Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms

Chi Jin CHIJ@PRINCETON.EDU
Princeton University

Qinghua Liu QINGHUAL@PRINCETON.EDU
Princeton University

Sobhan Miryoosefi MIRYOOSEFI@CS.PRINCETON.EDU
Princeton University

Abstract

Finding the minimal structural assumptions that empower sample-efficient learning is one of the most important research directions in Reinforcement Learning (RL). This paper advances our understanding of this fundamental question by introducing a new complexity measure—Bellman Eluder (BE) dimension. We show that the family of RL problems of low BE dimension is remarkably rich, which subsumes a vast majority of existing tractable RL problems including but not limited to tabular MDPs, linear MDPs, reactive POMDPs, low Bellman rank problems as well as low Eluder dimension problems. This paper further designs a new optimization-based algorithm—GOLF, and reanalyzes a hypothesis elimination-based algorithm—OLIVE (proposed in Jiang et al. (2017)). We prove that both algorithms learn the near-optimal policies of low BE dimension problems in a number of samples that is polynomial in all relevant parameters, but independent of the size of state-action space. Our regret and sample complexity results match or improve the best existing results for several well-known subclasses of low BE dimension problems.

1. Introduction

Modern Reinforcement Learning (RL) commonly engages practical problems with an enormous number of states, where *function approximation* must be deployed to approximate the true value function using functions from a prespecified function class. Function approximation, especially based on deep neural networks, lies at the heart of the recent practical successes of RL in domains such as Atari (Mnih et al., 2013), Go (Silver et al., 2016), robotics (Kober et al., 2013), and dialogue systems (Li et al., 2016).

Despite its empirical success, RL with function approximation raises a new series of theoretical challenges when comparing to the classic tabular RL: (1) *generalization*, to generalize knowledge from the visited states to the unvisited states due to the enormous state space. (2) *limited expressiveness*, to handle the complicated issues where true value functions or intermediate steps computed in the algorithm can be functions outside the prespecified function class. (3) *exploration*, to address the tradeoff between exploration and exploitation when above challenges are presented.

Consequently, most existing theoretical results on efficient RL with function approximation rely on relatively strong structural assumptions. For instance, many require that the MDP admits a linear approximation (Wang et al., 2019; Jin et al., 2020; Zanette et al., 2020a), or that the model is precisely Linear Quadratic Regulator (LQR) (Anderson and Moore, 2007; Fazel et al., 2018;

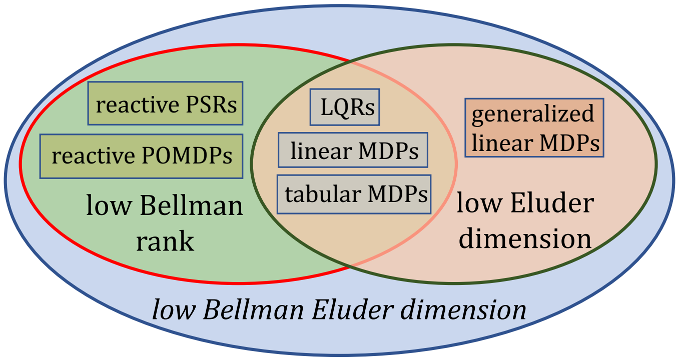

Figure 1: A schematic summarizing relations among families of RL problems

Dean et al., 2019). Most of these structural assumptions rarely hold in practical applications. This naturally leads to one of the most fundamental questions in RL.

What are the minimal structural assumptions that empower sample-efficient RL?

We advance our understanding of this grand question via the following two steps: (1) identify a rich class of RL problems (thus with weak structural assumption) that cover many practical applications of interests; (2) design sample-efficient algorithms that provably learn any RL problem in this rich class.

The attempts to find weak or minimal structural assumptions that allow statistical learning can be traced in supervised learning where VC dimension (Vapnik, 2013) or Rademacher complexity (Bartlett and Mendelson, 2002) is proposed, or in online learning where Littlestone dimension (Littlestone, 1988) or sequential Rademacher complexity (Rakhlin et al., 2010) is developed.

In the area of reinforcement learning, there are two intriguing lines of recent works that have made significant progress in this direction. To begin with, Jiang et al. (2017) introduces a generic complexity notion—Bellman rank, which can be proved small for many RL problems including linear MDPs (Jin et al., 2020), reactive POMDPs (Krishnamurthy et al., 2016), etc. Jiang et al. (2017) further propose an hypothesis elimination-based algorithm—OLIVE for sample-efficient learning of problems with low Bellman rank. On the other hand, recent work by Wang et al. (2020) considers general function approximation with low Eluder dimension (Russo and Van Roy, 2013), and designs a UCB-style algorithm with regret guarantee. Noticeably, the set of generalized linear MDPs (Wang et al., 2019) is a subclass of low Eluder dimension problems, but not low Bellman rank.

In this paper, we make the following three contributions.

- We introduce a new complexity measure for RL—Bellman Eluder (BE) dimension. We prove that the family of RL problems of low BE dimension is remarkably rich, which subsumes both low Bellman rank problems and low Eluder dimension problems—two arguably most generic tractable function classes so far in the literature (see Figure 1).

- We design a new optimization-based algorithm—GOLF, which provably learn the near-optimal policies of low BE dimension problems in a number of samples that is polynomial in all relevant parameters, but independent of the size of state-action space. Our regret or sample complexity guarantees match Zanette et al. (2020a) when specified to the linear setting, and improve upon Jiang et al. (2017); Wang et al. (2020) in low Bellman rank, and low Eluder dimension settings respectively.

- We reanalyze the hypothesis elimination-based algorithm—OLIVE proposed in Jiang et al. (2017). We show it can also learn RL problems with low BE dimension sample-efficiently, under slightly different assumptions but with worse sample complexity comparing to GOLF.

1.1. Related works

This section reviews prior theoretical works on RL, under Markov Decision Process (MDP) models.

We remark that there has been a long line of research on function approximation in the *batch RL* setting (see, e.g., Szepesvári and Munos, 2005; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Xie and Jiang, 2020). In this setting, agents are provided with exploratory data or simulator, so that they do not need to explicitly address the challenge in exploration. In this paper, we do not make such assumption, and attack exploration problem directly. In the following we focus exclusively on the RL results in the general setting that require exploration.

Tabular RL. Tabular RL concerns MDPs with a small number of states and actions, which has been thoroughly studied in recent years (see, e.g., Brafman and Tennenholtz, 2002; Jaksch et al., 2010; Dann and Brunskill, 2015; Agrawal and Jia, 2017; Azar et al., 2017; Zanette and Brunskill, 2019; Jin et al., 2018; Zhang et al., 2020). In the episodic setting with non-stationary dynamics, the best regret bound $\tilde{\mathcal{O}}(\sqrt{H^2|\mathcal{S}||\mathcal{A}|T})$ is achieved by both model-based (Azar et al., 2017) and model-free (Zhang et al., 2020) algorithms. Moreover, the bound is proved to be minimax-optimal by Jin et al. (2018). This minimax bound suggests that when the state-action space is enormous, RL is information-theoretically hard without further structural assumptions.

RL with linear function approximation. A recent line of work studies RL with linear function approximation (see, e.g., Jin et al., 2020; Wang et al., 2019; Cai et al., 2019; Zanette et al., 2020a,b; Agarwal et al., 2020; Neu and Pike-Burke, 2020). These papers assume certain completeness conditions, as well as the optimal value function can be well approximated by linear functions. Under one formulation of linear approximation, the minimax regret bound $\tilde{\mathcal{O}}(d\sqrt{T})$ is achieved by algorithm ELEANOR (Zanette et al., 2020a), where d is the ambient dimension of the feature space.

RL with general function approximation. Beyond the linear setting, there is a flurry line of research studying RL with general function approximation (see, e.g., Osband and Van Roy, 2014; Jiang et al., 2017; Sun et al., 2019; Dong et al., 2020; Wang et al., 2020; Yang et al., 2020; Foster et al., 2020). Among them, Jiang et al. (2017) and Wang et al. (2020) are the closest to our work.

Jiang et al. (2017) proposes a complexity measure named Bellman rank and design an algorithm OLIVE with sample-efficient PAC guarantee for problems with low Bellman rank. We note that low Bellman rank is a special case of low BE dimension. When specialized to the low Bellman rank setting, our sample complexity of OLIVE exactly matches the guarantee in Jiang et al. (2017). Our result for GOLF requires additional completeness assumption, but achieves better sample complexity.

Wang et al. (2020) proposes a UCB-type algorithm with a regret guarantee under the assumption that the function class has a low eluder dimension. Again, we will show that low Eluder dimension is a special case of low BE dimension. Comparing to Wang et al. (2020), our algorithm GOLF works under weaker completeness assumptions, with a better regret guarantee.

2. Preliminaries

We consider episodic Markov Decision Process (MDP), denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the number of steps in each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is the collection of transition measures with $\mathbb{P}_h(s' \mid s, a)$ equal to the probability of transiting to s' after taking action a at state s at the h^{th} step, and $r = \{r_h\}_{h\in[H]}$ is the collection of reward functions with $r_h(s, a)$ equal to the deterministic reward received after taking action a at state s at the h^{th} step. [1] Throughout this paper, we assume reward is non-negative, and $\sum_{h=1}^{H} r_h(s_h, a_h) \le 1$ for all possible sequence $(s_1, a_1, \ldots, s_H, a_H)$.

In each episode, the agent starts at a *fixed* initial state s_1. Then, at each step $h \in [H]$, the agent observes its current state s_h, takes action a_h, receives reward $r_h(s_h, a_h)$, and causes the environment to transit to $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. Without loss of generality, we assume there is a terminating state s_{end} which the environment will *always* transit to s_{end} at step $H+1$, and the episode terminates when s_{end} is reached.

Policy and value functions A (deterministic) policy π of an agent is a collection of H functions $\{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^{H}$. We denote $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as the value function at step h for policy π, so that $V_h^\pi(s)$ gives the expected sum of the remaining rewards received under policy π, starting from $s_h = s$, till the end of the episode. In symbol,

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right],$$

Similarly, we denote $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as the Q-value function at step h for policy π, where

$$Q_h^\pi(s, a) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right].$$

There exists an optimal policy $\pi^\star$, which gives the optimal value function for all states (Puterman, 2014), in the sense, $V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s)$ for all $h \in [H]$ and $s \in \mathcal{S}$. For notational simplicity, we abbreviate $V^{\pi^\star}$ as $V^\star$. We similarly define the optimal Q-value function as $Q^\star$. Recall that $Q^\star$ satisfies the Bellman optimality equation:

$$Q_h^\star(s, a) = (\mathcal{T}_h Q_{h+1}^\star)(s, a) := r_h(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^\star(s', a'). \tag{1}$$

for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. We also call $\mathcal{T}_h$ the *Bellman operator* at step h.

ϵ-optimality and regret We say a policy π is ϵ-optimal if $V_1^\pi(s_1) \ge V_1^\star(s_1) - \epsilon$. Suppose an agent interacts with the environment for K episodes. Denote by π^k the policy the agent follows in episode $k \in [K]$. The (accumulative) regret is defined as

$$\mathrm{Reg}(K) := \sum_{k=1}^{K} \left[V^*(s_1) - V^{\pi^k}(s_1)\right].$$

The objective of reinforcement learning is to find an ϵ-optimal policy within a small number of interactions or achieve sublinear regret.

[1] We study deterministic reward functions for notational simplicity. Our results readily generalize to random rewards.

2.1. Function approximation

In this paper, we consider reinforcement learning with value function approximation. Formally, the learner is given a function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where $\mathcal{F}_h \subseteq \mathcal{S} \times \mathcal{A} \to [0,1]$ offers a set of candidate functions to approximate $Q^\star_h$—the optimal Q-value function at step h. Since no reward is collected in the $(H+1)^{\text{th}}$ steps, we always set $f_{H+1} = 0$.

Reinforcement learning with function approximation in general is extremely challenging without further assumptions (see, e.g., hardness results in Krishnamurthy et al. (2016); Weisz et al. (2020)). Below, we present two assumptions about function approximation that are commonly adopted in the literature.

Assumption 1 (Realizability) *$Q^\star_h \in \mathcal{F}_h$ for all $h \in [H]$.*

Realizability requires the function class is well-specified, i.e. function class $\mathcal{F}_h$ in fact contains the optimal Q-value function $Q^\star_h$ with no approximation error.

Assumption 2 (Completeness) *$\mathcal{T}_h \mathcal{F}_{h+1} \subseteq \mathcal{F}_h$ for all $h \in [H]$.*

Note $\mathcal{T}_h \mathcal{F}_{h+1}$ is defined as $\{\mathcal{T}_h f : f \in \mathcal{F}_{h+1}\}$. Completeness requires the function class $\mathcal{F}$ to be closed under the Bellman operator.

When function class $\mathcal{F}$ has finite elements, we can use its cardinality $|\mathcal{F}|$ to measure the "size" of function class $\mathcal{F}$. When we are addressing function classes with infinite elements, we require a notion similar to cardinality. We use the standard ϵ-covering number.

Definition 3 (ϵ-covering number) *The ϵ-covering number of a set $\mathcal{V}$ under metric ρ, denoted as $\mathcal{N}(\mathcal{V}, \epsilon, \rho)$, is the minimum integer n such that there exists a subset $\mathcal{V}_o \subset \mathcal{V}$ with $|\mathcal{V}_o| = n$, and for any $x \in \mathcal{V}$, there exists $y \in \mathcal{V}_o$ such that $\rho(x, y) \le \epsilon$.*

We refer readers to standard textbooks (see, e.g., Wainwright, 2019) for further properties of covering number. In this paper, we will always apply the covering number on function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_{H+1}$, and use metric $\rho(f, g) = \max_h \|f_h - g_h\|_\infty$. For notational simplicity, we omit the metric dependence and denote the covering number as $\mathcal{N}_{\mathcal{F}}(\epsilon)$.

2.2. Eluder dimension

One class of function highly related to this paper is the function class of low Eluder dimension (Russo and Van Roy, 2013).

Definition 4 (ϵ-independence between points) *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and z,x_1, x_2 ,. . .,x_n∈ $\mathcal{X}$. We say z is ϵ-independent of $\{x_1, x_2, \ldots, x_n\}$ with respect to $\mathcal{G}$ if there exist $g_1, g_2 \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^n (g_1(x_i) - g_2(x_i))^2} \le \epsilon$, but $g_1(z) - g_2(z) > \epsilon$.*

Intuitively, z is independent of $\{x_1, x_2, \ldots, x_n\}$ means if that there exist two "certifying" functions g_1 and g_2, so that their function values are similar at all point $\{x_i\}_{i=1}^n$, but the values are rather different at z. This independence relation naturally induces the following complexity measure.

Definition 5 (Eluder dimension) *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$. The Eluder dimension $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon)$ is the length of the longest sequence $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ such that there exists $\epsilon' \ge \epsilon$ where x_i is ϵ'-independent of $\{x_1, \ldots, x_{i-1}\}$ for all $i \in [n]$.*

Recall that a vector space has dimension d if and only if d is the length of the longest sequence of elements $\{x_1, \dots, x_d\}$ such that x_i is linearly independent of $\{x_1, \dots, x_{i-1}\}$ for all $i \in [n]$. Eluder dimension generalizes the linear independence relation in standard vector space dimension to capture both nonlinear independence and approximate independence, and thus is more general.

3. Bellman Eluder Dimension

In this section, we introduce our new complexity measure—Bellman Eluder (BE) dimension. As one of its most important properties, we will show that the class of functions with low BE dimension contains the two existing most general tractable function classes in RL—functions with low Bellman rank, and functions with low Eluder dimension (see Figure 1).

We start by developing a new distributional version of the original Eluder dimension proposed by Russo and Van Roy (2013) (See Section 2.2 for more details).

Definition 6 (ϵ-independence between distributions) *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \dots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ϵ-independent of $\{\mu_1, \mu_2, \dots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists $g \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^{n}(\mathbb{E}_{\mu_i}[g])^2} \le \epsilon$, but $|\mathbb{E}_{\nu}[g]| > \epsilon$.*

Definition 7 (Distributional Eluder (DE) dimension) *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and Π be a family of probability measures over $\mathcal{X}$. The distributional Eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \dots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \ge \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \dots, \rho_{i-1}\}$ for all $i \in [n]$.*

Definition 6 and Definition 7 generalize Definition 4 and Definition 5 to the distributional version, by inspecting the expected values of function instead of the function values at single points, and by restricting the candidate distributions to a certain family Π. The main advantage of this generalization is exactly in the statistical setting, when the estimation of the expected value of function with respect to certain distribution family can be easier than the estimation of the function at all single points (which is the case for RL in large state space).

It is clear that standard Eluder dimension is a special case of distribution Eluder dimension, because when choosing $\Pi = \{\delta_x(\cdot) | x \in \mathcal{X}\}$ where $\delta_x(\cdot)$ is the dirac measure centered at x, then $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) = \dim_{\mathrm{DE}}(\mathcal{G} - \mathcal{G}, \Pi, \epsilon)$ where $\mathcal{G} - \mathcal{G} = \{g_1 - g_2 : \ g_1, g_2 \in \mathcal{G}\}$.

Now we are ready to introduce our key notion in this paper—Bellman Eluder dimension.

Definition 8 (Bellman Eluder (BE) dimension) *Let $(I - \mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : \ f \in \mathcal{F}\}$ be the set of Bellman residuals induced by $\mathcal{F}$ at step h, and $\Pi = \{\Pi_h\}_{h=1}^{H}$ be a collection of H probability measure families over $\mathcal{S} \times \mathcal{A}$. The ϵ-Bellman Eluder of $\mathcal{F}$ with respect to Π is defined as* [2]

$$\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon) := \max_{h \in [H]} \dim_{\mathrm{DE}}\big((I - \mathcal{T}_h)\mathcal{F}, \Pi_h, \epsilon\big).$$

In short, Bellman Eluder dimension is simply the distribution Eluder dimension on the function class of Bellman residuals, maximizing over all steps. In additional to the function class $\mathcal{F}$ and error ϵ, the Bellman Eluder dimension also depends on the choice of distribution family Π. For the purpose of this paper, we focus on the following two specific choices.

[2] With a different choice of Bellman residual, we can alternatively define type-II Bellman Eluder dimension, where similar results can be obtained. For clean presentation, we defer all results of type-II BE dimension to Appendix A.

1. $\mathcal{D}_{\mathcal{F}} := \{\mathcal{D}_{\mathcal{F},h}\}_{h\in[H]}$, where $\mathcal{D}_{\mathcal{F},h}$ denotes the collection of all probability measures over $\mathcal{S}\times\mathcal{A}$ at the h^{th} step, which can be generated by executing the greedy policy π_f induced by any $f\in\mathcal{F}$, i.e., $\pi_{f,h}(\cdot) = \operatorname{argmax}_{a\in\mathcal{A}} f_h(\cdot, a)$ for all $h\in[H]$.
2. $\mathcal{D}_{\Delta} := \{\mathcal{D}_{\Delta,h}\}_{h\in[H]}$, where $\mathcal{D}_{\Delta,h} = \{\delta_{(s,a)}(\cdot) | s\in\mathcal{S}, a\in\mathcal{A}\}$, i.e. the collections of probability measures that put measure 1 on single state-action pairs.

We say a RL problem has low BE dimension if $\min_{\Pi\in\{\mathcal{D}_{\mathcal{F}},\mathcal{D}_{\Delta}\}} \dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$ is small.

3.1. Relations with existing tractable function classes in RL

Known tractable problem classes in RL include but not limited to tabular MDPs, linear MDPs (Jin et al., 2020), linear quadratic regulators (Anderson and Moore, 2007), generalized linear MDPs (Wang et al., 2019), reactive POMDPs (Krishnamurthy et al., 2016), reactive PSRs (Singh et al., 2012; Jiang et al., 2017). There are two existing generic tractable problem classes that jointly contain all the examples mentioned above: the set of RL problems with low Bellman rank, and the set of RL problems with low Eluder dimension. However, for these two generic sets, one does not contain the other.

In this section, we will show that our new class of RL problems with low BE dimension in fact contains both low Bellman rank problems and low Eluder dimension problems (see Figure 1). That is, our new problem class covers almost all existing tractable RL problems, and, to our best knowledge, is the most generic tractable function class so far.

Relation with low Bellman rank The seminar paper by Jiang et al. (2017) proposes the complexity measure—Bellman rank, and shows that a majority of RL examples mentioned above have low Bellman rank. They also propose a hypothesis elimination based algorithm—OLIVE, that learns any low Bellman rank problem within polynomial samples. Formally,

Definition 9 (Bellman rank) *The Bellman rank is the minimum integer d so that there exists $\phi_h : \mathcal{F}\to\mathbb{R}^d$ and $\psi_h : \mathcal{F}\to\mathbb{R}^d$ for each $h\in[H]$, such that for any $f, f'\in\mathcal{F}$, the average Bellman error*[3]

$$\mathcal{E}(f, \pi_{f'}, h) := \mathbb{E}_{\pi_{f'}}[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)] = \langle \phi_h(f), \psi_h(f')\rangle,$$

where $\|\phi_h(f)\|_2 \cdot \|\psi_h(f)\|_2 \le \zeta$, and ζ is the normalization parameter.

Recall that we use π_f to denote the greedy policy induced by value function f. Intuitively, a problem with Bellman rank says its average Bellman error can be decomposed as the inner product of two d-dimensional vectors, where one vector depends on the roll-in distribution π_f', while the other vector depends on the value function f. In a high level, it claims that the average Bellman error has a linear inner product structure.

Proposition 10 (low Bellman rank $\subset$ low BE dimension) *If an MDP with function class $\mathcal{F}$ has Bellman rank d with normalization parameter ζ, then*

$$\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \le \mathcal{O}(1 + d\log(1+\zeta/\epsilon)).$$

[3] The definition presented here is slightly different from the original version in Jiang et al. (2017). In this paper, we denote the original version as type-II Bellman rank. We can also show that low type-II Bellman rank $\subset$ low type-II BE dimension, and low type-II BE dimension problems can be sample-efficiently learned (see Appendix A).

Proposition 10 claims that problems with low Bellman rank also have low BE dimension, with a small multiplicative factor that is only logarithmic in ζ and ϵ^{-1}.

Next, we show that the set of low BE dimension problems is a strictly larger than the set of low Bellman rank problems. This is intuitively because, as a feature of Eluder dimension, problems with low BE dimension further allow the average Bellman error to have certain nonlinear structure.

Proposition 11 (low BE dimension $\not\subset$ low Bellman rank) *For any $m \in \mathbb{N}^+$, there exist an MDP and a function class $\mathcal{F}$ so that for all $\epsilon > 0$, $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \leq 9$, but the Bellman rank is at least m.*

Relation with low Eluder dimension Wang et al. (2020) study the setting where the function class $\mathcal{F}$ has low Eluder dimension, which includes generalized linear functions. They prove that, when the completeness assumption is satisfied,[4] low Eluder dimension problems can be efficiently learned in polynomial samples.

Proposition 12 (low Eluder dimension $\subset$ low BE dimension) *Assume $\mathcal{F}$ satisfies completeness (Assumption 2). Then for all $\epsilon > 0$,*

$$\dim_{\mathrm{BE}}\left(\mathcal{F}, \mathcal{D}_{\Delta}, \epsilon\right) \leq \max_{h\in[H]} \dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon).$$

Proposition 12 asserts that problems with low Eluder dimension also have low BE dimension, which is a natural consequence of completeness and the fact that Eluder dimension is a special case of distributional Eluder dimension.

Finally, similar to proposition 11, we can also show that the set of low BE dimension problems is a strictly larger than the set of low Eluder dimension problems.

Proposition 13 (low BE dimension $\not\subset$ low Eluder dimension) *For any $m \in \mathbb{N}^+$, there exist an MDP and a function class $\mathcal{F}$ so that for all $\epsilon \in (0, 1]$, $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\Delta}, \epsilon) \leq 9$, but $\min_{h\in[H]} \dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon)$ is at least m.*

4. Algorithm GOLF

Section 3 defines a new class of RL problems with low BE dimension, and shows that the new class is rich, containing almost all the existing known tractable RL problems so far. In this section, we propose a new simple optimization-based algorithm—**G**lobal **O**ptimism based on **L**ocal **F**itting (GOLF). We prove that, low BE dimension problems are indeed tractable, i.e., GOLF finds the near-optimal policies of these problems within a polynomial number of samples.

The pseudocode of GOLF is given in Algorithm 1. GOLF initializes datasets $\{\mathcal{D}_h\}_{h=1}^H$ to be empty sets, and confidence set $\mathcal{B}^0$ to be $\mathcal{F}$. Then, in each episode, GOLF performs two main steps:

- Line 3 (Optimistic planning): compute the most optimistic value function f^k from the confidence set $\mathcal{B}^{k-1}$ constructed in the last episode[5] , and choose π^k to be its greedy policy.
- Line 4-6 (Execute the policy and update the confidence set): execute policy π^k for one episode, collect data, and update the confidence set using the new data.

[4] Wang et al. (2020) assume for any function g (not necessarily in $\mathcal{F}$), $\mathcal{T}g \in \mathcal{F}$, which is stronger than the completeness assumption presented in this paper (Assumption 2).

[5] We remark that in general, the optimization problem in Line 3 of GOLF can not be solved computationally efficiently.

Algorithm 1 GOLF $(\mathcal{F}, K, \beta)$ — **G**lobal **O**ptimism based on **L**ocal **F**itting

```
Initialize: $\mathcal{D}_1, \dots, \mathcal{D}_H \leftarrow \emptyset$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
for episode $k$ from 1 to $K$ do
    Choose policy $\pi^k = \pi_{f^k}$, where $f^k = \mathrm{argmax}_{f \in \mathcal{B}^{k-1}} f(s_1, \pi_f(s_1))$.
    Collect a trajectory $(s_1, a_1, r_1, \dots, s_H, a_H, r_H, s_{H+1})$ by following $\pi^k$.
    Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(s_h, a_h, r_h, s_{h+1})\}$ for all $h \in [H]$.
    Update
        $\mathcal{B}^k = \left\{ f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta \text{ for all } h \in [H] \right\}$,
        where $\mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{(s,a,r,s') \in \mathcal{D}_h} [\xi_h(s,a) - r - \max_{a' \in \mathcal{A}} \zeta_{h+1}(s', a')]^2.$    (2)
Output $\pi^{\mathrm{out}}$ sampled uniformly at random from $\{\pi^k\}_{k=1}^K$.
```

At the heart of GOLF is the way we construct the confidence set $\mathcal{B}^k$. For each $h \in [H]$, GOLF maintains a *local* regression constraint using the collected transition data $\mathcal{D}_h$ at this step

$$\mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta, \tag{3}$$

where β is a confidence parameter, and $\mathcal{L}_{\mathcal{D}_h}$ is the squared loss defined in (2), which can be viewed as a proxy to the squared Bellman error at step h. Classic algorithm—Fitted Q-Iteration (FQI) (Szepesvári, 2010) simply updates $f_h \leftarrow \mathrm{argmin}_{g \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1})$. Our constraint (3) can be viewed as a relaxed version of this update, which allows f_h to be not only the minimizer of the loss $\mathcal{L}_{\mathcal{D}_h}(\cdot, f_{h+1})$, but also any function whose loss is only slightly larger then the optimal loss.

4.1. Theoretical guarantees

Now, we present the theoretical guarantees for GOLF.

Theorem 14 (Regret of GOLF) *Under Assumption 1, 2, there exists an absolute constant c such that for any $\delta \in (0, 1]$, $K \in \mathbb{N}$, if we choose parameter $\beta = c \log[\mathcal{N}_{\mathcal{F}}(1/K) \cdot KH/\delta]$ in* GOLF, *then with probability at least $1 - \delta$, for all $k \in [K]$, we have $\mathrm{Reg}(k) \le \mathcal{O}(H\sqrt{dk\beta})$, where $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_\mathcal{F}\}} \dim_{\mathrm{BE}}\left(\mathcal{F}, \Pi, 1/\sqrt{K}\right)$ is the BE dimension.*

Theorem 14 asserts that, under the realizability and completeness assumptions, the general class of RL problems with low BE dimension is indeed tractable: there exists an algorithm (GOLF) that can achieve $\sqrt{K}$ regret, whose multiplicative factor depends only polynomially on the horizon of MDP H, the BE dimension d, and the log covering number of the function class. Most importantly, the regret is independent of the number of the states, which is crucial for dealing with practical RL problems with function approximation, where the state space is typically exponentially large.

We remark that when function class $\mathcal{F}$ has a finite number of elements, the covering number is upper bounded by its cardinality $|\mathcal{F}|$. For a wide range of function classes in practice, the log ϵ'-covering number has only logarithmic dependence on ϵ'. Informally, we denote the log covering number as $\log \mathcal{N}_\mathcal{F}$ and omit its ϵ' dependency for clean presentation. Theorem 14 claims that the regret scales as $\tilde{\mathcal{O}}(H\sqrt{dK \log \mathcal{N}_\mathcal{F}})$, where $\tilde{\mathcal{O}}(\cdot)$ omits absolute constant and logarithmic terms.[6]

[6] We will not omit $\log \mathcal{N}_\mathcal{F}$ in $\tilde{\mathcal{O}}(\cdot)$ notation since for many function classes, $\log \mathcal{N}_\mathcal{F}$ is not small. For instance, for linear function class, $\log \mathcal{N}_\mathcal{F} = \tilde{\mathcal{O}}(\tilde{d})$ where $\tilde{d}$ is ambient dimension.

By standard online-to-batch argument, we also derive the sample complexity of GOLF.

Corollary 15 (Sample Complexity of GOLF) *Under Assumption 1, 2, there exists an absolute constant c such that for any $\epsilon \in (0, 1]$, if we choose $\beta = c\log[\mathcal{N}_{\mathcal{F}}(\epsilon^2/(dH^2)) \cdot HK]$ in* GOLF, *then the output policy π^{out} is $\mathcal{O}(\epsilon)$-optimal with probability at least $1/2$, if $K \geq \Omega((H^2d/\epsilon^2) \cdot \log[\mathcal{N}_{\mathcal{F}}(\epsilon^2/(dH^2)) \cdot Hd/\epsilon])$, where $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_{\mathcal{F}}\}} \dim_{\mathrm{BE}}\left(\mathcal{F}, \Pi, \epsilon/H\right)$ is the BE dimension.*

Corollary 15 claims that $\tilde{\mathcal{O}}(H^2 d\log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$ samples are enough for GOLF to learn a near-optimal policy of any low BE dimension problem. Our sample complexity scales linear in both the BE dimension d, and the log covering number $\log(\mathcal{N}_{\mathcal{F}})$.

To showcase the sharpness of our results, we compare them to the previous results when restricted to the corresponding settings. (1) For linear function class with d_{lin} ambient dimension, we have BE dimension $d = \tilde{\mathcal{O}}(d_{\text{lin}})$ and $\log(\mathcal{N}_{\mathcal{F}}) = \tilde{\mathcal{O}}(d_{\text{lin}})$. Our regret bound becomes $\tilde{\mathcal{O}}(Hd_{\text{lin}}\sqrt{K})$ which matches the best known result (Zanette et al., 2020a) up to logarithmic factors; (2) For function class with low Eluder dimension (Wang et al., 2020), our results hold under weaker completeness assumptions. Our regret scales with $\sqrt{d_{\text{E}}}$ in terms of dependency on Eluder dimension d_{E}, which improves the linear d_{E} scaling in the regret of Wang et al. (2020); (3) Finally, for low Bellman rank problems, our sample complexity scales linearly with Bellman rank, which improves upon the quadratic dependence in Jiang et al. (2017). We remark that all results mentioned above assume (approximate) realizability. All except Jiang et al. (2017) assume (approximate) completeness.

4.2. Key ideas in proving Theorem 14

In this subsection, we present a brief sketch for proving the regret bound of GOLF. We defer all the proof details to Appendix C. For simplicity, we only discusses the case of choosing $\mathcal{D}_{\mathcal{F}}$ as the distribution family Π in the definition of Bellman Eluder dimension (Definition 8). The proof for using $\mathcal{D}_\Delta$ as distribution family follows from similar arguments.

Our proof strategy has three main steps.

Step 1: Prove optimism. We firstly show that, with high probability, the optimal value function $Q^\star$ indeed lies in the confidence set $\mathcal{B}^k$ for all $k \in [K]$ (Lemma 25 in Appendix C.1), which is a natural consequence of martingale concentration and the specific form of the confidence set we constructed. Because of $Q^\star \in \mathcal{B}^k$, the optimistic planning step (Line 3) in GOLF guarantees that $V_1^\star(s_1) \leq \max_a f_1^k(s_1, a)$ for every episode k. This optimism allows the following upper bound on regret

$$\mathrm{Reg}(K) \leq \sum_{k=1}^{K} \left(\max_a f_1^k(s_1, a) - V_1^{\pi^k}(s_1)\right) = \sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{E}_{\pi^k}[(f_h^k - \mathcal{T} f_{h+1}^k)(s_h, a_h)], \tag{4}$$

where the right equality follows from the standard policy loss decomposition (see, e.g., Lemma 1 in Jiang et al. (2017)), and $\mathbb{E}_\pi$ denotes the expectation taken over the sequence $(s_1, a_1, \dots, s_H, a_H)$ when executing policy π.

Step 2: Utilize the sharpness of our constraint set. Recall that our construction of the confidence set in Line 6 of GOLF forces f^k computed in episode k to have a small loss $\mathcal{L}_{\mathcal{D}_h}$, which is a proxy for empirical squared Bellman error under data $\mathcal{D}_h$. Since data in $\mathcal{D}_h$ in episode k are collected by

Algorithm 2 OLIVE $(\mathcal{F}, \zeta_{\text{act}}, \zeta_{\text{elim}}, n_{\text{act}}, n_{\text{elim}})$

```
**Initialize**: $\mathcal{B}^0 \leftarrow \mathcal{F}$, $\mathcal{D}_h \leftarrow \emptyset$ for all $h, k$.
**for phase** $k = 1, 2, \ldots$ **do**
    **Choose policy** $\pi^k = \pi_{f^k}$, where $f^k = \text{argmax}_{f\in\mathcal{B}^{k-1}} f(s_1, \pi_f(s_1))$.
    **Execute** $\pi^k$ for $n_{\text{act}}$ episodes and *refresh* $\mathcal{D}_h$ to include the fresh $(s_h, a_h, r_h, s_{h+1})$ tuples.
    **Estimate** $\hat{\mathcal{E}}(f^k, \pi^k, h)$ for all $h \in [H]$, where
        $$\hat{\mathcal{E}}(g, \pi, h) = \frac{1}{|\mathcal{D}_h|} \sum_{(s,a,r,s')\in\mathcal{D}_h} \left( g_h(s,a) - r - \max_{a'\in\mathcal{A}} g_{h+1}(s', a') \right).$$
    **if** $\sum_{h=1}^{H} \hat{\mathcal{E}}(f^k, \pi^k, h) \le H\zeta_{\text{act}}$ **then**
        Terminate and output $\pi^k$.
    Pick any $t \in [H]$ for which $\hat{\mathcal{E}}(f^k, \pi^k, t) \ge \zeta_{\text{act}}$.
    **Execute** $\pi^k$ for $n_{\text{elim}}$ episodes and *refresh* $\mathcal{D}_h$ to include the fresh $(s_h, a_h, r_h, s_{h+1})$ tuples.
    **Estimate** $\hat{\mathcal{E}}(f, \pi^k, t)$ for all $f \in \mathcal{F}$.
    **Update** $\mathcal{B}^k = \left\{ f \in \mathcal{B}^{k-1} : \left| \hat{\mathcal{E}}(f, \pi^k, t) \right| \le \zeta_{\text{elim}} \right\}$.
```

executing each π^i for one episode for all $i < k$, by standard martingale concentration argument and the completeness assumption, we can show that with high probability (Lemma 24 in Appendix C.1)

$$\sum_{i=1}^{k-1} \mathbb{E}_{\pi^i}[(f_h^k - \mathcal{T} f_{h+1}^k)(s_h, a_h)]^2 \le \mathcal{O}(\beta), \text{ for all } (k,h) \in [K] \times [H]. \tag{5}$$

Step 3: Establish relations between (4) and (5). So far, we want to upper-bound (4), while we know (5). We note that the RHS of (4) is very similar to the LHS of (5), except that the latter is the squared Bellman error, and the expectation is taken under previous policy π^i for $i < k$. To establish the connection between these two, it turns out that we need the Bellman Eluder dimension to be small. Concretely, we have the following lemma.

Lemma 16 *Given a function class $\mathcal{G}$ defined on $\mathcal{X}$ with $|g(x)| \le 1$ for all $(g, x) \in \mathcal{G} \times \mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{g_k\}_{k=1}^K \subset \mathcal{G}$ and $\{\mu_k\}_{k=1}^K \subset \Pi$ satisfy that for all $k \in [K]$, $\sum_{i=1}^{k-1} (\mathbb{E}_{\mu_i}[g_k])^2 \le \beta$. Then for all $k \in [K]$, $\sum_{i=1}^{k} |\mathbb{E}_{\mu_i}[g_i]| \le \mathcal{O}(\sqrt{\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, 1/k)\beta k})$.*

Lemma 16 is a simplification of Lemma 26 in Appendix C, which is a modification of Lemma 2 in Russo and Van Roy (2013). Intuitively, Lemma 16 can be viewed as an analogue of the pigeon-hole principles for DE dimension. Choose $\mathcal{G}$ to be the function class of Bellman residuals, and μ_k to be the distribution under policy π^k, we finish the proof.

5. Algorithm OLIVE

In this section, we analyze the OLIVE algorithm proposed in Jiang et al. (2017), which is based on hypothesis elimination. We prove that, despite OLIVE was originally proposed for solving low Bellman rank problems, it naturally learns the RL problems with low BE dimension as well.

The pseudocode of OLIVE is presented in Algorithm 2, in each phase, the algorithm contains the following three main components.

- Line 3 (Optimistic planning): compute the most optimistic value function f^k from the candidate set $\mathcal{B}^{k-1}$, and choose π^k to be its greedy policy.
- Line 4-7 (Estimate Bellman error): estimate the Bellman error of f^k under π^k; output π^k if the estimated error is small, and otherwise activate the elimination procedure.
- Line 8-11 (Eliminate functions with large Bellman error): pick a step $t \in [H]$ whose estimated Bellman error exceeds the activation threshold ζ_{act}; eliminate all functions in the candidate set whose Bellman error at step t exceeds the elimination threshold ζ_{elim}.

We comment that OLIVE is computationally inefficient in general because implementing the optimistic planning part requires solving an NP-hard problem in the worst case (Dann et al., 2018).

5.1. Theoretical guarantees

Now, we are ready to present the theoretical guarantee for OLIVE.

Theorem 17 (OLIVE) *Under Assumption 1, there exists absolute constant c such that if we choose*

$$\zeta_{act} = \frac{2\epsilon}{H},\ \zeta_{elim} = \frac{\epsilon}{2H\sqrt{d}},\ n_{act} = \frac{H^2\iota}{\epsilon^2},\ \textit{and}\ n_{elim} = \frac{H^2 d \log(\mathcal{N}_{\mathcal{F}}(\zeta_{elim}/8)) \cdot \iota}{\epsilon^2}$$

where $d = \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H)$ and $\iota = c\log(Hd/\delta\epsilon)$, then with probability at least $1-\delta$, Algorithm 2 will output an $\mathcal{O}(\epsilon)$-optimal policy using at most $\mathcal{O}(H^3 d^2 \log[\mathcal{N}_{\mathcal{F}}(\zeta_{elim}/8)] \cdot \iota/\epsilon^2)$ episodes.

Theorem 17 claims that OLIVE learns an ϵ-optimal policy of an MDP with BE dimension d within $\tilde{\mathcal{O}}(H^3 d^2 \log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$ episodes. When specialized to low Bellman rank problems, our sample complexity has the same quadratic dependence on Bellman rank d as in Jiang et al. (2017).

Comparing to GOLF, the major advantage of OLIVE is that OLIVE does not require completeness assumption (Assumption 2) to work. Nevertheless, OLIVE only learns the RL problems that have low BE dimension with respect to the distribution family $\mathcal{D}_{\mathcal{F}}$, not $\mathcal{D}_{\Delta}$. The sample complexity of OLIVE is also worse than the sample complexity GOLF (as presented in Corollary 15).

Finally, we comment that interpreting OLIVE through the lens of BE dimension, makes the proof of Theorem 17 surprisingly natural, which follows from the definition of BE dimension along with some standard concentration arguments.

5.2. Interpret OLIVE with BE dimension

In this subsection, we explain the key idea behind OLIVE through the lens of BE dimension.

To provide a clean high-level view, let us assume all estimates are accurate for now, and the activation threshold ζ_{act} and the elimination threshold ζ_{elim} satisfy $\zeta_{\text{elim}}\sqrt{d} \le \zeta_{\text{act}}$, where $d = \dim_{\mathrm{BE}}\big(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \zeta_{\text{act}}\big)$. Since $\mathcal{E}(Q^\star, \pi, h) \equiv 0$ for any (π, h), $Q^\star$ is always in the candidate set. Therefore, the optimistic planning (Line 3) guarantees $\max_a f_1^k(s_1, a) \ge V_1^\star(s_1)$.

If the Bellman error summation is small (Line 6) i.e., $\sum_{h=1}^H \mathcal{E}(f^k, \pi^k, h) \le H\zeta_{\text{act}}$, then by simple policy loss decomposition (e.g., Lemma 1 in Jiang et al. (2017)) and the optimism of f^k, π^k is $H\zeta_{\text{act}}$-optimal. Otherwise, the elimination procedure is activated at some step t satisfying $\mathcal{E}(f^k, \pi^k, t) \ge \zeta_{\text{act}}$ and all f with $\mathcal{E}(f, \pi^k, t) \ge \zeta_{\text{elim}}$ get eliminated. The *key* observation here is:

> *If the elimination procedure is activated at step h in phase $k_1 < \ldots < k_m$, then the roll-in distribution of $\pi^{k_1}, \ldots, \pi^{k_m}$ at step h is an ζ_{act}-independent sequence with respect to the class of Bellman residual $(I - \mathcal{T}_h)\mathcal{F}$ at step h. Therefore, we should have $m \leq d$.*

For the sake of contradiction, assume $m \geq d+1$. Let us prove $\pi^{k_1}, \ldots, \pi^{k_{d+1}}$ is a ζ_{act}-independent sequence. Firstly, for any $j \in [d+1]$, since f^{k_j} is not eliminated in phase $k_1, \ldots, k_{j-1}$, we have

$$\sqrt{\sum_{i=1}^{j-1} \left(\mathcal{E}(f^{k_j}, \pi^{k_i}, h)\right)^2} \leq \sqrt{d} \times \zeta_{\text{elim}} \leq \zeta_{\text{act}}.$$

Besides, because the elimination procedure is activated at step h in phase k_j, we have $\mathcal{E}(f^{k_j}, \pi^{k_j}, h) \geq \zeta_{\text{act}}$. By Definition 6, we obtain that the roll-in distribution of π^{k_j} at step h is ζ_{act}-independent of those of $\pi^{k_1}, \ldots, \pi^{k_{j-1}}$ for $j \in [d+1]$, which contradicts the definition $d = \dim_{\text{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \zeta_{\text{act}})$. As a result, the elimination procedure can happen at most d times for each $h \in [H]$, which means the algorithm should terminate within $dH+1$ phases and output an $H\zeta_{\text{act}}$-optimal policy.

6. Conclusion

In this paper, we propose a new complexity measure—Bellman Eluder (BE) dimension for reinforcement learning with function approximation. Our new complexity measure identifies a new rich class of RL problems that subsume a majority of existing tractable problem classes in RL. We design a new optimization-based algorithm—GOLF, and provide a new analysis for algorithm OLIVE. Both algorithms show that the new rich class of RL problems we identified can be in fact learned within a polynomial number of samples. We hope our results shed light on the future research in finding the minimal structural assumptions that allow sample-efficient reinforcement learning.

References

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

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

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in neural information processing systems*, pages 1422–1432, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

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

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. 2010.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Satinder Singh, Michael James, and Matthew Rudary. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167*, 2012.

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

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pages 880–887, 2005.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33, 2020b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.

Appendix A. Type-II BE Dimension

The definition of Bellman rank, mentioned in Proposition 10 and Definiton 9, is slightly different from the original definition introduced by Jiang et al. (2017). We denote the former by **Type-I** and the latter (the original definition) by **Type-II**. In this section we introduce Type-II BE Dimension as well as Type-II variants of GOLF and OLIVE. We show that similar results can still be obtained for Type-II variants.

Definition 18 (Type-II Bellman rank) *The Type-II Bellman rank is the minimum integer d so that there exists $\phi_h : \mathcal{F} \to \mathbb{R}^d$ and $\psi_h : \mathcal{F} \to \mathbb{R}^d$ for each $h \in [H]$, such that for any $f, f' \in \mathcal{F}$, the average Type-II Bellman error*

$$\mathcal{E}_{\mathrm{II}}(f, \pi_{f'}, h) := \mathbb{E}[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h) \mid s_h \sim \pi_{f'}, a_h \sim \pi_f] = \langle \phi_h(f), \psi_h(f') \rangle,$$

where $\|\phi_h(f)\|_2 \cdot \|\psi_h(f)\|_2 \leq \zeta$, and ζ is the normalization parameter.

The only difference between these two definitions lies in the sampling of a_h. In Type-I definition we have $a_h \sim \pi_{f'}$, however in Type-II definition we have $a_h \sim \pi_f$ instead. It is worth mentioning that the Type-I and Type-II bellman error coincide whenever $f = f'$; namely, $\mathcal{E}(f, \pi_f, h) = \mathcal{E}_{\mathrm{II}}(f, \pi_f, h)$ for all $f \in \mathcal{F}$.

We can similarly define the Type-II variant of BE Dimension. At a high level, **Type-II BE dimension** $\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F}, \Pi, \epsilon)$ measures the complexity of finding a function in $\mathcal{F}$ such that its expected Bellman error under any state distribution in Π is smaller than ϵ.

Definition 19 (Type-II BE dimension) *Let $(I - \mathcal{T}_h)V_{\mathcal{F}} \subseteq \mathcal{S} \to \mathbb{R}$ be the state-wise Bellman residual class of $\mathcal{F}$ at step h which is defined as*

$$(I - \mathcal{T}_h)V_{\mathcal{F}} := \{s \mapsto (f_h - \mathcal{T}_h f_{h+1})(s, \pi_{f_h}(s)) : f \in \mathcal{F}\}.$$

*Let $\Pi = \{\Pi_h\}_{h=1}^H$ be a collection of H probability measure families over $\mathcal{S}$. The **Type-II ϵ-BE dimension** of $\mathcal{F}$ with respect to Π is defined as*

$$\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F}, \Pi, \epsilon) := \max_{h \in [H]} \dim_{\mathrm{DE}}\big((I - \mathcal{T}_h)V_{\mathcal{F}}, \Pi_h, \epsilon\big).$$

Relation with low Type-II Bellman rank Denote by $\mathcal{D}_{\mathcal{F},h}$ the collection of all probability measures over $\mathcal{S}$ at the h^{th} step, which can be generated by rolling in with a greedy policy π_f with $f \in \mathcal{F}$. Similar to Proposition 10, the following proposition claims that the BE dimension of $\mathcal{F}$ with respect to $\mathcal{D}_{\mathcal{F}} := \{\mathcal{D}_{\mathcal{F},h}\}_{h \in [H]}$ is always upper bounded by its Bellman rank up to some logarithmic factor.

Proposition 20 (low Type-II Bellman rank) *If an MDP with function class $\mathcal{F}$ has Type-II Bellman rank d with normalization parameter ζ, then*

$$\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \leq \mathcal{O}(1 + d \log(1 + \zeta/\epsilon)).$$

The proof of Proposition 20 is almost the same as that of Proposition 10 in Appendix B.1. We omit it here since the only modification is to replace Type-I Bellman rank with its Type-II variant wherever it's used.

Algorithm 3 Type-II OLIVE $(\mathcal{F}, \zeta_{\text{act}}, \zeta_{\text{elim}}, n_{\text{act}}, n_{\text{elim}})$

```
1:  **Initialize**: $\mathcal{B}^0 \leftarrow \mathcal{F}$, $\mathcal{D}_h \leftarrow \emptyset$ for all $h, k$.
2:  **for phase** $k = 1, 2, \ldots$ **do**
3:      **Choose policy** $\pi^k = \pi_{f^k}$, where $f^k = \operatorname{argmax}_{f \in \mathcal{B}^{k-1}} f(s_1, \pi_f(s_1))$.
4:      **Execute** $\pi^k$ for $n_{\text{act}}$ episodes and *refresh* $\mathcal{D}_h$ to include the fresh $(s_h, a_h, r_h, s_{h+1})$ tuples.
5:      **Estimate** $\tilde{\mathcal{E}}_{\text{II}}(f^k, \pi^k, h)$ for all $h \in [H]$, where
            $\tilde{\mathcal{E}}_{\text{II}}(g, \pi^k, h) = \frac{1}{|\mathcal{D}_h|} \sum_{(s,a,r,s') \in \mathcal{D}_h} \left( g_h(s,a) - r - \max_{a' \in \mathcal{A}} g_{h+1}(s', a') \right).$
6:      **if** $\sum_{h=1}^{H} \tilde{\mathcal{E}}_{\text{II}}(f^k, \pi^k, h) \le H\zeta_{\text{act}}$ **then**
7:          Terminate and output $\pi^k$.
8:      Pick any $t \in [H]$ for which $\tilde{\mathcal{E}}_{\text{II}}(f^k, \pi^k, t) > \zeta_{\text{act}}$.
9:      **Collect** $n_{\text{elim}}$ episodes by executing $\pi^k$ for step $1, \ldots, t-1$ and picking action uniform at
        random for step $t$. *Refresh* $\mathcal{D}_h$ to include the fresh $(s_h, a_h, r_h, s_{h+1})$ tuples.
10:     **Estimate** $\hat{\mathcal{E}}_{\text{II}}(f, \pi^k, t)$ for all $f \in \mathcal{F}$, where
            $\hat{\mathcal{E}}_{\text{II}}(g, \pi^k, h) = \frac{1}{|\mathcal{D}_h|} \sum_{(s,a,r,s') \in \mathcal{D}_h} \frac{\mathbf{1}[a = \pi_g(s)]}{1/|\mathcal{A}|} \left( g_h(s,a) - r - \max_{a' \in \mathcal{A}} g_{h+1}(s', a') \right).$
11:     **Update** $\mathcal{B}^k = \left\{ f \in \mathcal{B}^{k-1} : \left| \hat{\mathcal{E}}_{\text{II}}(f, \pi^k, t) \right| \le \zeta_{\text{elim}} \right\}.$
```

A.1. Algorithm Type-II OLIVE

In this section, we describe the original OLIVE (i.e., Type-II OLIVE) proposed by Jiang et al. (2017), and its theoretical guarantee in terms of Type-II BE dimension.

The pseudocode is provided in Algorithm 3. Its only difference from Algorithm 2 is Line 9-10: note that Type-II Bellman rank needs the action at step t to be greedy with respect to the function f instead of being picked by the roll-in policy π^k, so we choose action a_t uniformly at random and use the importance-weighted estimator to estimate the Bellman error for each f.

We have the following similar theoretical guarantee for Algorithm 3. Its proof is almost the same as that of Theorem 17 and can be found in Appendix E.1.

Theorem 21 (Type-II OLIVE) *Assume realizability (Assumption 1) holds and $\mathcal{F}$ is finite. There exists absolute constant c such that if we choose*

$$\zeta_{\text{act}} = \frac{2\epsilon}{H}, \; \zeta_{elim} = \frac{\epsilon}{2H\sqrt{d}}, \; n_{act} = \frac{H^2\iota}{\epsilon^2}, \; \textit{and } n_{elim} = \frac{H^2 d|\mathcal{A}|\log(|\mathcal{F}|)\cdot\iota}{\epsilon^2}$$

where $d = \dim_{\mathrm{BE}_{\mathrm{II}}}\big(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H\big)$ and $\iota = c\log[Hd|\mathcal{A}|/\delta\epsilon]$, then with probability at least $1-\delta$, Algorithm 3 will output an $\mathcal{O}(\epsilon)$-optimal policy using at most $\mathcal{O}(H^3 d^2 |\mathcal{A}| \log(|\mathcal{F}|)\cdot\iota/\epsilon^2)$ episodes.

For problems with Bellman rank d and finite function class $\mathcal{F}$, Theorem 21 together with Proposition 20 guarantees $\tilde{\mathcal{O}}(H^3 d^2 |\mathcal{A}|\log(|\mathcal{F}|)/\epsilon^2)$ samples suffice for finding an ϵ-optimal policy, which matches the result in Jiang et al. (2017). For function class $\mathcal{F}$ of infinite cardinality but with finite covering number, we can first compute an $\mathcal{O}(\zeta_{\text{elim}})$-cover of $\mathcal{F}$, which we denote as $\mathcal{Z}_\rho$, and then run Algorithm 3 on $\mathcal{Z}_\rho$. By following almost the same arguments in the proof of Theorem 21 (the only difference is to replace $Q^\star$ by its proxy in $\mathcal{Z}_\rho$), we can show Algorithm 3 will output an $\mathcal{O}(\epsilon)$-optimal policy using at most $\Omega(H^3 d^2 |\mathcal{A}|\log(Nt)/\epsilon^2)$ episodes where $N = \mathcal{N}_{\mathcal{F}}(\mathcal{O}(\zeta_{\text{elim}}))$.

Algorithm 4 Type-II GOLF $(\mathcal{F}, K, \beta)$

```
1: Initialize: $\mathcal{D}_1, \ldots, \mathcal{D}_H \leftarrow \emptyset$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
2: for epoch $k$ from 1 to $K$ do
3:   Choose policy $\pi^k = \pi_{f^k}$, where $f^k = \mathrm{argmax}_{f \in \mathcal{B}^{k-1}} f(s_1, \pi_f(s_1))$.
4:   for step $h$ from 1 to $H$ do
5:     Collect a tuple $(s_h, a_h, r_h, s_{h+1})$ by executing $\pi^k$ at step $1, \ldots, h-1$ and taking action
       uniformly at random at step $h$.
6:     Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(s_h, a_h, r_h, s_{h+1})\}$ for all $h \in [H]$.
7:   Update
       $$\mathcal{B}^k = \left\{ f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta \text{ for all } h \in [H] \right\},$$
       $$\text{where } \mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{(s,a,r,s') \in \mathcal{D}_h} [\xi_h(s,a) - r - \max_{a' \in \mathcal{A}} \zeta_{h+1}(s', a')]^2.$$
8: Output $\pi^{out}$ sampled uniformly at random from $\{\pi^k\}_{k=1}^K$.
```

A.2. Algorithm Type-II GOLF

In this section we describe the Type-II variant of GOLF. The pseudocode is provided in Algorithm 4. Its only difference from the Type-I analogue is in Line 5: for each $h \in [H]$, we roll in with policy π^k to sample s_h, and then instead of continuing following π^k we take random action at step h.

Now we present the theoretical guarantee for Algorithm 4. Its proof is almost the same as that of Corollary 15 and can be found in appendix E.2.

Theorem 22 (Type-II GOLF) *Assume realizability (Assumption 1) and completeness (Assumption 2) hold. There exists an absolute constant c_1, c_2 such that for any given $\epsilon > 0$, if we choose $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_\mathcal{F}\}} \dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F}, \Pi, \frac{\epsilon}{H})$, $K = c_1 H^2 d|\mathcal{A}| \log(Hd|\mathcal{A}|\mathcal{N}_\mathcal{F}(\frac{\epsilon^2}{d|\mathcal{A}|H^2})/\epsilon)/\epsilon^2$ and $\beta = c_1 \log[KH\mathcal{N}_\mathcal{F}(\frac{\epsilon^2}{d|\mathcal{A}|H^2})]$, then with probability at least 0.99, π^{out} is $c_2\epsilon$-optimal.*

Compared with Theorem 21 (type-II OLIVE), Theorem 22 (type-II GOLF) has the following two advantages.

- The sample complexity in Theorem 22 depends linearly on the type-II BE-dimension while the dependence in Theorem 21 is quadratic.
- Theorem 22 applies to RL problems of finite type-II BE dimension with respect to either $\mathcal{D}_\mathcal{F}$ or $\mathcal{D}_\Delta$. In comparison, Theorem 21 provides no guarantee for the $\mathcal{D}_\Delta$ case.

Appendix B. Proof for Section 3

In this section, we provide the formal proof for the results stated in Section 3.

B.1. Proof of Proposition 10

The proof is basically the same as that of Example 3 in Russo and Van Roy (2013) with minor modification.

Proof Without loss of generality, assume $\max\{\|\phi_h(f)\|_2, \|\psi_h(f)\|_2\} \le \sqrt{\zeta}$, otherwise we can satisfy this assumption by rescaling the feature mappings. Assume there exists $h \in [H]$ such that $\dim_{\mathrm{DE}}((I-\mathcal{T}_h)\mathcal{F},\mathcal{D}_{\mathcal{F},h},\epsilon)\ge m$. Let $\mu_1,\ldots,\mu_m\in\mathcal{D}_{\mathcal{F},h}$ be a an ϵ-independent sequence with respect to $(I-\mathcal{T}_h)\mathcal{F}$. By Definition 6, there exists $f^1,\ldots,f^m$ such that for all $i\in[m]$, $\sqrt{\sum_{t=1}^{i-1}(\mathbb{E}_{\mu_t}[f_h^i-\mathcal{T}_h f_{h+1}^i])^2}\le\epsilon$ and $|\mathbb{E}_{\mu_i}[f_h^i-\mathcal{T}_h f_{h+1}^i]|>\epsilon$. By the definition of Bellman rank, this is equivalent to: for all $i\in[m]$, $\sqrt{\sum_{t=1}^{i-1}(\langle\phi_h(f^i),\psi_h(f^t)\rangle)^2}\le\epsilon$ and $|\langle\phi_h(f^i),\psi_h(f^i)\rangle|>\epsilon$.

For notational simplicity, define $\mathbf{x}_i=\phi_h(f^i)$, $\mathbf{z}_i=\psi_h(f^i)$ and $\mathbf{V}_i=\sum_{t=1}^{i-1}\mathbf{z}_t\mathbf{z}_t^\top+\frac{\epsilon^2}{\zeta}\cdot\mathbf{I}$. The previous argument directly implies: for all $i\in[m]$, $\|\mathbf{x}_i\|_{\mathbf{V}_i}\le\sqrt{2}\epsilon$ and $\|\mathbf{x}_i\|_{\mathbf{V}_i}\cdot\|\mathbf{z}_i\|_{\mathbf{V}_i^{-1}}>\epsilon$. Therefore, we have $\|\mathbf{z}_i\|_{\mathbf{V}_i^{-1}}\ge\frac{1}{\sqrt{2}}$.

By the matrix determinant lemma,

$$\det[\mathbf{V}_m]=\det[\mathbf{V}_{m-1}](1+\|\mathbf{z}_m\|^2_{\mathbf{V}_m^{-1}})\ge\frac{3}{2}\det[\mathbf{V}_{m-1}]\ge\ldots\ge\det[\frac{\epsilon^2}{\zeta}\cdot\mathbf{I}](\frac{3}{2})^{m-1}=(\frac{\epsilon^2}{\zeta})^d(\frac{3}{2})^{m-1}.$$

On the other hand,

$$\det[\mathbf{V}_m]\le(\frac{\mathrm{trace}[\mathbf{V}_m]}{d})^d\le(\frac{\zeta(m-1)}{d}+\frac{\epsilon^2}{\zeta})^d.$$

Therefore, we obtain

$$(\frac{3}{2})^{m-1}\le(\frac{\zeta^2(m-1)}{d\epsilon^2}+1)^d.$$

Take logarithm on both sides,

$$m\le 4\left[1+d\log(\frac{\zeta^2(m-1)}{d\epsilon^2}+1)\right],$$

which, by simple calculation, implies

$$m\le\mathcal{O}\left(1+d\log(\frac{\zeta^2}{\epsilon^2}+1)\right).$$

■

B.2. Proof of Proposition 11

Proof Without loss of generality, assume m is even. Let $x_1,\ldots,x_m\in[0,\frac{1}{m}]$ be m mutually distinct numbers. Define $y_i=\sqrt{\frac{1}{m^2}-x_i^2}$. Consider the following linear bandits ($|\mathcal{S}|=H=1$) problem.

- The action set $\mathcal{A}=\{a_i=(1,x_i,y_i):\ i\in[m]\}$.
- The function set $\mathcal{F}_1=\{f_{\theta_i}(a)=(\langle a,\theta_i\rangle)^m:\ \theta_i=(1,x_i,y_i),\ i\in[m]\}$.
- The reward function is always zero, i.e., $r\equiv 0$.

Bellman rank First, note that the Bellman error $\mathcal{E}(f_\theta,\pi_{f_{\theta'}},h)$ is simply $(\langle\theta,\theta'\rangle)^m$ because (a) $\mathcal{T}_1\mathcal{F}_2=\{r\}$ (recall WLOG, we assume $\mathcal{F}_{H+1}=\{0\}$), (b) $r\equiv 0$, and (c) the greedy policy induced by $f_{\theta'}$ always picks action θ'. As a result, we only need to show the m by m matrix $\mathcal{E}:=(\Theta\Theta^\top)^{\odot m}\in\mathbb{R}^{m\times m}$ has rank m, where $\Theta=[\theta_1;\theta_2;\ldots;\theta_m]$ and $\odot m$ represents entry-wise power of m.

Define $z_i := \underbrace{\theta_i \otimes \cdots \otimes \theta_i}_{m \text{ times}} \in \mathbb{R}^{1\times 3^m}$ and $Z = [z_1; \ldots; z_m] \in \mathbb{R}^{m\times m}$, where $\otimes$ denotes Kronecker product. It is direct to see $\mathcal{E} = ZZ^\top$. Therefore, it suffices to show Z has rank m. Note that Z includes $V = [v_1; \ldots; v_m]$ as a submatrix, where $v_i = [C_m^0 1, C_m^1 x_i, C_m^2 x_i^2, \ldots, C_m^m x_i^m]$. Observe that V is essentially a rescaled Vandermonde matrix, so its rank is equal to m, which implies Z is also rank-m.

BE dimension First, note in this setting $\mathcal{D}_{\mathcal{F},1}$ is simply the collection of all Dirac distributions over $\mathcal{A}$, and $(I - \mathcal{T}_1)\mathcal{F}$ equals to $\mathcal{F}_1$ (because $\mathcal{F}_2 = \{0\}$ and $r \equiv 0$). So it suffices to show $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon) \le 9$.

Assume $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon) = k$. Then there exist $q_1, \ldots, q_k \in \mathcal{A}$ and $w_1, \ldots, w_k \in \mathcal{A}$ such that for all $i \in [k]$, $\sqrt{\sum_{t=1}^{i-1}(\langle q_i, w_t\rangle)^{2m}} \le \epsilon$ and $|\langle q_t, w_t\rangle|^m > \epsilon$. By simple calculation, we have $q_i^\top w_j \in [1, 1+\frac{1}{m}]$ for all $i, j \in [m]$. Therefore, if $\epsilon > e$, then $k = 0$ because $|\langle q_t, w_t\rangle|^m \le (1+\frac{1}{m})^m < e$; if $\epsilon \le e$, then $k < 10$ because $\sqrt{k-1} \le \sqrt{\sum_{t=1}^{k-1}(\langle q_i, w_k\rangle)^{2m}} \le \epsilon$. ■

Remark 23 *In this example, $Q^\star \equiv 0$ is not in $\mathcal{F}$. However, it is direct to see adding $Q^\star \equiv 0$ into $\mathcal{F}$ will not decrease the Bellman rank and will not increase the BE dimension.*

B.3. Proof of Proposition 12

Proof Assume $\delta_{z_1}, \ldots, \delta_{z_m}$ is an ϵ-independent sequence of distributions with respect to $(I - \mathcal{T}_h)\mathcal{F}$, where $\delta_{z_i} \in \mathcal{D}_\Delta$. By Definition 6, there exist functions $f^1, \ldots, f^m \in \mathcal{F}$ such that for all $i \in [m]$, we have $|(f_h^i - \mathcal{T}_h f_{h+1}^i)(z_i)| > \epsilon$ and $\sqrt{\sum_{t=1}^{i-1}|(f_h^i - \mathcal{T}_h f_{h+1}^i)(z_t)|^2} \le \epsilon$. Define $g_h^i = \mathcal{T}_h f_{h+1}^i$. Note that $g_h^i \in \mathcal{F}_h$ because $\mathcal{T}_h \mathcal{F}_{h+1} \subset \mathcal{F}_h$. Therefore, we have for all $i \in [m]$, $|(f_h^i - g_h^i)(z_i)| > \epsilon$ and $\sqrt{\sum_{t=1}^{i-1}|(f_h^i - g_h^i)(z_t)|^2} \le \epsilon$ with $f_h^i, g_h^i \in \mathcal{F}_h$. By Definition 4 and 5, this implies $\dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon) \ge m$, which completes the proof. ■

B.4. Proof of Proposition 13

Proof For any $m \in \mathbb{N}^+$, denote by $e_1, \ldots, e_m$ the basis vectors in $\mathbb{R}^m$, and consider the following linear bandits ($|\mathcal{S}| = H = 1$) problem.

- The action set $\mathcal{A} = \{a_i = (1; e_i) \in \mathbb{R}^{m+1} : \ i \in [m]\}$.
- The function set $\mathcal{F}_1 = \{f_{\theta_i}(a) = a^\top \theta_i : \ \theta_i = (1; e_i), \ i \in [m]\}$.
- The reward function is always zero, i.e., $r \equiv 0$.

Eluder dimension For any $\epsilon \in (0, 1]$, $a_1, \ldots, a_{m-1}$ is an ϵ-independent sequence of points because: (a) for any $t \in [m-1]$, $\sum_{i=1}^{t-1}(f_{\theta_t}(a_i) - f_{\theta_{t+1}}(a_i))^2 = 0$; (b) for any $t \in [m-1]$, $f_{\theta_t}(a_t) - f_{\theta_{t+1}}(a_t) = 1 \ge \epsilon$. Therefore, $\min_{h\in[H]} \dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon) = \dim_{\mathrm{E}}(\mathcal{F}_1, \epsilon) \ge m - 1$.

BE dimension First, note in this setting $(I - \mathcal{T}_1)\mathcal{F}$ is simply $\mathcal{F}_1$ (because $\mathcal{F}_2 = \{0\}$ and $r \equiv 0$). So it suffices to show $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon) \leq 5$.

Assume $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon) = k$. Then there exist $q_1, \ldots, q_k \in \mathcal{A}$ and $w_1, \ldots, w_k \in \mathcal{A}$ such that for all $t \in [k]$, $\sqrt{\sum_{i=1}^{t-1}(\langle q_t, w_i\rangle)^2} \leq \epsilon$ and $|\langle q_t, w_t\rangle| > \epsilon$. By simple calculation, we have $q_i^\top w_j \in [1, 2]$ for all $i, j \in [k]$. Therefore, if $\epsilon > 2$, then $k = 0$ because $|\langle q_t, w_t\rangle| \leq 2$; if $\epsilon \leq 2$, then $k \leq 5$ because $\sqrt{k-1} \leq \sqrt{\sum_{i=1}^{k-1}(\langle q_k, w_i\rangle)^2} \leq \epsilon$. ■

Appendix C. Proof for Section 4

In this section, we provide the formal proof for the results stated in Section 4.

C.1. Proof of Theorem 14

We start the analysis with the following two lemmas. The first lemma shows that with high probability any function in the confidence set has low Bellman-error over the collected datasets $\mathcal{D}_1, \ldots, \mathcal{D}_H$ as well as the distributions from which $\mathcal{D}_1, \ldots, \mathcal{D}_H$ are sampled.

Lemma 24 *Let $\rho > 0$ be an arbitrary fixed number. If we choose $\beta = c\big(\log[KH\mathcal{N}_{\mathcal{F}}(\rho)/\delta] + K\rho\big)$ with some large absolute constant c in Algorithm 1, then with probability at least $1-\delta$, for all $(k, h) \in [K] \times [H]$, we have*

(a) $\sum_{i=1}^{k-1} \mathbb{E}[\big(f_h^k(s_h, a_h) - (\mathcal{T} f_{h+1}^k)(s_h, a_h)\big)^2 \mid s_h, a_h \sim \pi^i] \leq \mathcal{O}(\beta)$.

(b) $\sum_{i=1}^{k-1} \big(f_h^k(s_h^i, a_h^i) - (\mathcal{T} f_{h+1}^k)(s_h^i, a_h^i)\big)^2 \leq \mathcal{O}(\beta)$,

where $(s_1^i, a_1^i, \ldots, s_H^i, a_H^i, s_{H+1}^i)$ denotes the trajectory sampled by following π^i in the i^{th} episode.

The second lemma guarantees that the optimal value function is inside the confidence with high probability. As a result, the selected value function f^k in each iteration shall be an upper bound of $Q^\star$ with high probability.

Lemma 25 *Under the same condition of Lemma 24, with probability at least $1-\delta$, we have $Q^\star \in \mathcal{B}^k$ for all $k \in [K]$.*

The proof of Lemma 24 and 25 relies on standard martingale concentration (e.g. Freedman's inequality) and can be found in Appendix C.3.

Step 1. Bounding the regret by Bellman error By Lemma 25, we can upper bound the cumulative regret by the summation of Bellman error with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\Big(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\Big) \leq \sum_{k=1}^{K}\Big(\max_a f_1^k(s_1, a) - V_1^{\pi^k}(s_1)\Big) \overset{(i)}{=} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h), \tag{6}$$

where (i) follows from standard policy loss decomposition (e.g. Lemma 1 in Jiang et al. (2017)).

Step 2. Bounding cumulative Bellman error using DE dimension Next, we focus on a fixed step h and bound the cumulative Bellman error $\sum_{k=1}^{K}\mathcal{E}(f^k,\pi^k,h)$ using Lemma 24. To proceed, we need the following lemma to control the accumulating rate of Bellman error.

Lemma 26 *Given a function class $\mathcal{G}$ defined on $\mathcal{X}$ with $|g(x)|\le C$ for all $(g,x)\in\mathcal{G}\times\mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{g_k\}_{k=1}^K\subset\mathcal{G}$ and $\{\mu_k\}_{k=1}^K\subset\Pi$ satisfy that for all $k\in[K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[g_k])^2\le\beta$. Then for all $k\in[K]$ and $\omega>0$,*

$$\sum_{t=1}^{k}|\mathbb{E}_{\mu_t}[g_t]|\le\mathcal{O}\left(\sqrt{\dim_{\mathrm{DE}}(\mathcal{G},\Pi,\omega)\beta k}+\min\{k,\dim_{\mathrm{DE}}(\mathcal{G},\Pi,\omega)\}C+k\omega\right).$$

Lemma 26 is a simple modification of Lemma 2 in Russo and Van Roy (2013) and its proof can be found in Appendix C.4. We provide two ways to apply Lemma 26, which can produce regret bounds in term of two different complexity measures. If we invoke Lemma 24 (a) and Lemma 26 with

$$\begin{cases}\rho=\frac{1}{K},\ \omega=\sqrt{\frac{1}{K}},\ C=1,\\ \mathcal{X}=\mathcal{S}\times\mathcal{A},\ \mathcal{G}=(I-\mathcal{T}_h)\mathcal{F},\ \Pi=\mathcal{D}_{\mathcal{F},h},\\ g_k=f_h^k-\mathcal{T}_hf_{h+1}^k \text{ and } \mu_k=\mathbb{P}^{\pi^k}(s_h=\cdot,a_h=\cdot),\end{cases}$$

we obtain

$$\sum_{t=1}^{k}\mathcal{E}(f^t,\pi^t,h)\le\mathcal{O}\left(\sqrt{k\cdot\dim_{\mathrm{BE}}(\mathcal{F},\mathcal{D}_{\mathcal{F}},\sqrt{1/K})\log[KH\mathcal{N}_{\mathcal{F}}(1/K)/\delta]}\right). \tag{7}$$

We can also invoke Lemma 24 (b) and Lemma 26 with

$$\begin{cases}\rho=\frac{1}{K},\ \omega=\sqrt{\frac{1}{K}},\ C=1,\\ \mathcal{X}=\mathcal{S}\times\mathcal{A},\ \mathcal{G}=(I-\mathcal{T}_h)\mathcal{F},\ \text{and } \Pi=\mathcal{D}_{\Delta,h},\\ g_k=f_h^k-\mathcal{T}_hf_{h+1}^k \text{ and } \mu_k=\mathbf{1}\{\cdot=(s_h^k,a_h^k)\},\end{cases}$$

and obtain

$$\begin{aligned}\sum_{t=1}^{k}\mathcal{E}(f^t,\pi^t,h)\le&\sum_{t=1}^{k}(f_h^t-\mathcal{T}f_{h+1}^t)(s_h^t,a_h^t)+\mathcal{O}\left(\sqrt{k\log(k)}\right)\\ \le&\mathcal{O}\left(\sqrt{k\cdot\dim_{\mathrm{BE}}(\mathcal{F},\mathcal{D}_{\Delta},\sqrt{1/K})\log[KH\mathcal{N}_{\mathcal{F}}(1/K)/\delta]}\right),\end{aligned} \tag{8}$$

where the first inequality follows from standard martingale concentration.

Plugging either equation (7) or (8) back into equation (6) completes the proof.

C.2. Proof of Corollary 15

Step 1. Bounding the regret by Bellman error By Lemma 25, we can upper bound the cumulative regret by the summation of Bellman error with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\left(V_1^\star(s_1)-V_1^{\pi^k}(s_1)\right)\le\sum_{k=1}^{K}\left(\max_a f_1^k(s_1,a)-V_1^{\pi^k}(s_1)\right)\overset{(i)}{=}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathcal{E}(f^k,\pi^k,h), \tag{9}$$

where (i) follows from standard policy loss decomposition (e.g. Lemma 1 in Jiang et al. (2017)).

Step 2. Bounding cumulative Bellman error using DE dimension Next, we focus on a fixed step h and bound the cumulative Bellman error $\sum_{k=1}^{K} \mathcal{E}(f^k, \pi^k, h)$ using Lemma 24.

If we invoke Lemma 24 (a) with

$$\rho = \frac{\epsilon^2}{H^2 \cdot \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H)},$$

and Lemma 26 with

$$\begin{cases} \omega = \frac{\epsilon}{H},\ C = 1, \\ \mathcal{X} = \mathcal{S} \times \mathcal{A},\ \mathcal{G} = (I - \mathcal{T}_h)\mathcal{F},\ \Pi = \mathcal{D}_{\mathcal{F},h}, \\ g_k = f_h^k - \mathcal{T}_h f_{h+1}^k \text{ and } \mu_k = \mathbb{P}^{\pi^k}(s_h = \cdot, a_h = \cdot), \end{cases}$$

we obtain with probability at least $1 - 10^{-3}$,

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}(f^k, \pi^k, h) \leq & \mathcal{O}\left(\sqrt{\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H)[\frac{\log[KH\mathcal{N}_{\mathcal{F}}(\rho)]}{K} + \rho]} + \frac{\epsilon}{H}\right) \\ \leq & \mathcal{O}\left(\frac{\epsilon}{H} + \sqrt{\frac{d\log[KH\mathcal{N}_{\mathcal{F}}(\rho)]}{K}}\right), \end{aligned} \tag{10}$$

where the second inequality follows from the choice of ρ and $d := \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H)$. Now we need to choose K such that

$$\sqrt{\frac{d\log[KH\mathcal{N}_{\mathcal{F}}(\rho)]}{K}} \leq \frac{\epsilon}{H}. \tag{11}$$

By simple calculation, one can verify it suffices to choose

$$K = \frac{H^2 d \log(Hd\mathcal{N}_{\mathcal{F}}(\rho)/\epsilon)}{\epsilon^2}. \tag{12}$$

Plugging equation (10) back into equation (9) completes the proof. We can similarly prove the bound in terms of the BE dimension with respect to $\mathcal{D}_\Delta$.

C.3. Proof of concentration lemmas

To begin with, recall the Freedman's inequality that controls the sum of martingale difference by the sum of their predicted variance.

Lemma 27 (Freedman's inequality (e.g., Agarwal et al. (2014))) *Let $(Z_t)_{t\leq T}$ be a real-valued martingale difference sequence adapted to filtration $\mathfrak{F}_t$, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathfrak{F}_t]$. If $|Z_t| \leq R$ almost surely, then for any $\eta \in (0, \frac{1}{R})$ it holds that with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} Z_t \leq \mathcal{O}\left(\eta\sqrt{\sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t^2]} + \frac{R\log(\delta^{-1})}{\eta}\right).$$

C.3.1. PROOF OF LEMMA 24

Proof We prove inequality (b) first.

Consider a fixed (k, h, f) tuple. Let

$$X_t(h,f) := (f(s_h^t, a_h^t) - r_h^t - f(s_{h+1}^t, \pi_f(s_{h+1}^t)))^2 - (\mathcal{T}f(s_h^t, a_h^t) - r_h^t - f(s_{h+1}^t, \pi_f(s_{h+1}^t)))^2$$

and $\mathfrak{F}_{t,h}$ be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1} \bigcup \{s_1^t, a_1^t, r_1^t, \ldots, s_h^t, a_h^t\}$. We have

$$\mathbb{E}[X_t(h,f) \mid \mathfrak{F}_{t,h}] = [(f - \mathcal{T}f)(s_h^t, a_h^t)]^2$$

and

$$\text{Var}[X_t(h,f) \mid \mathfrak{F}_{t,h}] \le \mathbb{E}[(X_t(h,f))^2 \mid \mathfrak{F}_{t,h}] \le 36[(f - \mathcal{T}f)(s_h^t, a_h^t)]^2 = 36\mathbb{E}[X_t(h,f) \mid \mathfrak{F}_{t,h}].$$

By Freedman's inequality, we have, with probability at least $1-\delta$,

$$\left|\sum_{t=1}^k X_t(h,f) - \sum_{t=1}^k \mathbb{E}[X_t(h,f) \mid \mathfrak{F}_{t,h}]\right| \le \mathcal{O}\left(\sqrt{\sum_{t=1}^k \mathbb{E}[X_t \mid \mathfrak{F}_{t,h}]} + \log(1/\delta)\right).$$

Let $\mathcal{Z}_\rho$ be a ρ-cover of $\mathcal{F}$. Now taking a union bound for all $(k,h,g) \in [K] \times [H] \times \mathcal{Z}_\rho$, we obtain that with probability at least $1-\delta$, for all $(k,h,g) \in [K] \times [H] \times \mathcal{Z}_\rho$

$$\left|\sum_{t=1}^k X_t(h,g) - \sum_{t=1}^k \mathbb{E}[(g - \mathcal{T}g)(s_h^t, a_h^t)]^2\right| \le \mathcal{O}\left(\sqrt{\sum_{t=1}^k [(g - \mathcal{T}g)(s_h^t, a_h^t)]^2} + \iota\right), \tag{13}$$

where $\iota = \log(HK|\mathcal{Z}_\rho|/\delta)$. From now on, we will do all the analysis conditioning on this event being true.

Consider an arbitrary $(h,k) \in [H] \times [K]$ pair. By the definition of $\mathcal{B}^k$ and $f^k \in \mathcal{B}^k$,

$$\begin{aligned}
\sum_{t=1}^{k-1} X_t(h, f^k) =& \sum_{t=1}^{k-1} (f^k(s_h^t, a_h^t) - r_h^t - f^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t)))^2 \\
&- \sum_{t=1}^{k-1} (\mathcal{T}f^k(s_h^t, a_h^t) - r_h^t - f^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t)))^2 \\
\le& \sum_{t=1}^{k-1} (f^k(s_h^t, a_h^t) - r_h^t - f^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t)))^2 \\
&- \inf_{g \in \mathcal{F}} \sum_{t=1}^{k-1} (g(s_h^t, a_h^t) - r_h^t - f^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t)))^2 \le \mathcal{O}(\iota + k\rho).
\end{aligned}$$

Define $g^k = \text{argmin}_{g \in \mathcal{Z}_\rho} \max_{h \in [H]} \|f_h^k - g_h^k\|_\infty$. By the definition of $\mathcal{Z}_\rho$, we have

$$\left|\sum_{t=1}^{k-1} X_t(h, f^k) - \sum_{t=1}^{k-1} X_t(h, g^k)\right| \le \mathcal{O}(k\rho).$$

Therefore,

$$\sum_{t=1}^{k-1} X_t(h, g^k) \leq \mathcal{O}(\iota + k\rho). \tag{14}$$

Recall inequality (13) implies

$$\left|\sum_{t=1}^{k-1} X_t(h, g^k) - \sum_{t=1}^{k-1}[(g^k - \mathcal{T}g^k)(s_h^t, a_h^t)]^2\right| \leq \mathcal{O}\left(\sqrt{\sum_{t=1}^{k-1}[(g^k - \mathcal{T}g^k)(s_h^t, a_h^t)]^2 + \iota}\right). \tag{15}$$

Putting (14) and (15) together, we obtain

$$\sum_{t=1}^{k-1}[(g^k - \mathcal{T}g^k)(s_h^t, a_h^t)]^2 \leq \mathcal{O}(\iota + k\rho).$$

Because g^k is an ρ-approximation to f^k, we conclude

$$\sum_{t=1}^{k-1}[(f^k - \mathcal{T}f^k)(s_h^t, a_h^t)]^2 \leq \mathcal{O}(\iota + k\rho).$$

Therefore, we prove inequality (b) in Lemma 24.

To prove inequality (a), we only need to redefine $\mathfrak{F}_{t,h}$ to be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1}$ and then repeat the arguments above verbatim. ■

C.3.2. PROOF OF LEMMA 25

Proof Consider an arbitrary fixed (k, h, f) tuple. Let

$W_t(h, f) := (f(s_h^t, a_h^t) - r_h^t - Q^\star(s_{h+1}^t, \pi_{Q^\star}(s_{h+1}^t)))^2 - (Q^\star(s_h^t, a_h^t) - r_h^t - Q^\star(s_{h+1}^t, \pi_{Q^\star}(s_{h+1}^t)))^2$

and $\mathfrak{F}_{t,h}$ be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1} \bigcup \{s_1^t, a_1^t, r_1^t, \ldots, s_h^t, a_h^t\}$. We have

$$\mathbb{E}[W_t(h, f) \mid \mathfrak{F}_{t,h}] = [(f - Q^\star)(s_h^t, a_h^t)]^2$$

and

$$\mathrm{Var}[W_t(h, f) \mid \mathfrak{F}_{t,h}] \leq \mathbb{E}[(W_t(h, f))^2 \mid \mathfrak{F}_{t,h}] \leq 36((f - Q^\star)(s_h^t, a_h^t))^2 = 36\mathbb{E}[W_t(h, f) \mid \mathfrak{F}_{t,h}].$$

By Freedman's inequality, with probability at least $1 - \delta$,

$$\left|\sum_{t=1}^{k} W_t(h, f) - \sum_{t=1}^{k}[(f - Q^\star)(s_h^t, a_h^t)]^2\right| \leq \mathcal{O}\left(\sqrt{\sum_{t=1}^{k}[(f - Q^\star)(s_h^t, a_h^t)]^2 + \log(1/\delta)}\right).$$

Taking a union bound over $[K] \times [H] \times \mathcal{Z}_\rho$ and note $\sum_{t=1}^{k}[(f - Q^\star)(s_h^t, a_h^t)]^2$ being nonnegative, we obtain that with probability at least $1 - \delta$, for all $(k, h, g) \in [K] \times [H] \times \mathcal{Z}_\rho$

$$-\sum_{t=1}^{k} W_t(h, g) \leq \mathcal{O}(\iota),$$

where $\iota = \log(HK|\mathcal{Z}_\rho|/\delta)$. This directly implies for all $(k, h, f) \in [K] \times [H] \times \mathcal{F}$

$$\begin{aligned}&\sum_{t=1}^{k-1}(Q^\star(s_h^t, a_h^t) - r_h^t - Q^\star(s_{h+1}^t, \pi_{Q^\star}(s_{h+1}^t)))^2 \\ \leq &\sum_{t=1}^{k-1}(f(s_h^t, a_h^t) - r_h^t - Q^\star(s_{h+1}^t, \pi_{Q^\star}(s_{h+1}^t)))^2 + \mathcal{O}(\iota + k\rho).\end{aligned}$$

Finally, by recalling the definition of $\mathcal{B}^k$, we conclude that with probability at least $1 - \delta$, $Q^\star \in \mathcal{B}^k$ for all $k \in [K]$. ■

C.4. Proof of Lemma 26 (modification of Appendix C in Russo and Van Roy (2013))

We first prove the following proposition which bounds the number of times $|\mathbb{E}_{\mu_t}[g_t]|$ can exceed a certain threshold.

Proposition 28 *Given a function class $\mathcal{G}$ defined on $\mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{g_k\}_{k=1}^K \subset \mathcal{G}$ and $\{\mu_k\}_{k=1}^K \subset \Pi$ satisfy that for all $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[g_k])^2 \leq \beta$. Then for all $k \in [K]$,*

$$\sum_{t=1}^{k} \mathbf{1}\{|\mathbb{E}_{\mu_t}[g_t]| > \epsilon\} \leq (\frac{\beta}{\epsilon^2} + 1)\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon).$$

Proof [Proof of Proposition 28] We first show that if for some k we have $|\mathbb{E}_{\mu_k}[g_k]| > \epsilon$, then μ_k is ϵ-independent of at most β/ϵ^2 disjoint subsequences in $\{\mu_1, \ldots, \mu_{k-1}\}$. By definition of DE dimension, if $|\mathbb{E}_{\mu_k}[g_k]| > \epsilon$ and μ_k is ϵ-dependent on a subsequence $\{\nu_1, \ldots, \nu_\ell\}$ of $\{\mu_1, \ldots, \mu_{k-1}\}$, then we should have $\sum_{t=1}^{\ell}(\mathbb{E}_{\nu_t}[g_k])^2 \geq \epsilon^2$. It implies that if μ_k is independent of L disjoint subsequences in $\{\mu_1, \ldots, \mu_{k-1}\}$, we have

$$\beta \geq \sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[g_k])^2 \geq L\epsilon^2$$

resulting in $L \leq \beta/\epsilon^2$.

Now we want to show that for any sequence $\{\nu_1, \ldots, \nu_\kappa\} \subseteq \Pi$, there exists $j \in [\kappa]$ such that ν_j is ϵ-dependent on at least $L = \lceil(\kappa - 1)/\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)\rceil$ disjoint subsequences in $\{\nu_1, \ldots, \nu_{j-1}\}$. We argue by the following mental procedure: we start with singleton sequences $B_1 = \{\nu_1\}, \ldots, B_L = \{\nu_L\}$ and $j = L + 1$. For each j, if ν_j is ϵ-dependent on $B_1, \ldots, B_L$ we already achieved our goal so we stop; otherwise, we pick an $i \in [L]$ such that ν_j is ϵ-independent of B_i and update $B_i = B_i \cup \{\nu_j\}$. Then we increment j by 1 and continue this process. By the definition of DE dimension, the size of each $B_1, \ldots, B_L$ cannot get bigger than $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ at any point in this process. Therefore, the process stops before or on $j = L\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon) + 1 \leq \kappa$.

Fix $k \in [K]$ and let $\{\nu_1, \ldots, \nu_\kappa\}$ be subsequence of $\{\mu_1, \ldots, \mu_k\}$, consisting of elements for which $|\mathbb{E}_{\mu_t}[g_t]| > \epsilon$. Using the first claim, we know that each ν_j is ϵ-dependent on at most β/ϵ^2

disjoint subsequences of $\{\nu_1, \ldots, \nu_{j-1}\}$. Using the second claim, we know there exists $j \in [\kappa]$ such that ν_j is ϵ-dependent on at least $(\kappa / \dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)) - 1$ disjoint subsequences of $\{\nu_1, \ldots, \nu_{j-1}\}$. Therefore, we have $(\kappa - 1)/\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon) \leq \beta/\epsilon^2$ which results in

$$\kappa \leq (\frac{\beta}{\epsilon^2} + 1) \dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$$

and completes the proof. ■

Proof [Proof of Lemma 26] Fix $k \in [K]$; let $d = \dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \delta)$. Sort the sequence $\{|\mathbb{E}_{\mu_1}[g_1]|, \ldots, |\mathbb{E}_{\mu_k}[g_k]|\}$ in a decreasing order and denote it by $\{e_1, \ldots, e_k\}$ $(e_1 \geq e_2 \geq \cdots \geq e_k)$.

$$\sum_{t=1}^{k} |\mathbb{E}_{\mu_t}[g_t]| = \sum_{t=1}^{k} e_t = \sum_{t=1}^{k} e_t \mathbf{1}\{e_t \leq \delta\} + \sum_{t=1}^{k} e_t \mathbf{1}\{e_t > \delta\} \leq k\delta + \sum_{t=1}^{k} e_t \mathbf{1}\{e_t > \delta\}.$$

For $t \in [k]$, we want to prove that if $e_t > \delta$, then we have $e_t \leq \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Assume $t \in [k]$ satisfies $e_t > \delta$. Then there exists α such that $e_t > \alpha \geq \delta$. By Proposition 28, we have

$$t \leq \sum_{i=1}^{k} \mathbf{1}\{e_i > \alpha\} \leq \big(\frac{\beta}{\alpha^2} + 1\big) \dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \alpha) \leq \big(\frac{\beta}{\alpha^2} + 1\big) \dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \delta),$$

which implies $\alpha \leq \sqrt{\frac{d\beta}{t-d}}$. Besides, recall $e_t \leq C$, so we have $e_t \leq \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$.

Finally, we have

$$\sum_{t=1}^{k} e_t \mathbf{1}\{e_t > \delta\} \leq \min\{d, k\} C + \sum_{t=d+1}^{k} \sqrt{\frac{d\beta}{t-d}} \leq \min\{d, k\} C + \sqrt{d\beta} \int_0^k \frac{1}{\sqrt{t}} dt$$
$$\leq \min\{d, k\} C + 2\sqrt{d\beta k},$$

which completes the proof. ■

Appendix D. Proof for Section 5

In this section, we provide the formal proof for the results stated in Section 5.

D.1. Full proof of Theorem 17

Proof [Proof of Theorem 17] By standard concentration arguments (Hoeffding's inequality plus union bound argument), with probability at least $1-\delta$, the following events hold for the first $dH+1$ phases (please refer to Appendix D.2 for the proof)

1. If the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\mathcal{E}(f^k, \pi^k, h) > \zeta_{\text{act}}/2$ and all $f \in \mathcal{F}$ satisfying $|\mathcal{E}(f, \pi^k, h)| \geq 2\zeta_{\text{elim}}$ get eliminated.
2. If the elimination procedure is not activated in the k^{th} phase, then, $\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h) < 2H\zeta_{\text{act}} = 4\epsilon$.

3. $Q^\star$ is not eliminated.

Therefore, if we can show OLIVE terminates within $dH+1$ phases, then with high probability the output policy is 4ϵ-optimal by the optimism of f^k and simple policy loss decomposition (e.g. Lemma 1 in Jiang et al. (2017)):

$$\Big(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\Big) \le \max_a f^k(s_1,a) - V^{\pi^k}(s_1) = \sum_{h=1}^{H} \mathcal{E}(f^k,\pi^k,h) \le 4\epsilon. \tag{16}$$

In order to prove that OLIVE terminates within $dH+1$ phases, it suffices to show that for each $h \in [H]$, we can activate the elimination procedure at the h^{th} step for at most d times.

For the sake of contradiction, assume that OLIVE does not terminate in $dH+1$ phases. Within these $dH+1$ phases, there exists some $h \in [H]$ for which the activation process has been activated for at least $d+1$ times. Denote by $k_1 < \cdots < k_{d+1} \le dH+1$ the indices of the phases where the elimination is activated at the h^{th} step. By the high-probability events, for all $i < j \le d+1$, we have $|\mathcal{E}(f^{k_j},\pi^{k_i},h)| < 2\zeta_{\text{elim}}$ and for all $l \le d+1$, we have $\mathcal{E}(f^{k_l},\pi^{k_l},h) > \zeta_{\text{act}}/2$. This means for all $l \le d+1$, we have both $\sqrt{\sum_{i=1}^{l-1}\big(\mathcal{E}(f^{k_l},\pi^{k_i},h)\big)^2} < \sqrt{d} \times 2\zeta_{\text{elim}} = \epsilon/H$ and $\mathcal{E}(f^{k_l},\pi^{k_l},h) > \zeta_{\text{act}}/2 = \epsilon/H$. Therefore, the roll-in distribution of $\pi^{k_1},\ldots,\pi^{k_{d+1}}$ at step h is an ϵ/H-independent sequence of length $d+1$, which contradicts with the definition of BE dimension. So OLIVE should terminate within $dH+1$ phases.

In sum, with probability at least $1-\delta$, Algorithm 2 will terminate and output a 4ϵ-optimal policy using at most

$$(dH+1)(n_{\text{act}}+n_{\text{elim}}) \le \frac{3cH^3d^2\log(\mathcal{N}(\mathcal{F},\zeta_{\text{elim}}/8))\cdot\iota}{\epsilon^2}$$

episodes. ■

D.2. Concentration arguments for Theorem 17

Recall in Algorithm 2 we choose

$$\zeta_{\text{act}} = \frac{2\epsilon}{H},\ \zeta_{\text{elim}} = \frac{\epsilon}{2H\sqrt{d}},\ n_{\text{act}} = \frac{cH^2\iota}{\epsilon^2},\ \text{and } n_{\text{elim}} = \frac{cH^2d\log(\mathcal{N}(\mathcal{F},\zeta_{\text{elim}}/8))\cdot\iota}{\epsilon^2},$$

where $d = \max_{h\in[H]} \dim_{\text{BE}}\big(\mathcal{F},\mathcal{D}_{\mathcal{F},h},\epsilon/H\big)$, $\iota = \log[Hd/\delta\epsilon]$ and c is a large absolute constant.Our goal is to prove with probability at least $1-\delta$, the following events hold for the first $dH+1$ phases

1. If the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\mathcal{E}(f^k,\pi^k,h) > \zeta_{\text{act}}/2$ and all $f \in \mathcal{F}$ satisfying $|\mathcal{E}(f,\pi^k,h)| \ge 2\zeta_{\text{elim}}$ get eliminated.

2. If the elimination procedure is not activated in the k^{th} phase, then, $\sum_{h=1}^H \mathcal{E}(f^k,\pi^k,h) < 2H\zeta_{\text{act}} = 4\epsilon$.

3. $Q^\star$ is not eliminated.

We begin with the activation procedure.

Concentration in the activation procedure Consider a fixed $(k,h) \in [dH+1] \times [H]$ pair. By Azuma-Hoefdding's inequality, with probability at least $1 - \frac{\delta}{8H(dH^2+1)}$, we have

$$|\hat{\mathcal{E}}(f^k, \pi^k, h) - \mathcal{E}(f^k, \pi^k, h)| \le \mathcal{O}\left(\sqrt{\frac{\iota}{n_{\text{act}}}}\right) \le \frac{\epsilon}{2H} \le \zeta_{\text{act}}/4,$$

where the second inequality follows from $n_{\text{act}} = C\frac{H^2\iota}{\epsilon^2}$ with C being chosen large enough.

Take a union bound for all $(k,h) \in [dH+1] \times [H]$, we have with probability at least $1 - \delta/4$, the following holds for all $(k,h) \in [dH+1] \times [H]$

$$|\hat{\mathcal{E}}(f^k, \pi^k, h) - \mathcal{E}(f^k, \pi^k, h)| \le \zeta_{\text{act}}/4.$$

By Algorithm 2, if the elimination procedure is not activated in the k^{th} phase, we have $\sum_{h=1}^{H} \hat{\mathcal{E}}(f^k, \pi^k, h) \le H\zeta_{\text{act}}$. Combine it with the concentration argument we just proved,

$$\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h) \le \sum_{h=1}^{H} \hat{\mathcal{E}}(f^k, \pi^k, h) + \frac{H\zeta_{\text{act}}}{4} < \frac{5H\zeta_{\text{act}}}{4}.$$

On the other hand, if the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\hat{\mathcal{E}}(f^k, \pi^k, h) > \zeta_{\text{act}}$. Again combine it with the concentration argument we just proved,

$$\mathcal{E}(f^k, \pi^k, h) \ge \hat{\mathcal{E}}(f^k, \pi^k, h) - \frac{\zeta_{\text{act}}}{4} > \frac{3\zeta_{\text{act}}}{4}.$$

Concentration in the elimination procedure Now, let us turn to the elimination procedure. First, let $\mathcal{Z}$ be an $\zeta_{\text{elim}}/8$-cover of $\mathcal{F}$ with cardinality $\mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8)$. With a little abuse of notation, for every $f \in \mathcal{F}$, define $\hat{f} = \operatorname{argmin}_{g\in\mathcal{Z}} \max_{h\in[H]} \|f_h - g_h\|_\infty$. By applying Azuma-Hoeffding's inequality to all $(k,g) \in [dH+1] \times \mathcal{Z}$ and taking a union bound, we have with probability at least $1 - \delta/4$, the following holds for all $(k,g) \in [dH+1] \times \mathcal{Z}$

$$|\hat{\mathcal{E}}(g, \pi^k, h_k) - \mathcal{E}(g, \pi^k, h_k)| \le \zeta_{\text{elim}}/4.$$

Recall that Algorithm 2 eliminates all f satisfying $|\hat{\mathcal{E}}(f, \pi^k, h_k)| > \zeta_{\text{elim}}$ when the elimination procedure is activated at the h_k^{th} step in the k^{th} phase. Therefore, if $|\mathcal{E}(f, \pi^k, h_k)| \ge 2\zeta_{\text{elim}}$, f will be eliminated because

$$\begin{aligned}
|\hat{\mathcal{E}}(f, \pi^k, h_k)| &\ge |\hat{\mathcal{E}}(\hat{f}, \pi^k, h_k)| - 2 \times \frac{\zeta_{\text{elim}}}{8} \\
&\ge |\mathcal{E}(\hat{f}, \pi^k, h_k)| - \frac{\zeta_{\text{elim}}}{2} \\
&\ge |\mathcal{E}(f, \pi^k, h_k)| - \frac{\zeta_{\text{elim}}}{2} - 2 \times \frac{\zeta_{\text{elim}}}{8} > \zeta_{\text{elim}}.
\end{aligned}$$

Finally, note that $\mathcal{E}(Q^\star, \pi, h) \equiv 0$ for any π and h. As a result, it will never be eliminated within the first $dH+1$ phases because we can similarly prove

$$|\hat{\mathcal{E}}(Q^\star, \pi^k, h_k)| \le |\mathcal{E}(Q^\star, \pi^k, h_k)| + \frac{3\zeta_{\text{elim}}}{4} < \zeta_{\text{elim}}.$$

Wrapping up: take a union bound for the activation and elimination procedure, and conclude that the three events, listed at the beginning of this section, hold for the the first $dH+1$ phases with probability at least $1 - \delta/2$.

Appendix E. Proof for Appendix A

In this section, we provide the formal proof for the results stated in Section A.

E.1. Proof of Theorem 21 (similar to Appendix D)

Proof [Proof of Theorem 21] By standard concentration arguments (Hoeffding's inequality, Bernstein's inequality, and union bound argument), with probability at least $1-\delta$, the following events hold for the first $dH+1$ phases (please refer to Appendix E.1.1 for the proof)

1. If the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\mathcal{E}_{\text{II}}(f^k,\pi^k,h) > \zeta_{\text{act}}/2$ and all $f\in\mathcal{F}$ satisfying $|\mathcal{E}_{\text{II}}(f,\pi^k,h)|\ge 2\zeta_{\text{elim}}$ get eliminated.
2. If the elimination procedure is not activated in the k^{th} phase, then, $\sum_{h=1}^{H}\mathcal{E}_{\text{II}}(f^k,\pi^k,h) < 2H\zeta_{\text{act}} = 4\epsilon$.
3. $Q^\star$ is not eliminated.

Therefore, if we can show OLIVE terminates within $dH+1$ phases, then with high probability the output policy is 4ϵ-optimal by the optimism of f^k and simple policy loss decomposition (e.g., Lemma 1 in Jiang et al. (2017)):

$$\left(V_1^\star(s_1)-V_1^{\pi^k}(s_1)\right)\le \max_a f^k(s_1,a)-V^{\pi^k}(s_1)=\sum_{h=1}^{H}\mathcal{E}_{\text{II}}(f^k,\pi^k,h)\le 4\epsilon. \tag{17}$$

In order to prove that OLIVE terminates within $dH+1$ phases, it suffices to show that for each $h\in[H]$, we can activate the elimination procedure at the h^{th} step for at most d times.

For the sake of contradiction, assume that OLIVE does not terminate in $dH+1$ phases. Within these $dH+1$ phases, there exists some $h\in[H]$ for which the activation process has been activated for at least $d+1$ times. Denote by $k_1<\cdots<k_{d+1}\le dH+1$ the indices of the phases where the elimination is activated at the h^{th} step. By the high-probability events, for all $i<j\le d+1$, we have $|\mathcal{E}_{\text{II}}(f^{k_j},\pi^{k_i},h)|<2\zeta_{\text{elim}}$ and for all $l\le d+1$, we have $\mathcal{E}_{\text{II}}(f^{k_l},\pi^{k_l},h)>\zeta_{\text{act}}/2$. This means for all $l\le d+1$, we have both $\sqrt{\sum_{i=1}^{l-1}\left(\mathcal{E}_{\text{II}}(f^{k_l},\pi^{k_i},h)\right)^2}<\sqrt{d}\times 2\zeta_{\text{elim}}=\epsilon/H$ and $\mathcal{E}_{\text{II}}(f^{k_l},\pi^{k_l},h)>\zeta_{\text{act}}/2=\epsilon/H$. Therefore, the roll-in distribution of $\pi^{k_1},\ldots,\pi^{k_{d+1}}$ at step h is an ϵ/H-independent sequence of length $d+1$ with respect to $(I-\mathcal{T}_h)V_{\mathcal{F}}$, which contradicts with the definition of BE dimension. So OLIVE should terminate within $dH+1$ phases.

In sum, with probability at least $1-\delta$, Algorithm 2 will terminate and output a 4ϵ-optimal policy using at most

$$(dH+1)(n_{\text{act}}+n_{\text{elim}})\le\frac{3cH^3d^2|\mathcal{A}|\log(|\mathcal{F}|)\cdot\iota}{\epsilon^2}$$

episodes. ■

E.1.1. CONCENTRATION ARGUMENTS FOR THEOREM 21

Recall in Algorithm 3 we choose

$$\zeta_{\text{act}}=\frac{2\epsilon}{H},\ \zeta_{\text{elim}}=\frac{\epsilon}{2H\sqrt{d}},\ n_{\text{act}}=\frac{cH^2\iota}{\epsilon^2},\ \text{and } n_{\text{elim}}=\frac{c|\mathcal{A}|H^2d\log(\mathcal{N}(\mathcal{F},\zeta_{\text{elim}}/8))\cdot\iota}{\epsilon^2},$$

where $d = \max_{h\in[H]} \dim_{\mathrm{BE_{II}}}\big(\mathcal{F}, \mathcal{D}_{\mathcal{F},h}, \epsilon/H\big)$, $\iota = \log[Hd/\delta\epsilon]$ and c is a large absolute constant. Our goal is to prove with probability at least $1-\delta$, the following events hold for the first $dH+1$ phases

1. If the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h) > \zeta_{\text{act}}/2$ and all $f \in \mathcal{F}$ satisfying $|\mathcal{E}_{\mathrm{II}}(f, \pi^k, h)| \ge 2\zeta_{\text{elim}}$ get eliminated.
2. If the elimination procedure is not activated in the k^{th} phase, then, $\sum_{h=1}^{H} \mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h) < 2H\zeta_{\text{act}} = 4\epsilon$.
3. $Q^\star$ is not eliminated.

We begin with the activation procedure.

Concentration in the activation procedure Consider a fixed $(k,h) \in [dH+1]\times[H]$ pair. By Azuma-Hoefdding's inequality, with probability at least $1 - \frac{\delta}{8H(dH+1)}$, we have

$$|\tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) - \mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h)| \le \mathcal{O}\left(\sqrt{\frac{\iota}{n_{\text{act}}}}\right) \le \frac{\epsilon}{2H} \le \zeta_{\text{act}}/4,$$

where the second inequality follows from $n_{\text{act}} = C\frac{H^2\iota}{\epsilon^2}$ with C being chosen large enough.

Take a union bound for all $(k,h) \in [dH+1]\times[H]$, we have with probability at least $1-\delta/4$, the following holds for all $(k,h) \in [dH+1]\times[H]$

$$|\tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) - \mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h)| \le \zeta_{\text{act}}/4.$$

By Algorithm 3, if the elimination procedure is not activated in the k^{th} phase, we have $\sum_{h=1}^{H} \tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) \le H\zeta_{\text{act}}$. Combine it with the concentration argument we just proved,

$$\sum_{h=1}^{H} \mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h) \le \sum_{h=1}^{H} \tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) + \frac{H\zeta_{\text{act}}}{4} \le \frac{5H\zeta_{\text{act}}}{4}.$$

On the other hand, if the elimination procedure is activated at the h^{th} step in the k^{th} phase, then $\tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) > \zeta_{\text{act}}$. Again combine it with the concentration argument we just proved,

$$\mathcal{E}_{\mathrm{II}}(f^k, \pi^k, h) \ge \tilde{\mathcal{E}}_{\mathrm{II}}(f^k, \pi^k, h) - \frac{\zeta_{\text{act}}}{4} > \frac{3\zeta_{\text{act}}}{4}.$$

Concentration in the elimination procedure Now, let us turn to the elimination procedure. We start by bounding the the second moment of

$$\frac{\mathbf{1}[\pi_f(s_h) = a_h]}{1/|\mathcal{A}|}\big(f_h(s_h, a_h) - r_h - \max_{a'\in\mathcal{A}} f_{h+1}(s_{h+1}, a')\big)$$

for all $f \in \mathcal{F}$. Let $y(s_h, a_h, r_h, s_{h+1}) = f_h(s_h, a_h) - r_h - \max_{a'\in\mathcal{A}} f_{h+1}(s_{h+1}, a') \in [-2, 1]$, then we have

$$\begin{aligned}&\mathbb{E}[\big(|\mathcal{A}|\mathbf{1}[\pi_f(s_h) = a_h] y(s_h, a_h, r_h, s_{h+1})\big)^2 \mid s_h \sim \pi^k, a_h \sim \text{Uniform}(\mathcal{A})]\\ \le& 4|\mathcal{A}|^2 \mathbb{E}[\mathbf{1}[\pi_f(s_h) = a_h] \mid s_h \sim \pi^k, a_h \sim \text{Uniform}(\mathcal{A})] = 4|\mathcal{A}|.\end{aligned}$$

For a fixed $(k,f) \in [dH+1] \times \mathcal{F}$, by applying Azuma-Bernstein's inequality, with probability at least $1 - \frac{\delta}{8(dH+1)|\mathcal{F}|}$ we have

$$|\hat{\mathcal{E}}_{\mathrm{II}}(f, \pi^k, h_k) - \mathcal{E}_{\mathrm{II}}(f, \pi^k, h_k)| \le \mathcal{O}\left(\sqrt{\frac{|\mathcal{A}|\iota'}{n_{\text{elim}}}} + \frac{|\mathcal{A}|\iota'}{n_{\text{elim}}}\right) \le \mathcal{O}\left(\sqrt{\frac{|\mathcal{A}|\iota'}{n_{\text{elim}}}}\right) \le \zeta_{\text{elim}}/2,$$

where $\iota' = \log[8(dH+1)|\mathcal{F}|/\delta]$, and the third inequality follows from $n_{\text{elim}} = C|\mathcal{A}|\iota/\zeta_{\text{elim}}^2$ with C being chosen large enough.

Taking a union bound over $[dH+1]\times\mathcal{F}$, we have with probability at least $1-\delta/4$, the following holds for all $(k,f) \in [dH+1] \times \mathcal{F}$

$$|\hat{\mathcal{E}}_{\mathrm{II}}(f, \pi^k, h_k) - \mathcal{E}_{\mathrm{II}}(f, \pi^k, h_k)| \le \zeta_{\text{elim}}/2.$$

Recall that Algorithm 3 eliminates all f satisfying $|\hat{\mathcal{E}}_{\mathrm{II}}(f, \pi^k, h_k)| > \zeta_{\text{elim}}$ when the elimination procedure is activated at the h_k^{th} step in the k^{th} phase. Therefore, if $|\mathcal{E}_{\mathrm{II}}(f, \pi^k, h_k)| \ge 2\zeta_{\text{elim}}$, f will be eliminated because

$$|\hat{\mathcal{E}}_{\mathrm{II}}(f, \pi^k, h_k)| \ge |\mathcal{E}_{\mathrm{II}}(f, \pi^k, h_k)| - \frac{\zeta_{\text{elim}}}{2} > \zeta_{\text{elim}}.$$

Finally, note that $\mathcal{E}_{\mathrm{II}}(Q^\star, \pi, h) \equiv 0$ for any π and h. As a result, it will never be eliminated within the first $dH+1$ phases because we can similarly prove

$$|\hat{\mathcal{E}}_{\mathrm{II}}(Q^\star, \pi^k, h_k)| \le |\mathcal{E}_{\mathrm{II}}(Q^\star, \pi^k, h_k)| + \frac{\zeta_{\text{elim}}}{2} < \zeta_{\text{elim}}.$$

Wrapping up: take a union bound for the activation and elimination procedure, and conclude that the three events, listed at the beginning of this section, hold for the the first $dH+1$ phases with probability at least $1-\delta/2$.

E.2. Proof of Theorem 22 (similar to Appendix C.2)

The proof is basically the same as that of Theorem 14 in Appendix C.

To begin with, we have the following lemma (akin to Lemma 24 and 25) showing that with high probability: (i) any function in the confidence set has low Bellman-error over the collected Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_H$ as well as the distribution from which $\mathcal{D}_1, \ldots, \mathcal{D}_H$ are sampled; (ii) the optimal value function is inside the confidence set. Its proof is almost identical to that of Lemma 24 and 25 which can be found in Appendix C.3.

Lemma 29 (Akin to Lemma 24 and 25) *Let $\rho > 0$ be an arbitrary fixed number. If we choose $\beta = c\big(\log[KH\mathcal{N}_{\mathcal{F}}(\rho)/\delta] + K\rho\big)$ with some large absolute constant c in Algorithm 4, then with probability at least $1-\delta$, for all $(k,h) \in [K] \times [H]$, we have*

(a) $\sum_{i=1}^{k-1} \mathbb{E}[\big(f_h^k(s_h,a_h) - (\mathcal{T} f_{h+1}^k)(s_h,a_h)\big)^2 \mid s_h \sim \pi^i, a_h \sim \mathrm{Uniform}(\mathcal{A})] \le \mathcal{O}(\beta)$,

(b) $\frac{1}{|\mathcal{A}|} \sum_{i=1}^{k-1} \sum_{a\in\mathcal{A}} \big(f_h^k(s_h^i,a) - (\mathcal{T} f_{h+1}^k)(s_h^i,a)\big)^2 \le \mathcal{O}(\beta)$,

(c) $Q^\star \in \mathcal{B}^k$ *for all* $k \in [K]$,

where s_h^i denotes the state at step h collected according to Line 5 in Algorithm 4 following π^i.

Proof [Proof of Lemma 29] To prove inequality (a), we only need to redefine the filtration $\mathfrak{F}_{t,h}$ in Appendix C.3.1 to be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1}$ and repeat the arguments there verbatim.

To prove inequality (b), we only need to redefine the filtration $\mathfrak{F}_{t,h}$ in Appendix C.3.1 to be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1} \bigcup \{s_1^t, a_1^t, r_1^t, \ldots, s_h^t\}$ and repeat the arguments there verbatim.

The proof of (c) is the same as that of Lemma 25 in Appendix C.3.2. ■

Step 1. Bounding the regret by Bellman error By Lemma 29 (c), we can upper bound the cumulative regret by the summation of Bellman error with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\left(V_1^{\star}(s_1)-V_1^{\pi^k}(s_1)\right)\le\sum_{k=1}^{K}\left(\max_a f_1^k(s_1,a)-V_1^{\pi^k}(s_1)\right)\overset{(i)}{=}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathcal{E}_{\mathrm{II}}(f^k,\pi^k,h), \tag{18}$$

where (i) follows from standard policy loss decomposition (e.g. Lemma 1 in Jiang et al. (2017)).

Step 2. Bounding cumulative Bellman error using DE dimension Next, we focus on a fixed step h and bound the cumulative Bellman error $\sum_{k=1}^{K}\mathcal{E}_{\mathrm{II}}(f^k,\pi^k,h)$ using Lemma 29.

Invoking Lemma 29 (a) with

$$\rho=\frac{\epsilon^2}{H^2\cdot\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F},\mathcal{D}_{\mathcal{F}},\epsilon/H)\cdot|\mathcal{A}|}$$

implies that with probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, we have

$$\sum_{i=1}^{k-1}\mathbb{E}\left[\left(f_h^k(s_h,\pi_{f_h^k}(s_h))-(\mathcal{T}f_{h+1}^k)(s_h,\pi_{f_h^k}(s_h))\right)^2\mid s_h\sim\pi^i\right]\le\mathcal{O}(|\mathcal{A}|\beta).$$

Further invoking Lemma 26 with

$$\begin{cases}\omega=\dfrac{\epsilon}{H},\ C=1,\\ \mathcal{X}=\mathcal{S},\ \mathcal{G}=(I-\mathcal{T}_h)V_{\mathcal{F}},\ \Pi=\mathcal{D}_{\mathcal{F},h},\\ g_k(s):=(f_h^k-\mathcal{T}_hf_{h+1}^k)(s,\pi_{f_h^k}(s))\text{ and }\mu_k=\mathbb{P}^{\pi^k}(s_h=\cdot),\end{cases}$$

we obtain

$$\frac{1}{K}\sum_{t=1}^{K}\mathcal{E}_{\mathrm{II}}(f^t,\pi^t,h)\le\mathcal{O}\left(\sqrt{\frac{\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F},\mathcal{D}_{\mathcal{F}},\epsilon/H)|\mathcal{A}|\log[KH\mathcal{N}_{\mathcal{F}}(\rho)/\delta]}{K}}+\frac{\epsilon}{H}\right).$$

Plugging in the choice of K completes the proof.

Similarly, for $\mathcal{D}_{\Delta}$, we can invoke Lemma 29 (b) witht

$$\rho=\frac{\epsilon^2}{H^2\cdot\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F},\mathcal{D}_{\Delta},\epsilon/H)\cdot|\mathcal{A}|},$$

and Lemma 26 with

$$\begin{cases}\omega = \dfrac{\epsilon}{H},\ C = 1,\\ \mathcal{X} = \mathcal{S},\ \mathcal{G} = (I - \mathcal{T}_h)V_{\mathcal{F}},\ \Pi = \mathcal{D}_{\Delta,h},\\ g_k(s) := (f_h^k - \mathcal{T}_h f_{h+1}^k)(s, \pi_{f_h^k}(s)) \text{ and } \mu_k = \mathbf{1}\{\cdot = s_h^k\},\end{cases}$$

and obtain

$$\begin{aligned}\frac{1}{K}\sum_{t=1}^{K}\mathcal{E}_{\mathrm{II}}(f^t,\pi^t,h) \le& \frac{1}{K}\sum_{t=1}^{K}(f_h^t - \mathcal{T} f_{h+1}^t)(s_h^t, \pi_{f_h^t}(s_h^t)) + \mathcal{O}\left(\sqrt{\frac{\log K}{K}}\right)\\ \le& \mathcal{O}\left(\sqrt{\frac{\dim_{\mathrm{BE}_{\mathrm{II}}}(\mathcal{F},\mathcal{D}_\Delta,\epsilon/H)|\mathcal{A}|\log[KH\mathcal{N}_{\mathcal{F}}(\rho)/\delta]}{K}} + \frac{\epsilon}{H} + \sqrt{\frac{\log K}{K}}\right),\end{aligned}$$

where the first inequality follows from standard martingale concentration.

Plugging in the choice of K completes the proof.

Appendix F. Auxiliary Results

In this section, we state and prove some auxiliary results.

F.1. Review of relevant definitions in previous works

In this section, we review the definition of some existing function approximation settings in ascending order of generality. We start with the definition of linear MDPs (e.g., Jin et al., 2020).

Definition 30 (linear MDPs) *We say an MDP is linear of dimension d if for each $h \in [H]$, there exist feature mappings $\phi_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, $\psi_h : \mathcal{S} \to \mathbb{R}^d$ and vector $\theta_h^r \in \mathbb{R}^d$ such that $\mathbb{P}_h(s' \mid s,a) = \phi_h(s,a)^\top \psi_h(s')$ and $r_h(s,a) = \phi_h(s,a)^\top\theta_h^r$.*

We remark that existing works (e.g., Jin et al., 2020) usually assume ϕ is *known* in advance while ψ and θ^r are *unknown*. Next, we review the linear function approximation setting (e.g., Zanette et al., 2020a).

Definition 31 (linear realizability and linear completeness) *We say an MDP satisfies d-dimensional linear realizability with respect to feature mapping $\phi = \{\phi_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d\}_{h\in[H]}$ if for each $h \in [H]$, there exists a vector $\theta_h^\star \in \mathbb{R}^d$ such that $Q_h^\star(\cdot) = \phi_h(\cdot)^\top\theta_h^\star$. Moreover, we say it satisfies linear completeness with respect to ϕ if for each $h \in [H]$ and $\theta \in \mathbb{R}^d$, there exists $\theta' \in \mathbb{R}^d$ such that $(\mathcal{T} f_{\theta,h+1})(s,a) = \phi_h(s,a)^\top\theta'$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $f_{\theta,h+1}(\cdot) = \phi_{h+1}(\cdot)^\top\theta$.*

We make two comments here. Firstly, previous works (Jin et al., 2020; Zanette et al., 2020a) prove that linear MDPs always satisfy linear realizability and linear completeness with the same ambient dimension. Secondly, only assuming linear realizability is insufficient for sample-efficient learning because exponential lower bounds for sample complexity is known in that case (Weisz et al., 2020).

Finally, we briefly discuss the Bellman rank proposed by Jiang et al. (2017) (see Definition 9). Unlike the setting of linear MDPs or linear realizability and completeness, the feature mappings ϕ and ψ in the definition of Bellman rank are assumed to be *unknown*. The following lemma claims that any RL problems admitting d-dimensional linear realizability and completeness have Bellman rank at most d.

Lemma 32 *Suppose MDP $\mathcal{M}$ satisfies d-dimensional linear realizability and completeness with respect to feature mapping ϕ. Define $\theta := (\theta_1, \ldots, \theta_H) \in \mathbb{R}^{d\times H}$, and $f_{\theta,h}(s,a) := \phi_h(s,a)^\top \theta_h$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$. Then $\mathcal{M}$ and $\mathcal{F} := \{f_\theta = (f_{\theta,1}, \ldots, f_{\theta,H}, 0) : \theta \in \mathbb{R}^{d\times H}\}$ have Bellman rank at most d.*

Proof The proof follows directly from the definitions of linear completeness and Bellman rank.

Consider an arbitrary $\theta \in \mathbb{R}^{d\times H}$ and $h \in [H]$. By linear completeness, there exists $\hat\theta_h \in \mathbb{R}^d$ such that $(\mathcal{T} f_{\theta,h+1})(s,a) = \phi_h(s,a)^\top \hat\theta_h$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Let π be an arbitrary policy. We have

$$\mathcal{E}(f_\theta, \pi, h) = \mathbb{E}_\pi[(f_{\theta,h} - \mathcal{T} f_{\theta,h+1})(s_h,a_h)] = \left\langle \mathbb{E}_\pi[\phi_h(s_h,a_h)], \theta_h - \hat\theta_h \right\rangle. \tag{19}$$

We conclude the proof by noting that $\mathbb{E}_\pi[\phi_h(s_h,a_h)] \in \mathbb{R}^d$ only depends on (π,h) and $\theta_h - \hat\theta_h \in \mathbb{R}^d$ is fully determined by (f_θ, h). ■

F.2. $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon)$ versus $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$

In this paper, we have mainly focused on the BE dimension induced by two special distribution families: (a) $\mathcal{D}_\mathcal{F}$ — the roll-in distributions produced by executing the greedy policies induced by the functions in $\mathcal{F}$, (b) $\mathcal{D}_\Delta$ — the collection of all Dirac distributions. And we prove that both low $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon)$ and low $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$ can imply sample-efficient learning. As a result, it is natural to ask what is the relation between $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon)$ and $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$? Is it possible that one of them is always no larger than the other so that we only need to use the smaller one? We answer this question with the following proposition, showing that either of them can be arbitrarily larger than the other.

Proposition 33 *There exists absolute constant c such that for any $m \in \mathbb{N}^+$,*

(a) there exist an MDP and a function class $\mathcal{F}$ satisfying for all $\epsilon \in (0,1/2]$, $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon) \le c$ while $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon) \ge m$.

(b) there exist an MDP and a function class $\mathcal{F}$ satisfying for all $\epsilon \in (0,1/2]$, $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon) \le c$ while $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon) \ge m$.

Proof We prove (a) first. Consider the following contextual bandits problem ($H = 1$).

- There are m states $s_1, \ldots, s_m$ but the agent always starts at s_1. This means the agent can never visit other states because each episode contains only one step ($H = 1$).
- There are two actions a_1 and a_2. The reward function is zero for any state-action pair.
- The function class $\mathcal{F}_1 = \{f_i(s,a) = \mathbf{1}(s = s_i) + \mathbf{1}(a = a_1) : i \in [m]\}$.

First of all, note in this setting $\mathcal{D}_\Delta$ is the collection of all Dirac distributions over $\mathcal{S}\times\mathcal{A}$, $\mathcal{D}_{\mathcal{F},1}$ is a singleton containing only $\delta_{(s_1,a_1)}$, and $(I - \mathcal{T}_1)\mathcal{F}$ is simply $\mathcal{F}_1$ because $H = 1$ and $r \equiv 0$. Since $\mathcal{D}_{\mathcal{F},1}$ has cardinality one, it follows directly from definition that $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$ is at most 1. Moreover, it is easy to verify that $(s_1,a_2), (s_2,a_2), \ldots, (s_m,a_m)$ is a 1-independent sequence with respect to $\mathcal{F}$ because we have $f_i(s_j,a_2) = \mathbf{1}(i = j)$ for all $i,j \in [m]$. As a result, we have $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon) \ge m$ for all $\epsilon \in (0,1]$.

Now we come to the proof of (b). Consider the following contextual bandits problem ($H = 1$).

- There are 2 states s_1 and s_2. In each episode, the agent starts at s_1 or s_2 uniformly at random.
- There are m actions $a_1, \ldots, a_m$. The reward function is zero for any state-action pair.
- The function class $\mathcal{F}_1 = \{f_i(s,a) = (2 \cdot \mathbf{1}(s = s_1) - 1) + 0.5 \cdot \mathbf{1}(a = a_i) : \ i \in [m]\}$.

First of all, note in this setting $(I - \mathcal{T}_1)\mathcal{F}$ is simply $\mathcal{F}_1$ and the roll-in distribution induced by the greedy policy of f_i is the uniform distribution over (s_1, a_i) and (s_2, a_i), which we denote as μ_i. It is easy to verify that $\mu_1, \ldots, \mu_m$ is a 0.5-independent sequence with respect to $\mathcal{F}$ because $\mathbb{E}_{(s,a)\sim\mu_i}[f_j(s,a)] = 0.5 \cdot \mathbf{1}(i = j)$. Therefore, for all $\epsilon \in (0, 0.5]$, $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \geq m$.

Next, we upper bound $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$ which is equivalent to $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon)$ in this problem. Assume $\dim_{\mathrm{DE}}(\mathcal{F}_1, \mathcal{D}_\Delta, \epsilon) = k$. Then there exist $g_1, \ldots, g_k \in \mathcal{F}_1$ and $w_1, \ldots, w_k \in \mathcal{S} \times \mathcal{A}$ such that for all $i \in [k]$, $\sqrt{\sum_{t=1}^{i-1}(g_i(w_i))^2} \leq \epsilon$ and $|g_i(w_i)| > \epsilon$. Note that we have $|f(s,a)| \in [0.5, 1.5]$ for all $(s, a, f) \in \mathcal{S} \times \mathcal{A} \times \mathcal{F}_1$. Therefore, if $\epsilon > 1.5$, then $k = 0$; if $\epsilon \leq 1.5$, then $k \leq 10$ because $0.5 \times \sqrt{k-1} \leq \sqrt{\sum_{t=1}^{k-1}(g_k(w_t))^2} \leq \epsilon \leq 1.5$. ■