A Sufficient Condition of Sample-Efficient Reinforcement Learning with General Function Approximation

by

Wei Xiong

A Thesis Submitted to

The Hong Kong University of Science and Technology

in Partial Fulfillment of the Requirements for

the Degree of Master of Philosophy

in Mathematics

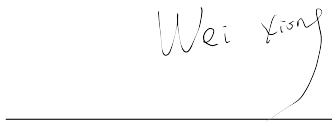
29 July 2023, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.



Wei XIONG

29 July 2023

Sample-Efficient Reinforcement Learning with General Function Approximation

by

Wei XIONG

This is to certify that I have examined the above MPhil thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Tong ZHANG

Prof. Kun XU
Head of Department

Department of Mathematics

29 July 2023

Acknowledgment

Writing a dissertation at the end of the master's study provided me with an excellent opportunity to rethink the past two years. Research is a long journey for years, with countless challenges and failures. However, I always consider myself fortunate to have received invaluable help and support from my amazing research mentors, reliable co-authors, and lovely friends.

My first and deepest thanks go to my supervisor Tong Zhang, for his invaluable guidance and support throughout the journey. Tong led me into the fabulous world of learning theory from nowhere. I am particularly grateful that Tong shared his excellent book on learning theory with me when I was a fresh student so that I could have an overview of the advances in this area and be equipped with a wide range of research techniques in the early stage of my career. But what I have learned from him is far beyond the specific research techniques. Tong has influenced me in the past two years with his patient guidance, persistence in science, and great academic ethics. Tong basically satisfies all the factors in my mind that make a good advisor: he has a great depth and breadth of knowledge, innovative research ideas, and excellent research taste. He has been kind and patient in guiding me during my study and has provided endless support in both life and career. Tong definitely sets a good model as both researcher and advisor for me in my future career. I am deeply grateful for his mentorship and support, and I hope to continue to learn from him in the future.

I would like to express my gratitude to the many amazing members and faculties of the Department of Mathematics at HKUST for their support throughout my studies. I extend my appreciation to Prof. Dong Xia and Prof. Xinzhou Guo for serving as my committee members and for their invaluable guidance during the preparation of this paper. I am also grateful to Prof. Tianling Jin and Prof. Zhigang Bao for their help in the reading course organization and my life

at HKUST. Finally, I would like to thank Elina Chiu for her great assistance during my study.

One of the most enjoyable aspects of academia for me is the opportunity to collaborate with great senior mentors and talented peers. I am grateful to have had the privilege of working under the guidance of many senior mentors. I worked closely with Prof. Cong Shen at the University of Virginia and Prof. Jing Yang from Pennsylvania State University starting from my sophomore year as an undergraduate student, and I am fortunate to have continued to study with them in the years since. I appreciate the guidance and mentorship of Prof. Zhuoran Yang from Yale University, Prof. Zhaoran Wang from Northwestern University, and also Prof. Liwei Wang from Peking University. Their contributions have greatly expanded my horizons in theoretical reinforcement learning. I was fortunate to work with Prof. Haishan Ye from Xian Jiaotong University in the early stage of my study on decentralized optimization. In addition, I want to thank Prof. Quanquan Gu from UCLA for sharing his inspiring algorithmic ideas for online decision-making problems. I would like to thank the collaborations from my coauthors, Chengshuai Shi, Han Zhong, Chenlu Ye, Hanze Dong, Jiyuan Tan, Rui Pan, Shizhe Diao, Jipeng Zhang, Sirui Zheng, and Kashun Shum. I appreciate all their contributions to our joint projects. Finally, I would like to express my appreciation for the constructive and insightful discussions with Tengyang Xie, Qinghua Liu when preparing the draft of this paper; I thank Fan Chen for many discussions on the DEC and its variant; and I also thank Nan Jiang for the valuable feedback and ideas in the algorithmic design and also the discussion of completeness and realizability. I am grateful for everything I have learned from all these great minds.

I enjoyed the time with my friends. I would like to thank Zhiwei Wang, Chenlu Ye, Yifan Hao, Chuqi Chen, Hanze Dong, Yong Lin, Wanteng Ma, Mengyue Zha, Zetao Fei, Xinyu Liu, Kai Han, Zhongyi Shi, Tianyu Jin, Tiankai Hu, Yueyan

Jiang, Chen Liu, Min Zeng, Wen Ji, and many others. I cherish the wonderful moments we have together in Hong Kong. I also want to thank Chengxi Yang, Shuyu Zhang, and Qiang Sun for the various conversations about our lives, and all the encouragement from them. Finally, I would like to thank Xinyu You for her encouragement and support that helped me overcome the difficult moments at the beginning of my master's study.

Finally, I want to thank my parents, for their unconditional love.

Contents

List of Figures

List of Tables

A Sufficient Condition of Sample-Efficient Reinforcement Learning with General Function Approximation

Wei Xiong

Department of Mathematics

Abstract

In this paper, we study reinforcement learning (RL) with general function approximation, where either the value function or the model dynamics is approximated by a given abstract hypothesis space. We propose the generalized eluder coefficient (GEC), which measures the hardness of generalization from the historical in-sample error to the prediction error, and further serves to measure the hardness of learning an RL problem. In terms of the algorithmic design, we propose an optimization-based framework for RL with general function approximation, following the general principle of "Optimism in the Face of Uncertainty" (OFU). Compared to existing algorithms, the proposed framework does not explicitly maintain the confidence set, and neatly handles both model-free and model-based problems with a low GEC. Theoretical analysis shows that our regret results match those provided by existing frameworks.

Chapter 1

Introduction

1.1 Introduction

In a single-agent episodic Markov decision process (MDP), an agent interacts with the environment by executing an action at each step after observing the current state, collects an immediate reward, and receives an observation emitted from the environment. Then, the next step begins and the game ends until the agent reaches step $H + 1$ (referred to as the episode length or horizon). The goal of the agent is to find an optimal policy, which maximizes her long-turn cumulative reward. The process of finding such a (near-optimal) policy is referred to as the learning in reinforcement learning (RL) [1]. In such a decision-making problem, we face the trade-off between **exploiting** the current knowledge about the environment from the data observed so far and **exploring** the unknown by taking the decision that seems to be sub-optimal, betting on the fact that observed data are not sufficient to truly identify the best option.

One of the core problems in RL is to identify the structural assumption that permits sample-efficient learning, in the sense that we can find a near-optimal policy in polynomial number of interactions with the environment. While the

tabular MDP with finite state space and action space has been well studied [2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12], the minimax regret bound of tabular setting depends on the number of states S. However, in the modern RL, the state space is usually extremely large or even infinite (e.g. Go with $S = 10^{170}$, and Xiangqi with $S = 10^{40}$). This suggests that MDPs with large state space cannot be handled without further structural assumptions. However, there are also many applications showing that practical deep reinforcement learning (DRL) can be quite sample-efficient (e.g. [13; 14; 15]). Motivated by these empirical successes of DRL, a line of work is devoted to the function approximation setting where we approximate the value function, policy, or model dynamic by an abstract hypothesis set $\mathcal{H}$ (typically the neural network) [16; 17; 18; 19], first in the linear case [20; 21; 22; 17; 23; 24; 25; 26] and is later extended to the general function approximation [16; 27; 18; 19; 28]. However, the algorithmic design and theoretical analysis in the function approximation scenario are still largely in a case-by-case manner, and whether we can describe the complexity of the solvable RL problems in a unified notion remains open.

One can generally summarize the exploitation-exploration trade-off in function approximation into three steps: (i) estimate the hypotheses based on the data collected so far (typically in a supervised manner); (ii) plan according to step (i) and pick one hypothesis as the approximate ground truth, and (iii) explore the environment to collect new data. Motivated by the classic result of supervised learning, one may expect that (1) realizability (the true model $f^* \in \mathcal{H}$) and (2) bounded statistical complexity (e.g., VC dimension or the covering number of $\mathcal{H}$) are sufficient for RL because they jointly ensure a supervised guarantee in step (i). Unfortunately, due to the distinct challenges arising from the online nature of RL, we have a negative result, stating that learning a good policy is statistically hard even though the hypothesis class is realizable (see Proposition 1.2.9). Therefore, we require certain additional assumptions for RL to solve the under-

lying MDP, at least in the worst case. In this paper, we focus on the MDP with general function approximation and propose *generalized eluder coefficient* (GEC), a unified complexity measure that generalizes the eluder coefficient considered in [28] and captures nearly all known solvable RL problems. Meanwhile, we propose a unified optimization-based algorithmic framework, following the general principle of *"Optimism in the Face of Uncertainty"* (OFU) [29; 30; 31], which solves all problems with a low GEC.

From a high level, the GEC states that the problems have a certain structure that on average, we can bound the prediction error on the next unseen sample by the in-sample training error on the samples collected so far over the hypothesis class $\mathcal{H}$, in an online manner. This allows us to reduce the online RL problem to a relatively well-studied in-sample supervised error estimation, which shares similar spirits with [19], and also [32; 33] but with different reduction targets. In terms of the algorithmic design, the proposed framework simply takes the in-sample loss minimizer at each iteration but with a "feel-good" modification in the objective function as inspired by [32]. Compared to the existing works, the optimization subroutine is constraint-free and unifies many famous algorithms with an elegant interpretation from GEC. Moreover, the analysis is standard and simple, regardless of the considered problems.

The rest of this section is devoted to the problem setup and a comprehensive review of related work. Then, we motivate and develop the GEC in Chapter 2 and propose our algorithm in Chapter 3. We then compare GEC with the existing frameworks and discuss several challenges, and potential extensions of GEC in Chapter 4. For a better presentation, some of the proofs in the main text are compactly provided in Chapter 5. Finally, we conclude in Chapter 6.

1.2 Problem Setup

Markov decision process (MDP). A MDP is specified by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the episode length, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ and $r = \{r_h\}_{h=1}^H$ are the state transition kernels and reward functions, respectively. For each $h \in [H]$, $\mathbb{P}_h(\cdot|x, a)$ is the distribution of the next state given the state-action pair (x, a) at step h, $r_h(x, a) \in [0, 1]$ is the deterministic reward given the state-action pair (x, a) at step h[1]. The key property of the MDP is that the transition kernel satisfies the Markov property, i.e., $\mathbb{P}_h(x_{h+1} \mid x_1, a_1, \cdots, x_h, a_h) = \mathbb{P}_h(x_{h+1} \mid x_h, a_h)$ for any $h \in [H]$ and $(x_1, a_1 \cdots x_h, a_h, x_{h+1}) \in \mathcal{S}^{h+1} \times \mathcal{A}^h$.

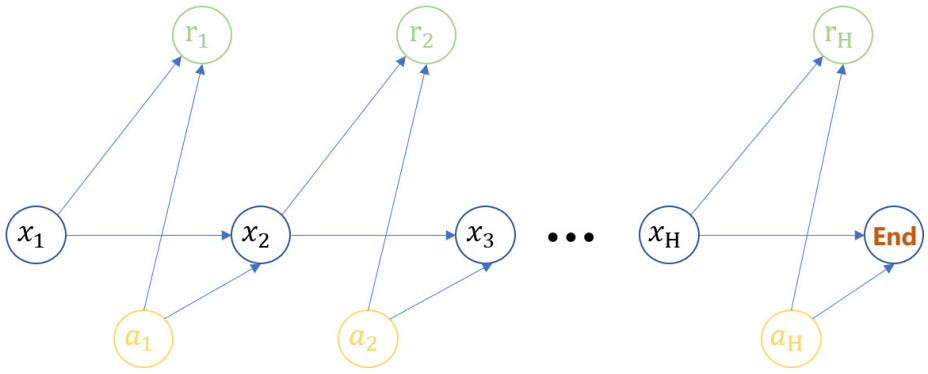


Figure 1.1: An illustration of MDP with episode length H.

A Markovian policy $\pi = \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h \in [H]}$ maps each state to a distribution over actions. Given a Markovian policy π, its Q-function and value function at step h are defined as expected cumulative rewards, given the current state (or

[1]The results readily generalize to the stochastic reward as the uncertainty of reward is non-dominating compared with that of state transition.

state-action pair):

$$V_h^\pi(x) = \mathbb{E}_\pi\Big[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) \,\Big|\, x_h = x\Big],$$
$$Q_h^\pi(x,a) = \mathbb{E}_\pi\Big[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) \,\Big|\, x_h = x, a_h = a\Big].$$

It immediately follows that the following *Bellman equation* holds:

$$Q_h^\pi(x,a) = (\mathcal{T}_h V_{h+1}^\pi)(x,a) := r_h(x,a) + \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a)} V_{h+1}^\pi(x'), \qquad \forall \pi, x, a. \tag{1.2.1}$$

Here $\mathcal{T}_h$ is referred to as the *Bellman operator* at step h. We also use $\pi^* = \{\pi_h^*\}_{h\in[H]}$, $V^* = \{V_h^*\}_{h\in[H]}$, and $Q^* = \{Q_h^*\}_{h\in[H]}$ to denote the optimal (Markovian) policy, optimal value function and optimal Q-function, respectively, where they satisfy the following properties [1]:

$$\begin{aligned} Q_h^{\pi^*}(x,a) = Q_h^*(x,a) = \sup_\pi Q_h^\pi(x,a) \qquad \forall (x,a) \in \mathcal{S}\times\mathcal{A}, \\ V_h^{\pi^*}(x) = V_h^*(x) = \sup_\pi V_h^\pi(x) \qquad \forall x \in \mathcal{S}, \end{aligned} \tag{1.2.2}$$

and one optimal policy π^* is the greedy policy induced by Q^*. We remark that the existence of an optimal Markovian policy is because of the Markovian property of MDP. For general decision-making problems (e.g. partially observable MDP), the optimal policy depends on the whole history. It is also well known that (Q^*, V^*) satisfies the *Bellman optimality equation* for any $(h,x,a) \in [H]\times\mathcal{S}\times\mathcal{A}$:

$$Q_h^*(x,a) = (\mathcal{T}_h V_{h+1}^*)(x,a), \qquad V_h^*(x) = \max_{a\in\mathcal{A}} Q_h^*(x,a). \tag{1.2.3}$$

For simplicity, we additionally assume that the first state is a fixed one x_1, which does not hurt the generality because one can assume that there is a null state at step 0 and then transits to step 1 according to the fixed state distribution. We give an example of tabular MDP below.

Example 1.2.1 (Tabular MDP)**.** A tabular MDP is a MDP with finite state and action spaces. In this case, the transition kernel $\mathbb{P}_h$ can be represented by a table of size $|\mathcal{S}\times\mathcal{A}| \times |\mathcal{S}|$, where the (x,a,x')-entry is $\mathbb{P}_h(x'|x,a)$.

Function Approximation. Following [19], we assume that we have access to a hypothesis class $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_H$, which can be either *model-based* or *value-based*, and we detail them as follows.

Example 1.2.2 (Model-based Hypothesis)**.** A model-based hypothesis class $\mathcal{H}$ is a set of models (transition kernel $\mathbb{P}$ and reward function r). In this case, for any $f = (\mathbb{P}_f, r_f) \in \mathcal{H}$, we denote $\pi_f = \{\pi_{h,f}\}_{h\in[H]}$ and $Q_f = \{Q_{h,f}\}_{h\in[H]}, V_f = \{V_{h,f}\}_{h\in[H]}$ as the optimal policy and optimal value functions corresponding to the model f, respectively. We also denote the real model by f^*.

Example 1.2.3 (Value-based Hypothesis for MDP)**.** A value-based hypothesis class $\mathcal{H}$ is a set of Q-function, that is, $\mathcal{H} = \{\mathcal{H}_h\}_{h\in[H]}$, where $\mathcal{H}_h = \{Q_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}\}$. For any $f = \{Q_h\}_{h\in[H]}$, let $Q_f = \{Q_{h,f} = Q_h\}_{h\in[H]}$, $V_f = \{V_{h,f}(\cdot) = \max_{a\in\mathcal{A}} Q_{h,f}(\cdot, a)\}_{h\in[H]}$, and $\pi_f = \{\pi_{h,f}(\cdot) = \operatorname{argmax}_{a\in\mathcal{A}} Q_h(\cdot, a)\}_{h\in[H]}$. We also denote $f^* = Q^*$, where Q^* is the optimal Q-function.

To further improve readability, sometimes we will also use $\mathcal{F}$ for a value-based hypothesis, and $\mathcal{M}$ for a model-based hypothesis to distinguish them. We remark that the main difference between value-based and model-based hypothesis spaces is whether we use or learn the information of the transition kernel. Accordingly, when the algorithm does not explicitly use the information of the transition kernel, it is referred to as a *model-free* approach. The model-based algorithm has a meaning similar to the model-based hypothesis. For each $f \in \mathcal{H}$, we define the Bellman residual as

$$\mathcal{E}_h(f, x, a) := Q_{h,f}(x, a) - (\mathcal{T}_h V_{h+1,f})(x, a). \tag{1.2.4}$$

By (1.2.3), we know that $\mathcal{E}_h(f^*, x, a) = 0$ for all $(h, x, a) \in [H]\times\mathcal{S}\times\mathcal{A}$. Throughout this paper, we will assume that $\mathcal{H}$ contains f^* (c.f. Example 1.2.2 or Example 1.2.3), which is standard in the literature [e.g., 34; 18; 19; 28].

Assumption 1.2.4 (Realizability)**.** We assume $f^* \in \mathcal{H}$.

Remark 1.2.5 (Notions of realizability.)**.** For the model-based hypothesis set $\mathcal{M}$, Assumption 1.2.4 means that the true model $M^* \in \mathcal{M}$. For the value-based hypothesis $\mathcal{F}$, Assumption 1.2.4 means that the $Q^* \in \mathcal{F}$. Clearly, model-based realizability implies value-based realizability (if we consider the induced value class $\{Q_{h,M} : (h, M) \in [H] \times \mathcal{M}\}$). We will see that we can obtain a sharper result with the model-based hypothesis under realizability in Chapter 3.

We now illustrate the notion of function approximation by the case of linear MDP [17].

Example 1.2.6 (Linear MDP)**.** MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a (known) feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown signed measures $\mu_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$, we have $\mathbb{P}_h(\cdot \mid x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle, r_h(x, a) = \langle \phi(x, a), \theta_h \rangle$. Without loss of generality, we assume that $||\phi(x, a)|| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{||\mu_h(\mathcal{S})||, ||\theta_h||\} \leq \sqrt{d}$ for all $h \in [H]$.

For linear MDP, both the transition kernel and the reward function are linear in a known feature. As a special case, we can parameterize the tabular MDP by taking the one-hot mapping for the state-action pair. :

$$\phi(x_1, a_1) = (1, 0, \cdots, 0)^\top \qquad \phi(x_i, a_j) = (0, \cdots, 0, \underbrace{1}_{(i-1)\times|\mathcal{S}|+j}, 0 \cdots, 0)^\top. \tag{1.2.5}$$

In this case, the linear feature will be of dimension $d = |\mathcal{S}||\mathcal{A}|$. But the linear MDP can further handle the infinite state-action space. For linear MDP, we have the following result.

Lemma 1.2.7. For any function $V : \mathcal{S} \to [0, H-1]$ and $h \in [H]$, there exist vectors $\beta_h, w_h \in \mathbb{R}^d$ with $\max\{||\beta_h||, ||w_h||\} \leq \sqrt{d}H$, such that $\forall (x, a) \in \mathcal{S} \times \mathcal{A}$, the conditional expectation and Bellman update are both linear in the feature:

$$(\mathbb{P}_h V)(x, a) = \phi(x, a)^\top \beta_h, \quad \text{and} \quad (\mathcal{T}_h V)(x, a) = \phi(x, a)^\top w_h, \tag{1.2.6}$$

where $(\mathbb{P}_h V)(x, a) := \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V(x')$.

We will prove this lemma as part of the proof of Example 2.2.5 for completeness. By the Bellman equation (1.2.1) and the linearity of the Bellman operator in linear MDP, one can take the linear function space as the hypothesis space for the linear MDP, i.e., $\mathcal{H}_h = \{Q_{h,f}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \theta_{h,f} : \|\theta_{h,f}\| \le \sqrt{d}H\}$ so that $Q_h^* \in \mathcal{H}_h$.

For a clear presentation, we additionally assume that $|\mathcal{H}|$ is finite. We note that this is only for simplicity as it can be readily extended to the infinite hypothesis space with a mild covering number by standard discretization technique.

Assumption 1.2.8 (Finite Hypothesis)**.** We assume $|\mathcal{H}| < \infty$.

Learning process. For each time step t, the agent first picks a hypothesis $f^t \in \mathcal{H}$. Then she collects a new trajectory $\zeta_h^t = \{x_1^t, a_1^t, r_1^t, \cdots, x_H^t, a_H^t, r_H^t\}$, by following the greedy policy π_{f^t} induced by f^t. Then, a new iteration begins.

Learning objective. We consider the following regret minimization problem for T iterations in total:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V_1^*(x_1) - V_1^{\pi^t}(x_1).$$

The goal is to design an algorithm to solve the underlying MDP with a sub-linear (in T) regret. The following result shows that realizability itself is not sufficient for sample-efficient learning.

Proposition 1.2.9 (Realizability is not sufficient [35])**.** For any $S \in \mathbb{N}$ and $H \in \mathbb{N}$, there exists a class of horizon-H MDPs $\mathcal{M}$ with $|\mathcal{S}| = S$, $|\mathcal{A}| = 2$, and $\log|\mathcal{M}| = \log(S)$. We suppose that the true model $M^* \in \mathcal{M}$ (realizability), yet any algorithm must have

$$\mathbb{E}\mathrm{Reg}(T) \ge \sqrt{\min\{S, 2^H\} \cdot T}.$$

It is known that the regret guarantee can be transformed into the sample complexity bound.

Lemma 1.2.10 (Online-to-batch conversion)**.** If an algorithm has a sublinear regret of $c^\dagger T^{1-\alpha}$ with $0 < \alpha \leq 1$, then the algorithm finds an ϵ-optimal policy with at most $(c^\dagger/\epsilon)^{1/\alpha}$ samples. Here $c^\dagger$ is a problem-dependent constant.

Proof. We denote the policy sequence as $\{\pi^1, \cdots, \pi^T\}$. Then, by definition of regret, we know

$$\begin{aligned}\mathrm{Reg}(T) &= T \cdot V_1^*(x_1) - \sum_{t=1}^{T} V_1^{\pi^t}(x_1) \\ &\leq c^\dagger T^{1-\alpha}.\end{aligned}$$

We consider the uniform policy $\widetilde{\pi} := \mathrm{Uniform}(\pi^1, \cdots, \pi^T)$. It follows that

$$V_1^*(x_1) - V_1^{\widetilde{\pi}}(x_1) = V_1^*(x_1) - \frac{1}{T}\sum_{t=1}^{T} V_1^{\pi^t}(x_1) \leq c^\dagger T^{-\alpha} := \epsilon,$$

which implies that $T = (c^\dagger/\epsilon)^{1/\alpha}$. □

Additional notations. We use $\Delta_{\mathcal{X}}$ to denote the space of all distributions over the set $\mathcal{X}$. For some $n \in \mathbb{N}^+$, we use the convention that $[n] = \{1, \cdots, n\}$. We also let $x_{1:n} = \{x_1, \cdots, x_n\}$. For two distributions $P, Q \in \Delta_{\mathcal{X}}$, the Hellinger divergence $D^2_{\mathrm{H}}(P, Q)$ is given by:

$$D^2_{\mathrm{H}}(P, Q) = \frac{1}{2}\int_{\mathcal{X}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 dx = 1 - \int_{\mathcal{X}} \sqrt{P(x)Q(x)}dx, \tag{1.2.7}$$

where P and Q are probability mass functions or probability density functions. We use $f = \mathcal{O}(g)$ or $f \lesssim g$ to hide the constant factor, i.e., $f \leq c \cdot g$ for some constant $c > 0$. We also use $\widetilde{\mathcal{O}}$ to further omit logarithmic factors. To improve the readability, we provide a summary of notations in Table 1.1.

Table 1.1: Summary of notations used in this paper.

Notation	Explanation
$\mathcal{S}, \mathcal{A}, H$	State space, action space, and episode length
S, A	$\lvert\mathcal{S}\rvert$ and $\lvert\mathcal{A}\rvert$
$\mathcal{H}$	Hypothesis space, $\mathcal{H}_1 \times \cdots \times \mathcal{H}_H$
$\mathcal{E}_h(f, x_h, a_h)$	Bellman residual of hypothesis f at step h, (1.2.4)
$\langle a, b \rangle$	Inner product of a and b
π_f	The greedy policy induced by f.
$\mathbf{1}(E)$	Indicator function of event E
$\mathbf{I}$	Identity matrix
$\mathrm{Unif}(\mathcal{X})$	Uniform distribution over set $\mathcal{X}$
$\lvert\lvert \cdot \rvert\rvert$	The 2-norm by default.
$x[i]$	The i-th entry of the vector x.

1.3 Related Work

The central problem in theoretical RL is to identify the structural assumption that permits sample-efficient learning. We now present a comprehensive review of the attempts and also the results we have collected so far in the literature.

Tabular MDP. For a tabular MDP, we assume that the state space $\mathcal{S}$ and action space $\mathcal{A}$ are small. But we do not impose any structural assumption across states and actions. The goal in the tabular case is to design algorithms that achieve a regret depending polynomially on S, A and also the horizon H. The tabular MDP has been extensively studied in the literature [2; 3; 4; 5; 6; 7; 36; 8; 9; 10; 11; 12]. Among them, [3] designs a model-based algorithm UCB-VI that explicitly models the transition matrix of the MDP, and attains the minimax-optimal regret bound $\mathcal{O}(\sqrt{H^2SAT})$. After this, [5] proposes an optimistic variant

of Q-learning UCB-B, which is model-free because it directly learns the optimal Q-value instead of the model dynamic, and attains a regret of $\mathcal{O}(\sqrt{H^3SAT})$ with Bernstein-type bonus. This is later improved by [36] to close the gap to the lower bound by leveraging the idea of variance reduction by a reference function [37]. In comparison, the model-free algorithms typically require less time space and storage space as compared to the model-based counterparts. Since both the model-based and model-free algorithms attain the minimax-optimal regret bound, the tabular settings are well-studied. However, since the lower bound depends on the $\sqrt{SA}$, we cannot handle modern RL problems with large state space without further structural assumptions.

To handle the large or even infinite state/action space, we need to impose additional structural assumptions across different states and actions. Motivated by the empirical success of DRL, we will approximate either the model dynamics (the transition kernel and the reward function, referred to as the model-based approach) or the value functions (e.g. Q^*, V^*, Q^π, referred to as the model-free approach) by an abstract hypothesis space $\mathcal{H}$. The primary goal is to design algorithms that generalize across the large state-action space well and attain a sub-linear regret bound in T and also with a mild dependence on H and other problem-dependent parameters (e.g. d, dimension of the feature). In particular, instead of depending on the number of states S, we expect that the regret bounds scales with the statistical complexity of the function class (e.g. $\log|\mathcal{H}|$ for finite class or the log covering number $\log\mathcal{N}$ for infinite class).

Linear function approximation and learnability. Linear function approximation is arguably the most fundamental one [20; 21; 22; 17; 23; 24; 25; 26; 38; 39; 40] in function approximation. Typically, we will assume that we have access to a d-dimensional feature map of the state-action pair $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. A natural idea is to assume that the optimal Q-value Q^* is linear in this feature, where we refer it as the *linear Q^* condition*, in the sense that there exists a $\theta_h \in \mathbb{R}^d$

and $||\theta_h^*|| \leq B$:

$$Q_h^*(x,a) := \langle \phi(x,a), \theta_h^* \rangle, \qquad \forall h \in [H]. \tag{1.3.1}$$

The main technical consideration is that due to the extra linear structure, we can generalize from the visited states to the unseen states by the standard analysis of linear regression (in comparison, the tabular MDP does not impose any structure across states and actions so such a generalization is impossible). Then, we can approximate the Q^* by $\mathcal{H} := \{Q_h(x,a) = \langle \phi(x,a), \theta_h \rangle : ||\theta_h|| \leq B, h \in [H]\}$, which satisfies the realizability assumption (Assumption 1.2.4). Unfortunately, there exists a negative result [41], saying that this is not sufficient for sample-efficient learning.

Proposition 1.3.1 (Linear-realizability is not sufficient [41])**.** There exists an MDP with feature map ϕ that satisfies (1.3.1) but any algorithms must have

$$\mathbb{E}\mathrm{Reg}(T) \gtrsim \min\{2^{\Omega(d)}, 2^{\Omega(H)}\}.$$

To bypass this hardness result, [17] imposes a stronger assumption that both the transition kernel and the reward function are linear in the feature, where we refer to this condition as the linear MDP (Example 1.2.6). [17] designs an optimistic variant of LSVI (Least Squares Value Iteration), referred to as the LSVI-UCB, that achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{d^3H^4T})$. This result is recently improved by [39; 40] to $\widetilde{\mathcal{O}}(\sqrt{d^2H^3T})$, which matches the minimax lower bound in [26]. Another line of work [24; 25; 22; 26] study the linear mixture MDP, where the transition kernel is a linear mixture of a number of basis kernels and also has designed algorithms based on linear regression to achieve the minimax-optimal regret bound. As we will show in Chapter 2, the linear MDP and the linear mixture MDP essentially limit the generalization from the historical data samples to the newly arrived trajectory by limiting the freedom of the transition kernels and reward functions. Such a generalization ability is $\widetilde{O}(d)$ in the complexity measure proposed in this paper. We note that both the linear MDP and linear mixture MDP are rather

strong assumptions. There are also some works that consider linear realizability different from (1.3.1). For instance, [42] assumes that the Q-value of any policy is linear:

$$Q_h^\pi(x,a) = \langle \phi(x,a), \theta_h^\pi \rangle, \qquad \forall h \in [H].$$

It is shown that if we can query a simulator with (x_h, a_h) to get $x' \sim \mathbb{P}_h(x'|x_h,a_h)$ and $r_h(x_h,a_h)$, the problem is sample-efficient. On the other hand, in the standard online setting where we always play the episode starting from x_1, the learnability of Q_h^π-realizability remains open.

Sample-efficient RL with general function approximation. While the linear MDP assumption permits sample-efficient learning, it is rather limited in practice. A long line of work extends to the general *non-linear* function approximation and designs algorithms to effectively solve these RL problems. Generally speaking, these works can be largely grouped into two categories (i) impose certain low-rank structures so that some results in the linear MDP can be generalized; (ii) limit the sequence length of effective state (and/or action) distributions with respect to the hypothesis space. We now review them as follows.

A line of work imposes a low-rank structure on certain parts of the RL problems. The seminal work [34] considers the model-free approach with value-based hypothesis $\mathcal{F}$ and proposes the *Bellman rank*, which is imposed on the class of induced Bellman residual: $\{\mathbb{E}_{\pi_f}\mathcal{E}_h(g,x_h,a_h) : f,g \in \mathcal{F}\}$[2]. If we view this set as a matrix of $\mathbb{R}^{|\Pi|\times|\mathcal{F}|}$ with $\Pi := \{\pi_f : f \in \mathcal{F}\}$, the Bellman rank d is the rank of the matrix, maximized over $h \in [H]$. In this case, the linearly independent rows of the matrix one can find are at most d, thus limiting the generalization of the problems. More generally, one can extend the finite-dimensional case by consid-

[2]The example presented here is referred to as the Q-type Bellman rank, which is different from the V-type one considered in [34]. We choose the Q-type one for a clear presentation and will discuss the V-type variant in Chapter 4.

ering a bilinear structure. Specifically, we assume that we have two unknown embeddings $W_h : \mathcal{F} \to \mathcal{V}$ and $X_h : \mathcal{F} \to \mathcal{V}$ where $\mathcal{V}$ is a Hilbert space such that

$$\mathbb{E}_{\pi_f} \mathcal{E}_h(g, x_h, a_h) = \langle W_h(g) - W_h(f^*), X_h(f) \rangle \,, \tag{1.3.2}$$

where the Bellman rank can be described in terms of the information gain [43]. [34] also proposes OLIVE, based on the OFU principle and hypothesis elimination, to solve the problems with a low Bellman rank. After Bellman rank and OLIVE, [27] shows that there exists an exponential separation between the model-free approach and model-based approach, in the sense that the Bellman rank of factored MDPs [44] can be exponentially large. Alternatively, [27] extends the idea of Bellman rank to the model-based setting, and proposes the *witness rank*, to capture the factored MDPs. [19] generalizes them by proposing the bilinear class, which assumes that the average Bellman error and a *discrepancy loss* have a special bilinear structure. Moreover, the bilinear class allows a flexible choice of "discrepancy function" to capture both the model-free and model-based problems. [19] also proposes BiLin-UCB, which is more similar to the optimism-based algorithms that have been analyzed in contextual bandit [45; 46; 47]. Specifically, BiLin-UCB maintains a confidence set $\mathcal{H}^t$ at each iteration where $f^* \in \mathcal{H}^t$ with high probability. Then, the agent chooses the estimator with the highest value estimations (that is why we say it is optimistic) such that the estimation is higher than that of the ground truth to encourage exploration:

$$f^t = \operatorname*{argmax}_{f \in \mathcal{H}^t} V_{1,f}(x_1). \tag{1.3.3}$$

Another line of work focuses on explicitly limiting the length of the longest sequence of effective distributions with respect to the hypothesis space. [48] proposes the notion of the eluder dimension, which generalizes the notion of linear independence in $\mathbb{R}^d$. The eluder dimension is later leveraged to RL by [49], which includes the linear MDP [17] as a special example. However, [49] only characterizes the eluder dimension of the hypothesis space $\mathcal{H}$, and the covered problems are

rather limited. [18] further considers the distributional eluder dimension on the induced Bellman residual space (referred to as the Bellman eluder dimension) so that the eluder dimension is imposed on the interplay between the function class and the underlying MDP, and captures more RL problems. [17] also proposes a model-free OFU-based algorithm, GOLF, which is also based on confidence sets and optimism (1.3.3). The main difference between BiLin-UCB and GOLF is that GOLF additionally assumes the Bellman completeness condition (see Assumption 3.1.3) thus leveraging a minimax formulation [50] to give a more efficient estimation of the Bellman errors (see Chapter 3 for a detailed interpretation of these algorithmic choices). Consequently, GOLF achieves a $\sqrt{T}$-regret, while BiLin-UCB only achieves a $T^{2/3}$-regret (by online-to-bach conversion). We also note that [51] generalizes the Bellman completeness assumption to the more general discrepancy function and applies the minimax formulation to achieve a better regret bound.

It is known that neither the Bellman eluder dimension nor bilinear class captures each other (see comments on page 6 of [19]). Attempts have been made since then to unify these two rich tackable RL problems. [28] introduces the notion of the *eluder coefficient*[3] and studies the Q-type model-free problems. The eluder coefficient $d(\mu)$ explicitly relates the *out-of-sample* average Bellman residual to the *in-sample* average (squared) Bellman error:

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\pi_{f^t}}\mathcal{E}_h(f^t,x_h,a_h) \le \mu\sum_{h=1}^{H}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}[\mathcal{E}_h(f^t,x_h,a_h)^2] + \frac{d(\mu)}{4\mu}. \tag{1.3.4}$$

Intuitively, the eluder coefficient quantifies the rate at which the prediction error can grow in comparison to the cumulative training error on average in an online manner, thus explicitly limiting the generalization from the visited state-action distributions to the unseen part. Technically, we note that the eluder coefficient

[3]The eluder coefficient is referred to as the decoupling coefficient in [4]. However, we note that the decoupling coefficient proposed in [32] is distinctly different from the eluder coefficient in terms of intuition and technical consideration so we use the name eluder coefficient here.

serves to reduce the RL problems into an in-sample error estimation problem over the hypothesis space $\mathcal{H}$ in a supervised learning manner, which is relatively well studied in the literature. Our complexity measure is mostly related to this eluder coefficient, and we will extend it to more general RL problems in Chapter 2, which can unify the Bellman eluder dimension and bilinear class. [28] also proposes conditional posterior sampling with an optimistic modification in the prior, and develops new analysis techniques for sampling-based algorithms.

There is also another line of work with distinct technical considerations compared to the above-mentioned works. [33; 52] propose the decision estimation coefficient (DEC) to unify the complexity measures in interactive decision-making, which takes the MDPs as a special example. Given a model class $\mathcal{M}$ and a reference model $\widehat{M}$, the DEC is given by

$$\mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[\underbrace{V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D_{\mathrm{H}}^2(M(\pi), \widehat{M}(\pi))}_{\text{Easy to control}}], \tag{1.3.5}$$

where $V_{1,M}^{\pi}(x_1)$ is the V-value of policy π when M is the underlying model and $M(\pi)$ denotes the trajectory distribution jointly determined by the model M and the executed policy π. The technical consideration is to convert the RL problems into an online learning problem, by reducing the out-of-sample regret to another out-of-sample divergence $D_{\mathrm{H}}^2(M(\pi), \widehat{M}(\pi))$. We note such a technical treatment also arises in the decoupling coefficient proposed in [32; 53; 54]. The idea of such a conversion may date back to the *information ratio* [55]. DEC is a more general complexity measure that captures both the bilinear class and Bellman eluder dimension. The most appealing part of DEC is the matching lower bound in terms of DEC in some decision-making problems, which suggests that a low DEC is necessary for sample-efficient learning. However, the vanilla DEC (1.3.5) cannot be applied in a model-free manner. To address this issue, [52] adopt an optimistic modification as in [32], and extend it to the model-free scenario. However, the regret bound of the model-free E2D obtained in [52] is

inferior compared to that of [19] under only realizability. We mention in passing that [56] also independently studies the optimistic variant of DEC. A notable feature of DEC is that it involves a minimax operation, which accounts for the minimax subroutine in their algorithms, Estimation-to-Decisions (E2D). Such a black box minimax subroutine may lack efficient computation guidelines in practice and the DEC may not be applied to analyze the classic OFU-based or sampling-based algorithms. In comparison, the decoupling coefficient [32; 53] does not require solving a minimax problem, and can be applied to analyze the optimistic algorithms based on posterior sampling or Maximum Likelihood Estimation (MLE).

Decision-making problems with general function approximation are still an active research direction and are still developing rapidly toward a deeper understanding of the learnability and the goal of guiding the design of practical algorithms.

Chapter 2

Complexity Measure, and Examples

In this chapter, we introduce a complexity measure, called the generalized eluder coefficient, that characterizes the hardness of an RL problem.

2.1 Regret Decomposition and Optimism

We start with the following value decomposition lemma.

Lemma 2.1.1 (Regret decomposition)**.** Suppose that we execute π_{f^t} (i.e., the greedy policy of f^t) for each iteration. Then, it holds that[1]:

$$\begin{aligned}\sum_{t=1}^{T} V_1^*(x_1) - V_1^{\pi_{f^t}}(x_1) &= \sum_{t=1}^{T}\Big[\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\Big] - \sum_{t=1}^{T}\Big[\underbrace{\left(V_{1,f^t}(x_1) - V_1^*(x_1)\right)}_{\Delta V_{1,f^t}(x_1)}\Big] \\ &\le \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] - \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right].\end{aligned} \tag{2.1.1}$$

The primary goal in this chapter is to control the first term under the expectation

[1] We defer the proof to Section 5.1.

of π_{f^t}, which is tackable as we have access to both the sequence of $\{f^t\}_{t=1}^T$ and will also make certain assumptions on $\mathcal{H}$. In contrast, the second term is hard to deal with because we have no idea about the optimal policy π^*. This is the main technical reason why in the literature, we mainly consider algorithms that follow the optimistic principle, such that $\Delta V_{1,f^t}(x_1) \geq 0$ for all $t \in [T]$ so the second term can be eliminated. We will return to this algorithmic design in Chapter 3. With this lemma in hand, our target moves from the regret to the cumulative Bellman residuals $\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi_{f^t}}[\mathcal{E}_h(f^t, x_h^t, a_h^t)]$.

2.2 Exploitation is Safe When the Generalization is Limited

We motivate the complexity measure by analyzing the exploration-exploitation trade-off in the context of RL. At the beginning of iteration t, the agent has collected the data for the first $t-1$ iterations: $\{\zeta_h^s\}_{s=1}^{t-1}$. Then, the agent needs to make decisions based on these historical samples, aiming to perform well on the *unseen* trajectory at iteration t so as to achieve a low regret in the long run. This requires certain extrapolation from the states already visited to the unseen part of the state space. In other words, we shall be able to infer the trajectory at iteration t by the knowledge of the history.

As an illustrative example, we first consider a class of linear function $\mathcal{F} = \{f(\cdot) = \phi(\cdot)^\top \theta_f : ||\theta_f|| \leq 1\}$ where $\phi(z) \in \mathbb{R}^d$. For simplicity, we also assume that $||\phi(z)|| \leq 1$ for all $z \in \mathcal{Z}$, which also implies that $|f(z)| \leq 1$ for all $z \in \mathcal{Z}$. With the notation $\Sigma_t = \lambda \mathbf{I} + \sum_{s=1}^{t-1} \phi(z_s)\phi(z_s)^\top$, for any $f, g \in \mathcal{F}$, we relate their

difference in the next unseen $z_t \in \mathcal{Z}$ with the regularized in-sample error:

$$\begin{aligned} |f(z_t) - g(z_t)|^2 = |\langle \phi(z_t), \theta_f - \theta_g \rangle|^2 &\leq ||\phi(z_t)||^2_{\Sigma_t^{-1}} ||\theta_f - \theta_g||^2_{\Sigma_t} \\ &\leq ||\phi(z_t)||^2_{\Sigma_t^{-1}} \Big(\lambda + \sum_{s=1}^{t-1} |f(z_s) - g(z_s)|^2\Big), \end{aligned} \tag{2.2.1}$$

where we use Cauchy-Schwarz inequality and expand Σ_t in the last inequality. Therefore, the out-of-sample prediction error on the unseen z_t can be bounded by the (regularized) in-sample training error up to a factor of $||\phi(z_t)||^2_{\Sigma_t^{-1}}$, where $||\phi(z_t)||^2_{\Sigma_t^{-1}}$ is referred to as the elliptical potential in the literature [47]. The following lemma shows that for the linear model, the generalization is limited, in the sense that the elliptical potential is small on average.

Lemma 2.2.1 (Exploitation is safe for linear model)**.** We consider $\mathcal{F} = \{f(\cdot) = \phi(\cdot)^\top \theta_f : ||\theta_f|| \leq 1\}$ where $\phi(z) \in \mathbb{R}^d$ and $||\phi(z)|| \leq 1$ for all $z \in \mathcal{Z}$. For any sequence of $\{f_t, g_t, z_t\}_{t=1}^T$, we have

$$\sum_{t=1}^T |f_t(z_t) - g_t(z_t)| \leq \widetilde{\mathcal{O}}\Big(\Big[d \cdot \sum_{t=1}^T \Big[\lambda + \sum_{s=1}^{t-1} (f_t(z_s) - g_t(z_s))^2\Big]\Big]^{1/2}\Big).$$

Proof. Following the idea in (2.2.1), we decompose the prediction error into the in-sample error and potential:

$$\begin{aligned} \sum_{t=1}^T |f_t(z_t) - g_t(z_t)| &= \sum_{t=1}^T |f_t(z_t) - g_t(z_t)| \{\mathbf{1}(||\phi(z_t)||_{\Sigma_t^{-1}} \leq 1) + \mathbf{1}(||\phi(z_t)||_{\Sigma_t^{-1}} > 1)\} \\ &\leq \sum_{t=1}^T \min\{||\phi(z_t)||_{\Sigma_t^{-1}}, 1\} ||\theta_{f^t} - \theta_{g^t}||_{\Sigma_t} + \sum_{t=1}^T \mathbf{1}(||\phi(z_t)||_{\Sigma_t^{-1}} > 1) \\ &\lesssim \sqrt{\sum_{t=1}^T \min\{||\phi(z_t)||^2_{\Sigma_t^{-1}}, 1\}} \sqrt{\sum_{t=1}^T ||\theta_{f^t} - \theta_{g^t}||^2_{\Sigma_t} + d \log\big(1 + \frac{1}{\lambda}\big)} \\ &\leq \widetilde{O}\Big(\Big[d \cdot \sum_{t=1}^T \Big[\lambda + \sum_{s=1}^{t-1} (f_t(z_s) - g_t(z_s))^2\Big]\Big]^{1/2}\Big), \end{aligned}$$

where the first inequality uses $|f_t(z_t) - g_t(z_t)| \leq 1$, and the second inequality holds because of the Cauchy-Schwarz inequality and some calculations. Finally, we invoke Lemma A.1.3 to bound the summation of elliptical potentials in the second and last inequalities. □

The above lemma ensures that when the loss function ($|f(\cdot) - g(\cdot)|$ in this example) has a special linear structure, we can reduce the prediction error to the in-sample error in an online manner. We can generalize this idea in the context of RL.

Definition 2.2.2 (Eluder Coefficient)**.** Given a MDP and a hypothesis class $\mathcal{H}$, the eluder coefficient $d(\epsilon)$ is the smallest d ($d \geq 0$) such that for any sequence of hypotheses $\{f^t \in \mathcal{H}\}_{t=1}^T$, it holds that

$$\sum_{t=1}^{T} \underbrace{V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1)}_{\text{prediction error}} = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \mathcal{E}_h(f^t, x_h, a_h)$$

$$\leq \Bigg[d \sum_{h=1}^{H}\sum_{t=1}^{T} \underbrace{\sum_{s=1}^{t-1} \Big(\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, x_h, a_h)\Big)^2}_{\text{training error}} \Bigg]^{1/2} + \underbrace{2\min\{Hd, H^2T\} + \epsilon B^\dagger T}_{\text{burn-in cost}},$$

where $B^\dagger > 0$ is some problem-dependent constant for regularization.

The equality follows from the value decomposition lemma A.1.1 (also c.f. [16; 28]). Ignoring the burn-in cost, which is typically non-dominating, the eluder coefficient suggests that the prediction error can be upper bounded by the cumulative training error on average, although the training error is amplified by the eluder coefficient. Therefore, the eluder coefficient can be used to measure the hardness of such a generalization, thus further serving to measure the hardness of learning the MDP. We make several remarks before continuing.

Remark 2.2.3. The definition presented here is similar to (1.3.4) from [28] (up to a Cauchy-Schwarz inequality), except that the expectation is inside the square in the training error. It turns out that this allows a more flexible choice of algorithmic design, as we will detail in Section 3.

Remark 2.2.4. We introduce an expectation in the notion of loss, instead of evaluating the loss at a specific point as in the Lemma 2.2.1. Therefore, the implicit structure assumption is now imposed on the interplay between the MDP and the hypothesis class, which allows this formulation to capture more problems.

As a motivating example, we show that the linear MDP has a low eluder coefficient. The following proof heavily relies on the linear structure, but can be generalized to analyze the bilinear class [19].

Example 2.2.5 (Linear MDP has a low eluder coefficient)**.** For the linear MDP defined in Example 1.2.6, if we take $\mathcal{H}_h = \{Q_{h,f}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \theta_{h,f} : \|\theta_{h,f}\| \leq \sqrt{d}H\}$, then it has an eluder coefficient of $d(\epsilon) = \mathcal{O}\big(Hd\log(1+\frac{T}{\epsilon})\big)$.

Proof. The first step is to show that the $\mathbb{E}\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)$ is linear, which can be further controlled by the techniques presented in Lemma 2.2.1. We first prove Lemma 1.2.7. For any $V : \mathcal{S} \to [0, H-1]$, we have

$$\begin{aligned}\mathcal{T}_h V(x,a) = r_h(x,a) + (\mathbb{P}_h V)(x,a) &= \phi(x,a)^\top \theta_h + \int_{\mathcal{S}} V(x_{h+1}) \left\langle \phi(x,a), d\mu_h(x_{h+1}) \right\rangle \\ &= \left\langle \phi(x,a), \theta_h + \int_{\mathcal{S}} V(x_{h+1}) d\mu_h(x_{h+1}) \right\rangle := \left\langle \phi(x,a), w_h \right\rangle.\end{aligned}$$

Therefore, the Bellman update of any V is linear in the feature $\phi(\cdot,\cdot)$ and $||w_h|| \leq \sqrt{d} \cdot H$ by the regularization condition. The proof of $\mathbb{P}_h V$ follows from setting $r_h = 0$. We are ready to prove the following lemma.

Lemma 2.2.6. For linear MDP with hypothesis class $\mathcal{H}_h = \{Q_{h,f}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \theta_{h,f} : \|\theta_{h,f}\| \leq \sqrt{d}H\}$, for any $f \in \mathcal{H}$ and $h \in [H]$, there exists a $X_h(\cdot) : \mathcal{H} \to \mathbb{R}^d$ such that for any $f, g \in \mathcal{H}$, $\mathbb{E}_{\pi_g}\mathcal{E}_h(f, x_h, a_h) = \langle X_h(g), \theta_{h,f} - w_{h,f} \rangle$, where $Q_{h,f}(x,a) = \phi(x,a)^\top \theta_{h,f}$ and $\mathcal{T}_h V_{h+1,f}(x,a) = \phi(x,a)^\top w_{h,f}$. Moreover, by Definition 1.2.7, it holds

$$\sup_{h,f \in [H]\times\mathcal{H}} \max\{||\theta_{h,f}||, ||w_{h,f}||\} \leq \sqrt{d}H, \text{ and } \sup_{h,f\in[H]\times\mathcal{H}} ||X_h(f)|| \leq 1.$$

Proof. By Lemma 1.2.7, we know that for any $V_{h+1,f}$ associated with $f \in \mathcal{H}$, we can assume that there exists a $w_{h,f} \in \mathbb{R}^d$ such that $\mathcal{T}_h V_{h+1,f}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top w_{h,f}$. Meanwhile, for any $Q_{h,f}$, it can be represented by $Q_{h,f}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \theta_{h,f}$ for some $\theta_{h,f} \in \mathbb{R}^d$. As a result, for any $f \in \mathcal{H}$, the Bellman residual is also linear:

$$\mathcal{E}_h(f, x_h, a_h) = Q_{h,f}(x_h, a_h) - \mathcal{T}_h V_{h+1,f}(x_h, a_h) = \langle \phi(x_h, a_h), \theta_{h,f} - w_{h,f} \rangle.$$

Therefore, we know that $X_h(g) = \mathbb{E}_{\pi_g}\phi(x_h, a_h)$ satisfies the condition. $\square$

We now invoke the above result to establish the eluder coefficient of linear MDP. Before continuing, we introduce the notation $\Sigma_{t;h} = \lambda\mathbf{I} + \sum_{s=1}^{t-1} X_h(f^s)X_h(f^s)^\top$, which is an estimation of the covariance matrix. It follows that

$$\begin{aligned}
&\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi^t}(x_1) = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]\left(\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} + \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} > 1\}\right) \\
&\le H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\min\left\{|\langle X_h(f^t), \frac{\theta_{h,f^t} - w_{h,f^t}}{H}\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} \\
&\quad + H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} > 1\} \\
&\le H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\min\left\{|\langle X_h(f^t), \frac{\theta_{h,f^t} - w_{h,f^t}}{H}\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} + \min\{H\widetilde{d}, H^2T\},
\end{aligned} \tag{2.2.2}$$

where $\widetilde{d} = \frac{3Hd}{\log 2}\log\left(1 + \frac{T}{\lambda\log 2}\right)$. Here the last inequality uses the fact that $||X_h(f^t)||_{\Sigma_{t;h}^{-1}}$ cannot exceed 1 too much times as detailed in Lemma A.1.3. We now fix a (t, h) in the first summation and proceed as follows:

$$\begin{aligned}
&\min\left\{|\left\langle X_h(f^t), \frac{\theta_{h,f^t} - w_{h,f^t}}{H}\right\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} \\
&\le ||\frac{\theta_{h,f^t} - w_{h,f^t}}{H}||_{\Sigma_{t;h}}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\} \\
&= \frac{1}{H}\left[\lambda||\theta_{h,f^t} - w_{h,f^t}||^2 + \sum_{s=1}^{t-1}|\langle X_h(f^s), \theta_{h,f^t} - w_{h,f^t}\rangle|^2\right]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\} \\
&\le \left[\lambda d + \frac{1}{H^2}\sum_{s=1}^{t-1}\left(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t, x_h, a_h)\right)^2\right]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\}
\end{aligned} \tag{2.2.3}$$

where the equality uses $\Sigma_{t;h} = \lambda\mathbf{I} + \sum_{s=1}^{t-1} X_h(f^s)X_h(f^s)^\top$, and the last inequality uses $||\theta_{h,f^t} - w_{h,f^t}|| \le \sqrt{d}H$ and

$$\begin{aligned}
\langle X_h(f^s), \theta_{h,f^t} - w_{h,f^t}\rangle &= \mathbb{E}_{\pi_{fs}}\phi(x_h, a_h)^\top(\theta_{h,f^t} - w_{h,f^t}) \\
&= \mathbb{E}_{\pi_{fs}}\left(Q_{h,f^t}(x_h, a_h) - \mathcal{T}_h V_{h+1,f^t}(x_h, a_h)\right).
\end{aligned}$$

Plugging (2.2.3) into (2.2.2), we obtain that

$$
\begin{aligned}
&\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi^t}(x_1) \\
&\le H \cdot \sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\lambda d + \frac{1}{H^2}\sum_{s=1}^{t-1}\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h)\big)^2\Big]^{1/2}\cdot \min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\} + \min\{H\widetilde{d},H^2T\} \\
&\le \Big(\sum_{t=1}^{T}\sum_{h=1}^{H}\sqrt{\lambda d}H + \sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\sum_{s=1}^{t-1}\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h)\big)^2\Big]^{1/2}\Big)\cdot \min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\} + \min\{H\widetilde{d},H^2T\} \\
&\le \Big[\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s=1}^{t-1}\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h)\big)^2\Big]^{1/2}\Big[\sum_{t=1}^{T}\sum_{h=1}^{H}\min\{||X_h(f^t)||^2_{\Sigma_{t;h}^{-1}},1\}\Big]^{1/2} \\
&\quad + \Big[\sum_{t=1}^{T}\sum_{h=1}^{H}\lambda d H^2\Big]^{1/2}\Big[\sum_{t=1}^{T}\sum_{h=1}^{H}\min\{||X_h(f^t)||^2_{\Sigma_{t;h}^{-1}},1\}\Big]^{1/2} + \min\{H\widetilde{d},H^2T\} \\
&\le \Big[\widetilde{d}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s=1}^{t-1}\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h)\big)^2\Big]^{1/2} + 2\min\{H\widetilde{d},H^2T\} + H^2dT\lambda.
\end{aligned}
$$

We conclude that linear MDP has an eluder coefficient of $\mathcal{O}\big(Hd\log\big(1+\frac{T}{\lambda}\big)\big)$. □

As a direct corollary, we know that the tabular MDP also has a low GEC.

Example 2.2.7 (Tabular MDP has a low eluder coefficient)**.** If we take the feature map as in (1.2.5), then we know that the tabular MDP has a eluder coefficient of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|)$.

2.3 Generalized Eluder Coefficient

To capture more generality, we allow a more flexible choice of the loss function and distribution family. As a motivating example, we first introduce the following linear mixture MDP [24; 22; 25].

Example 2.3.1. We say an MDP is a liner mixture model if there exists (known) feature $\phi : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}^d$ [2] and $\psi : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$; and (unknown) $\theta^*\in\mathbb{R}^d$, such that for all $h\in[H]$ and $(x,a,x')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, we have

$$
\mathbb{P}_h(x' \mid x,a) = \langle\theta_h^*,\phi(x,a,x')\rangle, \qquad r_h(x,a) = \langle\theta_h^*,\psi(x,a)\rangle. \tag{2.3.1}
$$

[2] The case of $\mathbb{R}^d$ can be readily generalized to the general Hilbert space with the notion of "effective dimension" [19].

For regularization, we assume that $||\theta_h^*|| \leq B$ for some constant $B > 0$.

Unfortunately, one cannot obtain a low eluder coefficient for linear mixture MDP by taking the loss function as $\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, x_h, a_h)$. However, it achieves a low eluder coefficient by considering a different loss function. To this end, we define the following generalized eluder coefficient.

Definition 2.3.2 (Generalized eluder coefficient)**.** Given a MDP and a hypothesis class $\mathcal{H}$, the generalized eluder coefficient $d(\epsilon)$ is the smallest d ($d \geq 0$) such that for any sequence of hypotheses $\{f^t \in \mathcal{H}\}_{t=1}^T$, it holds that

$$\sum_{t=1}^{T} \underbrace{V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1)}_{\text{prediction error}} \leq \left[d \sum_{t=1}^{T} \underbrace{\sum_{h=1}^{H} \sum_{s=1}^{t-1} \ell_h^s(f^t)}_{\text{training error}} \right]^{1/2} + \underbrace{2 \min\{Hd, H^2 T\} + \epsilon B^\dagger T}_{\text{burn-in cost}}.$$

We assume that $\ell_h^s(f^*) = 0$ holds for any $(s, h) \in [T] \times [H]$.

As compared to Definition 2.2.2, the generalized version allows different choices of in-sample error. In general, $\ell_h^s(\cdot)$ is determined by some specific function (e.g. Bellman error) and also a distribution induced by f^s. One can take $\ell_h^s(f) := \mathbb{E}_{\pi^s} \mathcal{E}_h(f, x_h, a_h)$ to recover the eluder coefficient. The following example shows that such a formulation strictly enriches the covered RL problems.

Example 2.3.3 (Linear mixture MDP has a low generalized eluder coefficient)**.** We consider the hypothesis space $\mathcal{H} = \{f = (\theta_{1,f}, \cdots, \theta_{H,f}) : \forall h \in [H], ||\theta_{h,f}|| \leq B\}$ and adopt the following loss function for linear mixture MDP:

$$\ell_h^s(f) := \mathbb{E}_{\pi_{f^s}} \left[\theta_{h,f}^\top \left[\psi(x_h, a_h) + \sum_{x' \in \mathcal{S}} \phi(x_h, a_h, x') V_{h+1,f^s}(x') \right] - r_h(x_h, a_h) - V_{h+1,f^s}(x_{h+1}) \right].$$

Then, for linear mixture MDP, its GEC satisfies $d(\epsilon) = \widetilde{\mathcal{O}}(Hd)$.

Proof. We defer the proof to Section 5.2. □

We can expand the loss function $\ell_h^s(f)$ for a clear interpretation:

$$\ell_h^s(f) = \mathbb{E}_{\pi_{f^s}}\Big[r_{h,f}(x_h,a_h) - r_{h,f^*}(x_h,a_h) \\ + \mathbb{E}_{x_{h+1}\sim\mathbb{P}_{h,f}(\cdot|x_h,a_h)}V_{h+1,f^s}(x_{h+1}) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}V_{h+1,f^s}(x_{h+1})\Big].$$

Clearly, the true model f^* achieves a loss of zero regardless of f^s. On the other hand, a hypothesis f suffers from a non-zero loss due to the difference in reward function and transition kernel. Another interesting example is the following Q^*-state abstraction model, where the agent faces many similar states so the "effective" state may be small.

Example 2.3.4 (Q^*-state abstraction model [57])**.** A MDP is said to be a Q^*-state abstraction model if there exists $\xi : \mathcal{S} \to \mathcal{K}$ so that for any $h \in [H]$,

$$\xi(x) = \xi(x') \Longrightarrow Q_h^*(x,a) = Q_h^*(x',a), \forall x, x', a \in \mathcal{S}\times\mathcal{S}\times\mathcal{A}.$$

Then, we first choose the following feature maps: $\phi(\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{|\mathcal{K}||\mathcal{A}|}$ and $\psi(\cdot) : \mathcal{S} \to \mathbb{R}^{|\mathcal{K}|}$,

$$\phi(x,a)[z,a'] = \mathbf{1}(\xi(x) = z, a = a'), \qquad \psi(x)[z] = \mathbf{1}(\xi(x) = z).$$

Accordingly, we can choose $\theta_h \in \mathbb{R}^{|\mathcal{K}||\mathcal{A}|}, w_h \in \mathbb{R}^{|\mathcal{K}|}$ and set the hypothesis class:

$$\mathcal{H}_h = \{||\theta_h||_2 \le B, ||w_h|| \le B : \max_{a\in\mathcal{A}} \phi(x,a)^\top\theta_h = \psi(x)^\top w_{h+1}, \forall x \in \mathcal{S}\}.$$

Then, the Q^*-state abstraction model has a low GEC of $\widetilde{O}(|\mathcal{K}||\mathcal{A}|)$ with

$$\ell_h^s(f) = \left(\mathbb{E}_{\pi_{f^s}}[\phi(x,a)^\top\theta_{h,f} - r_h - \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a)}\psi(x')^\top w_{h+1,f}]\right)^2.$$

Proof. We defer the proof to Section 5.2. □

In contrast to the tabular MDP which admits a GEC of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|)$, the similarity across different states leads to an easier game as we expect.

In [27], the authors show that there is an exponential separation between the value-based method and model-based method, where the latter explicitly uses or

studies the information of the transition kernel. To this end, in what follows, we show that we can handle the model-based approach in our framework.

We now introduce the witness rank as an example of model-based RL with function approximation. [27] studies the case where $\mathcal{H}_h$ is the hypothesis class of $\mathbb{P}_h$ (we assume that the reward function is known for simplicity). For witness rank, we adopt a discriminator class $\mathcal{V} = \{\mathcal{V}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]\}_{h\in[H]}$.

Definition 2.3.5 (Q-type Witness Rank)**.** We say an MDP has witness rank d if given two models $f, g \in \mathcal{H}$, there exists $X_h : \mathcal{H} \to \mathbb{R}^d$ and $W_h : \mathcal{H} \to \mathbb{R}^d$ such that

$$\begin{aligned}&\max_{v\in\mathcal{V}_h} \mathbb{E}_{\pi_f}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} v(x_h,a_h,x')] \geq \langle W_h(g), X_h(f)\rangle \\ &\kappa_{\text{wit}} \cdot \mathbb{E}_{\pi_f}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} V_{h+1,g}(x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} V_{h+1,g}(x')] \leq \langle W_h(g), X_h(f)\rangle\,,\end{aligned} \tag{2.3.2}$$

where $\kappa_{\text{wit}} \in (0,1]$ is a constant. Moreover, we assume that $\sup_{f\in\mathcal{H},h\in[H]} ||W_h(f)||_2 \leq B$ and $\sup_{f\in\mathcal{H},h\in[H]} ||X_h(f)||_2 \leq B$.

The definition presented here is slightly different from that of [27] and we will discuss the extension in Chapter 4. Intuitively, the left-hand side of (2.3.2) is the difference between some functions under the true model and the tested hypothesis model. Such a difference can be naturally reduced to the difference (distance) between the underlying distributions of these hypotheses when the involved functions are bounded, as verified by the following result.

Example 2.3.6 (Witness rank is controlled by GEC)**.** For the MDPs with a low witness rank d, if we take $\ell_h^s(f) = \mathbb{E}_{\pi_{fs}} D_H^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)$, we have

$$d(\epsilon) = \mathcal{O}\Big(dH \cdot \log\Big(1 + \frac{T}{\epsilon\kappa_{\text{wit}}^2}\Big)/\kappa_{\text{wit}}^2\Big).$$

Proof. We defer the proof to Section 5.2. □

Remark 2.3.7 (Q-type problem and V-type problem)**.** The witness rank can be defined for the Q-type problem and V-type problem. For the Q-type problem, we

mean that the expectation in the training loss $\ell_h^s(f^t)$ is taken with respect to the distribution used to collect the samples in s-th iteration, which is also referred to as the on-policy strategy [19]. Meanwhile, the V-type witness rank means that in $\ell_h^s(f^t)$, while the state follows the distribution at iteration s: $x_h \sim \pi_{f^s}$, the action may not follow π_{f^s}. Typically, the action is taken by following f^t: $a_h \sim \pi_{f^t}$ or $a_h \sim \mathrm{Unif}(\mathcal{A})$. Such a formulation is unified in the formulation of GEC but differs in the algorithmic design as we will detail in the next chapter.

To illustrate the idea of model-based function approximation, we introduce the factored MDPs [27].

Example 2.3.8 (Factored MDPs)**.** In factored MDPs, the state admits a factored structure. Specifically, we have $\mathcal{S} \subset \mathcal{O}^d$ where $\mathcal{O}$ is a discrete set and we use $x[i]$ to denote the i-th entry of the state. For each dimension, only a subset $\mathrm{pa}_i \subset [d]$ of entries will influence it and we call them the parent set of the i-th dimension. Mathematically, the transition probability is given by

$$\mathbb{P}_h(x'|x,a) = \prod_{i=1}^{d} \mathbb{P}_h^{(i)}(x'[i]|x[\mathrm{pa}_i], a), \qquad \forall (h,x,a,x') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S},$$

where $\mathbb{P}^{(i)}$ is the transition kernel from $x[\mathrm{pa}_i], a$ to x'. If we let the hypothesis space $\mathcal{H}$ contain all possible transitions, we can show that the factored MDP admits a GEC of $\widetilde{O}\big(H^3A^3\sum_{i=1}^d |\mathcal{O}|^{|\mathrm{pa}_i|}\big)$ with

$$\ell_h^s(f) = \mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\mathrm{Unif}(\mathcal{A})} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big).$$

Proof. We defer the proof to Appendix 5.2.3. □

[27] showed that there is an exponential separation between model-based and model-free RL in this rich-observation setting. Specifically, the OLIVE algorithm [34] requires $\Omega(2^H)$ samples to solve factored MDP in the worst case, which means that the Bellman rank or Bellman eluder dimension must be exponential in H.

With the above examples in hand, we observe that the framework of GEC can naturally handle both value-based (also referred to as the model-free) and model-based problems. However, there are still two problems before we can eventually solve the problems with a low GEC. First, on the left-hand-side of GEC, the prediction error is $\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1)$, while we care about the cumulative error $\sum_{t=1}^{T} V_1^*(x_1) - V_1^{\pi_{f^t}}(x_1)$. We briefly mentioned at the beginning of this chapter that in the literature we mainly addressed this mismatch by optimism, i.e., ensuring that $V_{1,f^t}(x_1) \geq V_1^*(x_1)$. But the implementation of such a "hard" optimism may lead to several disadvantages from the viewpoint of practice. Second, on the right-hand side of GEC, the training error is typically in an expected form, which is not directly available from the dataset and calls for the construction of appropriate estimators. We will handle these two issues in the next chapter with corresponding algorithmic designs.

Chapter 3

Maximize to EXplore: Algorithmic Design and Theoretical Guarantee

We introduce the Maximize to EXplore (MEX) in this chapter, which is inspired by the seminal work of feel-good Thompson sampling [32]. We start by addressing the two issues mentioned at the end of the last chapter.

3.1 Loss Estimation

We recall the definition of the generalized eluder coefficient:

$$\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1) \lesssim \left[d(\epsilon) \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{s=1}^{t-1} \ell_h^s(f^t) \right]^{1/2}$$

where we omit the burn-in cost for a clearer presentation. To achieve a low prediction error, it suffices to achieve a low cumulative in-sample error. To this end, we introduce the following loss estimator based on the collected samples.

Definition 3.1.1 (Loss estimator with batch sampling)**.** We consider a general

sampling strategy where we sample m i.i.d. trajectories $\{\zeta_{i,h}^k\}_{i=1}^m$ [1] at each stage k and assume that T is divisible by m without loss of generality. With $K := T/m$, for each iteration $k \in [K]$, we suppose that we have access to a loss estimator $L_h^{1:k-1}(\cdot) : \mathcal{H} \to \mathbb{R}$, which only depends on the history:

$$\{f^1, (\zeta_{i,h}^1)_{i=1}^m, f^2, (\zeta_{i,h}^2)_{i=1}^m, \cdots, f^{k-1}, (\zeta_{i,h}^{k-1})_{i=1}^m\}.$$

Moreover, it satisfies the following estimation error bound: with probability at least $1-\delta$, it holds that for all $(k, h, f) \in [K] \times [H] \times \mathcal{H}$

$$\sum_{s=1}^{k-1} \ell_h^s(f) \le L_h^{1:k-1}(f) + \Delta_h^k, \tag{3.1.1}$$

and

$$L_h^{1:k-1}(f^*) \le \Delta_h^k. \tag{3.1.2}$$

Intuitively speaking, (3.1.1) states that the loss estimator is an upper bound of the in-sample training loss if we introduce a suitable confidence interval. Meanwhile, the loss estimator should well approximate the in-sample loss of the ground truth up to the confidence interval.

The existence of such an estimator is trivial. For instance, when $\ell_h^s(\cdot)$ is bounded by C^2, one can always take $L_h^{1:k-1}(f) = 0$ and $\Delta_h^k = C^2 k$. On the other hand, we can do much better than this naive estimator, whose proofs are standard applications of concentration inequalities. We remark that the loss estimators introduced in this section are independently studied in the literature widely. The main purpose here is to provide a new interpretation of these algorithmic designs which better fit our framework.

We first focus on the value-based case, where the loss function is $\ell_h^s(f) = \big(\mathbb{E}_{x_h \sim \pi_{f^s}, a_h \sim \pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\big)^2$. As a motivating example, if we want to estimate $\sum_{s=1}^{k-1} (\mathbb{E} X_s)^2$ with a collection of samples $X_1, \cdots, X_{k-1}$ from the underlying dis-

[1] When $m = 1$, we omit the subscript i for simplicity, which should be clear from the context.

tribution sequence, we cannot directly use the loss estimator $\sum_{s=1}^{k-1} X_s^2$ because

$$\mathbb{E}\sum_{s=1}^{k-1} X_s^2 = \underbrace{\sum_{s=1}^{k-1}(\mathbb{E}X_s)^2}_{\text{Goal}} + \underbrace{\sum_{s=1}^{k-1}\sigma_s^2}_{\text{Sampling variance}} .$$

The error term of sampling variance grows linearly in the time steps and makes it unaffordable. To address this issue, one straightforward idea is to replace X_s with a sample mean to achieve a low variance.

Lemma 3.1.2 (In-sample error estimation with trajectory average [34; 19])**.** Suppose that $\ell_h^s(f) = \big(\mathbb{E}_{x_h\sim\pi_{f^s},a_h\sim\pi_{f^s}}\mathcal{E}_h(f,x_h,a_h)\big)^2$. We can independently sample m trajectories $\{\zeta_{i,h}^k\}_{i=1}^m$ by following π_{f^k} for each $k\in[K]$ and take

$$L_h^{1:k-1}(f) = \sum_{s=1}^{k-1} L_h^s(f) := 2\sum_{s=1}^{k-1}\Big[\frac{1}{m}\sum_{i=1}^{m}\Big(Q_{h,f}(x_{i,h}^s,a_{i,h}^s) - r_{i,h}^s - V_{h+1,f}(x_{i,h+1}^s)\Big)\Big]^2,$$

where it satisfies (3.1.1) and (3.1.2) with $\Delta_h^k = \frac{4(k-1)H^2\iota_h}{m}$, and $\iota_h = O(\log(KH|\mathcal{H}_h|/\delta))$.

Proof. We denote $\epsilon_h^s(f) = \frac{1}{m}\sum_{i=1}^m(Q_{h,f}(x_{i,h}^s,a_{i,h}^s) - r_{i,h}^s - V_{h+1,f}(x_{i,h+1}^s))$ for notation simplicity. For each fixed s,h,f_h, the Azuma-Hoeffding inequality implies that with probability at least $1-\delta/(KH|\mathcal{H}_h|)$, we have

$$\Big|\epsilon_h^s(f) - \mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f,x_h,a_h)\Big| \le H\sqrt{\frac{2\log(KH|\mathcal{H}_h|/\delta)}{m}}.$$

By $(a+b)^2\le 2a^2+2b^2$, we further have

$$\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f,x_h,a_h)\big)^2 \le 2\big(\epsilon_h^s(f)\big)^2 + \frac{4H^2\log(KH|\mathcal{H}_h|/\delta)}{m}.$$

Taking a union bound over $[K]$, $\mathcal{H}_h$ and then $h\in[H]$, with probability at least $1-\delta$, the inequality holds for all $(s,h,f)\in[K]\times[H]\times\mathcal{H}$. Therefore, it satisfies that

$$\sum_{s=1}^{k-1}\ell_h^s(f) \le L_h^{1:k-1}(f) + \frac{4(k-1)H^2\log(KH|\mathcal{H}_h|/\delta)}{m}.$$

To prove (3.1.2), we note $\mathcal{E}_h(f^*,x_h,a_h)=0$ for any (x_h,a_h). □

The guarantee provided by Lemma 3.1.2 can be sub-optimal since we use m samples for one hypothesis choice f^t. We should view the Lemma 3.1.2 as

$$m \cdot \sum_{s=1}^{k-1} \ell_h^s(f) \leq m \cdot L_h^{1:k-1}(f) + 4(k-1)H^2 \cdot \log(KH|\mathcal{H}_h|/\delta).$$

For a value-based approach, we can obtain a sharper estimator with the following Bellman completeness condition.

Assumption 3.1.3 (Bellman Completeness)**.** We consider a value-based hypothesis $\mathcal{H}_h = \mathcal{F}_h \subset \{f_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. The hypothesis class is said to be Bellman complete, if for each $h \in [H]$, $\mathcal{T}_h^* \mathcal{F}_{h+1} \subset \mathcal{F}_h$, where $\mathcal{T}_h^* \mathcal{F}_{h+1} = \{\mathcal{T}_h f_{h+1} : f_{h+1} \in \mathcal{F}_{h+1}\}$ and

$$(\mathcal{T}_h^* f_{h+1})(x,a) := r_h(x,a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} \max_{a'} f_{h+1}(x',a').$$

Bellman completeness is stronger than the realizability since by taking $f_{H+1} = 0$, we can show that $f^* \in \mathcal{F}$ by Bellman completeness. However, completeness itself is not desirable because adding a new function into the function class can destroy such a property.

Lemma 3.1.4 (In-sample error estimation with minimax formulation [50; 18; 28; 51])**.** Suppose that $\ell_h^s(f) = \left(\mathbb{E}_{x_h \sim \pi_{f^s}, a_h \sim \pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. We set $m = 1$ so $T = K$ in this case. For each $t \in [T]$, we can collect the trajectory ζ^t by following π_{f^t} and take $L_h^{1:t}(f) := \sum_{s=1}^t L_h^s(f)$ where

$$L_h^s(f) = \left(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2 - \inf_{f_h' \in \mathcal{H}_h} \left(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2,$$

where it satisfies (3.1.1) and (3.1.2) with $\Delta_h^t = O(H^2 \iota_h)$ and $\iota_h = \log(H|\mathcal{H}_h|T/\delta)$.

Proof. We defer the proof to Section 5.3. □

The main intuition of the minimax formulation is that it allows us to consider the following loss function in the theoretical analysis:

$$\widetilde{L}_h^s(f) = \left(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2 - \left(\mathcal{T}_h V_{h+1,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2,$$

where the conditional expectation of the second term is exactly the sampling variance, thus canceling the variance term. The reason is that we can approximate the second term (which is not available to the actually executed algorithm) by $\inf_{f'_h \in \mathcal{H}_h} \left(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2$ and the Bellman completeness condition. The loss estimator in Lemma 3.1.4 is referred to as minimax formulation [50] because it can be also written as

$$\min_{f \in \mathcal{H}} \max_{f' \in \mathcal{H}} \sum_{h=1}^{H} \left[\left(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2 - \left(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2 \right].$$

Remark 3.1.5. Both Lemma 3.1.2 and Lemma 3.1.4 can ensure a small average loss $\sum_{s=1}^{t-1} \left(\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. However, the minimax formulation further ensures a small $\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left(\mathcal{E}_h(f, x_h, a_h)\right)^2$, while the trajectory average cannot.

In addition to the model-free approach, we also handle the model-based case in the following lemma.

Lemma 3.1.6 (In-sample error estimation of model-based method)**.** Suppose that $\ell_h^s(f) = \mathbb{E}_{\pi_{f^s}} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)$. We set $m = 1$ so $T = K$ in this case. For each $h \in [H]$, we can take $L_h^{1:t}(f) := \sum_{s=1}^{t} L_h^s(f) := \frac{1}{2} \sum_{s=1}^{t} -\log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)$. However, implementing Algorithm 1 with $L_h^s(\cdot)$ does not lead to a condition as required in Definition 3.1.1 for $L_h^s(\cdot)$ itself. The key observation is that adopting $L_h^s(\cdot)$ is equivalent to implementing Algorithm 1 with $\widetilde{L}_h^s(f) = -\frac{1}{2} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s) + \frac{1}{2} \log \mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)$ because we subtract the same amount of loss for all hypotheses. Therefore, we can use it in our theoretical analysis, and it satisfies (3.1.1) and (3.1.2) with $\Delta_h^t = \log(H|\mathcal{H}_h|/\delta)$.

Proof. We defer the proof to Section 5.3. □

The idea of using an equivalent $\widetilde{L}_h^s(\cdot)$ can be viewed as introducing a baseline in the loss estimator. As long as the baseline is fixed for all $f \in \mathcal{H}$, it leads to an equivalent algorithm.

Based on the plug-in principle, one can adopt the following choice of hypothesis for each iteration k:

$$f^k := \operatorname*{argmin}_{f \in \mathcal{H}} L_h^{1:k-1}(f). \tag{3.1.3}$$

The model-free version is referred to as the fitted Q-iteration (FQI) in the literature [58]. We may call (3.1.3) as Maximize to EXplore (MEX) since it only involves an optimization sub-routine to balance the exploration and exploitation simultaneously. MEX picks the hypothesis that best approximates the ground truth. Unfortunately, this does not ensure a low regret bound or a mild sample complexity. This is because a low cumulative prediction error does not necessarily lead to a low regret because of the difference between $V_{1,f}$ and V_1^*.

To alleviate this issue, we need to adopt the principle of optimism into the algorithmic framework, as detailed in the next subsection.

3.2 The Power of Optimism

In the literature, the terminology optimism is referred to as the inequality $V_{1,f^t}(x_1) \geq V_1^*(x_1)$, which means that our choice of the hypothesis always takes an optimistic estimation about the initial value. The technical reason for such an algorithmic choice is that (we omit x_1):

$$\sum_{t=1}^{T} V_1^* - V_1^{\pi_{f^t}} = \sum_{t=1}^{T} \left(V_{1,f^t} - V_1^{\pi_{f^t}}\right) + \sum_{t=1}^{T} \left(V_1^* - V_{1,f^t}\right) \leq \sum_{t=1}^{T} \left(V_{1,f^t} - V_1^{\pi_{f^t}}\right), \tag{3.2.1}$$

where the last term can be further related to the in-sample training error via the GEC. To address the mismatch between $V_{1,f}$ and V_1^*, instead of achieving optimism directly, we modify the MEX by adding another "feel-good" term into the objective:

$$f^t = \operatorname*{argmax}_{f \in \mathcal{H}} \left[V_{1,f}(x_1) - \eta \sum_{s=1}^{t-1} \sum_{h=1}^{H} L_h^s(f)\right], \tag{3.2.2}$$

where $\eta > 0$ is a hyper-parameter to control the relative importance of the feel-good term. The technical consideration of this modification is inspired by [32]. The difference is that we analyze an optimization-based framework, while [32] focuses on the sampling-based one. Combining this new objective with the error estimation introduced in the last section, we obtain the MEX in Algorithm 1, where we further consider the more general batch sampling strategy.

Algorithm 1 Maximize to EXplore (MEX)

1: **Input:** Hypothesis space $\mathcal{H}$, $\eta > 0$, batch size $m > 0$.

2: **for** $k = 1, 2, \cdots, K := T/m$ **do**

3: Select f^k by solving

$$f^k = \operatorname*{argmax}_{f \in \mathcal{H}} \Big[V_{1,f}(x_1) - \eta \sum_{h=1}^{H} L_h^{1:k-1}(f) \Big]. \tag{3.2.3}$$

4: For each $h \in [H]$, collect a batch of dataset $\{\zeta_{i,h}^k\}_{i=1}^m$ by following π_{f^k} m times.

5: **end for**

Comparison to existing OFU-based algorithms. The algorithm can be viewed as a unification of the BiLin-UCB [19], GOLF [18], and OMLE [59], where the generality is mainly from the flexible choice of the loss function. However, the main difference is that we do not explicitly maintain a confidence set and perform a constraint optimization subroutine on it. Instead, the optimism is achieved *implicitly* via the feel-good term in the objective function. Meanwhile, such a formulation can be viewed as the Lagrange relaxation of the constraint optimization:

$$\begin{aligned} &\max_{f \in \mathcal{F}} V_{1,f} \\ &\text{subject to } \sum_{h=1}^{H} L_h^{1:k-1}(f) \le \beta^k, \end{aligned} \tag{3.2.4}$$

where $\beta^k > 0$ is the confidence radius of the confidence set. To the best of our knowledge, such an implicit formulation is new in the literature of online

RL and we believe that it helps to better illustrate the role of optimism in the algorithmic design. We mention in passing that [60] has a similar algorithmic idea in the pessimism-based offline setting. Combining this with the in-sample error estimation bound, we can establish the main result of this paper.

Theorem 3.2.1. Under Assumptions 1.2.4 and 1.2.8, we consider the problem with a low GEC $d(\epsilon)$ and the loss estimator given in Definition 3.1.1 with estimation interval Δ_h^k. Then, with $\eta = \sqrt{d(\epsilon)/(2\sum_{h=1}^H \sum_{k=1}^K \Delta_h^k)} > 0$, Algorithm 1 satisfies that with probability at least $1-\delta$,

$$\mathrm{Reg}(T) \le 2\sqrt{2}m\sqrt{d(\epsilon)\sum_{h=1}^H\sum_{k=1}^K \Delta_h^k} + 2m\cdot\min\{Hd, H^2K\} + \epsilon B^\dagger T,$$

where $T = mK$.

Proof. We recall that the batch size is m, and the total iteration is $K := T/m$. We denote $\Delta V_{1,f}(x_1) = V_{1,f}(x_1) - V_1^*(x_1)$. It follows that

$$\begin{aligned}
&\sum_{k=1}^K V_1^*(x_1) - V_1^{\pi^k}(x_1) := \sum_{k=1}^K V_{1,f^k}(x_1) - V_1^{\pi^k}(x_1) - \Delta V_{1,f^k}(x_1)\\
&\le -\sum_{k=1}^K \Delta V_{1,f^k}(x_1) + \eta\sum_{h=1}^H\sum_{k=1}^K\Big(\sum_{s=1}^{k-1}\ell_h^s(f^k)\Big) + \frac{1}{\eta}\cdot d + 2\min\{Hd, H^2K\} + \epsilon B^\dagger K\\
&\le -\sum_{k=1}^K \Delta V_{1,f^k}(x_1) + \eta\sum_{h=1}^H\sum_{k=1}^K\Big(L_h^{1:k-1}(f^k)\Big) + \eta\sum_{h=1}^H\sum_{k=1}^K\Delta_h^k\\
&\quad + \frac{1}{\eta}\cdot d + 2\min\{Hd, H^2K\} + \epsilon B^\dagger K,
\end{aligned} \tag{3.2.5}$$

where the first inequality follows from the definition of GEC and Cauchy-Schwarz inequality with a tuning parameter $\eta > 0$. The second inequality follows from the definition of loss estimator (3.1.1). By the selection rule of Algorithm 1 (i.e., (3.2.3)), we know that

$$-\Delta V_{1,f^k}(x_1) + \eta\cdot\sum_{h=1}^H L_h^{1:k-1}(f^k) \le \eta\cdot\sum_{h=1}^H L_h^{1:k-1}(f^*) \le \eta\sum_{h=1}^H\Delta_h^k, \tag{3.2.6}$$

where the last inequality uses the definition of loss estimator (3.1.2) and $\ell_h^s(f^*) \equiv 0$. We plug (3.2.6) into (3.2.5) to obtain that

$$\sum_{k=1}^{K} V_1^*(x_1) - V_1^{\pi^k}(x_1) \le 2\eta \sum_{h=1}^{H}\sum_{k=1}^{K} \Delta_h^k + \frac{1}{\eta}\cdot d + 2\min\{Hd, H^2K\} + \epsilon B^\dagger K$$
$$= 2\sqrt{2d\sum_{h=1}^{H}\sum_{k=1}^{K}\Delta_h^k + 2\min\{Hd, H^2K\}} + \epsilon B^\dagger K,$$

where we take $\eta = \sqrt{d/(2\sum_{h=1}^{H}\sum_{k=1}^{K}\Delta_h^k)} > 0$. Since for each iteration, we sample m trajectories in total, the final regret is given by:

$$\mathrm{Reg}(T) \le 2\sqrt{2}m\sqrt{d\sum_{h=1}^{H}\sum_{k=1}^{K}\Delta_h^k + 2m\min\{Hd, H^2K\}} + \epsilon B^\dagger T.$$

□

The theorem shows that the regret heavily relies on two components: (i) generalized eluder coefficient d, which measures the cost of transforming the out-of-sample error to the in-sample error; (ii) Δ_h^k, which is the in-sample estimation error over the hypothesis space $\mathcal{H}$ (more specifically, a trade-off between the batch size m and the estimation error).

We can combine the in-sample error bounds to obtain the following corollaries.

Corollary 3.2.2 (Regret bound of model-free approach with trajectory average)**.** Under the same condition of Theorem 3.2.1, we suppose that we have access to a loss estimator in Lemma 3.1.2. Then, when $HK > d$, by taking the batch size as $m = T^{\frac{2}{3}}\iota^{\frac{1}{3}}d^{-\frac{1}{3}}$ and iteration $K = T^{\frac{1}{3}}\iota^{-\frac{1}{3}}d^{\frac{1}{3}}$, with probability at least $1-\delta$, it holds that

$$\mathrm{Reg}(T) \lesssim d\big(1/(T^{1/3}B^\dagger)\big)^{\frac{2}{3}} \cdot HT^{\frac{2}{3}}\iota^{\frac{1}{3}},$$

where $\iota = \mathcal{O}(\log(|\mathcal{H}|KH/\delta))$.

Corollary 3.2.3 (Regret bound of model-free approach with minimax formulation)**.** Under the same condition of Theorem 3.2.1, we suppose that we have

access to a loss estimator in Lemma 3.1.4. Then, when $HK > d$, by taking the batch size as $m = 1$, with probability at least $1 - \delta$, it holds that

$$\mathrm{Reg}(T) \lesssim \sqrt{d\big(1/(\sqrt{T}B^{\dagger})\big) \cdot H^2 T\iota}.$$

where $\iota = O(\log(|\mathcal{H}|TH/\delta))$.

Corollary 3.2.4 (Regret bound of model-based approach)**.** Under the same condition of Theorem 3.2.1, we suppose that we have access to a loss estimator in Lemma 3.1.6. Then, when $HK > d$, by taking the batch size as $m = 1$, with probability at least $1 - \delta$, it holds that

$$\mathrm{Reg}(T) \lesssim \sqrt{d\big(1/(\sqrt{T}B^{\dagger})\big) \cdot T\iota}.$$

where $\iota = O(\log(|\mathcal{H}|H/\delta))$.

We note that the model-based approach gives a sharper in-sample training error estimation than the model-free approach under Assumption 1.2.4, thus a sharper regret bound in T (we remark that some H-dependence is hidden in the GEC of witness rank). However, we would like to remark that the model-based realizability is indeed much stronger than the model-free one. Suppose that we are given a model class $\mathcal{M}$ such that the true model $M^* \in \mathcal{M}$. We can take $\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_H$ and $\mathcal{H}_h = \{Q_{h,M} : M \in \mathcal{M}\} \cup \{(\mathcal{T}_h^M \max_{a' \in \mathcal{A}} f_{h+1}(\cdot, a')) : f_{h+1} \in \mathcal{H}_{h+1}\}$ for all $M \in \mathcal{M}$, where $\mathcal{T}_h^M$ is the Bellman operator under model M. By doing so, we construct a value-based hypothesis $\mathcal{H}$ satisfying realizability and Bellman completeness assumptions and $|\mathcal{H}| = |\mathcal{M}|^2$.

Chapter 4

Discussion, Potential Extension, and Limitations

In this chapter, we compare our results with existing works, and discuss the potential extensions and the limitations.

4.1 Relationship with Eluder Dimension

The eluder coefficient is closely related to the notion of the (distributional) eluder dimension. In the context of RL, [18] applies the distributional eluder dimension to the value-based hypothesis space and proposes the Bellman eluder dimension. We now comment on the similarities and differences between them as follows.

We start by introducing the eluder dimension. [48], which is a generalization of the linear independence for measuring the complexity of a general value-based function class $\mathcal{F}$. Different from the reductions presented in Chapter 2, whose proof heavily relies on the linear structure, we discuss the relationship between the eluder dimension and eluder coefficient in this section. We start with the

definition of ϵ-dependence.

Definition 4.1.1 (ϵ-independence between distributions)**.** Let $\mathcal{G}$ be a function class defined on $\mathcal{Z}$, and $\nu, \mu_1, \cdots, \mu_n$ be probability measures over $\mathcal{Z}$. We say ν is ϵ-independent of $\{\mu_1, \mu_2, \cdots, \mu_n\}$ 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 4.1.2 (Distributional eluder (DE) dimension)**.** Let $\mathcal{G}$ be a function class defined on $\mathcal{Z}$, and Π be a family of probability measures over $\mathcal{Z}$. The distributional eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \cdots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \geq \epsilon$ with ρ_i being ϵ'-independent of $\{\rho_1, \cdots, \rho_{i-1}\}$ for all $i \in [n]$.

Intuitively, for any $f \in \mathcal{F}$, the ϵ-dependence of the sequence means that if it is relatively consistent on the historical distributions $\{\mu_1, \cdots \mu_n\}$ (i.e., $(\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2 \leq \epsilon^2)$, the error on the new test distribution ν will also be small. On the other hand, independence means that while g is consistent on the historical dataset, it can suffer from a large prediction error in the newly arrived point z, which is not desirable for our needs. The distributional eluder dimension simply says that independence cannot happen too many times. The following lemma shows that a problem with a low distributional eluder dimension also has a low eluder coefficient.

Lemma 4.1.3. Suppose that a problem has a distributional eluder dimension of $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ and suppose that $|g|$ is bounded by H for all $g \in \mathcal{G}$. Then, the eluder coefficient satisfies $d(\epsilon) \leq \mathcal{O}(\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon) \log T)$ in the following sense: for arbitrary sequence of $\{(d_t, g_t) \in \Pi \times \mathcal{G}\}_{t=1}^T$, we have

$$\sum_{t=1}^T |\mathbb{E}_{d_t} g_t| \leq \sqrt{d(\epsilon) \sum_{t=1}^T \sum_{s=1}^{t-1} \left(\mathbb{E}_{d_s} g_t\right)^2} + \min\{Hd(\epsilon), H^2 T\} + \epsilon HT.$$

The proof basically follows from [28] except that we need to use a more abstract notion of function class and distribution to make the proof work in a more general

sense. In particular, [61] uses the techniques from [28] to additionally handle the V-type problems. We omit the proof for simplicity. In contrast, we note that the problems with a low eluder coefficient can have a large eluder dimension, as shown by the following lemma adapted from [62].

Lemma 4.1.4. Fix the time horizon $T > 0$ and $H = 2$. Let $\mathcal{G}_h := \{Q_{h,f} - \mathcal{T}_h V_{h+1,f} : f \in \mathcal{H}\}$ be the set of Bellman residuals induced by $\mathcal{H}$ at step h, and Π_h be a collection of probability measure families over $\mathcal{S} \times \mathcal{A}$ induced by following π_f, with $f \in \mathcal{H}$. Then, there exists a class of MDPs such that the distributional eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}_h, \Pi_h, 1/T^{1/3})$ is lower bounded by $\Omega(T^{1/3})$, while the eluder coefficient $d(\epsilon)$ is upper bounded by $\mathcal{O}(\log T)$, regardless of the ϵ, in the sense of: for an arbitrary sequence of $\{(f^t) \in \mathcal{H}\}_{t=1}^T$, we have

$$\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1) \le \sqrt{d(\epsilon) \sum_{t=1}^{T} \sum_{h=1}^{2} \sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, x_h, a_h)^2} + \min\{Hd(\epsilon), H^2 T\} + \epsilon HT.$$

Comparison with eluder dimension in terms of generality. Both the eluder coefficient and eluder dimension limit the degree to which we can be surprised by the unseen trajectory $\zeta^t \sim \pi_{f^t}$ given the historical dataset $\{\zeta^1, \cdots, \zeta^{t-1}\}$. However, the eluder coefficient further takes the magnitude of the prediction error into consideration, while the eluder dimension only considers the frequency. In particular, Lemma 4.1.4 shows that there exists an exponential separation between the eluder coefficient and eluder dimension in certain cases.

Comparison with eluder dimension in terms of applicability. It is worth noting that in the literature of eluder dimension [48; 49; 18], to apply the eluder dimension in analysis, the algorithms must attain an increasing sequence of upper bounds for the in-sample error (e.g. $\sum_{s=1}^{t-1} (\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, x_h, a_h)^2 \le \beta_t$, β_t is increasing in t). This prevent the eluder dimension from being used to analyze Algorithm 1 and also the posterior sampling [28; 53; 61]. On the other hand, the eluder coefficient can be used to analyze all of these algorithms.

4.2 Comparison with Other Existing Works

Comparison with Bellman eluder dimension [18] and bilinear class [19]. We have present the detailed comparison with the eluder dimension in last section so we focus on the bilinear class here.

Definition 4.2.1 (Bilinear Class)**.** Given an MDP, a hypothesis class $\mathcal{H}$, and a discrepancy function $l = \{l_f : \mathcal{H} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \times \mathcal{H} \to \mathbb{R}\}_{f \in \mathcal{H}}$, we say the RL problem is in a bilinear class if there exist functions $W_h : \mathcal{H} \to \mathbb{R}^d$ and $X_h : \mathcal{H} \to \mathbb{R}^d$, such that for all $f \in \mathcal{H}$ and $h \in [H]$, we have

$$\begin{aligned} \left|\mathbb{E}_{\pi_f}\left[Q_{h,f}\left(x_h, a_h\right) - r\left(x_h, a_h\right) - V_{h+1,f}\left(x_{h+1}\right)\right]\right| &\leq \left|\left\langle W_h(f) - W_h\left(f^*\right), X_h(f)\right\rangle\right|, \\ \left|\mathbb{E}_{x_h \sim \pi_f, a_h \sim \widetilde{\pi}}\left[l_f\left(g, x_h, a_h, r_h, x_{h+1}\right)\right]\right| &= \left|\left\langle W_h(g) - W_h\left(f^*\right), X_h(f)\right\rangle\right|, \end{aligned} \tag{4.2.1}$$

where $\widetilde{\pi}$ is either π_f (Q-type) or π_g (V-type). Moreover, it is required that $\sup_{f \in \mathcal{H}, h \in [H]} ||W_h(f)||_2 \leq B$ and $\sup_{f \in \mathcal{H}, h \in [H]} ||X_h(f)||_2 \leq B$.

Similar to the witness rank [27], the bilinear class also assumes that the MDP admits certain bilinear structure so can be reduced to the GEC with similar techniques. We omit the proof to avoid repetition and refer interested readers to [61] for the complete proof. For bilinear class, the GEC satisfies $d(\epsilon) = \widetilde{O}(dH)$ with the discrepancy loss $\ell_h^s(f^t) := (\mathbb{E}_{x_h \sim \pi_{f^s}, a_h \sim \widetilde{\pi}} l_{f^s}(f^t, x_h, a_h, r_h, x_{h+1}))^2$. Therefore, GEC subsumes the bilinear class as a subset.

Comparison with DEC [33]. The technical consideration of DEC (1.3.5) is to convert the RL problems into an online learning problem, by reducing the out-of-sample regret to another out-of-sample divergence $D^2_{\mathrm{H}}(M(\pi), \widehat{M}(\pi))$. In contrast, GEC reduces the out-of-sample regret to an in-sample training error over the past $t-1$ iterations. Both GEC and DEC can capture most of the known tackable RL problems so far. But since there exists the matching lower bound in terms of DEC in some decision-making problems, it is possible to construct some

instance that is covered by DEC but not GEC. In terms of algorithmic design, however, DEC is mainly used to analyzed the E2D algorithm proposed in [33] and cannot apply to the classic posterior sampling algorithms and OFU-based algorithms. In contrast, GEC can be used to analyze both of them, and also the MEX algorithm proposed in this paper. Meanwhile, the E2D algorithm requires to solve a minimax optimization problem so lacking general implementation guidance in practice, while the MEX algorithm presented in this paper can be readily approximated as shown in [63]. Moreover, the regret bounds obtained from the framework of DEC are often sub-optimal. For instance, according to [52], for the bilinear class with only realizability, DEC only gives a bound of order $T^{3/4}$, while GEC gives a bound of order $T^{2/3}$.

Comparison to the sampling-based algorithmic frameworks. It is possible to apply GEC to the posterior sampling framework. [61] extends the conditional posterior sampling proposed in [28] to a more generic algorithmic framework and can also handle all the problems studied in this paper. In comparison, the sampling-based framework in [61] requires knowledge of the environment by assuming that we have access to a good prior distribution that approximately satisfies the Assumption 1.2.4 and the Assumption 3.1.3. In contrast, the optimization-based framework presented in this paper does not require knowledge of the environment. However, we do remark that while the algorithms in [61] and this paper achieve similar theoretical guarantees, empirical studies show that sampling-based algorithms are usually superior in practice (including [64] for bandit, and [65] for RL).

4.3 Extension and Challenges

We discuss several extensions and limitations of the framework presented in this paper.

4.3.1 V-type Variant

In the literature (e.g. [34; 18]), the examples introduced in the previous sections are referred to as the "Q-type" problems, which means that the expectation in the training loss $\ell_h^s(f^t)$ is taken with respect to the distribution used to collect the samples in s-th iteration. Meanwhile, in the literature, we also consider the V-type problems where main difference is that for the V-type problem, in $\ell_h^s(f^t)$, while the state follows the distribution at iteration s: $x_h \sim \pi_{f^s}$, the action is taken by following f^t: $a_h \sim \pi_{f^t}$. Such a formulation is unified in the formulation of GEC by an abstract choice of the loss function. To illustrate the idea, we consider the value-based case with

$$\ell_h^s(f^t) = \left(\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\pi_{f^s}}\mathcal{E}_h(f^t, x_h, a_h)\right)^2 \qquad \text{v.s.} \qquad \left(\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\pi_{f^t}}\mathcal{E}_h(f^t, x_h, a_h)\right)^2.$$

However, the construction of the loss estimator $L_h^{1:t}(\cdot)$ can be different, and in particular, we need to perform some special sample collection techniques. At iteration $s < t$, we collect trajectories with some exploration policy $\pi_{\text{exp}}(f^s, h)$, and we need to use these trajectories to estimate $\ell_h^s(f)$ in the future (e.g. t-th iteration with $t > s$). The problem is that, the action a_h^s sampled at iteration s may not be identical to the greedy action of f^t taken in the future. To address this "mismatch", we need to adopt some special exploration strategy so that the distribution of (x_h^s, a_h^s) can be used to estimate $\ell_h^s(f^t)$, even though f^t is not known at iteration s. When the action space is finite, we can adopt one step of uniform exploration in the action space: $x_h^s \sim \pi_{f^s}, a_h^s \sim \text{Unif}(\mathcal{A})$. In this case, we modify the loss estimation in Lemma 3.1.2 by

$$L_h^s(f) = \left[\frac{1}{m}\sum_{i=1}^m \frac{\mathbf{1}(a_{i,h} = \pi_f(x_{i,h}))}{1/A}\left(Q_{h,f}(x_{i,h}^s, a_{i,h}^s) - r_{i,h}^s - V_{h+1,f}(x_{i,h+1}^s)\right)\right]^2, \tag{4.3.1}$$

where the loss estimator satisfies the condition in Definition 3.1.1 with $\Delta_h^k = \frac{4(k-1)H^2A^2\iota_h}{m}$. The proof is almost identical to that of Lemma 3.1.2 except for a different range so we omit it. On the other hand, for the model-based approach,

one can use the same error estimator for the V-type problems but rescaled by a factor of A, mainly due to the fact that the expectation is at the outside of the square:

$$\begin{aligned}\ell_h^s(f) &= \mathbb{E}_{x_h\sim\pi_{f^s},a_h\sim\pi_f} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h,a_h),\mathbb{P}_{h,f^*}(\cdot \mid x_h,a_h)\big)\\ &\le \mathbb{E}_{x_h\sim\pi_{f^s}} \sum_{a_h\in\mathcal{A}} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h,a_h),\mathbb{P}_{h,f^*}(\cdot \mid x_h,a_h)\big)\\ &= A\cdot \mathbb{E}_{x_h\sim\pi_{f^s}}\mathbb{E}_{a_h\sim\mathrm{Unif}(\mathcal{A})} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h,a_h),\mathbb{P}_{h,f^*}(\cdot \mid x_h,a_h)\big).\end{aligned}$$

We note that when the expectation is at the outside of the square, the target is non-negative, which is the key to change-of-measure arguments. Then, we modify the loss estimator in Lemma 3.1.6 as follows:

$$L_h^{1:t}(f) = \frac{A}{2}\sum_{s=1}^{t} -\log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s),$$

which satisfies Definition 3.1.1 with $\Delta_h^t = A\log(H|\mathcal{H}_h|/\delta)$. We also include the proof in Section 5.3. A similar observation holds for the Bellman complete case, which also results from the non-negative of $\mathcal{E}_h(f,x_h,a_h)^2$. We refer readers to [61] for a more detailed discussion.

When the action space is infinite, additional structural assumptions (referred to as the linearly embeddable Bellman error) and a more advanced exploration strategy are required to handle the V-type problems. These techniques are presented in [54].

4.3.2 Multi-agent Variant

We can extend the GEC to the multi-agent case. We illustrate the idea in the case of a two-player zero-sum Markov game, as done by [66]. We focus on the equilibrium computation case, where there exists a central controller that decides the behaviors of all players. For simplicity, we consider the value-based setting.

Markov Games (MGs) generalize the MDPs to the multi-agent setting. We

consider the episodic two-player zero-sum MG in this subsection, denoted as $MG(H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$, which additionally incorporates the action space of the second player ($\mathcal{B}$). The two players are referred to as the max-player and the min-player, respectively, and now the transition kernel $\mathbb{P}_h(\cdot|x, a, b)$ and the reward function (for the max-player) $r_h(x, a, b) \in [0, 1]$ are jointly determined by these two players. For the max-player, a Markov policy is a map from $\mathcal{S}$ to a distribution over $\mathcal{A}$ and we define it similarly for the min-player but with action space $\mathcal{B}$. We can similarly define the value functions for a policy pair (μ, ν) as

$$V_h^{\mu,\nu}(x) = \mathbb{E}_{\mu,\nu}\Big[\sum_{h'=h}^{H} r_{h'}\left(x_{h'}, a_{h'}, b_{h'}\right) \mid x_h = x\Big]$$
$$Q_h^{\mu,\nu}(x, a, b) = \mathbb{E}_{\mu,\nu}\Big[\sum_{h=h}^{H} r_{h'}\left(x_{h'}, a_{h'}, b_{h'}\right) \mid (x_h, a_h, b_h) = (x, a, b)\Big].$$

Best Response. For any policy of max-player μ, a corresponding best response for the min-player can be found, denoted as $\nu^\dagger(\mu)$, such that $V_h^{\mu,\nu^\dagger(\mu)}(x) = \inf_\nu V_h^{\mu,\nu}(x)$ for all (x, h). Similarly, for a min-player policy ν, there exists a best response for the max-player, denoted as $\mu^\dagger(\nu)$, such that $V_h^{\mu^\dagger,\nu}(x) = \sup_\mu V_h^{\mu,\nu}(x)$ for all (x, h). To simplify the notation, we use

$$V_h^{\mu,\dagger}(x) := V_h^{\mu,\nu^\dagger(\mu)}(x), Q_h^{\mu,\dagger}(x, a, b) := Q_h^{\mu,\nu^\dagger(\mu)}(x, a, b);$$
$$V_h^{\dagger,\nu}(x) := V_h^{\mu^\dagger(\nu),\nu}(x), Q_h^{\dagger,\nu}(x, a, b) := Q_h^{\mu^\dagger(\nu),\nu}(x, a, b).$$

Nash Equilibrium. Moreover, there exists a set of Nash equilibrium (NE) policies (μ^*, ν^*) [67] that are optimal against their best response such that

$$V_h^{\mu^*,\dagger}(x) = \sup_\mu V_h^{\mu,\dagger}(x), \quad V_h^{\dagger,\nu^*}(x) = \inf_\nu V_h^{\dagger,\nu}(x),$$

for all $(x, h) \in \mathcal{S} \times [H]$. For this NE, the following famous minimax equation holds:

$$\sup_\mu \inf_\nu V_h^{\mu,\nu}(x) = V_h^{\mu^*,\nu^*}(x) = \inf_\nu \sup_\mu V_h^{\mu,\nu}(x)$$

for all $(x, h) \in \mathcal{S} \times [H]$. For simplicity, we denote $V_h^*(x) := V_h^{\mu^*,\nu^*}(x)$ and

$Q_h^*(x) := Q_h^{\mu^*,\nu^*}(x)$. Note that although there might exist multiple NE policies, the NE value function is unique for a zero-sum MG.

Performance metrics with the exploiter. We analyze a special setting where we adopt a non-symmetric structure in the max-player and min-player so that the min-player (exploiter) serves as an exploiter to exploit the weakness of the max-player (main agent). This innovative idea is from [68; 69] and is motivated the practical self-play training. To this end, we aim to minimize the following regret for the main agent:

$$\mathrm{Reg}^{\mathrm{MG}}(T) := \sum_{t=1}^{T} \left[V_1^* (x_1) - V_1^{\mu_t,\dagger} (x_1) \right],$$

where μ_t is the policy adopted by the max-player for episode t. Note that we can switch the roles of two players to learn a policy ν for the min-player.

For simplicity, we only consider the value-based approach with the function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subset (\mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R})$. Each $f \in \mathcal{F}$ induces a NE and a max-player's policy:

$$\mu_{h,f}(x) = \mathrm{argmax}_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x, \cdot, \cdot)\nu. \tag{4.3.2}$$

The induced value function for all (x, h) is then given by

$$V_{h,f}(x) = \max_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x, \cdot, \cdot)\nu. \tag{4.3.3}$$

Moreover, for a fixed max-player policy μ_f, we can approximate the value of the best response via a function g by

$$V_{h,g}^{\mu_f}(x) = \min_{\nu \in \Delta_{\mathcal{B}}} \mu_{h,f}(x)^\top g_h(x, \cdot, \cdot)\nu. \tag{4.3.4}$$

The correspondingly approximate best response is given by

$$\nu_{f,g,h}(x) = \mathrm{argmin}_{\nu \in \Delta_{\mathcal{B}}} \mu_{h,f}^\top g_h(x, \cdot, \cdot)\nu. \tag{4.3.5}$$

We define two different Bellman operators as in [70; 68; 69]:

$$\begin{aligned}(\mathcal{T}_h f)(x,a,b) &:= r_h(x,a,b) + \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)} V_{h,f+1}(x');\\ (\mathcal{T}_h^{\mu} f)(x,a,b) &:= r_h(x,a,b) + \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)} V^{\mu}_{h,f+1}(x').\end{aligned}$$

The corresponding Bellman residual are denoted as

$$\begin{aligned}\mathcal{E}_h(f;x,a,b) &= \mathcal{E}_h(f_h, f_{h+1};\zeta) = f_h(x,a,b) - (\mathcal{T}_h f)(x,a,b);\\ \mathcal{E}_h^{\mu}(f;x,a,b) &= \mathcal{E}_h^{\mu}(f_h, f_{h+1};\zeta) = f_h(x,a,b) - (\mathcal{T}_h^{\mu} f)(x,a,b),\end{aligned} \tag{4.3.6}$$

where we use the subscript and the trajectory ζ to denote the dependency on the state-action pairs for notation simplicity. We can decompose the regret into two parts:

$$\mathrm{Reg}^{\mathrm{MG}}(T) = \underbrace{\left(\sum_{t=1}^{T} V_1^{*}(x_1) - V_1^{\mu_t,\nu_t}(x_1)\right)}_{\text{main agent}} + \underbrace{\left(\sum_{t=1}^{T} V_1^{\mu_t,\nu_t}(x_1) - V_1^{\mu_t,\dagger}(x_1)\right)}_{\text{exploiter}}. \tag{4.3.7}$$

Similar to the MDP case, we first decompose each part of regret into the Bellman residuals, where the proofs are deferred to Section 5.1.

Lemma 4.3.1 (Value decomposition for the main agent)**.** Let $\mu = \mu_f$ (as in (4.3.2)) and ν be an arbitrary policy taken by the min-player. It holds that

$$V^{*}(x_1) - V_1^{\mu,\nu}(x_1) \le \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu}\mathcal{E}_h(f_h, f_{h+1};\zeta) + V^{*}(x_1) - V_{1,f}(x_1).$$

The innovative idea ([68; 69]) is that we may adopt a non-symmetric structure in the max-player and min-player so that the min-player approximates the best response for the max-player and serves as an exploiter to exploit its weakness.

Lemma 4.3.2 (Value decomposition for the exploiter)**.** Suppose that $\mu = \mu_f$ is taken by the max-player and g is chosen by the min-player. Let ν be taken according to (4.3.5). Then, it holds that

$$V_1^{\mu,\nu}(x_1) - V_1^{\mu,\dagger}(x_1) = -\sum_{h=1}^{H} \mathbb{E}_{\mu,\nu}\mathcal{E}_h^{\mu}(g_h, g_{h+1},\zeta) + V_{1,g}^{\mu}(x_1) - V_1^{\mu,\dagger}(x_1).$$

We can relate the prediction error to the training error via the following version of eluder coefficient.

Definition 4.3.3 (Eluder coefficient for two-player zero-sum MGs)**.** Given an $MG(H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$, a function class $\mathcal{F}$, a time horizon T, the eluder coefficient is the smallest $d(\epsilon)$ such that

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\left[\mathbb{E}_{\pi_t}\mathcal{E}_h^{\mu_{f^t}}\left(g^t; x^h, a^h, b^h\right)\right] \le \sqrt{d(\epsilon)\sum_{h=1}^{H}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\left[\mathbb{E}_{\pi_s}\mathcal{E}_h^{\mu_{f^t}}\left(g^t; x^h, a^h, b^h\right)\right]^2} + 2\min\{Hd, H^2T\} + \epsilon B^\dagger T,$$

for any sequence of $\{f^t, g^t \in \mathcal{F}\}_{t=1}^T$, where π_s is a policy pair $(\mu_{f_s}, \nu_{f_s,g_s})$ induced by (f_s, g_s) via (4.3.2) and (4.3.5).

Then, it suffices to apply the optimistic modification and loss estimator for the main agent and exploiter separately. We present the algorithm in Algorithm 2, where the loss estimators $L_h^{1:k-1}(\cdot)$ and $L_{h,\mu}^{1:k-1}(\cdot)$ can be constructed similarly to Lemma 3.1.2 and Lemma 3.1.4. Then, the analyses basically follow Chapter 3 by looking at the main agent and exploiter separately. We do not dive into details here to avoid repetition.

Nevertheless, most of the techniques and ideas are similar for the two-player zero-sum MGs and MDPs. It is interesting to see whether we can extend the GEC to capture a more general multi-agent formulation.

4.3.3 Limitations

Computational efficiency. In this paper, we mainly care about the statistical efficiency of the proposed algorithms, and the proposed algorithms are computationally inefficient in general. In contrast, [49] studies the general function approximation under the LSVI-based framework (Least Squares Value Iteration) with the eluder dimension, which is known to be computationally efficient as

Algorithm 2 MEX for Two-player Zero-sum MG

1: **Input:** Hypothesis space $\mathcal{H}$, $\eta > 0$, batch size $m > 0$.

2: **for** $k = 1, 2, \cdots, K := T/m$ **do**

3: For main agent, select f^k by solving

$$f^k = \operatorname*{argmax}_{f \in \mathcal{H}} \Big[V_{1,f}(x_1) - \eta \sum_{h=1}^{H} L_h^{1:k-1}(f) \Big]. \tag{4.3.8}$$

4: For exploiter, select g^k by solving

$$g^k = \operatorname*{argmax}_{g \in \mathcal{H}} \Big[- V_{1,g}^{\mu_{f^k}}(x_1) - \eta \sum_{h=1}^{H} L_{h,\mu_{f^k}}^{1:k-1}(g) \Big]. \tag{4.3.9}$$

5: Get the policies $(\mu_{f^k}, \nu_{f^k,g^k})$ according to (4.3.2) and (4.3.5);

6: For each $h \in [H]$, collect a batch of dataset $\{\zeta_{i,h}^k\}_{i=1}^m$ by following $(\mu_{f^k}, \nu_{f^k,g^k})$ m times.

7: **end for**

long as we can efficiently solve

$$\operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \Big(f(x_i, a_i) - y_i \Big)^2,$$

for any $\{x_i, a_i, y_i\}_{i=1}^n \in \mathcal{S}^n \times \mathcal{A}^n \times \mathbb{R}^n$ with respect to the function class $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to [0, H]\}$ for function approximation. However, the problems covered by the framework in [49] are rather limited and the regret bounds are also usually inferior.

Suboptimal statistical efficiency. The regret bounds presented in this paper are usually suboptimal in terms of the dependency on the episode length H compared to the minimax lower bound. For instance, combining Example 2.2.5, Lemma 3.1.4, and Theorem 3.2.1 [1], our framework suggests that we can achieve

[1]Linear MDP is Bellman complete because the Bellman update of any V is linear in the known feature according to Lemma 1.2.7. Meanwhile, we use the fact that the covering number of the hypothesis space if $\widetilde{\mathcal{O}}(d)$.

a high-probability regret bound of $\widetilde{\mathcal{O}}(H^2 d\sqrt{T})$. In comparison, [39; 40] give a regret bound of $\widetilde{\mathcal{O}}(H^{3/2} d\sqrt{T})$. These works utilize the information of variance to achieve the sharp horizon dependence and it would be interesting to see whether we can achieve similar results in the framework presented in this paper.

Chapter 5

Missing Proof

5.1 Proof of Regret Decomposition Lemmas

We prove the value decomposition lemmas in this section.

Proof of Lemma 2.1.1. The proof relies on the following lemma from [22].

Lemma 5.1.1 (Extended Value Difference Lemma)**.** Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be two arbitrary policies and let $\{\widehat{Q}_h\}_{h=1}^H$ be any given Q-functions. Then define $\widehat{V}_h(x) := \left\langle \widehat{Q}_h(x,\cdot), \pi_h(\cdot|x) \right\rangle$ for all $x \in \mathcal{S}$. Then, for all $x \in \mathcal{S}$,

$$\begin{aligned}\widehat{V}_1(x) - V_1^{\pi'}(x) = &\sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\left\langle \widehat{Q}_h\left(x_h,\cdot\right), \pi_h\left(\cdot \mid x_h\right) - \pi'_h\left(\cdot \mid x_h\right)\right\rangle \mid x_1 = x\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\widehat{Q}_h\left(x_h,a_h\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(x_h,a_h\right) \mid x_1 = x\right].\end{aligned}$$

Then, for each $t \in [T]$, we can decompose the immediate regret as follows:

$$V_1^*(x_1) - V_1^{\pi_{f^t}}(x_1) = \underbrace{V_1^*(x_1) - V_{1,f^t}(x_1)}_{\text{(i)}} + \underbrace{V_{1,f^t}(x_1) - V_1^{\pi_{f^t}}(x_1)}_{\text{(ii)}}. \tag{5.1.1}$$

In what follows, we omit the condition on the initial state x_1 for notation simplicity. We invoke Lemma 5.1.1 with $\pi = \pi_{f^t}$ and $\pi' = \pi^*$ to obtain that

$$\begin{aligned}\text{(i)} &= -\mathbb{E}_{\pi^*}\sum_{h=1}^{H}\mathcal{E}_h(f^t, x_h, a_h) + \sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\langle Q_{h,f^t}(x_h,\cdot), \pi_h^*(\cdot|x_h) - \pi_{h,f^t}(\cdot|x_h)\rangle\right] \\ &\le -\mathbb{E}_{\pi^*}\sum_{h=1}^{H}\mathcal{E}_h(f^t, x_h, a_h),\end{aligned} \tag{5.1.2}$$

where the inequality follows from π_{h,f^t} is greedy in terms of Q_{h,f^t}. We further invoke Lemma 5.1.1 with $\pi = \pi_{f^t}$ and $\pi' = \pi_{f^t}$ to obtain that

$$\text{(ii)} = \mathbb{E}_{\pi_{f^t}}\sum_{h=1}^{H}\mathcal{E}_h(f^t, x_h, a_h). \tag{5.1.3}$$

Plugging (5.1.2) and (5.1.3) into (5.1.1), we obtain the desired result. □

We now prove the decomposition lemma for the MGs.

Proof of Lemma 4.3.1. For a clean presentation, we use the notation $\mathbb{D}_\pi$ so that $[\mathbb{D}_\pi Q](x) := \mathbb{E}_{(a,b)\sim\pi(\cdot,\cdot|x)}Q(x,a,b)$, for any policy pair $\pi = (\mu,\nu)$ and action-value function Q. With these notations, we have

$$V_h^{\mu,\nu}(x) = [\mathbb{D}_{\mu_h\times\nu_h}Q_h^{\mu,\nu}](x).$$

Let $\mu = \mu_f$ and ν be an arbitrary policy taken by the min-player.

$$\begin{aligned}&V_1^*(x_1) - V_1^{\mu,\nu}(x_1) \\ &\quad= \sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}V_{f,h}(x_h) - r_h(x_h,a_h,b_h) - V_{h+1,f}(x_{h+1}) + V_1^*(x_1) - V_{1,f}(x_1) \\ &\quad= \sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\min_{\nu'}\mathbb{D}_{\mu,\nu'}f(x_h) - r_h(x_h,a_h,b_h) - V_{h+1,f}(x_{h+1}) + V_1^*(x_1) - V_{1,f}(x_1) \\ &\quad\le \sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\mathbb{D}_{\mu,\nu}f^h(x_h) - r_h(x_h,a_h,b_h) - V_{h+1,f}(x_{h+1}) + V_1^*(x_1) - V_{1,f}(x_1) \\ &\quad= \sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}f^h(x_h,a_h,b_h) - r_h(x_h,a_h,b_h) - V_{h+1,f}(x_{h+1}) + V_1^*(x_1) - V_{1,f}(x_1) \\ &\quad= \sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\mathcal{E}_h(f^h, f^{h+1}, \zeta) + V_1^*(x_1) - V_{1,f}(x_1),\end{aligned}$$

where the first equality comes from the value-decomposition Theorem [34] (can be verified easily by telescope sum and $V^{H+1} = 0$); the second equality is because of the definition of $\mu = \mu_{h,f}(x) = \operatorname{argmax}_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x, \cdot, \cdot)\nu$; the inequality comes from the fact that $\mu = \mu_f$ and ν may not be $\operatorname{argmin}_{\nu'} \mathbb{D}_{\mu,\nu'} f(x_h)$. □

Proof of Lemma 4.3.2. Suppose that $\mu = \mu_f$ is taken by the max-player and g is sampled from the posterior by the booster agent. Let ν be given by $\nu = \operatorname{argmin}_{\nu'} V_h^\mu(x)$ for all (x, h). Then, we have:

$$
\begin{aligned}
&V_1^{\mu,\dagger}(x_1) - V_1^{\mu,\nu}(x_1) \\
&\quad = V_{1,g}^{\mu}(x_1) - V_1^{\mu,\nu}(x_1) + V_1^{\mu_t,\dagger}(x_1) - V_{1,g}^{\mu}(x_1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} \mathbb{D}_{\mu,\nu} g(x_h) - r_h(x_h, a_h, b_h) - V_{h+1,g}^{\mu}(x_{h+1}) + V_1^{\mu,\dagger}(x_1) - V_{1,g}^{\mu}(x_1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} g^h(x_h, a_h, b_h) - r_h(x_h, a_h, b_h) - V_{h+1,g}^{\mu}(x_{h+1}) + V_1^{\mu,\dagger}(x_1) - V_{1,g}^{\mu}(x_1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} \mathcal{E}_h^{\mu}(g^h, g^{h+1}, \zeta) + V_1^{\mu,\dagger}(x_1) - V_{1,g}^{\mu}(x_1).
\end{aligned}
$$

□

5.2 Proof of Reductions

In this section, we prove the reduction of the problems for the linear mixture MDP (Example 2.3.3) and witness rank (Example 2.3.6).

5.2.1 Reduction of Linear Mixture MDP

We recall that

$$
\ell_h^s(f) := \mathbb{E}_{\pi_{f^s}} \left[\theta_{h,f}^\top \left[\psi(x_h, a_h) + \sum_{x' \in \mathcal{S}} \phi(x_h, a_h, x') V_{h+1,f^s}(x') \right] - r_h - V_{h+1,f^s}(x_{h+1}) \right].
$$

Proof. Similar to the linear MDP case, we first observe that

$$\begin{aligned}\mathbb{E}_{\pi_{f^t}}\mathcal{E}_h(f^t, x_h, a_h) &= \mathbb{E}_{\pi_{f^t}}Q_{h,f^t}(x_h, a_h) - \mathcal{T}_h V_{h+1,f^t}(x_h, a_h)\\ &= \mathbb{E}_{\pi_{f^t}}\left(\theta_{h,f^t}^\top - \theta_{h,f^*}^\top\right)\left(\psi(x_h, a_h) + \sum_{x'\in\mathcal{S}}\phi(x_h, a_h, x')V_{h+1,f^t}(x')\right)\\ &= \left\langle X_h(f^t), \theta_{h,f^t}^\top - \theta_{h,f^*}^\top\right\rangle,\end{aligned}$$

where $X_h(f^t) = \mathbb{E}_{\pi_{f^t}}\left(\psi(x_h, a_h) + \sum_{x'\in\mathcal{S}}\phi(x_h, a_h, x')V_{h+1,f^t}(x')\right)$. With the same proof in Example 2.2.5, we can show that

$$\begin{aligned}&\sum_{t=1}^T V_{1,f^t} - V_1^{\pi^t}\\ &\le H\cdot\sum_{t=1}^T\sum_{h=1}^H \min\left\{|\left\langle X_h(f^t), \frac{\theta_{h,f^t}-\theta_{h,f^*}}{H}\right\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}}\le 1\} + \min\{H\widetilde{d}, H^2T\},\end{aligned} \tag{5.2.1}$$

where $\widetilde{d} = \frac{3Hd}{\log 2}\log\left(1+\frac{1}{\lambda\log 2}\right)$. We now fix a (t,h) in the first summation and proceed as follows:

$$\begin{aligned}&\min\left\{|\left\langle X_h(f^t), \frac{\theta_{h,f^t}-\theta_{h,f^*}}{H}\right\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}}\le 1\}\\ &\le ||\frac{\theta_{h,f^t}-\theta_{h,f^*}}{H}||_{\Sigma_{t;h}}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\}\\ &= \frac{1}{H}\left[\lambda||\theta_{h,f^t}-\theta_{h,f^*}||^2 + \sum_{s=1}^{t-1}|\left\langle X_h(f^s), \theta_{h,f^t}-\theta_{h,f^*}\right\rangle|^2\right]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\}\\ &\le \sqrt{\frac{\lambda}{H^2}||\theta_{h,f^t}-\theta_{h,f^*}||^2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\} + \frac{1}{H}\left[\sum_{s=1}^{t-1}\left(\ell_h^s(f^t)\right)^2\right]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\},\end{aligned} \tag{5.2.2}$$

where the equality uses $\Sigma_{t;h} = \lambda\mathbf{I} + \sum_{s=1}^{t-1}X_h(f^s)X_h(f^s)^\top$, and the last inequality is because

$$\begin{aligned}&\langle X_h(f^s), \theta_{h,f^t}-\theta_{h,f^*}\rangle\\ &= \mathbb{E}_{\pi_{f^s}}\left(\psi(x_h, a_h) + \sum_{x'\in\mathcal{S}}\phi(x_h, a_h, x')V_{h+1,f^s}(x')\right)^\top(\theta_{h,f^t}-\theta_{h,f^*})\\ &= \mathbb{E}_{\pi_{f^s}}\left[\theta_{h,f^t}^\top\left(\psi(x_h, a_h) + \sum_{x'\in\mathcal{S}}\phi(x_h, a_h, x')V_{h+1,f^s}(x')\right) - r_h(x_h, a_h) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}V_{h+1,f^s}(x_{h+1})\right]\\ &= \mathbb{E}_{\pi_{f^s}}\left[\theta_{h,f^t}^\top\left(\psi(x_h, a_h) + \sum_{x'\in\mathcal{S}}\phi(x_h, a_h, x')V_{h+1,f^s}(x')\right) - r_h(x_h, a_h) - V_{h+1,f^s}(x_{h+1})\right]\\ &= \ell_h^s(f^t).\end{aligned}$$

We also note that by regularization condition that $\sup_{f\in\mathcal{H},h\in[H]}||\theta_{h,f}|| \le B$, we

have

$$
\begin{aligned}
&H \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \sqrt{\frac{\lambda}{H^2}||\theta_{h,f^t} - \theta_{h,f^*}||^2} \cdot \min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\} \\
&\lesssim \sqrt{\lambda B^2} \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}}, 1\} \\
&\leq \sqrt{T\lambda B^2}\sqrt{H\sum_{t=1}^{T}\sum_{h=1}^{H} \min\{||X_h(f^t)||^2_{\Sigma_{t;h}^{-1}}, 1\}} \\
&\leq \sqrt{T\lambda B^2}\sqrt{\min\{H\widetilde{d}, H^2T\}}.
\end{aligned} \tag{5.2.3}
$$

where we denote $\widetilde{d} = \frac{3Hd}{\log 2}\log\big(1+\frac{1}{\lambda \log 2}\big)$. Plugging (5.2.2) and (5.2.3) into (5.2.1), we obtain that

$$
\begin{aligned}
&\sum_{t=1}^{T} V_{f^t} - V^{\pi^t} \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\sum_{s=1}^{t-1}\big(\ell_h^s(f^t)\big)^2\Big]^{1/2} \cdot \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \leq 1\} \\
&\quad + \sqrt{T\lambda B^2}\sqrt{\min\{H\widetilde{d}, H^2T\}} + \min\{H\widetilde{d}, H^2T\} \\
&\leq \Big[\widetilde{d}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s=1}^{t-1}\big(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t, x_h, a_h)\big)^2\Big]^{1/2} + 2\min\{H\widetilde{d}, H^2T\} + B^2T\lambda.
\end{aligned}
$$

Therefore, we conclude that linear MDP has a low eluder coefficient of $O\big(Hd\log\big(1+\frac{1}{\lambda}\big)\big)$. □

5.2.2 Reduction of Witness Rank

We recall the (2.3.2) and write it here for the reader's convenience.

$$
\max_{v\in\mathcal{V}_h} \mathbb{E}_{\pi_f}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)}v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}v(x_h,a_h,x')] \geq \langle W_h(g), X_h(f)\rangle
$$

$$
\kappa_{\mathrm{wit}} \cdot \mathbb{E}_{\pi_f}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)}V_{h+1,g}(x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}V_{h+1,g}(x')] \leq \langle W_h(g), X_h(f)\rangle\,.
$$

Proof of Example 2.3.6. To begin with, we note that for any $(h,x,a,g) \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{H}$, we have

$$
\begin{aligned}
&\max_{v\in\mathcal{V}}[(\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)}v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}v(x_h,a_h,x'))^2] \\
&\quad \leq \mathrm{TV}\big(\mathbb{P}_{h,g}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)^2 \leq 2D_{\mathrm{H}}^2\big(\mathbb{P}_{h,g}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big),
\end{aligned} \tag{5.2.4}
$$

where $\mathrm{TV}(\cdot,\cdot)$ is the total variation distance and $D^2_{\mathrm{H}}(\cdot,\cdot)$ is the Hellinger divergence. By the Bellman equation under the hypothesis model f, we have $Q_{h,f}(x_h,a_h) = r_h(x_h,a_h) + \mathbb{E}_{x'\sim\mathbb{P}_{h,f}(\cdot|x_h,a_h)}V_{h+1,f}(x')$ for any $f\in\mathcal{H}$. Combining this with the second condition of witness rank, we have

$$|\mathbb{E}_{\pi_f}[Q_{h,f}(x_h,a_h) - r_h(x_h,a_h) - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}V_{h+1,f}(x')]| \le \frac{1}{\kappa_{\mathrm{wit}}}\langle W_h(f), X_h(f)\rangle\,. \tag{5.2.5}$$

Let $\Sigma_{t;h} = \lambda\mathbf{I} + \sum_{s=1}^{t-1} X_h(f^s)X_h(f^s)^\top$. Then, by the value decomposition lemma (Lemma A.1.1), we have

$$\begin{aligned}
&\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t,x_h,a_h\right)\right]\\
&\le H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\min\{\frac{1}{H\kappa_{\mathrm{wit}}}|\left\langle W_h(f^t),X_h(f^t)\right\rangle|,1\}\\
&= H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\min\{\frac{1}{H\kappa_{\mathrm{wit}}}|\left\langle W_h(f^t),X_h(f^t)\right\rangle|,1\}\cdot\Big(\mathbf{1}\{\left\|X_h\left(f^t\right)\right\|_{\Sigma_{t;h}^{-1}}\le 1\} + \mathbf{1}\{\left\|X_h\left(f^t\right)\right\|_{\Sigma_{t;h}^{-1}} > 1\}\Big),\\
&\le H\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{H\kappa_{\mathrm{wit}}}||W_h(f^t)||_{\Sigma_{t;h}}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\} + \min\{\widetilde{d}H, H^2T\},
\end{aligned} \tag{5.2.6}$$

where $\widetilde{d} = \frac{3Hd}{\log 2}\log\left(1+\frac{1}{\lambda\log 2}\right)$. Here the last inequality is because of the Lemma A.1.3. We now fix a (t,h) in the first summation and proceed as follows:

$$\begin{aligned}
&\frac{1}{H\kappa_{\mathrm{wit}}}||W_h(f^t)||_{\Sigma_{t;h}}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}\\
&= \frac{1}{H\kappa_{\mathrm{wit}}}\Big[\lambda\cdot||W_h(f^t)||_2^2 + \sum_{s=1}^{t-1}\left|\left\langle W_h(f^t),X_h(f^s)\right\rangle\right|^2\Big]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}\\
&\le \frac{1}{H\kappa_{\mathrm{wit}}}\Big(\sqrt{\lambda B^2} + \Big[\sum_{s=1}^{t-1}\left|\left\langle W_h(f^t),X_h(f^s)\right\rangle\right|^2\Big]^{1/2}\Big)\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}\\
&\le \frac{1}{H\kappa_{\mathrm{wit}}}\Big(\sqrt{\lambda B^2} + \Big[\sum_{s=1}^{t-1}2\mathbb{E}_{x_h\sim\pi_f}D^2_{\mathrm{H}}\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h),\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)\Big]^{1/2}\Big)\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\},
\end{aligned} \tag{5.2.7}$$

where we use $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ in the second-last inequality and use $||W_h(f^t)||\le B$; the last inequality is because the first condition of witness rank and (5.2.4). We also note that

$$H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{H\kappa_{\mathrm{wit}}}\sqrt{\lambda B^2}\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\} \le \sqrt{T\lambda B^2/\kappa^2_{\mathrm{wit}}}\sqrt{\min\{H\widetilde{d},H^2T\}}. \tag{5.2.8}$$

Plugging (5.2.7) and (5.2.8) into (5.2.5), we obtain that

$$
\begin{aligned}
&\sum_{t=1}^{T} V_{f^t} - V^{\pi^t} \\
&\le \frac{1}{\kappa_{\mathrm{wit}}} \sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\sum_{s=1}^{t-1} 2\mathbb{E}_{\pi_f} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)\Big]^{1/2} \cdot \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} \\
&\quad + \sqrt{T\lambda B^2/\kappa_{\mathrm{wit}}^2}\sqrt{\min\{H\widetilde{d}, H^2T\}} + \min\{H\widetilde{d}, H^2T\} \\
&\le \Big[\frac{2\widetilde{d}}{\kappa_{\mathrm{wit}}^2}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{s=1}^{t-1}\mathbb{E}_{\pi_f} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)\Big]^{1/2} + 2\min\{H\widetilde{d}, H^2T\} + B^2T\lambda/\kappa_{\mathrm{wit}}^2.
\end{aligned}
$$

By rescaling $\lambda = \lambda'\kappa_{\mathrm{wit}}^2$, we conclude that problems with a witness rank of d have a low eluder coefficient of $O\big(Hd\log\big(1+\frac{1}{\lambda'\kappa_{\mathrm{wit}}^2}\big)/\kappa_{\mathrm{wit}}^2\big)$. □

5.2.3 Reduction of Factored MDP

Proof of Example 2.3.8. To begin with, we introduce the following discriminator class

$$\mathcal{V}_h = \{w_1 + w_2 + \cdots w_d : w_i \in \mathcal{W}_i\},$$

where $\mathcal{W}_i = \{\mathcal{O}^{|\mathrm{pa}_i|\times\mathcal{A}\times\mathcal{O}} \to \{-1,+1\}\}$. Note that w_i indeed takes x, a, x' as input, but it only looks at $(x[\mathrm{pa}_i], a, x'[i])$ to determine whether it will output 1 or -1. According to Lemma 23 of [27], we know that

$$
\begin{aligned}
&\max_{v\in\mathcal{V}_h} \mathbb{E}_{x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} v(x_h,a_h,x')] \\
&= \mathbb{E}\Big[\sum_{i=1}^{d} ||\mathbb{P}_{h,g}^{(i)}(\cdot|x_h[\mathrm{pa}_i], a_h) - \mathbb{P}_{h,f^*}^{(i)}(\cdot|x_h[\mathrm{pa}_i], a_h)||_{TV} | x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})\Big].
\end{aligned}
$$

We denote $\mathbb{P}_h^{\pi_f}(x_h)$ as the probability of x_h by following π_f. Then, we have

$$
\begin{aligned}
&\max_{v\in\mathcal{V}_h} \mathbb{E}_{x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} v(x_h,a_h,x')] \\
&= \frac{1}{A}\sum_{i=1}^{d}\sum_{x_h\in\mathcal{S}}\sum_{a_h\in\mathcal{A}} \mathbb{P}_h^{\pi_f}(x_h)||\mathbb{P}_{h,g}^{(i)}(\cdot|x_h[\mathrm{pa}_i], a_h) - \mathbb{P}_{h,f^*}^{(i)}(\cdot|x_h[\mathrm{pa}_i], a_h)||_{TV} \\
&= \frac{1}{A}\sum_{i=1}^{d}\sum_{z\in\mathcal{O}^{|\mathrm{pa}_i|}}\sum_{a_h\in\mathcal{A}} \mathbb{P}_h^{\pi_f}(x_h[\mathrm{pa}_i]=z)||\mathbb{P}_{h,g}^{(i)}(\cdot|z, a_h) - \mathbb{P}_{h,f^*}^{(i)}(\cdot|z, a_h)||_{TV} \\
&= \langle X_h(f), W_h(g)\rangle,
\end{aligned}
$$

where $X_h(f)[i,a,z] = \frac{1}{A}\mathbb{P}_h^{\pi_f}(x_h[\mathrm{pa}_i] = z)$ and $W_h(g)[i,a,z] = ||\mathbb{P}_{h,g}^{(i)}(\cdot|z,a_h) - \mathbb{P}_{h,f^*}^{(i)}(\cdot|z,a_h)||_{TV}$. Therefore, we know that there exists $X_h : \mathcal{H} \to \mathbb{R}^L$ and $W_h : \mathcal{H} \to \mathbb{R}^L$ with $L = A\sum_{i=1}^d |\mathcal{O}|^{|\mathrm{pa}_i|}$.

On the other hand, we have

$$
\begin{aligned}
&\mathbb{E}_{x_h\sim\pi_f, a_h\sim\pi_g}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)}V_{h+1,g}(x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)}V_{h+1,g}(x')] \\
&\le H\mathbb{E}\Big[\sum_{a\in\mathcal{A}}\pi_g(a|x_h)||\mathbb{P}_{h,g}(\cdot|x_h,a) - \mathbb{P}_{h,f^*}(\cdot|x_h,a)||_{TV}|x_h\sim\pi_f\Big] \\
&\le AH\cdot\mathbb{E}\Big[\sum_{a\in\mathcal{A}}\pi_g(a|x_h)||\mathbb{P}_{h,g}(\cdot|x_h,a) - \mathbb{P}_{h,f^*}(\cdot|x_h,a)||_{TV}|x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})\Big] \\
&\le AH\cdot\mathbb{E}\Big[\sum_{a\in\mathcal{A}}\pi_g(a|x_h)\sum_{i=1}^d||\mathbb{P}_{h,g}^{(i)}(\cdot|x_h[\mathrm{pa}_i],a_h) - \mathbb{P}_{h,f^*}^{(i)}(\cdot|x_h[\mathrm{pa}_i],a_h)||_{TV}|x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})\Big] \\
&= AH\cdot\langle X_h(f), W_h(g)\rangle\,.
\end{aligned}
$$

With the same proof of Example 2.3.6, we can show that

$$
\begin{aligned}
&\sum_{t=1}^T V_{f^t} - V^{\pi^t} \\
&\le \Big[\frac{2\widetilde{d}}{\kappa_{\mathrm{wit}}^2}\sum_{t=1}^T\sum_{h=1}^H\sum_{s=1}^{t-1}\mathbb{E}_{x_h\sim\pi_f, a_h\sim\mathrm{Unif}(\mathcal{A})}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)\Big]^{1/2} \\
&\qquad + 2\min\{H\widetilde{d}, H^2T\} + B^2T\lambda/\kappa_{\mathrm{wit}}^2,
\end{aligned}
$$

where $\kappa_{\mathrm{wit}} = \frac{1}{AH}$, $\widetilde{d} = \widetilde{O}(LH)$, and $B = \sup_f ||W_h(f)||_2 = O(\sqrt{L})$. By rescaling $\lambda = \lambda'\kappa_{\mathrm{wit}}^2$, we conclude that the factored MDP has a low eluder coefficient of $\widetilde{O}\big(H^3A^2L\big)$ with $\ell_h^s(f) = \mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\mathrm{Unif}(\mathcal{A})}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)$. □

5.2.4 Reduction of Q^*-state abstration model

Proof of Example 2.3.4. It suffices to prove that with the feature maps and also the hypothesis class specified in Example 2.3.4, the space of Bellman residual is a linear one with dimension $d = (|\mathcal{K}|+1)|\mathcal{A}|$. Then, the remaining proof is the same as that of linear MDP (Example 2.2.5).

To start with, we note that with the definitions of the feature maps $\phi(\cdot,\cdot)$ and

$\psi(\cdot)$, we have

$$Q_h^*(x,a) = \phi(x,a)^\top \theta_{h,f^*}, \qquad \theta_{h,f^*}[z,a] = Q_h^*(x,a) \text{ with } \xi(x) = z,$$
$$V_{h+1}^*(x) = \psi(x)^\top w_{h+1,f^*}(x), \qquad w_{h+1,f^*}[z] = V_{h+1}^*(x) \text{ with } \xi(x) = z.$$

Therefore, we have $f^* \in \mathcal{H}$. Moreover, for any $f, g \in \mathcal{H}$, we can obtain that

$$\begin{aligned}
&\mathbb{E}_{\pi_g} \mathcal{E}_h(f, x_h, a_h) \\
&= \mathbb{E}_{\pi_g}\left[\phi(x_h,a_h)^\top \theta_{h,f} - r_h(x_h,a_h) - \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x_h,a_h)} \psi(x')^\top w_{h+1,f}\right] \\
&= \mathbb{E}_{\pi_g}\left[\phi(x_h,a_h)^\top \theta_{h,f} - Q_h^*(x_h,a_h) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x_h,a_h)} V_{h+1}^*(x') - \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x_h,a_h)} \psi(x')^\top w_{h+1,f}\right] \\
&= \mathbb{E}_{\pi_g}\left[\phi(x_h,a_h)^\top (\theta_{h,f} - \theta_{h,f^*}) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x_h,a_h)} \psi(x')^\top (w_{h+1,f^*} - w_{h+1,f})\right] \\
&= \langle X_h(g), W_h(f) \rangle,
\end{aligned}$$

where the second step uses (1.2.1). Here $X_h, W_h : \mathcal{H} \to \mathbb{R}^d$ are defined as follows:

$$X_h(g) = \begin{bmatrix} \mathbb{E}_{\pi_g} \phi(x_h, a_h) \\ \mathbb{E}_{\pi_g} \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x_h,a_h)} \psi(x') \end{bmatrix}, \qquad W_h(f) = \begin{bmatrix} \theta_{h,f} - \theta_{h,f^*} \\ w_{h+1,f^*} - w_{h+1,f} \end{bmatrix}.$$

□

5.3 Proof of In-sample Error Estimation

5.3.1 Proof for Model-based Approach

In this section, we provide the proof of in-sample error estimation for the model-based approach (Lemma 3.1.6) and also its V-type variant discussed in Chapter 4.

Proof of Lemma 3.1.6. We recall that we will consider the following discrepancy function:

$$\widetilde{L}_h^s(f) = -\frac{1}{2} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s) + \frac{1}{2} \log \mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s),$$

which is equivalent to use $-\frac{1}{2} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)$ in the executed algorithm. The proof itself is a standard application of Cramer Chernoff's method. To

this end, we first estimate the moment-generating function. We denote $\mathbb{E}_t := \mathbb{E}[\cdot|(f^1, \zeta^1, \cdots, f^{t-1}, \zeta^{t-1}, f^t)]$. In what follows, $\widetilde{\pi}_t$ is either π_{f^t} (Q-type) or $\mathrm{Unif}(\mathcal{A})$ (V-type). It follows that

$$
\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{s=1}^{t}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)}\Big)\Big] \\
&= \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{s=1}^{t-1}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)}\Big)\Big]\mathbb{E}_t\sqrt{\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^t, a_h^t)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^t, a_h^t)}} \\
&= \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{s=1}^{t-1}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)}\Big)\Big]\mathbb{E}_{x_h\sim\pi_{f^t}, a_h\sim\widetilde{\pi}_t}\int_{x\in\mathcal{S}}\sqrt{\mathbb{P}_{h,f}(x \mid x_h, a_h)\cdot\mathbb{P}_{h,f^*}(x \mid x_h, a_h)} \\
&= \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{s=1}^{t-1}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)}\Big)\Big]\Big(1-\mathbb{E}_{x_h\sim\pi_{f^t}, a_h\sim\widetilde{\pi}_t}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)\Big) \\
&= \cdots \\
&= \prod_{s=1}^{t}\Big(1-\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\widetilde{\pi}_s}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)\Big)
\end{aligned}
$$

where we use the equivalent definitions of Hellinger distance (1.2.7) in the second last equality. We now invoke Lemma A.1.4 to obtain that for any fixed (h, f), we have

$$
\begin{aligned}
1-\frac{\delta}{H|\mathcal{H}_h|} \le \mathbb{P}\Big[\forall t>0 : \frac{1}{2}\sum_{s=1}^{t}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)} \le \log(H|\mathcal{H}_h|/\delta) \\
+\sum_{s=1}^{t}\log\Big(1-\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\widetilde{\pi}_s}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)\Big)\Big] \\
\le \mathbb{P}\Big[\forall t>0 : \frac{1}{2}\sum_{s=1}^{t}\log\frac{\mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)} \le \log(H|\mathcal{H}_h|/\delta) \\
-\sum_{s=1}^{t}\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\widetilde{\pi}_s}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)\Big]
\end{aligned}
$$

where we use $\log(1-x) \le -x$ for $x \le 1$. With a union bound over cH_h and then $[H]$, we conclude that with probability at least $1-\delta$, we have for all $t \in [T]$ and $h \in [H]$,

$$
\sum_{s=1}^{t}\mathbb{E}_{x_h\sim\pi_{f^s}, a_h\sim\widetilde{\pi}_s}D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)\Big] \le \sum_{s=1}^{t}\widetilde{L}_h^s(f)+\log(H|\mathcal{H}_h|/\delta).
$$

Therefore, for the Q-type problem, (3.1.1) is satisfied with $\Delta_h^t = \log(H|\mathcal{H}_h|/\delta)$ and for the V-type problem, (3.1.1) is satisfied with $\Delta_h^t = A\cdot\log(H|\mathcal{H}_h|/\delta)$. (3.1.2) naturally holds because $\widetilde{L}_h^s(f^*) = 0$. □

5.3.2 Proof for Bellman-complete Case

Inspired by [18; 28], we use the discrepancy function

$$L_h^s(f) = \big(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 - \inf_{f_h' \in \mathcal{H}_h} \big(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2.$$

Proof. We define the auxiliary discrepancy loss function

$$\widetilde{L}_h^s(f) = \big(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 - \big(\mathcal{T}_h V_{h+1,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2.$$

By completeness, we know that $\mathcal{T}_h V_{h+1,f} \in \mathcal{H}_h$, which implies that

$$\widetilde{L}_h^s(f) \le L_h^s(f). \tag{5.3.1}$$

Therefore, we can prove (3.1.1) for $\widetilde{L}_h^s(f)$ and it works for $L_h^s(f)$. The proof is a standard application of concentration inequality, where we choose the Freedman's inequality as in [18] (Lemma A.1.5).

The first key observation here is that

$$\mathbb{E}_{x_{h+1}^s \sim \mathbb{P}_h(\cdot|x_h^s, a_h^s)} \widetilde{L}_h^s(f) = \big(\mathcal{E}_h(f, x_h^s, a_h^s)\big)^2,$$

because the expectation of the second term is exactly the conditional variance. Therefore, we know that

$$\mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \widetilde{L}_h^s(f) = \mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \big(\mathcal{E}_h(f, x_h^s, a_h^s)\big)^2. \tag{5.3.2}$$

We proceed to control the second moment of $\widetilde{L}_h^s(f)$:

$$\begin{aligned}
&\mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} [\widetilde{L}_h^s(f)^2] \\
=&\mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \Big[\mathcal{E}_h(f, x_h^s, a_h^s)^2 \Big(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) + \mathcal{T}_h V_{h+1,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\Big)^2\Big] \\
\le& 4H^2 \mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \mathcal{E}_h(f, x_h^s, a_h^s)^2.
\end{aligned}$$

By Lemma A.1.5, for any fixed $(t, h, f) \in [T] \times [H] \times \mathcal{H}$, we know that with probability at least $1 - \frac{\delta}{H|\mathcal{H}_h|T}$, we have

$$\begin{aligned}
\Big|\sum_{s=1}^t \mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \big(\mathcal{E}_h(f, x_h, a_h)\big)^2 - \sum_{s=1}^t \widetilde{L}_h^s(f)\Big| \le \mathcal{O}\Big(&\sqrt{H^2 \cdot \log(H|\mathcal{H}_h|T/\delta) \sum_{s=1}^t \mathbb{E}_{x_h^s, a_h^s \sim \pi_{f^s}} \big(\mathcal{E}_h(f, x_h, a_h)\big)^2} \\
&+ H^2 \cdot \log(H|\mathcal{H}_h|T/\delta)\Big).
\end{aligned} \tag{5.3.3}$$

With a union bound over $[T] \times \mathcal{H}_h$ and then $[H]$, we know that with probability at least $1-\delta$, for any $(t, h, f) \in [T] \times [H] \times \mathcal{H}$, we have

$$-\sum_{s=1}^{t} \widetilde{L}_h^s(f) \leq -0.5 \sum_{s=1}^{t} \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(\mathcal{E}_h(f, x_h, a_h)\right)^2 + c \cdot H^2 \log(H|\mathcal{H}_h|T/\delta),$$

where the inequality is due to Cauchy-Schwarz inequality and a careful adjustment of the constant. By (5.3.1), we know that

$$\sum_{s=1}^{t} \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(\mathcal{E}_h(f, x_h, a_h)\right)^2 \lesssim \sum_{s=1}^{t} L_h^s(f) + H^2 \log(H|\mathcal{H}_h|T/\delta).$$

This concludes the proof of the first condition. We proceed to prove the second condition. We define the following auxiliary random variable:

$$W_h^s(f) = \left(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2 - \left(Q_h^*(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2.$$

Similarly, we first note that

$$\begin{aligned} \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} W_h^s(f) &= \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(Q_{h,f}(x_h^s, a_h^s) - \mathcal{T}_h V_{h+1}^*(x_h^s, a_h^s)\right)^2 \\ &= \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(Q_{h,f}(x_h^s, a_h^s) - Q_h^*(x_h^s, a_h^s)\right)^2, \end{aligned} \tag{5.3.4}$$

where the second equality follows from Bellman optimality equation. We can also obtain the following bound for the second moment:

$$\mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} [W_h^s(f)^2] \leq 4H^2 \mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(Q_{h,f}(x_h^s, a_h^s) - Q_h^*(x_h^s, a_h^s)\right)^2.$$

By Freedman's inequality and a union bound, with probability at least $1-\delta$, for any $(t, h, f) \in [T] \times [H] \times \mathcal{H}$, we have

$$-\sum_{s=1}^{t} W_h^s(f) \lesssim H^2 \log(H|\mathcal{H}_h|T/\delta),$$

where we use the fact that $\mathbb{E}_{x_h^s, a_h^s \sim \pi_{fs}} \left(Q_{h,f}(x_h^s, a_h^s) - Q_h^*(x_h^s, a_h^s)\right)^2 \geq 0$. This implies that for any $(t, h, f) \in [T] \times [H] \times \mathcal{H}$, we have

$$\left(Q_h^*(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2 - \left(Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2 \lesssim H^2 \log(H|\mathcal{H}_h|T/\delta).$$

Since f is arbitrary, we conclude that

$$\left(Q_h^*(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2 - \inf_{f_h' \in \mathcal{H}_h} \left(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1}^*(x_{h+1}^s)\right)^2 \lesssim H^2 \log(H|\mathcal{H}_h|T/\delta).$$

This proves (3.1.2) for $L_h^s(\cdot)$.

□

Chapter 6

Conclusion

In this paper, we study the reinforcement learning with general function approximation. We identify a structural measure, generalized eluder coefficient (GEC), which serves to reduce the prediction error to the historical in-sample training error in an online manner. We show that GEC captures a rich class of known solvable problems. In terms of the algorithmic design, we propose an optimization-based framework, whose target is modified with the "feel-good" term [32]. The proposed algorithm is neat and unified that can solve problems with a low GEC. Finally, we discuss the extensions of GEC to multi-agent scenarios.

Appendix A

Appendix chapter

A.1 Technical Lemmas

Lemma A.1.1 (Value Decomposition Lemma [34; 27])**.** For any hypothesis f and the induced greedy policy π_f, it holds that

$$V_{1,f}(x_1) - V_1^{\pi_f}(x_1) = \sum_{h=1}^{H} \mathbb{E}_{\pi_f}\left[\mathcal{E}_h(f, x_h, a_h)\right]. \tag{A.1.1}$$

Lemma A.1.2 (Elliptical Potential Lemma [45; 71; 47])**.** Let $\{x_i\}_{i\in[T]}$ be a sequence of vectors in $\mathbb{R}^d$ with $||x_i||_2 \leq L < \infty$ for all $t \in [T]$. Let Λ_0 be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1} x_i x_i^\top$. It holds that

$$\sum_{i=1}^{T} \min\{1, \|x_i\|^2_{\Lambda_i^{-1}}\} \leq 2\log\left(\frac{\det(\Lambda_{T+1})}{\det(\Lambda_1)}\right) \leq 2d\log\left(\frac{\mathrm{trace}(\Lambda_0) + TL^2}{d\det(\Lambda_0)^{1/d}}\right)$$

Lemma A.1.3 (Elliptical Potential is small for most of the time [72])**.** Given $\lambda > 0$ and $\{X_t\}_{t=1}^T \subset \mathbb{R}^d$ with $||X_t|| \leq L$ for all $t \in [T]$, if we denote $\Sigma_t = \lambda\mathbf{I} + \sum_{s=1}^{t} X_s X_s^\top$, then $||X_t||_{\Sigma_{t-1}^{-1}} \geq 1$ happens for at most

$$\frac{3d}{\log 2}\log\left(1 + \frac{L^2}{\lambda\log 2}\right).$$

Lemma A.1.4 (Martingale Exponential Inequalities)**.** Consider a sequence of random functions $\xi_1(\mathcal{Z}_1), \cdots, \xi_t(\mathcal{Z}_t), \cdots$ with respect to filtration $\{\mathcal{F}_t\}$. We have

for any $\delta \in (0,1)$ and $\lambda > 0$:

$$\mathbb{P}\Big[\exists n > 0 : -\sum_{i=1}^{n} \xi_i \geq \frac{\log(1/\delta)}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{n} \log \mathbb{E}_{Z_i^{(y)}} \exp(-\lambda \xi_i)\Big] \leq \delta,$$

where $Z_t = (Z_t^{(x)}, Z_t^{(y)})$ and $\mathcal{Z}_t = (Z_1, \cdots, Z_t)$.

Proof. See e.g., Theorem 13.2 of [73] for a detailed proof. □

Lemma A.1.5 (Freedman's inequality)**.** Let $\{X_t\}_{t \leq T}$ be a real-valued martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t \leq T}$. If $|X_t| \leq R$ almost surely, then for any $\eta \in (0, 1/R)$ it holds that with probability at least $1 - \delta$,

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

Proof. See [74] for detailed proof. □

Bibliography

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. 2018.

[2] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," *Advances in neural information processing systems*, vol. 21, 2008.

[3] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *International Conference on Machine Learning*, pp. 263–272, PMLR, 2017.

[4] C. Dann, T. Lattimore, and E. Brunskill, "Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is q-learning provably efficient?," *Advances in neural information processing systems*, vol. 31, 2018.

[6] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] A. Zanette and E. Brunskill, "Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function

bounds," in *International Conference on Machine Learning*, pp. 7304–7312, PMLR, 2019.

[8] Z. Zhang, X. Ji, and S. Du, "Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon," in *Conference on Learning Theory*, pp. 4528–4531, PMLR, 2021.

[9] P. Ménard, O. D. Domingues, X. Shang, and M. Valko, "Ucb momentum q-learning: Correcting the bias without forgetting," in *International Conference on Machine Learning*, pp. 7609–7618, PMLR, 2021.

[10] G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi, "Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17762–17776, 2021.

[11] T. Wu, Y. Yang, H. Zhong, L. Wang, S. Du, and J. Jiao, "Nearly optimal policy optimization with stable at any time guarantee," in *International Conference on Machine Learning*, pp. 24243–24265, PMLR, 2022.

[12] Z. Zhang, X. Ji, and S. Du, "Horizon-free reinforcement learning in polynomial time: the power of stationary policies," in *Conference on Learning Theory*, pp. 3858–3904, PMLR, 2022.

[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

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

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

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

[18] C. Jin, Q. Liu, and S. Miryoosefi, "Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang, "Bilinear classes: A structural framework for provable generalization in rl," in *International Conference on Machine Learning*, pp. 2826–2836, PMLR, 2021.

[20] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy, "Optimism in reinforcement learning with generalized linear function approximation," *arXiv preprint arXiv:1912.04136*, 2019.

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

[22] Q. Cai, Z. Yang, C. Jin, and Z. Wang, "Provably efficient exploration in policy optimization," in *International Conference on Machine Learning*, pp. 1283–1294, PMLR, 2020.

[23] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric, "Frequentist regret bounds for randomized least-squares value iteration," in

International Conference on Artificial Intelligence and Statistics, pp. 1954–1964, PMLR, 2020.

[24] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang, "Model-based reinforcement learning with value-targeted regression," in *International Conference on Machine Learning*, pp. 463–474, PMLR, 2020.

[25] A. Modi, N. Jiang, A. Tewari, and S. Singh, "Sample complexity of reinforcement learning using linearly combined model ensembles," in *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020, PMLR, 2020.

[26] D. Zhou, Q. Gu, and C. Szepesvari, "Nearly minimax optimal reinforcement learning for linear mixture markov decision processes," in *Conference on Learning Theory*, pp. 4532–4576, PMLR, 2021.

[27] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford, "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches," in *Conference on learning theory*, pp. 2898–2933, PMLR, 2019.

[28] C. Dann, M. Mohri, T. Zhang, and J. Zimmert, "A provably efficient model-free posterior sampling method for episodic reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12040–12051, 2021.

[29] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[30] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[31] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.

[32] T. Zhang, "Feel-good thompson sampling for contextual bandits and reinforcement learning," *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 2, pp. 834–857, 2022.

[33] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, "The statistical complexity of interactive decision making," *arXiv preprint arXiv:2112.13487*, 2021.

[34] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, "Contextual decision processes with low Bellman rank are PAC-learnable," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713, PMLR, 06–11 Aug 2017.

[35] D. J. Foster and A. Rakhlin, "Statistical reinforcement learning and decision making: Course notes," *Course Note*.

[36] Z. Zhang, Y. Zhou, and X. Ji, "Almost optimal model-free reinforcement learningvia reference-advantage decomposition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15198–15207, 2020.

[37] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.

[38] H. Zhong and T. Zhang, "A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes," *arXiv preprint arXiv:2305.08841*, 2023.

[39] A. Agarwal, Y. Jin, and T. Zhang, "Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation," *arXiv preprint arXiv:2212.06069*, 2022.

[40] J. He, H. Zhao, D. Zhou, and Q. Gu, "Nearly minimax optimal reinforcement learning for linear markov decision processes," *arXiv preprint arXiv:2212.06132*, 2022.

[41] Y. Wang, R. Wang, and S. Kakade, "An exponential lower bound for linearly realizable mdp with constant suboptimality gap," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9521–9533, 2021.

[42] S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang, "Is a good representation sufficient for sample efficient reinforcement learning?," *arXiv preprint arXiv:1910.03016*, 2019.

[43] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[44] M. Kearns and D. Koller, "Efficient reinforcement learning in factored mdps," in *IJCAI*, vol. 16, pp. 740–747, 1999.

[45] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," 2008.

[46] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

[47] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.

[48] D. Russo and B. Van Roy, "Eluder dimension and the sample complexity of optimistic exploration," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[49] R. Wang, R. R. Salakhutdinov, and L. Yang, "Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6123–6135, 2020.

[50] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path," *Machine Learning*, vol. 71, pp. 89–129, 2008.

[51] Z. Chen, C. J. Li, A. Yuan, Q. Gu, and M. I. Jordan, "A general framework for sample-efficient function approximation in reinforcement learning," *arXiv preprint arXiv:2209.15634*, 2022.

[52] D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari, "A note on model-free reinforcement learning with the decision-estimation coefficient," *arXiv preprint arXiv:2211.14250*, 2022.

[53] A. Agarwal and T. Zhang, "Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity," *arXiv preprint arXiv:2206.07659*, 2022.

[54] A. Agarwal and T. Zhang, "Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling," *arXiv preprint arXiv:2203.08248*, 2022.

[55] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[56] F. Chen, S. Mei, and Y. Bai, "Unified algorithms for rl with decision-

estimation coefficients: No-regret, pac, and reward-free learning," *arXiv preprint arXiv:2209.11745*, 2022.

[57] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," 2019.

[58] C. Szepesvári, "Algorithms for reinforcement learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 4, no. 1, pp. 1–103, 2010.

[59] Q. Liu, A. Chung, C. Szepesvári, and C. Jin, "When is partially observable reinforcement learning not scary?," *arXiv preprint arXiv:2204.08967*, 2022.

[60] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal, "Bellman-consistent pessimism for offline reinforcement learning," *Advances in neural information processing systems*, vol. 34, pp. 6683–6694, 2021.

[61] H. Zhong, W. Xiong, S. Zheng, L. Wang, Z. Wang, Z. Yang, and T. Zhang, "A posterior sampling framework for interactive decision making," *arXiv preprint arXiv:2211.01962*, 2022.

[62] T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade, "The role of coverage in online reinforcement learning," *arXiv preprint arXiv:2210.04157*, 2022.

[63] Z. Liu, M. Lu, W. Xiong, H. Zhong, H. Hu, S. Zhang, S. Zheng, Z. Yang, and Z. Wang, "One objective to rule them all: A maximization objective fusing estimation and planning for exploration," *arXiv preprint arXiv:2305.18258*, 2023.

[64] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," *Advances in neural information processing systems*, vol. 24, 2011.

[65] I. Osband, B. Van Roy, and Z. Wen, "Generalization and exploration via randomized value functions," in *International Conference on Machine Learning*, pp. 2377–2386, PMLR, 2016.

[66] W. Xiong, H. Zhong, C. Shi, C. Shen, and T. Zhang, "A self-play posterior sampling algorithm for zero-sum Markov games," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 24496–24523, PMLR, 17–23 Jul 2022.

[67] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[68] C. Jin, Q. Liu, and T. Yu, "The power of exploiter: Provable multi-agent rl in large state spaces," *arXiv preprint arXiv:2106.03352*, 2021.

[69] B. Huang, J. D. Lee, Z. Wang, and Z. Yang, "Towards general function approximation in zero-sum markov games," *arXiv preprint arXiv:2107.14702*, 2021.

[70] J. Perolat, B. Scherrer, B. Piot, and O. Pietquin, "Approximate dynamic programming for two-player zero-sum markov games," in *International Conference on Machine Learning*, pp. 1321–1329, PMLR, 2015.

[71] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

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

[73] T. Zhang, *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

[74] D. A. Freedman, "On tail probabilities for martingales," *the Annals of Probability*, pp. 100–118, 1975.