A General Framework for Sequential Decision-Making under Adaptivity Constraints

Nuoya Xiong [1] **Zhaoran Wang** [2] **Zhuoran Yang** [3]

Abstract

We take the first step in studying general sequential decision-making under two adaptivity constraints: rare policy switch and batch learning. First, we provide a general class called the Eluder Condition class, which includes a wide range of reinforcement learning classes. Then, for the rare policy switch constraint, we provide a generic algorithm to achieve a $\widetilde{\mathcal{O}}(\log K)$ switching cost with a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret on the EC class. For the batch learning constraint, we provide an algorithm that provides a $\widetilde{\mathcal{O}}(\sqrt{K}+K/B)$ regret with the number of batches B. This paper is the first work considering rare policy switch and batch learning under general function classes, which covers nearly all the models studied in the previous works such as tabular MDP (Bai et al., 2019; Zhang et al., 2020), linear MDP (Wang et al., 2021; Gao et al., 2021), low eluder dimension MDP (Kong et al., 2021; Velegkas et al., 2022), generalized linear function approximation (Qiao et al., 2023), and also some new classes such as the low D_Δ-type Bellman eluder dimension problem, linear mixture MDP, kernelized nonlinear regulator and undercomplete partially observed Markov decision process (POMDP).

1 Introduction

Reinforcement Learning (RL) provides a systematic framework for solving large-scale sequential decision-making problems and has demonstrated striking empirical successes across various domains (Li, 2017), including games (Silver et al., 2016; Vinyals et al., 2019), robotic control (Akkaya et al., 2019), healthcare (Yu et al., 2021), hardware device placement (Mirhoseini et al., 2017), recommender systems (Zou et al., 2020), and so on.

In the online setting, an RL algorithm iteratively finds the optimal policy of the sequential decision-making problem by (i) deploying the current policy to gather data and (ii) using the collected data to learn an improved policy. Most of provably sample efficient algorithms in the existing literature consider an ideal setting where policy updates can be *fully adaptive*, i.e., the policy can be updated after each episode, using the data sampled from this newly finished episode. From a practical perspective, however, updating the policy after each episode can be unrealistic, especially when computation resources are limited, or the cost of policy switching is prohibitively high, or the data is not fully serial. For example, in recommender systems, it is unrealistic to change the policy after each instantaneous data such as a click of one of the customers. Moreover, the customers might not come in a serial manner – it is possible that multiple customers arrive at the same time and we need to make simultaneous decisions. Similarly, when the RL algorithms are deployed on the large-scale hardwares, changing a policy may need recompiling the code or changing the physical placement for devices, incurring considerable switching costs. Thus, when it comes to designing RL algorithms in these scenarios, in addition to achieving sample efficiency, we also aim to reduce or limit the number of policy switches.

Such an additional restriction is known as the *adaptivity constraints* (Wang et al., 2021). There are two common types of adaptivity constraints: the rare policy switch constraint (Perchet et al., 2016; Gu et al., 2021) and the batch learning constraint (Abbasi-Yadkori et al., 2011). With the rare policy switch constraint, the agent adaptively decides when to update the policy during the course of the online reinforcement learning, and the goal is to achieve the sample efficiency that is comparable to the fully adaptive setting, while minimizing the number of policy switches. With the batch learning constraint, the total number of batches B is pre-determined, and the agent has to follow the same policy within each batch. In other words, the number of policy switches is limited by the number of batches. In addition to designing the policies used in each batch, the agent addi-

*Equal contribution [1]IIIS, Tsinghua University, China [2]Department of Industrial Engineering and Management Sciences, Northwestern University, USA [3]Department of Statistics and Data Science, Yale University, USA. Correspondence to: Nuoya Xiong <nuoyaxiong@gmail.com>, Zhuoran Yang <zhuoran.yang@yale.edu>, Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

tionally needs to decide how to split the total K episodes in to B batches before interacting with the environment.

Moreover, in many real applications of RL such as recommender systems, the state space can be extremely large or even infinite (Chen et al., 2019). Function approximation is an effective tool for handling such a challenge and has been extensively studied in the literature under the fully adaptive setting (Jiang et al., 2017; Sun et al., 2019; Foster et al., 2021; Jin et al., 2021; Zhong et al., 2022; Chen et al., 2022). Accordingly, a few previous works provide provably sample-efficient RL algorithms under adaptivity constraints for MDPs with linear and generalized linear structures (Gao et al., 2021; Wang et al., 2021; Qiao et al., 2023) and low-eluder-dimension MDPs (Kong et al., 2021; Velegkas et al., 2022). However, it remains open when considering more general classes such as Bellman Eluder dimension (Jin et al., 2021; Qiao et al., 2023). This motivates us to consider the following question:

Can we design sample-efficient RLs algorithm under adaptivity constraints in the context of general function approximation?

In this work, we establish the first algorithmic framework under adaptivity constraints for a general function class named **E**luder-**C**ondition (EC) class. Our framework applies to both single-agent MDP and one player in a zero-sum Markov game. Besides, the EC class contains many popular MDP and Markov game models studied in the previous literature. Some examples contained in our framework and the comparison of previous works are shown in Table 1. We also provide some additional examples in §E.

For the rare policy switch problem, motivated by the optimistic algorithms such as GOLF (Jin et al., 2021) and OPERA (Chen et al., 2022), our algorithm constructs an optimal confidence set and computes the optimal policy via optimistic planning based on the confidence set. Rather than updating the confidence set at each episode, we use a more delicate strategy to reduce the number of policy switches. As we employ optimistic planning, switching a policy essentially means that we update the confidence set that contains the true hypothesis. To reduce the number of policy switches, we update the confidence set only when the update provide considerable improvement in terms of the estimation. In particular, in each episode, we first estimate the improvement provided by the new confidence set, and then only decide to update the confidence set and optimal policy when the estimated improvement exceeds a certain threshold. Our lazy switching strategy can reduce the number of policy switches from $\mathcal{O}(K)$ to $\mathcal{O}(\text{poly}(\log K))$, leading to an exponential improvement in terms of policy switches. We also refer to the number of policy switches as *the switching cost*. Meanwhile, for the batch learning problem, we use a fixed uniform grid that divides K episodes into B batches. While the uniform grid is an intuitive and common choice (Wang et al., 2021; Han et al., 2020), analyzing the regret of this approach under general function classes requires new techniques. Our work takes the first step in studying the rare policy switch problem and the batch learning problem with general function approximation.

In summary, we make the following contributions:

• We provide a general function classes called the ℓ_2-Eluder Condition (EC) class and the ℓ_1-EC class, and then show that the EC class contains a wide range of previous RL models with general function approximation, such as low D_Δ-type Bellman eluder dimension model, linear mixture MDP, KNR and Generalized Linear Bellman Complete MDP.

• We develop a generic algorithm ℓ_2-EC-Rare Switch (RS) for the ℓ_2-type EC class. The algorithm uses optimistic estimation to achieve a $\widetilde{\mathcal{O}}(H\sqrt{dK})$ regret, updates the confidence set, and changes the policy by a delicate strategy to achieve a $\widetilde{\mathcal{O}}(dH\log K)$ switching cost, where d is the parameter in the EC class and $\widetilde{\mathcal{O}}$ contains the logarithmic term except for $\log K$. In Appendix F, we also provide ℓ_1-EC-RS algorithm for the ℓ_1-type EC class. We apply our results to some specific examples, showing that our method is sample-efficient with a low switching cost.

• For batch learning problems, we also develop an intuitive and generic algorithm ℓ_2-EC-Batch that achieves a $\widetilde{\mathcal{O}}(\sqrt{dHK} + dHK/B)$ regret, where B is the number of batches. Our regret is comparable to the existing works for batch learning in the linear MDP (Wang et al., 2021) and also matches their lower bound.

Related Works. Our paper is closely related to the prior research on RL with general function approximation and RL with adaptivity constraints. A comprehensive summary of the related literature can be found in §A.

2 Preliminaries

Episodic MDP A finite-horizon, episodic Markov decision process (MDP) is represented by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and action space; H is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is the transition kernel, where $\mathbb{P}_h(s_{h+1} \mid s_h, a_h)$ represents the probability to arrive state s_{h+1} when taking action a_h on state s_h at step h; $r = \{r_h(s,a)\}_{h\in[H]}$ denotes the deterministic reward function after taking action a at state s and step h. We assume $\sum_{h=1}^{H} r_h(s_h, a_h) \in [0,1]$ for all possible sequences $\{s_h, a_h\}_{h=1}^{H}$. A deterministic Markov policy $\pi = \{\pi_h\}_{h\in[H]}$ is a set of H functions, where $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a mapping from state to an action. For any policy π, its action value function $Q_h^\pi(s,a)$ and

Table 1: Comparison of previous representative work for adaptivity constraints and our works. For function classes with low eluder dimension, (Kong et al., 2021) needs a strong value-closeness assumption, while we do not need that assumption.

	(Wang et al., 2021)	(Kong et al., 2021)	(Qiao et al., 2023)	Ours
Tabular MDP	✓	✓	✓	✓
Linear MDP (Jin et al., 2020)	✓	✓	✓	✓
Low Eluder Dimension (Russo and Van Roy, 2013)	✗	✓	✗	✓
Low Inherent Bellman Error (Zanette et al., 2020)	✗	✗	✓	✗
Low D_Δ-type BE Dimension (Jin et al., 2021)	✗	✗	✗	✓
Linear Mixture MDP (Ayoub et al., 2020)	✗	✗	✗	✓
Kernelized Nonlinear Regulator (Kakade et al., 2020)	✗	✗	✗	✓
SAIL Condition (Liu et al., 2022b)	✗	✗	✗	✓
Undercomplete POMDP (Liu et al., 2022a)	✗	✗	✗	✓
Zero-Sum Markov Games with Low Minimax BE Dimension (Huang et al., 2021)	✗	✗	✗	✓

state value function $V_h^\pi(s)$ are defined as

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

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

To simplify the presentation, without loss of generality, we assume the initial state is fixed at s_1. The optimal policy π^* maximizes the value function, i.e., $\pi^* = \operatorname{argmax}_{\pi\in\Pi} V_1^\pi(s_1)$. We also denote $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$. Note that the function Q^* is the unique solution to the Bellman equations $Q_h^*(s,a) = \mathcal{T}_h Q_{h+1}^*(s,a)$, where the Bellman operator $\mathcal{T}$ is defined by

$$(\mathcal{T}_h Q_{h+1})(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim\mathbb{P}_h(s'|s,a)} \max_{a\in\mathcal{A}} Q_{h+1}(s,a). \tag{2.1}$$

We also study the low switching-cost problems under zero-sum Markov Games, and we put the definitions, learning objective, algorithm and results into §D.

Covering Number We provide the definition of ρ-covering number. In this work, we mainly consider the covering number with respect to the distance ℓ_∞ norm.

Definition 2.1 (ρ-Covering Number). The ρ-covering number of a function class $\mathcal{F}$ is the minimum integer t that satisfies the following property: There exists $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| = t$, and for any $f_1 \in \mathcal{F}$ we can find $f_2 \in \mathcal{F}'$ such that $\|f_1 - f_2\|_\infty \le \rho$.

Function Approximation Generally speaking, under the function approximation setting, we can access to a hypothesis class $\mathcal{F}$ which captures the key feature of the value functions (in the model-free setting) or the transition kernels and the reward functions (in the model-based setting) of the RL problem. In specific, let M denote the MDP instance, which is clear from the context. We assume that we have access to a hypothesis class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where a hypothesis function $f = \{f_1, \ldots, f_H\} \in \mathcal{F}$ either represents an action-value function $Q_f = \{Q_{h,f}\}_{h\in[H]}$ in the model-free setting, or the environment model of MDP $M_f = \{\mathbb{P}_{h,f}, r_{h,f}\}_{h\in[H]}$ in the model-based setting. Moreover, for any $f \in \mathcal{F}$, let π_f denote the optimal policy corresponding to the hypothesis f. That is, under the model-free setting, π_f is the greedy policy with respect to Q_f, i.e., $\pi_{h,f}(s) = \arg\max_{a\in\mathcal{A}} Q_{h,f}(s,a)$. Moreover, given Q_f and π_f, we define state-value function V_f by letting $V_{h,f}(s) = \mathbb{E}_{a\sim\pi_{h,f}(s)}[Q_{h,f}(s,a)]$ in the MDP. Furthermore, under the model-based setting, let $M_f = (\mathbb{P}_f, r_f)$ be the transition kernel and reward function associated with the hypothesis f. For the function approximation for zero-sum Markov Games, we put the definition in §D.

Similar to the previous works (Chen et al., 2022), we impose the following realizability assumption to make sure the true MDP or MG model M is captured by the hypothesis class $\mathcal{F}$.

Assumption 2.2 (Realizability). A hypothesis class $\mathcal{F}$ satisfies the realizability condition if there exists a hypothesis function $f^* \in \mathcal{F}$ such that $Q_{h,f^*} = Q_h^*, V_{h,f^*} = V_h^*$ for all $h \in [H]$.

Learning Goal In this paper, we aim to design online reinforcement learning algorithms for the rare policy switches problem and the batched learning problem. Assume the agent executes the policy π^k in the k-th episode for all $k \in [K]$, the regret of the agent is defined as

$$R(K) = \sum_{t=1}^{K} (V_1^*(s_1) - V_1^{\pi^k}(s_1)). \quad (2.2)$$

Moreover, for zero-sum MGs, we aim to design online reinforcement learning algorithms for the player P1 (max-player). In other words, we only control P1 and let P2 play arbitrarily. The goal is to design an RL algorithm such that P1's expected total return is close to the value of the game, namely $V_1^*(s_1)$. For any $k \in [K]$, in the k-th episode, players P1 and P2 executes policy pair $\pi^k = (\nu^k, \mu^k)$ and P1's expected total return is given by $V_1^{\pi^k}(s_1)$. The regret of P1 is also given by (2.3).

$$\begin{aligned} R(K) &= \sum_{t=1}^{K} (V_1^*(s_1) - V_1^{\pi^k}(s_1)) \\ &= \sum_{t=1}^{K} (V_1^*(s_1) - V_1^{\upsilon^k, \mu^k}(s_1)). \end{aligned} \quad (2.3)$$

The switching cost is the number of policy switches during the interactive process. For the single-agent setting, assume the agent uses the policy π^t in t-th episode, the switch cost at T-th episodes is:

$$N_{\text{switch}}(K) = \sum_{k=1}^{K} \mathbb{I}\{\pi^k \neq \pi^{k+1}\}.$$

For MGs with the decoupled setting, since we can only control the player P1, the switching cost is only defined on the action of player P1 υ, i.e.

$$N_{\text{switch}}(K) = \sum_{k=1}^{K} \mathbb{I}\{\upsilon^k \neq \upsilon^{k+1}\}.$$

In this paper, for the rare policy switch problem, we aim to achieve a logarithmic switching cost and maintain a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. In other words, we aim to design an algorithm such that $R(K) = \widetilde{\mathcal{O}}(\sqrt{K})$ while $N_{\text{switch}}(K) = \text{poly}\log(K)$, where we ignore problem dependent quantities and $\widetilde{\mathcal{O}}(\cdot)$ omits logarithmic terms.

For the batch learning problem, let $B \in [K]$ be a fixed integer. The agent of the MDP or the max-player in the zero-sum MG *pre-determines* a grid $1 = k_1 < k_2 < \cdots < k_{B+1} = K + 1$ with $B + 1$ points that split the K episodes into B batches $\{k_1, \ldots, k_2 - 1\}, \ldots, \{k_2, \ldots, k_3 - 1\}, \ldots, \{k_{B-1}, \ldots k_B - 1\}, \{k_B, \ldots, K\}$. In an MDP, the agent can only execute the same policy within a batch and change the policy only at the end of a batch. In a zero-sum MG, batch learning requires that the max-player can only change her policies at the end of each batch. Meanwhile, the min-player is free to change the policy after each episode. Similarly, the agent (max-player) aims to minimize the regret in (2.2) by (a) selecting the batching grid at the beginning of the algorithm and (b) designing the B policies that are executed in each batch. Furthermore, in the case of $B = K$, the problem is reduced to a standard online reinforcement learning problem.

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

3 Eluder-Condition Class

To handle the RL problems with adaptivity constraints, we propose a general class called Eluder-Condition (EC) class, which has a stronger eluder assumption and thus helps us to control the adaptivity constraints. There are two types of EC class: ℓ_2-EC class and ℓ_1-EC class. We mainly discuss the ℓ_2-EC class in the main text, and introduce the ℓ_1-EC class in §F. Before introducing the concepts of ℓ_2-EC class, we first consider the function class with low D_Δ-type BE dimension (Jin et al., 2021) as a primary example to show our stronger eluder assumption. Define the Bellman residual $\mathcal{E}_h(f)(s_h, a_h) = (f_h - \mathcal{T}(f_{h+1}))(s_h, a_h)$ for all $h \in [H]$. In the eluder argument (Lemma 17) of (Jin et al., 2021), it is proven that for any sequence $\{f^k\}_{k=1}^K$, if the Bellman error of f^k and historical data $\{s_h^i, a_h^i\}_{i=1}^{k-1}$ satisfy

$$\sum_{i=1}^{k-1} \left(\mathcal{E}(f_h^k)(s_h^i, a_h^i)\right)^2 \leq \beta,$$

then the sum of discrepancy over all episodes can be bounded by $\widetilde{\mathcal{O}}(\sqrt{K})$, i.e.

$$\sum_{i=1}^{k} \left|\mathcal{E}(f_h^i)(s_h^i, a_h^i)\right| \leq \mathcal{O}(\sqrt{d\beta k}), \ \forall k \in [K]. \quad (3.1)$$

However, (3.1) is not enough to control the adaptivity constraints such as the switching cost. Instead, we find that a slightly stronger assumption in (3.2) below helps us reduce the switching cost.

$$\sum_{i=1}^{k}\left(\mathcal{E}(f_h^i)(s_h^i,a_h^i)\right)^2 \le \mathcal{O}(d\beta\log k),\ \forall k\in[K]. \quad (3.2)$$

It is easy to show that (3.2) is slightly stronger than (3.1) by Cauchy's inequality. However, this stronger assumption enables help us to achieve a low switching cost through some additional analyses. Moreover, in Section E, we show that (3.2) also holds for a wide range of tractable RL problems studied in the previous work such as linear mixture MDP, D_Δ-type BE dimension and KNR.

Now we provide the formal definition of the ℓ_2-EC class. To provide a unified treatment for both MDP and MG, we let $\{\zeta_h,\eta_h\}_{h\in[H]}$ be subsets of the trajectory. In particular, we let $\eta_h=\{s_h,a_h\}$ and $\zeta_h=\{s_{h+1}\}$ in a single-agent MDP, and let $\eta_h=\{s_h,a_h,b_h\}$ and $\zeta_h=\{s_{h+1}\}$ in a two-player zero-sum MG.

Definition 3.1 (ℓ_2-type EC Class)**.** Given a MDP or MG instance M, let $\mathcal{F}$ and $\mathcal{G}$ be two hypothesis function classes satisfying the realizability Assumption 2.2 with $\mathcal{F}\subset\mathcal{G}$. For any $h\in[H]$ and $f'\in\mathcal{F}$, let $\ell_{h,f'}(\zeta_h,\eta_h,f,g)$ be a vector-valued and bounded loss function which serves as a proxy of the Bellman error at step h, where $f,f'\in\mathcal{F},g\in\mathcal{G}$, and ζ_h,η_h are subsets of trajectory defined above. Moreover, we assume that $\|\ell_{h,f'}(\zeta_h,\eta_h,f,g)\|_2$ is upper bounded by a constant R for all h, (f',f,g), and (ζ_h,η_h). For parameters d and κ, we say that $(M,\mathcal{F},\mathcal{G},\ell,d,\kappa)$ is a ℓ_2-type EC class if the following two conditions hold for any $\beta\ge R^2$ and $h\in[H]$:

(i). (ℓ_2-type Eluder Condition) For any K hypotheses $f^1,\dots,f^K\in\mathcal{F}$, if

$$\sum_{i=1}^{k-1}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^k,f^k)\right]\right\|_2^2\le\beta \quad (3.3)$$

holds for any $k\in[K]$, then we have

$$\sum_{i=1}^{k}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)\right]\right\|_2^2\le\mathcal{O}(d\beta\log k) \quad (3.4)$$

for all $k\in[K]$, where we consider R as a constant and ignore it in $\mathcal{O}(\cdot)$.

(ii). (κ-Dominance) There exists a parameter κ such that, for any $k\in[K]$, with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{i=1}^{k}\left(V_{1,f^i}(s_1)-V^{\pi_i}(s_1)\right)\\&\le\kappa\cdot\left(\sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{\eta_h\sim\pi_i}\left[\left\|\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h,\eta_h,f^i,f^i)]\right\|_2\right]\right).\end{aligned} \quad (3.5)$$

In this definition, the κ-dominance property (3.5) shows that the final regret is upper bounded by the cumulative expectation of in-sample loss, which is standard in many previous works (Du et al., 2021; Chen et al., 2022). The ℓ_2-type eluder condition is a generalized version of (3.2). Indeed, when we choose $\zeta_h=\{s_{h+1}\},\eta_h=\{s_h,a_h\}$, and $\ell_{h,f'}(\zeta_h,\eta_h,f,g)=Q_{h,g}(s_h,a_h)-r(s_h,a_h)-V_{h+1,f}(s_{h+1})$, the ℓ_2-type condition in (3.4) can be regarded as the condition involving the Bellman error, as shown in (3.1). Intuitively, the term $\sum_{i=1}^{k-1}\|\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^k,f^k)]\|^2$ in Eq.(3.3) represents the discrepancy between the function f^k and the previous data. This term can be regarded as the estimation error after $k-1$ episodes. The term $\sum_{i=1}^{k}\|\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)]\|_2^2$ in Eq.(3.4) represents the discrepancy between f^i and the data of i-th episode, which serves as an upper bound of the regret incurred in the first k episodes due to the κ-dominance condition Eq.(3.5). Hence EC class connects these two terms which has the following implication: The regret of an optimistic algorithm is small as long as it generates a sequence of functions $\{f^k\}_{k\in[K]}$ such that the estimation error of f^k on the data given by the previous $k-1$ episodes is small. The parameter d quantifies the hardness of achieving low regret via a small estimation error. From the previous discussion, it is easy to show that our assumption is stricter than the previous works, and this stricter assumption can help us to reduce the switching cost by some additional original analyses. As we will show later in Section E, it is satisfied by many previous important models like D_Δ-type BE dimension (Jin et al., 2021), which includes low eluder dimension (Kong et al., 2021; Velegkas et al., 2022) and linear MDP (Gao et al., 2021; Wang et al., 2021).

Moreover, in the ℓ_2-type EC class, we consider the decomposable loss function (Chen et al., 2022). The decomposable property generalizes one of the properties of Bellman error and implies the completeness assumption in previous work (Jin et al., 2021).

Definition 3.2 (Decomposable Loss Function (DLF) (Chen et al., 2022))**.** The loss function $\ell_{h,f'}(\zeta_h,\eta_h,f,g)$ is decomposable if there exists an operator $\mathcal{T}:\mathcal{F}\to\mathcal{G}$, such that

$$\begin{aligned}&\ell_{h,f'}(\zeta_h,\eta_h,f,g)-\mathbb{E}_{\zeta_h}\left[\ell_{h,f'}(\zeta_h,\eta_h,f,g)\right]\\&\quad=\ell_{h,f'}(\zeta_h,\eta_h,f,\mathcal{T}(f)).\end{aligned} \quad (3.6)$$

Also, the operator $\mathcal{T}$ satisfies that $\mathcal{T}(f^*)=f^*$.

The decomposable property claims that for any $f\in\mathcal{F},g\in\mathcal{G}$, there exists a function $\mathcal{T}(f)\in\mathcal{G}$ that is independent of g satisfying (3.6), which can be regarded as a generalized completeness assumption. For example, in the function classes with low D_Δ-type BE dimension for single-agent MDP, the operator $\mathcal{T}$ is selected as the Bellman operator for single-agent MDP, which is given in (2.1). In this case, we can choose when we choose $\zeta_h = \{s_{h+1}\}, \eta_h = \{s_h, a_h\}$, and $\ell_{h,f'}(\zeta_h, \eta_h, f, g) = Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$, then we have

$$\begin{aligned} &\ell_{h,f'}(\zeta_h, \eta_h, f, g) - \mathbb{E}_{\zeta_h}\left[\ell_{h,f'}(\zeta_h, \eta_h, f, g)\right] \\ &\quad = (\mathcal{T}_h V_{h+1}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})) \\ &\quad = \ell_{h,f'}(\zeta_h, \eta_h, f, \mathcal{T}(f)), \end{aligned} \tag{3.7}$$

where $\mathcal{T}$ is the Bellman operator. The following lemma shows that D_Δ-type Bellman eluder dimension model belongs to ℓ_2-type EC class with parameter $d = d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{T})$. The proof is provided in §I.5.

Lemma 3.3 (Low D_Δ-type Bellman Eluder Dimension $\subset$ ℓ_2-type EC class)**.** *Suppose the function class $\mathcal{F}$ with a low D_Δ-type Bellman eluder dimension with auxiliary function class $\mathcal{G}$ (Jin et al., 2021), then for any MDP model M, choose $\zeta_h = \{s_{h+1}\}, \eta_h = \{s_h, a_h\}$ and $(M, \mathcal{F}, \mathcal{G},)$, $\kappa = 1$, and DLF ℓ as in 3.7, then $(M, \mathcal{F}, \mathcal{G}, \ell, d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{T}), \kappa)$ is a ℓ_2-type EC class by ℓ_2 condition: If $\sum_{i=1}^{k-1}[\mathcal{E}(f^k, s_h^i, a_h^i)^2] \le \beta$ holds for any $k \in [K]$ and $\beta \ge 9$, then for any $k \in [K]$ we have*

$$\sum_{i=1}^{k}[\mathcal{E}(f^i, s_h^i, a_h^i)^2] \le \mathcal{O}(d\beta \log K), \tag{3.8}$$

where $d = d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{T})$ and we choose the upper bound of the Bellman residual $\mathcal{E}$ as $R = 3$. The dominance can be derived by Lemma 1 in (Jiang et al., 2017).

For some other particular examples like linear mixture MDP and KNR, the operator $\mathcal{T}$ is chosen as the optimal operator $\mathcal{T}(f) = f^*$. The detailed definition and theoretical results are provided in §E.

In recent years, the ℓ_1-eluder argument is proposed in (Liu et al., 2022a) and followed by (Liu et al., 2022b) to provide another way for the sample-efficient algorithm of POMDP. In §F, we also provide a similar EC class named ℓ_1-type EC class. Compared to the ℓ_2-EC class, ℓ_1-EC class replaces the square sum in the ℓ_2-type EC property (Eq. (3.4)) by a standard sum. By considering a particular model-based loss function, the ℓ_1-type EC class can reduce to the assumption in (Liu et al., 2022b). We provide a sample-efficient algorithm for ℓ_1-EC class with low switching cost in §F.2 and a batch learning algorithm in Appendix H.3.

4 Rare Policy Switch Problem

In this section, we propose an algorithm for the ℓ_2-type EC class that achieves a low switching cost. Our algorithm extends the optimistic-based exploration algorithm (Jin et al., 2021; Chen et al., 2022) with a lazy policy switches strategy. The optimistic-based exploration algorithm calculates a confidence set using historical data and performs optimistic planning within this set to determine the optimal model f^k and policy π^k at each episode k. Unlike the previous algorithm, we choose to update the confidence set only when a specific condition holds. This modification helps reduce the frequency of policy switches and lowers the associated cost.

Algorithm 1 ℓ_2-EC-RS

1: **Initialize:** $D_1, D_2, \cdots, D_H = \emptyset, \mathscr{B}_1 = \mathcal{F}$.
2: **for** $k = 1, 2, \cdots, K$ **do**
3: (MDP): Compute the greedy policy $\pi^k = \pi_{f^k}$, where $f^k = \arg\max_{f \in \mathscr{B}^{k-1}} V_f^{\pi_f}(s_1)$.
4: (Zero-Sum MG): Compute $\upsilon^k = \upsilon_{f^k}$, where $f^k = \arg\max_{f \in \mathscr{B}_{k-1}} V_f^{\upsilon_f, \mu_f}(s_1)$. The adversary chooses strategy μ^k, then we let $\pi^k = (\upsilon^k, \mu^k)$.
5: Execute policy π^k to collect the trajectory, update $D_h = D_h \cup \{\zeta_h^k, \eta_h^k\}, \forall h \in [H]$.
6: **if** $L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^k, g) \ge 5\beta$ for some $h \in [H]$ **then**
7: Update $\mathscr{B}^k = \Big\{f \in \mathcal{F} : L_h^{1:k}(D_h^{1:k}, f, f) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f, g) \le \beta, \forall h \in [H]\Big\}$.
8: **else**
9: $\mathscr{B}^k = \mathscr{B}^{k-1}$.
10: **end if**
11: **end for**

Optimistic Exploration with Low-Switching Cost. At episode k, the agent first computes the optimal policy $\pi^k = \pi_{f^k}$, where f^k is the optimal model in confidence set $\mathscr{B}_k$ and π_f is the greedy policy with respect to Q_f. Then it executes the policy π^k (or υ^k for zero-sum MG) and collects the data D_h for each step $h \in [H]$. Line 6 - 10 compute the optimistic confidence set $\mathscr{B}^{k+1}$ for next episode $k+1$. Define the loss function

$$L_h^{a:b}(D_h^{a:b}, f, g) = \sum_{i=a}^{b} \|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f, g)\|_2^2,$$

and calculate the confidence sets in Line 7 based on history data. Unlike the previous algorithm, our algorithm provides a novel policy switching condition in Line 6, which is the following inequality:

$$L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^k, g) \ge 5\beta \tag{4.1}$$

for some $h \in [H]$, where β is a logarithmic confidence parameter. In fact, the left-hand side of (4.1) represents the in-sample discrepancy between f^k and the historical data $D_h^{1:k}$ at step h after the first k episodes. When (4.1) does not hold, then we have

$$L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^k, g) \le 5\beta \tag{4.2}$$

for all $h \in [H]$. Moreover, for any $k \in [K]$, we have $f^k \subseteq \mathscr{B}^{k-1}$. Moreover, let t_{k-1} be the index of the episode after which $\mathscr{B}^{k-1}$ is constructed. That is, t_{k-1} is the smallest t such that $\mathscr{B}^t = \mathscr{B}^{k-1}$. Then by the construction of the confidence set $\mathscr{B}^{k-1}$, the discrepancy between f^k and the historical data $D_h^{1:t_{k-1}}$ satisfies

$$L_h^{1:t_{k-1}}(D_h^{1:t_{k-1}}, f^k, f^k) - \inf_{g \in \mathcal{G}} L_h^{1:t_{k-1}}(D_h^{1:t_{k-1}}, f^k, g) \leq \beta. \tag{4.3}$$

Comparing (4.2) and (4.3), we observe that, when adding the new data from $(t_{k-1}+1)$-th to the k-th episode, the discrepancy between the collected data and f^k remains relatively small for all steps $h \in [H]$. In this case, the improvement brought from adding new data limited, and thus we choose not to update the policy to save computation. Instead, when (4.1) holds, this means that the discrepancy between f^k and the offline data $D_h^{1:k}$ is significant. In light of (4.3), the newly added data from $(t_{k-1}+1)$-th to the k-th episode brings considerable new information from newly collected data, and thus we update the confidence set and hence update the policy. Furthermore, in the following theorem, we prove that such a lazy policy switching scheme achieves both sample efficiency while incurring a small switching cost, assuming the underlying model belongs to the ℓ_2-type EC class. The detailed proof of the theorem is provided in §G.1.

Theorem 4.1. *Given an EC class $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ with two hypothesis classes $\mathcal{F}, \mathcal{G}$ and a decomposable loss function ℓ satisfying Eq. (3.4), Eq. (3.5) and Definition 3.2. Set $\beta = c(R^2\iota + R)$ for a large constant c with $\iota = \log(HK^2\mathcal{N}_{\mathcal{L}}(1/K)/\delta)$, in which $\mathcal{N}_{\mathcal{L}}(1/K)$ is the $1/K$-covering number for DLF class $\mathcal{L} = \{\ell_{h,f'}(\cdot,\cdot,f,g) : (h, f', f, g) \in [H] \times \mathcal{F} \times \mathcal{F} \times \mathcal{G}\}$ with norm $\|\cdot\|_\infty$ (Defined in §B). With probability at least $1-\delta$, Algorithm 1 achieves a sublinear regret*

$$R(K) \leq \widetilde{\mathcal{O}}(\kappa H \sqrt{d\beta K} \cdot \mathrm{poly}(\log K)),$$

Also, Algorithm 1 has a logarithmic switching cost

$$N_{switch}(K) \leq \mathcal{O}(dH \cdot \log K). \tag{4.4}$$

The theorem above gives us an upper bound for both a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and a logarithmic switching cost. When applying to the specific examples such as function class with low D_Δ-type BE dimension $d = d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{K})$, Combining Lemma 3.3 and Theorem 4.1, Algorithm 1 achieves a $\widetilde{\mathcal{O}}(H\sqrt{d\beta K}\log(K)) = \widetilde{\mathcal{O}}(H\sqrt{d\log(N_{\mathcal{L}}(1/K)/\delta)K} \cdot \mathrm{poly}(\log K))$ regret and a $\mathcal{O}(dH \cdot \log K)$ switching cost. Some additional examples of ℓ_2-type EC classes such as linear mixture MDP and KNR, and the corresponding theoretical results are provided in Section E.

Comparison with Previous Algorithms The natural idea to solve low switching-cost problems is to measure the information gain and change the policy only when the gained information is large enough. However, the previous techniques to represent the gained information cannot apply to more general RL problems. For the tabular MDP and linear MDP (Gao et al., 2021; Wang et al., 2021), the gain of new information can be explicitly formulated as the determinant of the Hessian matrix of the least-squares loss function. For the function classes with low eluder dimension (Kong et al., 2021), their algorithm requires the construction of the bonus function and a sensitivity-based subsampling approach, which cannot be extended beyond their setting. Moreover, they require a value closeness assumption: For each function $V : \mathcal{S} \to [0, H]$, they assume the function class $\mathcal{F}$ satisfies that $r(s,a) + \sum_{s'} \mathbb{P}(s' \mid s, a)V(s') \in \mathcal{F}$ for all (s, a). This assumption is very stringent and is not satisfied by many general classes such as linear mixture MDP and KNR. All of these approaches cannot be applied to our EC class. The primary difficulty for the low switching cost problems under general function approximation is to quantify the information gain for general nonlinear models. We quantify the information by the discrepancy between the estimated model and the historical data, and find a delicate condition for policy switches. Compared to (Kong et al., 2021), we can derive a $\sqrt{K}$ regret and logarithmic switching cost *without a restrictive value-closeness assumption*.

The computational complexity mainly depends on the Line 3 or 4 in Algorithm 1. Previous works often assume there exists an oracle that approximately solves Line 3, e.g., (Jin et al., 2021; Chen et al., 2022). Such an oracle is queried in each episode to update the policy. Thus, these works incur an $\mathcal{O}(K)$ oracle complexity. In contrast, with the lazy update scheme specified in Lines 7–9, the oracle complexity of Algorithm 1 is $\mathcal{O}(\log K)$, which leads to an exponential improvement in terms of the computational cost.

To be more specific, checking the switching condition is always easier than implementing the oracle, since it only needs to calculate a term $L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g \in G} L_h^{1:k}(D_h^{1:k}, f^k, g)$, which is also required for constructing the constrained set B^k needed for the oracle. However, implementing the oracle requires solving an additional optimization problem $f^k = \arg\max_{f \in B^{k-1}} V_f^{\pi_f}(s_1)$ with constraint set B^{k-1}, which is much harder.

In the experiment, the execution time of our algorithm is 20 times faster than the algorithm without lazy policy switches, while maintaining a similar performance.

5 Batch Learning Problem

In this section, we provide an algorithm for the batch learning problem. Recall that in the batch learning problem, the agent selects the batch before the algorithm starts, then she uses the same policy within each batch. Denote the number of batches as B, Algorithm 2 try to divide each batch equally and choose the batch as $[k_i, k_{i+1})$, where $k_i = i \cdot \lfloor K/B \rfloor + 1$. This selection is intuitive and common in many previous works of batch learning (Han et al., 2020; Wang et al., 2021; Gu et al., 2021). After setting the batches, the agent adopts optimistic planning for policy updates, and only updates the policies in episodes $\{k_i, i \in [B-1]\}$. The details of the algorithm is presented in Algorithm 2.

In the following theorem, we provide a regret upper bound for Algorithm 2. The detailed proof is in §H.1.

Theorem 5.1. *Given an EC class $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ with two hypothesis classes $\mathcal{F}, \mathcal{G}$ and a decomposable loss function ℓ satisfying 3.4 and 3.5. Set $\beta = c(R^2\iota + R)$ for a large constant c with $\iota = \log(HK^2\mathcal{N}_{\mathcal{L}}(1/K)/\delta)$, in which $\mathcal{N}_{\mathcal{L}}(1/K)$ is the $1/K$-covering number for DLF class $\mathcal{L} = \{\ell_{h,f'}(\cdot, \cdot, f, g) : (h, f', f, g) \in [H] \times \mathcal{F} \times \mathcal{F} \times \mathcal{G}\}$ with norm $\|\cdot\|_\infty$. With probability at least $1-\delta$ the Algorithm 2 will achieve a sublinear regret*

$$R(T) \le \widetilde{\mathcal{O}}\left(\kappa H\sqrt{d\beta K}\log K + \kappa \cdot \frac{dHK}{B} \cdot (\log K)^2\right).$$

Hence if we choose $B = \Omega(\sqrt{K/d})$, we can get a sublinear regret $\widetilde{\mathcal{O}}(H\sqrt{d\beta K})$. In particular, for the linear MDP with dimension d_{lin}, we have $d = \widetilde{\mathcal{O}}(d_{\text{lin}})$ and $\beta = \widetilde{\mathcal{O}}(d_{\text{lin}} \cdot \text{poly}(\log K))$, Theorem 5.1 achieves a $\widetilde{\mathcal{O}}(Hd_{\text{lin}}\sqrt{K} + d_{\text{lin}}HK/B)$ regret upper bound. Note that (Wang et al., 2021) provide a regret lower bound $\Omega(d\sqrt{HK} + dHK/B)$ after rescaling the reward to $\sum_{h=1}^H r_h(s_h, a_h) \in [0, 1]$. Thus, the first term of our result has an additional $\sqrt{H}$ term, and our result matches the lower bound in terms of d, B, and K.

For function classes with low D_Δ-type Eluder dimension, Algorithm 2 provides a $\widetilde{\mathcal{O}}(H\sqrt{d_{\text{BE}}(\mathcal{F}, D_\Delta, 1/\sqrt{T})\beta K}\log K + d_{\text{BE}}(\mathcal{F}, D_\Delta, 1/\sqrt{T})HK(\log K)^2/B)$ regret upper bound. More specific examples and the corresponding results are provided in Section E. We also provide the batch learning results of ℓ_1-EC class in §F.2.

Now we provide the proof sketch. First, for a batch j, we consider the maximum in-sample error brought by this batch: $\max_{k \in [k_j, k_{j+1}-1]} L_h^{k_j:k}(D_h^{k_j:k}, f^{k_j}, f^{k_j}) - L_h^{k_j:k}(D_h^{k_j:k}, f^{k_j}, \mathcal{T}(f^{k_j})) \triangleq c_j\beta$. Indeed, this term represents the maximum fitting error for the data within the data of this batch and the model f^{k_j}. Then for all batches $[k_j, k_{j+1}-1]$ with a small in-sample error, namely, $c_j = O(1)$, we can still deploy the optimism mechanism and control the regret, thus the final regret can vary in magnitude by at most a constant. Moreover, for these batches with $c_j \le 5$, we call them the "Good" batches, meaning that the regret caused by these batches can still be upper bounded by $\mathcal{O}(\sqrt{K})$. For batch j with $c_j > 5$, we called them the "Bad" batches. Then we can show a fact that the number of "Bad" batches is at most $\mathcal{O}((\log K)^2)$. In fact, we can divide all the $c_j \in [5, K]$ into $\mathcal{O}(\log K)$ intervals $[5 \cdot 2^i, 5 \cdot 2^{i+1})_{i \ge 0}$, and use the ℓ_2-type eluder condition to bound that $|\{j \mid C/2 \le c_j \le C\}| \le \mathcal{O}(\log K)$ for any constant C.

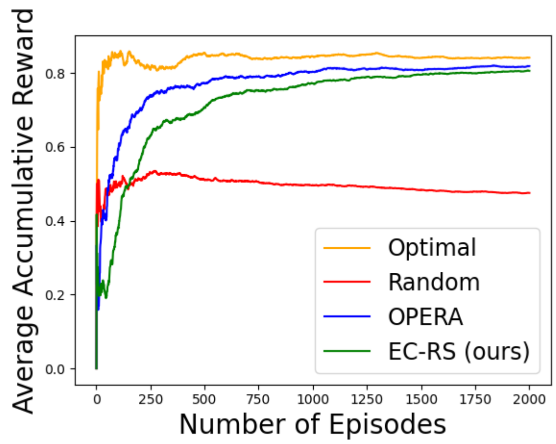


Figure 1: The average accumulative reward for optimal policy, random policy, OPERA algorithm (Chen et al., 2022) and EC-RS (Algorithm 1)

Once the fact is proven, the regret can be derived by adding "Good" batches and "Bad" batches. All "Good" batches will lead to at most a $\mathcal{O}(\sqrt{K})$ regret, and all "Bad" batches will lead to at most $\mathcal{O}((K/B) \cdot (\log K)^2)$ regret. Combining two types of batches, we can get Theorem 5.1.

Moreover, we consider another batch learning setting called "the adaptive batch setting" that was studied in (Gao et al., 2019). In this setting, the agent can select the batch size adaptively during the algorithm. At the end of each batch, the agent observes the reward feedback of this batch, and she can select the next batch size according to the historical information and change the policy. We show that in this setting, $\mathcal{O}(\text{poly}(\log K))$ batches are sufficient for a $\mathcal{O}(\sqrt{K})$ regret. The proof employs an extra doubling trick performed on the low switching cost Algorithm 1 and we discuss it in §H.2.

6 Experiment

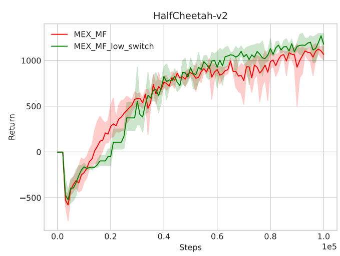
6.1 Linear Mixture MDP

We experimented in the linear mixture MDP with the same setting as Appendix H in (Chen et al., 2022). We choose $T = 2000$ and $\beta = 0.3 \log T$ in the experiment. We compare our ℓ_2-EC-RS algorithm with OPERA (Chen et al.,

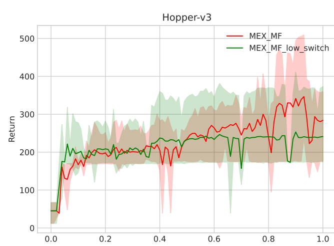
Table 2: The number of policy switches for MEX_MF (Liu et al., 2024) and our MEX_MF_low_switch algorithm on different Mujoco tasks.

	HalfCheetah-v2	Hopper-v3	Walker2d-v3
MEX_MF (Liu et al., 2024)	100000	100000	100000
MEX_MF_low_switch **(Ours)**	10052	29930	10742

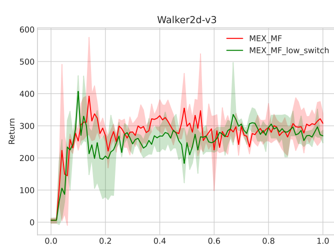
2022), optimal policy and the random policy. The cumulative reward curves show that our algorithm converges to the optimal value slightly slower than OPERA. However, the average number of strategy transitions and calls to the optimization tool decreases from 2000 to 92.8 times over 10 simulations, decreasing the average execution time from 321.6 seconds to 15.9 seconds.



(a) HalfCheetah-v2



(b) Hopper-v3



(c) Walker2d-v3

Figure 2: Model-Free Mujoco Problems

6.2 Model-Free Mujoco Problem

To show the potential insight of our theoretic discovery, we also implement a rare policy switch "MEX_MF_low_switch" algorithm on Mujoco problems based on our theoretical insight, which changes the policy only when the discrepancy between previous data and the current estimate is large enough. Based on the model-free algorithm "MEX_MF" in Section 7 of (Liu et al., 2024), we modify it by changing the policy only when the loss (Equation (7.1) in (Liu et al., 2024)) is large enough.

We execute two algorithms on three different tasks "Hopper-v3", "HalfCheetah-v2" and "Walker2d-v3" for 100000 episodes, and the setting is the same as Section 7 of (Liu et al., 2024). The comparisons of the rewards and the number of policy switches are shown in Figure 2 and Table 2. The results show that the rare policy switch algorithm maintains a similar performance compared to the original algorithm MEX_MF. However, there is a huge reduction in the number of policy switches. Also, the rare policy switch algorithm performs even better under some settings than the previous one. This is likely due to the implementation of a conservative policy-switching algorithm, which prevents premature and potentially detrimental changes. Consequently, it enhances the stability of the MEX_MF algorithm.

7 Conclusion

In this paper, we study the general sequential decision-making problem under general function approximation with two adaptivity constraints: the rare policy switch constraint and the batch learning constraint. Motivated by the eluder argument, we first introduce a general class named EC class that includes various previous RL models, and then provide algorithms for both two adaptivity constraints. For the rare policy switch problem, we propose a lazy policy switch strategy to achieve a logarithmic switching cost while maintaining a sublinear regret on the EC class. For the batch learning problem, we analyze the regret when the batch is a uniform grid on EC class. The result matches the lower bound under the linear MDP (Wang et al., 2021) in terms of d, B and T. To the best of our knowledge, this paper is the first work to systematically investigate these two adaptivity constraints under a general framework that contains a wide range of RL problems.

Impact Statement

The goal of this paper is to advance the field of theoretical reinforcement learning under general function approximation with adaptivity constraints. None of the potential impact must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR.

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

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. (2019). Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32.

Bradtke, S. (1992). Reinforcement learning applied to linear quadratic regulation. *Advances in neural information processing systems*, 5.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. H. (2019). Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464.

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

Cui, Q., Zhang, K., and Du, S. S. (2023). Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. *arXiv preprint arXiv:2302.03673*.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022). Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR.

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

Foster, D. J., Foster, D. P., Golowich, N., and Rakhlin, A. (2023). On the complexity of multi-agent decision making: From learning in games to partial monitoring. *arXiv preprint arXiv:2305.00684*.

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

Gao, M., Xie, T., Du, S. S., and Yang, L. F. (2021). A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. (2019). Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32.

Gu, Q., Karbasi, A., Khosravi, K., Mirrokni, V., and Zhou, D. (2021). Batched neural bandits. *arXiv preprint arXiv:2102.13028*.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

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

Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., and Yang, L. (2021). Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

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

Jin, C., Liu, Q., and Yu, T. (2022). The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. (2020). Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325.

Kong, D., Salakhutdinov, R., Wang, R., and Yang, L. F. (2021). Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. (2022a). When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. (2022b). Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. (2023). One objective to rule them all: A maximization objective fusing estimation and planning for exploration.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. (2024). Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36.

Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., and Dean, J. (2017). Device placement optimization with reinforcement learning. In *International Conference on Machine Learning*, pages 2430–2439. PMLR.

Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. (2016). Batched bandit problems. *The Annals of Statistics*, pages 660–681.

Qiao, D., Yin, M., Min, M., and Wang, Y.-X. (2022). Sample-efficient reinforcement learning with loglog (t) switching cost. In *International Conference on Machine Learning*, pages 18031–18061. PMLR.

Qiao, D., Yin, M., and Wang, Y.-X. (2023). Logarithmic switching cost in reinforcement learning beyond linear mdps. *arXiv preprint arXiv:2302.12456*.

Qiu, S., Ye, J., Wang, Z., and Yang, Z. (2021). On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

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

Velegkas, G., Yang, Z., and Karbasi, A. (2022). Reinforcement learning with logarithmic regret and policy switches. *Advances in Neural Information Processing Systems*, 35:36040–36053.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

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

Wang, T., Zhou, D., and Gu, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. (2023). Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. (2020). Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. (2022). Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*.

Zhang, Z., Jiang, Y., Zhou, Y., and Ji, X. (2022). Near-optimal regret bounds for multi-batch reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24586–24596.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207.

Zhao, H., He, J., and Gu, Q. (2023). A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *arXiv preprint arXiv:2311.15238*.

Zhao, Y., Tian, Y., Lee, J., and Du, S. (2022). Provably efficient policy optimization for two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761. PMLR.

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

Zhong, H., Yang, Z., Wang, Z., and Jordan, M. I. (2021). Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopic followers? *arXiv preprint arXiv:2112.13521*.

Zou, L., Xia, L., Du, P., Zhang, Z., Bai, T., Liu, W., Nie, J.-Y., and Yin, D. (2020). Pseudo dyna-q: A reinforcement learning framework for interactive recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 816–824.

Appendix

A Related Work

RL with General Function Approximation To solve the large-state RL problems, many works consider capturing the special structures of the MDP models. (Jiang et al., 2017) consider the RL problems with low Bellman rank; (Jin et al., 2020) consider a particular linear structure of MDP models named *linear* MDP; (Wang et al., 2020) consider RL problems with bounded *Eluder dimension* using sensitivity sampling, and (Ishfaq et al., 2021) use a simpler optimistic reward sampling to combine the optimism principle and Thompson sampling. (Jin et al., 2021) capture an extension of Eluder dimension called *Bellman Eluder dimension*; (Du et al., 2021) consider a particular model named *bilinear* model; Very recently, (Chen et al., 2022) consider a more extensive class ABC that contains many previous models. (Foster et al., 2021; Agarwal and Zhang, 2022; Zhong et al., 2022) provide the posterior-sampling style algorithm for sequential decision-making. Previous works also consider the Markov Games with multiple players under the function approximation setting. (Xie et al., 2020; Huang et al., 2021; Qiu et al., 2021; Jin et al., 2022; Zhao et al., 2022; Liu et al., 2023) study the two-player zero-sum Markov Games with linear or general function approximations. (Zhong et al., 2021; Ding et al., 2022; Cui et al., 2023; Wang et al., 2023; Foster et al., 2023) further consider the general-sum Markov Games with function approximations. Among these works, our paper is particularly related to the works that use the eluder dimension to capture the complexity of a function class (Jin et al., 2021; Chen et al., 2022; Liu et al., 2022a;b). Compared to them, our EC class requires a property that is slightly stricter than the normal eluder argument, which can help us to deal with the additional adaptivity constraints. More details are provided in Section 3.

RL with Adaptivity Constraints The rare policy switch problem and the batch learning problem are the two main adaptivity constraints considered in the previous works.

(Abbasi-Yadkori et al., 2011; Auer et al., 2002) give algorithms to achieve a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and a $\widetilde{\mathcal{O}}(\log\log K)$ switching cost in the bandit problem. (Bai et al., 2019) study the rare policy switch problem for tabular MDP and (Zhang et al., 2020) improve their results. (Qiao et al., 2022; Zhang et al., 2022) provide $\widetilde{\mathcal{O}}(\log\log T)$ switching cost algorithms for tabular MDP. (Wang et al., 2021; Gao et al., 2021) provide low-switching cost algorithms for linear MDP. (Kong et al., 2021) considers this problem for the function classes with low eluder dimension, and (Velegkas et al., 2022) extends their results to be gap-dependent. (Qiao et al., 2023) gives an algorithm to achieve logarithmic switching cost in the linear complete MDP with low inherent Bellman error and generalized linear function approximation. They leave the low switching cost problem for the function classes with low BE dimension as an open problem, where our paper solves a part of this open problem (D_Δ-type BE dimension). Concurrently, (Zhao et al., 2023) considers the rare policy switch algorithms under the function classes with low general eluder dimension. However, the connection between the general eluder dimension and the Bellman-Eluder dimension or linear mixture MDP is still unclear, while our result is more general to solve the rare policy switch problem under these function classes. Also, they need a stricter completeness assumption. $\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[r_h(s,a)+V(s')]\in F$ and $\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[(r_h(s,a)+V(s'))^2]\in F$ for any function $V: S\mapsto[0,1]$ (Assumption 2.2 in (Zhao et al., 2023)).

For the batch learning problem, (Perchet et al., 2016) considers this problem in 2-armed bandits, and (Gao et al., 2019) further considers this problem in the multi-armed bandit with both fixed batch size and adaptive batch size. (Han et al., 2020) study batch learning problem in the linear contextual bandits. (Gu et al., 2021) also study this problem in the neural bandit setting. (Wang et al., 2021) considers this problem in the linear MDP, and gives a lower bound of the regret.

B Definition of Bracketing Number

In the Theorem 4.1, Theorem 5.1, Theorem F.4 and Theorem F.5, the regret results contain the logarithmic term of the $1/K$-covering number of the function classes $\mathcal{N}_{\mathcal{F}}(1/K)$ or the $1/K$-bracketing number $\mathcal{B}_{\mathcal{F}}(1/K)$. Both of them can be regarded as a surrogate cardinality of the function class $\mathcal{F}$. The definition of the covering number is provided in 2.1. In this subsection, we provide the definition of the bracketing number, which is commonly used in the previous model-based RL works (Zhan et al., 2022; Zhong et al., 2022).

Definition B.1 (ρ-Bracket Number)**.** A ρ-bracket with size N contains $2N$ functions $\{f_1^i, f_2^i\}_{i=1}^N$ that maps a policy π and a trajectory τ to a real value such that $\|f_1^i(\pi,\cdot)-f_2^i(\pi,\cdot)\|_1\le\rho$. Moreover, for any $f\in\mathcal{F}$, there exists $i\in[N]$ such that $f_1^i(\pi,\tau)\le\mathbb{P}_f^\pi(\tau)\le f_2^i(\pi,\tau)$. The ρ-bracket number of a function class $\mathcal{F}$, denoted as $\mathcal{B}_{\mathcal{F}}(1/K)$, is the minimum size N of a ρ-bracket.

Algorithm 2 ℓ_2-EC-Batch

```
Input D_1, D_2, ..., D_H = ∅, 𝓑_1 = 𝓕.
for k = 1, 2, ..., K do
    (MDP): Compute π^k = π_{f^k}, where f^k = argmax_{f∈𝓑^{k-1}} V_f^{π_f}(s_1).
    (Zero-Sum MG): Compute υ^k = υ_{f^k}, where f^k = argmax_{f∈𝓑_{k-1}} V_f^{υ_f,μ_f}(s_1). The adversary chooses strategy μ^k,
    then we let π^k = (υ^k, μ^k).
    Execute policy π^k to collect the trajectory, update D_h = D_h ∪ {ζ_h^k, η_h^k}, ∀h ∈ [H].
    if k = i · ⌊K/B⌋ + 1 for some i ≥ 0 then
        Update
            𝓑^k = { f ∈ 𝓕 : L_h^{1:k}(D_h^{1:k}, f, f) − inf_{g∈𝒢} L_h^{1:k}(D_h^{1:k}, f, g) ≤ β, ∀h ∈ [H] }.
    else
        𝓑^k = 𝓑^{k-1}.
    end if
end for
```

As shown in (Zhan et al., 2022), the logarithm of the $1/K$-bracket number $\log(\mathcal{B}_{\mathcal{F}}(1/K))$ usually scales polynomially with respect to the parameters of the problem.

C Pseudo-Code of Algorithm for Batch Learning

In this subsection, we provide the pseudo-code of our algorithm ℓ_2-EC-batch for batch learning, which divides the entire episode into a uniform grid. The pseudo-code of the algorithm is provided in Algorithm 2.

D Switching-Cost for Zero-Sum Markov Games

D.1 Definition

Markov Games A zero-sum Markov Game (MG) consists of two players, while max-player P1 wants to maximize the reward and min-player P2 wants to minimize it. The model is represented by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}(\cdot \mid s, a, b), r(s, a, b))$, where $\mathcal{A}$ and $\mathcal{B}$ denote the action space of player P1 and P2 respectively. Similar to the episode MDP, we also assume $\sum_{h=1}^{H} r_h(s_h, a_h, b_h) \in [0, 1]$ for all possible sequences $\{s_h, a_h, b_h\}_{h=1}^{H}$ in this paper. The policy pair $(\upsilon, \mu) = \{\upsilon_h, \mu_h\}_{h\in[H]}$ consists of $2H$ functions $\upsilon_h : \mathcal{S} \to \Delta_{\mathcal{A}}, \mu_h : \mathcal{S} \to \Delta_{\mathcal{B}}$. For any policy (υ, μ), the action value function and state value function can be represented by

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

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

Given the policy of P1 υ, the *best response policy* of P2 is $\mu_\upsilon^* = \arg\min_\mu V_1^{\upsilon,\mu}(s_1)$. Similarly, the best response of P1 given μ is $\upsilon_\mu^* = \arg\max_\upsilon V_1^{\upsilon,\mu}(s_1)$. The Nash Equilibrium (NE) of an MG is a policy pair (υ^*, μ^*) such that

$$V_1^{\upsilon^*,\mu^*_{\upsilon^*}}(s_1) = V_1^{\upsilon^*,\mu^*}(s_1) = V_1^{\upsilon^*_{\mu^*},\mu^*}(s_1).$$

We denote $V_h^{\upsilon^*,\mu^*}$ and $Q_h^{\upsilon^*,\mu^*}$ by V_h^* and Q_h^* respectively in the sequel. In addition, to simplify the notation, we let $\pi = (\upsilon, \mu)$ denote the joint policy of the two players. Then we write $V_h^\pi(s_1) = V_h^{\upsilon,\mu}(s_1)$. Similar to the MDP, we can define the Bellman operator $\mathcal{T}$ for a MG by letting

$$(\mathcal{T}_h Q_{h+1})(s_h, a_h, b_h) = r_h(s_h, a_h, b_h) + \mathbb{E}_{s_{h+1}}\big[\max_\upsilon \min_\mu Q_{h+1}(s_{h+1}, \upsilon, \mu)\big], \tag{D.1}$$

where we denote $Q_{h+1}(s, \upsilon, \mu) = \mathbb{E}_{a\sim\upsilon, b\sim\mu}[Q_{h+1}(s, a, b)]$. By definition, $Q^* = \{Q^*_h\}_{h\in[H]}$ is the unique fixed point of $\mathcal{T}$, i.e., $Q^*_h = \mathcal{T}_h Q^*_{h+1}$ for all $h \in [H]$.

Function Approximation of Zero-Sum Markov Games In specific, let M denote the MG instance, which is clear from the context. We assume that we have access to a hypothesis class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where a hypothesis function $f = \{f_1, \ldots, f_H\} \in \mathcal{F}$ either represents an action-value function $Q_f = \{Q_{h,f}\}_{h\in[H]}$ in the model-free setting, or the environment model of zero-sum MG $M_f = \{\mathbb{P}_{h,f}, r_{h,f}\}_{h\in[H]}$ in the model-based setting. Similar to the function approximation of the single-agent MDP, under model-free zero-sum MG setting, we denote $\pi_f = (\upsilon_f, \mu_f)$, where (υ_f, μ_f) is a NE policy pair with respect to $Q_f(s, \cdot, \cdot)$. Moreover, given Q_f and π_f, we define state-value function V_f by letting $V_{h,f}(s) = \mathbb{E}_{a\sim\pi_{h,f}(s)}[Q_{h,f}(s, a)]$ in the MDP and $V_{h,f}(s) = \mathbb{E}_{a,b\sim\pi_{h,f}(s)}[Q_{h,f}(s, a, b)]$ in the MG.

D.2 Learning Goal of Zero-Sum Markov Games

For zero-sum MGs, we aim to design online reinforcement learning algorithms for the player P1 (max-player). In other words, we only control P1 and let P2 play arbitrarily. The goal is to design an RL algorithm such that P1's expected total return is close to the value of the game, namely $V^*_1(s_1)$. For any $k \in [K]$, in the k-th episode, players P1 and P2 executes policy pair $\pi^k = (\nu^k, \mu^k)$ and P1's expected total return is given by $V_1^{\pi^k}(s_1)$. The regret of P1 is also given by (2.2).

For MGs with the decoupled setting, since we can only control the player P1, the switching cost is only defined on the action of player P1 υ, i.e.

$$N_{\text{switch}}(K) = \sum_{k=1}^{K} \mathbb{I}\{\upsilon^k \neq \upsilon^{k+1}\}.$$

E Concrete Examples and Theoretical Results of EC Class

Now we provide a large amount of RL problems that are contained in the EC class with a decomposable loss function and provide corresponding theoretical results for them.

E.1 Linear Mixture MDP

Example E.1 (Linear Mixture MDP)**.** The transition kernel of the Linear Mixture MDP (Ayoub et al., 2020) is a linear combination of several basis kernels. In this model, the transition kernel can be represented by $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a, s'), \theta_h \rangle$, where the feature $\phi(s, a, s') \in \mathbb{R}^d$ is known and the weight $\theta_h \in \mathbb{R}^d$ is unknown. We assume $\|\theta_h\|_2 \leq 1$ and $\|\phi(s, a, s')\|_2 \leq 1$. The reward function can be written as $r_h = \langle \psi(s, a), \theta_h \rangle$ with mapping $\psi(s, a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$. The following lemma shows that ℓ_2-type EC class contains the linear mixture MDP as a special case. The proof is provided in I.6.

Lemma E.2 (Linear Mixture MDP $\subset$ ℓ_2-type EC Class)**.** *The linear mixture model belongs to ℓ_2-type EC class. Indeed, if we choose the model-based function approximation $\mathcal{F} = \mathcal{G} = \{\theta_h\}_{h=1}^H$, $\zeta_h = \{s_{h+1}\}$, $\eta_h = \{s_h, a_h\}$, $\kappa = 1$, and*

$$\ell_{h,f'}(\zeta_h, \eta_h, f, g) = \theta_{h,g}^T \left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right] - r_h - V_{h+1,f'}(s_{h+1}),$$

then we have

$$\mathbb{E}_{\zeta_h}\Big[\ell_{h,f'}(\zeta_h, \eta_h, f, g)\Big] = (\theta_{h,g} - \theta^*_h)^T \left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right], \tag{E.1}$$

and the loss function satisfies the dominance, decomposable property with $\mathcal{T}(f) = f^*$ and ℓ_2-type condition. Hence $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ is a ℓ_2-type EC class.

Combining Theorem 4.1, Theorem 5.1 and Lemma E.2, Algorithm 1 provides a $\widetilde{\mathcal{O}}(H\sqrt{d\beta \log K})$ regret and a $\widetilde{\mathcal{O}}(dH \log K)$ switching cost. Also, Algorithm 2 satisfies a $\widetilde{\mathcal{O}}(H\sqrt{d\beta K} \log K + dHK(\log K)^2/B)$ regret upper bound, where B is the number of batches.

E.2 Kernelized Nonlinear Regulator

Kernelized Nonlinear Regulator (Kakade et al., 2020) (KNR) models a nonlinear control system as an unknown function in a RKHS. When we consider the finite dimension RKHS, the model represents the dynamic as $s_{h+1} = U_h^* \phi(s_h, a_h) + \varepsilon_{h+1}$, where $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\phi}$ is a mapping from a state-action pair to a feature with dimension d_ϕ and $\|\phi(\cdot,\cdot)\|_2 \le 1$. $U_h^* \in \mathbb{R}^{d_s} \times \mathbb{R}^{d_\phi}$ is a linear mapping such that $\|U_h^*\|_2 \le R$ and $\varepsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$ is a normal distribution noise. Following (Chen et al., 2022), we choose $\mathcal{F}_h = \mathcal{G}_h = \{U \in \mathbb{R}^{d_s} \times \mathbb{R}^{d_\phi} : \|\mathcal{U}\|_2 \le R\}$ and $\ell_{h,f'}(\zeta_h, \eta_h, f, g) = U_{h,g}\phi(s_h, a_h) - s_{h+1}$[1] is a DLF and satisfies the ℓ_2-type condition. Denote $\mathbb{E}_{\zeta_h}[\ell_{h,f'}(\zeta_h, \eta_h, f, g)] = (U_{h,g} - U^*)\phi(s_h, a_h)$. The following lemma shows that ℓ_2-type EC class contains the KNR.

Lemma E.3 (KNR $\subset$ ℓ_2-type EC class)**.** *If we choose $\mathcal{F}, \mathcal{G}, \ell$ as defined above, $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ is a ℓ_2-type EC class, where $d = \widetilde{\mathcal{O}}(d_\phi)$ is a parameter and $\kappa = 2H/\sigma$. Indeed, fix a parameter $\beta \ge R^2$, the KNR model belongs to ℓ_2-type EC class by:*

$$\sum_{i=1}^{k-1} \|(U_{h,f^k} - U^*)\phi(s_h^i, a_h^i)\|_2^2 \le \beta, \ \ \forall\, k \in [K], \tag{E.2}$$

$$\Rightarrow \sum_{i=1}^{k} \|(U_{h,f^i} - U^*)\phi(s_h^i, a_h^i)\|_2^2 \le \mathcal{O}(d\beta \log K), \ \ \forall\, k \in [K]. \tag{E.3}$$

The proof of decomposable property and the dominance property with $\kappa = 2H/\sigma$ are provided in Proposition 11 of (Chen et al., 2022). If we choose $X_h(f^k) = (U_{h,f^k} - U^*)$ and $W_{h,f^k}(s_h, a_h) = \phi(s_h, a_h)$, we can apply the similar argument in Section I.6 to prove the ℓ_2-type condition Eq.(E.3)*.*

The detailed proof is provided in I.8.

Example E.4 (Decoupled Zero-Sum Markov Games with Low Minimax BE Dimension)**.** The ℓ_2-type EC class can be extended to the multi-agent setting. We consider the zero-sum MGs with the decoupled setting (Huang et al., 2021), which means that the agent can only control the max-player P1, while an adversary can control the min-player P2. In this case, we let $\eta_h = \{s_h, a_h, b_h\}$ and $\zeta_h = \{s_{h+1}\}$, choose the loss function as

$$\begin{aligned} \ell_{h,f'}(\zeta_h, \eta_h, f, g) &= Q_{h,g}(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) - V_{h+1,f}(s_{h+1}) \\ &= Q_{h,g}(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) - \max_{\upsilon} \min_{\mu} Q_{h+1,f}(s_{h+1}, \upsilon, \mu) \\ &\triangleq \mathcal{E}(f, s_h, a_h, b_h). \end{aligned} \tag{E.4}$$

The Bellman operator for MGs is defined as

$$\mathcal{T}_h' f(s_h, a_h, b_h) = r_h(s_h, a_h, b_h) + \mathbb{E}_{s_{h+1}} \max_{\upsilon} \min_{\mu} f(s_{h+1}, \upsilon, \mu),$$

where we denote $f(s, \upsilon, \mu) = \mathbb{E}_{a \sim \upsilon, b \sim \mu}[f(s, a, b)]$. We prove the decoupled MG belongs to the ℓ_2-type EC class.

Lemma E.5 (Decoupled Zero-Sum MGs $\subseteq$ ℓ_2-type EC Class)**.** *If we choose $\mathcal{F}, \mathcal{G}$ such that $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{G}$, then for any two-player zero-sum MG M, $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ is a ℓ_2-type EC class, where $\eta_h = \{s_h, a_h, b_h\}$, $\zeta_h = \{s_{h+1}\}$, ℓ is chosen as in E.4, $\kappa = 1$, and the parameter d is the minimax BE dimension $d_{\mathrm{ME}}(\mathcal{F}, 1/\sqrt{T})$. The dominance and decomposable property of the loss function (Eq.* (E.4)*) with the Bellman operator $\mathcal{T}$ for MGs are provided in (Huang et al., 2021). The ℓ_2-eluder condition holds by replacing a_h^i in Lemma 3.3 to (a_h^i, b_h^i): If $\sum_{i=1}^{k-1} [\mathcal{E}(f^k, s_h^i, a_h^i, b_h^i)^2] \le \beta$ holds for any $k \in [K]$ and $\beta \ge R^2$, then for any $k \in [K]$ we have*

$$\sum_{i=1}^{k} \left[\mathcal{E}(f^i, s_h^i, a_h^i, b_h^i)^2\right] \le \mathcal{O}(d\beta \log K). \tag{E.5}$$

The detailed proof is provided in I.7. Similarly, the previous theorems and Lemma E.5 give a $\widetilde{\mathcal{O}}(H\sqrt{d\beta \log K})$ regret and $\widetilde{\mathcal{O}}(dH \log K)$ switching cost for the decoupled zero-sum MG, where d is the minimax BE dimension $d_{\mathrm{ME}}(\mathcal{F}, 1/\sqrt{T})$. Also Algorithm 2 gives a $\widetilde{\mathcal{O}}(H\sqrt{d\beta K} \log K + dHK(\log K)^2/B)$ regret. We mainly consider the decoupled setting because it can be naturally contained in our ℓ_2-type EC class.

[1]Note that the loss function can be arbitrary large because of normal distribution noise ε, so it does not satisfy the bounded requirement. However, we can regard it as a bounded loss function because it can be upper bounded by $\widetilde{\mathcal{O}}(\sigma)$ with high probability (Chen et al., 2022).

E.3 Generalized Linear Bellman Complete

Generalized Linear Bellman Complete is introduced in (Du et al., 2021; Chen et al., 2022), which consists of a link function $\sigma : \mathbb{R} \to \mathbb{R}^+$ with $\sigma'(x) \in [L, U]$ for $0 < L < U$, and a hypothesis class $\mathcal{F} = \{\mathcal{F}_h = \sigma(\theta_h^T \phi(s,a)) : \theta_h \in \mathcal{H}_h\}$ with $\|\theta_h\|_2 \leq 1$. Also, for any $f \in \mathcal{F}$, the Bellman complete condition holds:

$$r(s,a) + \mathbb{E}_{s'}\left[\max_{a' \in \mathcal{A}} \sigma(\theta_{h+1,f}^T \phi(s', a'))\right] \in \mathcal{F}_h.$$

Hence we know there is a mapping $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ such that

$$r(s,a) + \mathbb{E}_{s'}\left[\max_{a' \in \mathcal{A}} \sigma(\theta_{h+1,f}^T \phi(s', a'))\right] = \sigma(\mathcal{T}(\theta_{h+1,f})^T \phi(s,a)).$$

If we let

$$\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) = \sigma(\theta_{h,g}^T \phi(s_h, a_h)) - r_h(s_h, a_h) - \max_{a' \in \mathcal{A}} \theta_{h+1,f}^T \phi(s_{h+1}, a'),$$

then the expectation can be written as

$$\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g)\Big] = \sigma(\theta_{h,g}^T \phi(s_h, a_h)) - \sigma(\mathcal{T}(\theta_{h+1,f})^T \phi(s_h, a_h)).$$

Thus we have

$$\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g)\Big]^2 \in [L^2(\theta_{h,g}^T - \theta_{h+1,f}^T)\phi(s_h, a_h))^2, U^2(\theta_{h,g}^T - \theta_{h+1,f}^T)\phi(s_h, a_h))^2].$$

Now following the similar analyses in Section I.6 with $X_h(f^k) = \theta_{h,f}^T - \theta_{h+1,f}^T, W_{h,f^k}(s_h, a_h) = \phi(s_h, a_h)$, we can show that it is a DLF and satisfies the ℓ_2-type condition and dominance property. The constant L and U will only influence the final ℓ_2-type condition by a constant and could be ignored in the $\mathcal{O}(\cdot)$ notation.

E.4 Linear Q^*/V^*

The Linear Q^*/V^* model is proposed in (Du et al., 2021). In this model, the optimal Q-value and V-value functions have a linear structure: There are two known features $\phi(s,a)$ and $\psi(s')$ with unknown parameters ω^*, θ^* such that

$$Q^*(s,a) = \langle \phi(s,a), \omega_h^* \rangle, \ V^*(s) = \langle \psi(s), \theta_h^* \rangle.$$

Denote our hypothesis class as $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where $\mathcal{F}_h$ is defined as

$$\{f = (\omega, \theta) : \max_{a \in \mathcal{A}} \omega^T \phi(s,a) = \theta^T \psi(s), \ \forall \, s \in \mathcal{S}\}.$$

Then we denote the loss function as

$$\begin{aligned} &\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) \\ &\quad = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) = \omega_{h,g}^T \phi(s_h, a_h) - r_h - \theta_{h+1,f}^T \psi(s_{h+1}), \end{aligned}$$

and we can calculate the expectation by

$$\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g)\Big] = (\omega_{h,g} - \omega^*, \theta_{h+1,f} - \theta^*)^T \mathbb{E}_{s_{h+1}}\Big[\phi(s_h, a_h), \psi(s_{h+1})\Big].$$

Note that the expectation has a bilinear structure that is similar to the linear mixture MDP (E.1), then if we choose $X_h(f^k) = (\omega_{h,f^k} - \omega^*, \theta_{h+1,f^k} - \theta^*)$ and $W_{h,f^k}(s_h, a_h) = \mathbb{E}_{s_{h+1}}[\phi(s_h, a_h), \psi(s_{h+1})]$, we can apply an argument similar to Section I.6 to prove the ℓ_2-type condition, dominance property and decomposable property.

E.5 Linear Quadratic Regulator

In the Linear Quadratic Regulator (LQR) model (Bradtke, 1992), we consider d_s-dimensional state space $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and d_a-dimensional action space $\mathcal{A} \subseteq \mathbb{R}^{d_a}$, then a LQR model consists of unknown matrix $A \in \mathbb{R}^{d_s \times d_s}, B \in \mathbb{R}^{d_s \times d_a}$ and $P \in \mathbb{R}^{d_s \times d_s}$ such that

$$s_{h+1} = As_h + Ba_h + \varepsilon_h, \;\; r_h(s_h, a_h) = s_h^T Q s_h + a_h^T a_h + \varepsilon_h',$$

where $\varepsilon_h, \varepsilon_h'$ are zero-centered random noises with $\mathbb{E}[s_h s_h^T] = \Sigma$ and $\mathbb{E}[(\varepsilon_h')^2] = \sigma^2$.

The LQR model has been extensively analyzed (Du et al., 2021; Chen et al., 2022). By Lemma A.3 in (Du et al., 2021), the hypothesis class are $\mathcal{F} = \{(C_h, \Lambda_h, O_h) : C_h \in \mathbb{R}^{d_a \times d_s}, \Lambda_h \in \mathbb{R}^{d_s \times d_s}, O_h \in \mathbb{R}\}$, and

$$\pi_f(s_h) = C_{h,f}(s_h), \;\; V_{h,f}(s_h) = s_h^T \Lambda_{h,f} s_h + O_{h,f}.$$

Let

$$\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}),$$

then the expectation

$$\begin{aligned}
&\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g)\Big] \\
&\quad = \Big\langle \text{vec}(\Lambda_{h,f} - Q - C_{h,f}^T C_{h,f} - (A + BC_{h,f})^T \Lambda_{h+1,f}(A + BC_{h,f})), \\
&\qquad O_{h,f} - O_{h+1,f} - \text{tr}(\Lambda_{h+1,f}\Sigma))^T \text{vec}(s_h s_h^T, 1)\Big\rangle
\end{aligned}$$

has a bilinear structure that is similar to the linear mixture MDP (E.1), then if we choose

$$X_h(f^k) = \text{vec}(\Lambda_{h,f} - Q - C_{h,f}^T C_{h,f} - (A + BC_{h,f})^T \Lambda_{h+1,f}(A + BC_{h,f})), O_{h,f} - O_{h+1,f} - \text{tr}(\Lambda_{h+1,f}\Sigma))$$

and

$$W_{h,f^k}(s_h, a_h) = \text{vec}(s_h s_h^T, 1)$$

in the analyses of Section I.6, we can apply an argument similar to Section I.6 to prove the ℓ_2-type condition, dominance property and decomposable property.

F ℓ_1-type EC Class

F.1 Definition of ℓ_1-type EC Class

In recent years, the ℓ_1-eluder argument has been proposed in (Liu et al., 2022a) for the sample-efficient algorithm of POMDP, and (Liu et al., 2022b) generalize it to the more general classes. Similar to ℓ_2-type EC class, we provide the definition of the ℓ_1-type EC class based on (Liu et al., 2022b). The ℓ_1-type class has two assumptions, which are similar to the ℓ_2-type EC class. To provide a consistent treatment of ℓ_2-type EC class, we let $\{\zeta_h, \eta_h\}_{h\in[H]}$ be subsets of the trajectory. In particular, we let $\eta_h = \{\mathcal{T}_H\}$ and $\zeta_h = \emptyset$, and only consider the single-agent MDP in the ℓ_1-type EC class.

Definition F.1. Given an MDP or POMDP instance (Example F.2) M, let $\mathcal{F}$ and $\mathcal{G}$ be two hypothesis function classes satisfying the realizability Assumption 2.2 with $\mathcal{F} \subseteq \mathcal{G}$. For any $h \in [H]$ and $f' \in \mathcal{F}$, let $\ell_{h,f'}(\zeta_h, \eta_h, f, g)$ be a vector-valued loss function at step h, where ζ_h, η_h are subsets of trajectory that defined above. For parameters d and κ, we say that $(M, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ is a ℓ_1-type EC class if the following two conditions hold for any β and $h \in [H]$:

(i). (ℓ_1-type Condition) For any K hypotheses $f^1, \cdots, f^K \in \mathcal{F}$, if

$$\sum_{i=1}^{k-1} \mathbb{E}_{\eta_h \sim \pi^i, \zeta_h}\Big[\ell_{h,f^i}(\zeta_h, \eta_h, f^k, f^k)\Big] \leq \sqrt{\beta k} \tag{F.1}$$

holds for any $k \in [K]$, then for any $k \in [K]$, we have

$$\sum_{i=1}^{k} \mathbb{E}_{\eta_h \sim \pi^i, \zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^i, f^i) \right] \leq \widetilde{\mathcal{O}}(\text{poly}\log(k)(\sqrt{d\beta k} + d \cdot \text{poly}(H))). \tag{F.2}$$

When we choose $\beta \geq 1$, the right side of Eq.(F.2) can be simplified as $\widetilde{\mathcal{O}}\left(\sqrt{d\beta k} \cdot \text{poly}\log(k)\right)$.

(ii). (κ-Dominance) For any fixed $k \in [K]$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{k} (V_{1,f^i}(s_1) - V^{\pi_i}(s_1)) \leq \kappa \cdot \left(\sum_{h=1}^{H} \sum_{i=1}^{k} \mathbb{E}_{\eta_h \sim \pi^i, \zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^i, f^i) \right] \right). \tag{F.3}$$

Moreover, in this work we only consider a particular loss function

$$\ell_{h,f'}(\zeta_h, \eta_h, f, g) = \ell_{h,f'}(\tau_H, f, g) = |\mathbb{P}_f(\tau_H)/\mathbb{P}_{f^*}(\tau_H) - 1| \tag{F.4}$$

in the ℓ_1-type EC class, where f^* is the true model in realizability Assumption 2.2, and

$$\mathbb{P}_f(\tau_H) = \prod_{h=1}^{H} \mathbb{P}_f(s_h \mid \tau_{h-1})$$

is the product of transition probability in τ_H under the model f. Then

$$\begin{aligned}
&\mathbb{E}_{\eta_h \sim \pi, \zeta_h}[\ell_{h,f'}(\zeta_h, \eta_h, f, g)] \\
&\quad = \mathbb{E}_{\eta_h \sim \pi, \zeta_h} \left(\frac{\prod_{h=1}^{H} \mathbb{P}_f(s_h \mid \tau_{h-1})}{\prod_{h=1}^{H} \mathbb{P}_{f^*}(s_h \mid \tau_{h-1})} - 1 \right) \\
&\quad = \mathbb{E}_{\tau_H \sim \pi} \left[\frac{\prod_{h=1}^{H} (\mathbb{P}_f(s_h \mid \tau_{h-1}) \pi(a_h \mid s_h, \tau_{h-1}))}{\prod_{h=1}^{H} (\mathbb{P}_{f^*}(s_h \mid \tau_{h-1}) \pi(a_h \mid s_h, \tau_{h-1}))} - 1 \right] \\
&\quad = \sum_{\tau_H} \left[\prod_{h=1}^{H} (\mathbb{P}_f(s_h \mid \tau_{h-1}) \pi(a_h \mid s_h, \tau_{h-1})) - \prod_{h=1}^{H} (\mathbb{P}_{f^*}(s_h \mid \tau_{h-1}) \pi(a_h \mid s_h, \tau_{h-1})) \right] s \\
&\quad = d_{\text{TV}}(\mathbb{P}_f^\pi, \mathbb{P}_{f^*}^\pi),
\end{aligned}$$

which is the total variation difference between the trajectory distribution under model f and the true model f^* with policy π. By this particular selection of loss function, the κ-Dominance property is satisfied by $\kappa = 1/H$ and the following inequality:

$$\sum_{i=1}^{k} (V_{1,f^i}(s_1) - V^{\pi_i}(s_1)) \leq \sum_{i=1}^{k} d_{\text{TV}}(\mathbb{P}_{f^i}^{\pi_i}, \mathbb{P}_{f^*}^{\pi_i}).$$

Compared to the ℓ_2-type condition, the primary difference is that the precondition of ℓ_1-type condition (Eq.(F.1)) requires the sum of ℓ_1 norm of the loss function can be controlled by $\mathcal{O}(\sqrt{k})$, while the precondition of ℓ_2-type condition (Eq.(3.3)) requires the square sum of the loss function is controlled by $\mathcal{O}(\beta)$. Second, the selection of ζ_h and η_h are different to the ℓ_2-type EC class, and the left side of Eq.(F.2) contains an extra expectation on $\eta_h = \tau_H \sim \pi^i$. Moreover, since we consider a particular scalar loss function $\ell_{h,f'}(\zeta_h, \eta_h, f, g) = |\mathbb{P}_f(\tau_H)/\mathbb{P}_{f^*}(\tau_H) - 1|$, we do not use the norm on the loss function like ℓ_2-type condition. With this selection of loss function, the ℓ_1-type Condition Eq.(F.2) is similar to the generalized eluder-type condition (Condition 3.1) in (Liu et al., 2022b). We provide two examples in the ℓ_1-type EC class, which are also introduced in the previous works (Liu et al., 2022a;b).

Example F.2 (Undercomplete POMDP (Liu et al., 2022a))**.** A partially observed Markov decision process (POMDP) is represented by a tuple

$$(\mathcal{S}, \mathscr{O}, \mathcal{A}, H, s_1, \mathbb{T} = \{\mathbb{T}_{h,a}\}_{(h,a) \in [H] \times \mathcal{A}}, \mathbb{O} = \{\mathbb{O}_h\}_{h \in [H]}, r = \{r_h\}_{h \in [H]}),$$

where $\mathbb{T}_{h,a} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ represents the transition matrix for latent state of the action a at step h, $\mathbb{O}_h : \mathcal{S} \times \mathscr{O} \mapsto \mathbb{R}$ denotes the probability of generating the observation $o \in \mathscr{O}$ conditioning on the latent state $s \in \mathcal{S}$. $r_h : \mathscr{O} \mapsto \mathbb{R}^+$ is the reward function at step h with observation. We assume $\sum_{h=1}^{H} r_h(s_h, a_h) \in [0, 1]$ for all possible sequences $\{s_h, a_h\}_{h\in[H]}$. During the interactive process, at each step, the agent can only receive the observation and reward without information about the latent state. In POMDP, we consider the general policy $\pi = \{\pi_h\}_{h\in[H]}$, where $\pi_h : \tau_{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}}$, which can be history-dependent. At step h, the agent can only see her observations o_h with probability $\mathbb{O}_h(s_h, o_h)$, take her action a_h with policy $\pi(\tau_{h-1} \times s_h)$, and receive the reward $r_h(s_h, a_h)$. Then the agent arrives to the next state s' with probability $\mathbb{T}_{h,a_h}(\cdot \mid s_h)$. For POMDP, the transition kernel $\mathbb{P}_f$ consists of $\{\mathbb{T}_f, \mathbb{O}_f\}$, and the model is represented by $M_f = \{\mathbb{T}_f, \mathbb{O}_f, r_f\}$.

Undercomplete POMDP (Liu et al., 2022a) is a special case of POMDP such that $S = |\mathcal{S}| \le |\mathscr{O}|$ and there exists a constant $\alpha > 0$ with $\min_h \sigma_S(\mathbb{O}_h) \ge \alpha$. This assumption implies that the observation contains enough information to distinguish two states. In this paper, we only consider undercomplete POMDP because only in this setting we can have a sublinear regret result.[2] The undercomplete POMDP with the model classes $\mathcal{F}$ and $\min_h \sigma_S(\mathbb{O}_h) \ge \alpha$ belongs to the ℓ_1-type EC class by

$$\sum_{i=1}^{k-1} d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i}) \le \sqrt{\beta k},\ \forall\, k \in [K],$$
$$\Rightarrow \sum_{i=1}^{k} d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i}) \le \widetilde{\mathcal{O}}(\mathrm{poly}\log(k)(\sqrt{d\beta k} + \sqrt{d})),\ \ \forall\, k \in [K], \tag{F.5}$$

where $d = S^4A^2O^2H^6 \cdot \alpha^{-4}$. The proof is provided at step E.1, step E.2 and E.3 of Theorem 24 in (Liu et al., 2022a).

Example F.3 (Q-type SAIL condition (Liu et al., 2022b)**).** Q-type SAIL condition provided in (Liu et al., 2022b) is satisfied by many RL models such as witness condition, factor MDPs and sparse linear bandits. A model class $\mathcal{F}$ satisfies the Q-type (d, c, B)-SAIL condition if there exists two sets of mapping functions $\{p_{h,i} : \mathcal{F} \to \mathbb{R}^{d_{\mathcal{F}}}\}_{(h,i)\times[H]\times[m]}$ and $\{q_{h,i} : \mathcal{F} \to \mathbb{R}^{d_{\mathcal{F}}}\}_{(h,i)\times[H]\times[n]}$ such that for $f, f' \in \mathcal{F}$ with optimal policy $\pi^f, \pi^{f'}$, we have

$$d_{\mathrm{TV}}(\mathbb{P}_{f'}^{\pi^f}, \mathbb{P}_{f^*}^{\pi^f}) \ge c^{-1}\sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j=1}^{n} |\langle p_{h,i}(f), q_{h,i}(f')\rangle|,$$
$$d_{\mathrm{TV}}(\mathbb{P}_{f}^{\pi^f}, \mathbb{P}_{f^*}^{\pi^f}) \le \sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j=1}^{n} |\langle p_{h,i}(f), q_{h,i}(f)\rangle|,$$
$$\left(\sum_{i=1}^{m} \|p_{h,i}(f)\|_1\right) \cdot \left(\sum_{j=1}^{n} \|q_{h,i}(f')\|_\infty\right) \le B.$$

From Lemma 6.3 in (Liu et al., 2022b), the Q-type SAIL condition also satisfies the Eq. (F.5) if we choose $d = \mathrm{poly}(H) \cdot \max\{c^2, B^2\} \cdot d_{\mathcal{F}}^2$.

F.2 Rare Policy Switch Algorithm for ℓ_1-type EC Class

In this subsection, we provide an algorithm for the ℓ_1-type EC class with the particular loss function Eq.(F.4). We only consider the MDP model for ℓ_1-type EC class, and leave the zero-sum MG or multi-player general-sum MG as the future work. Our algorithm achieves a logarithmic switching cost while still maintaining a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. The pseudo-code of the algorithm is in Algorithm 3.

In Algorithm 3, the discrepancy function L is selected as the negative log-likelihood function

$$L^{1:k-1}(D^{1:k-1}, f) = -\sum_{i=1}^{k-1} \log \mathbb{P}_f(\tau_H^i),$$

then the Line 7 in Algorithm 3 is equivalent to the OMLE algorithm (Liu et al., 2022b). Unlike OMLE, we change the policy only when the TV distance between f^k and estimated optimal policy $g^k = \inf_{g\in\mathcal{F}} L^{1:k}(D_{1:k}, g)$ is relatively large.

[2] In the previous works studying sample-efficient POMDP, they only provide sample complexity or "pseudo-regret" (defined in (Liu et al., 2022b), (Zhong et al., 2022)) for overcomplete POMDP.

Intuitively, this distance measures the possible improvement based on the historical data. Only when we can get enough new information from the data, we recompute the confidence set and switch the policy.

Algorithm 3 Modified ℓ_1 ABC-Rare switch

Input $D = \emptyset, \mathscr{B}_1 = \mathcal{F}$, constant c in Lemma G.4.
for $k = 1, 2, \cdots, K$ **do**
  Compute $\pi^k = \pi_{f^k}$, where $f^k = \arg\max_{f \in \mathscr{B}^{k-1}} V_f^{\pi_f}(s_1)$.
  Execute policy π^k to collect τ^k, update $D = D \cup \{\tau_H\}$.
  Calculate $g^k = \inf_{g \in \mathcal{F}} L^{1:k}(D^{1:k}, g)$.
  if $\sum_{i=1}^k d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi_i}, \mathbb{P}_{g^k}^{\pi_i}) \le 5c\sqrt{\beta k}$ **then**
    Update
$$\mathscr{B}^k = \left\{ f \in \mathcal{F} : L^{1:k}(D^{1:k}, f) - L^{1:k}(D^{1:k}, g_h^k) \le \beta \right\}.$$
  else
    $\mathscr{B}^k = \mathscr{B}^{k-1}$.
  end if
end for

Now we state our results for ℓ_1-type EC class under both the rare policy switch problem and the batch learning problem. The proof of them is provided in G.2 and H.3.

Theorem F.4. *Given the hypothesis class $\mathcal{F}$, we choose $\eta_h = \tau_H, \zeta_h = \emptyset$ and the loss function $\ell_{h,f'}(\zeta_h, \eta_h, f, g) = \ell_{h,f'}(\tau_H, f, g) = |\mathbb{P}_f(\tau_H)/\mathbb{P}_{f^*}(\tau_H) - 1|$. Denote $\mathcal{B}_{\mathcal{F}}(\rho)$ as the ρ-bracketing number for hypothesis class $\mathcal{F}$ that defined in §B. By setting $\beta = c\log(TB_{\mathcal{F}}(1/K)/\delta) \ge 1$, in which with probability at least $1-\delta$ the Algorithm 1 will achieve sublinear regret*

$$R(K) = \widetilde{\mathcal{O}}(H\kappa\sqrt{d\beta K} \cdot \mathrm{poly}\log(K))$$

with switch cost

$$N_{switch}(K) = \mathcal{O}\left(\sqrt{d} \cdot \mathrm{poly}\log(K)\right).$$

Theorem F.5. *Under the same condition as F.4, if we choose the position of batches as $[k_j, k_{j+1})$ with $k_j = j \cdot \lfloor K/B \rfloor + 1$, then with probability at least $1-\delta$ we can get the following regret*

$$R(K) = \widetilde{\mathcal{O}}\left(\mathrm{poly}\log(K)\left(\sqrt{d} \cdot \frac{K}{B} + \sqrt{d\beta K}\right)\right).$$

By applying Theorem F.4 and Theorem F.5 to the examples in Section F, we can get a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and a logarithmic switching cost in the rare policy switch problem, and about $\widetilde{\mathcal{O}}(\sqrt{d}K/B + \sqrt{dK})$ regret in the batch learning problem for the examples of ℓ_1-type EC class such as undercomplete POMDP and SAIL condition, where d is the parameter that is specific to the concrete examples.

G Proof of the Rare Policy Switch Problem

G.1 Proof of Theorem 4.1

First, by choosing β the same as Theorem 4.1, we provide the following lemma, which shows that $L_h^{1:k}(D_h^{1:k}, f, \mathcal{T}(f))$ is close to the optimal value $\inf_{g\in\mathcal{G}} L_h^{1:k}(D_h^{1:k}, f, g)$. The proof is provided in I.1.

Lemma G.1. *For any $f \in \mathcal{F}$, let $\ell_{h,f^i}(\zeta_h, \eta_h^i, f, g)$ be a DLF, then with probability at least $1-\delta$, we have*

$$0 \ge \inf_{g\in\mathcal{G}} L_h^{a:b}(D_h^{a:b}, f, g) - L_h^{a:b}(D_h^{a:b}, f, \mathcal{T}(f)) \ge -\beta \tag{G.1}$$

for all $1 \le a \le b \le K$. Moreover, by choosing $f = f^*$ in Eq.(G.1), we can get $f^* \in \mathcal{B}^k$ for all $k \in [K]$.

Now we provide two lemmas to show that $L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k))$ is an estimate of

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k) \right] \right\|_2^2.$$

The proof of them are provided in I.2 and I.3.

Lemma G.2. *If*

$$L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)) \leq C\beta \tag{G.2}$$

for some constant $100 \geq C \geq 1$*, then with probability at least* $1 - 2\delta$*,*

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k) \right] \right\|_2^2 \leq (C+1)\beta. \tag{G.3}$$

Moreover, we have

$$\sum_{i=1}^{k-1} \mathbb{E}_{\eta_h \sim \pi^i} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^k, f^k) \right] \right\|_2^2 \leq (C+1)\beta. \tag{G.4}$$

Also, if all constant $C \geq 2$*, we have*

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k) \right] \right\|_2^2 \leq (2C)\beta, \tag{G.5}$$

and

$$\sum_{i=1}^{k-1} \mathbb{E}_{\eta_h \sim \pi^i} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^k, f^k) \right] \right\|_2^2 \leq (2C)\beta. \tag{G.6}$$

Lemma G.3. *If we have*

$$L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)) \geq C\beta$$

for some constant $100 \geq C \geq 2$*, then with probability at least* $1 - 2\delta$

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k) \right] \right\|_2^2 \geq (C-1)\beta. \tag{G.7}$$

Moreover, we have

$$\sum_{i=1}^{k-1} \mathbb{E}_{\eta_h \sim \pi^i} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^k, f^k) \right] \right\|_2^2 \geq (C-1)\beta. \tag{G.8}$$

Also, if all constant $C \geq 2$*, we have*

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k) \right] \right\|_2^2 \geq (C/2)\beta, \tag{G.9}$$

and

$$\sum_{i=1}^{k-1} \mathbb{E}_{\eta_h \sim \pi^i} \left\| \mathbb{E}_{\zeta_h} \left[\ell_{h,f^i}(\zeta_h, \eta_h, f^k, f^k) \right] \right\|_2^2 \geq (C/2)\beta. \tag{G.10}$$

Combining Lemma G.2 and Lemma G.3, we can claim that the term $L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k))$ for $h \in [H]$ in Algorithm 3 is a good estimate for the expectation of loss function.

Proof of Regret First, we claim that for each episode $k \in [K]$,

$$L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)) \leq 6\beta. \tag{G.11}$$

If the policy changes at episode $k-1$, $L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - \inf_{g\in\mathcal{G}} L_h^{1:k-1}(D_h^{1:k-1}, f^k, g) \leq \beta$ for the construction of confidence set. Combining with Eq. (G.1) we can get Eq. (G.11). If the policy has not been changed and $f^{k-1} = f^k$, $L_h^{1:k-1}(D_h^{1:k-1}, f^{k-1}, f^{k-1}) - \inf_{g\in\mathcal{G}} L_h^{1:k-1}(D_h^{1:k-1}, f^{k-1}, g) \leq 5\beta$. Combining with Eq. (G.1), we can get

$$L_h^{1:k-1}(D_h^{1:k-1}, f^{k-1}, f^{k-1}) - L_h^{1:k-1}(D_h^{1:k-1}, f^{k-1}, \mathcal{T}(f^{k-1})) \leq 6\beta.$$

Then Eq. (G.11) can be derived by the fact $f^{k-1} = f^k$. Now based on Lemma G.2 and Eq. (G.11), we have

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right]\right\|_2^2 \leq 7\beta. \tag{G.12}$$

Now by the ℓ_2-type eluder condition and Cauchy's inequality, we have

$$\sum_{i=1}^{k} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right]\right\|_2 \leq \mathcal{O}(\sqrt{d\beta k} \cdot \log k). \tag{G.13}$$

Also, by the dominance property,

$$\begin{aligned}
\sum_{i=1}^{k}(V_{1,f^*}(s_1) - V^{\pi_i}(s_1)) &\leq \sum_{i=1}^{k}(V_{1,f^i}(s_1) - V^{\pi_i}(s_1)) \\
&\leq \kappa \sum_{h=1}^{H}\sum_{i=1}^{k} \mathbb{E}_{\eta_h}\left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h, f^i, f^i)\right]\right\|_2 \\
&= \kappa \cdot \sum_{h=1}^{H}\left(\sum_{i=1}^{k}\left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right]\right\|_2 + \widetilde{\mathcal{O}}\left(\sqrt{K}\log K\right)\right) \qquad (G.14) \\
&= \widetilde{\mathcal{O}}(\kappa H\sqrt{d\beta K} \cdot \log K),
\end{aligned}$$

where the first inequality is derived from Lemma G.1 and $\mathcal{T}(f^*) = f^*$, which implies $f^* \in \mathscr{B}^k$ for all $k \in [K]$ by $L_h^{1:k}(D_h^{1:k}, f^*, f^*) - \inf_{g\in\mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^*, g) \leq \beta$. Eq. (G.14) is derived from the Azuma-Hoeffding's inequality and the boundness property of the loss function ℓ.

Proof of Switch Cost Fixed a step $h \in [H]$, assume the policy changes at episode $b_1^h, b_2^h, \cdots, b_l^h$ because the in-sample error at step h is larger than the threshold,

$$L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g\in\mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^k, g) \geq 5\beta,$$

where l is the number of the policy switch because the error at step h is larger than the threshold 5β. Then by Lemma G.1, we have

$$L_h^{1:k}(D_h^{1:k}, f^k, f^k) - L_h^{1:k}(D_h^{1:k}, f^k, \mathcal{T}(f^k)) \geq 4\beta \tag{G.15}$$

for all $k = b_i^h, 1 \leq i \leq l$. Define $b_0^h = 0$ for simplicity. Fixed an $1 \leq j \leq l$ and consider the latest time b' that changes the policy before episode b_j^h, we will get $b' \geq b_{j-1}^h$ and

$$\begin{aligned}
L_h^{1:b'}(D_h^{1:b'}, f^{b'+1}, f^{b'+1}) - \inf_{g\in\mathcal{G}} L_h^{1:b'}(D_h^{1:b'}, f^{b'+1}, g) &\leq \beta, \\
L_h^{1:b'}(D_h^{1:b'}, f^{b'+1}, f^{b'+1}) - L_h^{1:b'}(D_h^{1:b'}, f^{b'+1}, \mathcal{T}(f^{b'+1})) &\leq \beta. \qquad (G.16)
\end{aligned}$$

Since at episode $b'+1, \cdots, b_j^h - 1$ the confidence set is not changed, we have $\mathscr{B}^{b'} = \mathscr{B}^{b'+1} = \cdots = \mathscr{B}^{b_j^h - 1}$ and $f^{b'+1} = f^{b'+2} = \cdots = f^{b_j^h}$. Then combining Eq. (G.15) and Eq. (G.16), we can get

$$L_h^{b'+1:b_j^h}(D_h^{b'+1:b_j^h}, f^{b'+1}, f^{b'+1}) - L_h^{b'+1:b_j^h}(D_h^{b'+1:b_j^h}, f^{b'+1}, \mathcal{T}(f^{b'+1})) \geq 3\beta.$$

By Lemma G.3, with probability at least $1-\delta$, $\sum_{i=b'+1}^{b_j^h} \|\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)]\|_2^2 \geq 2\beta$. By $b' \geq b_{j-1}^h$, we can see that

$$\sum_{i=b_{j-1}^h+1}^{b_j^h} \|\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)]\|_2^2 \geq 2\beta.$$

Now sum over all $1 \leq i \leq l$, we can get

$$\sum_{i=1}^{K} \left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right]\right\|_2^2 \geq \sum_{j=1}^{l-1} \sum_{i=b_{j-1}^h+1}^{b_j^h} \left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right]\right\|_2^2 \geq 2(l-1)\beta, \tag{G.17}$$

where l is the number of switches corresponding to step $h \in [H]$.

Now by Eq. (G.12) and ℓ_2-type eluder condition,

$$\sum_{i=1}^{K} \left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right]\right\|_2^2 \leq \mathcal{O}(d\beta \log K).$$

Combining with Eq. (G.17), $l = \mathcal{O}(d \log K)$ and the total switching cost can be bounded by $\mathcal{O}(dH \log K)$.

G.2 Proof of Theorem F.4

Since $L^{1:k-1}(D^{1:k-1}, g) = -\sum_{i=1}^{k-1} \log \mathbb{P}_g(\tau_H^i)$, we first show that $g^k = \arg\inf_g L^{1:k-1}(D^{1:k-1}, g)$ is closed to f^* with respect to TV distance. Since we choose $\beta \geq 1$, we simplify the right side of Eq.(F.2) as $\widetilde{\mathcal{O}}(\sqrt{d\beta k} \cdot (\log k)^2)$.

Lemma G.4. *For all $k \in [K]$, let $g_k = \arg\max_g L^{1:k}(D^{1:k}, g)$, then with probability at least $1-\delta$,*

$$\sum_{i=1}^{k} d_{\mathrm{TV}}(\mathbb{P}_{f^*}^{\pi^i}, \mathbb{P}_{g^k}^{\pi^i}) \leq c\sqrt{\beta k}.$$

Proof. By proposition 14 in (Liu et al., 2022a), there is a constant $c > 0$ such that

$$\sum_{i=1}^{k} d_{\mathrm{TV}}^2(\mathbb{P}_{f^*}^{\pi^i}, \mathbb{P}_{g^k}^{\pi^i}) \leq c^2\beta$$

with probability at least $1-\delta$ for all $k \in [K]$. Then by Cauchy's inequality, we have

$$\sum_{i=1}^{k} d_{\mathrm{TV}}(\mathbb{P}_{f^*}^{\pi^i}, \mathbb{P}_{g^k}^{\pi^i}) \leq c\sqrt{\beta k}$$

for all $k \in [K]$. □

We then prove that

$$\sum_{i=1}^{k-1} d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i}) \leq 6c\sqrt{\beta k}. \tag{G.18}$$

Case 1: If at episode $k-1$ the confidence set are not changed, it implies that

$$\sum_{i=1}^{k-1} d_{\mathrm{TV}}(\mathbb{P}_{f^{k-1}}^{\pi^i}, \mathbb{P}_{g^{k-1}}^{\pi^i}) \leq 5c\sqrt{\beta(k-1)}.$$

Thus since $f^{k-1} = f^k$,

$$\sum_{i=1}^{k-1} d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi^i}, \mathbb{P}_{g^{k-1}}^{\pi^i}) \leq 5c\sqrt{\beta k}.$$

Combining with Lemma G.4, we can get Eq. (G.18)

Case 2: If the confidence set are changed at episode $k-1$, then

$$\sum_{i=1}^{k-1} d_{\mathrm{TV}}(\mathbb{P}_{f^k}^{\pi^i}, \mathbb{P}_{g^{k-1}}^{\pi^i}) \leq c\sqrt{\beta k}.$$

Combining with Lemma G.4, we can get Eq. (G.18).

Now since $\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)] = d_{\mathrm{TV}}(\mathbb{P}_{f^i}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i})$, by ℓ_1-type eluder condition with Eq. (G.18), there is a constant c' such that

$$\sum_{i=1}^{k} d_{\mathrm{TV}}(\mathbb{P}_{f^i}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i}) \leq c'(\sqrt{d\beta k} \cdot \mathrm{poly}\log(k)). \tag{G.19}$$

Now since from Proposition 13 in (Liu et al., 2022a), by choosing $\beta = \mathcal{O}(\log(KN_{\mathcal{F}}(1/K)/\delta))$ with a sufficiently large constant, we know $f^* \in \mathscr{B}^k$. Then

$$\begin{aligned} R(K) &= \sum_{i=1}^{K} (V_{1,f^*}^{\pi^*}(s_1) - V_{1,f^*}^{\pi^i}(s_1)) \\ &\leq \sum_{i=1}^{K} (V_{1,f^i}^{\pi^i}(s_1) - V_{1,f^*}^{\pi^i}(s_1)) \\ &\leq \sum_{i=1}^{K} H \cdot d_{\mathrm{TV}}(\mathbb{P}_{f^i}^{\pi^i}, \mathbb{P}_{f^*}^{\pi^i}) \\ &\leq c'H(\sqrt{d\beta k} \cdot \mathrm{poly}\log(k)), \end{aligned}$$

where the first inequality holds because $f^* \in \mathcal{B}^k$ for all $k \in [K]$, and f^k is the optimal policy within the confidence set $\mathcal{B}^k$.

Switch Cost Now assume the policy changed at time $b_1, \cdots, b_l$ and define $b_0 = 0$, then

$$\sum_{i=1}^{b_j} d_{\mathrm{TV}}(\mathbb{P}_{f^{b_j}}^{\pi_i}, \mathbb{P}_{g^{b_j}}^{\pi_i}) \geq 5c\sqrt{\beta b_j}$$

for all $1 \leq j \leq l$. By the triangle inequality, we can get

$$\sum_{i=1}^{b_j} d_{\mathrm{TV}}(\mathbb{P}_{f^{b_j}}^{\pi_i}, \mathbb{P}_{f^*}^{\pi_i}) \geq \sum_{i=1}^{b_j} (d_{\mathrm{TV}}(\mathbb{P}_{f^{b_j}}^{\pi_i}, \mathbb{P}_{g^{b_j}}^{\pi_i}) - d_{\mathrm{TV}}(\mathbb{P}_{f^*}^{\pi_i}, \mathbb{P}_{g^{b_j}}^{\pi_i})) \geq 4c\sqrt{\beta b_j} \tag{G.20}$$

for all $1 \leq j \leq l$. Now by the construction of confidence set $\mathscr{B}^k$, we have $f^{b_{j-1}+1} = \cdots = f^{b_j}$ and

$$\sum_{i=1}^{b_{j-1}} d_{\mathrm{TV}}(\mathbb{P}_{f^{b_{j-1}+1}}^{\pi_i}, \mathbb{P}_{f^*}^{\pi_i}) \leq c\sqrt{\beta(b_{j-1})}. \tag{G.21}$$

Thus combining with Eq. (G.20) and Eq. (G.21),

$$\sum_{i=b_{j-1}+1}^{b_j} d_{\mathrm{TV}}(\mathbb{P}^{\pi_i}_{f^i}, \mathbb{P}^{\pi_i}_{f^*}) = \sum_{i=b_{j-1}+1}^{b_j} d_{\mathrm{TV}}(\mathbb{P}^{\pi_i}_{f^{b_j}}, \mathbb{P}^{\pi_i}_{f^*}) \geq 3c\sqrt{\beta b_j}, \tag{G.22}$$

and for all $k \in [K]$,

$$\sum_{i=1}^{k} d_{\mathrm{TV}}(\mathbb{P}^{\pi_i}_{f^i}, \mathbb{P}^{\pi_i}_{f^*}) \geq \sum_{b_j \leq k} \left(\sum_{i=b_{j-1}+1}^{b_j} d_{\mathrm{TV}}(\mathbb{P}^{\pi_i}_{f^i}, \mathbb{P}^{\pi_i}_{f^*}) \right) \geq 3c \sum_{b_j \leq k} \sqrt{\beta b_j}.$$

The first inequality is because we divide the time interval $[1, k]$ to some intervals $[b_{j-1}+1, b_j]$ for all $b_j \leq k$, and the second inequality is from Eq. (G.22). Now fixed a $k \in [K]$, by Eq. (G.19), we have

$$3c \sum_{b_j \leq k} \sqrt{\beta b_j} \leq c'(\sqrt{d\beta k} \cdot \mathrm{poly}\log(k)).$$

Denote the number of j such that $b_j \in (k/2, k]$ are s_k, i.e. $s_k = |\{j : b_j \in (k/2, k]\}|$, then

$$\begin{aligned} 3cs_k \cdot \sqrt{\beta k/2} \leq 3c \sum_{b_j \leq k} \sqrt{\beta b_j} \leq c'(\sqrt{d\beta k} \cdot \mathrm{poly}\log(k)), \\ s_k \leq \frac{c'\sqrt{2}}{3c}(\sqrt{d} \cdot \mathrm{poly}\log(K)). \end{aligned} \tag{G.23}$$

Now we divide the interval $[1, K]$ into $(\lceil K/2 \rceil, K], (\lceil K/4 \rceil, \lfloor K/2 \rfloor], \cdots, (\lceil \frac{K}{2^m} \rceil, \lfloor \frac{K}{2^{m-1}} \rfloor], [1]$ with $\lceil K/2^m \rceil = 1$, and $m = \mathcal{O}(\log K)$. Then the number of b_j in each interval is upper bounded by $\mathcal{O}(\sqrt{d}(\log K)^2)$ from Eq. (G.23), because $\frac{c'\sqrt{2}}{3c}$ does not depend on the selection of k. Then the total number of policy switch is upper bounded by $\mathcal{O}(\sqrt{d} \cdot \mathrm{poly}\log(K))$.

H Proof of the Batch Learning Problem

H.1 Proof of Theorem 5.1

Proof. We first fix an $h \in [H]$ in the proof. Since we change our policy at time $k_j = j \cdot \lfloor K/B \rfloor + 1$ for $j \geq 0$, we can know that

$$\begin{aligned} &L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, f^{k_j}) - L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, \mathcal{T}(f^{k_j})) \\ &\leq L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, f^{k_j}) - \inf_{g \in \mathcal{G}} L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, g) \\ &\leq \beta. \end{aligned}$$

We denote

$$c_j := \max_{k \in [k_j : k_{j+1}-1]} \left(L_h^{k_j:k}(D_h^{k_j:k}, f^{k_j}, f^{k_j}) - L_h^{k_j:k}(D_h^{k_j:k}, f^{k_j}, \mathcal{T}(f^{k_j})) \right) / \beta.$$

The parameter c_j represents the maximum fitting error for data of this batch and the model f^{k_j} determined by previous batches. If the error is small, the regret can be easily bounded. Thus we only need to prove that the number of batches with large in-sample error is small. Denote $S = \{j \geq 0 \mid c_j > 5\}$ are all "Bad" batches with relatively large in-sample error, then we prove that $|S| \leq \widetilde{\mathcal{O}}(d(\log K)^2)$.

We will prove the following lemma to upper bound $|S|$.

Lemma H.1. *For a fixed $C \geq 10$, with probability at least $1 - \delta$, we will have*

$$\left| \left\{ j \in S \,\middle|\, \frac{C}{2} \leq c_j \leq C \right\} \right| \leq \widetilde{\mathcal{O}}(d \log K).$$

Proof. Denote $\{j \in S \mid \frac{C}{2} \le c_j \le C\} = \{i_1, \cdots, i_M\}$ with $M = \left|\{j \in S \mid \frac{C}{2} \le c_j \le C\}\right|$ and $i_1 \le i_2 \le \cdots \le i_M$. Then for $m \in [M]$ and $k \in [k_{i_m}, k_{i_m+1} - 1]$, we have

$$L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)) \le (1 + c_{i_m})\beta \le (1 + C)\beta.$$

Then by Lemma G.2, we can get

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right] \right\|_2^2 \le \widetilde{\mathcal{O}}((2 + 2C)\beta)$$

for any $m \in [M]$ and $k \in [k_{i_m}, k_{i_m+1} - 1]$. Then we will have

$$\begin{aligned}\sum_{\substack{1 \le i \le k-1 \\ i \in [k_{i_m}, k_{i_m+1}-1], m \in [M]}} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right] \right\|_2^2 &\le \sum_{1 \le i \le k-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right] \right\|_2^2 \\ &\le \widetilde{\mathcal{O}}(2 + 2C)\beta.\end{aligned}$$

By using the ℓ_2-type eluder condition for all such $i \in [k_{i_m}, k_{i_m+1} - 1]$, we can have

$$\sum_{\substack{1 \le i \le k \\ i \in [k_{i_m}, k_{i_m+1}-1], m \in [M]}} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right] \right\|_2^2 \le \widetilde{\mathcal{O}}(d(2 + 2C)\beta \log K). \tag{H.1}$$

Also, by Lemma G.3 and the fact that $c_{i_m} \ge C/2$, with probability at least $1 - \delta$, for any $m \in [M]$,

$$\begin{aligned}&\sum_{i=k_{i_m}}^{k_{i_m+1}-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^i, f^i)\right] \right\|_2^2 && \text{(H.2)}\\ &= \sum_{i=k_{i_m}}^{k_{i_m+1}-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^{k_{i_m}}, f^{k_{i_m}})\right] \right\|_2^2 && \text{(H.3)}\\ &= \max_{k \in [k_{i_m}, k_{i_m+1}-1]} \sum_{i=k_{i_m}}^{k} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^{k_{i_m}}, f^{k_{i_m}})\right] \right\|_2^2 && \text{(H.4)}\\ &\ge \max_{k \in [k_{i_m}, k_{i_{m+1}}-1]} \Big(L_h^{k_{i_m}:k}(D_h^{k_{i_m}:k}, f^{k_{i_m}}, f^{k_{i_m}}) \\ &\qquad - L_h^{k_{i_m}:k}(D_h^{k_{i_m}:k}, f^{k_{i_m}}, \mathcal{T}(f^{k_{i_m}})) - \beta\Big) && \text{(H.5)}\\ &\ge (C/4)\beta. && \text{(H.6)}\end{aligned}$$

Hence from (H.1), (H.2) and $C \ge 10$, we can see that

$$M = \widetilde{\mathcal{O}}(d \log K).$$

□

Now by Lemma H.1, we can divide $S = \{j \ge 0 \mid c_j > 5\}$ as $S^{(1)}, S^{(2)}, \cdots$ that $S^{(i)} = \{j \ge 0 \mid 5 \cdot 2^{i-1} \le c_j \le 5 \cdot 2^i\}$. Then for each i, $|S^{(i)}| \le \widetilde{\mathcal{O}}(d \log K)$. Since we have a trivial upper bound $c_i \le (K/B)$ for all $0 \le i < B$, we know the number of sets $S^{(i)}$ is at most $\log_2(K/B)$, then $|S| \le \widetilde{\mathcal{O}}(d(\log K)^2)$.

Now, since for any $j \notin S$ and $k \in [k_j, k_{j+1} - 1]$, we have

$$\begin{aligned}&L_h^{1:k-1}(D_h^{1:k-1}, f^k, f^k) - L_h^{1:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)) \\ &\quad = (L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, f^{k_j}) + L_h^{k_j:k-1}(D_h^{1:k-1}, f^k, f^k)) \\ &\qquad - (L_h^{1:k_j-1}(D_h^{1:k_j-1}, f^{k_j}, \mathcal{T}(f^{k_j}) + L_h^{k_j:k-1}(D_h^{1:k-1}, f^k, \mathcal{T}(f^k)))\end{aligned}$$

$$\leq (1+c_j)\beta$$
$$\leq 6\beta.$$

By Lemma G.2, we can get

$$\sum_{\substack{1\leq i\leq k-1\\ i\in[k_j,k_{j+1}-1],j\notin S}} \left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^k,f^k)\right]\right\|_2^2 \leq \sum_{1\leq i\leq k-1}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^k,f^k)\right]\right\|_2^2 \leq 7\beta.$$

By the ℓ_2-type eluder condition, the regret caused by "Good" batches can be upper bounded by

$$\sum_{\substack{1\leq i\leq k\\ i\in[k_j,k_{j+1}-1],j\notin S}} \left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)\right]\right\|_2^2 \leq \widetilde{\mathcal{O}}(d\beta\log K). \tag{H.7}$$

Then by the dominance property and Azuma-Hoeffding's inequality, we have

$$\begin{aligned}
&\sum_{i=1}^{K}(V_{1,f^i}(s_1)-V_1^{\pi_i}(s_1))\\
&\leq \kappa\left(\sum_{h=1}^{H}\sum_{i=1}^{k}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)\right]\right\|_2+\widetilde{\mathcal{O}}(\sqrt{HK}\log K)\right) &\text{(H.8)}\\
&\leq \kappa\sum_{h=1}^{H}\left(\sum_{\substack{1\leq i\leq k\\ i\in[k_j,k_{j+1}-1],j\notin S}}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)\right]\right\|_2\right.\\
&\qquad\left.+\sum_{\substack{1\leq i\leq k\\ i\in[k_j,k_{j+1}-1],j\in S}}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)\right]\right\|_2\right)+\kappa\widetilde{\mathcal{O}}(\sqrt{HK}\log K)\\
&=\kappa\sum_{h=1}^{H}\left(\widetilde{\mathcal{O}}(\sqrt{d\beta K}\log K)+R\cdot\lfloor K/B\rfloor\widetilde{\mathcal{O}}(d(\log K)^2)\right)+\kappa\widetilde{\mathcal{O}}(\sqrt{HK}\log K) &\text{(H.9)}\\
&=\widetilde{\mathcal{O}}\left(\kappa H\sqrt{d\beta K\log K}+\frac{dHK}{B}(\log K)^2\right).
\end{aligned}$$

The inequality Eq.(H.8) is derived by the dominance property and Azuma-Hoeffding's inequality, and the first equality Eq.(H.9) holds by Cauchy's inequality, Eq.(H.7), $|S|\leq\widetilde{\mathcal{O}}(d(\log K)^2)$ and $\mathbb{E}_{\zeta_h}[\ell_{h,f^i}(\zeta_h,\eta_h^i,f^i,f^i)]\|_2\leq R$. □

H.2 Discussion about Adaptive Batch Setting

For the adaptive batch setting, to achieve a $O(\sqrt{K})$ regret, we want to let every step have a small in-sample error. That is, $L_h^{1:k}(D_h^{1:k},f,f)-\inf_{g\in\mathcal{G}}L_h^{1:k}(D_h^{1:k},f,g)\leq\mathcal{O}(\beta)$ for all episode k. Then the regret can be easily bounded by $\mathcal{O}(\sqrt{K})$ using previous analyses in Theorem 4.1.

To guarantee this, we modify the rare policy switch Algorithm 1. Note that the Algorithm 1 guarantees that each step has $\mathcal{O}(\beta)$ in-sample error by the updating rule. However, in the adaptive batch setting, we cannot receive the feedback of the current batch. To solve this problem, we can use a simple double trick: We observe the feedback when the length of the batch doubles, and check whether the in-sample error $L_h^{1:k}(D_h^{1:k},f,f)-\inf_{g\in\mathcal{G}}L_h^{1:k}(D_h^{1:k},f,g)$ is greater than 5β. Whenever we observe this error is greater than 5β, we change our policy and begin to choose a batch with length 1. The entire algorithm is presented in Algorithm 4.

We show that for MDP, with this double trick, we can still maintain

$$L_h^{1:k}(D_h^{1:k},f,f)-\inf_{g\in\mathcal{G}}L_h^{1:k}(D_h^{1:k},f,g)\leq\mathcal{O}(\beta)$$

Algorithm 4 ℓ_2-EC-Adaptive Batch

Input $D_1, D_2, \cdots, D_H = \emptyset, \mathscr{B}_1 = \mathcal{F}, length = 1$.
for $k = 1, 2, \cdots, K$ **do**
  (MDP): Compute $\pi^k = \pi_{f^k}$, where $f^k = \arg\max_{f \in \mathscr{B}^{k-1}} V_f^{\pi_f}(s_1)$.
  Execute policy π^k.
  if $length = 2^i$ for some $i \geq 0$ **then**
    Observe the feedback and update $D_h^{1:k} = \{\zeta_h^{1:k}, \eta_h^{1:k}\}$.
    if $L_h^{1:k}(D_h^{1:k}, f^k, f^k) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f^k, g) \geq 5\beta$ for some $h \in [H]$ **then**
      Update

$$\mathscr{B}^k = \left\{ f \in \mathcal{F} : L_h^{1:k}(D_h^{1:k}, f, f) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f, g) \leq \beta, \forall h \in [H] \right\}.$$

      $length = 1$.
    else
      $\mathscr{B}^k = \mathscr{B}^{k-1}$.
      $length = length + 1$.
    end if
  end if
end for

for each round $k \in [K]$, thus achieve the $\mathcal{O}(\sqrt{K})$ regret.

Lemma H.2. *In Algorithm 4, for any $k \in [K]$, we have $L_h^{1:k}(D_h^{1:k}, f, f) - \inf_{g \in \mathcal{G}} L_h^{1:k}(D_h^{1:k}, f, g) \leq \mathcal{O}(\beta)$.*

The proof is provided in I.4. It is worth noting that in Algorithm 4 we choose not to consider the zero-sum MG in the adaptive batch setting. The primary reason is that the policy $\pi^i = (v^i, \mu^i)$ can change within a batch, which makes the double trick fail to work. Indeed, this nature introduces technical difficulties when proving the Lemma H.2.

Now given that each step has a low in-sample error, we can show Algorithm 4 have $\mathcal{O}(\sqrt{K})$ regret, and the number of batches is at most $\mathcal{O}((\log K)^2)$. The square term arises from the extra division of batches for a double trick.

Theorem H.3. *Under the adaptive batch setting and the same condition as Theorem 4.1, with high probability at least $1 - \delta$ the Algorithm 4 will achieve a sublinear regret*

$$R(K) \leq \widetilde{\mathcal{O}}(\kappa H \sqrt{d\beta K} \cdot poly(\log K)).$$

Moreover, the number of batches is at most $\mathcal{O}(dH \cdot poly(\log K))$.

Proof. By Lemma H.2, at each episode $k \in [K]$, we have a small $\mathcal{O}(\beta)$ error of the previous data. Then applying the proof of Theorem 4.1, the upper bound of regret is $\widetilde{\mathcal{O}}(\kappa H \sqrt{d\beta K} \cdot \text{poly}(\log K))$.

Now consider the number of batches. Since we change the policy only when the error is larger than 5β, the policy will be changed at most $\mathcal{O}(dH \cdot \text{poly}(\log K))$ times by Theorem 4.1. In addition, suppose the agent changes the policy at k_1 and k_2 while keeps the change unchanged at episode $k_1 \leq k < k_2$, the number of batches between k_1 and k_2 is at most $\mathcal{O}(\log K)$. Hence the total number of batches can be upper bounded by $\mathcal{O}(dH \cdot \text{poly}(\log K)) \cdot \log K = \mathcal{O}(dH \cdot \text{poly}(\log K))$. $\square$

H.3 Proof of Theorem F.5

Proof. The proof for the batch learning problem under ℓ_1-EC class is straightforward. We fix a $h \in [H]$ in the proof. Since we change our policy at time $k_j = j \cdot \lfloor K/B \rfloor + 1$ for $j \geq 0$, we can get

$$\sum_{i=1}^{k_j - 1} d_{\mathrm{TV}}(\mathbb{P}_{f^{k_j}}^{\pi^i}, \mathbb{P}_{g^{k-1}}^{\pi^i}) \leq c\sqrt{\beta k_j}.$$

Then by Eq.(G.18) in §G.2, for all $j \geq 0$, we have

$$\sum_{i=1}^{k_j-1} d_{\mathrm{TV}}(\mathbb{P}^{\pi^i}_{f^{k_j}}, \mathbb{P}^{\pi^i}_{f^*}) \leq 6c\sqrt{\beta k_j}. \tag{H.10}$$

By our batch learning algorithm, for $k_j \leq k < k_{j+1}$, π^k is the same policy. Thus, we can transform Eq.(H.10) to

$$\sum_{i=0}^{j-1} d_{\mathrm{TV}}(\mathbb{P}^{\pi^{k_i}}_{f^{k_j}}, \mathbb{P}^{\pi^{k_i}}_{f^*}) \leq 6c\sqrt{\beta k_j} \cdot \frac{B}{K} \leq 12c\sqrt{\frac{\beta B}{K} j}. \tag{H.11}$$

Then by the ℓ_1-type eluder condition, we can get

$$\sum_{i=0}^{j-1} d_{\mathrm{TV}}(\mathbb{P}^{\pi^{k_i}}_{f^{k_i}}, \mathbb{P}^{\pi^{k_i}}_{f^*}) \leq \widetilde{\mathcal{O}}\left(\mathrm{poly}\log(K) \cdot \left(\sqrt{\frac{d\beta B}{K} j} + \sqrt{d}\right)\right). \tag{H.12}$$

Then we have

$$\begin{aligned} R(K) &= \sum_{i=1}^{K} (V^{\pi^*}_{1,f^*}(s_1) - V^{\pi^i}_{1,f^*}(s_1)) \leq \sum_{i=1}^{K} (V^{\pi^i}_{1,f^i}(s_1) - V^{\pi^i}_{1,f^*}(s_1)) \\ &\leq \sum_{i=1}^{K} H \cdot d_{\mathrm{TV}}(\mathbb{P}^{\pi^i}_{f^i}, \mathbb{P}^{\pi^i}_{f^*}) \leq H \cdot \left\lfloor \frac{K}{B} \right\rfloor \cdot \sum_{i=0}^{\lceil K/B \rceil} d_{\mathrm{TV}}(\mathbb{P}^{\pi^{k_i}}_{f^{k_i}}, \mathbb{P}^{\pi^{k_i}}_{f^*}), \end{aligned}$$

where the first inequality is because $f^* \in \mathcal{B}^k$ for all $k \in [K]$ and f^i is the optimal policy within the confidence set $\mathcal{B}^i$. The second inequality holds because of the definition of TV distance, and the last inequality holds because the f^j and π^j are the same within the same batch $j \in [k_i, k_{i+1})$.

Then by the Eq.(H.12), we can get

$$\begin{aligned} R(K) &\leq H \cdot \left\lfloor \frac{K}{B} \right\rfloor \cdot \sum_{i=0}^{\lceil K/B \rceil} d_{\mathrm{TV}}(\mathbb{P}^{\pi^{k_i}}_{f^{k_i}}, \mathbb{P}^{\pi^{k_i}}_{f^*}) \\ &\leq \widetilde{\mathcal{O}}\left(H \cdot \left\lfloor \frac{K}{B} \right\rfloor \cdot \mathrm{poly}\log(K)\left(\sqrt{\frac{d\beta B}{K} B} + \sqrt{d}\right)\right) \\ &= \widetilde{\mathcal{O}}\left(H \cdot \left\lfloor \frac{K}{B} \right\rfloor \cdot \mathrm{poly}\log(K)\left(B\sqrt{\frac{d\beta}{K}} + \sqrt{d}\right)\right) \\ &\leq \widetilde{\mathcal{O}}\left(H \cdot \mathrm{poly}\log(K)\left(\sqrt{d} \cdot \frac{K}{B} + \sqrt{d\beta K}\right)\right). \end{aligned}$$

We complete the proof of Theorem F.5. □

I Proof of Lemmas

I.1 Proof of Lemma G.1

Proof. First note that we choose $\zeta_h = s_{h+1}$, $\eta_h = \{s_h, a_h\}$ for MDP and $\eta_h = \{s_h, a_h, b_h\}$ for zero-sum Markov Games. Define the auxillary variable $X_{i,f'}(h,f,g) = \|\ell_{h,f'}(s^i_{h+1}, \eta^i_h, f, g)\|_2^2 - \|\ell_{h,f'}(s^i_{h+1}, \eta^i_h, f, \mathcal{T}(f))\|_2^2$, then we know $|X_{i,f'}(h,f,g)| \leq 2R^2$ for all $1 \leq i \leq k$, where $R = \sup \|\ell_{h,f'}(\zeta_h, \eta_h, f, g)\|_2$. Now we can have

$$\begin{aligned} &\mathbb{E}_{s_{h+1}}\Big[X_{i,f'}(h,f,g)\Big] \\ &= \mathbb{E}_{s_{h+1}}\Big[\langle \ell_{h,f'}(s_{h+1}, \eta^i_h, f, g) - \ell_{h,f'}(s_{h+1}, \eta^i_h, f, \mathcal{T}(f)), \\ &\quad \ell_{h,f'}(s_{h+1}, \eta^i_h, f, g) + \ell_{h,f^i}(\zeta_h, \eta^i_h, f, \mathcal{T}(f))\rangle\Big] \end{aligned}$$

$$= \mathbb{E}_{s_{h+1}}\Big[\langle \mathbb{E}_{s_{h+1}}[\ell_{h,f'}(s_{h+1},\eta_h^i,f,g)], \ell_{h,f'}(s_{h+1},\eta_h^i,f,g)\rangle\Big] \tag{I.1}$$

$$= \Big\|\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f'}(s_{h+1},\eta_h^i,f,g)\Big]\Big\|_2^2. \tag{I.2}$$

The Eq.(I.1) holds from the decomposable property of ℓ. Then we have

$$\begin{aligned}
&\mathbb{E}_{s_{h+1}}\Big[(X_{i,f'}(h,f,g))^2\Big] \\
&= \mathbb{E}_{s_{h+1}}\Big[\|\ell_{h,f'}(s_{h+1},\eta_h^i,f,g) - \ell_{h,f^i}(\zeta_h,\eta_h^i,f,\mathcal{T}(f))\|_2^2 \\
&\quad \cdot \|\ell_{h,f'}(s_{h+1},\eta_h^i,f,g) + \ell_{h,f'}(s_{h+1},\eta_h^i,f,\mathcal{T}(f))\|_2^2\Big] \\
&\le 4R^2\mathbb{E}_{s_{h+1}}\Big[\|\ell_{h,f'}(s_{h+1},\eta_h^i,f,g) - \ell_{h,f'}(s_{h+1},\eta_h^i,f,\mathcal{T}(f))\|_2^2\Big] && \text{(I.3)} \\
&= 4R^2\mathbb{E}_{s_{h+1}}\Big[\|\mathbb{E}_{s_{h+1}}[\ell_{h,f'}(s_{h+1},\eta_h^i,f,g)]\|_2^2\Big] && \text{(I.4)} \\
&= 4R^2\Big\|\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f^i}(\zeta_h,\eta_h^i,f,g)\Big]\Big\|_2^2 \\
&= 4R^2\mathbb{E}_{s_{h+1}}\Big[X_{i,f'}(h,f,g)\Big], && \text{(I.5)}
\end{aligned}$$

where the inequality (I.3) is because $\|\ell_{h,f'}(s_{h+1},\eta_h^i,f,g)\|_2 \le R$ for any h, (f',f,g) and s_{h+1,η_h^i}. The Eq.(I.4) holds from the decomposable property, and the Eq.(I.5) holds from the Eq.(I.2). Thus by Freedman's inequality (Agarwal et al., 2014; Jin et al., 2021; Chen et al., 2022), with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{i=a}^{b} X_{i,f'}(h,f,g) - \sum_{i=a}^{b}\mathbb{E}_{\zeta_h}[X_{i,f'}(h,f,g)]\right| \\
&\quad \le \mathcal{O}\left(R\sqrt{\log(1/\delta)\sum_{i=a}^{b}\mathbb{E}_{\zeta_h}\Big[X_{i,f'}(h,f,g)\Big]} + 2R^2\log(1/\delta)\right).
\end{aligned}$$

Now we consider a ρ-cover $L_{h,\rho} = (\widetilde{\mathcal{F}}_\rho, \widetilde{\mathcal{F}}_\rho, \widetilde{\mathcal{G}}_\rho)$ for $(\mathcal{F},\mathcal{G})$: For any $f,f' \in \mathcal{F}, g \in \mathcal{G}$, there exists a pair of function $(\widetilde{f}',\widetilde{f},\widetilde{g}) \in (\widetilde{\mathcal{F}}_\rho,\widetilde{\mathcal{F}}_\rho,\widetilde{\mathcal{G}}_\rho)$ such that $\|\ell_{h,\widetilde{f}'}(\cdot,\widetilde{f},\widetilde{g}) - \ell_{h,f'}(\cdot,f,g)\|_\infty \le \rho$. By taking a union bound over $L_{h,\rho}$, $a\in[K], b\in[K], h\in[H]$, we can have

$$\begin{aligned}
&\left|\sum_{i=a}^{b} X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g}) - \sum_{i=a}^{b}\mathbb{E}_{\zeta_h}\Big[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})\Big]\right| \\
&\quad \le \mathcal{O}\left(R\sqrt{\iota\sum_{i=a}^{b}\mathbb{E}_{\zeta_h}\Big[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})\Big]} + 2R^2\iota\right),
\end{aligned}$$

where $\iota = \log(HK^2\mathcal{N}_{\mathcal{L}}(1/K)/\delta)$ and $\mathcal{N}_{\mathcal{L}}(1/K) = \max_h |L_{h,\rho}|$ is the maximum ρ-covering number of $(\mathcal{F},\mathcal{F},\mathcal{G})$ for loss function $\ell_{h,f'}(\cdot,\cdot,f,g)$ for $h\in[H]$.

Since $\mathbb{E}_{\zeta_h}[X_{i,f'}(h,\widetilde{f},\widetilde{g})] \ge 0$ and

$$\begin{aligned}
&-\sum_{i=a}^{b} X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g}) \le \mathcal{O}(R^2\iota), \;\; \forall\,(\widetilde{f},\widetilde{g}) \in L_{h,\rho}, \\
&-\sum_{i=a}^{b} X_{i,f'}(h,f,g) \le \mathcal{O}(R^2\iota + R) \le \beta, \;\; \forall\, f\in\mathcal{F}, g\in\mathcal{G}.
\end{aligned}$$

where $\beta = c(R^2\iota + R)$ for some large enough constant c. □

I.2 Proof of Lemma G.2

Proof. The proof is similar to Lemma G.1. Apply the same covering argument and concentration inequality, for any $(\widetilde{f}', \widetilde{f}, \widetilde{g}) \in L_{h,\rho}$, we have

$$\left| \sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g}) - \sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g})\right] \right| \leq \mathcal{O}\left(R\sqrt{\iota \sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g})\right] + 2R^2\iota} \right), \tag{I.6}$$

where $\iota = \log(HK^2\mathcal{N}_{\mathcal{L}}(1.T)/\delta)$ and $\mathcal{N}_{\mathcal{L}}(\rho) = \max_h |L_{h,\rho}|$ is the maximum ρ-covering number of $(\mathcal{F}, \mathcal{F}, \mathcal{G})$ for loss function $\ell_{h,f'}(\cdot, \cdot, f, g)$ for $h \in [H]$. Now note that

$$\begin{aligned} &\sum_{i=1}^{k-1} X_{i,f^i}(h, f^k, f^k) \\ &= \sum_{i=1}^{k-1} (\|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f^k, f^k)\|_2^2 - \|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f^k, \mathcal{T}(f^k))\|_2^2) \\ &\leq C\beta. \end{aligned}$$

Then there is a pair $(\widetilde{f}', \widetilde{f}, \widetilde{g}) \in L_{h,1/K}$ such that

$$\left| \sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g}) - \sum_{i=1}^{k-1} X_{i,f^i}(h, f^k, f^k) \right| \leq \mathcal{O}(R)$$

and

$$\sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g}) \leq C\beta + \mathcal{O}(R).$$

Combining with Eq. (I.6), when $\beta = c(R^2\iota + R)$ for sufficiently large constant c, if $C \leq 100$ is a small constant, we get

$$\sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g})\right] \leq \left(C + \frac{1}{2}\right)\beta.$$

Since $(\widetilde{f}', \widetilde{f}, \widetilde{g})$ is the $\rho-$approximation of (f^i, f^k, f^k),

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right] \right\|_2^2 = \sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,f^i}(h, f^k, f^k)\right] \leq (C+1)\beta. \tag{I.7}$$

The first equality of Eq. (I.7) is derived from Eq. (I.2).

If $C \geq 100$, similarly we can get

$$\sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h, \widetilde{f}, \widetilde{g})\right] \leq (2C - 1)\,\beta$$

and

$$\sum_{i=1}^{k-1} \left\| \mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right] \right\|_2^2 = \sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}\left[X_{i,f^i}(h, f^k, f^k)\right] \leq (2C)\beta. \tag{I.8}$$

Similar to (Jin et al., 2021), to prove the second inequality Eq.(G.4) we can add the η_h to expectation by replacing the $\mathbb{E}_{s_{h+1}}[X_{i,f'}(h, f, g)]$ to $\mathbb{E}_{s_h, a_h \sim \pi^i}\mathbb{E}_{s_{h+1}}[X_{i,f'}(h, f, g)]$. □

I.3 Proof of Lemma G.3

Proof. The proof is similar to Lemma G.1. Now apply the same argument, for any $(\widetilde{f}', \widetilde{f}, \widetilde{g}) \in L_{h,\rho}$, with probability at least $1-\delta$, we have

$$\left|\sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g}) - \sum_{i=1}^{k-1} \mathbb{E}_{\zeta_h}[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})]\right| \\ \leq \mathcal{O}\left(R\sqrt{\iota\sum_{i=1}^{k-1}\mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})\right]} + 2R^2\iota\right), \tag{I.9}$$

where $\iota = \log(HK^2\mathcal{N}_{\mathcal{L}}(1/K)/\delta)$ and $\mathcal{N}_{\mathcal{L}}(\rho) = \max_h |L_{h,\rho}|$ is the maximum ρ-covering number of $(\mathcal{F}, \mathcal{F}, \mathcal{G})$ for loss function $\ell_{h,f'}(\cdot,\cdot,f,g)$ for $h \in [H]$. Now note that

$$\sum_{i=1}^{k-1} X_{i,f'}(h, f^k, f^k) \\ = \sum_{i=1}^{k-1}(\|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f^k, f^k)\|_2^2 - \|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f^k, \mathcal{T}(f^k))\|_2^2) \\ \geq C\beta.$$

Then there is a pair $(\widetilde{f}', \widetilde{f}, \widetilde{g}) \in L_{h,1/K}$ such that

$$\left|\sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g}) - \sum_{i=1}^{k-1} X_{i,f'}(h, f^k, f^k)\right| \leq \mathcal{O}(R),$$

and

$$\sum_{i=1}^{k-1} X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g}) \geq C\beta - \mathcal{O}(R).$$

Combining with Eq. (I.9), when $\beta = c(R^2\iota + R)$ for sufficiently large constant c, if $C \leq 100$ we get

$$\sum_{i=1}^{k-1}\mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})\right] \geq \left(C - \frac{1}{2}\right)\beta.$$

Since $(\widetilde{f}', \widetilde{f}, \widetilde{g})$ is a $\rho-$approximation of (f^i, f^k, f^k),

$$\sum_{i=1}^{k-1}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right]\right\|_2^2 = \sum_{i=1}^{k-1}\mathbb{E}_{\zeta_h}\left[X_{i,f^i}(h, f^k, f^k)\right] \\ \geq \left(C - \frac{1}{2}\right)\beta - \mathcal{O}(R) \\ \geq (C-1)\beta. \tag{I.10}$$

the first equality of Eq. (I.10) is derived from Eq. (I.2).

Similarly, if $C \geq 100$, we can get

$$\sum_{i=1}^{k-1}\mathbb{E}_{\zeta_h}\left[X_{i,\widetilde{f}'}(h,\widetilde{f},\widetilde{g})\right] \geq (C/2+1)\,\beta$$

and

$$\sum_{i=1}^{k-1}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h, \eta_h^i, f^k, f^k)\right]\right\|_2^2 = \sum_{i=1}^{k-1}\mathbb{E}_{\zeta_h}\left[X_{i,f^i}(h, f^k, f^k)\right]$$

$$\geq (C/2+1)\,\beta - \mathcal{O}(R)$$
$$\geq (C/2)\beta. \tag{I.11}$$

Similar to Lemma G.2, to prove the second inequality Eq.(G.8), we can add the η_h to expectation by replacing the $\mathbb{E}_{s_{h+1}}[X_{i,f'}(h,f,g)]$ to $\mathbb{E}_{\eta_h\sim\pi^i}\mathbb{E}_{s_{h+1}}[X_{i,f'}(h,f,g)]$. □

I.4 Proof of Lemma H.2

Assume that at episode k we have $L_h^{1:k}(D_h^{1:k},f^k,f^k) - \inf_{g\in\mathcal{G}} L_h^{1:k}(D_h^{1:k},f^k,g) \geq 24\beta$, and it belongs to a batch with length 2^i. First by Lemma G.1,

$$L_h^{1:k}(D_h^{1:k},f^k,f^k) - L_h^{1:k}(D_h^{1:k},f^k,\mathcal{T}(f^k)) \geq 23\beta.$$

Suppose $i=0$ and this batch starts at k', thus at episode k' we have

$$\begin{aligned}L_h^{1:k'}(D_h^{1:k'},f^k,f^k) - L_h^{1:k'}(D_h^{1:k},f^k,\mathcal{T}(f^k)) &\leq L_h^{1:k'}(D_h^{1:k'},f^k,f^k) - \inf_{g\in\mathcal{G}} L_h^{1:k'}(D_h^{1:k},f^k,g)\\ &\leq \beta,\end{aligned}$$

where the second inequality is the property of optimistic confidence set. Then by these two inequalities,

$$L_h^{k'+1:k}(D_h^{k'+1:k},f^k,f^k) - L_h^{k'+1:k}(D_h^{k'+1:k},f^k,\mathcal{T}(f^k)) \geq 22\beta.$$

This is impossible because $i=0$ and thus $k=k'+2^i=k'+1$, and $\beta \geq 2R^2 \geq L_h^{k'+1:k}(D_h^{k'+1:k},f^k,f^k) - L_h^{k'+1:k}(D_h^{k'+1:k},f^k,\mathcal{T}(f^k))$, where R is the upper bound of the norm of loss function ℓ.

Now suppose $i\geq 1$, then

$$\begin{aligned}L_h^{1:k'}(D_h^{1:k'},f^k,f^k) - L_h^{1:k'}(D_h^{1:k},f^k,\mathcal{T}(f^k)) &\leq L_h^{1:k'}(D_h^{1:k'},f^k,f^k) - \inf_{g\in\mathcal{G}} L_h^{1:k'}(D_h^{1:k},f^k,g)\\ &\leq 5\beta,\end{aligned}$$

where the second inequality is derived by the updating rule with $i\neq 0$, and the fact that $f^k = f^{k'}$. Then we can also get

$$L_h^{k'+1:k}(D_h^{k'+1:k},f^k,f^k) - L_h^{k'+1:k}(D_h^{k'+1:k},f^k,\mathcal{T}(f^k)) \geq 24\beta - 5\beta = 19\beta.$$

Note that we only consider the MDP problem here, then π^i, f^i are the same for $i\in[k'+1,k]$. By combining this fact and Lemma G.3, we can get

$$\begin{aligned}(k-k'-1)\cdot\mathbb{E}_{\eta_h\sim\pi^k}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^k}(\zeta_h,\eta_h,f^k,f^k)\right]\right\|_2^2 &= \sum_{i=k'+1}^{k}\mathbb{E}_{\eta_h\sim\pi^i}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h,f^k,f^k)\right]\right\|_2^2\\ &\geq 19\beta - \beta\\ &= 18\beta.\end{aligned} \tag{I.12}$$

However, since $k\in[k'+2^i,k'+2^{i+1})$ with $i\geq 1$, we know for $t\in[k'+2^{i-1},k'+2^i)$, $f^t=f^k$ because at episode $k'+2^{i-1}$, the agent does not change the policy by the definition of k'. Thus, denote $k_1 = k'+2^{i-1}+1$, $k_2=k'+2^i$, we have

$$L_h^{k_1:k_2}(D_h^{k_1:k_2},f^{k_2},f^{k_2}) - L_h^{k_1:k_2}(D_h^{k_1:k_2},f^{k_2},\mathcal{T}(f^{k_2})) \leq 5\beta,$$

and

$$\begin{aligned}(k_2-k_1)\cdot\mathbb{E}_{\eta_h\sim\pi^k}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^k}(\zeta_h,\eta_h,f^k,f^k)\right]\right\|_2^2 &= \sum_{i=k_1}^{k_2}\mathbb{E}_{\eta_h\sim\pi^i}\left\|\mathbb{E}_{\zeta_h}\left[\ell_{h,f^i}(\zeta_h,\eta_h,f^{k_2},f^{k_2})\right]\right\|_2^2\\ &\leq 6\beta.\end{aligned} \tag{I.13}$$

Note that $3\cdot(k_2-k_1) = 3\cdot(2^{i-1}-1) > 2^i \geq k-k'-1$, we know Eq.(I.12) and Eq.(I.13) cannot both hold. Hence we have done the proof by contradiction.

I.5 Proof of Lemma 3.3

First, we introduce the definition of D_Δ-type Bellman eluder dimension.

Definition I.1 (Bellman eluder dimension)**.** Given a function class $\mathcal{F}$, the D_Δ Bellman eluder dimension $d(\mathcal{F}, D_\Delta, \varepsilon)$ is the length n of longest sequence $((s_h^1, a_h^1), \cdots, (s_h^n, a_h^n))$ such that for some $\varepsilon' \geq \varepsilon$ and any $j \in [n]$, there exists a $f^j \in \mathcal{F}$ and $\sqrt{\sum_{i=1}^{j-1} \mathcal{E}_h(f^j, s_h^i, a_h^i)^2} \leq \varepsilon$ and $\mathcal{E}_h(f^j, s_h^j, a_h^j) > \varepsilon$. The term $\mathcal{E}_h(f, s, a)$ is the Bellman error $\mathcal{E}_h(f, s, a) = (f_h - \mathcal{T}(f_{h+1}))(s, a)$.

Proof. Now we begin to prove the Lemma 3.3. First, we restate the Proposition 43 in (Jin et al., 2021) with $\Pi = D_\Delta$.

Proposition I.2 (Proposition 43 in (Jin et al., 2021) with $\Pi = D_\Delta$)**.** *Given a function class Φ defined on $\mathcal{X}$, suppose given sequence $\{\phi_i\}_{1 \leq i \leq K} \subset \Phi$ and sequences $\{(s_h^i, a_h^i)\}_{i \leq [K]}$ such that for all $k \in [K]$, $\sum_{i=1}^{k} (\mathbb{E}_{\mu_i}[\phi_k])^2 \leq \beta$, then for all $k \in [K]$,*

$$\sum_{i=1}^{k} \mathbb{I}\{|\mathbb{E}_{\mu_i}[\phi_i]| > \varepsilon\} \leq \left(\frac{\beta}{\varepsilon^2} + 1\right) d_{\mathrm{BE}}(\Phi, D_\Delta, \varepsilon).$$

Now, we first fixed a $h \in [H]$, then choosing $\Pi = D_\Delta$, $\phi_i = \mathcal{E}_h(f_i, s, a)$ and $\mu_i = \mathbb{I}\{(s, a) = (s_h^i, a_h^i)\}$ in proposition 43, based on since $\sum_{i=1}^{k-1} \mathcal{E}_h(f^k, s_h^i, a_h^i) \leq \beta$ for all h, k, we have

$$\sum_{i=1}^{k} \mathbb{I}\{\mathcal{E}_h(f^i, s_h^i, a_h^i)^2 > \varepsilon^2\} \leq \left(\frac{\beta}{\varepsilon^2} + 1\right) d_{\mathrm{BE}}(\mathcal{F}, D_\Delta, \varepsilon).$$

Then by replacing ε^2 to ε,

$$\sum_{i=1}^{k} \mathbb{I}\{\mathcal{E}_h(f^i, s_h^i, a_h^i)^2 > \varepsilon\} \leq \left(\frac{\beta}{\varepsilon} + 1\right) d_{\mathrm{BE}}(\mathcal{F}, D_\Delta, \sqrt{\varepsilon}).$$

Now sort the sequence $\{\mathcal{E}_h(f^1, s_h^1, a_h^1)^2, \mathcal{E}_h(f^2, s_h^2, a_h^2)^2, \cdots, \mathcal{E}_h(f^k, s_h^k, a_h^k)^2\}$ in a decreasing order and denote them by $\{e_1, \cdots, e_k\}$, for any ω we can have

$$\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} e_i \mathbb{I}\{e_i \leq \omega\} + \sum_{i=1}^{k} e_i \mathbb{I}\{e_i > \omega\} \leq k\omega + \sum_{i=1}^{k} e_i \mathbb{I}\{e_i > \omega\}.$$

Assume $e_t > \omega$ and there exists a parameter $\alpha \in (\omega, e_t)$, then

$$t \leq \sum_{i=1}^{k} \mathbb{I}\{e_t > \alpha\} \leq \left(\frac{\beta}{\alpha} + 1\right) d_{\mathrm{BE}}(\mathcal{F}, D_\Delta, \sqrt{\alpha}) \leq \left(\frac{\beta}{\alpha} + 1\right) d_{\mathrm{BE}}(\mathcal{F}, D_\Delta, \sqrt{\omega}).$$

Now denote $d = d_{\mathrm{BE}}(\mathcal{F}, D_\Delta, \sqrt{\omega})$, we can get $\alpha \leq \frac{d\beta}{t-d}$. Since α is arbitrarily chosen, we have $e_t \leq \frac{d\beta}{t-d}$. Also, recall that $e_t \leq R^2$, we can get

$$\sum_{i=1}^{k} e_i \mathbb{I}\{e_i > \omega\} \leq \min\{d, k\} R^2 + \sum_{i=d+1}^{k} \left(\frac{d\beta}{t-d}\right) \leq \min\{d, k\} R^2 + 2d\beta \log K = \mathcal{O}(d\beta \log K),$$

where the last equality derived from the condition $\beta \geq R^2$. By choosing $\omega = 1/K$, the equation Eq. (3.8) holds. □

I.6 Proof of Lemma E.2

Proof. In this subsection, we prove that linear mixture MDP belongs to ℓ_2-type EC class with $\mathcal{T}(f) = f^*$ for any $f \in \mathcal{F}$ and loss function

$$\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g)$$

$$= \theta_{h,g}^T \left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right] - r_h - V_{h+1,f'}(s_{h+1}).$$

It is easy to show that the loss function above is bounded. The expectation of the loss function can be calculated by

$$\mathbb{E}_{s_{h+1}} \Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) \Big] = (\theta_{h,g} - \theta_h^*)^T \left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right].$$

Now we prove the loss function satisfies the dominance, decomposable property and ℓ_2-type condition.

1. Dominance

$$\begin{aligned} &\sum_{i=1}^{k} (V_{1,f^i}(s_1) - V_{\pi^i}(s_1)) \\ &\quad \le \sum_{h=1}^{H} \sum_{i=1}^{k} \mathbb{E}_{s_h, a_h \sim \pi^i} [Q_{h,f^i}(s_h, a_h) - r_h - V_{h+1,f^i}(s_{h+1})] \\ &\quad = \sum_{h=1}^{H} \sum_{i=1}^{k} \mathbb{E}_{s_h, a_h \sim \pi^i} \Big[(\theta_{h,f^i} - \theta_h^*)^T \Big[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f^i}(s') \Big] \Big] \\ &\quad = \sum_{h=1}^{H} \sum_{i=1}^{k} \mathbb{E}_{s_h, a_h \sim \pi^i, s_{h+1}} \Big[\ell_{h,f^i}(s_{h+1}, \{s_h, a_h\}, f^i, f^i) \Big]. \end{aligned}$$

2. Decomposable Property

$$\begin{aligned} &\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) - \mathbb{E}_{s_{h+1}} \Big[\ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, g) \Big] \\ &\quad = (\theta_h^*)^T \left[\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1}(s') \right] \\ &\quad = \ell_{h,f'}(s_{h+1}, \{s_h, a_h\}, f, f^*). \end{aligned}$$

3. ℓ_2-type Eluder Condition First, for any h and $\eta_h = \{s_h, a_h\}$, we have

$$\begin{aligned} &\sum_{i=1}^{k-1} \Big\| \mathbb{E}_{s_{h+1}} \Big[\ell_{h,f^i}(s_{h+1}, \eta_h^i, f^k, f^k) \Big] \Big\|_2^2 \\ &\quad = \sum_{i=1}^{k-1} \left((\theta_{h,f^k} - \theta_h^*)^T \left[\psi(s_h^i, a_h^i) + \sum_{s'} \phi(s_h^i, a_h^i, s') V_{h+1,f^i}(s') \right] \right)^2. \end{aligned}$$

Denote $\psi(s_h^i, a_h^i) + \sum_{s'} \phi(s_h^i, a_h^i, s') V_{h+1,f^k}(s') = W_{h,f^i}(s_h^i, a_h^i)$, $(\theta_{h,f^i} - \theta_h^*) = X_h(f^i)$ and

$$\Sigma_k = I + \sum_{i=1}^{k-1} W_{h,f^i}(s_h^i, a_h^i) W_{h,f^i}(s_h^i, a_h^i)^T,$$

then

$$\begin{aligned} \|\theta_{h,f^k} - \theta_h^*\|_{\Sigma_i}^2 = \|X_h(f^k)\|_{\Sigma_i}^2 &= \sum_{i=1}^{k-1} \Big\| \mathbb{E}_{s_{h+1}} \Big[\ell_{h,f^i}(s_{h+1}, \eta_h^i, f^k, f^k) \Big] \Big\|_2^2 + 4 \\ &\le \beta + 4, \end{aligned}$$

where $\|\theta\|_2 \le 1$. Now note that $\|\psi(s,a)\|_2 \le 1$ and $\|\sum_{s'} \phi(s,a,s')V_{h+1,f}(s')\|_2 \le 1$, we can get

$$\begin{aligned}
&\sum_{i=1}^{k} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,f^i}(s_{h+1}, \eta_h^i, f^i, f^i) \right] \right\|_2^2 \\
&= \sum_{i=1}^{k} \left((\theta_{h,f^i} - \theta_h^*)^T \left[\psi(s_h^i, a_h^i) + \sum_{s'} \phi(s_h^i, a_h^i, s') V_{h+1,f^i}(s') \right] \right)^2 \\
&= \sum_{i=1}^{k} 4 \wedge \left((\theta_{h,f^i} - \theta_h^*)^T \left[\psi(s_h^i, a_h^i) + \sum_{s'} \phi(s_h^i, a_h^i, s') V_{h+1,f^i}(s') \right] \right)^2 \\
&\le \sum_{i=1}^{k} 4 \wedge \|X_h(f^i)\|_{\Sigma_i}^2 \|W_{h,f^i}(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 \\
&\le \sum_{i=1}^{k} 4 \wedge (\beta + 4) \|W_{h,f^i}(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 \\
&\le (\beta + 4) \sum_{i=1}^{k} \left(1 \wedge \|W_{h,f^i}(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 \right).
\end{aligned}$$

By the Elliptical Potential Lemma (Dani et al., 2008; Abbasi-Yadkori et al., 2011),

$$\begin{aligned}
\sum_{i=1}^{k} \left(1 \wedge \|W_{h,f^i}(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 \right) &\le \sum_{i=1}^{k} 2\log(1 + \|W_{h,f^i}(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2) \\
&\le 2 \log \frac{\det(\Sigma_{k+1})}{\det(\sigma_0)}.
\end{aligned}$$

Now note that $\det(\Sigma_{k+1}) \le \left(\frac{\mathrm{tr}(\Sigma_{k+1})}{d} \right)^d$, then

$$\begin{aligned}
2 \log \frac{\det \Sigma_{k+1}}{\det \Sigma_0} &= 2 \log \det(\Sigma_{k+1}) - 2 \log \det(\Sigma_0) \\
&\le 2d \log \left(1 + \frac{\sum_{i=1}^{k} \mathrm{tr}(W_{h,f^i}(s_h^i, a_h^i) W_{h,f^i}(s_h^i, a_h^i)^T)}{d} \right) \\
&\le 2d \log \left(1 + \frac{\sum_{i=1}^{k} \|W_{h,f^i}(s_h^i, a_h^i)\|_2^2}{d} \right) \\
&\le 2d \log \left(1 + \frac{4k}{d} \right).
\end{aligned}$$

So we can get

$$\sum_{i=1}^{k} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,f^i}(s_{h+1}, \eta_h^i, f^i, f^i) \right] \right\|_2^2 \le 2d(\beta+4) \log \left(1 + \frac{4k}{d} \right) = \mathcal{O}(d\beta \log k),$$

where we ignore all the terms that are independent with k. □

I.7 Proof of Lemma E.5

In this subsection, we prove that decoupled Markov Games belong to ℓ_2-type EC class with $\mathcal{T}_h Q_{h+1}(s,a,b) = r_h(s,a,b) + \mathbb{E}_{s'|\mathbb{P}_h(s'|s,a,b)} \max_{\upsilon} \min_{\mu} Q_h(s', \upsilon, \mu)$.

1. Dominance With probability at least $1 - \delta$,

$$\sum_{i=1}^{k} (V_{1,f^i}(s_1) - V_{\pi^i}(s_1))$$

$$
\begin{aligned}
&= \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\mathbb{E}_{s_h\sim\pi^i}\Big[V_{h,f^i}(s_h) - r_h - V_{h+1,f^i}(s_{h+1})\Big] \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\mathbb{E}_{\pi^i}\Big[\min_{\mu}\mathbb{P}_{\upsilon^i,\mu}Q_{h,f^i}(s_h,\upsilon^i,\mu) - r_h - V_{h+1,f^i}(s_{h+1})\Big] \qquad \text{(I.14)} \\
&\le \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\mathbb{E}_{\pi^i}\Big[\mathbb{P}_{\upsilon^i,\mu^i}Q_{h,f^i}(s_h,\upsilon^i,\mu^i) - r_h - V_{h+1,f^i}(s_{h+1})\Big] \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\mathbb{E}_{\pi^i}\Big[Q(s_h,a_h,b_h) - r_h - V_{h+1,f^i}(s_{h+1})\Big] \qquad \text{(I.15)} \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\Big[Q_{h,f^i}(s_h^i,a_h^i,b_h^i) - r_h - V_{h+1,f^i}(s_{h+1})\Big] + \mathcal{O}(\sqrt{KH}\log(KH/\delta)) \qquad \text{(I.16)} \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k}\mathbb{E}_{s_{h+1}}\Big[\ell_{h,f^i}(s_{h+1},\{s_h^i,a_h^i,b_h^i\},f^i,f^i)\Big] + \mathcal{O}(\sqrt{KH}\log(KH/\delta)),
\end{aligned}
$$

where the Eq.(I.14) holds because the greedy policy $\upsilon^i = \upsilon_{f^i}$ satisfies that

$$
V_{h,f^i}(s_h) = \min_{\mu}\mathbb{P}_{\upsilon^i,\mu}Q_{h,f^i}(s_h,\upsilon^i,\mu).
$$

Eq.(I.15) holds because $\pi_h^i = (\upsilon^i,\mu^i)$, and Eq.(I.16) follows from Azuma-Hoeffding's inequality.

2. Decomposable Property

$$
\begin{aligned}
&\ell_{h,f'}(\zeta_h,\eta_h,f,g) - \mathbb{E}_{\zeta_h}\Big[\ell_{h,f'}(\zeta_h,\eta_h,f,g)\Big] \\
&\quad = \Big[Q_{h,g}(s_h,a_h,b_h) - r(s_h,a_h,b_h) - V_{h+1,f}(s_{h+1})\Big] \\
&\qquad - \Big[Q_{h,g}(s_h,a_h,b_h) - (\mathcal{T}_h V_{h+1,f})(s_h,a_h,b_h)\Big] \\
&\quad = \Big[(\mathcal{T}_h V_{h+1,f})(s_h,a_h,b_h) - r(s_h,a_h,b_h) - V_{h+1,f}(s_{h+1})\Big] \\
&\quad = \ell_{h,f'}(\zeta_h,\eta_h,f,\mathcal{T}(f)).
\end{aligned}
$$

3. ℓ_2-type eluder Condition Note that $\mathbb{E}_{\zeta_h}[\ell_{h,f'}(\zeta_h,\eta_h,f,g)] = [f_h - \mathcal{T}_h f_{h+1}](s_h^i,a_h^i,b_h^i)$ is the Bellman residual of Markov Games. The proof can be derived similarly to Lemma 3.3 by replacing $\{a_h\}$ and the Bellman operator for single-agent MDP to $\{a_h,b_h\}$ and the Bellman operator for the two-player zero-sum MG.

I.8 Proof of Lemma E.3

Proof. We proof this Lemma by the classical ℓ_2 eluder argument. First, denote $U_{h,f,j}$ with $j\in[d_s]$ as the j-th row of $U_{h,f}$, then

$$
\|(U_{h,f} - U_h^*)\phi(s,a)\|_2^2 = \sum_{j=1}^{d_s}\|(U_{h,f,j} - U_{h,j}^*)\phi(s,a)\|_2^2.
$$

Then, denote $\Sigma_k = \sum_{i=1}^{k-1}\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^T + \lambda I$, we can get

$$
\begin{aligned}
\sum_{j=1}^{d_s}\|U_{h,f^k,j} - U_{h,j}^*\|_{\Sigma_k}^2 &= \sum_{i=1}^{k-1}\sum_{j=1}^{d_s}((U_{h,f^k,j} - U_{h,j}^*)\phi(s_h^i,a_h^i))^2 + \lambda\sum_{j=1}^{d_s}\|U_{h,f^k,j} - U_{h,j}^*\|_2^2 \\
&\le \beta + \lambda\cdot d_s R^2,
\end{aligned}
$$

where $\|U_{h,f,j}\|_2^2 \le \|U_{h,f}\|_2^2 \le R^2$ for any $f \in \mathcal{F}$. Now, recall that $\|\phi(s,a)\|_2 \le 1$, then $\|(U_{h,f^i} - U_h^*)\phi(s,a)\|_2^2 \le 4R^2$. Hence, choosing $\lambda = \frac{4}{d_s}$,

$$\begin{aligned}
\sum_{i=1}^k \|(U_{h,f^i} - U_h^*)\phi(s_h^i, a_h^i)\|_2^2 &= \sum_{i=1}^k \left(\|(U_{h,f^i} - U_h^*)\phi(s_h^i, a_h^i)\|_2^2 \wedge 4R^2\right) \\
&= \sum_{i=1}^k \left(\left(\sum_{j=1}^{d_s} \|(U_{h,f^i,j} - U_{h,j}^*)\phi(s_h^i, a_h^i)\|_2^2\right) \wedge 4R^2\right) \\
&\le \sum_{i=1}^k \left(\left(\sum_{j=1}^{d_s} \|(U_{h,f^i,j} - U_{h,j}^*)\|_{\Sigma_i}^2 \|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2\right) \wedge 4R^2\right) \\
&\le \sum_{i=1}^k \left(\left(\|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 \sum_{j=1}^{d_s} \|U_{h,f^k,j} - U_{h,j}^*\|_{\Sigma_k}^2\right) \wedge 4R^2\right) \\
&\le \sum_{i=1}^k \left(\left(\|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2 (\beta + 4R^2)\right) \wedge 4R^2\right) \\
&\le \sum_{i=1}^k (\beta + 4R^2) \left(1 \wedge \|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2\right).
\end{aligned}$$

By the Elliptical Potential Lemma (Dani et al., 2008; Abbasi-Yadkori et al., 2011), we have

$$\begin{aligned}
\sum_{i=1}^k \left(1 \wedge \|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2\right) &\le \sum_{i=1}^k 2\log(1 + \|\phi(s_h^i, a_h^i)\|_{\Sigma_i^{-1}}^2) \\
&\le 2\log \frac{\det \Sigma_{k+1}}{\det \Sigma_0}.
\end{aligned}$$

Now note that $\det(\Sigma_{k+1}) \le \left(\frac{\mathrm{tr}(\Sigma_{k+1})}{d_\phi}\right)^{d_\phi}$, then

$$\begin{aligned}
2\log \frac{\det \Sigma_{k+1}}{\det \Sigma_0} &= 2\log\det(\Sigma_{k+1}) - 2\log\det(\Sigma_0) \\
&\le 2d_\phi \log\left(\lambda + \frac{\sum_{i=1}^k \mathrm{tr}(\phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^T)}{d_\phi}\right) \\
&\le 2d_\phi \log\left(\lambda + \frac{\sum_{i=1}^k \|\phi(s_h^i, a_h^i)\|_2^2}{d_\phi}\right) \\
&\le 2d_\phi \log\left(\lambda + \frac{k}{d_\phi}\right).
\end{aligned}$$

Thus

$$\sum_{i=1}^k \|(U_{h,f^i} - U_h^*)\phi(s_h^i, a_h^i)\|_2^2 \le 2d_\phi(\beta + 4R^2)\log\left(\frac{4}{d_s} + \frac{k}{d_\phi}\right) = \mathcal{O}(d_\phi \beta \log k),$$

where we ignore all terms independent with k, and regard R as a constant. □