The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

Theoretical Reinforcement Learning - Multi Agent Function Approximation and Effective Classes

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
by

Dolev Danino
Email: dolevd404@gmail.com

This work was carried out under the supervision of
Prof. Yishay Mansour

October 2024

The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University

Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Yishay Mansour, for his continuous support, guidance, and encouragement throughout my research and thesis writing.

I would also like to thank the faculty members and staff of the Raymond and Beverly Sackler Faculty of Exact Sciences at Tel Aviv University for their support and the excellent facilities provided.

Special thanks to my friends and colleagues in Check-Point building in the 4^{th} floor, for their insightful discussions, support, and friendship during this journey.

Finally, I am deeply grateful to my family and my wife, Sapir, for their endless support and encouragement throughout my journey and studies.

Dolev Danino

Abstract

Function approximation has become an increasingly important topic in reinforcement learning (RL), driven by advances in machine learning and the need to handle large, non-linear function classes. Despite recent progress, no existing algorithms fully address this challenge efficiently. This thesis explores way to improve RL's effectiveness in common scenarios while maintaining theoretical guarantees.

This work contributes with introducing the concept of the *effective horizon*, extending the Eluder dimension to define the *expected Eluder dimension*, and leveraging multi-agent systems and generative models.

Our results improve scalability, reduce sample complexity, and demonstrate regret bounds with potentially lower dimensional dependence. These contributions collectively advance RL, making it more efficient and applicable to real-world scenarios.

Contents

1 Introduction and Preliminary

1.1 Introduction

Reinforcement Learning (RL) has emerged as a powerful framework for modeling and solving complex decision-making problems. At its core, RL involves an agent interacting with an environment to learn optimal behaviors through trial and error, guided by the maximization of cumulative rewards. This iterative process of learning from feedback has seen remarkable successes in diverse applications ranging from robotics (Kober et al., 2013) and game playing (Silver et al., 2016) to personalized recommendations (Zheng et al., 2018) and autonomous driving (Kiran et al., 2021).

One of the cornerstone concepts in RL theory is the Markov Decision Process (MDP), which provides a formalism for modeling sequential decision-making problems. MDPs are characterized by a state space, action space, transition dynamics, and a reward function, serving as the basis for many RL algorithms (Agarwal et al., 2021; Mannor et al., 2022; Puterman, 2014). The goal of reinforcement learning is to find an optimal policy, which is a strategy that defines the action an agent should take in each state of a MDP to maximize the expected cumulative reward over time.

Probably Approximately Correct (PAC) learning (Valiant, 1984) in supervised learning provides a framework to evaluate the performance of learning algorithms by ensuring that, with high probability, the learned hypothesis is approximately correct within a specified error margin. This approach offers theoretical guarantees on the number of samples needed to achieve a certain level of accuracy, making it a foundational concept in the study of machine learning efficiency.

The definition of PAC in RL, following Kearns and Singh (1999), is similar to supervised learning. An algorithm is designed to learn a policy that is approximately optimal with high probability while minimizing the number of samples. The primary goal is to ensure that after a polynomial number of samples. The number of samples is relative to the relevant parameters of the MDP such as the size of the state and action spaces, the horizon, and the approximation parameters. This framework is crucial for understanding the sample efficiency of RL algorithms, as it provides bounds on the number of interactions with the environment required to achieve near-optimal performance.

Function approximation for large MDPs (Sutton and Barto, 2018; Mannor et al., 2022; Agarwal et al., 2021), mainly large state spaces, in reinforcement learning involves estimating the value function, on its shapes, or the policy function using techniques such as linear models, neural networks, or kernel methods. This approach enables RL algorithms to generalize from limited data to vast or continuous state-action spaces, making it possible to handle large-scale or complex environments efficiently.

Sample-efficient RL algorithms (Chen et al., 2022; Jiang et al., 2017; Jin et al., 2021a; Agarwal et al., 2022) aim to maximize learning performance while minimizing the number of interactions required with the environment. These algorithms employ techniques such as efficient exploration strategies, function approximation, and leveraging prior knowledge to achieve near-optimal policies with fewer samples, making them suitable for real-world applications where data collection is costly or time-consuming.

In multi-agent reinforcement learning (MARL) (see, Busoniu et al. (2008); Zhang et al. (2021); Hernandez-Leal et al. (2019); Foerster et al. (2018)), sample efficiency becomes even more critical as multiple agents interact within a shared environment, each learning policies to optimize their individual or collective rewards. Efficient MARL algorithms must account for the added complexity of agent interactions and coordination, employing advanced strategies to ensure that all agents learn effectively with minimal samples.

This thesis makes several significant contributions to the field of reinforcement learning (RL), particularly in the context of function approximation. By enhancing sample efficiency and practical implementation through multi-agent systems and generative models, we advance the foundational algorithms in meaningful

ways. The key contributions are as follows:

1. **Improvement of Sample Efficiency**:
    - We introduce the concept of the *effective horizon*, a novel definition that refines the understanding of Iterations Complexity in RL problems. By utilizing this concept, we develop more sample-efficient algorithms that require fewer interactions with the environment to achieve near-optimal performance.
    - We propose the *expected Eluder dimension*, extending the Eluder dimension framework to account common distributions. This new dimension provides a more nuanced measure of complexity for function approximation in RL, leading to algorithms that are better at balancing exploration and exploitation.
2. **Multi-Agent Systems for Practical Implementability**:
    - We leverage multi-agent systems to enhance the practical implementability of RL algorithms. By coordinating multiple agents within a shared environment, we develop algorithms that are not only theoretically sound but also more realistic and applicable to complex real-world scenarios.
    - Our multi-agent framework improves the scalability and robustness of RL algorithms, making them suitable for tasks that involve cooperation or competition among multiple entities.
3. **Generative Models for Enhanced Sample Efficiency**:
    - We employ the concept of *generative models* to further improve sample efficiency. Generative models allow for the simulation of environment dynamics, enabling agents to learn optimal policies with fewer real-world interactions.
    - By integrating generative models into our RL algorithms, we demonstrate significant reductions in sample complexity, making the learning process more efficient and feasible for practical applications.

Overall, the contributions of this thesis advance the fundamental algorithms in function approximation for reinforcement learning by introducing new theoretical concepts, leveraging multi-agent systems for practical implementation, and utilizing generative models to achieve better sample efficiency. These advancements bring us closer to developing RL algorithms that are both highly efficient and applicable to a wide range of real-life problems.

1.2 Related Work

The field of reinforcement learning (RL) has a rich theoretical foundation, particularly in the study of Markov Decision Processes (MDPs) with small state spaces. Early works, such as those by Kearns and Singh (2002), Brafman and Tennenholtz (2002), and Strehl et al. (2006), established PAC-MDP bounds that are polynomial in the number of states S and actions A, focusing on sophisticated exploration techniques to find near-optimal policies in a sample-efficient manner. However, extending these techniques to large state spaces has proven challenging, as evidenced by the works of Kakade et al. (2003), Jong and Stone (2007), and Pazis and Parr (2016), which often struggle to address the complexities of practical scenarios involving high-dimensional sensory inputs.

As RL research has evolved, there has been a growing emphasis on RL under nonlinear function approximation to model complex function spaces, such as neural networks. This shift has led to the development of rank-based measures that capture the hardness of RL in these settings, with frameworks like Bellman rank (Jiang et al., 2017), low Eluder dimension (Wang et al., 2020), Bellman-Eluder dimension (Jin et al., 2021a), and Bilinear classes (Du et al., 2021). These measures have paved the way for designing algorithms that provide PAC guarantees, such as the OLIVE algorithm, which leverages Bellman rank, and the GOLF algorithm, which incorporates the Bellman operator into complexity measures.

Eluder dimension, introduced by Russo and Van Roy (2013), and its various extensions have further refined the understanding of informational independence among decisions, influencing the balance between exploration and exploitation in RL. Notable contributions in this area include the Generalized Eluder dimension (Agarwal et al., 2022), which extends these concepts to broader function classes. When aiming to learn PAC policies, it is crucial that the algorithm's complexity is polynomial relative to the Bellman rank or the Eluder dimension. Theoretical analyses of algorithms such as OPERA (Chen et al., 2022) have demonstrated significant improvements in sample complexity, making RL more feasible for real-world applications where data collection is costly or time-consuming.

In the context of multi-agent reinforcement learning (MARL), significant strides have been made, particularly in Markov Games (MGs), where research focuses on convergence to an equilibrium (Littman, 1994; Jin et al., 2021b; Qiao and Wang, 2024). Decentralized MARL, which emphasizes independent learning and convergence without central coordination, has also been extensively studied (Kretchmar, 2002; Zhang et al., 2018; Zhuo et al., 2020; Zhang et al., 2021; Dubey and Pentland, 2021a; Mao et al., 2022). For instance, works like Qi et al. (2021) have developed unified versions of TD and Q-learning, while Fan et al. (2022) explored policy gradients with fault tolerance.

This thesis focuses on cooperative MARL, particularly in function approximation settings. Cooperative MARL, where all agents collaborate to achieve a shared goal, represents a significant portion of MARL research. In homogeneous settings, where the underlying MDP is the same for every agent, we explore algorithms with low regret guarantees, extending beyond existing works that address non-stationary environments and heterogeneity (Lowe et al., 2017; Yu et al., 2021; Kuba et al., 2022; Liu et al., 2022; Jin et al., 2022; Dubey and Pentland, 2021b; Min et al., 2023).

Several key advancements have shaped our understanding of reinforcement learning and multi-agent systems. Lancewicki et al. (2022) made notable contributions by addressing the challenges of fresh/non-observations in tabular settings, which is crucial for cooperative multi-agent systems. In the realm of rare policy switching (RPS), Zhao et al. (2023) extended the framework of the Generalized Eluder Dimension (Agarwal et al., 2022), providing deeper insights into RPS complexities and its practical applications in RL.

Batch learning has also seen significant advancements, with Xiong et al. (2023) exploring the $\mathcal{L}_2$-EC (Euclidean norm - Eluder Condition) and optimizing batch processing for better learning outcomes. Moreover, asynchronous communication—a critical aspect of distributed systems—has been effectively addressed by Min et al. (2023), particularly in the context of linear function approximation, demonstrating how asynchronous methods can maintain performance despite communication delays.

2 Preliminaries

In this section, we introduce the foundational concepts and formal definitions that are essential for understanding the reinforcement learning (RL) framework discussed in this thesis. We begin by defining the core model used in RL, known as the Markov Decision Process (MDP). This model provides the mathematical structure required to describe the environment in which an RL agent operates.

2.1 Reinforcement Learning Model

To formalize the interaction between an agent and its environment, we use the concept of a Markov Decision Process (MDP). The MDP framework encapsulates the dynamics of the environment, the actions available to the agent, and the rewards the agent can receive.

Definition 1 (Markov Decision Process (MDP))**.** *A Markov Decision Process (MDP) is a mathematical framework used to describe an environment in reinforcement learning. An MDP is defined by a tuple* $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$*, where:*

- $\mathcal{S}$ *is a finite set of states.*
- $\mathcal{A}$ *is a finite set of actions.*
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ *is the state transition probability function, where* $\mathcal{P}(s'|s, a)$ *denotes the probability of transitioning to state* s' *from state* s *after taking action* a*.*
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is the reward function, where* $\mathcal{R}(s, a)$ *denotes the immediate reward received after taking action* a *in state* s*.*
- $\gamma \in [0, 1]$ *is the discount factor, which represents the difference in importance between future rewards and immediate rewards.*

The MDP framework serves as the basis for defining the optimal behavior of an RL agent. The goal of the agent is to learn a policy that maximizes the expected cumulative reward over time.

Definition 2 (Policy in Reinforcement Learning)**.** *In reinforcement learning (RL), a* ***policy*** *is a strategy used by an agent to determine its actions based on the current state of the environment. Formally, a policy* π *is a mapping from states* $\mathcal{S}$ *to actions* $\mathcal{A}$*. There are two main types of policies:*

- ***Deterministic Policy****: A deterministic policy maps each state to a specific action, i.e.,* $\pi : \mathcal{S} \to \mathcal{A}$*. For any state* $s \in \mathcal{S}$*,* $\pi(s)$ *specifies the action to be taken when the agent is in state* s*.*
- ***Stochastic Policy****: A stochastic policy defines a probability distribution over actions for each state, i.e.,* $\pi(a|s)$ *is the probability of taking action* a *when the agent is in state* s*. This allows for a randomized decision-making process, which can be useful in environments with inherent uncertainty or where exploration is necessary.*

The goal of reinforcement learning is to find an optimal policy $\pi^\star$ that maximizes the expected cumulative reward for the agent over time. The performance of a policy is typically evaluated using value functions, such as the state-value function $V^\pi(s)$ or the action-value function $Q^\pi(s, a)$, which estimate the expected return starting from state s (and action a) and following policy π thereafter. Finding an optimal policy involves solving the underlying Markov Decision Process (MDP) by optimizing these value functions through various algorithms, such as policy iteration, value iteration, or actor-critic methods.

Bellman Equations The Bellman equations are fundamental to dynamic programming and reinforcement learning. They describe the relationship between the value of a state and the values of its successor states. There are two primary forms of the Bellman equation: the Bellman Expectation Equation and the Bellman Optimality Equation. We will focus on the Bellman Optimality Equation.

Bellman Optimality Equation. The optimal value function $V^*(s)$ is defined as the maximum expected cumulative reward that can be obtained from state s. The Bellman Optimality Equation for $V^*(s)$ is:

$$V^*(s) = \max_a \left[\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^*(s') \right].$$

Similarly, the action-value function $Q^*(s,a)$ is given by:

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \max_{a'} Q^*(s',a').$$

These equations form the basis for many RL algorithms some of them we will see in the next sections.

PAC Learning in Reinforcement Learning Probably Approximately Correct (PAC) learning is a framework in reinforcement learning (RL) that provides theoretical guarantees on the performance of learning algorithms. In PAC learning, an algorithm is designed to learn a policy that is approximately optimal with high probability, within a specified number of samples. The primary goal is to ensure that, with probability at least $1-\delta$, the learned policy's performance is within ϵ of the optimal policy, after a polynomial number of samples in terms of the relevant parameters such as the size of the state and action spaces, the horizon, $1/\epsilon$, and $1/\delta$. This framework is crucial for understanding the sample efficiency of RL algorithms, as it provides bounds on the number of interactions with the environment required to achieve near-optimal performance.

Recent advancements have extended PAC learning results to more complex settings, including function approximation and multi-agent systems, demonstrating the robustness and applicability of PAC guarantees in practical RL scenarios. These theoretical insights help in designing algorithms that are both effective and efficient, ensuring that they can be applied to real-world problems where sample efficiency is paramount.

To further improve the sample efficiency and performance of RL algorithms, many approaches leverage the concept of generative models. These models allow the agent to simulate interactions with the environment, providing additional samples without the need for real-world interactions. This concept is formalized in the following definition:

Definition 3 (Generative Model)**.** *Assume we have access to a generative model or a sampler, which can provide us with samples from the environment. Specifically, given any state-action pair (s,a), the generative model can generate a next state $\hat{s} \sim P(\cdot|s,a)$ and reward $\hat{r} \sim R(s,a)$ according to the transition dynamics and reward function of the MDP. This capability allows the algorithm to perform planning and learning with potentially fewer real-world interactions, significantly enhancing sample efficiency.*

2.2 Function Approximation

Value-based RL and Function Approximation We consider value-based RL with function approximation. The agent is given a set of functions $\mathcal{F} \subseteq \mathcal{X} \times \mathcal{A} \to [0,1]$ and uses it to approximate an action-value function (or Q-value function). Without loss of generality, we assume that $f(x_{H+1}, a) \equiv 0$.

For the purpose of presentation, we assume that $\mathcal{F}$ is a finite space with $|\mathcal{F}| = N < \infty$ for beginning and later from Section 4, we relax this assumption and allow infinite function classes with finite ϵ-cover (Definition 10).

As in typical value-based RL, the goal is to identify $f \in \mathcal{F}$ which respects a particular set of Bellman equations and achieves a high value with its greedy policy $\pi_f(x) = \operatorname{argmax}_{a \in \mathcal{A}} f(x,a)$. We next set up the appropriate extensions of Bellman equations to MDPs and the optimal value $V^\star$ through a series of definitions. Unlike typical definitions in MDPs, these involve both the MDP and function approximator $\mathcal{F}$.

Definition 4 (Average Bellman error)**.** *Given any policy $\pi : \mathcal{X} \to \mathcal{A}$ and a function $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$, the* average Bellman error *of f under roll-in policy π at level h is defined as*

$$\mathcal{E}(f, \pi, h) = \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \,\middle|\, a_{1:h-1} \sim \pi,\ a_{h:h+1} \sim \pi_f\right]. \tag{1}$$

In words, the average Bellman error measures the self-consistency of a function f between its predictions at levels h and $h + 1$ when all the previous actions are taken according to some policy π.

Next, we introduce the concept of Bellman factorization, which provides a structured way to analyze the complexity of MDPs when using function approximation. This leads to the definition of Bellman rank, a measure that captures the difficulty of learning in these settings.

Definition 5. *(Bellman factorization and Bellman rank).*
We say that a MDP $(\mathcal{X}, A, H, P)$ and $\mathcal{F} \subseteq \mathcal{X} \times A \to [0, 1]$ admit Bellman factorization with Bellman rank d and norm parameter ζ, if there exists $\nu_h : \mathcal{F} \to \mathbb{R}^d$, $\xi_h : \mathcal{F} \to \mathbb{R}^d$ for each $h \in [H]$, such that for any $f, f' \in \mathcal{F}, h \in [H]$,

$$\mathcal{E}(f, \pi_{f'}, h) = \langle \nu_h(f'), \xi_h(f) \rangle, \tag{2}$$

Where $\mathcal{E}(f, \pi_{f'}, h)$ is the average Bellman error, and $\|\nu_h(f')\|_2 \cdot \|\xi_h(f)\|_2 \leq \zeta < \infty$.

This factorization not only decomposes the average Bellman error into more manageable components but also introduces the Bellman rank as a key parameter that influences the learning complexity in RL tasks. A low Bellman rank indicates that the problem is easier to solve with fewer samples, making it an important metric in designing efficient RL algorithms.

Finally, we formalize the notion of the validity of a function f with respect to the Bellman equation. This concept is central to determining whether a function is a good fit for modeling the value function in an MDP.

Definition 6 (Bellman equations and validity of f)**.** *Given an (f, π, h) triple, a Bellman equation posits $\mathcal{E}(f, \pi, h) = 0$. We say $f \in \mathcal{F}$ is valid if the Bellman equation on $(f, \pi_{f'}, h)$ holds for every $f' \in \mathcal{F}, h \in [H]$.*

Fact 7 ($Q^\star$ is always valid)**.** *Given an MDP and a space of functions $\mathcal{F} : \mathcal{S} \times [H] \times \mathcal{A} \to [0, 1]$, if the optimal Q-value function of the MDP $Q^\star$ lies in $\mathcal{F}$, then in the corresponding MDP with $\mathcal{X} = \mathcal{S} \times [H]$, $Q^\star$ is valid.*

The fact that $Q^\star$ is always valid within the function class $\mathcal{F}$ highlights the importance of selecting a function class that includes $Q^\star$. This leads us to the assumption of realizability, which formalizes the condition that such an optimal function exists within the chosen hypothesis class.

Assumption 8 (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.

Realizability is a critical assumption that ensures the function class $\mathcal{F}$ is rich enough to include an optimal hypothesis. However, for reinforcement learning algorithms to perform well, it's not only important that $Q^\star$ is realizable, but also that the function class $\mathcal{F}$ is closed under the Bellman operator. This leads to the completeness assumption:

Assumption 9. *[Completeness] For a Function class $\mathcal{F}$, the completeness assumption define if for another function class $\mathcal{G}$, holds that $\mathcal{F}$ is closed under bellman operator. i.e. $\{g | g = \mathcal{T} f \ \ \forall f \in \mathcal{F}\} \subset \mathcal{G}$*

To measure the capacity of the function class $\mathcal{F}$, we introduce the concept of the ϵ-covering number, which helps in quantifying the richness of the hypothesis space.

Definition 10 (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 $\rho(f, f') \leq \epsilon$.*

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 and g represent 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 corresponds to the maximal (squared) Hellinger distance between two probability distribution sequences.

We now define the structure of the function class representation for the finite horizon setup, which is essential for understanding how the learner approximates the optimal Q-value function over time.

Definition 11 (Function Class Representation)**.** *For finite horizon setup, 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 $(H + 1)^{th}$ steps, we always set $f_{H+1} = 0$.*

This representation ensures that the function class is appropriately structured to capture the value function across all time steps in the finite horizon setup, thereby enabling the learning algorithm to effectively approximate the optimal policy.

2.3 Key Concepts in Multi-Agent Reinforcement Learning

In multi-agent reinforcement learning (MARL), multiple agents interact within a shared environment, learning to optimize their individual or collective rewards through cooperation or competition. This setting introduces additional complexities beyond single-agent RL, such as coordinating actions, sharing information, and ensuring that learning remains stable and efficient across all agents. In this subsection, we outline the learning objectives in the multi-agent context and describe how agents interact within the environment.

Learning Objective Our goal is to achieve a PAC-guarantee while ensuring sample-efficient learning in Markov Decision Processes (MDPs) where the value function can be approximated by a structured function class. Specifically, we seek to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ whose value function can be effectively represented by a function approximator $f \in \mathcal{F}$. The learning algorithm is designed to use a polynomial number of trajectories, $Poly(d, H, N, \frac{1}{\epsilon}, \delta)$, to identify an ϵ-suboptimal policy, where d is the dimension capturing the complexity of $\mathcal{F}$, H is the planning horizon, N is the cardinality of $\mathcal{F}$, and δ represents the allowable probability of failure.

Multi-Agent Interaction. A team I of m agents interacts with the Markov Decision Process (MDP) $\mathcal{M}$ with the collective goal of optimizing a shared objective, such as maximizing the cumulative reward across the team. At the beginning of each episode t, each agent $A_i \in I$ selects a policy $\pi^{t,i}$ and begins in the initial context x_1^t. At each time step $h = 1, \ldots, H$, every agent observes its current context $x_h^{t,i}$ and selects an action $a_h^{t,i} \sim \pi^{t,i}(\cdot|x_h^{t,i})$, where the chosen action is based on the agent's current policy and observed state. The agents' interactions and decisions influence the state transitions and the overall outcome, requiring coordination and strategy to achieve the common goal effectively.
In the **fresh randomness** model, the next state is sampled independently for each agent, i.e., $s_{h+1}^{t,i} \sim$

$p^h(\cdot|s_h^{t,i}, a_h^{t,i})$. For **non-fresh randomness**, the next state $s_{h+1}^{t,i}$ is the same for each state-action pair $S_h^k(s,a) \sim p^h(\cdot|s,a)$ ahead of the episode, and then every agent i that takes action a in s at time h transitions to the same state $S_h^k(s,a)$, i.e., $s_{h+1}^{t,i} = S_h^k(s_h^{t,i}, a_h^{t,i})$. Similarly, the reward $R_h^{t,i}$ suffered by the agent is either sampled independently when randomness is fresh, or sampled once for each state-action pair (s,a) ahead of the episode when randomness is non-fresh. At the end of the episode, the team observes the trajectories and costs of all agents $\left\{s_h^{t,i}, a_h^{t,i}, R_h^{t,i}\right\}_{h=1,i=1}^{H,m}$.

Switching Cost The concept of switching cost is used to quantify the adaptability of reinforcement learning algorithms. The main focus of our work is the global switching cost, which counts the number of policy changes in the running of the algorithm in K episodes, namely:

$$N_{\text{switch}} \triangleq \sum_{k=1}^{K-1} \mathbb{I}\left\{\pi_k \neq \pi_{k+1}\right\} \tag{3}$$

Rare Policy Switching (RPS) in RL refers to a strategy where the learning agent updates its policy infrequently and only under certain conditions, typically when enough significant data has been gathered to justify a change. This approach is designed to limit the frequency of policy changes to reduce the computational and learning overhead associated with each switch, aiming to maintain a stable learning trajectory over longer periods. The infrequency of the switches contributes to the switching cost, balancing the need for policy improvement against the stability of the learning process. Rare policy switching is particularly valuable in environments where the act of switching itself incurs a cost or when the learning algorithm benefits from longer periods of consistent policy application to adequately assess performance.

Batch Learning [Multi-Batch] in the context of RL is a learning paradigm where the decision-making agent divides the learning process into discrete, sequential segments called "batches". Within each batch, the agent commits to a single policy, only updating or changing this policy at the conclusion of the batch. This approach contrasts with more continuous learning strategies where policies may be updated more frequently.

The difference between the rare policy switch setting and batch learning setting is that, in the RPS case, the algorithm can adaptively decide when to switch the policy based on the data, whereas in the batch learning case, the episodes where the agent adopts a new policy are deterministically decided before the first episode. In other words, in reinforcement learning with rare policy switches, we are confident to achieve a sublinear $\tilde{O}(\sqrt{K})$ regret, e.g., using an online reinforcement learning algorithm that switches the policy after each episode. The goal is to attain the desired regret with a small number of policy switches. In contrast, in the batch learning setting, with B fixed, we aim to minimize the regret, under the restriction that the number of policy switches is no more than B.

Problem Equivalence. We observed that Multi-Batch and Multi-Agent frameworks exhibit equivalent characteristics, suggesting that a solution developed for one can be effectively adapted to address the other. On the other hand, RPS aligns in parallel with these frameworks is still unknown. This complexity largely stems from the integrated feedback gathered at the end of each episode, which directly influences policy switch.

Asynchronous Communication In the multi-agent setting, the agents need to communicate (i.e., share data) to collaboratively learn the underlying optimal policy while minimizing the regret. Without communication, the problem would reduce to m separate single-agent MDP problems. In general case, this would lead to a worst-case regret of order $\tilde{O}(m\sqrt{K/m})$, which suffers from an extra $\sqrt{m}$ factor as compared to the regret in the single-agent setting. In the following sections we will show that this extra factor can be avoided at the cost of a small number of communication rounds.

3 Olive: Multi-Agent

3.1 Introduction

Building upon the pioneering work of Jiang et al. (2017) our exploration into the complexities of reinforcement learning (RL) benefits significantly from their foundational contributions. They introduced Contextual Decision Processes (CDPs), a sophisticated framework designed to navigate decision-making scenarios where agents process complex, rich sensory information and historical context. Additionally, they developed the Bellman factorization structure, a critical advancement that systematizes the decision-making process. To solve the challenges posed by CDPs, they proposed the OLIVE algorithm, a novel approach that significantly enhances the understanding and application of reinforcement learning in environments laden with intricate sensory inputs.

Our research extends their innovative work into new territories, specifically focusing on multi-agent environments and the utilization of generative models, where the learner can query any state-action pair and observe a next state and reward. By adapting the principles underlying CDPs and the Bellman factorization, along with the foundational strategies of the OLIVE algorithm, we present novel contributions tailored for the complexities of interacting agents and dynamic, generative settings. Our work not only demonstrates the versatility and applicability of the framework in broader contexts but also showcases our advancements in deploying these concepts to address and solve new, emergent challenges in reinforcement learning. Through this, we aim to push the boundaries of what's possible in RL, leveraging the solid foundation laid by to explore and innovate within the realms of multi-agent systems and generative environments.

3.1.1 Contextual Decision Processes (CDPs)

A Contextual Decision Process s make minimal assumptions to capture a very general class of RL problems and are defined as follows.

Definition 12 (Contextual Decision Process (CDP))**.** *A (finite-horizon) Contextual Decision Process (CDP for short) is defined as a tuple $(\mathcal{X}, \mathcal{A}, H, P)$, where $\mathcal{X}$ is the context space, $\mathcal{A}$ is the action space, and H is the horizon of the problem. $P = (P_\Phi, P_+)$ is the system descriptor, where $P_\Phi \in \Delta(\mathcal{X})$ is a distribution over initial contexts, that is $x_1 \sim P_\Phi$, and $P_+ : (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^* \times \mathcal{X} \times \mathcal{A} \to \Delta(\mathbb{R} \times \mathcal{X})$ elicits the next reward and context from the interactions so far $x_1, a_1, r_1, \ldots, x_h, a_h$:*

$$(r_h, x_{h+1}) \sim P_+(x_1, a_1, r_1, \ldots, x_h, a_h).$$

In a CDP , the agent's interaction with the environment proceeds in episodes. In each episode, the agent observes a context x_1, takes action a_1, receives reward r_1 and observes x_2, repeating H times. A policy $\pi : \mathcal{X} \to \mathcal{A}$ specifies the decision-making strategy of an agent, that is $a_h = \pi(x_h),\ \forall h \in [H]$, and induces a distribution over the trajectory $(x_1, a_1, r_1, \ldots, x_H, a_H, r_H, x_{H+1})$ according to the system descriptor P.

The value of a policy, V^π, is defined as

$$V^\pi = \mathbb{E}_P \left[\sum_{h=1}^H r_H \;\middle|\; a_{1:H} \sim \pi \right], \tag{4}$$

where $a_{1:H} \sim \pi$ abbreviates for $a_1 = \pi(x_1), \ldots, a_H = \pi(x_H)$. Here, and in the sequel, the expectation is always taken over contexts and rewards drawn according to the system descriptor P, so we suppress the subscript P for brevity. The goal of the agent is to find a policy π that attains the largest value.

Below we show that CDPs capture classical RL models, including MDPs and POMDPs, and the optimal policies can be expressed as a function of appropriately chosen contexts.

Example 13 (*MDPs with states as contexts*)**.** *Consider a finite-horizon MDP $(\mathcal{S}, \mathcal{A}, H, \Gamma_1, \Gamma, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the horizon, $\Gamma_1 \in \Delta(\mathcal{S})$ is the initial state distribution, $\Gamma : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition function, $R : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ is the reward function, and an episode takes the form of $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)$. We can convert the MDP to a CDP $(\mathcal{X}, \mathcal{A}, H, P)$ by letting $\mathcal{X} = \mathcal{S} \times [H]$ and $x_h = (s_h, h)$, which allows the set of policies $\{\mathcal{X} \to \mathcal{A}\}$ to contain the optimal policy (Puterman, 2014). The system descriptor is $P = (P_\Phi, P_+)$, where $P_\Phi(x_1) = \Gamma_1(s_1)$, and $P_+(r_h, x_{h+1} \,|\, x_1, a_1, r_1, \ldots, x_h, a_h) = R(r_h|s_h, a_h)\, \Gamma(s_{h+1}|s_h, a_h)$.*

Besides this example, we refer readers to the paper for additional examples that demonstrate the flexibility of keeping contexts separate from states. We will employ the realizabllity assumption, noting that, in general, $\mathcal{F}$ may not contain Q^* or any valid functions at all. It is desirable to have an algorithm that is robust to such scenarios. Our improvement of the algorithm, like its predecessor, requires only an approximate notion of validity, ensuring graceful degradation in the results.

Finally, we introduce a regularity assumption on the rewards.

Assumption 14 (Boundedness of rewards)**.** *We assume that regardless of how actions are chosen, for any $h = 1, \ldots, H$, $r_h \geq 0$ and $\sum_{h=1}^H r_h \leq 1$ almost surely.*

We aim to improve the dependency on H, by working on all the layers simultaneously. therefore we will use the next definition:

Definition 15. *Let $\eta > 0$, $t \in [T]$, we define h^t to be the number of layers $h \in [H]$ where $|\mathcal{E}(f_t, \pi_t, h_j)| \geq \eta$. In other words, denote $\mathcal{H}^t = \left\{h | h \in [H] \;\&\; |\mathcal{E}(f_t, \pi_t, h_j)| \geq \eta\right\}$ and define $h^t = |\mathcal{H}^t|$. In addition, let denote $\tilde{h} = \frac{1}{T} \times \sum_{t=1}^T h^t$, the average amount of layers that have a large exception.*

3.1.2 Section Overview

In this section we present two algorithms. The first, $OLIVE - Multi - Agent$, will use the power of multi agent to improve simultaniously H layers. The second, $OLIVE - Generative - Multi - Agent$, will use the generative model (Definition 3) to improve the sample complexity in multi agent.

In addition to $OLIVE$ inputs, we have now a set I of agents and we will show how it is affect the sample complexity. The algorithm $OLIVE - Multi - Agent$ achieve sample complexity of:

$$\tilde{\mathcal{O}}\left(\frac{Hd\log(Hd\zeta/\epsilon)}{\tilde{h}} \times \left(1 + \frac{d\tilde{h}H^2A}{|I|\epsilon^2} \log(HN\zeta/\delta)\right)\right), \tag{5}$$

per agent. Furthermore, the algorithm $OLIVE - Generative - Multi - Agent$ which achieve a better sample complexity per agent:

$$\tilde{\mathcal{O}}\left(\frac{Hd\log(Hd\zeta/\epsilon)}{\tilde{h}} \times \left(1 + \frac{dH^2A}{|I|\epsilon^2} \log(N\zeta/\delta)\right)\right). \tag{6}$$

3.2 Olive: Multi-Agent

Theorem 16. *For any $\epsilon, \delta \in (0,1)$, any CDP and function class $\mathcal{F}$ that admits a Bellman factorization with parameters d, ζ, run $OLIVE - Multi - Agent$ with I as set of agents and the following parameters:*

$$\phi = \frac{\epsilon}{12H\sqrt{d}}, \qquad \eta = 5\epsilon/8H, \qquad n_{est} = \frac{32}{\epsilon^2}\log(6N/\delta),$$
$$n_{eval} = \frac{288H^2}{\epsilon^2}\log\left(\frac{12H^2 d\log(6H\sqrt{d}\zeta/\epsilon)}{\delta}\right),$$
$$n_{train} = \frac{4608H^2 dA}{\epsilon^2}\log\left(\frac{12NHd\log(6H\sqrt{d}\zeta/\epsilon)}{\delta}\right).$$

Then, with probability at least $1-\delta$, $OLIVE - Multi - Agent$ halts and returns a policy π_{out} that satisfies $V_{\pi_{out}} \geq V^{\star}_{\mathcal{F}} - \epsilon$, and the number of episodes required is at most

$$\tilde{\mathcal{O}}\left(\frac{Hd}{\tilde{h}} \times \left(1 + \frac{d\tilde{h}H^2 A}{|I|\epsilon^2}\right)\right), \tag{9}$$

We split the proof into three parts:

1. Subsection 3.2.1, bound the number of phases the algorithm executes before termination.
2. Subsection 3.2.2, bound the number samples per agent in each phase.
3. Subsection 3.2.3, prove optimality of π_{out}.

3.2.1 Iteration Complexity

In this subsection, our goal is to establish an upper limit on the number of phases the algorithm executes before termination. In order to create a common language that will help us analyze the algorithm more easily we will introduce the following term:

Definition 17. *A layer $h \in \mathcal{H}^t$ is called **corrected**. (Recall that $h^t = |S^t|$.)*

Based on this definition, we make the following distinction:

Observation 18. *In every phase $t > 0$, we corrected $h^t > 0$ layers, or we terminate.*

Therefore, we are left to understand how many times during the algorithm we make a correction to some layer.

Lemma 19. *Every layer $h \in [H]$, can be corrected at most $\tilde{\mathcal{O}}(d)$ times.*

The proof will be found in Lemma 26. Next we bound the iteration complexity of the $OLIVE - Multi - Agent$:

Theorem 20. *Algorithm $OLIVE - Multi - Agent$ terminates after at most $\tilde{\mathcal{O}}(\frac{Hd}{\tilde{h}})$ phases.*

Algorithm 1 OLIVE-Multi-Agent$(\mathcal{F}, d, I, \zeta, \epsilon, \delta)$

Collect n_{est} trajectories with actions taken in an arbitrary manner; save initial contexts $\{x_1^{(i)}\}_{i=1}^{n_{\text{est}}}$.
Estimate the predicted value for each $f \in \mathcal{F}$: $\hat{V}_f = \frac{1}{n_{\text{est}}} \sum_{i=1}^{n_{\text{est}}} f(x_1^{(i)}, \pi_f(x_1^{(i)}))$.
$\mathcal{F}_0 \leftarrow \mathcal{F}$.
for $t = 1, 2, \ldots$ **do**
  Choose policy $f_t = \operatorname{argmax}_{f \in \mathcal{F}_{t-1}} \hat{V}_f$, $\pi_t = \pi_{f_t}$.
  for Agent $A_i \in [I]$ **do**
    Collect $n = \lceil \frac{n_{\text{eval}}}{|I|} \rceil$ trajectories by following π_t.
    Add to data set: $\mathcal{D}^t_{eval} = \mathcal{D}^t_{eval} \cup \{(x_1^{(i)}, a_1^{(i)}, r_1^i, \ldots, x_H^{(i)}, a_H^{(i)}, r_H^{(i)})\}_{i=1}^n$.
  Estimate $\forall h \in [H]$,

$$\tilde{\mathcal{E}}(f_t, \pi_t, h) = \frac{1}{|\mathcal{D}^t_{eval}|} \sum_{i \in \mathcal{D}^t_{eval}} \left[f_t(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_t(x_{h+1}^{(i)}, a_{h+1}^{(i)}) \right]. \tag{7}$$

  if $(1/H) \sum_{h=1}^H |\tilde{\mathcal{E}}(f_t, \pi_t, h)| \leq \eta$ **then**
    Terminate and output $\pi_{out} = \pi_t$.
  else Define the set $\mathcal{H}^t = \left\{h | h \in [H] \ \& \ |\tilde{\mathcal{E}}(f_t, \pi_t, h_j)| \geq \eta\right\}$ {Definition 15}
  for Agent $A_i \in [I]$ **do**
    Sample uniformly at random $h_i^t \in \mathcal{H}^t$.
    Collect $N = \lceil \frac{h^t n_{\text{train}} \log(H/\delta)}{|I|} \rceil$ trajectories by following
    $\pi_t^{h_i^t} := \begin{cases} \pi_t & : h \neq h_i^t \\ \pi_{\text{UNIF}} & : h = h_i^t \end{cases}$. where π_{UNIF} is the uniform distrubtion over actions
    Add to data set: $\mathcal{D}^t_{h_i^t} = \mathcal{D}^t_{h_i^t} \cup \{(x_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \ldots, x_{h_i^t}^{(i)}, \hat{a}_{h_i^t}^{(i)}, \hat{r}_{h_i^t}^{(i)}, \hat{x}_{h_i^t+1}^{(i)})\}_{i=1}^n$.
  Estimate $\forall h \in \mathcal{H}^t$ for which $|\tilde{\mathcal{E}}(f_t, \pi_t, h)| \geq \eta$: {One is guaranteed to exist }

$$\hat{\mathcal{E}}(f, \pi_t, h) = \frac{1}{|\mathcal{D}^t_h|} \sum_{i \in \mathcal{D}^t_h} \frac{\mathbf{1}[\hat{a}_h^{(i)} = \pi_f(x_h^{(i)})]}{1/A} \Big(f(x_h^{(i)}, \hat{a}_h^{(i)}) - \hat{r}_h^{(i)} - f(\hat{x}_{h+1}^{(i)}, \pi_f(\hat{x}_{h+1}^{(i)})) \Big). \tag{8}$$

  Learn

$$\mathcal{F}_t = \left\{ f \in \bigcap_{h \in [H]} \left\{ f \in \mathcal{F}_{t-1} : \left| \hat{\mathcal{E}}(f, \pi_t, h) \right| \leq \phi \right\} \right\}.$$

Proof. Let T be the number of phases until termination. By Lemma 19, each level $h \in [H]$, we corrected it at most $d' = \tilde{\mathcal{O}}(d)$ times. Therefore, for any h:

$$\sum_{i=1}^{T} \mathbf{I}\{|\tilde{\mathcal{E}}(f_t, \pi_t, h)| > \eta\} \leq d'.$$

On the other hand, in phase t we improve h^t layers simultaneously, and we get that:

$$\sum_{i=1}^{T} h_t = \sum_{i=1}^{T} \sum_{h \in [H]} \mathbf{I}\{|\tilde{\mathcal{E}}(f_t, \pi_t, h)| > \eta\} \leq Hd'.$$

Finally, using the Definition 15 we get that:

$$T\tilde{h} = \sum_{i=1}^{T} h_t \leq Hd' \Longrightarrow T \leq \frac{Hd'}{\tilde{h}}.$$

□

3.2.2 Sample Complexity

Theorem 21 (Sample Complexity Per Agent)**.** *The average sample complexity in each phase of the algorithm, for an agent is given by:*

$$\tilde{\mathcal{O}}\left((1 + \frac{d\tilde{h}H^2A}{|I|\epsilon^2})\right), \tag{10}$$

where d is a Bellman rank dimensionality parameter, H is the horizon, A is the number of actions, $|I|$ is the number of agents, ϵ is the accuracy parameter, and δ is the confidence parameter.

Proof. According to the definition provided for $\mathcal{D}^t_{eval}$ within $OLIVE - Multi - Agent$, it is guaranteed that the cardinality of $\mathcal{D}^t_{eval}$ as specified in line 7 is at least n_{eval}. Furthermore, in each round we sample at least $\frac{h^t n_{\text{train}} \log(H/\delta)}{|I|}$. Therefore, using n_{eval} and n_{train} as like in Theorem 16, we get that at the average iteration $t \in [T]$, agent $A_i \in [I]$ have a sample complexity of:

$$\frac{1}{T}\sum_{t=1}^{T}\left(1 + \frac{dh^tH^2A}{|I|\epsilon^2}\right) = \tilde{\mathcal{O}}\left(1 + \frac{d\tilde{h}H^2A}{|I|\epsilon^2}\right), \tag{11}$$

□

Corollary 22. *Therefore, using Theorem 21, Theorem 20, the sample complexity of the algorithm is the number of samples in each episode time the amount of episode:*

$$\tilde{\mathcal{O}}\left(\frac{Hd}{\tilde{h}} \times (1 + \frac{d\tilde{h}H^2A}{|I|\epsilon^2})\right),$$

which is what we had to proof in Theorem 16.

3.2.3 Near optimality of π_{out}

We left to show that with probability at least $1-\delta$, $OLIVE-Multi-Agent$ halts and returns a policy π_{out} that satisfies $V_{\hat{\pi}} \geq V^{\star}_{\mathcal{F}} - \epsilon$.

Lemma 23 (Optimism drives exploration)**.** *Suppose $f^{\star}$ is never eliminated. Also, suppose the estimates $\hat{V}_f$ and $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ in Line 2 and 9 always satisfy*

$$|\hat{V}_f - V_f| \leq \epsilon/8, \qquad and \qquad |\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \leq \frac{\epsilon}{8H} \tag{12}$$

throughout the execution of the algorithm.
Then in any phase t, one of the following two statements holds:

(i) the algorithm does not terminate and for $h_t \in \mathcal{H}^t$ holds that:

$$\mathcal{E}(f_t, \pi_t, h_t) \geq \eta - \frac{\epsilon}{8H} = \frac{\epsilon}{2H}, \tag{13}$$

(ii) the algorithm terminates and the output policy π_t satisfies $V_{\pi_t} \geq V^{\star}_{\mathcal{F}} - \epsilon$.

Proof. If we do not terminate, the first claim follows from the algorithm and the assumption that $|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \leq \frac{\epsilon}{8H}$.

We now show that if the algorithm terminates after phase t, and outputs $\pi_{out} = \pi_t$, then it is a near optimal policy. Recall that f_t was the selected function approximation.

$$\begin{aligned}
V_{\pi_t} &= V_{f_t} - \sum_h \mathcal{E}(f_t, \pi_t, h) \\
&\geq \hat{V}_{f_t} - \epsilon/8 - \sum_h (\tilde{\mathcal{E}}(f_t, \pi_t, h) + \epsilon/8H) \\
&\geq \hat{V}_{f_t} - \epsilon/8 - H\eta - \epsilon/8 \\
&\geq \hat{V}_{f_{max}} - 7\epsilon/8 \quad (f_t \text{ is the maximizer of } \hat{V}_f) \\
&\geq V_{f_{max}} - \epsilon/8 - 7\epsilon/8 \\
&\geq V^{\star}_{\mathcal{F}} - \epsilon.
\end{aligned}$$

The first identity is due to Lemma 27. The inequalities is based on Equation 12 and the termination criterion. Since $f^{\star}$ is never eliminated, the optimisms guarantees that $\hat{V}_{f_{max}} \geq \hat{V}_{f^{\star}}$. The proof that $f^{\star}$ is never eliminated is part of 26. □

The sample sizes of n_{est} and n_{eval} were determined based on a deviation bound to ensure sufficient statistical confidence in the observed outcomes for Equation 12.

3.2.4 Complete The Proof

To complete the proof we are left to show three things:

- $f^{\star}$ is never eliminated.
- Proof for Lemma 19 which bound the number of iteration we are doing.

- Deviation bounds for all of our estimations.

We now proceed to examine the proof.

Lemma 24. *[Algebraic Lemma] Given $c, D > 0$, and two sets $V_0, P \subseteq \mathbb{R}^d$ such that $\max_{p,v \in P,V_0} ||p^T v|| \leq D$. We define a game where for each round $t > 0$, the first player pick $p_t \in P$ s.t. $\exists v \in V_{t-1}$ s.t. $|p_t^T v| > c\sqrt{d}$. The second player returns an ellipsoid B_t s.t. for $V_t = \left\{v \in V_{t-1} \middle| |p_t^T v| \leq c\right\}$, such that $V_t \subset B_t$. Then we get that the process ends after $T \leq d \log(\frac{D}{2c})$.*

The proof for that base on Todd (1982), minimum volume enclosing ellipsoid(MVEE) convergence, and on Lemma 11 in Jiang et al. (2017) which we add here for completeness:

Lemma 25. *Consider a closed and bounded set $V \subset \mathbb{R}^d$ and a vector $p \in \mathbb{R}^d$. Let B be any enclosing ellipsoid of V that is centered at the origin, and we abuse the same symbol for the symmetric positive definite matrix that defines the ellipsoid, i.e., $B = \{v \in \mathbb{R}^d : v^\top B^{-1} v \leq 1\}$. Suppose there exists $v \in V$ with $|p^\top v| \geq \kappa$ and define B_+ as the minimum volume enclosing ellipsoid of $\{v \in B : |p^\top v| \leq \gamma\}$. If $\gamma/\kappa \leq 1/\sqrt{d}$, we have*

$$\frac{vol(B_+)}{vol(B)} \leq \sqrt{d}\frac{\gamma}{\kappa}\left(\frac{d}{d-1}\right)^{(d-1)/2}\left(1 - \frac{\gamma^2}{\kappa^2}\right)^{(d-1)/2}. \tag{14}$$

Therefore, using lemma 24, we can bound our iteration complexity finally, and furthermore, prove that $f^\star$ is never eliminate.

Lemma 26 (Iteration complexity)**.** *For an arbitrary $h_t \in \mathcal{H}^t$, if $\hat{\mathcal{E}}(f, \pi_t, h_t)$ in Eq. (8) always satisfies*

$$|\hat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \leq \phi \tag{15}$$

throughout the execution of the algorithm (ϕ is the threshold in the elimination criterion), then $f^\star$ is never eliminated.
Furthermore, for any particular level h, if whenever $h \in \mathcal{H}^t$ we have (base on 13)

$$|\mathcal{E}(f_t, \pi_t, h)| \geq 6\sqrt{d}\phi, \tag{16}$$

then the number of iterations that $h \in \mathcal{H}^t$ is at most $d \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3}$.

Proof. The first claim that $f^\star$ is never eliminated can be directly inferred from the condition $|\mathcal{E}(f, \pi_t, h_t)| = 0$. The deviation bound in Lemma 31, which assume the high probability event of Claim 30, confirms that $f^\star$ always satisfies Equation 15 (w.p $> 1 - \delta$). Similar, the selection of the sample size n_{train} is grounded on a deviation bound that assures a high degree of statistical confidence in the outcomes relevant to Equation 15. The second claim is a direct result of the Lemma 24 with the following parameters:

- $p_1, \ldots, p_T$. $p_\tau := \nu_h(f_{i_\tau})$ where $\nu_h(\cdot)$ is given in Definition 6. Recall that f_{i_τ} is the optimistic function used for exploration in iteration $t = i_\tau$.
- $\mathcal{U}(\mathcal{F}_{i_0}), \mathcal{U}(\mathcal{F}_{i_1}), \ldots, \mathcal{U}(\mathcal{F}_{i_T})$. $\mathcal{U}(\mathcal{F}_{i_\tau}) = \{\xi_h(f) : f \in \mathcal{F}_{i_\tau}\}$ where $\xi_h(f) \in \mathbb{R}^d$ is given in Definition 6.
- $\Psi = \sup_{f \in \mathcal{F}} \|\nu_h(f)\|_2$, and $\Phi = \sup_{f \in \mathcal{F}} \|\xi_h(f)\|_2$. By Definition 6, $\Psi \cdot \Phi \leq \zeta$.

- $V_0, V_1, \ldots, V_T$. $V_0 := \{v : \|v\|_2 \le \Phi\}$, and $V_\tau := \{v \in V_{\tau-1} : |p_\tau^\top v| \le 2\phi + \theta + \eta\}$.
- $B_0, B_1, \ldots, B_T$. B_τ is a minimum volume enclosing ellipsoid (MVEE) of V_τ.

□

Lemma 27. *[Bellman Decomposition] With $V_f = \mathbb{E}[f(x_1, \pi_f(x_1))]$, we have*

$$V_f - V_{\pi_f} = \sum_{h=1}^{H} \mathcal{E}(f, \pi_f, h). \tag{17}$$

Proof. Recall from Definition 4 that the average Bellman errors are defined as

$$\mathcal{E}(f, \pi, h) = \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:h-1} \sim \pi,\ a_{h:h+1} \sim \pi_f\right].$$

Expanding RHS of Eq. (17), we get

$$\sum_{h=1}^{H} \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:h-1} \sim \pi_f,\ a_{h:h+1} \sim \pi_f\right].$$

Since all H expected values share the same distribution over trajectories, which is the one induced by $a_{1:H} \sim \pi_f$, the above expression is equal to

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:H} \sim \pi_f\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \Big(f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})\Big) \;\middle|\; a_{1:H} \sim \pi_f\right] \\
&= \mathbb{E}\left[f(x_1, \pi_f(x_1))\right] - \mathbb{E}\left[r_h \;\middle|\; a_{1:H} \sim \pi_f\right] = V_f - V_{\pi_f}.
\end{aligned}$$

□

Lemma 28 (Deviation Bound for $\hat{V}_f$)**.** *With probability at least $1 - \delta$,*

$$|\hat{V}_f - V_f| \le \sqrt{\frac{1}{2n_{est}} \log \frac{2N}{\delta}}$$

holds for all $f \in \mathcal{F}$ simultaneously. Hence, we can set $n_{est} \ge \frac{32}{\epsilon^2} \log \frac{2N}{\delta}$ to guarantee that $|\hat{V}_f - V_f| \le \epsilon/8$.

Proof. The bound follows from a straight-forward application of Hoeffding's inequality and the union bound, and we only need to verify that the V_f is the expected value of the $\hat{V}_f$, and the range of the random variables is $[0, 1]$. □

Lemma 29 (Deviation Bound for $\tilde{\mathcal{E}}(f_t, \pi_t, h)$)**.** *For any fixed f_t, with probability at least $1 - \delta$,*

$$|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le 3\sqrt{\frac{1}{2n_{eval}} \log \frac{2H}{\delta}}$$

holds for all $h \in [H]$ simultaneously. Hence, for any $n_{eval} \ge \frac{288H^2}{\epsilon^2} \log \frac{2H}{\delta}$, with probability at least $1 - \delta$ we have $|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \frac{\epsilon}{8H}$.

Proof. This bound is another straight-forward application of Hoeffding's inequality and the union bound, except that the random variables that go into the average have range $[-1, 2]$, and we have to realize that $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ is an unbiased estimate of $\mathcal{E}(f_t, \pi_t, h)$. □

Claim 30 (The High Probability Event)**.** *For all $t \in [T]$ and $h \in [H]$ with high probability, $|\mathcal{D}_h^t| > n_{train}$.*

In each round we sample at least $\frac{h^t n_{\text{train}} \log(HT/\delta)}{|I|}$ and split them uniformly to $\{\mathcal{D}_h^t\}_{h \in \mathcal{H}^t}$. Therefore, We get that with a probability exceeding greater than $1 - \delta$,the cardinality of each $\mathcal{D}_h^t$ for every $h \in \mathcal{H}^t$ as established in line 15, is greater than n_{train}.

Lemma 31 (Deviation Bound for $\hat{\mathcal{E}}(f, \pi_t, h_j^t)$)**.** *Under the high probability event (of Claim 30), with probability at least $1 - \delta$, for any fixed π_t and for any $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$, we have*

$$|\hat{\mathcal{E}}(f, \pi_t, h) - \mathcal{E}(f, \pi_t, h)| \leq \sqrt{\frac{8A \log \frac{2NH}{\delta}}{n_{train}}} + \frac{2A \log \frac{2NH}{\delta}}{n_{train}} \tag{18}$$

holds for all $f \in \mathcal{F}$ and $h \in [H]$ simultaneously. Hence, for any $n_{train} \geq \frac{32A}{\phi^2} \log \frac{2NH}{\delta}$ and $\phi \leq 4$, with probability at least $1 - \delta$ we have $|\hat{\mathcal{E}}(f, \pi_t, h) - \mathcal{E}(f, \pi_t, h)| \leq \phi$.

Proof. We first show that $\hat{\mathcal{E}}(f, \pi_t, h)$ is an average of i.i.d. random variables with mean $\mathcal{E}(f, \pi_t, h)$. We will show that for fix $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$ the statement in Equation 18 hold, and with union bound we conclude it for any $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$. We use μ as a shorthand for the distribution over trajectories induced by $a_1, \ldots, a_{h-1} \sim \pi_t, \hat{a}_h \sim \text{unif}(\mathcal{A})$, which is the distribution of data used to estimate $\hat{\mathcal{E}}(f, \pi_t, h)$. On the other hand, let μ' denote the distribution over trajectories induced by $a_1, \ldots, a_{h-1} \sim \pi_t, \hat{a}_h \sim \pi_f$. The importance weight used in Eq. (8) essentially converts the distribution from μ to μ', hence the expected value of $\hat{\mathcal{E}}(f, \pi_t, h)$ can be written as

$$\begin{aligned}
&\mathbb{E}_\mu \left[A \, \mathbf{1}[a_h = \pi_f(x_h)] \left(f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \right) \right] \\
&\quad = \mathbb{E}_{\mu'} \left[f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \right] = \mathcal{E}(f, \pi_t, h_t).
\end{aligned}$$

Now, we apply Bernstein's inequality. We first analyze the 2nd-moment of the random variable. Defining $Y(x_h, a_h, r_h, x_{h+1}) = f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \in [-2, 1]$, the 2nd-moment is

$$\begin{aligned}
&\mathbb{E}_\mu \left[\left(A\mathbf{1}[a_h = \pi_f(x_h)] Y(x_h, a_h, r_h, x_{h+1}) \right)^2 \right] \\
&= \mathbb{P}_\mu[a_h = \pi_f(x_h)] \cdot \mathbb{E}_\mu \left[\left(AY(x_h, a_h, r_h, x_{h+1}) \right)^2 \,\middle|\, a_h = \pi_f(x_h) \right] + \mathbb{P}_\mu[a_h \neq \pi_f(x_h)] \cdot 0 \\
&\leq \frac{1}{A} \mathbb{E}_\mu \left[A^2 \cdot 4 \,\middle|\, a_h = \pi_f(x_h) \right] = 4A.
\end{aligned}$$

Next we check the range of the centered random variable. The uncentered variable lies in $[-2A, A]$, and the expected value is in $[-2, 1]$, so the centered variable lies in $[-2A - 1, A + 2] \subseteq [-3A, 3A]$. Applying Bernstein's inequality, we have with probability at least $1 - \delta$,

$$\begin{aligned}
|\hat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| &\leq \sqrt{\frac{2 \operatorname{Var}\left[A\mathbf{1}[a_h = \pi_f(x_h)] y(x_h, a_h, r_h, x_{h+1}) \right] \log \frac{2HN}{\delta}}{n_{\text{train}}}} + \frac{6A \log \frac{2HN}{\delta}}{3n_{\text{train}}} \\
&\leq \sqrt{\frac{8A \log \frac{2HN}{\delta}}{n_{\text{train}}}} + \frac{2A \log \frac{2HN}{\delta}}{n_{\text{train}}}. \qquad \text{(variance is bounded by 2nd-moment)}
\end{aligned}$$

As long as $\frac{2A\log\frac{2HN}{\delta}}{n_{\text{train}}} \leq 1$, the above is bounded by $2\sqrt{\frac{8A\log\frac{2HN}{\delta}}{n_{\text{train}}}}$. The choice of n_{train} follows from solving $2\sqrt{\frac{8A\log\frac{2N}{\delta}}{n_{\text{train}}}} = \phi$ for n_{train}, which indeed guarantees that $\frac{2A\log\frac{2HN}{\delta}}{n_{\text{train}}} \leq 1$ as $\phi \leq 4$. □

3.3 Olive-Generative: Multi-Agent

Improve Sample Complexity To achieve an improvement of $\tilde{h}$ in iteration complexity, we estimates $\hat{\mathcal{E}}(f, \pi_t, h_t)$ for any layer, as presented in Equation (8). That estimation correct h^t layers every phase. We want to achieve this improvement also in sample complexity. Naively, with sampling n_{train} trajectories and greedily trying to correct all the $h \in \mathcal{H}^t$, could potentially reduce the sample complexity of each iteration to $n_{\text{eval}} + n_{\text{train}}$ and spare the factor of $\tilde{h}$. The problem with this method is that Lemma 31 would yield an estimate for $\hat{\mathcal{E}}(f, \pi_{\text{UNIF}}, h_t)$, which is not what we needed. This estimation is bad because it does not necessarily fall within the sets of interest indicated in Lemma 26 and therefore we won't shrink the functions class in the confidence set.

Thus, in this subsection 3.3, We propose a method that performs the correction in place while overcoming the problem of wrong policy estimation and maintaining the current estimation accuracy. Specifically, we sample all n_{train} trajectories initially, and for each layer in $\mathcal{H}^t$, we introduce uniformly sampled n_{train} new states. Then perform a backward estimation of $\hat{\mathcal{E}}(f, \pi_t, h_t)$ for all $f \in \mathcal{F}_t$. This approach allows us to keep the previously sampled states from π_t while limiting the number of new samples to Hn_{train}, thereby maintaining our desired sample complexity.

We now turn to the main result, which guarantees that OLIVE-Generative-Multi-Agent PAC-learns *Contextual Decision Processes* with polynomial sample complexity.

Theorem 32. *Running OLIVE-Generative-Multi-Agent with the same parameters as like in Theorem 16, we get with probability at least* $1-\delta$, $OLIVE-Generative-Multi-Agent$ *halts and returns a policy* $\hat{\pi}$ *that satisfies* $V_{\hat{\pi}} \geq V^{\star}_{\mathcal{F}} - \epsilon$, *and the number of episodes required is at most*

$$\tilde{\mathcal{O}}\left(\frac{Hd\log(Hd\zeta/\epsilon)}{\tilde{h}} \times \left(1 + \frac{dH^2A}{|I|\epsilon^2}\log(N\zeta/\delta)\right)\right). \tag{21}$$

In this subsection as like previously, we get the iteration complexity and the optimality by the same way. But in this case using Generative Model as defined in 3, we improves our sample complexity. This improvement is effecting the amount of sample we need to re-estimate for correct a layer, as like in Line 9, and change our deviation bound for $\hat{\mathcal{E}}(f, \pi_t, h_j^t)$.

Lemma 33. *Using* n_{eval} *and* n_{train} *as like in Theorem 16, we get that for each agent* $A_i \in [I]$, *the sample complexity in each round* $t > 0$ *is:*

$$\tilde{\mathcal{O}}\left(\left(1 + \frac{dH^2A}{|I|\epsilon^2}\log(N\zeta/\delta)\right)\right). \tag{22}$$

Proof. Following Line 9 and Line 11 in $OLIVE-Generative-Multi-Agent$ we use $2\times\max\{1, \frac{n_{\text{train}}}{|I|}\} \leq 2 \times \left(1 + \frac{n_{\text{train}}}{|I|}\right)$, trajectories in each time step $t > 0$. And also, from the fact that $n_{\text{eval}} \leq n_{\text{train}}$, we get that $OLIVE - Generative - Multi - Agent$ sample complexity is $\leq \tilde{\mathcal{O}}\left(3 \times \left(1 + \frac{n_{\text{train}}}{|I|}\right)\right) = \tilde{\mathcal{O}}\left(\left(1 + \frac{dH^2A}{|I|\epsilon^2}\log(N\zeta/\delta)\right)\right)$. □

Algorithm 2 OLIVE-Generative-Multi-Agent$(\mathcal{F}, d, I, \zeta, \epsilon, \delta)$

Collect n_{est} trajectories with actions taken in an arbitrary manner; save initial contexts $\{x_1^{(i)}\}_{i=1}^{n_{\text{est}}}$.
Estimate the predicted value for each $f \in \mathcal{F}$: $\hat{V}_f = \frac{1}{n_{\text{est}}} \sum_{i=1}^{n_{\text{est}}} f(x_1^{(i)}, \pi_f(x_1^{(i)}))$.
$\mathcal{F}_0 \leftarrow \mathcal{F}$.
for $t = 1, 2, \ldots$ **do**
  Choose policy $f_t = \text{argmax}_{f \in \mathcal{F}_{t-1}} \hat{V}_f$, $\pi_t = \pi_{f_t}$.
  for Agent $A_i \in [I]$ **do**
    Collect $n = \lceil \frac{n_{\text{eval}}}{|I|} \rceil$ trajectories by following π_t.
    Add to data set: $\mathcal{D}^t_{eval} = \mathcal{D}^t_{h[A_i]} \cup \{(x_1^{(i)}, a_1^{(i)}, r_1^i, \ldots, x_H^{(i)}, a_H^{(i)}, r_H^{(i)})\}_{i=1}^n$.
    Collect $n = \lceil \frac{n_{\text{train}}}{|I|} \rceil$ trajectories by following π_t.
    Add to data set: $\mathcal{D}^t_{train} = \{(x_1^{(i)}, a_1^{(i)}, r_1^i, \ldots, x_H^{(i)}, a_H^{(i)}, r_H^{(i)})\}_{i=1}^n$.
    Collect another n tuples there $\hat{a}_h^{(i)}$ is drawn uniformly at random for $h = H, \ldots, 1$.
    Add to data set: $\mathcal{D}^t_h = \mathcal{D}^t_h \cup \{x_h^{(i)}, \hat{a}_h^{(i)}, \hat{r}_h^{(i)}, \hat{x}_{h+1}^{(i)}\}_{i=1}^{n_{\text{train}}}$, where $x_h^{(i)}$ sampled in 9 and store on $\mathcal{D}^t_{train}$.
  Estimate $\forall h \in [H]$,

$$\tilde{\mathcal{E}}(f_t, \pi_t, h) = \frac{1}{|\mathcal{D}^t_{eval}|} \sum_{i \in \mathcal{D}^t_{eval}} \left[f_t(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_t(x_{h+1}^{(i)}, a_{h+1}^{(i)}) \right]. \tag{19}$$

  if $(1/H) \sum_{h=1}^H |\tilde{\mathcal{E}}(f_t, \pi_t, h)| \leq \eta$ **then**
    Terminate and output $\pi_{out} = \pi_t$.
  else $\forall h \in [H]$ for which $|\tilde{\mathcal{E}}(f_t, \pi_t, h)| \geq \eta$: {One is guaranteed to exist }
    Estimate

$$\hat{\mathcal{E}}(f, \pi_t, h) = \frac{1}{|\mathcal{D}^t_h|} \sum_{i \in \mathcal{D}^t_h} \frac{\mathbf{1}[\hat{a}_h^{(i)} = \pi_f(x_h^{(i)})]}{1/A} \Big(f(x_h^{(i)}, \hat{a}_h^{(i)}) - \hat{r}_h^{(i)} - f(\hat{x}_{h+1}^{(i)}, \pi_f(\hat{x}_{h+1}^{(i)})) \Big). \tag{20}$$

  Learn

$$\mathcal{F}_t = \left\{ f \in \bigcap_{h \in [H]} \left\{ f \in \mathcal{F}_{t-1} : \left| \hat{\mathcal{E}}(f, \pi_t, h) \right| \leq \phi \right\} \right\}.$$

Lemma 34 (Deviation Bound for $\hat{\mathcal{E}}(f, \pi_t, h_j^t)$)**.** *With probability at least $1-\delta$, for any fixed π_t and for any $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$, we have*

$$|\hat{\mathcal{E}}(f, \pi_t, h_j^t) - \mathcal{E}(f, \pi_t, h_j^t)| \le \sqrt{\frac{8A \log \frac{2NH}{\delta}}{n_{train}}} + \frac{2A \log \frac{2NH}{\delta}}{n_{train}}$$

holds for all $f \in \mathcal{F}$ and $h \in [H]$ simultaneously. Hence, for any $n_{train} \ge \frac{32A}{\phi^2} \log \frac{2NH}{\delta}$ and $\phi \le 4$, with probability at least $1-\delta$ we have $|\hat{\mathcal{E}}(f, \pi_t, h) - \mathcal{E}(f, \pi_t, h)| \le \phi$.

Proof. We first show that $\hat{\mathcal{E}}(f, \pi_t, h)$ is an average of i.i.d. random variables with mean $\mathcal{E}(f, \pi_t, h_j^t)$. We will show that for fix $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$ the statement in Equation 18 hold, and with union bound we conclude it for any $h \in [H]$, s.t. $|\tilde{\mathcal{E}}(f, \pi_t, h)| > \eta$. In addition, we use μ as a shorthand for the distribution over trajectories induced by $a_1, \ldots, a_{h-1} \sim \pi_t, \hat{a}_h \sim \text{unif}(\mathcal{A})$, which is the distribution of data used to estimate $\hat{\mathcal{E}}(f, \pi_t, h)$. We sample $\hat{a}_h$ using Generative model after we draw at first n trajectories following π_t. thus, we are indeed estimating $\hat{\mathcal{E}}(\cdot, \pi_t, h)$ as required. On the other hand, let μ' denote the distribution over trajectories induced by $a_1, \ldots, a_{h-1} \sim \pi_t, \hat{a}_h \sim \pi_f$. The importance weight used in Eq. (20) essentially converts the distribution from μ to μ', hence the expected value of $\hat{\mathcal{E}}(f, \pi_t, h)$ can be written as

$$\begin{aligned}
&\mathbb{E}_\mu \left[A\mathbf{1}[a_h = \pi_f(x_h)] \left(f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \right) \right] \\
&= \mathbb{E}_{\mu'} \left[f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \right] = \mathcal{E}(f, \pi_t, h_t).
\end{aligned}$$

Now, we apply Bernstein's inequality. We first analyze the 2nd-moment of the random variable. Defining $y(x_h, a_h, r_h, x_{h+1}) = f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \in [-2, 1]$, the 2nd-moment is

$$\begin{aligned}
&\mathbb{E}_\mu \left[\left(A\mathbf{1}[a_h = \pi_f(x_h)] y(x_h, a_h, r_h, x_{h+1}) \right)^2 \right] \\
&= \mathbb{P}_\mu[a_h = \pi_f(x_h)] \cdot \mathbb{E}_\mu \left[\left(A y(x_h, a_h, r_h, x_{h+1}) \right)^2 \,\middle|\, a_h = \pi_f(x_h) \right] + \mathbb{P}_\mu[a_h \ne \pi_f(x_h)] \cdot 0 \\
&\le \frac{1}{A} \mathbb{E}_\mu \left[A^2 \cdot 4 \,\middle|\, a_h = \pi_f(x_h) \right] = 4A.
\end{aligned}$$

Next we check the range of the centered random variable. The uncentered variable lies in $[-2A, A]$, and the expected value is in $[-2, 1]$, so the centered variable lies in $[-2A-1, A+2] \subseteq [-3A, 3A]$. Applying Bernstein's inequality, we have with probability at least $1-\delta$,

$$\begin{aligned}
|\hat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| &\le \sqrt{\frac{2 \operatorname{Var}\left[A\mathbf{1}[a_h = \pi_f(x_h)] y(x_h, a_h, r_h, x_{h+1}) \right] \log \frac{2HN}{\delta}}{n_{\text{train}}}} + \frac{6A \log \frac{2HN}{\delta}}{3n_{\text{train}}} \\
&\le \sqrt{\frac{8A \log \frac{2HN}{\delta}}{n_{\text{train}}}} + \frac{2A \log \frac{2HN}{\delta}}{n_{\text{train}}}. \qquad \text{(variance is bounded by 2nd-moment)}
\end{aligned}$$

As long as $\frac{2A \log \frac{2HN}{\delta}}{n_{\text{train}}} \le 1$, the above is bounded by $2\sqrt{\frac{8A \log \frac{2HN}{\delta}}{n_{\text{train}}}}$. The choice of n_{train} follows from solving $2\sqrt{\frac{8A \log \frac{2N}{\delta}}{n_{\text{train}}}} = \phi$ for n_{train}, which indeed guarantees that $\frac{2A \log \frac{2HN}{\delta}}{n_{\text{train}}} \le 1$ as $\phi \le 4$. □

4 The Many Faces of Eluder Dimension

This section explores various definitions of the Eluder Dimension, originally from Russo and Van Roy (2013), and their respective utilities in different contexts of reinforcement learning theory. In general, Eluder Dimensions quantify the informational independence among decisions, influencing the balance between exploration and exploitation. From now on we will work with the next definition for value-based function class.

Definition 35 (Function Approximation Representation)**.** *For finite horizon setups, 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. A function $f \in \mathcal{F}$ can be formalized as $f = (f_1, \ldots, f_H)$. Since no reward is collected in the $(H+1)^{th}$ step, we always set $f_{H+1} = 0$.*

At the core of these definitions lies the concept of *ϵ-independence between points*.

Definition 36 (ϵ-independence between points)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $z, x_1, x_2, \ldots, x_n \in \mathcal{X}$. 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.

4.1 Standard Definition of Eluder Dimension

Definition 37 (Eluder dimension)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$. The Eluder dimension $d = \dim_{\mathrm{E}}(\mathcal{G}, \epsilon)$ 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.

4.2 Distributional Eluder Dimension

The Distributional Eluder Dimension measures the uncertainty of functions whose values are random variables. In this context, the concept of ϵ-independence between points is adapted to accommodate the distributional nature of the functions. This definitions shows in Jin et al. (2021a).

Definition 38 (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 $\sqrt{\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2} \leq \epsilon$, but $|\mathbb{E}_\nu[g]| > \epsilon$.*

Definition 39 (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 $\epsilon' \geq \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \ldots, \rho_{i-1}\}$ for all $i \in [n]$.*

Definition 38 and Definition 39 generalize Definition 36 and Definition 37 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 $\Pi_{dirac} = \{\delta_x(\cdot) \mid x \in \mathcal{X}\}$ where $\delta_x(\cdot)$ is the dirac measure centered at x, then $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) = \dim_{\mathrm{DE}}(\Delta(\mathcal{G}), \Pi_{dirac}, \epsilon)$ where $\Delta(\mathcal{G}) = \{g_1 - g_2 : \ g_1, g_2 \in \mathcal{G}\}$.
Therefore we have this relation between the dimensions:

$$\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) \geq \min_{\Pi} \dim_{\mathrm{DE}}(\Delta(\mathcal{G}), \Pi, \epsilon). \tag{23}$$

Bellman Eluder Dimension

Definition 40 (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*

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

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. The GOLF algorithm (Jin et al., 2021a) focused on the following two specific choices.

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

4.3 Functional Eluder Dimension

We introduce another complexity measure, shown in Chen et al. (2022). The *functional eluder dimension*, which generalizes the concept of the Eluder dimension and its successors, capturing a broader class of functions. Here, we present its definition:

Definition 41 (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 longest sequence $f_1, \ldots, f_n \in \mathcal{F}$ satisfying for all $t \leq n$, exists $\epsilon_t \geq \epsilon$, such that there exists $g \in \mathcal{F}$ holding $\sqrt{\sum_{i=1}^{t-1} (G(g, f_i))^2} \leq \epsilon_t$ while $|G(g, f_t)| > \epsilon_t$. Function G is dubbed as the* coupling function.

When the coupling function is simply $G(f_1, f_2) = f_1 - f_2$, we have $\dim_{\mathrm{DE}}(\Delta(\mathcal{F}), \Pi_{dirac}, \epsilon) = \dim_{\mathrm{FE}}(\mathcal{F}, \Delta, \epsilon)$. If the coupling function is chosen as the expected Bellman error $G_h^{\Pi}(g, f) := \mathbb{E}_{\pi_{h,f} \sim \Pi}(Q_{h,g} - \mathcal{T}_h Q_{g,h+1})$, where $\mathcal{T}_h$ denotes the Bellman operator, and we recover the definition of BE dimension (Jin et al., 2021a), i.e., $\dim_{\mathrm{FE}}(\mathcal{F}, G, \epsilon) = \dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon)$.

4.4 Generalized Eluder Dimension

The previous definitions raise an important question: How do we determine if we have explored enough or if more exploration is needed? The generalized Eluder dimension, defined in Agarwal et al. (2022); Zhao et al. (2023), addresses this by extending previous definitions to weighted regression settings.

Definition 42 (Generalized Eluder dimension)**.** *Let $\lambda > 0$, a sequence of state-action pairs $Z = \{z_i\}_{i\in[T]}$ and $\sigma = \{\sigma_i\}_{i\in[T]}$ be given. The generalized Eluder dimension of a function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, L]$ is given by $dim_{\alpha,\lambda}(\mathcal{F}) := \sup_{Z,\sigma:|Z|=T,\sigma\geq\alpha} dim(\mathcal{F}, Z, \sigma)$, where*

$$dim(\mathcal{F}, Z, \sigma) := \sum_{i=1}^{T} \min\left(1, \frac{1}{\sigma_i^2} D^2_{\mathcal{F}}(z_i; z_{i-1}, \sigma_{i-1})\right),$$

$$\text{where } D^2_{\mathcal{F}}(z; z', \sigma') := \sup_{f_1,f_2\in\mathcal{F}} \frac{(f_1(z) - f_2(z))^2}{\sup_{s\in[t-1]} \frac{1}{\sigma_s} (f_1(z_s) - f_2(z_s))^2 + \lambda}.$$

Relation to standard Eluder dimension When $\sigma \equiv 1$, we have:

$$max_{Z:|Z|=T} \dim(\mathcal{F}, Z, \mathbf{1}) \leq \dim_E(\mathcal{F}, \sqrt{\lambda/T}) + 1, \tag{24}$$

where $\dim_E(\mathcal{F}, \epsilon)$ is the standard Eluder dimension of $\mathcal{F}$ as defined in Russo and Van Roy (2013). We refer to the paper for full proof.

4.5 Summary of Benefits

The **Standard Eluder Dimension** measures informational independence among decisions in reinforcement learning, aiding the balance between exploration and exploitation. It provides a foundational tool for quantifying information gain from exploring new actions and states. Building on this, the **Bellman Eluder Dimension** incorporates the Bellman operator to measure the complexity of value functions in MDPs, enhancing sample efficiency evaluation. The **Functional Eluder Dimension** generalizes these concepts further, capturing a broader class of functions and measuring the complexity of hypothesis classes in both model-free and model-based RL, thus offering a more comprehensive framework. Lastly, the **Generalized Eluder Dimension** define to base on these previous definitions and generalize them by including weighted regression settings, adding the benefit of quantifying exploration sufficiency, and providing practical guidance on the need for further exploration. Each successive definition refines and extends its predecessor, broadening the scope of applicable function classes and enhancing the measure's utility in assessing exploration strategies in reinforcement learning.

5 Expected Eluder Dimension

5.1 Introduction

In this section we will revisit the problem of *Stochastic Contextual Bandits*. Many modern applications involve the presence of additional side information, or context, that impacts the environment. A naive approach to handling the context is to extend the state space of the environment to include it. However, this method increases the complexity of learning and policy representation.

In the Stochastic Contextual Bandits setting, a *contextual model* predicts the expected reward for different actions based on the context available at the time of decision-making.

Definition 43 (Contextual Models)**.** *A contextual model estimates the expected reward R for each action A, given the observed context X. The agent's goal is to select an action A from a set of possible actions, conditioned on the context X, to maximize the expected reward.*

Mathematically, this is represented as:

$$A^* = \arg\max_A \mathbb{E}[R|X, A]$$

where X is the context, A is the action, and R is the reward.

Types of Contextual Models

1. **Linear Models**:

$$\mathbb{E}[R|X, A] = \theta_A^\top X$$

where θ_A is a parameter vector for action A, and X is the context.
2. **Non-linear Models**: These can include neural networks, decision trees, or other non-linear mappings from context to reward:

$$\mathbb{E}[R|X, A] = f(X, A)$$

where $f(X, A)$ is a non-linear function that predicts the reward based on context X and action A.

Components The model is time dependent, and the key components of a contextual model are:

- **Context** $\mathcal{X}$: A set of contexts that describe the environment, X_t descibe it at time t.
- **Action** $\mathcal{A}$: set of state-action pairs $\mathcal{A} := \{(x, a) : x \in \mathcal{X}, a \in \mathcal{A}(x)\}$, and the set of available actions to be $A_t = \{(X_t, a) : a \in \mathcal{A}(X_t)\}$.
- **Reward** R: The outcome (or feedback) from taking action A in context X.
- **Model** $f(X, A)$: A function that predicts the expected reward at time t is given by $f(a, X_t)$.

Eluder Dimension in Contextual Models

One class of functions highly related to this section is the function class of **Eluder dimension** (see Definition 37). The role of the Eluder dimension in this work is to avoid direct dependence on the size of the context space, which could be potentially large, while also maintaining computational efficiency.

Our objective is to extend the traditional framework of the Eluder dimension to incorporate probabilistic elements that reflect the reality of sampling state-action pairs from a distribution. Specifically, we aim to introduce a nuanced aspect to the Eluder dimension by considering not only the structure of the function class, but also the context distribution.

We define the *Expected Eluder Dimension* to provide a more accurate representation of the Eluder Dimension in stochastic environments, where not all state-action pairs are equally likely or relevant.

Definition 44 (Eluder Predicate)**.** *Let $\mathcal{G}$ be a function class and Z a sequence of n-context and actions. For $I \subseteq [n]$, a subsequence of Z, we define a predicate $EP_{\mathcal{G}}(I, j, \epsilon)$, which returns 1 if (x_j, a_j) is ϵ-independent from $(x_{i_1}, a_{i_1}), \ldots, (x_{i_{|I|}}, a_{i_{|I|}})$, respect to the function class $\mathcal{G}$.*

Definition 45 (Expected Eluder dimension)**.** *Let $\epsilon > 0$, and $\mathcal{D}$ a context distribution. Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$. Let W be a random variable of a sequence of n contexts drawn from $\mathcal{D}$. Given a sequence $Z = \{(x_1, a_1), \ldots, (x_n, a_n)\}$ and $X(Z) = (x_1, \ldots, x_n)$. The Expected Eluder dimension $\tilde{d} = \dim_{\mathbb{E}}(\mathcal{G}, \epsilon, \mathcal{D}^n)$ is*

$$\mathbb{E}_{W \sim \mathcal{D}^n}\big[\max_{Z:X(Z)=W} \max_{I \subseteq [n]} EP_{\mathcal{G}}(Z, I, \epsilon)\big], \tag{25}$$

where $EP_{\mathcal{G}}(Z, I, \epsilon) := \sum_{j \in I} EP_{\mathcal{G}}(\{z_{i_1}, \ldots, z_{i_{j-1}}\}, j, \epsilon)$, and $\mathcal{D}^n = \mathcal{D} \overset{n \text{ times}}{\times \cdots \times} \mathcal{D}$

5.2 Effective Dimension Reduce

The main application of the Eluder dimension is to design algorithms and prove regret guarantees for contextual bandits and reinforcement learning. When the Eluder dimension first appeared, it was proposed to handle Linear Bandits, as discussed in (Russo and Van Roy, 2013). They assume that the context follows a time-dependent model that samples from a function class $\mathcal{F}$, which holds the Eluder condition and has a dimension d. We focus specifically on contextual bandits similar to (Russo and Van Roy, 2013), and (Foster et al., 2020). The Eluder dimension is used to measure functions that take both contexts and actions as parameters, and the definition of the Eluder dimension is tailored to the adversarial setup.

In contrast, we return to stochastic settings and challenge this assumption. In our framework, the context is sampled from a distribution $\mathcal{D}$, where $\mathcal{F}$ has an Eluder dimension d, but the key difference is that the context is sampled from $\mathcal{D}$, leading to an Expected Eluder dimension $\tilde{d} \leq d$. This distinction allows us to model real-world scenarios more effectively, as it accounts for the stochastic nature of the context distribution.

5.2.1 Applications

Scenario: Linear Functions with Distribution of One Dominant Context

Consider a linear function class $\mathcal{F}$, where each function f maps features in $\mathbb{R}^d$ to real numbers ($\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$). For action set $\mathcal{A}$, and context class $\mathcal{X}$, we assume that $\exists \phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, such that the input vector for $f \in \mathcal{F}$ is $\phi(x, a)$.

Context and Setup

- **Actions**: Assume $\mathcal{A} = \mathcal{B}(d, 1)$ is the unit ball in $\mathbb{R}^d$.
- **Probability Model**: For the standard base in $\mathbb{R}^d$, $B = \{e_1, \ldots, e_d\}$, let the distribution $\mathcal{D}$ over $\mathcal{X}$, be define as follows, $\exists \delta > 0, i \in [d]$ s.t. $\forall j \neq i \in [d]$ for any context x hold :

$$x = \begin{cases} e_i & w.p.\ \ 1 - \delta \\ e_j & w.p.\ \ \delta/d \end{cases}.$$

- **Function**: We have a domain $\Theta \subseteq \mathcal{B}(d, 1)$. Each function $f_\theta \in \mathcal{F}$ is index by some $\theta \in \Theta$. The definition of f_θ is as follows, $f_\theta(x, a) = \langle \theta, \phi(x, a) \rangle$.
  And let $\phi(x, a) = x \odot a$ element-wise product, i.e. $\phi(x, a)_k = x_k \cdot a_k$.

Eluder Dimension When ignoring the specific distribution, all entries have an impact on the value of the functions in $\mathcal{F}$. Therefore, the Eluder dimension in this scenario considers the entire function, context, and action class, denoted as $\mathcal{G} = \{\mathcal{F}, \mathcal{X}, \mathcal{A}\}$. In general, this dimension can be expressed as $d\log(1/\epsilon)$. However, the number of contexts inherently limits the dimension. Since there are only d distinct contexts, the Eluder dimension is trivially bounded, giving us $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) = d$.

Expected Eluder Dimension Despite the general result that the Eluder dimension is bounded by d, it does not necessarily scale significantly with d itself. Instead, the Eluder dimension depends on the actual number of vectors or contexts that we encounter in practice. As the effective dimension is influenced by the specific set of vectors used, it can often be much smaller than the theoretical bound, especially when the number of relevant vectors is limited. This toy example highlights that the complexity of the problem is driven more by the structure and quantity of the vectors rather than just the raw dimensionality d

Claim 46. *The Expected Eluder dimension of this problem is δT.*

Corollary 47. *For $\delta = \mathcal{O}(\frac{1}{T})$, we $\dim_{\mathbb{E}}(\mathcal{F}, \epsilon, \mathcal{D}^T) = \mathcal{O}(1)$.*

Proof. . Let $\delta > 0$, and $\mathcal{D}$ the distribution as defined above. Let x be context sampled from $\mathcal{D}$.

$$\forall f_\theta \in \mathcal{F}, a \in \mathcal{A}, f_\theta(x, a) = \langle \theta, x \odot a \rangle = \begin{cases} \langle \theta, e_i \cdot a \rangle & w.p.\ 1 - \delta \\ \langle \theta, e_j \cdot a \rangle & w.p.\ \delta/d \text{ for } j \in [d] \end{cases}$$

Denote $Z = (w, a) : w \sim D^T, a \in \mathcal{A}^T$, where D^T defined in (25), and $\mathcal{A}^T$ defined in the same way. For any phase $t \in [T]$ and $\theta_1, \theta_2 \in \mathcal{B}(d, 1)$:

$$\begin{aligned} \mathbb{E}_{w \sim \mathcal{D}^T} \left[\max_{A^T} f_{\theta_1}(w^t, a^t) - f_{\theta_2}(w^t, a^t) \right] &\leq \mathbb{E}_{w \sim \mathcal{D}^T} \left[\sum_{i=1}^{d} \mathbf{1}\{e_i \in w\} \right] = \sum_{i=1}^{d} \mathbb{E}_{w \sim \mathcal{D}^T} \left[\mathbf{1}\{e_i \in w\} \right] \\ &\leq \sum_{i=1}^{d} \min\{T\mathbb{P}(e_i), 1\} \leq \min\{1 + T\delta, d\}. \end{aligned}$$

Therefore, $\dim_{\mathbb{E}}(\mathcal{F}, \epsilon, \mathcal{D}^T) \leq \min\{\delta T + 1, d\}$. □

Scenario: Linear Functions with Exponentially Decaying Distribution

Consider a scenario where the dimensionality d is very large, such that for any possible number of samples T, the ratio $T/d < 1$. This setup reflects a situation where the number of samples is significantly smaller than the dimension of the space, which leads to interesting challenges in learning and generalization.

Consider a linear function class $\mathcal{F}$, where each function f maps features in $\mathbb{R}^d$ to real numbers ($\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$). For action set $\mathcal{A}$, and context class $\mathcal{X}$, we assume that $\exists \phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, such that the input vector for $f \in \mathcal{F}$ is $\phi(x, a)$.

Context and Setup

- **Actions**: Assume $\mathcal{A} = \mathcal{B}(d, 1)$ is the unit ball in $\mathbb{R}^d$.

- **Probability Model**: For the standard base in $\mathbb{R}^d$, $B = \{e_1, \ldots, e_d\}$, let the distribution $\mathcal{D}$ over $\mathcal{X}$, be define as follows, $\forall i \in [d]$ any context x hold :

$$x = e_i \;\; \text{w.p} \;\; 2^{-i}.$$

- **Function**: We have a domain $\Theta \subseteq \mathcal{B}(d, 1)$. Each function $f_\theta \in \mathcal{F}$ is index by some $\theta \in \Theta$. The definition of f_θ is as follows, $f_\theta(x, a) = \langle \theta, \phi(x, a) \rangle$.
  And let $\phi(x, a) = x \odot a$ element-wise product, i.e. $\phi(x, a)_k = x_k \cdot a_k$.

Eluder Dimension There are two natural bounds on the Eluder dimension in this setup: the number of distinct contexts and the number of samples. Since the number of distinct contexts is d and the number of samples is at most T, the Eluder dimension is bounded by both, resulting in $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) = \min(d, T)$.

Expected Eluder Dimension

Claim 48. *The Expected Eluder dimension of this problem is* $\dim_{\mathbb{E}}(\mathcal{F}, \epsilon, \mathcal{D}^T) = \log(T)$.

Proof. Given a distribution D over a sequence $x \sim D$, where each x takes the value e_i with probability 2^{-i}, we want to compute the expected number of distinct elements observed in a sequence of T samples.

Denote $Z = (w, a) : w \sim D^T, a \in \mathcal{A}^T$. Let $I_i = \mathbf{1}\{e_i \in w\}$ be an indicator random variable that equals 1 if the element e_i appears at least once in the T samples, and 0 otherwise. The expected number of distinct elements seen in the T samples is given by:

$$\mathbb{E}_{w \sim \mathcal{D}^T}\left[\sum_{i=1}^{\infty} I_i\right] = \sum_{i=1}^{\infty} \mathbb{E}_{w \sim \mathcal{D}^T}[I_i].$$

Since I_i is an indicator variable, $\mathbb{E}_{w \sim \mathcal{D}}[I_i] = \mathbb{P}(I_i = 1)$, which is the probability that e_i appears at least once in the T samples.

The probability that e_i does not appear in a single sample is $1 - 2^{-i}$. Thus, the probability that e_i does not appear in all T samples is $(1 - 2^{-i})^T$. Consequently, the probability that e_i appears at least once in the T samples is:

$$\mathbb{P}(I_i = 1) = 1 - (1 - 2^{-i})^T.$$

Therefore, the expected number of distinct elements is:

$$\mathbb{E}_{w \sim \mathcal{D}^T}\left[\sum_{i=1}^{\infty} I_i\right] = \sum_{i=1}^{\infty} \left[1 - (1 - 2^{-i})^T\right].$$

For large T, if $2^{-i}T$ is small, we can use the approximation $(1 - 2^{-i})^T \approx e^{-T2^{-i}}$. Therefore, the probability that e_i appears at least once becomes:

$$1 - (1 - 2^{-i})^T \approx 1 - e^{-T2^{-i}}.$$

Substituting this into the sum, we obtain:

$$\mathbb{E}_{w \sim \mathcal{D}^T}\left[\sum_{i=1}^{\infty} I_i\right] \approx \sum_{i=1}^{\infty} \left(1 - e^{-T2^{-i}}\right).$$

- For small i where $2^{-i}T$ is large, $1 - e^{-T2^{-i}} \approx 1$. Thus, these terms contribute approximately 1 to the sum.
- For large i where $2^{-i}T$ is small, $1 - e^{-T2^{-i}} \approx T2^{-i}$, which decays exponentially with i.

The sum can be split into two parts based on an index k such that $2^{-k}T \approx 1$. For $i \leq k$, $1 - e^{-T2^{-i}} \approx 1$. For $i > k$, $1 - e^{-T2^{-i}} \approx T2^{-i}$. Thus,

$$\mathbb{E}_{w \sim \mathcal{D}^T}\left[\sum_{i=1}^{\infty} I_i\right] = \mathcal{O}\left(k + \sum_{i=k+1}^{\infty} T2^{-i}\right).$$

Since $\sum_{i=k+1}^{\infty} T2^{-i} \approx T2^{-k}$, and by choosing $k \approx \log(T)$, we get $T2^{-k} \approx 1$. Therefore, the expected number of distinct elements is approximately:

$$\mathbb{E}_{w \sim \mathcal{D}^T}\left[\sum_{i=1}^{\infty} I_i\right] = \mathcal{O}\left(\log(T)\right).$$

Thus, the Expected Eluder dimension is the expected number of different x values observed in T samples is approximately $\log(T)$. This result indicates that the number of distinct elements grows logarithmically with the number of samples T. □

Scenario: Sparse Linear Regression with Gaussian Features

Consider a linear regression model in a high-dimensional setting where only a sparse subset of features is non-zero. Each active feature is drawn from a Gaussian distribution.

Context and Setup

- **Action Set**: Let $\mathcal{A} = \mathcal{B}(d, 1)$ denote the unit ball in $\mathbb{R}^d$.
- **Feature Representation**: Define the context class $\mathcal{X} \subseteq \mathbb{R}^d$. Each function $f \in \mathcal{F}$ maps contexts in $\mathbb{R}^d$ to real numbers.
- **Sparsity of Features**: At any given instance, only same r features ($r \ll d$) are non-zero.
- **Distribution of Active Features**: The non-zero features are sampled from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

Eluder Dimension Since the context vectors can span the entire space $\mathcal{B}(d, 1)$, the Eluder dimension is bounded by $d\log(1/\epsilon)$.

Expected Eluder Dimension Given that only r features are likely to be non-zero with higher probability, the effective dimension reduces to $r\log(1/\epsilon)$.

Thus, the complexity of the learning problem depends primarily on r rather than the full dimensionality d. For instance, if $r = 1$ and $d = 100$, the expected Eluder dimension simplifies to 1, analogous to a classic stochastic bandit problem.

Scenario: Sparse Linear Regression with Mixture of k Gaussians

Consider a linear regression model in a high-dimensional setting, where the non-zero features follow a mixture of k Gaussian distributions. Each mixture component is defined with specific sparsity properties.

Context and Setup

- **Action Set**: Let $\mathcal{A} = \mathcal{B}(d, 1)$ denote the unit ball in $\mathbb{R}^d$.
- **Feature Representation**: Define the context class $\mathcal{X} \subseteq \mathbb{R}^d$. Each function $f \in \mathcal{F}$ maps contexts in $\mathbb{R}^d$ to real numbers.
- **Active Features**: The non-zero features follow a mixture of k Gaussian distributions, $\mathcal{G}(\mu_i, \Sigma_i)$ for $i \in [k]$. For each Gaussian component i, the covariance matrix Σ_i has $\text{Rank}(\Sigma_i) \leq r_i$ where $r_i \ll d$ for all $i \in [k]$.

Eluder Dimension Since the context vectors can span the full space $\mathcal{B}(d, 1)$, the Eluder dimension is initially bounded by $d \log(1/\epsilon)$.

Expected Eluder Dimension When the non-zero features belong to a mixture of k Gaussians with low-rank covariance matrices, the effective dimension is bounded by $\sum_{i=1}^{k} r_i \log(1/\epsilon)$. This bound captures the sparsity within each Gaussian component and the overall mixture structure.

For a mixture of k-Gaussians $\{\mathcal{G}(w_i, \mu_i, \Sigma_i)\}_{i=1}^{k}$, where w_i are the component weights, we identify non-zero features as those that, with high probability, exceed a threshold ϵ within T samples. Formally, for any feature x, we define a condition based on w, μ, Σ, and T to determine if its absolute value is likely to be greater than ϵ. For each Gaussian i, let r_i represent the maximum number of independent features that satisfy this condition.

Claim 49.

$$\dim_{\mathbb{E}}(\mathcal{F}, \epsilon, \mathcal{G}_k^T) \leq \sum_{i=1}^{k} r_i \log(1/\epsilon).$$

Remark 50. *Sampling from a mixture of Gaussians is equivalent to selecting one Gaussian component according to the mixture's weights and then sampling from that component. This property allows the analysis of the mixture model by considering each Gaussian's contribution separately and summing their effects.*

Proof. Let $\mathcal{F}$ be a linear function class as defined previously, and let $\mathcal{D} = \mathcal{G}_k$ be the mixture of k Gaussians. Assume, for simplicity, that all Gaussians have $\{e_j\}_{j=1}^{d}$ as eigenvectors for its covariance matrix. Let $i \in [k]$ denote the Gaussian component with weight w_i. For each feature $j \in [d]$, let e_j be an eigenvector of Σ_i with eigenvalue v_j.

Let $j \in [d]$ such that $v_j \leq \delta$ and $|\mu_i[j]| = \varepsilon_j \leq \epsilon$, if $\sqrt{2\delta_j \ln(2w_i T)} + \varepsilon_j < \epsilon$, we will show that the expected number of steps, in a sequence of T samples, where the j-th coordinate exceeds ϵ is less than 1. Therefore, for any feature $x \sim \mathcal{G}_k^T$, such that does not holds the condition above, for any pair of functions $f_{\theta_1}(\cdot, \cdot), f_{\theta_2}(\cdot, \cdot) \in \mathcal{F}$, we have

$$\mathbb{E}_{x \sim \mathcal{G}_k^T} \max_A \{f_{\theta_1}(x, a) - f_{\theta_2}(x, a)\} \leq \epsilon.$$

Therefore, covering the sub-space that the k-Gaussians are spans by those features who exceed the threshold ϵ within T samples, i.e. for each Gaussian this amount is

$$r_i = \left| \left\{ j : \sqrt{2\delta_j \ln(2w_i T)} + \varepsilon_j > \epsilon \right\} \right|, \tag{26}$$

This way measures the effective dimensionality influenced by the mean, the eigenvalues and the Gaussian mixture weights. Thus, implying that

$$\dim_{\mathbb{E}}(\mathcal{F}, \epsilon, \mathcal{G}_k^T) \leq \sum_{i=1}^{k} r_i \log(1/\epsilon).$$

Now we will see that the condition in Eq (26) indeed says that. Given e_j as an eigenvector of the covariance matrix Σ_i with eigenvalue δ, the probability that one of the T independent samples from a Gaussian component with weight w_i deviates from the mean by more than ϵ can be bounded using Gaussian tail bounds. For a single sample:

$$\mathbb{P}(|x - \mu_i[j]| > \epsilon - \varepsilon) \leq 2 \exp\left(-\frac{(\epsilon - \varepsilon)^2}{2\delta} \right).$$

Including the component weight w_i, this probability becomes:

$$\mathbb{P}(|x - \mathbb{E}[x]| > \epsilon) \leq 2w_i \exp\left(-\frac{(\epsilon - \varepsilon)^2}{2\delta} \right).$$

Using the union bound for T samples, the probability that at least one sample exceeds ϵ is:

$$\mathbb{P}\left(\bigcup_{i=1}^{T} |x_i| > \epsilon - \varepsilon \right) \leq 2w_i T \exp\left(-\frac{(\epsilon - \varepsilon)^2}{2\delta} \right).$$

We want this probability to be less than 1:

$$2w_i T \exp\left(-\frac{(\epsilon - \varepsilon)^2}{2\delta} \right) < 1.$$

Taking the natural logarithm of both sides:

$$\ln(2w_i T) - \frac{(\epsilon - \varepsilon)^2}{2\delta} < 0.$$

Rearranging the inequality:

$$\frac{(\epsilon - \varepsilon)^2}{2\delta} > \ln(2w_i T),$$

which gives us

$$\sqrt{2\delta \ln(2w_i T)} + \varepsilon < \epsilon$$

Therefore, the probability that at least one of the T samples deviates from the mean by more than ϵ is less than 1 if $\delta < \frac{(\epsilon-\varepsilon)^2}{2\ln(2w_i T)}$. □

This result implies that the complexity of the learning problem is primarily influenced by the total rank of the mixture components, $\sum_{i=1}^{k} r_i$, rather than the full dimensionality d.

Implications for Learning The reduced expected Eluder dimension allows the learning algorithm to focus on the relevant subspaces defined by the mixture of Gaussian components. This focus accelerates learning, improves data efficiency, and reduces computational costs, especially in cases where the mixture components have significantly lower ranks compared to the overall dimensionality d.

5.2.2 Practical Implications

Personalized Medicine Consider the application of this model in personalized medicine, where predictive models are used to tailor treatments based on genetic information. Typically, only a small subset of genetic markers (features) might be relevant for predicting the efficacy of a specific treatment, despite the presence of thousands of possible genetic markers. Efficiently identifying and focusing on these relevant markers using a model with a reduced Expected Eluder Dimension can significantly improve the speed and accuracy of personalized treatment plans.

Predictive Financial Modeling In finance, predictive models for stock returns can focus on a small subset of economic indicators and financial metrics, streamlining the analysis and improving forecast accuracy by concentrating on features with the highest predictive power.

Recommendation Systems In recommendation systems, leveraging a sparse feature model allows for efficiently predicting user preferences by concentrating on a small subset of user and item features that have the highest predictive power. This approach enhances the accuracy and efficiency of recommendations, particularly in systems with vast numbers of users and items.

In practical terms, if you are developing or analyzing algorithms for contextual bandits, considering the Eluder dimension can lead to more robust algorithms that handle the exploration-exploitation trade-off more effectively. It allows for a more nuanced understanding of how algorithm performance scales with the complexity of the environment and the diversity of contexts encountered. A more precise estimation of the Eluder dimension helps optimize our exploration strategy, potentially leading to an improved balance on the bias-variance trade-off. By accurately determining the necessary level of exploration, we can avoid overfitting and underfitting, enhancing the overall performance of learning algorithms.

5.3 The Algorithm

The LinUCB Algorithm for Contextual Bandits The Linear UCB (Upper Confidence Bound) algorithm is a popular method for solving contextual bandit problems with linear reward functions. It extends the classical UCB approach to contexts, balancing exploration and exploitation by maintaining uncertainty estimates for each action's expected reward. LinUCB assumes that the expected reward for each action is a linear function of the context.

Algorithm Outline

Algorithm 3 Linear UCB

Initialize: Select d linearly independent actions.

1: **for** $t = 1, 2, \ldots$ **do**

2: **Update Statistics:**

$$\hat{\theta}_t \leftarrow \min_{\theta} \sum_{k=1}^{t-1} \|f_\theta(a_k, x_k) - r_k\|_2$$

$$\Phi_t \leftarrow \sum_{k=1}^{t-1} \phi(a_k, x_k)\phi(a_k, x_k)^\top$$

$$\Theta_t \leftarrow \left\{ \rho : \left\| \rho - \hat{\theta}_t \right\|_{\Phi_t} \leq \beta \sqrt{d \log t} \right\}$$

3: **Select Action:**

1. Observe the context $x_t \in \mathbb{R}^d$.
2. Select the action a_t that maximizes $\hat{r}_a(t)$:

$$a_t \in \arg\max_{a \in A} \left\{ \max_{\rho \in \Theta_t} \langle \phi(a, x_t), \rho \rangle \right\}$$

3. Observe the reward $r_t(a_t)$ for the chosen action a_t.

LinUCB uses a confidence-bound approach where actions with higher uncertainty in their reward estimates are more likely to be selected. The parameter α controls how much weight is given to this uncertainty, balancing exploration (trying less certain actions) and exploitation (choosing actions with high expected rewards).

The linear assumption allows the algorithm to efficiently update the model after observing each reward, making it computationally efficient even in high-dimensional spaces, especially when combined with dimension reduction techniques.

5.3.1 The Expected Eluder Dimension

Theorem 51. *Given $\epsilon > 0$, and $\mathcal{D}$ a context distribution. For a sequence of confidence sets $\{\mathcal{F}_t : t \in \mathbb{N}\}$ used in LinUCB algorithm, if actions are selected such that $A_t \in \arg\max_{a \in \mathcal{A}} \left\{ \sup_{f \in \mathcal{F}_t} f(a) \right\}$ at each time t, then the regret is bounded by*

$$\mathcal{O}\left(\sqrt{\dim_{\mathbb{E}}(\mathcal{F}, \epsilon) \log\left(\mathcal{N}\left(F, \epsilon, \|\cdot\|_\infty\right)\right) T} \right)$$

Here, $\dim_{\mathbb{E}}$ denotes the expected eluder dimension respect to the distribution the context samples from ($\mathcal{D}$), and $\mathcal{N}$ denotes the covering number.

Definition 52 (Set Widths). *Define the width of a subset $\bar{\mathcal{F}} \subset \mathcal{F}$ at an action $a \in \mathcal{A}$ by*

$$w(\bar{\mathcal{F}}, a) = \sup_{f, f' \in \bar{\mathcal{F}}} (f(a) - f'(a)). \tag{27}$$

This is a worst-case measure of the uncertainty about the payoff $f_\theta(a)$ at a given that $f_\theta \in \bar{\mathcal{F}}$. We would like to look on $\bar{\mathcal{F}} = \mathcal{F}_t := \{f \in \tilde{\mathcal{F}} : \|f - \hat{f}_t^{LS}\|_{2,E_t} \leq \sqrt{\beta_t}\}$ where β_t is an appropriately chosen confidence parameter, and the empirical 2-norm $\|\cdot\|_{2,E_t}$ is defined by

$$\|g\|_{2,E_t}^2 = \sum_{t=1}^{t-1} g^2(A_k). \tag{28}$$

Hence $\|f - f_\theta\|_{2,E_t}^2$ measures the cumulative discrepancy between the previous predictions of f and f_θ.

Regret Decomposition

Proposition 53. *Fix any sequence $\{\mathcal{F}_t : t \in \mathbb{N}\}$, where $\mathcal{F}_t \subset \tilde{\mathcal{F}}$ is measurable with respect to $\sigma(H_t)$. Then for any $T \in \mathbb{N}$, with probability 1,*

$$R(T, \pi_{\mathcal{F}_1:T}) \leq \sum_{t=1}^{T} [w_t(\mathcal{F}_t, A_t) + C\mathbf{I}\{f_\theta \notin \mathcal{F}_t\}] \tag{29}$$

$$\mathbb{E}[R(T, \pi_{\mathcal{F}_1:T})] \leq \mathbb{E}\left[\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t)\right] + C\mathbf{I}\{f_\theta \notin \mathcal{F}_t\}. \tag{30}$$

Bounding the sum of widths To clarify the proof, we adopt a broader representation: for all $t > 0$, we let $A_t = \langle x_t, a_t \rangle$, as defined in Definition 43.

Claim 54. *Let B be a ϵ-independent sequence of context and actions, $B = \{\langle x_1, a_1 \rangle, \ldots, \langle x_n, a_n \rangle\}$. Then B holds that, EP$(B, \epsilon) = |B| = n$.*

Proof. Assume that B is ϵ-independent sequence of context and actions. We prove that EP$(B, \epsilon) = |B|$ by induction.

Statement: For any sequence with ϵ-eluder predicate valued n, if we add one ϵ-independent element to the sequence, the eluder predicate of the group becomes $n + 1$.

Base Case: The base case is trivial.

Inductive Step: Assume that the statement is true for a ϵ-independent sequence of size n. That is, when we add n ϵ-independent element in a sequence the eluder predicate of the sequence becomes n (inductive hypothesis). Now consider a ϵ-independent sequence length $n + 1$. We need to show that a ϵ-independent sequence results eluder predicate value $n + 1$. By our inductive hypothesis, for this sequence a prefix length n is still ϵ-independents sequence, then by the inductive hypothesis the value of ϵ-eluder predicate for that sequence is n. Then by adding an ϵ-independent element we get that the eluder predicate becomes $n + 1$. Finally we get that EP$(B, \epsilon) = |B|$.

□

Claim 55. *Let $T > 0$ and $x \sim \mathcal{D}^T$. For a sequence of T pairs of context and action $\{\langle x_1, a_1 \rangle, \ldots, \langle x_T, a_T \rangle\}$. ,the largest cardinality of subsequence B of ϵ-independent pairs, $\mathbb{E}_{x \sim \mathcal{D}^T}|B| \leq \tilde{d}$.*

Proof. Let $T > 0$ and let $x \sim \mathcal{D}^T$ we define

$$\tilde{A}^x = \{\tilde{A}_1, \ldots, \tilde{A}_T\} : \tilde{A}^x_t = \langle x_t, \alpha_t \rangle \text{ s.t. } \tilde{A}^x = argmax_{A:X(A)=x} \text{EP}(A, \epsilon).$$

Where **E**luder **P**redicate denote: $\text{EP}(A, \epsilon) := \max_{I \subseteq [n]} \text{EP}(A, I, \epsilon)$.

Fix $j > 0$ a set of j indexes $J \subset [T]$. Let $B \subseteq \tilde{A}^x$, then $\exists J \subseteq [T] : \tilde{A}^x[J] = \langle \tilde{A}_i \rangle_{i \in J} = B_{x,J}$. Therefore, $\text{EP}(B, \epsilon) = \max_{K \subseteq J} \text{EP}(B, K, \epsilon) = \text{EP}(\tilde{A}^x, K, \epsilon) \leq \max_{I \subseteq [T]:|I|=|K|} \text{EP}(\tilde{A}^x, I, \epsilon) \leq \max_{I \subseteq [T]} \text{EP}(\tilde{A}^x, I, \epsilon)$. Thus we get that for any subsequence we get that $\text{EP}(B, \epsilon) \leq \text{EP}(\tilde{A}^x, \epsilon)$. Next, assume that B is ϵ-independent subset of $\tilde{A}^x$. We proved in Claim 54 that,

$$\text{EP}(B, \epsilon) = \max_{K \subseteq J} \text{EP}(B, K, \epsilon) = \text{EP}(B, J, \epsilon) = |B|.$$

As a conclusion, we have for any subsequence $B \subseteq \tilde{A}^x$ that is ϵ-independent, that $|B| = \text{EP}(B, \epsilon) \leq \text{EP}(\tilde{A}^x, \epsilon)$.
Then we get our claim $\mathbb{E}_{x \sim \mathcal{D}^T} |B| = \mathbb{E}_{x \sim \mathcal{D}^T} |B_{x,J}| \leq \mathbb{E}_{x \sim \mathcal{D}^T} \text{EP}(\tilde{A}^x, \epsilon) = \mathbb{E}_{x \sim \mathcal{D}^T} [\max_{A:X(A)=x} \text{EP}(A, \epsilon)]$ $= \tilde{d}$

□

Proposition 56. *Let $i > 0$, $x^i \sim \mathcal{D}^T$, a sequence of T contexts that sample in run i. The expected regret over the runs of the algorithm is:*

$$\mathbb{E}_i R(T, \pi^i_{\mathcal{F}_1:T}) = \mathbb{E}_{x \sim \mathcal{D}^T} R(T, \pi_{\mathcal{F}_1:T})$$

where, $\pi_{\mathcal{F}_1:T}$ is the expected policies respect to the expected $x \sim \mathcal{D}^T$.

Corollary 57. $\mathbb{E}[R(T, \pi_{\mathcal{F}_1:T})] = \mathbb{E}_{x \sim \mathcal{D}^T}[R(T, \pi_{\mathcal{F}_1:T})] \leq \mathbb{E}_{x \sim \mathcal{D}^T} \left[\sum_{t=1}^T w_t(\mathcal{F}_t, A_t) \right] + C\mathbf{I}\{f_\theta \notin \mathcal{F}_t\}$.

Lemma 58 (Bounding the number of large widths)**.** *If $(\beta_t \geq 0 | t \in \mathbb{N})$ is a non-decreasing sequence and $\mathcal{F}_t := \{f \in \tilde{\mathcal{F}} : ||f - f^{LS}_t||_{2,E_t} \leq \sqrt{\beta_t}\}$ then*

$$\mathbb{E}_{x \sim \mathcal{D}^T} \big[\sum_{t=1}^T \mathbf{I}(w(\mathcal{F}_t, A_t) > \epsilon) \big] \leq \left(\frac{4\beta_T}{\epsilon^2} + 1 \right) \dim_{\mathbb{E}}(\mathcal{F}, \epsilon)$$

for all $T \in \mathbb{N}$ and $\epsilon > 0$.

Proof. We begin by denote $A = (A_1, \ldots, A_T)$ and showing that if $w_t := w(\mathcal{F}_t, A_t) > \epsilon$ then A_t is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_1, \ldots, A_{t-1})$ for $T > t$. To see this, note that if $w_t(A_t) > \epsilon$ there are $f, f' \in \mathcal{F}_t$, such that $f(A_t) - f'(A_t) > \epsilon$. By definition, since $f(A_t) - f'(A_t) > \epsilon$, if A_t is ϵ-dependent on a subsequence $(A_{i_1}, \ldots, A_{i_k})$ of $(A_1, \ldots, A_{t-1})$ then $\sum_{j=1}^k (f(A_{i_j}) - f'(A_{i_j}))^2 > \epsilon^2$. It follows that, if A_t is ϵ-dependent on K disjoint subsequences of $(A_1, \ldots, A_{t-1})$ then $||f - f'||_{2,E_t} \geq K\epsilon^2$. By the triangle inequality, we have

$$\begin{aligned} ||f - f'||_{2,E_t} &\leq ||f - f^{LS}_t||_{2,E_t} + ||f' - f^{LS}_t||_{2,E_t} \\ &\leq 2\sqrt{\beta_t} \leq 2\sqrt{\beta_T}. \end{aligned}$$

and it follows that $K < 4\beta_T/\epsilon^2$.

Next, we show that for sequence of context $(x_1, \ldots, x_T)$ and in any action sequence $(a_1, \ldots, a_T)$, there is some element $\alpha_\tau = \langle x_\tau, a_\tau \rangle$ that is ϵ-dependent on at least $\lfloor \tau / \tilde{d} \rfloor$ disjoint subsequences of $(\alpha_1, \ldots, \alpha_{\tau-1})$,

where $\tilde{d} := \dim_{\mathbb{E}}(\mathcal{F}, \epsilon)$. Now to show this, for an integer K satisfying $K\tilde{d}+1 \leq T \leq K\tilde{d}+\tilde{d}$, we will construct K disjoint subsequences $B_1, \ldots, B_K$. First let $B_i = (\alpha_i)$ for $i = 1, \ldots, K$. If α_{K+1} is ϵ-dependent on each subsequence $B_1, \ldots, B_K$, our claim is established. Otherwise, select a subsequence B_i such that α_{K+1} is ϵ-dependent and append α_{K+1} to B_i. Repeat this process for elements with indices $j > K+1$ until α_j is ϵ-dependent on each subsequence or α_T is. In the latter scenario when $\sum_{i=1}^{K} |B_i| \geq K\tilde{d}$, we will denote each B_i respect to the indexes of A. Thus for any $i \leq K$ $\exists I_i \subseteq [T]$ s.t. $B_i = \langle A_j \rangle_{j \in I_i}$, we will denote each of these subsequnce according to that $B_{x,I_i} = B_i$, where x is T contexts that sample in he algorithm. Since each element of a subsequence B_{x,I_i} is ϵ-dependent of its predecessors, we have:

$$|B_{x,I_i}| = \text{EP}(B_{x,I_i}, \epsilon).$$

Therefore we have:

$$\sum_{i=1}^{K} |B_{x,I_i}| \leq \sum_{i=1}^{K} \text{EP}(B_{x,I_i}, \epsilon) \leq K \cdot \text{EP}(A, \epsilon)$$

Eventually, we can see that the expected amount of elements in these subsequences is:

$$\mathbb{E}_{x \sim \mathcal{D}^T}\big[\sum_{i=1}^{K} |B_{x,I_i}|\big] \leq K \cdot \mathbb{E}_{x \sim \mathcal{D}^T}\big[\text{EP}(A, \epsilon)\big] = K\tilde{d}. \tag{31}$$

Now consider taking $(\alpha_1, \ldots, \alpha_T)$ to be the sequence $(A_{\tau_1}, \ldots, A_{\tau_T})$ of $(A_1, \ldots, A_T)$ consisting of elements A_t for which $w_t(A_t) > \epsilon$. As we have established earlier, α_t is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_{\tau_1}, \ldots, A_{\tau_{t-1}})$. It follows that each α_j is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(\alpha_1, \ldots, \alpha_{j-1})$. Combining this with the fact we have established that there is α_j that is ϵ-dependent on at least $\tau/\tilde{d} - 1$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$, we have $T/\tilde{d} - 1 \leq 4\beta_T/\epsilon^2$. It follows that $T \leq (4\beta_T/\epsilon^2 + 1)\tilde{d}$, which is our desired result. □

Lemma 59 (Bounding the sum of widths)**.** *If $(\beta_t \geq 0 | t \in \mathbb{N})$ is a nondecreasing sequence and $\mathcal{F}_t := \{f \in \tilde{\mathcal{F}} : ||f - f_t^{LS}||_{2,E_t} \leq \sqrt{\beta_t}\}$ then with probability 1,*

$$\mathbb{E}_{x \sim \mathcal{D}^T}\big[\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t)\big] \leq T\epsilon + \min\left\{\dim_{\mathbb{E}}(\mathcal{F}, \epsilon), T\right\} C + 4\sqrt{\dim_{\mathbb{E}}(\mathcal{F}, \epsilon)\beta_T T} \tag{32}$$

for all $T \in \mathbb{N}$.

Proof. Let $x \sim \mathcal{D}^T$, to reduce notation, write $A = (A_1, \ldots, A_T)$, $d_{x,\epsilon} = \text{EP}(A, \epsilon)$ and $w_t = w(\mathcal{F}_t, A_t)$. Rearrange the sequence $(w_1, \ldots, w_T)$ so that $w_{i_1} \geq w_{i_2} \geq \ldots \geq w_{i_T}$. We have

$$\begin{aligned}\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t) = \sum_{t=1}^{T} w_{i_t} &= \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} \leq \epsilon\} + \mathbb{E}\big[\sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \epsilon\} \\ &\leq T \cdot \epsilon + \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \epsilon\}\end{aligned}$$

Now, we know $w_{i_t} \leq C$. In addition, $w_{i_t} > \alpha \Rightarrow \sum_{t=1}^{T} \mathbf{1}(w(\mathcal{F}_t, A_k) > \alpha) \geq t$. By Proposition 3, this can only occur if $t < (\frac{4\beta_T}{\alpha^2} + 1)d_{x,\alpha}$. For $\alpha \geq \epsilon$, $d_{x,\alpha} \leq d_{x,\epsilon}$, since $\text{EP}(A, \cdot)$ is nonincreasing in ϵ. Therefore, when $w_{i_t} > \alpha \geq \epsilon$, $t \leq (\frac{4\beta_T}{\alpha^2} + 1)d_{x,\epsilon}$ which implies

$$\alpha \leq \sqrt{\frac{4\beta_T d_{x,\epsilon}}{t - d_{x,\epsilon}}}\,.$$

This shows that if $w_{i_t} > \alpha$, then $w_{i_t} \leq \min\{C, \frac{4\beta_T d_{x,\epsilon}}{T-t+d_{x,\epsilon}}\}$. Therefore,

$$\sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \alpha\} = \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \epsilon\} \leq d_{x,\epsilon}C + \sum_{t=\tilde{d}+1}^{T} \sqrt{\frac{4\beta_T d_{x,\epsilon}}{t - d_{x,\epsilon}}}$$
$$\leq d_{x,\epsilon}C + 2\sqrt{\beta_T d_{x,\epsilon}}\left(\sum_{t=0}^{T} \frac{1}{\sqrt{t}}\right) = d_{x,\epsilon}C + 2\sqrt{\beta_T d_{x,\epsilon} T}.$$

□

To complete the proof, we will see that the expected regret respect to x is:

$$\mathbb{E}_{x\sim\mathcal{D}^T}\big[\sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \epsilon\}\big] \leq \mathbb{E}_{x\sim\mathcal{D}^T}\big[d_{x,\epsilon}C + 2\sqrt{\beta_T d_{x,\epsilon} T}\big] = \tilde{d}C + 2\sqrt{\beta_T \tilde{d} T}.$$

We combine this with the fact that the sum of widths is always bounded by CT. This implies:

$$\mathbb{E}_{x\sim\mathcal{D}^T}\big[\sum_{t=1}^{T} w_{i_t}\big] \leq \min\left\{TC, T\epsilon + \tilde{d}C + 2\sqrt{\tilde{d}\beta_T T}\right\}$$
$$\leq T\epsilon + \min\{\dim_{\mathbb{E}}(\mathcal{F}, \epsilon), T\}C + 2\sqrt{\dim_{\mathbb{E}}(\mathcal{F}, \epsilon)\beta_T T}.$$

5.3.2 The High Probability Dimension

Another option to define the Eluder dimension in stochastic setting, is if we have a set of contexts that received most of the time and they induce a smaller dimension. Therefore we will look on the next definition.

Definition 60 (high-probability Eluder dimension)**.** *Let $\epsilon, \delta > 0$, and $\mathcal{D}$ a context distribution. Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$. Let W be a random variable of a sequence of n contexts drawn from $\mathcal{D}$. Given a sequence $Z = \{(x_1, a_1), \ldots, (x_n, a_n)\}$ and $X(Z) = (x_1, \ldots, x_n)$. We define a predicate $EP_{\mathcal{G}}(Z, i, \epsilon)$, which returns 1 if (x_i, a_i) is ϵ-independent from $(x_1, a_1), \ldots, (x_{i-1}, a_{i-1})$. The high-probability Eluder dimension $\tilde{d} = \dim_{\mathbb{E}}^{hp}(\mathcal{G}, \epsilon, \delta)$ is*

$$\min_{\mathcal{N}(U)} \max_{Z:X(Z)\subseteq U} EP(Z, \epsilon), \tag{33}$$

where $\mathcal{D}(U) \geq 1 - \delta$, and $\mathcal{N}$ denotes the covering number.

Claim 61. *Let $\epsilon, \delta, n > 0$, and $\mathcal{D}$ a context distribution.*

$$d = \dim_{\mathrm{E}}(\mathcal{G}, \epsilon) \geq \dim_{\mathbb{E}}^{hp}(\mathcal{G}, \epsilon, \delta) = \tilde{d}_{hp}$$
$$d = \dim_{\mathrm{E}}(\mathcal{G}, \epsilon) \geq \dim_{\mathbb{E}}(\mathcal{G}, \epsilon, \mathcal{D}^n) = \tilde{d}$$

Proof.

- $\dim_{\mathrm{E}}(\mathcal{G}, \epsilon) \geq \dim_{\mathbb{E}}^{hp}(\mathcal{G}, \epsilon, \delta)$: This inequality is easy to see due to $U \subseteq \mathcal{X}$.

- $\dim_{\mathrm{E}}(\mathcal{G},\epsilon) \geq \dim_{\mathbb{E}}(\mathcal{G},\epsilon)$:

$$\begin{aligned}\dim_{\mathrm{E}}(\mathcal{G},\epsilon) &= \max_{Z} \sum_{i=1}^{\infty} \mathrm{EP}(Z,i,\epsilon) \\ &\geq \max_{Z} \sum_{i=1}^{n} \mathrm{EP}(Z,i,\epsilon) \geq \max_{Z} \max_{I\subseteq[n]} \mathrm{EP}(Z,I,\epsilon) \\ &\geq \mathbb{E}_{W\sim\mathcal{D}^n}\big[\max_{Z:X(Z)=W} \max_{I\subseteq[n]} \mathrm{EP}(Z,I,\epsilon)\big] = \dim_{\mathbb{E}}(\mathcal{G},\epsilon).\end{aligned}$$

□

Note that if the support of $\mathcal{D}$ contains sequences with Eluder dimension d, then as $n \to \infty$, the Expected Eluder dimension will converge to $\tilde{d}$. Then $\tilde{d} \geq \tilde{d_{hp}}$, and it may be more effective to bound our regret with the high-probability version.

Theorem 62. *Given confidence parameter $\delta > 0$, $\epsilon > 0$, and $\mathcal{D}$ a context distribution. For a sequence of confidence sets $\{\mathcal{F}_t : t \in \mathbb{N}\}$ used in UCB algorithm, if actions are selected such that $A_t \in \arg\max_{a\in\mathcal{A}} \left\{\sup_{f\in\mathcal{F}_t} f(a)\right\}$ at each time t, then the regret is bounded by*

$$\mathcal{O}\left(\sqrt{\min\left\{\dim_{\mathbb{E}}^{hp}(\mathcal{F},\epsilon,\delta), \dim_{\mathbb{E}}(\mathcal{F},\epsilon)\right\} \log\left(\mathcal{N}\left(F,\epsilon,\|\cdot\|_\infty\right)\right) T} + \delta T + \epsilon T\right)$$

Here, $\dim_{\mathbb{E}}^{hp}$ denotes the high probability eluder dimension respect to the distribution the context samples from, and $\mathcal{N}$ denotes the covering number.

Definition 63 (Effective Class)**.** *Given distribution $\mathcal{D} : \mathcal{X} \to \mathbb{R}$ Define the set*

$$\tilde{\mathcal{F}_\delta} = \{f_{x_i} \in \mathcal{F} | \min_n \mathcal{D}(\bigcup_{i=1}^{n} B_i) > 1-\delta \text{ s.t. } B_i = Ball(x_i,\epsilon,\|\cdot\|_\infty)\}, \tag{34}$$

as the Effective class of $\mathcal{F}$ respect to $\mathcal{D}$.

For clearness we will denote $\tilde{\mathcal{F}} := \tilde{\mathcal{F}_\delta}$.

Claim 64. *The Effective Class eluder dimension is equal to the high probability eluder dimension of the class.*

$$\dim_{\mathbb{E}}^{hp}(\mathcal{F},\epsilon,\delta) = \dim_{\mathrm{E}}(\tilde{\mathcal{F}_\delta},\epsilon). \tag{35}$$

Proof Outlines We will bound our regret on two disjoint subset of $[T]$, the good iteration $G_T = \{t \leq T | \theta_t \in \tilde{\mathcal{F}}\}$ and the bad iterations $B_T = \{t \leq T | \theta_t \notin \tilde{\mathcal{F}}\}$. Its worth to noting that $\mathbb{P}(t \in G_t) = \mathbb{P}(\theta_t \in \tilde{\mathcal{F}}) > 1-\delta$. A key to our analysis is recent observation that the regret of UCB algorithm can be decomposed in terms of confidence sets.

Definition 65 (Set Widths)**.** *Define the width of a subset $\bar{\mathcal{F}} \subset \mathcal{F}$ at an action $a \in \mathcal{A}$ by*

$$w(\bar{\mathcal{F}},a) = \sup_{f,f'\in\bar{\mathcal{F}}} (f(a) - f'(a)). \tag{36}$$

This is a worst-case measure of the uncertainty about the payoff $f_\theta(a)$ at a given that $f_\theta \in \bar{\mathcal{F}}$. We would like to look on $\bar{\mathcal{F}} = \tilde{\mathcal{F}}$.

Regret Decomposition

Proposition 66. *The total regret of the algorithms is*

$$R(T, \pi_{\mathcal{F}_1:T}) = R(T, \pi_{\mathcal{F}_1:T}|G_T) + R(T, \pi_{\mathcal{F}_1:T}|B_T)$$

We can bound $R(T, \pi_{\mathcal{F}_1:T}|B_T) \leq \sum_{t=1}^{T} C\mathbf{I}\{\theta_t \notin \tilde{\mathcal{F}}\} \leq CT\delta$.

Proposition 67. *Fix any sequence $\{\mathcal{F}_t : t \in \mathbb{N}\}$, where $\mathcal{F}_t \subset \tilde{\mathcal{F}}$ is measurable with respect to $\sigma(H_t)$. Then for any $T \in \mathbb{N}$, with probability 1,*

$$R(T, \pi_{\mathcal{F}_1:T}|G_T) \leq \sum_{t=1}^{T} [w_t(\mathcal{F}_t, A_t) + C\mathbf{I}\{f_\theta \notin \mathcal{F}_t\}] \tag{37}$$

$$\mathbb{E}[R(T, \pi_{TS})|G_T] \leq \mathbb{E}\left[\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t)\right] + C\mathbf{I}\{f_\theta \notin \mathcal{F}_t\}. \tag{38}$$

Bounding the sum of widths The abstract confidence sets we construct are centered around least squares estimates $\hat{f}_t^{LS} \in \arg\min_{f \in \tilde{\mathcal{F}}} L_{2,t}(f)$ where $L_{2,t}(f) = \sum_{t=1}^{t-1}(f(A_t) - R_t)^2$ is the cumulative squared prediction error. The sets take the form $F_t := \{f \in \tilde{\mathcal{F}} : ||f - \hat{f}_t^{LS}||_{2,E_t} \leq \sqrt{\beta_t}\}$ where β_t is an appropriately chosen confidence parameter, and the empirical 2-norm $||\cdot||_{2,E_t}$ is defined by $||g||_{2,E_t}^2 = \sum_{t=1}^{t-1} g^2(A_k)$. Hence $||f - f_\theta||_{2,E_t}^2$ measures the cumulative discrepancy between the previous predictions of f and f_θ.

Lemma 68 (Bounding the number of large widths)**.** *If $(\beta_t \geq 0 | t \in \mathbb{N})$ is a non-decreasing sequence and $\mathcal{F}_t := \{f \in \tilde{\mathcal{F}} : ||f - f_t^{LS}||_{2,E_t} \leq \sqrt{\beta_t}\}$ then*

$$\sum_{t=1}^{T} \mathbf{I}(w(\mathcal{F}_t, A_t) > \epsilon) \leq \left(\frac{4\beta_T}{\epsilon^2} + 1\right) \dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$$

for all $T \in \mathbb{N}$ and $\epsilon > 0$.

Proof. We begin by showing that if $w_t := w(\mathcal{F}_t, A_t) > \epsilon$ then A_t is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_1, \ldots, A_{t-1})$ for $T > t$. To see this, note that if $w_t(A_t) > \epsilon$ there are $f, f' \in \mathcal{F}_t$, such that $f(A_t) - f'(A_t) > \epsilon$. By definition, since $f(A_t) - f'(A_t) > \epsilon$, if A_t is ϵ-dependent on a subsequence $(A_{i_1}, \ldots, A_{i_k})$ of $(A_1, \ldots, A_{t-1})$ then $\sum_{j=1}^{k}(f(A_{i_j}) - f'(A_{i_j}))^2 > \epsilon^2$. It follows that, if A_t is ϵ-dependent on K disjoint subsequences of $(A_1, \ldots, A_{t-1})$ then $||f - f'||_{2,E_t} \geq K\epsilon^2$. By the triangle inequality, we have

$$\begin{aligned} ||f - f'||_{2,E_t} &\leq ||f - f_t^{LS}||_{2,E_t} + ||f' - f_t^{LS}||_{2,E_t} \\ &\leq 2\sqrt{\beta_t} \leq 2\sqrt{\beta_T}. \end{aligned}$$

and it follows that $K < 4\beta_T/\epsilon^2$.

Next, we show that in any action sequence $(a_1, \ldots, a_T)$, there is some element a_j that is ϵ-dependent on at least $\tau/\tilde{d} - 1$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$, where $\tilde{d} := \dim_{\mathrm{E}}(\mathcal{F}, \epsilon)$. By Claim 64, we have, $\tilde{d} = \dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$. Now to show this, for an integer K satisfying $K\tilde{d}+1 \leq T \leq K\tilde{d}+\tilde{d}$, we will construct K disjoint subsequences $B_1, \ldots, B_K$. First let $B_i = (a_i)$ for $i = 1, \ldots, K$. If a_{K+1} is ϵ-dependent on each subsequence $B_1, \ldots, B_K$, our claim is established. Otherwise, select a subsequence B_i such that a_{K+1} is

ϵ-dependent and append a_{K+1} to B_i. Repeat this process for elements with indices $j > K+1$ until a_i is ϵ-dependent on each subsequence or a_T is. In the latter scenario $\sum_{i=1}^{K} |B_i| \geq K\tilde{d}$, and since each element of a subsequence B_i is ϵ-dependent of its predecessors, and each element that take in count if it from the interval G_T, we have $|B_i| = \tilde{d}$. In this case, a_T must be ϵ-dependent on each subsequence, by the definition of $\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$.

Now consider taking $(a_1, \ldots, a_T)$ to be the subsequence $(A_1, \ldots, A_T)$ of $(A_1, \ldots, A_T)$ consisting of elements A_t for which $w_t(A_t) > \epsilon$. As we have established earlier, a_t is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_1, \ldots, A_{t-1})$. It follows that each a_j is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$. Combining this with the fact we have established that there is some a_j that is ϵ-dependent on at least $\tau/\tilde{d} - 1$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$, we have $T/\tilde{d} - 1 \leq 4\beta_T/\epsilon^2$. It follows that $T \geq 4\beta_T/\epsilon^2 + \tilde{d}$, which is our desired result. □

Lemma 69 (Bounding the sum of widths)**.** *If $(\beta_t \geq 0 | t \in \mathbb{N})$ is a nondecreasing sequence and $\mathcal{F}_t := \{f \in \tilde{\mathcal{F}} : ||f - f_t^{LS}||_{2,E_t} \leq \sqrt{\beta_t}\}$ then with probability 1,*

$$\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t) \leq \epsilon T + \min\left\{\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta), T\right\} C + 4\sqrt{\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta)\beta_T T} \tag{13}$$

for all $T \in \mathbb{N}$.

Proof. To reduce notation, write $\tilde{d} = \dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$ and $w_t = w(\mathcal{F}_t, A_t)$. Reorder the sequence $(w_1, \ldots, w_T)$ so that $w_{i_1} \geq w_{i_2} \geq \ldots \geq w_{i_T}$. We have

$$\begin{aligned}\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t) = \sum_{t=1}^{T} w_{i_t} &= \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} \leq \alpha\} + \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \alpha\} \\ &\leq \epsilon T + \sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \alpha\}.\end{aligned}$$

The final step in the above inequality uses that either $w_{i_T} = \epsilon$ and $\sum_{t=1}^{T-1} w_{i_t} = \epsilon T$ or w_{i_T} is set below the smallest possible width and hence $\mathbf{1}\{w_{i_t} \leq \alpha\}$ never occurs.

Now, we know $w_{i_t} \leq C$. In addition, $w_{i_t} > \epsilon \Rightarrow \sum_{t=1}^{T} \mathbf{1}(w(\mathcal{F}_t, A_k) > \epsilon) \geq t$. By Proposition 3, this can only occur if $t < (\frac{4\beta_T}{\epsilon^2} + 1)\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$. For $t \leq \alpha$, $\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta) \leq \dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta) = \tilde{d}$, since $\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \epsilon, \delta)$ is nonincreasing in ϵ'. Therefore, when $w_{i_t} > \epsilon\alpha$, $t \leq (\frac{4\beta_T}{\epsilon^2} + 1)\tilde{d}$ which implies $t \leq \frac{4\beta_T \tilde{d}}{\epsilon^2}$. This shows that if $w_{i_t} > \alpha$, then $w_{i_t} \leq \min\{C, \frac{4\beta_T \tilde{d}}{T - t + \tilde{d}}\}$. Therefore,

$$\begin{aligned}\sum_{t=1}^{T} w_{i_t}\mathbf{1}\{w_{i_t} > \alpha\} &\leq dC + \sum_{t=\tilde{d}+1}^{T} \frac{4d\beta_T}{T - t - \tilde{d}} + 2\sqrt{\beta_T}\sum_{t=0}^{\sqrt{T}} \frac{1}{\sqrt{t}} \\ &\leq \tilde{d}C + 4\sqrt{\beta_T}\left(\sum_{t=0}^{\sqrt{T}} \frac{1}{\sqrt{t}}\right) = \tilde{d}C + 4\sqrt{\beta_T T}.\end{aligned}$$

□

To complete the proof, we combine this with the fact that the sum of widths is always bounded by CT. This implies:

$$\sum_{t=1}^{T} w_t(\mathcal{F}_t, A_t) \leq \min\left\{TC, \frac{1}{T} + \dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta)C\right\} + 4\sqrt{\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta)\beta_T T}$$
$$\leq \epsilon T + \min\{\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta), TC\} + 4\sqrt{\dim_{\mathbb{E}}^{hp}(\mathcal{F}, \alpha, \delta)\beta_T T}.$$

Confidence Bound The development of the confidence bound in this proof follows the approach presented in (Russo and Van Roy, 2013). For further details and a comprehensive explanation of the underlying principles, the reader is referred to that work.

6 Improved Analysis of the GOLF Algorithm

Bellman Eluder dimension a new complexity measure for reinforcement learning. A notable contribution to sample complexity, provides a unified complexity measure that subsumes many previously studied RL problems, including those characterized by low Bellman rank and low Eluder dimension which presented in the previous sections (3, 5). The paper demonstrates that RL problems with a low BE dimension can be addressed efficiently, subsuming a majority of existing tractable RL problems such as tabular MDPs, linear MDPs, and reactive POMDPs. Additionally, the authors propose a new optimization-based algorithm called GOLF and reanalyze OLIVE (Jiang et al. (2017)). Both algorithms are proven to learn near-optimal policies for low BE dimension problems with a sample complexity that is polynomial in relevant parameters but independent of the state-action space size. This characteristic ensures their applicability to a broad range of RL settings, providing improved regret and sample complexity results compared to existing methods.

Building on this foundational work, our paper aims to further improve the sample complexity for RL problems characterized by a low BE dimension. We refined Theoretical Analysis and provide a comprehensive theoretical analysis of GOLF, establishing improved regret and sample complexity bounds.

6.1 Introduction

Algorithm 4 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 $f^k \in_R \left\{ f \in \mathcal{B}^{k-1} \middle| \max_{f' \in \mathcal{B}^{k-1}} f'(s_1, \pi_{f'}(s_1)) - f(s_1, \pi_f(s_1)) \leq \max\{2\epsilon, \beta_k\} \right\}$.
4: **Collect** a trajectory $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1})$ by following π^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**

$$\mathcal{B}^k = \left\{ f \in \mathcal{F} : \ \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \leq \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta \text{ for all } h \in [H] \right\},$$

$$\text{where } \mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{(s,a,r,s') \in \mathcal{D}_h} [\xi_h(s,a) - r - \max_{a' \in \mathcal{A}} \zeta_{h+1}(s', a')]^2. \tag{39}$$

7: **Output** π^{out} sampled uniformly at random from $\{\pi^k\}_{k=1}^K$.

The GOLF algorithm 4, introduced by Jin et al. (2021a), has garnered attention for its simplicity and efficiency. In their innovative work presented the GOLF algorithm, showcasing a regret guarantee of $\mathcal{O}\left(H\sqrt{K \cdot \dim_{\text{BE}}(\mathcal{F}, \sqrt{1/K}) \log[KH\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}(1/K)/\delta]}\right)$, where $\dim_{\text{BE}}$ represents the Bellman-Eluder dimension (see definition 40) of the function class $\mathcal{F}$, K is the number of episodes, and $\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}(1/K)$ denotes the size of the $1/K$-covering number(see definition 10) of $\mathcal{F}$ and $\mathcal{G}$. Note that the algorithm assume completeness, i.e, completeness requires the function class $\mathcal{F}$ to be closed under the Bellman operator and $\mathcal{G}$ is the completeness (see definition 9). We consider the function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, and also assuming realizability (see Assumption 8).

The regret analysis can be separated to two main steps:

Step 1. Bounding the regret by Bellman error: Upper bound to the cumulative regret by the summation of Bellman error, with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\left(V_1^\star(s_1)-V_1^{\pi^k}(s_1)\right)\le\sum_{k=1}^{K}\left(\max_a f_1^k(s_1,a)-V_1^{\pi^k}(s_1)\right)\overset{(i)}{=}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathcal{E}(f^k,\pi^k,h), \tag{40}$$

where (i) follows from standard policy loss decomposition (see [Lemma 1 in Jiang et al. 2017]).

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

The following the main lemma that used in [Appendix D in Jin et al. 2021a, "Proofs for GOLF"]:

Lemma 70. *Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)|\le C$ 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 $k\in[K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2\le\beta$. 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 70 focuses on a fixed step h and bounds the cumulative Bellman error $\sum_{k=1}^{K}\mathcal{E}(f^k,\pi^k,h)$ using Lemma 77, which proves the condition $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2\le\beta$. The bound on the accumulating rate of Bellman error in Lemma 70 can be divided into two main components:

- $\sqrt{\dim_{\mathrm{DE}}(\Phi,\Pi,\omega)\beta k}+\min\{k,\dim_{\mathrm{DE}}(\Phi,\Pi,\omega)\}C$
- $k\omega$

The first component addresses the error when the chosen policy incurs a large error ($>\omega$). The second component applies when the error is small ($\le\omega$). For instance, if the algorithm consistently selects a good policy for step h with no errors, the regret for this step is bounded by $\sum_{t=1}^{k}|\mathbb{E}_{\mu_{h,t}}[\phi_{h,t}]|\le\mathcal{O}(K\omega)$.

Sample Complexity

Remark 71 (From average regret to PAC guarantees and optimal policies.)**.** *There is some intimate connection between regret bounds and PAC guarantees that has been pointed out previously by Jin et al. (2018). For instance, by fixing the initial state distribution to be identical (e.g., $s_1^{(n)}=s$ for all $1\le n\le K$) and choosing the output policy $\hat{\pi}$ uniformly at random from $\{\pi^{(n)}|1\le n\le N\}$, one can easily verify that this output policy $\hat{\pi}$ is ϵ-optimal for state s, as long as $\frac{1}{K}$Regret$(K)\le\epsilon$.*

Following this approach, the algorithm guarantees that the resulting output will be an ϵ-optimal policy if we perform at least $K\ge\frac{H^2d\beta}{\epsilon^2}$ phases. Furthermore, after $K=\frac{H^2d\beta}{\epsilon^2}$ episodes, the regret will be:

$$\mathcal{R}(K)=\mathcal{O}\left(\frac{H^2d\beta}{\epsilon}\right). \tag{41}$$

In our work, we present a refined analysis of the GOLF algorithm, aiming to refine the regret guarantee to achieve an ϵ-optimal policy, providing enhancement in performance evaluation and robustness of the algorithm. We are also improving the guarantee to get such policy. We will work on all the layers simultaneously, and we will use the next definition:

Definition 72. *Let $\epsilon > 0$, $t \in [K]$, we define h_t to be the number of layers $h \in [H]$ where $|\mathcal{E}(f^t, \pi^t, h)| \geq \epsilon$. In other words, $h_t^\epsilon = \sum_{h \in [H]} \mathbf{I}\{|\mathbb{E}_{\mu_{\{h,t\}}}[\phi_{\{h,t\}}]| > \epsilon\}$.*

In our case, $\forall t, h : \mu_{\{h,t\}}$ is present the probability induced by the policy π_h^t, and $\phi_{\{h,t\}}$ presenting $f_h^t - \mathcal{T} f_{h+1}^t$, which is the bellman error.

Driven form it we will define also the average amount of layers that have a large error:

Definition 73. *Let mark in $\tilde{h} = \frac{1}{|\mathcal{T}^\epsilon|} \times \sum_{t=1}^K h_t^\epsilon$, the average amount of layers that have a large exception, where $\mathcal{T}^\epsilon = \{t \in [K] \mid h_t^\epsilon > 0\}$ set of the phases where there is at least one large exception.*

We know that every round the error is scaling down as function of $\varepsilon \leq \sqrt{\frac{d\beta}{t-d}}$ as shows in equation 45, that's give us the motivation for Lemma 74 presents in the next subsection. And as a conclusion from this lemma (74) and definition 73 we get that:

$$|\mathcal{T}| \leq \mathcal{O}\left(\frac{H}{\tilde{h}} \frac{\beta}{\epsilon^2} \cdot \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon)\right).$$

Following this approach, we develop the regret over the average $\tilde{h}$, as detailed in Lemma 75. Summing the errors over $|\mathcal{T}|$ episodes, Theorem 76 presents a significant improvement in the regret bounds of the GOLF algorithm. By setting $K = \mathcal{O}(|\mathcal{T}|)$, it guarantees that there exists a time step $t \in [K]$ such that the Bellman error of the policy π^t is bounded by ϵ for all steps $h \in [H]$. Furthermore, the total regret of the algorithm is bounded by:

$$\mathcal{R}(K) = \sum_{h=1}^{H} \sum_{t=1}^{K} |\mathbb{E}_{\mu_{\{h,t\}}}[\phi_t]| \leq \mathcal{O}\left(\frac{H^2}{\tilde{h}} \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) \beta \frac{1}{\epsilon}\right).$$

This result not only provides a tighter bound on the regret but also demonstrates the effectiveness of the GOLF algorithm in minimizing the cumulative Bellman error. By leveraging the structure of the Bellman Eluder dimension and optimizing the parameter settings, this theorem ensures that the GOLF algorithm achieves near-optimal performance with improved sample efficiency.

6.2 ϵ-Optimality: Improved GOLF Sample Complexity and Regret Analysis

According to the Bellman-Eluder dimension, we aspire to develop an algorithm that optimize the policy selection based on picking the best function in each dimension. GOLF eliminate the functions that have a high Bellman error and we expect to get functions that are ϵ-independent and therefore we will suffer a big error for them.

Fortunately, we won't get a large error many times, furthermore, we can bound the occurs of it by t^ϵ and with Proposition 79 we would bound this amount and get that:

$$t^\epsilon = \left(\frac{\beta}{\epsilon^2} + 1\right) \times \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon). \tag{42}$$

Therefore, we will have the following lemma regarding the total amount of errors that we get, using definition 72:

Lemma 74. *Let $\epsilon > 0$. Then, $\sum_{t=1}^K h_t^\epsilon \leq H \times \left(\frac{\beta}{\epsilon^2} + 1\right) \times \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)$.*

As demonstrated in Lemma 70, the cumulative error can be divided into two parts: cases of large errors and cases of small errors. There is two crucial ideas to minimize. The first, number of episodes with large errors, the second, the amount of phases the algorithm need to run in order to get near-optimal policy. Lemma 74 assists in achieving this by bounding the number of episodes that have large errors.

To clarify this concept, we remind $\mathcal{T}$ from definition 73, the set $\mathcal{T}^\epsilon = \{t \in [K] \mid h_t^\epsilon > 0\}$. This including Lemma 74 implies:

$$|\mathcal{T}^\epsilon| = \frac{\sum_t h_t}{\tilde{h}} \leq \mathcal{O}\left(\frac{H}{\tilde{h}}\frac{\beta}{\epsilon^2} \cdot \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon)\right). \tag{43}$$

Equipped with this we can present an improved Lemma 70.

Lemma 75. *Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \leq C$ 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 $k \in [K]$, $h \in [H]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_{\{h,t\}}}[\phi_k])^2 \leq \beta$. Then for all $k \in [K]$ and $\omega > 0$ and t^ω(42),*

$$\sum_{h=1}^{H}\sum_{t=1}^{k} |\mathbb{E}_{\mu_{\{h,t\}}}[\phi_t]| \leq \mathcal{O}\left(H\sqrt{\dim_{\mathrm{DE}}(\Phi, \Pi, \omega)\beta min\{k, t^\omega\}} + \min\{k, \dim_{\mathrm{DE}}(\Phi, \Pi, \omega)\}HC + Hk\omega\right).$$

And for K big enough, and with t^ω from (42), our regret, as function of ω, is

$$\sum_{h=1}^{H}\sum_{t=1}^{K} |\mathbb{E}_{\mu_{\{h,t\}}}[\phi_t]| \leq \mathcal{O}\left(H \dim_{\mathrm{DE}}(\Phi, \Pi, \omega)\beta\frac{1}{\omega} + HK\omega\right).$$

Therefore, we get our main result, the regret bound is as follow:

Theorem 76 (Effective Horizon Regret Bound For PAC Guarantee)**.** *By setting $\omega = \epsilon$ and $K = \mathcal{O}\big(|\mathcal{T}^\epsilon|\big)$,we have , $\exists t \in [K]$, s.t. $BE(\pi^t) \leq \epsilon$,$\forall h \in [H]$.*
Furthermore, the total regret of the algorithm:

$$\mathcal{R}(K) = \sum_{h=1}^{H}\sum_{t=1}^{K} |\mathbb{E}_{\mu_{\{h,t\}}}[\phi_t]| \leq \mathcal{O}\left(\frac{H^2}{\tilde{h}} \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\beta\frac{1}{\epsilon}\right). \tag{44}$$

This theorem presents a significant improvement in the regret bounds of the GOLF algorithm. By setting $K = \mathcal{O}(|\mathcal{T}|)$, it guarantees that there exists a time step $t \in [K]$ such that the Bellman error of the policy π^t is bounded by ϵ for all steps $h \in [H]$. Furthermore, the total regret of the algorithm, for such policy, is effective horizon dependent and bounded by:

$$\mathcal{R}(K) \leq \mathcal{O}\left(\frac{H^2}{\tilde{h}} \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\beta\frac{1}{\epsilon}\right).$$

This result not only provides a tighter bound on the regret but also demonstrates the effectiveness of the GOLF algorithm in minimizing the cumulative Bellman error. By leveraging the structure of the Bellman Eluder dimension and optimizing the parameter settings, this theorem ensures that the GOLF algorithm achieves near-optimal performance with improved sample efficiency.

Proof. In equation (40) we bound the regret as follows,

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

We invoke Lemma 77 (a) and Lemma 75 with

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

and we obtain

$$\begin{aligned} \sum_{t=1}^{k} \sum_{h=1}^{H} \mathcal{E}(f^t, \pi^t, h) \leq & \mathcal{O}\left(H \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) \beta \frac{1}{\epsilon} + H \frac{H}{\tilde{h}\epsilon^2} \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) \beta \epsilon \right) \\ \leq & \mathcal{O}\left(\frac{H^2}{\tilde{h}} \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) \log[KH\mathcal{N}_{\mathcal{F}\cup\mathcal{G}}(1/K)/\delta] \frac{1}{\epsilon} \right). \end{aligned}$$

Due to the fact that we run the algorithm for $K = \mathcal{O}(|\mathcal{T}^\epsilon|)$ phases, we know that one of the following two holds:

- We finish the exploration.
- We expected to run a policy that had no-contribute to increase $\tilde{h}$, and therefore it is ϵ-optimal policy.

Thus, in both cases we have policy that is ϵ-optimal. □

6.3 Sections' Proofs

Lemmas from GOLF without proof In this section, we quote several key lemmas from the GOLF paper, which will be used later in our proofs.

Lemma 77. *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 4, then with probability at least $1-\delta$, for all $(k,h) \in [K]\times[H]$, we have*

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

(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 in the confidence set, with high probability. As a result, the selected value function f^k, in each iteration k, shall be an upper bound of $Q^\star$, with high probability.

Lemma 78. *Under the same condition of Lemma 77, with probability at least $1-\delta$, we have $Q^\star \in \mathcal{B}^k$ for all $k \in [K]$.*

The proof of Lemma 77 and 78 relies on standard martingale concentration (e.g., Freedman's inequality) and can be found in Appendix D.3 in Jin et al. (2021a).

The next lemmas are highly-relevant to our work so we add their proofs for completeness.

Proof of Lemma 70 The proof in this subsection basically follows the same arguments as in Appendix C of Russo and Van Roy (2013), and as in Appendix D of Jin et al. (2021a) We firstly prove the following proposition which bounds the number of times $|\mathbb{E}_{\mu_t}[\phi_t]|$ can exceed a certain threshold.

Proposition 79. *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 $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2 \le \beta$. Then for all $k \in [K]$,*

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

***Proof** of Proposition 79.* 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 $\sum_{t=1}^{\ell}(\mathbb{E}_{\nu_t}[\phi_k])^2 \ge \epsilon^2$. It implies that if μ_k is ϵ-dependent on L disjoint subsequences in $\{\mu_1, \ldots, \mu_{k-1}\}$, we have

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

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

Now we want to show that for any sequence $\{\nu_1, \ldots, \nu_T\} \subseteq \Pi$, there exists $j \in [T]$ such that ν_j is ϵ-dependent on at least $L = \lceil (T-1)/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\rceil$ 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 $j = L\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) + 1 \le T$.

Fix $k \in [K]$ and let $\{\nu_1, \ldots, \nu_T\}$ 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 $j \in [T]$ such that ν_j is ϵ-dependent on at least $(T/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)) - 1$ disjoint subsequences of $\{\nu_1, \ldots, \nu_{j-1}\}$. Therefore, we have $T/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon) - 1 \le \beta/\epsilon^2$ which results in

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

and completes the proof. □

Proof of Lemma 74

Proof. Recall Proposition 79:
For given $h \in [H]$, given a function class Φ defined on $\mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{\phi_{h,k}\}_{k=1}^K \subset \Phi$ and $\{\mu_{h,k}\}_{k=1}^K \subset \Pi$ satisfy that for all $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_{h,t}}[\phi_{h,t}])^2 \le \beta$. Then for all $k \in [K]$,

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

Recall the definition of h_t (definition 72) and sum over $h \in [H]$,

$$\sum_{t=1}^{k} h_t = \sum_{h=1}^{H}\sum_{t=1}^{k} \mathbf{1}\{|\mathbb{E}_{\mu_{h,k}}[\phi_{h,k}]| > \epsilon\} \le H(\frac{\beta}{\epsilon^2}+1)\dim_{\mathrm{DE}}(\Phi,\Pi,\epsilon),$$

which completes the proof. □

Proof of Lemma 75

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

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

For $t \in [k]$, we want to prove that if $e_t > \omega$, then we have $e_t \le \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Assume $t \in [k]$ satisfies $e_t > \omega$. Then there exists α such that $e_t > \alpha \ge \omega$. By Proposition 79, we have

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

which implies:

$$\alpha \le \sqrt{\frac{d\beta}{t-d}}. \tag{45}$$

Besides, recall $e_t \le C$, so we have $e_t \le \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Also, we use Proposition 79 from the other way,

$$\sum_{t=1}^{k}\mathbf{1}\{|\mathbb{E}_{\mu_{h,k}}[\phi_{h,k}]| > \omega\} = \mathbf{1}\{e_t > \omega\} \le t^{\omega}.$$

Therefore, for $K_\omega = \min\{k, t^\omega\}$, we have

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

which completes the proof. □

7 Opera Improving Sample Complexity

7.1 Introduction

Reinforcement Learning (RL) has made significant strides in recent years, particularly with the development of algorithms capable of efficiently handling large state and action spaces. A notable contribution to this field is the framework proposed by Chen et al. in their paper "A General Framework for Sample-Efficient Function Approximation in Reinforcement Learning" Chen et al. (2022). This framework unifies model-based and model-free approaches through the Admissible Bellman Characterization (ABC) class and introduces the functional eluder dimension 41, a novel complexity measure that generalizes the eluder dimension from bandit literature.

The ABC class encompasses nearly all Markov Decision Process (MDP) models considered tractable for RL, providing a versatile platform for both theoretical analysis and practical algorithm design. The functional eluder dimension serves as a critical tool in assessing the sample complexity of RL algorithms, which is pivotal for developing methods that can efficiently learn in high-dimensional spaces.

Building on this foundation, the **OP**timization-based **E**xplo**R**ation with **A**pproximation (OPERA) algorithm proposed by Chen et al. demonstrates superior sample efficiency and improved regret bounds.The OPERA algorithm leverages an optimization-based approach to exploration, characterized by decomposable structural properties in the estimation functions, enabling more effective navigation through large state-action spaces.

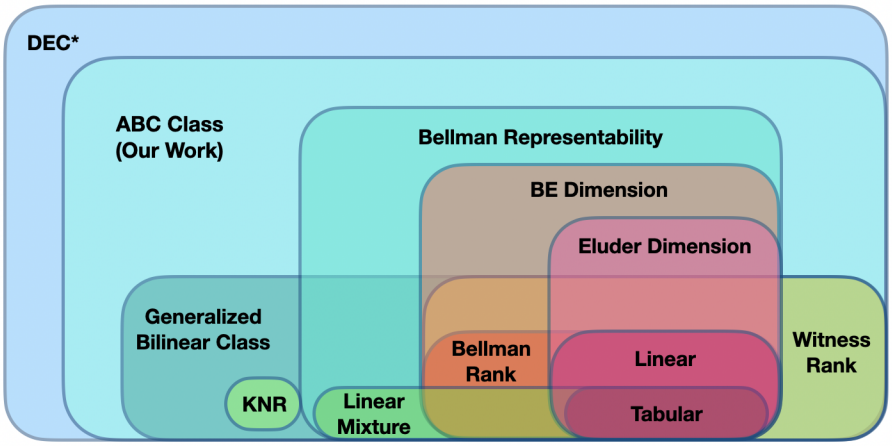

Figure 1: **Venn-Diagram Visualization of Prevailing Sample-Efficient RL Classes.** As by far the richest concept, the DEC framework is both a necessary and sufficient condition for sample-efficient interactive learning. BE dimension is a rich class that subsumes both low Bellman rank and low eluder 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 $Q^\star$ state-action aggregation and deterministic linear $Q^\star$ MDP models, which neither Bilinear Class nor our ABC class captures.

7.1.1 Assumptions

In this section in compare to the previous section the relevant assumptions are realizability Assumption 8, and a finite ϵ-cover (10). In addition we will use a Lipschitz assumption (Assumption 83) that will define later. We will assume that our hypothesis class has such function $\mathbf{D}$, and a bounded Functional Eluder Dimension. For that, recall Functional Eluder Dimension definition 41.

Definition 80 (Functional Eluder Dimension)**.** *For a given hypothesis class $\mathcal{F}$ and a function $\mathbf{D}$ defined on $\mathcal{F} \times \mathcal{F}$, the functional eluder dimension (FE dimension) $\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon)$ is the length of the longest sequence $f_1, \ldots, f_n \in \mathcal{F}$ satisfying for all $t \leq n$, exists $\epsilon_t \geq \epsilon$, such that there exists $g \in \mathcal{F}$ holding $\sqrt{\sum_{i=1}^{t-1} (\mathbf{D}(g, f_i))^2} \leq \epsilon_t$ while $|\mathbf{D}(g, f_t)| > \epsilon_t$.*

Such function $\mathbf{D}$ is dubbed as the *coupling function*.

7.2 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 (8), 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 Foster et al. (2023); Du et al. (2021). The *discriminator* v originates from the function class the Integral Probability MetriScal (IPM) Müller (1997) is taken with respect to (as a metric between two distributions), and is also used in the definition of Witness rank Sun et al. (2019). We use a coupling function $\mathbf{D}_{h,f^\star}(f, g)$ defined on $\mathcal{F} \times \mathcal{F}$ to characterize the interaction between two hypotheses $f, g \in \mathcal{F}$. The subscript $f^\star$ 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 81 (Admissible Bellman Characterization)**.** *Given an MDP M, two hypothesis classes $\mathcal{F}, \mathcal{G}$ satisfying the realizability assumption and $\mathcal{F} \subset \mathcal{G}$, an estimation function $\ell_{h,f'} : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \mathcal{F} \times \mathcal{G} \times \mathcal{V} \to \mathbb{R}^{d_s}$ and a constant $\kappa \in (0, 1]$, we say that $\mathbf{D}$ is an* admissible Bellman characterization *of $(M, \mathcal{F}, \mathcal{G}, \ell)$ if the following conditions hold:*

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

$$\max_{v \in \mathcal{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_g} \|\mathbb{E}_{s_{h+1}} [\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h]\|^2 \geq (\mathbf{D}_{h,f^\star}(f, g))^2 .$$

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

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

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

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

Definition 82 (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* $\mathcal{T}(\cdot) : \mathcal{F} \to \mathcal{G}$*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 [1]

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

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

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

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

Assumption 83 (Lipschitz Estimation Function)**.** *There exists a* $L > 0$*, under a well-defined metric* ρ*, 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*,*

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

Note that we have omitted the subscript h of hypotheses in Assumption 83 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 83, the covering number of the induced estimation function class $\mathcal{L}$ can be upper bounded as

$$N_{\mathcal{L}}(\epsilon) \leq N_{\mathcal{F}}^2(\frac{\epsilon}{4L}) N_{\mathcal{G}}(\frac{\epsilon}{4L}) N_{\mathcal{V}}(\frac{\epsilon}{4L}), \tag{46}$$

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.

7.3 Opera Algorithm

We first present the *OPtimization-based ExploRation with Approximation (OPERA)* algorithm in Algorithm 5, which finds an ϵ-optimal policy in polynomial time. Following earlier algorithmic art in the same vein e.g., GOLF Jin et al. (2021a), 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 Du et al. (2021).

Pertinent to the constrained optimization subproblem in Eq. (47) of Algorithm 5, we adopt the confidence region based on a general DEF, extending the Bellman-error-based confidence region used in Jin et al. (2021a). As a result of such an extension, our algorithm can deal with more complex models such as low Witness rank and KNR. We avoid unnecessary complications by forgoing the discussion on the computational efficiency of the optimization subproblem, aligning with recent literature on RL theory with general function approximations.

[1] The decomposability item (i) in Definition 82 directly implies that a Generalized Completeness condition similar to Completeness assumption in GOLF holds(Assumption 14 of Jin et al. (2021a)).

Algorithm 5 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 $\operatorname{argmax}_{f \in \mathcal{F}} Q_{f,1}(s_1, \pi_f(s_1))$ subject to

$$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \|\ell_{h,f^i}(o_h^i, f_{h+1}, f_h, v)\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \|\ell_{h,f^i}(o_h^i, f_{h+1}, g_h, v)\|^2 \right\} \leq \beta \quad \text{for all } h \in [H] \tag{47}$$

For any $h \in [H]$, collect tuple (r_h, s_h, a_h, s_{h+1}) by executing $s_h, a_h \sim \pi^t$

Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(r_h, s_h, a_h, s_{h+1})\}$

Output: π_{out} uniformly sampled from $\{\pi^t\}_{t=1}^T$

7.4 Regret Bounds

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

Theorem 84 (Regret Bound of OPERA)**.** *For an MDP M, hypothesis classes $\mathcal{F}, \mathcal{G}$, a Decomposable Estimation Function ℓ satisfying Assumption 83, 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 5, where $\mathcal{N}_{\mathcal{L}}$ defined as in Equation 46. Then with probability at least $1-\delta$, the regret is upper bounded by*

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

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

Step 1: Feasibility of $f^\star$. First, we show that the optimal hypothesis $f^\star$ lies within the confidence region defined by Eq. (47) with high probability. Given parameters $\rho > 0$ and $\delta > 0$, choosing $\beta = c(\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta) + T\rho)$ ensures that $f^\star$ stays within the defined confidence region.

Step 2: Policy Loss Decomposition. We upper bound the regret by the summation of Bellman errors using a policy loss decomposition lemma. This lemma helps in breaking down the policy loss into manageable components.

Step 3: Small ABC Value in the Confidence Region. We control the cumulative square of the Admissible Bellman Characterization (ABC) function. This ensures that the cumulative ABC value stays within the confidence region, using Freedman's inequality.

Step 4: Bounding the Cumulative Bellman Error by Functional Eluder Dimension. We translate the upper bound of the cumulative squared ABC to an upper bound of the cumulative ABC, using properties of the functional eluder dimension.

Step 5: Combining Everything. Finally, we combine the regret bound decomposition, the cumulative ABC bound, and the Bellman dominance property to derive our final regret guarantee. With high probability, the regret is bounded as:

$$\sum_{t=1}^{T} V_1^\star(s_1) - V_1^{\pi^t}(s_1) \leq \mathcal{O}\left(\frac{H}{\kappa}\sqrt{T \cdot \dim_{\text{FE}}\left(\mathcal{F}, \mathbf{D}, \sqrt{1/T}\right) \log\left(TH\mathcal{N}_{\mathcal{L}}(1/T)/\delta\right)}\right).$$

7.5 ϵ-Optimality: Improved Opera Sample Complexity and Regret Analysis

We present an advanced analysis of the Opera algorithm, aiming to refine the regret guarantee to achieve an ϵ-optimal policy, providing enhancement in performance evaluation and robustness of the algorithm. We are also improving the guarantee to get such policy. We will work on all the layers simultaneously, and we will refine the Definition 72 of corrected layers.

Definition 85. *Let $\epsilon > 0$, $t \in [K]$, For given decoupling function $\mathbf{D}$, we define h_t to be the amount of layers $h \in [H]$ where $|\mathbf{D}(f_h^t, g_h^t)| \geq \epsilon$. In other words, $h_t^\epsilon = \sum_{h\in[H]} \mathbf{I}\{|\mathbf{D}(f_h^t, g_h^t)| > \epsilon\}$.*

Here as well we won't get a large error many times. Bounding the occurs of it by t^ϵ and with Proposition 93 we would bound this amount and get that:

$$t^\epsilon = \left(\frac{\beta}{\epsilon^2} + 1\right) \times \dim_{\text{FE}}\left(\mathcal{F}, \mathbf{D}, \epsilon\right). \tag{48}$$

Now bounding the amount of errors we have:

Lemma 86. *Let $\epsilon > 0$, and given coupling function $\mathbf{D}$, we have $\sum_{t=1}^K h_t^\epsilon \leq H \times \left(\frac{\beta}{\epsilon^2}+1\right) \times \dim_{FE}\left(\mathcal{F}, \mathbf{D}, \epsilon\right)$.*

As we saw in the previous section in Lemma 70, the cumulative error can be separate to two parts, where we have a big error, and when we got a small error. This is relevant for this section also and we can see it in Lemma 92, in the next subsection. Therefore, We are very interested to bound as low as possible the amount of episode with big error, and keep the error low as possible. Here Lemma 86 comes to assist us, we will bound the number of episode that have big error. We will adapt the Definition 73 of $\mathcal{T}^\epsilon$, we will denote $\tilde{h} = \frac{1}{|\mathcal{T}|} \times \sum_{t=1}^K h_t$, the average amount of layers that have a large exception, which implies in this case:

$$|\mathcal{T}^\epsilon| \leq \mathcal{O}\left(\frac{H}{\tilde{h}}\frac{\beta}{\epsilon^2} \cdot \dim_{\text{FE}}\left(\mathcal{F}, \mathbf{D}, \epsilon\right)\right). \tag{49}$$

Bounding the Cumulative Bellman Error by Functional Eluder Dimension. Using the upper bound that we have from Step 3 (in subsection 7.4),

$$\sum_{i=1}^{t-1} \left(\mathbf{D}_{h,f^*}(f^t, f^i)\right)^2 \leq \mathcal{O}(\beta), \tag{50}$$

which will present in Lemma 91. We aim to translate the upper bound of the cumulative squared ABC at (f^t, f^i) in Eq. (54) to an upper bound of the cumulative ABC at (f^t, f^t). The following Lemma 87 is adapted from Lemma 75, which base on Lemma 70 both from Section 6, and similar to Lemma 58 in Section 5. Lemma 87 combine the previous properties we saw using Eq. (49) and controls the sum of ABC functions by properties of the functional eluder dimension.

Lemma 87. *For a hypothesis class $\mathcal{F}$ and a given coupling function $\mathbf{D}(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathcal{R}$ with bounded image space $|\mathbf{D}(\cdot,\cdot)| \leq C$. For any pair of sequences $\{f_t\}_{t\in[T]}, \{g_t\}_{t\in[T]} \subseteq \mathcal{F}$ satisfying for all $t \in [T]$, $\sum_{i=1}^{t-1}(\mathbf{D}(f_t, g_i))^2 \leq \beta$, the following inequality holds for all $t \in [T]$ and $\omega > 0$, and t^ω (Eq. 48) :*

$$\sum_{h=1}^{H}\sum_{i=1}^{t} |\mathbf{D}(f_i, g_i)| \leq \mathcal{O}\Big(H\sqrt{\dim_{FE}(\mathcal{F}, \mathbf{D}, \omega)\beta \min\{t, t^\omega\}} + HC \cdot \min\{t, \dim_{FE}(\mathcal{F}, \mathbf{D}, \omega)\} + Ht\omega\Big).$$

And for big enough k, and with t^ω from (48), our regret, as function of ω, is

$$\sum_{h=1}^{H}\sum_{t=1}^{K} |\mathbf{D}(f_i, g_i)| \leq \mathcal{O}\left(H \dim_{\mathrm{FE}}(\mathcal{F}, \mathbf{D}, \epsilon)\, \beta \frac{1}{\omega} + HK\omega\right).$$

Therefore, we get our main result, the regret bound is as follow:

Theorem 88 (Effective Horizon Regret Bound For General Eluder with PAC Guarantee)**.** *By setting $\omega = \epsilon$ and $K = \mathcal{O}(|\mathcal{T}|)$, we have , $\exists t \in [K]$, s.t. $BE(\pi^t) \leq \epsilon$,$\forall h \in [H]$.*
Furthermore, the total regret of the algorithm:

$$\mathcal{R}(K) = \sum_{h=1}^{H}\sum_{t=1}^{K} |\mathbb{E}_{\mu_{\{h,t\}}}[\phi_t]| \leq \mathcal{O}\left(\frac{H^2}{\tilde{h}\kappa} \dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon)\, \beta \frac{1}{\epsilon}\right). \tag{51}$$

Proof. In Lemma (90) we bound the regret as follows,

$$\mathcal{R}(K) = \sum_{k=1}^{K}\left(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\right) \leq \sum_{k=1}^{K}\left(\max_a f_1^k(s_1, a) - V_1^{\pi^k}(s_1)\right) = \sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h).$$

With the Bellman dominance property Definition 81 property (ii), we have that:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^t, \pi^t, h) \leq \frac{1}{\kappa}\sum_{h=1}^{H}\sum_{k=1}^{K} |\mathbf{D}(f_i, g_i)|$$

Combine the regret bound decomposition argument Lemma 87, the cumulative ABC bound, derive our regret guarantee with probability at least $1 - \delta$,

$$\begin{aligned}
\mathcal{R}(K) &\leq \frac{1}{\kappa}\sum_{h=1}^{H}\sum_{k=1}^{K} |\mathbf{D}(f_i, g_i)| \\
&\leq \frac{1}{\kappa}\mathcal{O}\left(H \dim_{\mathrm{FE}}(\mathcal{F}, \mathbf{D}, \epsilon)\, \beta \frac{1}{\epsilon} + H\frac{H}{\tilde{h}\epsilon^2} \dim_{\mathrm{FE}}(\mathcal{F}, \mathbf{D}, \epsilon)\, \beta\epsilon\right) \\
&\leq \mathcal{O}\left(\frac{H^2}{\tilde{h}\kappa} \dim_{\mathrm{FE}}(\mathcal{F}, \mathbf{D}, \epsilon) \log[KH\mathcal{N}_{\mathcal{L}}(1/K)/\delta] \frac{1}{\epsilon}\right).
\end{aligned}$$

Due to the fact that we run the algorithm for $K = \mathcal{O}(|\mathcal{T}^\epsilon|)$ phases, we know that one of the following two holds:

- We finish the exploration.
- We expected to run a policy that had no-contribute to increase $\tilde{h}$, and therefore it is ϵ-optimal policy.

Thus, in both cases we have policy that is ϵ-optimal. □

7.6 Sections' Proofs

In this subsection we will shows the proofs for the lemmas we used in order to prove our main theorem. The proof steps generalizes those who presented previously in 6.3 but admits general DEF and ABCs.

Lemmas from OPERA without proof In this section, we quote several key lemmas from the GOLF paper, which will be used later in our proofs. Our first lemma is to show that the optimal hypothesis $f^\star$ lies within the confidence region defined by Eq. (47) with high probability:

Lemma 89 (Feasibility of $f^\star$ - Lemma 21 in Chen et al. 2022)**.** *In Algorithm 5, 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 $1-\delta$, $f^\star$ satisfies for any $t \in [T]$:*

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

Lemma 89 shows that at each round of updates the optimal hypothesis $f^\star$ stays in the confidence region depicted by Eq. (47) with radius $\mathcal{O}(\beta)$. We refer to the proof of Lemma 89 to Lemma 21 in appendix F.2 in Chen et al. 2022. Together with the optimism procedure Line 3 of Algorithm 5 implies an upper bound of $V_1^\star(s_1) - V_1^{\pi^t}(s_1)$ with probability at least $1-\delta$ as follows:

$$V_1^\star(s_1) - V_1^{\pi^t}(s_1) \le V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1). \tag{52}$$

The second lemma we will present here is a upper bound the regret by the summation of Bellman errors. We apply the policy loss decomposition lemma in Jiang et al. (2017).

Lemma 90 (Regret Decomposition - Lemma 1 in Jiang et al. 2017)**.** $\forall f \in \mathcal{H}$,

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

Combining Lemma 90 with Eq. (52) we have the following:

$$V_1^\star(s_1) - V_1^{\pi^t}(s_1) \le V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1) = \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim\pi^t}\left[Q_{h,f^t}(s_h,a_h) - r_h - V_{h+1,f^t}(s_{h+1})\right]. \tag{53}$$

The third (and last) lemma we only quoting 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. (47), we arrive at the following Lemma 91:

Lemma 91 (Confidence Bound - Lemma 23 in Chen et al. 2022)**.** *In Algorithm 5, 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*

$$\sum_{i=1}^{t-1} \left(\mathbf{D}_{h,f^\star}(f^t, f^i)\right)^2 \le \mathcal{O}(\beta). \tag{54}$$

The proof of Lemma 91 makes use of Freedman's inequality (the precise version as in Agarwal et al. (2014)) and we refer for Lemma 23 in Appendix F.1 in Chen et al. 2022.

Lemma 92. *For a hypothesis class $\mathcal{F}$ and a given coupling function $\mathbf{D}(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathcal{R}$ with bounded image space $|\mathbf{D}(\cdot,\cdot)| \leq C$. For any pair of sequences $\{f_t\}_{t\in[T]}, \{g_t\}_{t\in[T]} \subseteq \mathcal{F}$ satisfying for all $t \in [T]$, $\sum_{i=1}^{t-1}(\mathbf{D}(f_t, g_i))^2 \leq \beta$, the following inequality holds for all $t \in [T]$ and $\omega > 0$:*

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

The proof defer for Lemma 24 in Chen et al. 2022.

proof for Lemma 87

Proof of Lemma 87. The proof basically follows Appendix C of Russo and Van Roy (2013) and Appendix D of Jin et al. (2021a). We first prove that for all $t \in [T]$,

Proposition 93. *For a hypothesis class $\mathcal{F}$ and a given coupling function $\mathbf{D}(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathcal{R}$. For any pair of sequences $\{f_t\}_{t\in[T]}, \{g_t\}_{t\in[T]} \subseteq \mathcal{F}$ satisfying for all $t \in [T]$, $\sum_{k=1}^{t-1}(\mathbf{D}(f_t, g_k))^2 \leq \beta$, Then for all $k \in [K]$,*

$$\sum_{k=1}^{t} \mathbf{1}(|\mathbf{D}(f_k, g_k)| > \epsilon) \leq (\beta/\epsilon^2 + 1) \dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon). \tag{55}$$

Proof. Let $m := \sum_{k=1}^{t} \mathbf{1}(|\mathbf{D}(f_k, g_k)| > \epsilon)$, then there exists $\{s_1, \ldots, s_m\}$ which is a subsequence of $[t]$ such that $\mathbf{D}(f_{s_1}, g_{s_1}), \ldots, \mathbf{D}(f_{s_m}, g_{s_m}) > \epsilon$.

We first show that for the sequence $\{f_{s_1}, \ldots, f_{s_m}\} \subseteq \mathcal{F}$, there exists $j \in [m]$ such that f_{s_j} is ϵ-independent on at least $L = \lceil (m-1)/\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon) \rceil$ disjoint sequences in $\{f_{s_1}, \ldots, f_{s_{j-1}}\}$ (Russo and Van Roy, 2013). We will prove this by following procedure. Starting with singleton sequences $B_1 = \{f_{s_1}\}, \ldots, B_L = \{f_{s_L}\}$ and $j = L + 1$. For each j, if f_{s_j} is ϵ-dependent on $B_1, \ldots, B_L$ we already achieved our goal and the process stops. Otherwise, there exist $i \in [L]$ such that f_{s_j} is ϵ-dependent of B_i and update $B_i = B_i \cup \{f_{s_j}\}$. Then we add increment j by 1 and continue the process. By the definition of FE dimension, the cardinally of each set $B_1, \ldots, B_L$ cannot larger than $\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon)$ at any point in this process. Therefore, by pigeonhole principle the process stops by step $j = L \dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon) + 1 \leq m$.

Therefore, we have proved that there exists j such that $|\mathbf{D}(f_{s_j}, g_{s_j})| > \epsilon$ and f_{s_j} is ϵ-independent with at least $L = \lceil (m-1)/\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon) \rceil$ disjoint sequences in $\{f_{s_1}, \ldots, f_{s_{j-1}}\}$. For each of the sequences $\{\hat{f}_1, \ldots, \hat{f}_l\}$, by definition of the FE dimension in Definition 37 we have that

$$\sum_{k=1}^{l} \big(\mathbf{D}(\hat{f}_k, g_{s_j})\big)^2 \geq \epsilon^2. \tag{56}$$

Summing all of bounds (56) for L disjoint sequences together we have that

$$\sum_{k=1}^{s_j-1} \big(\mathbf{D}(f_t, g_{s_j})\big)^2 \geq L\epsilon^2 = \lceil (m-1)/\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon) \rceil \cdot \epsilon^2. \tag{57}$$

The left hand side of (57) can be upper bounded by β^2 due to the condition of lemma. Therefore, we have proved that $\beta^2 \geq \lceil (m-1)/\dim_{FE}(\mathcal{F}, \mathbf{D}, \epsilon) \rceil \cdot \epsilon^2$ which completes the proof of (93). □

Now let $d = \dim_{FE}(\mathcal{F}, \mathbf{D}, \omega)$ and sort $|\mathbf{D}(f_1, g_1)|, \ldots, |\mathbf{D}(f_t, g_t)|$ in a nonincreasing order, denoted by $e_1, \ldots, e_t$. Then we have that

$$\sum_{k=1}^{t} |\mathbf{D}(f_k, g_k)| = \sum_{k=1}^{t} e_k = \sum_{k=1}^{t} e_k \mathbf{1}(e_k \le \omega) + \sum_{i=1}^{t} e_k \mathbf{1}(e_k > \omega) \le t\omega + \sum_{i=1}^{t} e_k \mathbf{1}(e_k > \omega). \tag{58}$$

For $k \in [t]$, we want to give an upper bound for those $e_k \mathbf{1}(e_k > \omega)$. Assume $e_k > \omega$, then for any α such that $e_k > \alpha \ge \omega$, by (55), we have that

$$k \le \sum_{i=1}^{t} \mathbf{1}(e_i > \omega) \le (\beta/\alpha^2 + 1) \dim_{FE}(\mathcal{F}, \mathbf{D}, \alpha) \le (\beta/\alpha^2 + 1)d,$$

which implies that $\alpha \le \sqrt{d\beta/(k-d)}$. Taking the limit $\alpha \to e_k^-$, we have that $e_k \le \min\{\sqrt{d\beta/(k-d)}, C\}$. Also, we use Proposition 93 from the other way,

$$\sum_{t=1}^{k} \mathbf{1}\{|\mathbf{D}(f_t, g_t)| > \omega\} = \mathbf{1}\{e_t > \omega\} \le t^{\omega}.$$

Therefore, for $K_\omega = \min\{k, t^\omega\}$, we have

$$\begin{aligned} \sum_{t=1}^{k} e_t \mathbf{1}\{e_t > \omega\} &= \sum_{t=1}^{K_\omega} e_t \mathbf{1}\{e_t > \omega\} \le \min\{d, K_\omega\} C + \sum_{t=d+1}^{K_\omega} \sqrt{\frac{d\beta}{t-d}} \\ &\le \min\{d, K_\omega\} C + \sqrt{d\beta} \int_0^{K_\omega} \frac{1}{\sqrt{t}} dt \le \min\{d, K_\omega\} C + 2\sqrt{d\beta K_\omega} \\ &= \min\{d, k\} C + 2\sqrt{d\beta \min\{k, t^\omega\}}, \end{aligned} \tag{59}$$

which completes the proof. Plugging (59) into (58) completes the proof. □

proof for Lemma 86

Proof. Recall Proposition 93:
For a hypothesis class $\mathcal{F}$ and a given coupling function $\mathbf{D}(\cdot, \cdot) : \mathcal{F} \times \mathcal{F} \to \mathcal{R}$. For any pair of sequences $\{f_t\}_{t\in[T]}, \{g_t\}_{t\in[T]} \subseteq \mathcal{F}$ satisfying for all $t \in [T]$, $\sum_{k=1}^{t-1} (\mathbf{D}(f_t, g_k))^2 \le \beta$, Then for all $k \in [K]$,

$$\sum_{t=1}^{k} \mathbf{1}\{|\mathbf{D}(f_t, g_t)| > \epsilon\} \le (\frac{\beta}{\epsilon^2} + 1) \dim_{\text{FE}}(\mathcal{F}, \mathbf{D}, \epsilon).$$

Recall the definition of h_t (definition 72) and sum over $h \in [H]$,

$$\sum_{t=1}^{k} h_t = \sum_{h=1}^{H} \sum_{t=1}^{k} \mathbf{1}\{|\mathbf{D}(f_t, g_t)| > \epsilon\} \le H(\frac{\beta}{\epsilon^2} + 1) \dim_{\text{FE}}(\mathcal{F}, \mathbf{D}, \epsilon),$$

which completes the proof. □

References

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.

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

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *WWW*, pages 167–176. ACM, 2018. URL http://dblp.uni-trier.de/db/conf/www/www2018.html#ZhengZZXY0L18.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey, 2021. URL https://arxiv.org/abs/2002.00444.

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. -, 2021. URL https://rltheorybook.github.io/.

Shie Mannor, Yishay Mansour, and Aviv Tamar. *Reinforcement Learning: Foundations*. -, 2022. URL https://sites.google.com/view/rlfoundations/home.

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

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*, pages 996–1002, 1999.

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

Zixiang Chen, Chris Junchi Li, Angela Yuan, Quanquan Gu, and Michael I. Jordan. A general framework for sample-efficient function approximation in reinforcement learning, 2022.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms, 2021a.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo*q*l: Towards optimal regret in model-free rl with nonlinear function approximation, 2022.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221, 2008.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. URL https://arxiv.org/abs/1911.10635.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, October 2019. ISSN 1573-7454. doi: 10.1007/s10458-019-09421-1. URL http://dx.doi.org/10.1007/s10458-019-09421-1.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002. URL http://jmlr.org/papers/v3/brafman02a.html.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 881–888, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143955. URL https://doi.org/10.1145/1143844.1143955.

Sham M. Kakade, Michael J. Kearns, and John Langford. Exploration in metric state spaces. In Tom Fawcett and Nina Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 306–312. AAAI Press, 2003. URL http://www.aaai.org/Library/ICML/2003/icml03-042.php.

Nicholas K. Jong and Peter Stone. Model-based exploration in continuous state spaces. In Ian Miguel and Wheeler Ruml, editors, *Abstraction, Reformulation, and Approximation, 7th International Symposium, SARA 2007, Whistler, Canada, July 18-21, 2007, Proceedings*, volume 4612 of *Lecture Notes in Computer Science*, pages 258–272. Springer, 2007. doi: 10.1007/978-3-540-73580-9_21. URL https://doi.org/10.1007/978-3-540-73580-9_21.

Jason Pazis and Ronald Parr. Efficient pac-optimal exploration in concurrent, continuous state mdps with delayed updates. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 1977–1985. AAAI Press, 2016. doi: 10.1609/AAAI.V30I1.10307. URL https://doi.org/10.1609/aaai.v30i1.10307.

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

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl, 2021. URL https://arxiv.org/abs/2103.10897.

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

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent rl, 2021b. URL https://arxiv.org/abs/2110.14555.

Dan Qiao and Yu-Xiang Wang. Near-optimal reinforcement learning with self-play under adaptivity constraints, 2024. URL https://arxiv.org/abs/2402.01111.

R Matthew Kretchmar. Parallel reinforcement learning. In *The 6th World Conference on Systemics, Cybernetics, and Informatics*, page 4. Citeseer, 2002.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International conference on machine learning*, pages 5872–5881. PMLR, 2018.

Hankz Hankui Zhuo, Wenfeng Feng, Yufeng Lin, Qian Xu, and Qiang Yang. Federated deep reinforcement learning, 2020. URL https://arxiv.org/abs/1901.08277.

Abhimanyu Dubey and Alex Pentland. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021a.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15007–15049. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mao22a.html.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence & Robotics*, 2021. doi: 10.20517/ir.2021.02. URL http://dx.doi.org/10.20517/ir.2021.02.

Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee, 2022. URL https://arxiv.org/abs/2110.14074.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent deep deterministic policy gradient. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

Tianxing Yu, Shuangqing Wang, and Haibo He. Policy learning in networked multi-agent systems with imperfect communication. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11971–11981. PMLR, 2021.

Filip Kuba, Krzysztof Choromanski, and Ryota Nakae. Heterogeneous agent reinforcement learning: A review and perspectives. *arXiv preprint arXiv:2201.09967*, 2022.

Xiao Liu, Weiyan Xu, and Peng Zhang. Decentralized reinforcement learning with networked agents: Theory and applications. In *Proceedings of the 39th International Conference on Machine Learning*, pages 4862–4871. PMLR, 2022.

Chi Jin, Kaiqing Zhang, Weiyan Xu, Zhuoran Yang, and Tamer Basar. V-learning: Decentralized multi-agent reinforcement learning with communication. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Abhishek Dubey and Alex Pentland. Heterogeneous parallel linear mdps: Efficient algorithms and lower bounds. *arXiv preprint arXiv:2110.14447*, 2021b.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Cooperative multi-agent reinforcement learning: Asynchronous communication and linear function approximation, 2023.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Cooperative online learning in stochastic and adversarial mdps, 2022.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation, 2023.

Nuoya Xiong, Zhaoran Wang, and Zhuoran Yang. A general framework for sequential decision-making under adaptivity constraints, 2023.

Michael Todd. On minimum volume ellipsoids containing part of a given ellipsoid. *Mathematics of Operations Research - MOR*, 7:253–261, 05 1982. doi: 10.1287/moor.7.2.253.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient?, 2018.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making, 2023. URL https://arxiv.org/abs/2112.13487.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429 – 443, 1997. URL https://api.semanticscholar.org/CorpusID:124648603.

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, 2019. URL https://arxiv.org/abs/1811.08540.

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

תמצית

קירוב פונקציות הפך לנושא חשוב יותר ויותר בלמידה מחיזוקים, בזכות ההתקדמות בלמידת מכונה והצורך להתמודד עם מחלקות גדולות של פונקציות לא ליניאריות. בשל ההתקדמות זאת, הדרישה לאלגוריתם הפותר את האתגר הזה באופן יעיל גברה.

עבודה זו בוחנת דרכים לשפר את היעילות של למידת חיזוק בתרחישים נפוצים תוך שמירה על קיום חסמים תיאורתים.

עבודה זו תורמת ומחדשת באמצעות הצגת המושג *האופק האפקטיבי*, יתרה על כך מרחיבה את ממד האלודר הנודע *ומגדירה את ממד האלודר הצפוי*. בנוסף משתמשים במערכות מרובות סוכנים ובמודלים גנרטיביים.

התוצאות בעבודה משפרות את העמידות בפני גודל המחקלות, מצמצמות את מורכבות הדגימה, ומציגות חסמי חרטה עם תלות פוטנציאלית נמוכה יותר בממד ובמקדם האופק.

תרומות אלו מקדמות את למידה מחיזוקים צעד נוסף לקראת יישום פרקטי, והופך אותה ליעילה וישימה יותר בתרחישים מציאותיים.

תיאוריה בלמידה מחיזוקים - על קירוב פונקציות בריבוי סוכנים ועל מחלקות אפקטיביות

עבודת גמר לתואר מוסמך במדעים מוגשת כחלק מהדרישות לשם קבלת התואר "מוסמך אוניבריסטה" באוניברסיטת תל אביב.

מוגש על ידי

דולב דנינו

עבודת המחקר בוצעה בהנחיית

פרופסור ישי מנצור

אלול תשפ"ד – ספטמבר 2024.

הפקולטה למדעים מדויקים על שם ריימונד ובברלי סאקלר אוניברסיטת תל אביב