$.

Applying Lemma 49, we have $\left|\mathcal{G}_{1,X'}^+\right| \le (n+1)^{d_{\mathcal{G}_1}}$, $\left|\mathcal{G}_{2,X''}^+\right| \le (n+1)^{d_{\mathcal{G}_2}}$. Therefore,

$$\begin{aligned}
\log\left|\mathcal{H}_{1,X}^+\right| &\le d_{\mathcal{G}_1}\log(n+1) + d_{\mathcal{G}_2}\log(n+1) \\
&\le (d_{\mathcal{G}_1} + d_{\mathcal{G}_2})\log(32(d_{\mathcal{G}_1}\log(d_{\mathcal{G}_1}) + d_{\mathcal{G}_2}\log(d_{\mathcal{G}_2})) + 1) \\
&\le (d_{\mathcal{G}_1} + d_{\mathcal{G}_2})\log(64(d_{\mathcal{G}_1}^2 + d_{\mathcal{G}_2}^2)) \le 2(d_{\mathcal{G}_1} + d_{\mathcal{G}_2})\log(64(d_{\mathcal{G}_1} + d_{\mathcal{G}_2})) \\
&\le 32(d_{\mathcal{G}_1}\log(d_{\mathcal{G}_1}) + d_{\mathcal{G}_2}\log(d_{\mathcal{G}_2})) = n.
\end{aligned}$$

Part 2. Note that we have the decomposition $\mathbf{1}[g_1(x) + g_2(x) > \zeta] = \mathbf{1}[g_1(x) > \xi] + \mathbf{1}[g_2(x) > \zeta - \xi]$. The remaining steps can be similarly followed from Part 1.

Part 3. We use $\mathcal{H}_{3,X}^+$ to denote the restriction of $\mathcal{H}_3^+ \coloneqq \{(x,a,\zeta) \to \mathbf{1}[f(x,a)\mathbf{1}[a = \pi(x)]] \ge \zeta : f \in \mathcal{F}, \pi \in \Pi\}$ to $X = ((x_1,a_1,\zeta_1),(x_2,a_2,\zeta_2),\dots,(x_n,a_n,\zeta_n))$. It suffices to show an upper bound of $|\mathcal{H}_{3,X}^+|$.

For $h \in \mathcal{H}_3^+$, we can decompose it as $\mathbf{1}[a = \pi(x)]\mathbf{1}[f(x,a) \ge \zeta]$. Following the proof of Lemma 21 in Jiang et al. (2017), we can also get $|\mathcal{H}_{3,X}^+| \le |\Pi_X||\mathcal{F}_X^+|$, where Π_X denotes the restriction of Π to $(x_1,x_2,\dots,x_n)$ and $\mathcal{F}_X^+$ denotes the restriction of $\mathcal{F}^+$ to $((x_1,a_1,\zeta_1),(x_2,a_2,\zeta_2),\dots,(x_n,a_n,\zeta_n))$. Notice that here $\mathcal{H}_{3,X}^+$ can not be produced by the Cartesian product of Π_X and $\mathcal{F}_X^+$, but the upper bound still holds. The remaining steps similarly follow from Jiang et al. (2017). ■

References

A. Agarwal, M. Henaff, S. Kakade, and W. Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, 2020a.

A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, 2020b.

A. Antos, C. Szepesvári, and R. Munos. Fitted q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, 2007.

A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020.

L. C. Baird III. Residual algorithms: reinforcement learning with function approximation. In *International Conference on Machine Learning*, 1995.

M. Bellemare, W. Dabney, R. Dadashi, A. Ali Taiga, P. S. Castro, N. Le Roux, D. Schuurmans, T. Lattimore, and C. Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

A. Carpentier, C. Vernade, and Y. Abbasi-Yadkori. The elliptical potential lemma revisited. *arxiv:2010.10182*, 2020.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

J. Chen, A. Modi, A. Krishnamurthy, N. Jiang, and A. Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. In *Advances in Neural Information Processing Systems*, 2022.

B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, 2018.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the erm principle. In *Conference on Learning Theory*, 2011.

C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*, 2018.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

K. Dong, J. Peng, Y. Wang, and Y. Zhou. $\sqrt{n}$-regret for learning in Markov decision processes with function approximation and low Bellman rank. In *Conference on Learning Theory*, 2020.

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

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

S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019b.

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

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

A.-m. Farahmand, A. Barreto, and D. Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, 2017.

D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Advances in Neural Information Processing Systems*, 2020.

C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

B. Hao, T. Lattimore, C. Szepesvári, and M. Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2021.

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. In *The Mathematics of Generalization*, pages 37–116. CRC Press, 2018.

J. Huang, J. Chen, L. Zhao, T. Qin, N. Jiang, and T.-Y. Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*, 2021.

N. Jiang and A. Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, 2018.

N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020a.

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

E. Kaufmann, P. Ménard, O. D. Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, 2021.

T. Lattimore and C. Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, 2020.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

H. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, 2019.

J. Lee, A. Pacchiano, V. Muthukumar, W. Kong, and E. Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Z. Lin, G. Thomas, G. Yang, and T. Ma. Model-based adversarial meta-reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

P. Massart. Rates of convergence in the central limit theorem for empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 381–423, 1986.

P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, 2021.

Z. Mhammedi, A. Block, D. J. Foster, and A. Rakhlin. Efficient model-free exploration in low-rank mdps. *arXiv preprint arXiv:2307.03997*, 2023.

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

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *Conference on Artificial Intelligence and Statistics*, 2020.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

I. Osband and B. V. Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, 2014.

A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv:2012.13045*, 2020.

M. Papini, A. Tirinzoni, A. Pacchiano, M. Restelli, A. Lazaric, and M. Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. In *Advances in Neural Information Processing Systems*, 2021.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

D. Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

T. Ren, T. Zhang, C. Szepesvári, and B. Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, 2022.

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

W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, 2019a.

W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, 2019b.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

M. Uehara, X. Zhang, and W. Sun. Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*, 2021.

B. Van Roy and S. Dong. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv:1911.07910*, 2019.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, 2022.

R. Wang, S. S. Du, F. L. Yang, and R. Salakhutdinov. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2020a.

R. Wang, R. Salakhutdinov, and L. F. Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv:2005.10804*, 2020b.

L. F. Yang and M. Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2020.

Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv:2011.04622*, 2020.

A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, 2020a.

A. Zanette, A. Lazaric, M. J. Kochenderfer, and E. Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*, 2020b.

A. Zhang, R. T. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2020.

R. Zhang, B. Dai, L. Li, and D. Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2019.

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

W. Zhang, D. Zhou, and Q. Gu. Reward-free model-based reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2021b.

X. Zhang, Y. Song, M. Uehara, M. Wang, A. Agarwal, and W. Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, 2022.

Z. Zhang, S. Du, and X. Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, 2021c.