Optimizing Sample Efficiency in Reinforcement Learning: Bellman Eluder Dimension and Admissible Bellman Characterization

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 22, 2024

Abstract

This paper introduces the Bellman Eluder (BE) dimension, a novel complexity measure for reinforcement learning (RL) with general function approximation. The BE dimension generalizes existing concepts such as Bellman rank and Eluder dimension, capturing a broader class of RL problems under weaker assumptions. We propose Golf, a new optimization-based algorithm that leverages the BE dimension to achieve near-optimal sample complexity in RL tasks with large state-action spaces. Our results provide theoretical guarantees, improving both regret bounds and sample efficiency over prior works. Additionally, we offer a comprehensive framework, termed Admissible Bellman Characterization (ABC), which bridges the gap between model-free and model-based RL methods. Our findings advance the understanding of RL in environments where minimal structural assumptions can still enable efficient learning.

Keywords: Reinforcement Learning (RL); Bellman Eluder Dimension; Function Approximation; Admissible Bellman Characterization (ABC); Sample Complexity; Markov Decision Processes (MDPs); Regret Bounds; Eluder Dimension

1 Introduction

Reinforcement learning (RL) has rapidly progressed in recent years, demonstrating impressive success in various applications such as robotics, games, and autonomous systems. These advancements are largely driven by powerful function approximation techniques, particularly those utilizing deep learning, to estimate value functions or policies in environments with large, complex state spaces where traditional tabular methods fall short. However, despite these empirical achievements, the theoretical understanding of RL with general function approximation remains incomplete. Key challenges persist in terms of sample efficiency, generalization, and exploration, especially when the underlying assumptions on the function class deviate from simplistic models like linear approximation.

Classical RL approaches often rely on strong structural assumptions—such as linear function approximation or simplified models like Linear Quadratic Regulators (LQR)—to ensure theoretical guarantees. While these assumptions facilitate analysis, they rarely hold in complex real-world applications, which motivates a fundamental question: ***What are the minimal structural assumptions that allow RL algorithms to efficiently learn with general function approximation?***

To address this, we introduce a new complexity measure, the Bellman Eluder (BE) dimension, which generalizes existing concepts like Bellman rank and Eluder dimension. The BE dimension offers a unified framework for analyzing RL problems with fewer assumptions, subsuming both low

Bellman rank and low Eluder dimension problems while introducing new classes of tasks that can be solved in a sample-efficient manner. This measure allows RL algorithms to tackle more complex environments while maintaining near-optimal sample complexity.

In addition, we propose the Golf algorithm, designed specifically to solve RL problems characterized by low BE dimension. Golf is an optimization-based algorithm that leverages the BE dimension to achieve competitive sample complexity bounds across diverse RL settings. Our approach not only improves on existing regret and sample efficiency bounds for RL with general function approximation but also extends the applicability of RL methods to settings with large or infinite state-action spaces.

Brief Contributions Our main contributions are as follows: (i) We introduce the Bellman Eluder dimension, a new complexity measure that subsumes existing tractable RL problem classes and defines new ones. (ii) We propose the *Golf* algorithm, which provably learns policies in RL problems with low BE dimension and achieves near-optimal sample complexity. (iii) We provide theoretical guarantees for both algorithms, offering regret and sample complexity bounds that improve upon existing results for RL with general function approximation.

1.1 Contributions

Our main contributions are as follows:

- We introduce the Bellman Eluder dimension, a new complexity measure that unifies several existing RL problem classes and identifies new ones.
- We propose the Golf algorithm, which provably learns RL tasks with low BE dimension and achieves near-optimal sample complexity.
- We provide theoretical guarantees for our algorithm, offering improved regret and sample complexity bounds for RL with general function approximation.

2 Leveraging Bellman Eluder Dimension for Efficient Function Approximation in Reinforcement Learning

Function approximation in modern RL, especially based on deep neural networks, lies at the heart of the recent practical successes of RL in domains such as Atari [MKS$^+$13], Go [SHM$^+$16], robotics [KBP13], and dialogue systems [LMR$^+$16]. 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 present.

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 [WWDK21, JYWJ20, ZLKB20a], or that the model is precisely Linear Quadratic Regulator (LQR) [AM07, FGKM18, DMM$^+$19]. 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 (with weak structural assumptions) that cover many practical applications of interests; (2) design sample-efficient algorithms that provably learn any RL problem in this class.

The attempts to find weak or minimal structural assumptions that allow statistical learning can be traced in supervised learning where VC dimension [Vap13] or Rademacher complexity [BM02] is proposed, or in online learning where Littlestone dimension [Lit88] or sequential Rademacher complexity [RST10] 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, [JKA$^+$17] introduces a generic complexity notion—Bellman rank, which can be proved small for many RL problems including linear MDPs [JYWJ20], reactive POMDPs [KAL16], etc. [JKA$^+$17] further propose an hypothesis elimination-based algorithmfor sample-efficient learning of problems with low Bellman rank. On the other hand, recent work by [WSY20a] considers general function approximation with low Eluder dimension [RVR13], and designs a UCB-style algorithm with regret guarantee. Noticeably, generalized linear MDPs [WWDK21] and kernel MDPs (see Section A.2) are subclasses 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). The family of low BE dimension further includes new problems such as kernel reactive POMDPs (see Section A.2) which were not known to be sample-efficiently learnable.

- We design a new optimization-based algorithm—Golf, which provably learns 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 [ZLKB20a] which is minimax optimal when specified to the linear setting. Our rates further improve upon [JKA$^+$17, WSY20a] in low Bellman rank and low Eluder dimension settings, respectively.

2.1 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., [SM05, MS08, CJ19, XJ20]. In this setting, agents are provided with exploratory data or simulator, so that they do not need to explicitly address the challenge of exploration. In this paper, we do not make such assumption, and attack the exploration problem directly. In the following we focus exclusively on the RL results in the general setting where exploration is required.

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., [BT02, JOA10, DB15, AJ17, AOM17, ZB19, JAZBJ18, ZZJ20]. 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 [AOM17] and model-free [ZZJ20] algorithms. Moreover, the bound is proved to be minimax-optimal [JAZBJ18, DMKV21]. 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., [JYWJ20, WWDK21, CYJW20a, ZLKB20a, ZLKB20b, AKKS20, NPB20, SJK$^+$19] 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 [ZLKB20a], 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., [OVR14, JKA$^+$17, SJK$^+$19, DPWZ20, WSY20a, YJW$^+$20a, FRSLX20]. Among them, [JKA$^+$17] and [WSY20a] are the closest to our work.[1] [WSY20a] propose 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 [WSY20a], our algorithm Golf works under a weaker completeness assumption, with a better regret guarantee.

Relation to bilinear classes Concurrent to this work, [DKL$^+$21] propose a new general tractable class of RL problems—bilinear class with low effective dimension (also known as low critical information gain in [DKL$^+$21]). We comment on the similarities and differences between two works as follows.

In terms of algorithms, the algorithm proposed in [DKL$^+$21] are based on the algorithm originally proposed in [JKA$^+$17]. The two algorithms share similar guarantees in terms of assumptions and complexity results. More importantly, our work further develops a new type of algorithm for general function approximation—Golf, a natural and clean algorithm which can be viewed as an optimistic version of classical algorithm—Fitted Q-Iteration [Sze10]. Golf gives much sharper sample complexity guarantees compared to [DKL$^+$21] for various settings, and is minimax-optimal when applied to the linear setting [ZLKB20a].

In terms of richness of new classes identified, it depends on (a) what structure of MDP the complexity measures are applied to, and (b) what complexity measures are used. For (a), BE dimension applies to the Bellman error, while the bilinear class allows general surrogate losses of the Bellman error. For (b), this paper uses Eluder dimension while [DKL$^+$21] uses effective dimension. It can be shown that low effective dimension always implies low Eluder dimension (see Section A.2.2). In short, [DKL$^+$21] is more general in (a), while our work is more general in (b). As a result, neither work fully captures the other.

In particular, our BE framework covers a majority of the examples identified in [DKL$^+$21] including low occupancy complexity, linear $Q^\star/V^\star$, $Q^\star$ state aggregation, feature selection/FLAMBE. Nevertheless, our work can not address examples with model-based function approximation (e.g., low witness rank [SJK$^+$19]) while [DKL$^+$21] can. On the other hand, [DKL$^+$21] can not address the class of RL problems with low Eluder dimension [WSY20a] while our work can. Moreover, for several classes of RL problems that both works cover, our complexity measure is sharper. For

[1] [JKA$^+$17] propose a complexity measure named Bellman rank and design an algorithm Olive with PAC guarantees for problems with low Bellman rank. We note that low Bellman rank is a special case of low BE dimension. When specialized to the low Bellman rank setting, our result for Olive exactly matches the guarantee in [JKA$^+$17]. Our result for Golf requires an additional completeness assumption, but provides sharper sample complexity guarantee.

example, in the setting of function approximation with generalized linear functions, the BE dimension is $\tilde{O}(d)$ where d is the ambient dimension of the feature vectors, while the effective dimension under the generalized bilinear framework of [DKL$^+$21] is at least $\tilde{\Omega}(d^2)$.

2.2 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. [2] 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 at step $H + 1$, and the episode terminates when s_{end} is reached.

Policy and value functions A (deterministic) policy π 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,

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

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

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

There exists an optimal policy $\pi^\star$, which gives the optimal value function for all states [Put14], 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

$$\text{Reg}(K) := \sum_{k=1}^{K} [V_1^\star(s_1) - V_1^{\pi^k}(s_1)]$$

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

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

2.2.1 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 $Q_h^\star$—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 [KAL16, WAS20]). Below, we present two assumptions about function approximation that are commonly adopted in the literature.

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

Realizability requires the function class is well-specified, i.e., function class $\mathcal{F}$ in fact contains the optimal Q-value function $Q^\star$ 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 $\{\mathcal{T}_h f_{h+1} : f_{h+1} \in \mathcal{F}_{h+1}\}$. 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 addressing function classes with infinite elements, we need a notion similar to cardinality. We use the standard ϵ-covering number.

Definition 1 (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 $\rho(x, y) \leq \epsilon$.*

We refer readers to standard textbooks , see, e.g., [Wai19] for further properties of covering number. In this paper, we will always apply the covering number on function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, 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 $\mathcal{N}_{\mathcal{F}}(\epsilon)$.

2.2.2 Eluder dimension

One class of functions highly related to this paper is the function class of low Eluder dimension [RVR13].

Definition 2 (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 $\sqrt{\sum_{i=1}^n (g_1(x_i) - g_2(x_i))^2} \leq \epsilon$, 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 points $\{x_i\}_{i=1}^n$, but the values are rather different at z. This independence relation naturally induces the following complexity measure.

Definition 3 (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 $\epsilon' \geq \epsilon$ 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 $\{x_1, \ldots, x_d\}$ such that x_i is linearly independent of $\{x_1, \ldots, x_{i-1}\}$ for all $i \in [n]$. Eluder dimension generalizes the linear independence relation in standard vector space to capture both nonlinear independence and approximate independence, and thus is more general.

2.3 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 family of problems with low BE dimension contains the two existing most general tractable problem classes in RL—problems with low Bellman rank, and problems with low Eluder dimension (see Figure 1).

We start by developing a new distributional version of the original Eluder dimension proposed by [RVR13] (see Section 2.2.2 for more details).

Definition 4 (ϵ-independence between distributions)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \ldots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ϵ-independent of $\{\mu_1, \mu_2, \ldots, \mu_n\}$ 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 5 (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 $\{\rho_1, \ldots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \ge \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \ldots, \rho_{i-1}\}$ for all $i \in [n]$.*

Definition 4 and Definition 5 generalize Definition 2 and Definition 9 to their distributional versions, by inspecting the expected values of functions instead of the function values at points, and by restricting the candidate distributions to a certain family Π. The main advantage of this generalization is exactly in the statistical setting, where estimating the expected values of functions with respect to a certain distribution family can be easier than estimating function values at each point (which is the case for RL in large state spaces).

It is clear that the standard Eluder dimension is a special case of the distributional Eluder dimension, because if we choose $\Pi = \{\delta_x(\cdot) \mid x \in \mathcal{X}\}$ 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 the key notion in this paper—Bellman Eluder dimension.

Definition 6 (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*

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

Remark (Q-type v.s. V-type)**.** Definition 6 is based on the Bellman residuals functions that take a state-action pair as input, thus referred to as Q-type BE dimension. Alternatively, one can define V-type BE dimension using a different set of Bellman residual functions that depend on states only (see Section A.1). We focus on Q-type in the main paper, and present the results for V-type in Section A.1. Both variants are important, and they include different sets of examples (see Section A.1, A.2).

In short, Bellman Eluder dimension is simply the distributional Eluder dimension on the function class of Bellman residuals, maximizing over all steps. In addition to function class $\mathcal{F}$ and error ϵ, 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.

(i) $\mathcal{D}_{\mathcal{F}} := \{\mathcal{D}_{\mathcal{F},h}\}_{h\in[H]}$, 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., $\pi_{f,h}(\cdot) = \text{argmax}_{a\in\mathcal{A}} f_h(\cdot, a)$ for all $h \in [H]$.

(ii) $\mathcal{D}_{\Delta} := \{\mathcal{D}_{\Delta,h}\}_{h\in[H]}$, 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 a single state-action pair.

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

2.3.1 Relations with known tractable classes of RL problems

Known tractable problem classes in RL include but not limited to tabular MDPs, linear MDPs [JYWJ20], linear quadratic regulators [AM07], generalized linear MDPs [WWDK21], kernel MDPs (Section A.2), reactive POMDPs [KAL16], reactive PSRs [SJR12, JKA+17]. 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 seminal paper by [JKA+17] proposes the complexity measure—Bellman rank, and shows that a majority of RL examples mentioned above have low Bellman rank.[3] Formally,

Definition 7 (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.*

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

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

We remark that similar to Bellman Eluder dimension, Bellman rank also has two variants—Q-type (Definition 7) and V-type (see Section A.1). 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 policy $\pi_{f'}$, while the other vector depends on the value function f. At a high level, it claims that the average Bellman error has a linear inner product structure.

Proposition 1 (low Bellman rank $\subset$ low BE dimension). *If an MDP with function class $\mathcal{F}$ has Bellman rank d with normalization parameter ζ, then*

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

Proposition 1 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}.

[3] They also propose a hypothesis elimination based algorithm—OLIVE, that learns any low Bellman rank problem within polynomial samples.

Relation with low Eluder dimension [WSY20a] 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 2 (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 2 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, we show that the set of low BE dimension problems is strictly larger than the union of low Eluder dimension problems and low Bellman rank problems.

Proposition 3 (low BE dimension $\not\subset$ low Eluder dimension $\cup$ low Bellman rank)**.** *For any $m \in \mathbb{N}^+$, there exists an MDP and a function class $\mathcal{F}$ so that for all $\epsilon \in (0, 1]$, we have* $\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\mathcal{F}, \epsilon) = \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon) \leq 5$*, but* $\min\{\min_{h\in[H]} \dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon), \text{Bellman rank}\} \geq m$*.*

In particular, the family of low BE dimension includes new examples such as kernel reactive POMDPs (Section A.2), which can not be addressed by the framework of either Bellman rank or Eluder dimension.

2.4 Algorithm Golf

Section 2.3 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 can find near-optimal policies for these problems within a polynomial number of samples.

At a high level, Golf can be viewed as an optimistic version of the classic algorithm—Fitted Q-Iteration (FQI) [Sze10]. Golf generalizes the Eleanor algorithm [ZLKB20a] from the special linear setting to the general setting with arbitrary function classes.

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 , 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.

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

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

[4] [WSY20a] 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).

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

```
1: Initialize: $\mathcal{D}_1, \ldots, \mathcal{D}_H \leftarrow \emptyset$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
2: for episode $k$ 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:     Collect a trajectory $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1})$ by following $\pi^k$
5:     Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(s_h, a_h, r_h, s_{h+1})\}$ for all $h \in [H]$
6:     Update
           $\mathcal{B}^k = \left\{ f \in \mathcal{F} : \ \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta \text{ for all } h \in [H] \right\}$
           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)
7: end for
8: Output $\pi^{\mathrm{out}}$ sampled uniformly at random from $\{\pi^k\}_{k=1}^K$
```

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. We remark that FQI algorithm [Sze10] simply updates $f_h \leftarrow \mathrm{argmin}_{\phi \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(\phi, f_{h+1})$. 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 than the optimal loss over the auxiliary function class $\mathcal{G}_h$.

We remark that in general, the optimization problem in Line 3 of Golf can not be solved computationally efficiently.

2.4.1 Theoretical guarantees

In this subsection, we present the theoretical guarantees for Golf, which hold under Assumption 1 (realizability) and the following generalized completeness assumption introduced in [ASM08, CJ19]. Let $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_H$ be an auxiliary function class provided to the learner where each $\mathcal{G}_h \subseteq (\mathcal{S} \times \mathcal{A} \to [0,1])$. Generalized completeness requires the auxiliary function class $\mathcal{G}$ to be rich enough so that applying Bellman operator to any function in the primary function class $\mathcal{F}$ will end up in $\mathcal{G}$.

Assumption 3 (Generalized completeness)**.** *$\mathcal{T}_h \mathcal{F}_{h+1} \subseteq \mathcal{G}_h$ for all $h \in [H]$.*

If we choose $\mathcal{G} = \mathcal{F}$, then Assumption 3 is equivalent to the standard completeness assumption (Assumption 2). Now, we are ready to present the main theorem for Golf.

Theorem 1 (Regret of Golf)**.** *Under Assumption 1, 3, there exists an absolute constant c such that for any $\delta \in (0,1]$, $K \in \mathbb{N}$, if we choose parameter $\beta = c \log[\mathcal{N}_{\mathcal{F} \cup \mathcal{G}}(1/K) \cdot KH/\delta]$ in* Golf*, then with probability at least $1 - \delta$, for all $k \in [K]$, we have*

$$\mathrm{Reg}(k) = \sum_{t=1}^{k} \left[V_1^\star(s_1) - V_1^{\pi^t}(s_1) \right] \le \mathcal{O}(H\sqrt{dk\beta})$$

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 1 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 two function classes. 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 spaces are typically exponentially large.

We remark that when function class $\mathcal{F} \cup \mathcal{G}$ has finite number of elements, its covering number is upper bounded by its cardinality $|\mathcal{F} \cup \mathcal{G}|$. 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 $\log \mathcal{N}_{\mathcal{F}\cup\mathcal{G}}$ and omit its ϵ' dependency for clean presentation. Theorem 1 claims that the regret scales as $\tilde{\mathcal{O}}(H\sqrt{dK \log \mathcal{N}_{\mathcal{F}\cup\mathcal{G}}})$, where $\tilde{\mathcal{O}}(\cdot)$ omits absolute constants and logarithmic terms.[5]

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

Corollary 1 (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 $\beta = c\log[\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}(\epsilon^2/(dH^2)) \cdot HK]$ in* Golf*, then the output policy π^{out} is $\mathcal{O}(\epsilon)$-optimal with probability at least $1/2$, if*

$$K \geq \Omega\left(\frac{H^2 d}{\epsilon^2} \cdot \log\left[\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}\left(\frac{\epsilon^2}{H^2 d}\right) \cdot \frac{Hd}{\epsilon}\right]\right)$$

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 1 claims that $\tilde{\mathcal{O}}(H^2 d \log(\mathcal{N}_{\mathcal{F}\cup\mathcal{G}})/\epsilon^2)$ 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 $\log(\mathcal{N}_{\mathcal{F}\cup\mathcal{G}})$.

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 ambient dimension d_{lin}, we have BE dimension $d = \tilde{\mathcal{O}}(d_{\mathrm{lin}})$ and $\log(\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}) = \tilde{\mathcal{O}}(d_{\mathrm{lin}})$. Our regret bound becomes $\tilde{\mathcal{O}}(Hd_{\mathrm{lin}}\sqrt{K})$ which matches the best known result [ZLKB20a] up to logarithmic factors; (2) For function class with low Eluder dimension [WSY20a], our results hold under weaker completeness assumptions. Our regret scales with $\sqrt{d_{\mathrm{E}}}$ in terms of dependency on Eluder dimension d_{E}, which improves the linear d_{E} scaling in the regret of [WSY20a]; (3) Finally, for low Bellman rank problems, our sample complexity scales linearly with Bellman rank, which improves upon the quadratic dependence in [JKA+17]. We remark that all results mentioned above assume (approximate) realizability. All except [JKA+17] assume (approximate) completeness.

2.4.2 Key ideas in proving Theorem 1

In this subsection, we present a brief proof sketch for the regret bound of Golf. We defer all the details to Section 2.6. For simplicity, we only discuss the case of choosing $\mathcal{D}_\mathcal{F}$ as the distribution family Π in the definition of Bellman Eluder dimension (Definition 6). The proof for using $\mathcal{D}_\Delta$ as the distribution family follows from similar arguments.

Our proof strategy consists of three main steps.

[5] We will not omit $\log \mathcal{N}_{\mathcal{F}\cup\mathcal{G}}$ in $\tilde{\mathcal{O}}(\cdot)$ notation since for many function classes, $\log \mathcal{N}_{\mathcal{F}\cup\mathcal{G}}$ is not small. For instance, for a $\tilde{d}$-dimensional linear function class, $\log \mathcal{N}_{\mathcal{F}\cup\mathcal{G}} = \tilde{\mathcal{O}}(\tilde{d})$.

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 3 in Section 2.6.1), which is a natural consequence of martingale concentration and the properties of the confidence set we designed. Because of $Q^\star \in \mathcal{B}^k$, the optimistic planning step (Line 3) in Golf guarantees that $V_1^\star(s_1) \le \max_a f_1^k(s_1, a)$ for every episode k. This optimism allows the following upper bound on regret

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

where the right equality follows from the standard policy loss decomposition (see, e.g., Lemma 1 in [JKA$^+$17]), and $\mathbb{E}_\pi$ denotes the expectation taken over sequence $(s_1, a_1, \ldots, s_H, a_H)$ when executing policy π.

Step 2: Utilize the sharpness of our confidence 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 $\mathcal{D}_h$ in episode k are collected by executing each π^i for one episode for all $i < k$, by standard martingale concentration arguments and the completeness assumption, we can show that with high probability (Lemma 2 in Section 2.6.1)

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

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

Lemma 1 is a simplification of Lemma 4 in Section 2.6, which is a modification of Lemma 2 in [RVR13]. Intuitively, Lemma 1 can be viewed as an analogue of the pigeon-hole principle for DE dimension. Choose Φ to be the function class of Bellman residuals, and μ_k to be the distribution under policy π^k, we finish the proof.

2.5 Proofs for BE Dimension

In this section, we provide formal proofs for the results stated in Section 2.3.

2.5.1 Proof of Proposition 1

The proof is basically the same as that of Example 3 in [RVR13] 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 4, 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 $|\mathbb{E}_{\mu_i}[f_h^i - \mathcal{T}_h f_{h+1}^i]| > \epsilon$. Since $\mu_1, \ldots, \mu_n \in \mathcal{D}_{\mathcal{F},h}$, there exist $g^1, \ldots, g^n \in \mathcal{F}$ so that μ_i is generated by executing π_{g^i} for all $i \in [n]$.

By the definition of Bellman rank, this is equivalent to: for all $i \in [m]$, $\sqrt{\sum_{t=1}^{i-1}(\langle \phi_h(g^i), \psi_h(f^t)\rangle)^2} \le \epsilon$ and $|\langle \phi_h(g^i), \psi_h(f^i)\rangle| > \epsilon$.

For notational simplicity, define $\mathbf{x}_i = \phi_h(g^i)$, $\mathbf{z}_i = \psi_h(f^i)$ 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

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

□

2.5.2 Proof of Proposition 2

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 4, 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 2 and 9, this implies $\dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon) \ge m$, which completes the proof. □

2.5.3 Proof of Proposition 3

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 $t \in [m-1]$, $f_{\theta_t}(a_t) - f_{\theta_{t+1}}(a_t) = 1 \geq \epsilon$. Therefore, $\min_{h \in [H]} \dim_{\mathrm{E}}(\mathcal{F}_h, \epsilon) = \dim_{\mathrm{E}}(\mathcal{F}_1, \epsilon) \geq m - 1$.

Bellman rank It is direct to see the Bellman residual matrix is $\mathcal{E} := \Theta^\top \Theta \in \mathbb{R}^{m \times m}$ with rank m, where $\Theta = [\theta_1, \theta_2, \ldots, \theta_m]$. As a result, the Bellman rank is at least m.

BE dimension First, note in this setting $(I - \mathcal{T}_1)\mathcal{F}$ is simply $\mathcal{F}_1$ (because $\mathcal{F}_2 = \{0\}$ and $r \equiv 0$), and $\mathcal{D}_{\mathcal{F}}$ coincides with $\mathcal{D}_\Delta$, 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 $\sqrt{k-1} \leq \sqrt{\sum_{i=1}^{k-1}(\langle q_k, w_i \rangle)^2} \leq \epsilon$. □

2.6 Proofs for Golf

In this section, we provide formal proofs for the results stated in Section 2.4.

2.6.1 Proof of Theorem 1

We start the proof 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 2. *Let $\rho > 0$ be an arbitrary fixed number. If we choose $\beta = c\big(\log[KH\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}(\rho)/\delta] + K\rho\big)$ 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 3. *Under the same condition of Lemma 2, with probability at least $1-\delta$, we have $Q^\star \in \mathcal{B}^k$ for all $k \in [K]$.*

The proof of Lemma 2 and 3 relies on standard martingale concentration (e.g. Freedman's inequality) and can be found in Section 2.6.3.

Step 1. Bounding the regret by Bellman error By Lemma 3, we can upper bound the cumulative regret by the summation of Bellman error with probability at least $1-\delta$:

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

where (i) follows from standard policy loss decomposition (e.g. Lemma 1 in [JKA$^+$17]).

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 2. To proceed, we need the following lemma to control the accumulating rate of Bellman error.

Lemma 4. *Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \le C$ for all $(g,x) \in \Phi \times \mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{\phi_k\}_{k=1}^K \subset \Phi$ and $\{\mu_k\}_{k=1}^K \subset \Pi$ satisfy that for all $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2 \le \beta$. Then for all $k \in [K]$ and $\omega > 0$*

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

Lemma 4 is a simple modification of Lemma 2 in [RVR13] and its proof can be found in Section 2.6.5. We provide two ways to apply Lemma 4, which can produce regret bounds in term of two different complexity measures. If we invoke Lemma 2 (a) and Lemma 4 with

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

we obtain

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

We can also invoke Lemma 2 (b) and Lemma 4 with

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

and obtain

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

where the first inequality follows from standard martingale concentration.

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

2.6.2 Proof of Corollary 1

Step 1. Bounding the regret by Bellman error By Lemma 3, we can upper bound the cumulative regret by the summation of Bellman error with probability at least $1-\delta$:

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

where (i) follows from standard policy loss decomposition (e.g. Lemma 1 in [JKA$^+$17]).

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 2.

If we invoke Lemma 2 (a) with

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

and Lemma 4 with

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

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

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

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

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

By simple calculation, one can verify it suffices to choose

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

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 $\mathcal{D}_\Delta$.

2.6.3 Proofs of concentration lemmas: Proof of Lemma 2

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

Lemma 5 (Freedman's inequality , e.g., [AHK$^+$14])**.** *Let $(Z_t)_{t\le T}$ 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 $|Z_t| \le R$ almost surely, then for any $\eta \in (0, \frac{1}{R})$ it holds that with probability at least* $1-\delta$

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

Proof. We prove inequality (b) first.

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

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

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

$$\mathbb{E}[X_t(h,f) \mid \mathfrak{F}_{t,h}] = [(f_h - \mathcal{T} f_{h+1})(s_h^t, a_h^t)]^2$$

and

$$\mathrm{Var}[X_t(h,f) \mid \mathfrak{F}_{t,h}] \le \mathbb{E}[(X_t(h,f))^2 \mid \mathfrak{F}_{t,h}] \le 36[(f_h - \mathcal{T} f_{h+1})(s_h^t,a_h^t)]^2 = 36\mathbb{E}[X_t(h,f) \mid \mathfrak{F}_{t,h}].$$

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

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

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

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

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 Assumption 3

$$\begin{aligned}
\sum_{t=1}^{k-1} X_t(h,f^k) =& \sum_{t=1}^{k-1} [f_h^k(s_h^t,a_h^t) - r_h^t - f_{h+1}^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t))]^2 \\
& - \sum_{t=1}^{k-1} [(\mathcal{T} f_{h+1}^k)(s_h^t,a_h^t) - r_h^t - f_{h+1}^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t))]^2 \\
\le & \sum_{t=1}^{k-1} [f_h^k(s_h^t,a_h^t) - r_h^t - f_{h+1}^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t))]^2 \\
& - \inf_{g\in\mathcal{G}} \sum_{t=1}^{k-1} [g_h(s_h^t,a_h^t) - r_h^t - f_{h+1}^k(s_{h+1}^t, \pi_{f^k}(s_{h+1}^t))]^2 \le \beta
\end{aligned}$$

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

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

Therefore,

$$\sum_{t=1}^{k-1} X_t(h, \phi^k) \le \mathcal{O}(k\rho) + \beta \tag{14}$$

Recall inequality (13) implies

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

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

$$\sum_{t=1}^{k-1}[(\phi_h^k - \mathcal{T}\phi_{h+1}^k)(s_h^t, a_h^t)]^2 \le \mathcal{O}(\iota + k\rho + \beta)$$

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

$$\sum_{t=1}^{k-1}[(f_h^k - \mathcal{T}f_{h+1}^k)(s_h^t, a_h^t)]^2 \le \mathcal{O}(\iota + k\rho + \beta)$$

Therefore, we prove inequality (b) in Lemma 2.

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. □

2.6.4 Proofs of concentration lemmas: Proof of Lemma 3

Proof. Let $\mathcal{V}_\rho$ be a ρ-cover of $\mathcal{G}$.

Consider an arbitrary fixed tuple $(k, h, g) \in [K] \times [H] \times \mathcal{G}$. Let

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

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

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

and

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

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

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

By taking a union bound over $[K] \times [H] \times \mathcal{V}_\rho$ and the non-negativity of $\sum_{t=1}^{k}[(g_h - Q_h^\star)(s_h^t, a_h^t)]^2$, we obtain that with probability at least $1-\delta$, for all $(k, h, \psi) \in [K] \times [H] \times \mathcal{V}_\rho$

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

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

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

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 $k \in [K]$. □

2.6.5 Proof of Lemma 4

The proof in this subsection basically follows the same arguments as in Appendix C of [RVR13]. We firstly prove the following proposition which bounds the number of times $|\mathbb{E}_{\mu_t}[\phi_t]|$ can exceed a certain threshold.

Proposition 4. *Given a function class Φ defined on $\mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{\phi_k\}_{k=1}^K \subset \Phi$ and $\{\mu_k\}_{k=1}^K \subset \Pi$ satisfy that for all $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2 \leq \beta$. Then for all $k \in [K]$*

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

Proof of Proposition 4. We first show that if for some k we have $|\mathbb{E}_{\mu_k}[\phi_k]| > \epsilon$, then μ_k is ϵ-dependent on at most β/ϵ^2 disjoint subsequences in $\{\mu_1, \ldots, \mu_{k-1}\}$. By definition of DE dimension, if $|\mathbb{E}_{\mu_k}[\phi_k]| > \epsilon$ 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 $\sum_{t=1}^{\ell}(\mathbb{E}_{\nu_t}[\phi_k])^2 \geq \epsilon^2$. It implies that if μ_k is ϵ-dependent on L disjoint subsequences in $\{\mu_1, \ldots, \mu_{k-1}\}$, we have

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

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 $L = \lceil(\kappa - 1)/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\rceil$ 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 $\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)$ at any point in this process. Therefore, the process stops before or on $j = L\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) + 1 \leq \kappa$.

Fix $k \in [K]$ and let $\{\nu_1, \ldots, \nu_\kappa\}$ be subsequence of $\{\mu_1, \ldots, \mu_k\}$, consisting of elements for which $|\mathbb{E}_{\mu_t}[\phi_t]| > \epsilon$. 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 $(\kappa/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)) - 1$ disjoint subsequences of $\{\nu_1, \ldots, \nu_{j-1}\}$. Therefore, we have $\kappa/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) - 1 \leq \beta/\epsilon^2$ which results in

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

and completes the proof. □

Proof of Lemma 4. Fix $k \in [K]$; let $d = \dim_{\mathrm{DE}}(\Phi, \Pi, \omega)$. Sort the sequence $\{|\mathbb{E}_{\phi_1}[\phi_1]|, \ldots, |\mathbb{E}_{\mu_k}[\phi_k]|\}$ in a decreasing order and denote it by $\{e_1, \ldots, e_k\}$ $(e_1 \geq e_2 \geq \cdots \geq e_k)$

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

For $t \in [k]$, we want to prove that if $e_t > \omega$, then we have $e_t \leq \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Assume $t \in [k]$ satisfies $e_t > \omega$. Then there exists α such that $e_t > \alpha \geq \omega$. By Proposition 4, we have

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

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

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

which completes the proof. □

3 Admissible Bellman Characterization: A Comprehensive Framework for Sample-Efficient Reinforcement Learning with Function Approximation

Reinforcement learning (RL) is a decision-making process that seeks to maximize the expected reward when an agent interacts with the environment [SB18]. Over the past decade, RL has gained increasing attention due to its successes in a wide range of domains, including Atari games [MKS$^+$13], Go game [SHM$^+$16], autonomous driving [YLCT20], Robotics [KBP13], etc. Existing RL algorithms can be categorized into value-based algorithms such as Q-learning [Wat89] and policy-based algorithms such as policy gradient [SMSM99]. They can also be categorized as a model-free approach where one directly models the value function classes, or alternatively, a model-based approach where one needs to estimate the transition probability.

Due to the intractably large state and action spaces that are used to model the real-world complex environment, function approximation in RL has become prominent in both algorithm design and theoretical analysis. It is a pressing challenge to design sample-efficient RL algorithms with general function approximations. In the special case where the underlying Markov Decision Processes (MDPs) enjoy certain linear structures, several lines of works have achieved polynomial sample complexity and/or $\sqrt{T}$ regret guarantees under either model-free or model-based RL settings. For linear MDPs where the transition probability and the reward function admit linear structure, [YW19] developed a variant of Q-learning when granted access to a generative model, [JYWJ20] proposed an LSVI-UCB algorithm with a $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret bound and [ZLKB20a] further extended the MDP model and improved the regret to $\tilde{\mathcal{O}}(dH\sqrt{T})$. Another line of work considers linear mixture MDPs [YW20, MJTS20, JYSW20, ZGS21], where the transition probability can be represented by a mixture of base models. In [ZGS21], an $\tilde{\mathcal{O}}(dH\sqrt{T})$ minimax optimal regret was achieved with weighted linear regression and a Bernstein-type bonus. Other structural MDP models include the block MDPs [DLWZ19] and FLAMBE [AKKS20][6], to mention a few.

In a more general setting, however, there is still a gap between the plethora of MDP models and sample-efficient RL algorithms that can learn the MDP model with function approximation. The question remains open as to what constitutes minimal structural assumptions that admit sample-efficient reinforcement learning. To answer this question, there are several lines of work along this direction. [RVR13, OVR14] proposed an structural condition named eluder dimension, and [WSY20b] extended the LSVI-UCB for general linear function classes with small eluder dimension. Another line of works proposed low-rank structural conditions, including Bellman rank [JKA+17, DPWZ20] and Witness rank [SJK+19]. Recently, [JLM21] proposed a complexity called Bellman eluder (BE) dimension, which unifies low Bellman rank and low eluder dimension. Concurrently, [DKL+21] proposed Bilinear Classes, which can be applied to a variety of loss estimators beyond vanilla Bellman error. Very recently, [FKQR21] proposed Decision-Estimation Coefficient (DEC), which is a necessary and sufficient condition for sample-efficient interactive learning. To apply DEC to RL, they proposed a RL class named Bellman Representability, which can be viewed as a generalization of the Bilinear Class. Nevertheless, [SJK+19] is limited to model-based RL, and [JLM21] is restricted to model-free RL. The only frameworks that can unify both model-based and model-free RL are [DKL+21] and [FKQR21], but their sample complexity results when restricted to special MDP instances do not always match the best-known results. Viewing the above gap, we aim to answer the following question: ***Is there a unified framework that includes all model-free and model-based RL classes while maintaining sharp sample efficiency?***

In this paper, we tackle this challenging question and give a *nearly* affirmative answer to it. We summarize our contributions as follows:

- We propose a general framework called Admissible Bellman Characterization (ABC) that covers a wide set of structural assumptions in both model-free and model-based RL, such as linear MDPs, FLAMBE, linear mixture MDPs, kernelized nonlinear regulator [KKL+20], etc. Furthermore, our framework encompasses comparative structural frameworks such as the low Bellman eluder dimension and low Witness rank.

- Under our ABC framework, we design a novel algorithm, OPtimization-based ExploRation with Approximation (OPERA), based on maximizing the value function while constrained in a small confidence region around the model minimizing the estimation function.

- We apply our framework to several specific examples that are known to be not sample-efficient with value-based algorithms. For the kernelized nonlinear regulator (KNR), our framework is the

[6] In this paper, we use FLAMBE to refer to both the algorithm and the low-rank MDP with unknown feature mappings.

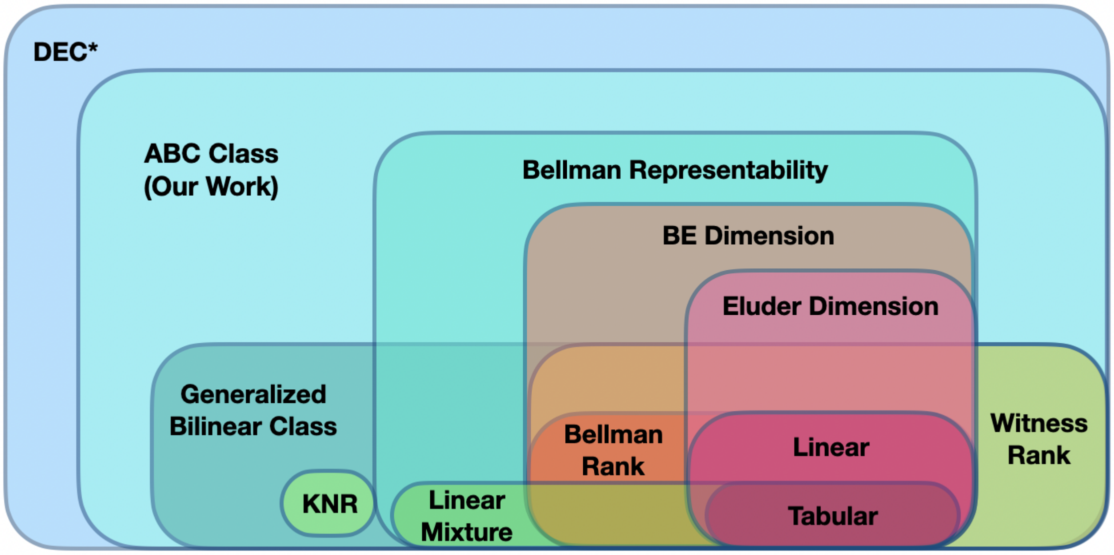

Figure 1. Venn-Diagram Visualization of Prevailing Sample-Efficient RL Classes. As by far the richest concept, the DEC framework is both a necessary and sufficient condition for sample-efficient interactive learning. BE dimension is a rich class that subsumes both low Bellman rank and low eluder dimension and addresses almost all model-free RL classes. The generalized Bilinear Class captures model-based RL settings including KNRs, linear mixture MDPs and low Witness rank MDPs, yet precludes some eluder-dimension based models. Bellman Representability is another unified framework that subsumes the vanilla bilinear classes but fails to capture KNRs and low Witness rank MDPs. Our ABC class encloses both generalized Bilinear Class and Bellman Representability and subsumes almost all known solvable MDP cases, with the exception of the Q^* state-action aggregation and deterministic linear Q^* MDP models, which neither Bilinear Class nor our ABC class captures.

first general framework to derive a $\sqrt{T}$ regret-bound result. For the witness rank, our framework yields a sharper sample complexity with a mild additional assumption compared to prior works.

We visualize and compare prevailing sample-efficient RL frameworks and ours in Figure 1. We can see that both the general Bilinear Class and our ABC frameworks capture most existing MDP classes, including the low Witness rank and the KNR models. In Table 1, we compare our ABC framework with other structural RL frameworks in terms of the model coverage and sample complexity.

3.1 Related Work

We discuss in this subsection the related work, providing comparisons with previous frameworks based on both coverage and sharpness of sample complexity.

Tabuler RL. Tabular RL considers MDPs with finite state space $\mathcal{S}$ and action space $\mathcal{A}$. This setting has been extensively studied [AJO08, DB15, BT02, AJ17, AOM17, ZB19, ZZJ20] and the minimax-optimal regret bound is proved to be $\tilde{O}(\sqrt{H^2|\mathcal{S}||\mathcal{A}|T})$ [JAZBJ18, DMKV21]. The minimax optimal bounds suggests that the tabular RL is information-theoretically hard for large $|\mathcal{S}|$ and $|\mathcal{A}|$. Therefore, in order to deal with high-dimensional state-action space arose in many real-world applications, more advanced structural assumptions that enable function approximation are in demand.

	Bilinear Class	Low BE Dimension	DEC and Bellman Representability	ABC Class (with Low FE Dimension)
Linear MDPs [YW19, JYWJ20]	d^3H^4/ϵ^2	d^2H^2/ϵ^2	d^3H^3/ϵ^2	d^2H^2/ϵ^2
Linear Mixture MDPs [MJTS20]	d^3H^4/ϵ^2	✘	d^3H^3/ϵ^2	d^2H^2/ϵ^2
Bellman Rank [JKA+17]	$d^2H^5\lvert\mathcal{A}\rvert/\epsilon^2$	$dH^2\lvert\mathcal{A}\rvert/\epsilon^2$	$d^2H^3\lvert\mathcal{A}\rvert/\epsilon^2$	$dH^2\lvert\mathcal{A}\rvert/\epsilon^2$
Eluder Dimension [WSY20b]	✘	$\dim_{\mathrm{E}} H^2/\epsilon^2$	$\dim_{\mathrm{E}}^2 H^3/\epsilon^2$	$\dim_{\mathrm{E}} H^2/\epsilon^2$
Witness Rank [SJK+19]	—	✘	—	$W_\kappa H^2\lvert\mathcal{A}\rvert/\epsilon^2$
Low Occupancy Complexity [DKL+21]	d^3H^4/ϵ^2	d^2H^2/ϵ^2	d^3H^3/ϵ^2	d^2H^2/ϵ^2
Kernelized Nonlinear Regulator [KKL+20]	—	✘	—	$d_\phi^2 d_s H^4/\epsilon^2$
Linear Q^*/V^* [DKL+21]	d^3H^4/ϵ^2	d^2H^2/ϵ^2	d^3H^3/ϵ^2	d^2H^2/ϵ^2

Table 1. Comparison of sample complexity for different MDP models under different RL frameworks. "—" indicates that the original work of framework does not provide an explicit sample complexity result for that model (although can be computed in principle), "✘" indicates the model is not included in the framework for complexity analysis. For models with the linear structure on a d-dimensional space, we present the sample complexity in terms of d. For models with their own complexity measures, we use W_κ to denote the witness rank, $\dim_{\mathrm{E}}$ the eluder dimension, d_ϕ the dimension of $\mathcal{H}$ in KNR and d_s the dimension number of the state space of KNR. The dependency on ρ-covering number is deliberately ignored for Bellman rank, eluder dimension, and the witness rank.

Complexity Measures for Statistical Learning. In classic statistical learning, a variety of complexity measures have been proposed to upper bound the sample complexity required for achieving a certain accuracy, including VC Dimension [Vap99], covering number [Pol12], Rademacher Complexity [BM02], sequential Rademacher complexity [RST10] and Littlestone dimension [Lit88]. However, for reinforcement learning, it is a major challenge to find such general complexity measures that can be used to analyze the sample complexity under a general framework.

RL with Linear Function Approximation. A line of work studied the MDPs that can be represented as a linear function of some given feature mapping. Under certain completeness conditions, the proposed algorithms can enjoy sample complexity/regret scaling with the dimension of the feature mapping rather than $|\mathcal{S}|$ and $|\mathcal{A}|$. One such class of MDPs is linear MDPs [JYWJ20, WWDK21, NPB20], where the transition probability function and reward function are linear in some feature mapping over state-action pairs. [ZLKB20a, ZLKB20b] studied MDPs under a weaker assumption called low inherent Bellman error, where the value functions are nearly linear w.r.t. the feature mapping. Another class of MDPs is linear mixture MDPs [MJTS20, JYSW20, AJS+20, ZHG21, CYJW20b], where the transition probability kernel is a linear mixture of a number of basis kernels. The above paper assumed that feature vectors are known in the MDPs with linear approximation while [AKKS20] studied a harder setting where both the feature and parameters are unknown in the linear model.

RL with General Function Approximation. Beyond the linear setting, a recent line of research attempted to unify existing sample-efficient approaches with general function approximation. [OVR14] proposed an structural condition named eluder dimension. [WSY20b] further proposed an efficient algorithm LSVI-UCB for general linear function classes with small eluder dimension. Another line of works proposed low-rank structural conditions, including Bellman rank [JKA+17, DPWZ20] and Witness rank [SJK+19]. [YJW+20b] studied the MDPs with a structure where the action-value function can be represented by a kernel function or an over-parameterized neural network. Recently, [JLM21] proposed a complexity called Bellman eluder (BE) dimension. The RL problems with low BE dimension subsume the problems with low Bellman rank and low eluder dimension. Simultaneously [DKL+21] proposed Bilinear Classes, which can be applied to a variety of loss estimators beyond vanilla Bellman error, but with possibly worse sample complexity. Very recently, [FKQR21] proposed Decision-Estimation Coefficient (DEC), which is a necessary and sufficient condition for sample-efficient interactive learning. To apply DEC to reinforcement learning, [FKQR21] further proposed a RL class named Bellman Representability, which can be viewed as a generalization of the Bilinear Class.

Notation. For a state-action sequence $s_1, a_1, \dots, s_H$ in our given context, we use $\mathscr{J}_h := \sigma(s_1, a_1, \dots, s_h)$ to denote the σ-algebra generated by trajectories up to step $h \in [H]$. Let π_f denote the policy of following the max-Q strategy induced by hypothesis f. When $f = f^i$ we write π_{f^i} as π^i for notational simplicity. We write $s_h \sim \pi$ to indicate the state-action sequence are generated by step $h \in [H]$ by following policy $\pi(\cdot \mid s)$ and transition probabilities $\mathbb{P}(\cdot \mid s, a)$ of the underlying MDP model M. We also write $a_h \sim \pi$ to mean $a_h \sim \pi(\cdot \mid s_h)$ for the h-th step. Let $\|\cdot\|_2$ denote the ℓ_2-norm and $\|\cdot\|_\infty$ the ℓ_∞-norm of a given vector. Other notations will be explained at their first appearances.

3.2 Preliminaries

We consider a finite-horizon, episodic Markov Decision Process (MDP) defined by the tuple $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$, where $\mathcal{S}$ is the space of feasible states, $\mathcal{A}$ is the action space. H is the horizon in each episode defined by the number of action steps in one episode, and $\mathbb{P} := \{\mathbb{P}_h\}_{h\in[H]}$ is defined for every $h \in [H]$ as the transition probability from the current state-action pair $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ to the next state $s_{h+1} \in \mathcal{S}$. We use $r_h(s, a) \geq 0$ to denote the reward received at step $h \in [H]$ when taking action a at state s and assume throughout this paper that for any possible trajectories, $\sum_{h=1}^{H} r_h(s_h, a_h) \in [0, 1]$.

A deterministic policy π is a sequence of functions $\{\pi_h : \mathcal{S} \mapsto \mathcal{A}\}_{h\in[H]}$, where each π_h specifies a strategy at step h. Given a policy π, the action-value function is defined to be the expected cumulative rewards where the expectation is taken over the trajectory distribution generated by $\{(\mathbb{P}_h(\cdot \mid s_h, a_h), \pi_h(\cdot \mid s_h))\}_{h\in[H]}$ as

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

Similarly, we define the state-value function for policy π as the expected cumulative rewards as

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

We use π^* to denote the optimal policy that satisfies $V_h^{\pi^*}(s) = \max_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ [Put14]. For simplicity, we abbreviate $V_h^{\pi^*}$ as V_h^* and $Q_h^{\pi^*}$ as Q_h^*. Moreover, for a sequence of value functions $\{Q_h\}_{h\in[H]}$, the Bellman operator at step h is defined as:

$$\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} \max_{a'\in\mathcal{A}} Q_{h+1}(s',a')$$

We also call $Q_h - (\mathcal{T}_h Q_{h+1})$ the Bellman error (or Bellman residual). The goal of an RL algorithm is to find an ϵ-optimal policy such that $V_1^\pi(s_1) - V_1^*(s_1) \le \epsilon$. For an RL algorithm that updates the policy π^t for T iterations, the cumulative regret is defined as

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \left[V_1^*(s_1) - V_1^{\pi^t}(s_1) \right]$$

Hypothesis Classes. Following [DKL$^+$21], we define the hypothesis class for both model-free and model-based RL. Generally speaking, a hypothesis class is a set of functions that are used to estimate the value functions (for model-free RL) or the transitional probability and reward (for model-based RL). Specifically, a hypothesis class $\mathcal{F}$ on a finite-horizon MDP is the Cartesian product of H hypothesis classes $\mathcal{F} := \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$ in which each hypothesis $f = \{f_h\}_{h\in[H]} \in \mathcal{F}$ can be identified by a pair of value functions $\{Q_f, V_f\} = \{Q_{h,f}, V_{h,f}\}_{h\in[H]}$. Based on the value function pair, it is natural to introduce the greedy policy $\pi_{h,f}(s) = \arg\max_{a\in\mathcal{A}} Q_{h,f}(s,a)$ at each step $h \in [H]$, and the corresponding $\pi_f(s)$ as the sequence of time-dependent policies $\{\pi_{h,f}\}_{h=0}^{H-1}$.

An example of a model-free hypothesis class is defined by a sequence of action-value function $\{Q_{h,f}\}_{h\in[H]}$. The corresponding state-value function is given by:

$$V_{h,f}(s) = \mathbb{E}_{a\sim\pi_{h,f}}\left[Q_{h,f}(s,a)\right]$$

In another example that falls under the model-based RL setting, where for each hypothesis $f \in \mathcal{F}$ we have the knowledge of the transition matrix $\mathbb{P}_f$ and the reward function r_f. We define the value function $Q_{h,f}$ corresponding to hypothesis f as the optimal value function following $M_f := (\mathbb{P}_f, r_f)$:

$$Q_{h,f}(s,a) = Q_{h,M_f}^*(s,a) \qquad \text{and} \quad V_{h,f}(s) = V_{h,M_f}^*(s)$$

We also need the following realizability assumption that requires the true model M_{f^*} (model-based RL) or the optimal value function f^* (model-free RL) to belong to the hypothesis class $\mathcal{F}$.

Assumption 4 (Realizability)**.** *For an MDP model M and a hypothesis class $\mathcal{F}$, we say that the hypothesis class $\mathcal{F}$ is* realizable with respect to M *if there exists a $f^* \in \mathcal{F}$ such that for any $h \in [H]$, $Q_h^*(s,a) = Q_{h,f^*}(s,a)$. We call such f^* an* optimal hypothesis.

This assumption has also been made in the Bilinear Classes [DKL$^+$21] and low Bellman eluder dimension frameworks [JLM21]. We also define the *ϵ-covering number* of $\mathcal{F}$ under a well-defined metric ρ of a hypothesis class $\mathcal{F}$:[7]

[7] For example for model-free cases where f, g are value functions, $\rho(f,g) = \max_{h\in[H]} \|f_h - g_h\|_\infty$. For model-based RL where f, g are transition probabilities, we adopt $\rho(\mathbb{P},\mathbb{Q}) = \max_{h\in[H]} \int(\sqrt{d\mathbb{P}_h} - \sqrt{d\mathbb{Q}_h})^2$ which is the maximal (squared) Hellinger distance between two probability distribution sequences.

Definition 8 (ϵ-Covering Number of Hypothesis Class)**.** *For any $\epsilon > 0$ and a hypothesis class $\mathcal{F}$, we use $N_{\mathcal{F}}(\epsilon)$ to denote the ϵ-covering number, which is the smallest possible cardinality of (an ϵ-cover) $\mathcal{F}_\epsilon$ such that for any $f \in \mathcal{F}$ there exists a $f' \in \mathcal{F}_\epsilon$ such that $\rho(f, f') \leq \epsilon$.*

Functional Eluder Dimension. We proceed to introduce our new complexity measure, *functional eluder dimension*, which generalizes the concept of *eluder dimension* firstly proposed in bandit literature [RVR13, RVR14]. It has since become a widely used complexity measure for function approximations in RL [WSY20b, AJS$^+$20, JLM21, FKQR21]. Here we revisit its definition:

Definition 9 (Eluder Dimension)**.** *For a given space $\mathcal{X}$ and a class $\mathcal{F}$ of functions defined on $\mathcal{X}$, the* eluder dimension $\dim_{\mathcal{E}}(\mathcal{F}, \epsilon)$ *is the length of the existing longest sequence $x_1, \ldots, x_n \in \mathcal{X}$ satisfying for some $\epsilon' \geq \epsilon$ and any $2 \leq t \leq n$, there exist $f_1, f_2 \in \mathcal{F}$ such that $\sqrt{\sum_{i=1}^{t-1} (f_1(x_i) - f_2(x_i))^2} \leq \epsilon'$ while $|f_1(x_t) - f_2(x_t)| > \epsilon'$.*

The eluder dimension is usually applied to the state-action space $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and the corresponding value function class $\mathcal{F} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ [JLM21, WSY20b]. We extend the concept of eluder dimension as a complexity measure of the hypothesis class, namely, the *functional eluder dimension*, which is formally defined as follows.

Definition 10 (Functional Eluder Dimension)**.** *For a given hypothesis class $\mathcal{F}$ and a function G defined on $\mathcal{F} \times \mathcal{F}$, the* functional eluder dimension (FE dimension) $\dim_{FE}(\mathcal{F}, G, \epsilon)$ *is the length of the existing longest sequence $f_1, \ldots, f_n \in \mathcal{F}$ satisfying for some $\epsilon' \geq \epsilon$ and any $2 \leq t \leq n$, there exists $g \in \mathcal{F}$ such that $\sqrt{\sum_{i=1}^{t-1} (G(g, f_i))^2} \leq \epsilon'$ while $|G(g, f_t)| > \epsilon'$. Function G is dubbed as the* coupling function*.*

The notion of functional eluder dimension introduced in Definition 10 is generalizable in a straightforward fashion to a sequence $G := \{G_h\}_{h \in [H]}$ of coupling functions: we simply set $\dim_{\mathrm{FE}}(\mathcal{F}, G, \epsilon) = \max_{h \in [H]} \dim_{\mathrm{FE}}(\mathcal{F}, G_h, \epsilon)$ to denote the FE dimension of $\{G_h\}_{h \in [H]}$. The Bellman eluder (BE) dimension recently proposed by [JLM21] is in fact a special case of FE dimension with a specific choice of coupling function sequence.[8] As will be shown later, our framework based on FE dimension with respect to the corresponding coupling function captures many specific MDP instances such as the kernelized nonlinear regulator (KNR) [KKL$^+$20] and the generalized linear Bellman complete model [WWDK21], which are not captured by the framework of low BE dimension. As we will see in later sections, introducing the concept of FE dimension allows the coverage of a strictly wider range of MDP models and hypothesis classes.

3.3 Admissible Bellman Characterization Framework

In this section, we first introduce the framework of Admissible Bellman Characterization (ABC) which covers a wide range of MDPs in Section 3.3.1, and then introduce the notion of Decomposable Estimation Function (DEF) which extends the Bellman error. We discuss MDP instances that belong to the ABC class with low FE dimension in Section 3.3.2.

[8] Indeed, when the coupling function is chosen as the expected Bellman error $G_h(g, f) := \mathbb{E}_{\pi_{h,f}}(Q_{h,g} - \mathcal{T}_h Q_{g,h+1})$ where $\mathcal{T}_h$ denotes the Bellman operator, we recover the definition of BE dimension [JLM21], i.e. $\dim_{\mathrm{FE}}(\mathcal{F}, G, \epsilon) = \dim_{\mathrm{BE}}(\mathcal{F}, G, \epsilon)$.

3.3.1 Admissible Bellman Characterization

Given an MDP M, a sequence of states and actions $s_1, a_1, \ldots, s_H$, two hypothesis classes $\mathcal{F}$ and $\mathcal{G}$ satisfying the realizability assumption (Assumption 4),[9] and a *discriminator function class* $\mathcal{V} = \{v(s,a,s') : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}\}$, the *estimation function* $\ell = \{\ell_{h,f'}\}_{h\in[H],f'\in\mathcal{F}}$ is an $\mathbb{R}^{d_s}$-valued function defined on the set consisting of $o_h := (s_h, a_h, s_{h+1}) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, $f \in \mathcal{F}$, $g \in \mathcal{G}$ and $v \in \mathcal{V}$ and serves as a surrogate loss function of the Bellman error. Note that our estimation function is a vector-valued function, and is more general than the scalar-valued estimation function (or discrepancy function) used in [FKQR21, DKL+21]. The *discriminator* v originates from the function class the Integral Probability Metrics (IPM) [Mül97] is taken with respect to (as a metric between two distributions), and is also used in the definition of Witness rank [SJK+19].

We use a coupling function $G_{h,f^*}(f,g)$ defined on $\mathcal{F}\times\mathcal{F}$ to characterize the interaction between two hypotheses $f, g \in \mathcal{F}$. The subscript f^* is an indicator of the *true model* and is by default unchanged throughout the context. When the two hypotheses coincide, our characterization of the coupling function reduces to the Bellman error.

Definition 11 (Admissible Bellman Characterization)**.** *Given an MDP M, two hypothesis classes $\mathcal{F}, \mathcal{G}$ satisfying the realizability assumption (Assumption 4) and $\mathcal{F} \subset \mathcal{G}$, an estimation function $\ell_{h,f'} : (\mathcal{S}\times\mathcal{A}\times\mathcal{S}) \times \mathcal{F}\times\mathcal{G}\times\mathcal{V} \to \mathbb{R}^{d_s}$, an operation policy π_{op} and a constant $\kappa \in (0,1]$, we say that G is an* admissible Bellman characterization *of $(M, \mathcal{F}, \mathcal{G}, \ell)$ if the following conditions hold:*

(i) ***(Dominating Average Estimation Function)*** *For any $f, g \in \mathcal{F}$*

$$\max_{v\in\mathcal{V}} \mathbb{E}_{s_h\sim\pi_g, a_h\sim\pi_{op}} \left\|\mathbb{E}_{s_{h+1}}\left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right]\right\|_2^2 \geq \left(G_{h,f^*}(f,g)\right)^2$$

(ii) ***(Bellman Dominance)*** *For any $(h,f) \in [H]\times\mathcal{F}$,*

$$\kappa \cdot \left|\mathbb{E}_{s_h,a_h\sim\pi_f}\left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1})\right]\right| \leq |G_{h,f^*}(f,f)|$$

We further say $(M, \mathcal{F}, \mathcal{G}, \ell, G)$ is an ABC class *if G is an admissible Bellman characterization of $(M, \mathcal{F}, \mathcal{G}, \ell)$.*

In Definition 11, one can choose either $\pi_{\mathrm{op}} = \pi_g$ or $\pi_{\mathrm{op}} = \pi_f$. We refer readers to Section 3.5.3 for further explanations on π_{op}. The ABC class is quite general and de facto covers many existing MDP models; see Section 3.3.2 for more details.

Comparison with Existing MDP Classes. Here we compare our ABC class with three recently proposed MDP structural classes: Bilinear Classes [DKL+21], low Bellman eluder dimension [JLM21], and Bellman Representability [FKQR21].

- *Bilinear Classes.* Compared to the structural framework of Bilinear Class in [DKL+21, Definition 4.3], Definition 11 of Admissible Bellman Characterization does not require a bilinear structure and recovers the Bilinear Class when we set $G_{h,f^*}(f,g) = \langle W_h(g) - W_h(f^*), X_h(f)\rangle$. Our ABC class is strictly broader than the Bilinear Class since the latter does not capture low eluder dimension models, and our ABC class does. In addition, the ABC class admits an estimation function that is *vector-valued*, and the corresponding algorithm achieves a $\sqrt{T}$-regret for KNR case while the BiLin-UCB algorithm for Bilinear Classes [DKL+21] does not.

[9] We assume $\mathcal{F} \subseteq \mathcal{G}$ throughout this paper and in the general case where $\mathcal{F} \not\subseteq \mathcal{G}$, we overload $\mathcal{G} := \mathcal{F} \cup \mathcal{G}$.

- *Low Bellman Eluder Dimension.* Definition 11 subsumes the MDP class of low BE dimension when $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1})$. Moreover, our definition unifies the V-type and Q-type problems under the same framework by the notion of π_{op}. We will provide a more detailed discussion on this in Section 3.3.2. Our extension from the concept of the Bellman error to estimation function (i.e. the surrogate of the Bellman error) enables us to accommodate model-based RL for linear mixture MDPs, KNR model, and low Witness rank.

- *Bellman Representability.* [FKQR21] proposed DEC framework which is another MDP class that unifies both the Bilinear Class and the low BE dimension. Indeed, our ABC framework introduced in Definition 11 shares similar spirits with the Bellman Representability Definition F.1 in [FKQR21]. Nevertheless, our framework and theirs bifurcate from the base point: our work studies an optimization-based exploration instead of the posterior sampling-based exploration in [FKQR21]. Structurally different from their DEC framework, our ABC requires estimation functions to be vector-valued, introduces the discriminator function v, and imposes the weaker Bellman dominance property (i) in Definition 11 than the corresponding one as in [FKQR21, Eq. (166)]. In total, this allows broader choices of coupling function G as well as our ABC class (with low FE dimension) to include as special instances both low Witness rank and KNR models, which are not captured in [FKQR21].

Decomposable Estimation Function. Now we introduce the concept of *decomposable estimation function*, which generalizes the Bellman error in earlier literature and plays a pivotal role in our algorithm design and analysis.

Definition 12 (Decomposable Estimation Function)**.** *A decomposable estimation function $\ell : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \mathcal{F} \times \mathcal{G} \times \mathcal{V} \to \mathbb{R}^{d_s}$ is a function with bounded ℓ_2-norm such that the following two conditions hold:*

(i) ***(Decomposability)*** *There exists an operator that maps between two hypothesis classes $\mathcal{T}(\cdot) : \mathcal{F} \to \mathcal{G}$*[10] *such that for any $f \in \mathcal{F}$, $(h, f', g, v) \in [H] \times \mathcal{F} \times \mathcal{G} \times \mathcal{V}$ and all possible o_h*

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

Moreover, if $f = f^*$, then $\mathcal{T}(f) = f^*$ holds.

(ii) ***(Global Discriminator Optimality)*** *For any $f \in \mathcal{F}$ there exists a global maximum $v_h^*(f) \in \mathcal{V}$ such that for any $(h, f', g, v) \in [H] \times \mathcal{F} \times \mathcal{G} \times \mathcal{V}$ and all possible o_h*

$$\left\|\mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v_h^*(f)) \mid s_h, a_h\right]\right\|_2 \geq \left\|\mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right]\right\|_2$$

Compared with the discrepancy function or estimation function used in prior work [DKL+21, FKQR21], our estimation function (EF) admits the unique properties listed as follows:

(a) Our EF enjoys a decomposable property inherited from the Bellman error — intuitively speaking, the decomposability can be seen as a property shared by all functions in the form of the difference of a $\mathcal{J}_h$-measurable function and a $\mathcal{J}_{h+1}$-measurable function;

(b) Our EF involves a discriminator class and assumes the global optimality of the discriminator on all (s_h, a_h) pairs;

[10] The decomposability item (i) in Definition 12 directly implies that a Generalized Completeness condition similar to Assumption 14 of [JLM21] holds.

(c) Our EF is a vector-valued function which is more general than a scalar-valued estimation function (or the discrepancy function).

We remark that when $f = g$, $\mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right]$ measures the discrepancy in optimality between f and f^*. In particular, when $f = f^*$, $\mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f^*_{h+1}, f^*_h, v) \mid s_h, a_h\right] = 0$. Consider a special case when $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$. Then the decomposability (i) in Definition 12 reduces to

$$\begin{aligned}&[Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})] - [Q_{h,g}(s_h, a_h) - (\mathcal{T}_h V_{h+1})(s_h, a_h)]\\ &= (\mathcal{T}_h V_{h+1})(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\end{aligned}$$

In addition, we make the following Lipschitz continuity assumption on the estimation function.

Assumption 5 (Lipschitz Estimation Function)**.** *There exists a $L > 0$ such that for any $(h, f', f, g, v) \in [H] \times \mathcal{F} \times \mathcal{F} \times \mathcal{G} \times \mathcal{V}$, $(\tilde f, \tilde g, \tilde v, \tilde f') \in \mathcal{F} \times \mathcal{G} \times \mathcal{V} \times \mathcal{F}$ and all possible o_h,*

$$\left\|\ell_{h,f'}(\cdot, f, g, v) - \ell_{h,f'}(\cdot, \tilde f, g, v)\right\|_\infty \le L\rho(f, \tilde f), \qquad \left\|\ell_{h,f'}(\cdot, f, g, v) - \ell_{h,f'}(\cdot, f, \tilde g, v)\right\|_\infty \le L\rho(g, \tilde g)$$
$$\left\|\ell_{h,f'}(\cdot, f, g, v) - \ell_{h,f'}(\cdot, f, g, \tilde v)\right\|_\infty \le L\left\|v - \tilde v\right\|_\infty, \quad \left\|\ell_{h,f'}(\cdot, f, g, v) - \ell_{h,\tilde f'}(\cdot, f, g, v)\right\|_\infty \le L\rho(f', \tilde f')$$

Note that we have omitted the subscript h of hypotheses in Assumption 5 for notational simplicity. We further define the induced estimation function class as $\mathcal{L} = \{\ell_{h,f'}(\cdot, f, g, v) : (h, f', f, g, v) \in [H] \times \mathcal{F} \times \mathcal{F} \times \mathcal{G} \times \mathcal{V}\}$. We can show that under Assumption 5, the covering number of the induced estimation function class $\mathcal{L}$ can be upper bounded as $N_{\mathcal{L}}(\epsilon) \le N^2_{\mathcal{F}}(\frac{\epsilon}{4L}) N_{\mathcal{G}}(\frac{\epsilon}{4L}) N_{\mathcal{V}}(\frac{\epsilon}{4L})$, where $N_{\mathcal{F}}(\epsilon), N_{\mathcal{G}}(\epsilon), N_{\mathcal{V}}(\epsilon)$ are the ϵ-covering number of $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{V}$, respectively. Later in our theoretical analysis in Section 3.4, our regret upper bound will depend on the growth rate of the covering number or the *metric entropy*, $\log N_{\mathcal{L}}(\epsilon)$.

3.3.2 MDP Instances in the ABC Class

In this subsection, we present a number of MDP instances that belong to ABC class with low FE dimension. As we have mentioned before, for all special cases with $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1})$, both conditions in Definition 11 are satisfied automatically with $G_{h,f^*}(f, g) = \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{\mathrm{op}}}\left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1})\right]$. The FE dimension under this setting recovers the the BE dimension. Thus, all model-free RL models with low BE dimension [JLM21] belong to our ABC class with low FE dimension. In the rest of this subsection, our focus shifts to the model-based RLs that belong to the ABC class: linear mixture MDPs, low Witness rank, and kernelized nonlinear regulator.

Linear Mixture MDPs. We start with a model-based RL with a linear structure called the *linear mixture MDP* [MJTS20, AJS+20, ZHG21]. For known transition and reward feature mappings $\phi(s, a, s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{H}$, $\psi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{H}$ taking values in a Hilbert space $\mathcal{H}$ and an unknown $\theta^* \in \mathcal{H}$, a linear mixture MDP assumes that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $h \in [H]$, the transition probability $\mathbb{P}_h(s' \mid s, a)$ and the reward function $r(s, a)$ are linearly parameterized as

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \theta^*_h, \phi(s, a, s') \right\rangle, \qquad r(s, a) = \left\langle \theta^*_h, \psi(s, a) \right\rangle$$

In this case, we choose $\mathcal{F}_h = \mathcal{G}_h = \{\theta_h \in \mathcal{H}\}$ and have the following proposition, which shows that linear mixture MDPs belong to the ABC class with low FE dimension:

Proposition 5 (Linear Mixture MDP ⊂ ABC with Low FE Dimension)**.** *The linear mixture MDP model belongs to the ABC class with estimation function*

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^{\top}\left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s')V_{h+1,f'}(s')\right] - r_h - V_{h+1,f'}(s_{h+1}) \tag{16}$$

and coupling function $G_{h,f^*}(f,g) = \left\langle \theta_{h,g} - \theta_h^*, \mathbb{E}_{s_h,a_h\sim\pi_f}\left[\psi(s_h,a_h) + \sum_{s'}\phi(s_h,a_h,s')V_{h+1,f}(s')\right]\right\rangle$*. Moreover, it has a low FE dimension.*

Low Witness Rank. The following definition is a generalized version of the witness rank in [SJK+19], where we require the discriminator class $\mathcal{V}$ to be *complete*, meaning that the assemblage of functions by taking the value at (s,a) from different functions also belongs to $\mathcal{V}$. We will elaborate this assumption later in Section B.1.2.

Definition 13 (Witness Rank)**.** *For an MDP M, a given symmetric and complete discriminator class $\mathcal{V} = \{\mathcal{V}_h\}_{h\in[H]}$, $\mathcal{V}_h \subset \mathcal{S}\times\mathcal{A}\times\mathcal{S} \mapsto \mathbb{R}$ and a hypothesis class $\mathcal{F}$, we define the Witness rank of M as the smallest d such that for any two hypotheses $f, g \in \mathcal{F}$, there exist two mappings $X_h : \mathcal{F} \to \mathbb{R}^d$ and $W_h : \mathcal{F} \to \mathbb{R}^d$ and a constant $\kappa \in (0,1]$, the following inequalities hold for all $h \in [H]$:*

$$\max_{v\in\mathcal{V}_h} \mathbb{E}_{s_h\sim\pi_f,a_h\sim\pi_g}\left[\mathbb{E}_{\tilde{s}\sim g_h} v(s_h,a_h,\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h,a_h,\tilde{s})\right] \geq \langle W_h(g), X_h(f)\rangle \tag{17}$$

$$\kappa\cdot\mathbb{E}_{s_h\sim\pi_f,a_h\sim\pi_g}\left[\mathbb{E}_{\tilde{s}\sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} V_{h+1,g}(\tilde{s})\right] \leq \langle W_h(g), X_h(f)\rangle \tag{18}$$

The following proposition shows that low Witness rank models belong to our ABC class with low FE dimension.

Proposition 6 (Low Witness Rank ⊂ ABC with Low FE Dimension)**.** *The low Witness rank model belongs to the ABC class with estimation function*

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s}\sim g_h} v(s_h,a_h,\tilde{s}) - v(s_h,a_h,s_{h+1}) \tag{19}$$

and coupling function $G_{h,f^*}(f,g) = \langle W_h(g), X_h(f)\rangle$. Moreover, it has a low FE dimension.

Kernelized Nonlinear Regulator. The *kernelized nonlinear regulator (KNR)* proposed recently by [MJR22, KKL+20] models a nonlinear control dynamics on an RKHS $\mathcal{H}$ of finite or countably infinite dimensions. Under the KNR setting, given current s_h, a_h at step $h \in [H]$ and a known feature mapping $\phi : \mathcal{S}\times\mathcal{A} \to \mathcal{H}$, the subsequent state obeys a Gaussian distribution with mean vector $U_h^*\phi(s_h,a_h)$ and homoskedastic covariance $\sigma^2 I$, where $\left\{U_h^* \in \mathbb{R}^{d_s}\times\mathcal{H}\right\}_{h\in[H]}$ are true model parameters and d_s is the dimension of the state space. Mathematically, we have for each $h = 1,\ldots,H$,

$$s_{h+1} = U_h^*\phi(s_h,a_h) + \epsilon_{h+1}, \quad \text{where } \epsilon_{h+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2 I) \tag{20}$$

Furthermore, we assume bounded reward $r \in [0,1]$ and uniformly bounded feature map $\|\phi(s,a)\|_2 \leq B$. The following proposition shows that KNR belongs to the ABC class with low FE dimension.

Proposition 7 (KNR $\subset$ ABC with Low FE Dimension)**.** *KNR belongs to the ABC class with estimation function*

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = U_{h,g}\phi(s_h, a_h) - s_{h+1} \tag{21}$$

and coupling function $G_{h,f^*}(f,g) := \sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g}\left\|(U_{h,f} - U_h^*)\phi(s_h,a_h)\right\|_2^2}$*. Moreover, it has a low FE dimension.*

Although the dimension of the RKHS $\mathcal{H}$ can be infinite, our complexity analysis depends solely on its effective dimension d_ϕ.

3.4 Algorithm and Main Results

In this section, we present an RL algorithm for the ABC class. Then we present the regret bound of this algorithm, along with its implications to several MDP instances in the ABC class.

3.4.1 OPERA Algorithm

We first present the *OPtimization-based ExploRation with Approximation (OPERA)* algorithm in Algorithm 2, which finds an ϵ-optimal policy in polynomial time. Following earlier algorithmic art in the same vein e.g., GOLF [JLM21], the core optimization step of OPERA is optimization-based exploration under the constraint of an identified confidence region; we additionally introduce an estimation policy π_{est} sharing the similar spirit as in [DKL+21]. Due to space limit, we focus on the Q-type analysis here and defer the V-type results to Section 3.5.3.[11]

Pertinent to the constrained optimization subproblem in Eq. (22) of our Algorithm 2, we adopt the confidence region based on a general DEF, extending the Bellman-error-based confidence region used in [JLM21]. As a result of such an extension, our algorithm can deal with more complex models such as low Witness rank and KNR. Similar to existing literature on RL theory with general function approximation, our algorithm is in general computationally inefficient. Yet OPERA is oracle efficient given the oracle for solving the optimization problem in Line 3 of Algorithm 2. We will discuss its computational issues in detail in Section B.1.1, Section B.1.2 and Section B.1.3.

3.4.2 Regret Bounds

We are ready to present the main theoretical results of our ABC class with low FE dimension:

Theorem 2 (Regret Bound of OPERA)**.** *For an MDP M, hypothesis classes $\mathcal{F}, \mathcal{G}$, a Decomposable Estimation Function ℓ satisfying Assumption 5, an admissible Bellman characterization G, suppose $(M, \mathcal{F}, \mathcal{G}, \ell, G)$ is an ABC class with low functional eluder dimension. For any fixed $\delta \in (0,1)$, we choose $\beta = \mathcal{O}\left(\log(THN_{\mathcal{L}}(1/T)/\delta)\right)$ in Algorithm 2. Then for the on-policy case when $\pi_{op} = \pi_{est} = \pi^t$, with probability at least $1-\delta$, the regret is upper bounded by*

$$\mathit{Regret}(T) = \mathcal{O}\left(\frac{H}{\kappa}\sqrt{T \cdot \dim_{FE}\left(\mathcal{F}, G, \sqrt{1/T}\right) \cdot \beta}\right).$$

[11] Here and throughout our paper we considers $\pi_{\text{est}} = \pi^t$ for Q-type models. For V-type models, we instead consider $\pi_{\text{est}} = U(\mathcal{A})$ to be the uniform distribution over the action space. Such a representation of estimation policy allows us to unify the Q-type and V-type models in a single analysis.

Algorithm 2 OPtimization-based ExploRation with Approximation (OPERA)

```
Initialize: $\mathcal{D}_h = \varnothing$ for $h = 1, \ldots, H$
for iteration $t = 1, 2, \ldots, T$ do
    Set $\pi^t := \pi_{f^t}$ where $f^t$ is taken as $\mathrm{argmax}_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$ subject to
        $$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left\| \ell_{h,f^i}(o_h^i, f_{h+1}, f_h, v) \right\|_2^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| \ell_{h,f^i}(o_h^i, f_{h+1}, g_h, v) \right\|_2^2 \right\} \le \beta \quad \text{for all } h \in [H] \qquad (22)$$
    For any $h \in [H]$, collect tuple $(r_h, s_h, a_h, s_{h+1})$ by rolling in $s_h \sim \pi^t$ and executing $a_h \sim \pi_{\text{est}}$
    Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(r_h, s_h, a_h, s_{h+1})\}$
end for
Output: $\pi_{\text{out}}$ uniformly sampled from $\{\pi^t\}_{t=1}^T$
```

We defer the proof of Theorem 2, together with a corollary for sample complexity analysis, to Section 3.5. We observe that the regret bound of the OPERA algorithm is dependent on both the functional eluder dimension $\dim_{\text{FE}}$ and the covering number of the induced DEF class $N_{\mathcal{L}}(\sqrt{1/T})$. In the special case when DEF is chosen as the Bellman error, the relation $\dim_{\text{FE}}(\mathcal{F}, G, \sqrt{1/T}) = \dim_{\text{BE}}(\mathcal{F}, \Pi, \sqrt{1/T})$ holds with Π being the function class induced by $\{\pi_f, f \in \mathcal{F}\}$, and our Theorem 2 reduces to the regret bound in [JLM21] (Theorem 15).

We will provide a detailed comparison between our framework and other related frameworks in Section 3.1 when applied to different MDP models.

3.4.3 Implication for Specific MDP Instances

Here we focus on comparing our results applied to model-based RLs that are hardly analyzable in the model-free framework in Section 3.3.2. We demonstrate how OPERA can find near-optimal policies and achieve a state-of-the-art sample complexity under our new framework.

We highlight that Algorithm 2 not only provides a simple optimization-based scheme, recovers previous near-optimal algorithms in literature (Algorithms 4 and 6 in Section B.1) when applied to specific MDP instances, but also reduces to a novel Algorithm 5 for low witness rank MDPs with improved sample complexity.

Low Witness Rank. We first provide a sample complexity result for the low Witness rank model structure. Let $|\mathcal{M}|$ and $|\mathcal{V}|$ be the cardinality of the model class[12] $\mathcal{M}$ and discriminator class $\mathcal{V}$, respectively, and W_κ be the witness rank (Definition 13) of the model. We have the following sample complexity result for low Witness rank models.

Corollary 2 (Finite Witness Rank)**.** *For an MDP model M with finite witness rank structure in Definition 13 and any fixed $\delta \in (0, 1)$, we choose $\beta = \mathcal{O}\left(\log(TH|\mathcal{M}||\mathcal{V}|/\delta)\right)$ in Algorithm 2. With probability at least $1-\delta$, Algorithm 2 outputs an ϵ-optimal policy π_{out} within $T = \tilde{\mathcal{O}}\left(H^2|\mathcal{A}|W_\kappa\beta/(\kappa^2\epsilon^2)\right)$ trajectories.*

[12] Hypothesis class reduces to model class [SJK$^+$19] when restricted to model-based setting.

Proof of Corollary 2 is delayed to Section B.1.4.[13] Compared with previous best-known sample complexity result of $\widetilde{O}\left(H^3 W_\kappa^2 |\mathcal{A}| \log(T|\mathcal{M}||\mathcal{V}|/\delta)/(\kappa^2\epsilon^2)\right)$ due to [SJK+19], our sample complexity is superior by a factor of dH up to a polylogarithmic prefactor in model parameters.

Kernel Nonlinear Regulator. Now we turn to the implication of Theorem 2 for learning *KNR models*. We have the following regret bound result for KNR.

Corollary 3 (KNR)**.** *For the KNR model in Eq.* (20) *and any fixed* $\delta \in (0,1)$*, we choose* $\beta = \mathcal{O}\left(\sigma^2 d_\phi d_s \log^2(TH/\delta)\right)$ *in Algorithm 2. With probability at least* $1-\delta$*, the regret is upper bounded by* $\tilde{\mathcal{O}}\left(H^2\sqrt{d_\phi T\beta}/\sigma\right)$*.*

We remark that neither the low BE dimension nor the Bellman Representability classes admit the KNR model with a sharp regret bound. Among earlier attempts, [DKL+21, Section 6] proposed to use a generalized version of Bilinear Classes to capture models including KNR, Generalized Linear Bellman Complete, and finite Witness rank. Nevertheless, their characterization requires imposing monotone transformations on the statistic and yields a suboptimal $\mathcal{O}(T^{3/4})$ regret bound. Our ABC class with low FE dimension is free of monotone operators, albeit that the coupling function for the KNR model is not of a bilinear form.

3.5 Proof of Main Results

In this section, we provide proofs of our main result Theorem 2 and a sample complexity corollary of the OPERA algorithm. Originated from proof techniques widely used in confidence bound based RL algorithms [RVR13] our proof steps generalizes that of the GOLF algorithm [JLM21] but admits general DEF and ABCs. Finally, Section 3.5.3 explains the V-type setting and the corresponding results.

3.5.1 Proof of Theorem 2

We prove our main result as follows:

Proof of Theorem 2. We recall that the objective of an RL problem is to find an ϵ-optimal policy satisfying $V_1^*(s_1) - V_1^{\pi^t}(s_1) \le \epsilon$. Moreover, the regret of an RL problem is defined as $\sum_{t=1}^T V_1^*(s_1) - V_1^{\pi^t}(s_1)$, where π^t is the output policy of an algorithm at time t.

Step 1: Feasibility of f^*. First of all, we show that the optimal hypothesis f^* lies within the confidence region defined by Eq. (22) with high probability:

Lemma 6 (Feasibility of f^*)**.** *In Algorithm 2, given* $\rho > 0$ *and* $\delta > 0$ *we choose* $\beta = c(\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta) + T\rho)$ *for some large enough constant* c*. Then with probability at least* $1-\delta$*,* f^* *satisfies for any* $t \in [T]$*:*

$$\max_{v\in\mathcal{V}}\left\{\sum_{i=1}^{t-1}\left\|\ell_{h,f_h^i}(o_h^i, f_{h+1}^*, f_h^*, v)\right\|_2^2 - \inf_{g_h\in\mathcal{G}_h}\sum_{i=1}^{t-1}\left\|\ell_{h,f_h^i}(o_h^i, f_{h+1}^*, g_h, v)\right\|_2^2\right\} \le \mathcal{O}(\beta)$$

[13] The definition of witness rank adopts a V-type representation and hence we can only derive the sample complexity of our algorithm. For detailed discussion on the V-type cases, we refer readers to Section 3.5.3.

Lemma 6 shows that at each round of updates the optimal hypothesis f^* stays in the confidence region depicted by Eq. (22) with radius $\mathcal{O}(\beta)$. We delay the proof of Lemma 6 to Section B.2.2. Lemma 6 together with the optimization procedure Line 3 of Algorithm 2 implies an upper bound of $V_1^*(s_1) - V_1^{\pi^t}(s_1)$ with probability at least $1-\delta$ as follows:

$$V_1^*(s_1) - V_1^{\pi^t}(s_1) \le V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1) \tag{23}$$

Step 2: Policy Loss Decomposition. The second step is to upper bound the regret by the summation of Bellman errors. We apply the policy loss decomposition lemma in [JKA+17].

Lemma 7 (Lemma 1 in [JKA+17]). *$\forall f \in \mathcal{H}$,*

$$V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1) = \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim\pi^t}\left[Q_{h,f^t}(s_h,a_h) - r_h - V_{h+1,f^t}(s_{h+1})\right]$$

Combining Lemma 7 with Eq. (23) we have the following:

$$V_1^*(s_1) - V_1^{\pi^t}(s_1) \le V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1) = \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim\pi^t}\left[Q_{h,f^t}(s_h,a_h) - r_h - V_{h+1,f^t}(s_{h+1})\right] \tag{24}$$

Step 3: Small ABC Value in the Confidence Region. The third step is devoted to controlling the cumulative square of Admissible Bellman Characterization function. Recalling that the ABC function is upper bounded by the average DEF, where each feasible DEF stays in the confidence region that satisfies Eq. (22), we arrive at the following Lemma 8:

Lemma 8. *In Algorithm 2, given $\rho > 0$ and $\delta > 0$ we choose $\beta = c(\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta) + T\rho)$ for some large enough constant c. Then with probability at least $1-\delta$, for all $(t,h) \in [T]\times[H]$, we have*

$$\sum_{i=1}^{t-1}\left(G_{h,f^*}(f^t,f^i)\right)^2 \le \mathcal{O}(\beta) \tag{25}$$

The proof of Lemma 8 makes use of Freedman's inequality (the precise version as in [AHK+14]) and we delay the proof to Section B.2.1.

Step 4: Bounding the Cumulative Bellman Error by Functional Eluder Dimension. In the fourth step, we aim to traslate the upper bound of the cumulative squared ABC at (f^t, f^i) in Eq. (25) to an upper bound of the cumulative ABC at (f^t, f^t). The following Lemma 9 is adapted from Lemma 41 in [JLM21] and Lemma 2 in [RVR13]. Lemma 9 controls the sum of ABC functions by properties of the functional eluder dimension.

Lemma 9. *For a hypothesis class $\mathcal{F}$ and a given coupling function $G(\cdot,\cdot) : \mathcal{F}\times\mathcal{F} \to \mathcal{R}$ with bounded image space $|G(\cdot,\cdot)| \le C$. For any pair of sequences $\{f_t\}_{t\in[T]}, \{g_t\}_{t\in[T]} \subseteq \mathcal{F}$ satisfying for all $t \in [T]$, $\sum_{i=1}^{t-1}(G(f_t,g_i))^2 \le \beta$, the following inequality holds for all $t\in[T]$ and $\omega > 0$:*

$$\sum_{i=1}^{t}|G(f_i,g_i)| \le \mathcal{O}\Big(\sqrt{\dim_{FE}(\mathcal{F},G,\omega)\beta t} + C\cdot\min\{t, \dim_{FE}(\mathcal{F},G,\omega)\} + t\omega\Big)$$

The proof of Lemma 9 is in Section B.2.3.

Step 5: Combining Everything. In the final step, we combine the regret bound decomposition argument, the cumulative ABC bound, and the Bellman dominance property together to derive our final regret guarantee.

For any $h \in [H]$, we take $G(\cdot,\cdot) = G_{h,f^*}(\cdot,\cdot)$, $g_i = f^i, f_t = f^t$ and $\omega = \sqrt{\frac{1}{T}}$ in Lemma 9. By Eq. (25) in Lemma 8, we have for any $h \in [H]$ and $t \in [T]$,

$$\begin{aligned}\sum_{i=1}^{t} |G_{h,f^*}(f^i, f^i))| &\leq \mathcal{O}\left(\sqrt{\dim_{\mathrm{FE}}(\mathcal{F}, G_{h,f^*}, \sqrt{1/T})\beta t} + C \cdot \min\{t, \dim_{\mathrm{FE}}(\mathcal{F}, G_{h,f^*}, \sqrt{1/T})\} + \sqrt{t}\right) \\ &\leq \mathcal{O}\Big(\sqrt{\dim_{\mathrm{FE}}(\mathcal{F}, G_{h,f^*}, \sqrt{1/T})\beta t}\Big)\end{aligned}$$

We recall our choice of $\beta = c\left(\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right) + T\rho\right)$. Taking $\rho = \frac{1}{T}$, we have

$$\begin{aligned}\sum_{i=1}^{t} |G_{h,f^*}(f^i, f^i))| &\leq \mathcal{O}\left(\sqrt{\dim_{\mathrm{FE}}\left(\mathcal{F}, G_{h,f^*}, \sqrt{1/T}\right)\log\left(TH\mathcal{N}_{\mathcal{L}}(1/T)/\delta\right) \cdot t}\right) \\ &\leq \mathcal{O}\left(\sqrt{\dim_{\mathrm{FE}}\left(\mathcal{F}, G, \sqrt{1/T}\right)\log\left(TH\mathcal{N}_{\mathcal{L}}(1/T)/\delta\right) \cdot t}\right)\end{aligned}$$

Combining this with property (ii) in Definition 11 and decomposition (24), we conclude our main result that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) &\leq \frac{1}{\kappa}\sum_{t=1}^{T}\sum_{h=1}^{H} |G_{h,f^*}(f^t, f^t)| \\ &\leq \mathcal{O}\left(\frac{H}{\kappa}\sqrt{T \cdot \dim_{\mathrm{FE}}(\mathcal{F}, G, \sqrt{1/T})\log\left(TH\mathcal{N}_{\mathcal{L}}(1/T)/\delta\right)}\right)\end{aligned}$$

This completes the whole proof of Theorem 2. □

3.5.2 Sample Complexity of OPERA

Corollary 4 (Sample Complexity of OPERA)**.** *For an MDP M with hypothesis classes $\mathcal{F}$, $\mathcal{G}$ that satisfies Assumption 4 and a Decomosable Estimation Function ℓ satisfying Assumption 5. If there exists an Admissible Bellman Characterzation G with low functional eluder dimension. For any $\epsilon \in (0,1]$, we choose $\beta = c\left(\log(THN_{\mathcal{L}}\left(\frac{\kappa^2\epsilon^2}{\dim_{FE}(\mathcal{F},G,\frac{\kappa\epsilon}{H})H^2}\right)/\delta) + T\frac{\kappa^2\epsilon^2}{\dim_{FE}(\mathcal{F},G,\frac{\kappa\epsilon}{H})H^2}\right)$ for some large enough constant c. For the on-policy case when $\pi_{op} = \pi_{est} = \pi^t$, with probability at least $1-\delta$ Algorithm 2 outputs a ϵ-optimal policy π_{out} within T trajectories where*

$$T = \frac{\dim_{FE}(\mathcal{F}, G, \frac{\kappa\epsilon}{H})\log\left(TH\mathcal{N}_{\mathcal{L}}\left(\frac{\kappa^2\epsilon^2}{\dim_{FE}(\mathcal{F},G,\frac{\kappa\epsilon}{H})H^2}\right)/\delta\right)H^2}{\kappa^2\epsilon^2}$$

Proof of Corollary 4. By the policy loss decomposition (24), (25) in Lemma 8 and Lemma 9, we have that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) &\leq \frac{1}{\kappa T}\sum_{t=1}^{T}\sum_{h=1}^{H} \left|G_{h,f^*}(f^t, f^t)\right| \\ &\leq \mathcal{O}\Big(\frac{H}{\kappa}\sqrt{\dim_{\mathrm{FE}}(\mathcal{F}, G, \omega)\left(\frac{\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)}{T} + \rho\right)} + \frac{H\omega}{\kappa}\Big) \qquad (26)\end{aligned}$$

Taking $\omega = \frac{\kappa\epsilon}{H}$ and $\rho = \frac{\kappa^2\epsilon^2}{\dim_{\mathrm{FE}}(\mathcal{F},G,\frac{\kappa\epsilon}{H})H^2}$, the above Eq. (26) becomes

$$\frac{1}{T}\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) \le \mathcal{O}\Big(\frac{H}{\kappa}\sqrt{\frac{\dim_{\mathrm{FE}}(\mathcal{F},G,\frac{\kappa\epsilon}{H})\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)}{T}} + \epsilon\Big)$$

Taking

$$T = \frac{\dim_{\mathrm{FE}}(\mathcal{F},G,\frac{\kappa\epsilon}{H})\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)H^2}{\kappa^2\epsilon^2}$$

yields the desired result. □

3.5.3 Q-type and V-type Sample Complexity Analysis

In Definition 11, we note that there are two ways to calculate the ABC of an MDP model depending on the different choices of the operating policy π_{op}. Specifically, if $\pi_{\mathrm{op}} = \pi_g$, we call it the Q-type ABC. Otherwise, if $\pi_{\mathrm{op}} = \pi_f$, we call it the V-type ABC. For example, when taking

$$G_{h,f^*}(f,g) = \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_g}\left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1})\right]$$

the FE dimension of $G_{h,f^*}(f,g)$ recovers the Q-type BE dimension (Definition 8 in [JLM21]. When taking

$$G_{h,f^*}(f,g) = \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_f}\left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1})\right]$$

the FE dimension of $G_{h,f^*}(f,g)$ recovers the V-type BE dimension (Definition 20 in [JLM21]. The algorithm for solving Q-type or V-type models slightly differs in the executing policy π_{est}. We use $\pi_{\mathrm{est}} = \pi^t$ for Q-type models in Algorithm 2, while $\pi_{\mathrm{est}} = U(\mathcal{A})$ is the uniform distribution on action set for V-type models.

The Q-type characterization and the V-type characterization have respective applicable zones. For example, the reactive POMDP model belongs to ABC with low FE dimension with respect to V-type ABC while inducing large FE dimension with respect to Q-type ABC. On the contrary, the low inherent bellman error problem in [ZLKB20a] is more suitable for using a Q-type characterization rather than a V-type characterization. For general RL models, we often prefer Q-type ABC because the sample complexity of V-type algorithms scales with the dimension of the action space $|\mathcal{A}|$. Due to the uniform executing policy, we will only be able to derive regret bound for Q-type characterizations, as is explained in [JLM21].

In Section 3.4 and Section 3.5, we have illustrated regret bound and sample complexity results for the Q-type cases where we let $\pi_{\mathrm{op}} = \pi_{\mathrm{est}} = \pi^t$ through Algorithm 2. In the following Corollary 5, we prove sample complexity result for V-type ABC models.

Corollary 5. *For an MDP M with hypothesis classes $\mathcal{F}$, $\mathcal{G}$ that satisfies Assumption 4 and a Decomposable Estimation Function ℓ satisfying Assumption 5. If there exists an Admissible Bellman Characterization G with low functional eluder dimension. For any $\epsilon \in (0,1]$, if we choose $\beta = \mathcal{O}\left(\log(THN_{\mathcal{L}}(\rho)/\delta) + T\rho\right)$. For V-type models when $\pi_{op} = \pi_{est} = \pi^t$, with probability at least $1-\delta$ Algorithm 2 outputs a ϵ-optimal policy π_{out} within $T = \frac{|\mathcal{A}|\dim_{FE}(\mathcal{F},G,\kappa\epsilon/H)\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta)H^2}{\kappa^2\epsilon^2}$ trajectories where $\rho = \frac{\kappa^2\epsilon^2}{\dim_{FE}(\mathcal{F},G,\frac{\kappa\epsilon}{H})H^2}$.*

Proof of Corollary 5. The proof of Corollary 5 basically follows the proof of Theorem 2 and Corollary 4. We again have feasibility of f^* and policy loss decomposition. However, due to different sampling policy, the proof of Lemma 8 differs at Eq. (57). Instead, we have

$$\begin{aligned}
&\sum_{i=1}^{t-1} \max_{v\in\mathcal{V}} \mathbb{E}_{s_h\sim\pi^i, a_h\sim\pi^t}\mathbb{E}_{s_{h+1}}\left[X_i(h, f^t, v) \mid s_h, a_h\right] \\
&= \sum_{i=1}^{t-1} \max_{v\in\mathcal{V}} \mathbb{E}_{s_h\sim\pi^i, a_h\sim U(\mathcal{A})}\frac{\mathbb{1}(a_h^i = \pi_f(s_h^i))}{1/|\mathcal{A}|}\mathbb{E}_{s_{h+1}}\left[X_i(h, f^t, v) \mid s_h, a_h\right] \\
&= \sum_{i=1}^{t-1} \max_{v\in\mathcal{V}} \mathbb{E}_{s_h\sim\pi^i, a_h\sim U(\mathcal{A})}\frac{\mathbb{1}(a_h^i = \pi_f(s_h^i))}{1/|\mathcal{A}|}\left\|\mathbb{E}_{s_{h+1}}\left[\ell_{h,f^i}(o_h, f_{h+1}^t, f_h^t, v) \mid s_h, a_h\right]\right\|_2^2 \\
&\leq \mathcal{O}(|\mathcal{A}|\left(\beta + Rt\rho + R^2\iota\right)) \qquad (27)
\end{aligned}$$

Thus, Eq. (25) in Lemma 8 becomes

$$\sum_{i=1}^{t-1}\left(G_{h,f^*}(f^t, f^i)\right)^2 \leq \mathcal{O}(|\mathcal{A}|\beta)$$

The rest of the proof follow the proof of Corollary 4 with an additional $|\mathcal{A}|$ factor. By the policy loss decomposition (24) and Lemma 9, we have that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) &\leq \frac{1}{\kappa T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left|G_{h,f^*}(f^t, f^t)\right| \\
&\leq \mathcal{O}\Big(\frac{H}{\kappa}\sqrt{|\mathcal{A}|\dim_{\mathrm{FE}}(\mathcal{F}, G, \omega)\left(\frac{\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)}{T} + \rho\right)} + \frac{H\omega}{\kappa}\Big) \qquad (28)
\end{aligned}$$

Taking $\omega = \frac{\kappa\epsilon}{H}$ and $\rho = \frac{\kappa^2\epsilon^2}{\dim_{\mathrm{FE}}(\mathcal{F}, G, \frac{\kappa\epsilon}{H})H^2}$, the above Eq. (28) becomes

$$\frac{1}{T}\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) \leq \mathcal{O}\Big(\frac{H}{\kappa}\sqrt{\frac{|\mathcal{A}|\dim_{\mathrm{FE}}(\mathcal{F}, G, \frac{\kappa\epsilon}{H})\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)}{T}} + \epsilon\Big)$$

Taking

$$T = \frac{|\mathcal{A}|\dim_{\mathrm{FE}}(\mathcal{F}, G, \frac{\kappa\epsilon}{H})\log\left(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta\right)H^2}{\kappa^2\epsilon^2}$$

yields the desired result. □

4 Conclusion

In this paper, we introduced the Bellman Eluder (BE) dimension, a novel complexity measure for reinforcement learning (RL) with general function approximation. The BE dimension unifies and extends existing complexity notions, such as Bellman rank and Eluder dimension, identifying a broad class of RL problems that are sample-efficiently solvable. This framework encompasses many well-known RL problems, including tabular MDPs, linear MDPs, and reactive POMDPs, offering a more comprehensive understanding of sample efficiency under minimal structural assumptions.

We proposed the GOLF algorithm, which efficiently learns near-optimal policies in RL tasks characterized by low BE dimension. GOLF achieves competitive sample complexity and regret bounds, independent of state-action space size, matching or improving upon existing results for several subclasses of RL problems. Additionally, we introduced the concept of Admissible Bellman Characterization (ABC), which bridges the gap between model-free and model-based RL paradigms by offering a unified framework for tackling function approximation in large state and action spaces.

These contributions advance the theoretical foundations of RL and offer practical tools for addressing the challenges of efficient learning in complex environments. While our framework addresses a wide range of RL problems, certain models, such as deterministic linear Q^* models and Q^* state-action aggregation, remain outside its scope, leaving room for further exploration.

Future research should focus on extending the BE dimension framework to even more general function classes and investigating its application in other RL settings, such as multi-agent RL and online learning. Further empirical studies will also be essential to assess the practical performance of GOLF and explore potential enhancements across diverse real-world domains.

References

[AHK$^+$14] 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. PMLR, 2014.

[AHKS20] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

[AJ17] 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.

[AJO08] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[AJS$^+$20] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[AKKS20] 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*, 33:20095–20107, 2020.

[AM07] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[ASM08] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[BM02] 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.

[BT02] 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.

[CJ19] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

[CYJW20a] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[CYJW20b] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[DB15] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 2137–2143, 2008.

[DKL+21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[DLWZ19] Simon S. Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8058–8068, 2019.

[DMKV21] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

[DMM+19] 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.

[DPWZ20] 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.

[FGKM18] 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.

[FKQR21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[FRSLX20] 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.

[JAZBJ18] 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.

[JKA+17] 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.

[JLM21] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[JYSW20] Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

[JYWJ20] 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. PMLR, 2020.

[KAL16] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

[KBP13] 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.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[KKL$^+$20] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[LMR$^+$16] 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.

[MJR22] Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022.

[MJTS20] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[MKS$^+$13] 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.

[MS08] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

[Mül97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[NPB20] Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, volume 27, pages 1466–1474, 2014.

[Pol12] David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[Put14] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23, 2010.

[RVR13] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, pages 2256–2264, 2013.

[RVR14] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction.* MIT Press, 2018.

[SHM$^+$16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[SJK$^+$19] 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. PMLR, 2019.

[SJR12] Satinder Singh, Michael James, and Matthew Rudary. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167*, 2012.

[SKKS09] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[SM05] 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.

[SMSM99] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[Sze10] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

[Vap99] Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 1999.

[Vap13] Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 2013.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, 2019.

[WAS20] 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.

[Wat89] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. *PhD Thesis, Cambridge University*, 1989.

[WAT15] Grady Williams, Andrew Aldrich, and Evangelos Theodorou. Model predictive path integral control using covariance variable importance sampling. *arXiv preprint arXiv:1509.01149*, 2015.

[WCSB19] Nolan Wagener, Ching-An Cheng, Jacob Sacks, and Byron Boots. An online learning approach to model predictive control. *arXiv preprint arXiv:1902.08967*, 2019.

[WSY20a] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

[WSY20b] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[WWDK21] Yining Wang, Ruosong Wang, Simon S. Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021.

[XJ20] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

[YJW$^+$20a] 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.

[YJW$^+$20b] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

[YLCT20] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

[YW19] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[YW20] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[ZB19] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

[ZGS21] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[ZHG21] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

[ZLKB20a] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[ZLKB20b] 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*, 33:11756–11766, 2020.

[ZZJ20] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.