AN INTRODUCTION TO ELUDER COEFFICIENT: THEORETICAL FRAMEWORK AND ALGORITHMIC DESIGN

Wei Xiong *

ABSTRACT

We are interested in online decision-making problems with general function approximation. We find that the existing techniques can be largely divided into two groups: 1) the eluder technique, which relates the out-of-sample target to the in-sample historical error; 2) the decoupling technique, which relates the out-of-sample target to another out-of-sample target but may be easier to handle. Technically, the first framework reduces the problem into an *in-sample* supervised learning problem, while the second framework reduces the problem into an online learning problem. We introduce the *eluder coefficient*, as a representative of the first type of complexity measure, which contains nearly all known tractable problems. Along the line, we also present some new algorithms to solve problems with a low eluder coefficient. We hope this note helps to provide a better understanding and interpretation of various algorithmic ideas in the literature, including optimism, attempts for handling the double-sampling issue, UCB, and posterior sampling. If you have any questions, feel free to contact me!

CONTENTS

*First draft: January 2022.
Second version: December 2022, added optimistic FLI.
Last update: May 2023, prepared for my thesis defense.

1 INTRODUCTION

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

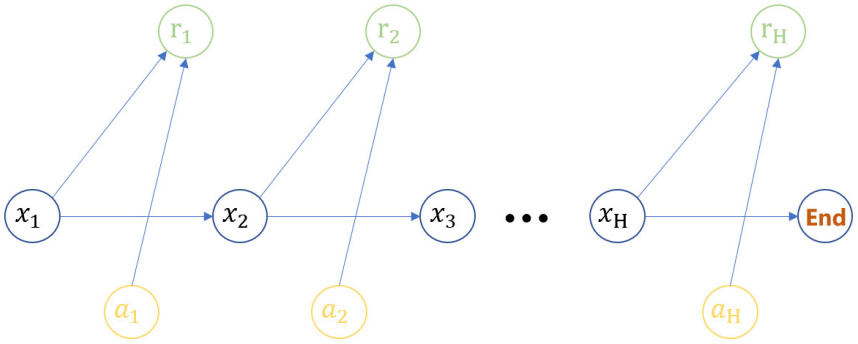

Figure 1: An illustration of MDP with episode length H.

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

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

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

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

Here $\mathcal{T}_h$ is referred to as the *Bellman operator* at step h. We also use $\pi^* = \{\pi_h^*\}_{h\in[H]}$, $V^* = \{V_h^*\}_{h\in[H]}$, and $Q^* = \{Q_h^*\}_{h\in[H]}$ to denote the optimal (Markovian) policy, optimal value function and optimal Q-function, respectively, where they satisfy the following properties Sutton & Barto (2018):

$$\begin{aligned} Q_h^{\pi^*}(x, a) = Q_h^*(x, a) = \sup_\pi Q_h^\pi(x, a) \qquad \forall (x, a) \in \mathcal{S} \times \mathcal{A}, \\ V_h^{\pi^*}(x) = V_h^*(x) = \sup_\pi V_h^\pi(x) \qquad \forall x \in \mathcal{S}, \end{aligned} \tag{2}$$

and one optimal policy π^* is the greedy policy induced by Q^*. It is also well known that (Q^*, V^*) satisfies the *Bellman optimality equation* for any $(h, x, a) \in [H] \times \mathcal{S} \times \mathcal{A}$:

$$Q_h^*(x, a) = (\mathcal{T}_h V_{h+1}^*)(x, a), \qquad V_h^*(x) = \max_{a\in\mathcal{A}} Q_h^*(x, a). \tag{3}$$

For simplicity, we additionally assume that the first state is a fixed one x_1

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

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

The goal is to design an algorithm to solve the underlying MDP with a sub-linear (in T) regret.

1.1 TABULAR MDP

For a tabular MDP, we assume that the state space $\mathcal{S}$ and action space $\mathcal{A}$ are small. But we do not impose any structural assumption across states and actions. The goal in the tabular case is to design algorithms that achieve a regret depending polynomially on S, A and also the horizon H. The tabular MDP has been extensively studied in the literature (Auer et al., 2008; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Agrawal & Jia, 2017; Zanette & Brunskill, 2019; Zhang et al., 2020; 2021; Ménard et al., 2021; Li et al., 2021; Wu et al., 2022; Zhang et al., 2022). Among them, Azar et al. (2017) designs a model-based algorithm UCB-VI that explicitly models the transition matrix of the MDP, and attains the minimax-optimal regret bound $\mathcal{O}(\sqrt{H^2SAT})$. After this, Zhang et al. (2020) closes the gap to the lower bound in the model-free setting by leveraging the idea of variance reduction by a reference function Johnson & Zhang (2013). Since both the model-based and model-free algorithms attain the minimax-optimal regret bound, the tabular settings are well-studied. However, since the lower bound depends on the $\sqrt{SA}$, we cannot handle modern RL problems with large state space without further structural assumptions.

The idea is to employ function approximation either the model dynamics (the transition kernel and the reward function), the value functions (e.g. Q^*, V^*, Q^π), or the policy by an abstract hypothesis space $\mathcal{H}$. We always assume that $\mathcal{H}$ is finite to illustrate the idea and it can be readily extended to the infinite class with a mild covering number by standard discretization technique.

1.2 SUPERVISED LEARNING

Before we dive into the details of RL, we first review some classic results from supervised learning. We assume (X, Y) is sampled from some unknown distribution $P(X, Y)$. For a fixed a hypothesis $f \in \mathcal{H}$, the *population risk* is given by

$$L(f) = \mathbb{E}_{(X,Y)\sim P}\ell(f(X), Y),$$

where $\ell(\cdot, \cdot)$ is a loss function, e.g., $\ell(f(x), y) = \frac{1}{2}(f(x) - y)^2$. Given a data set $\mathcal{D} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, we also define the *empirical risk* as

$$\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Clearly, we have $\mathbb{E}_P\widehat{L}(f) = L(f)$. According to the plug-in principle, we can adapt the ERM method (empirical risk minimization):

$$\widehat{f} := \operatorname*{argmin}_{f\in\mathcal{H}} \widehat{L}(f). \tag{4}$$

What do we care about? Suppose that the minimizer of $L(f)$ is $f^* \in \mathcal{H}$. Here we assume $0 \le \ell(\cdot, \cdot) \le b$ for simplicity (sub-Gaussian assumption). What we are really concerned about is the performance of $\widehat{f}$ on the *out-of-sample* samples, i.e., the population risk $L(\widehat{f})$. We can decompose the performance difference as follows.

$$L(\widehat{f}) - L(f^*) = \underbrace{\left(L(\widehat{f}) - \widehat{L}(\widehat{f})\right)}_{A} + \underbrace{\left(\widehat{L}(\widehat{f}) - \widehat{L}(f^*)\right)}_{B} + \underbrace{\left(\widehat{L}(f^*) - L(f^*)\right)}_{C}$$

We first note that $B \le 0$ because $\widehat{f}$ minimizes $\widehat{L}(\cdot)$. Term C can be directly controlled by Hoeffding's inequality. However, we cannot apply it to the term A because $\{\ell(\widehat{f}(X_i, Y_i) : i = 1, 2, \cdots, n\}$ are not independent since $\widehat{f}(\cdot)$ is obtained via a minimization problem over the data set $\mathcal{D}$ and consequently the i.i.d. assumption over $\mathcal{D}$ does not hold for $\ell(X_i, Y_i)$. We can use the following *uniform concentration argument* to solve this problem.

We first note that for a sequence of i.i.d. random variables $\{X_k\}_{k=1}^n$, $\{f(X_k)\}_{k=1}^n$ are also i.i.d. for any fixed function. We take $\delta_f = \frac{\delta}{|\mathcal{H}|}$ to construct a concentration bound for each $f \in \mathcal{H}$. Then,

applying a union bound over all $f \in \mathcal{H}$:

$$P\Big(\sup_{f\in\mathcal{H}} |L(f) - \widehat{L}(f)| > b\sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}|}}\Big) \le \sum_{f\in\mathcal{H}} P\Big(|L(f) - \widehat{L}(f)| > b\sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}|}}\Big)$$
$$\le |\mathcal{H}| \times \frac{\delta}{|\mathcal{H}|} = \delta, \tag{5}$$

where we use the boundedness assumption in the second inequality to apply Hoeffding's inequality. We can apply Hoeffding's inequality here because f in $P\left(|L(f) - \widehat{L}(f)| > \sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}|}}\right)$ is a fixed function instead of a function obtained via a minimization problem over $\mathcal{D}$. Then, we have

$$L(\widehat{f}) - L(f^*) \le \underbrace{\Big(L(\widehat{f}) - \widehat{L}(\widehat{f})\Big)}_{A} + \underbrace{\Big(\widehat{L}(f^*) - L(f^*)\Big)}_{C}$$
$$\le \sup_{f\in\mathcal{H}} |L(f) - \widehat{L}(f)| + \Big(\widehat{L}(f^*) - L(f^*)\Big)$$
$$\le 2\sup_{f\in\mathcal{H}} |L(f) - \widehat{L}(f)| \le \widetilde{O}\big(\frac{b}{\sqrt{n}}\big).$$

In particular, if the capacity of $\mathcal{H}$ is large enough so that $L(f^*) = 0$ (referred to as the realizability). We know that it takes $\widetilde{O}\big(\frac{b^2\log|\mathcal{H}|}{\epsilon^2}\big)$ samples to ensure that $L(\widehat{f}) \le \epsilon$ with high probability.

1.3 Function Approximation in RL

We now turn to the function approximation in RL where we approximate either the model dynamics (the transition kernel and the reward function), the value functions (e.g. Q^*, V^*, Q^π), or the policy by an abstract hypothesis space $\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_H$. For instance, the DQN (illustrated in 2) approximates the Q^* by a deep neural network.

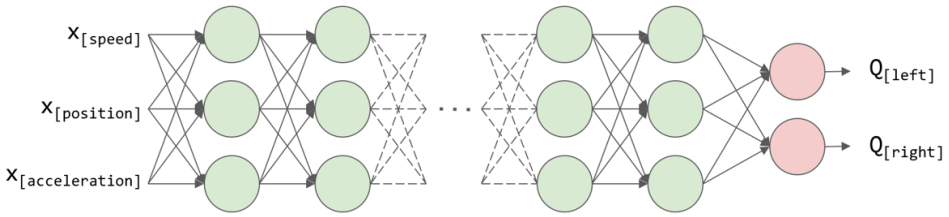

Figure 2: An illustration of DQN.

Linear function approximation and learnability. Linear function approximation is arguably the most fundamental one Wang et al. (2019); Yang & Wang (2019); Cai et al. (2020); Jin et al. (2020); Zanette et al. (2020); Ayoub et al. (2020); Modi et al. (2020); Zhou et al. (2021); Zhong & Zhang (2023); Agarwal et al. (2022); He et al. (2022). Typically, we will assume that we have access to a d-dimensional feature map of the state-action pair $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. A natural idea is to assume that the optimal Q-value Q^* is linear in this feature, in the sense that there exists a $\theta_h \in \mathbb{R}^d$ and $||\theta_h^*|| \le B$:

$$Q_h^*(x,a) := \langle \phi(x,a), \theta_h^* \rangle, \qquad \forall h \in [H]. \tag{6}$$

In this case, if we choose $\mathcal{H}_h = \{\theta \in \mathbb{R}^d : ||\theta|| \le B\}$, we know that $\theta^* \in \mathcal{H}$. One may expect that this is sufficient for sample-efficient learning as both realizability and a mild covering number hold for linear models. However, the following result shows that realizability itself is not sufficient for sample-efficient learning.

Proposition 1 (Linear-realizability is not sufficient Wang et al. (2021))**.** *There exists an MDP with feature map ϕ that satisfies equation 6 but any algorithms must have*

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

Challenges in online decision-making problems. We notice that supervised learning and RL are different in the following sense.

- Supervised learning: $\{x_i, y_i\}_{i=1}^n$ i.i.d. from a static distribution $\mathcal{D}_{\text{data}}$;
- RL: $\zeta_1 \sim \mathcal{D}_{\pi^1}, \zeta_2 \sim \mathcal{D}_{\pi^2}, \cdots, \zeta_T \sim \mathcal{D}_{\pi^T}$, distribution shifts all the time!

Therefore, the generalization from the *in-sample error* to the *out-of-sample* prediction error in supervised learning no longer holds in RL:

$$\mathcal{D}_1, \cdots, \mathcal{D}_{t-1} \xrightarrow{?} \mathcal{D}_t.$$

Intuitively, we need a different generalization guarantee in online decision-making problems, which is supposed to be incremental and also can handle the distribution shifts.

More notations. Following Du et al. (2021), we will assume that $\mathcal{H}$ can be either *model-based* or *value-based*, and we detail them as follows.

Example 1 (Model-based Hypothesis)**.** *A model-based hypothesis class $\mathcal{H}$ is a set of models (transition kernel $\mathbb{P}$ and reward function r). In this case, for any $f = (\mathbb{P}_f, r_f) \in \mathcal{H}$, we denote $\pi_f = \{\pi_{h,f}\}_{h\in[H]}$ and $Q_f = \{Q_{h,f}\}_{h\in[H]}, V_f = \{V_{h,f}\}_{h\in[H]}$ as the optimal policy and optimal value functions corresponding to the model f, respectively. We also denote the real model by f^*.*

Example 2 (Value-based Hypothesis for MDP)**.** *A value-based hypothesis class $\mathcal{H}$ is a set of Q-function, that is, $\mathcal{H} = \{\mathcal{H}_h\}_{h\in[H]}$, where $\mathcal{H}_h = \{Q_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. For any $f = \{Q_h\}_{h\in[H]}$, let $Q_f = \{Q_{h,f} = Q_h\}_{h\in[H]}$, $V_f = \{V_{h,f}(\cdot) = \max_{a\in\mathcal{A}} Q_{h,f}(\cdot, a)\}_{h\in[H]}$, and $\pi_f = \{\pi_{h,f}(\cdot) = \operatorname{argmax}_{a\in\mathcal{A}} Q_h(\cdot, a)\}_{h\in[H]}$. We also denote $f^* = Q^*$, where Q^* is the optimal Q-function.*

To further improve readability, sometimes we will also use $\mathcal{F}$ for a value-based hypothesis, and $\mathcal{M}$ for a model-based hypothesis to distinguish them. We remark that the main difference between value-based and model-based hypothesis spaces is whether we use or learn the information of the transition kernel. For each $f \in \mathcal{H}$, we define the Bellman residual as

$$\mathcal{E}_h(f, x, a) := Q_{h,f}(x, a) - (\mathcal{T}_h V_{h+1,f})(x, a). \tag{7}$$

By equation 3, we know that $\mathcal{E}_h(f^*, x, a) = 0$ for all $(h, x, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Throughout this paper, we will assume that $\mathcal{H}$ contains f^* (c.f. Example 1 or Example 2), which is standard in the literature (e.g., Jiang et al., 2017b; Jin et al., 2021a; Du et al., 2021; Dann et al., 2021).

Assumption 1 (Realizability)**.** *We assume $f^* \in \mathcal{H}$.*

Remark 1 (Notions of realizability.)**.** *For the model-based hypothesis set $\mathcal{M}$, Assumption 1 means that the true model $M^* \in \mathcal{M}$. For the value-based hypothesis $\mathcal{F}$, Assumption 1 means that the $Q^* \in \mathcal{F}$. Clearly, model-based realizability implies value-based realizability. We will see that we can obtain a sharper result with the model-based hypothesis under realizability.*

2 Generalization is limited in linear case

One rough intuition is that when the "complexity" of the problems is limited or there exists a certain underlying structure across different state-action pairs, the generalization cannot be infinite. We hope to quantify such an ability of generalization. To this end, we first consider a d-dimensional linear space

$$\mathcal{H} = \{f(\cdot) = \phi(\cdot)^\top \theta_f : ||\theta_f|| \le 1\}.$$

Given the historical sequence $\{f_1, g_1, z_1, \cdots, f_{t-1}, g_{t-1}, z_{t-1}\}$, we examine the prediction error on the unseen data: $|f_t(z_t) - g_t(z_t)|$. We let $\Sigma_t := \lambda \mathbf{I} + \sum_{s=1}^{t-1} \phi(z_s)\phi(z_s)^\top$. It follows that

$$\begin{aligned} |f(z_t) - g(z_t)|^2 = |\langle \phi(z_t), \theta_f - \theta_g \rangle|^2 &\le ||\phi(z_t)||^2_{\Sigma_t^{-1}} ||\theta_f - \theta_g||^2_{\Sigma_t} \\ &\le ||\phi(z_t)||^2_{\Sigma_t^{-1}} \Big(\lambda + \sum_{s=1}^{t-1} |f(z_s) - g(z_s)|^2\Big), \end{aligned} \tag{8}$$

where we use Cauchy-Schwarz inequality and expand Σ_t in the last inequality. Therefore, the out-of-sample prediction error on the unseen z_t can be bounded by the (regularized) in-sample training

error up to a factor of $||\phi(z_t)||^2_{\Sigma_t^{-1}}$, where $||\phi(z_t)||^2_{\Sigma_t^{-1}}$ is referred to as the elliptical potential in the literature Abbasi-Yadkori et al. (2011). The following lemma shows that the potential is small for most of the time.

Lemma 1 (Elliptical Potential Lemma (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011)). *Given $\lambda > 0$ and $\{X_t\}_{t=1}^T \subset \mathbb{R}^d$ with $||X_t|| \leq L$ for all $t \in [T]$, if we denote $\Sigma_t = \lambda \mathbf{I} + \sum_{s=1}^t X_s X_s^\top$, then $||X_t||_{\Sigma_{t-1}^{-1}} \geq 1$ happens for at most $\frac{3d}{\log 2} \log\big(1 + \frac{L^2}{\lambda \log 2}\big)$. times. It further holds that*

$$\sum_{t=1}^T \min\{1, \|x_t\|^2_{\Lambda_{t-1}^{-1}}\} \leq 2 \log\Big(\frac{\det(\Lambda_T)}{\det(\lambda \mathbf{I})}\Big) \leq 2d \log\Big(\frac{\mathrm{trace}(\lambda \mathbf{I}) + TL^2}{d \det(\lambda \mathbf{I})^{1/d}}\Big).$$

This implies the following reduction of prediction to the in-sample error.

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

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

Proof. Following the idea in equation 8, we decompose the prediction error into the in-sample error and potential:

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

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

This exactly matches our intuition as it proves that for linear function class, the prediction error can be controlled by the in-sample error on average, although it is amplified by dimension d.

We can extend the idea in the linear class to a more general setting to handle the RL problems. Before continuing, one may be curious why the linear Q^* is not sufficient. To answer this question, we need to specify the notion of prediction error in RL first.

3 Value Decomposition and Optimism

Lemma 2 (Regret decomposition). *Suppose that we execute π_{f^t} (i.e., the greedy policy of f^t) for each iteration. Then, it holds that:*

$$\begin{aligned}
\mathrm{Reg}(T) = \sum_{t=1}^T V_1^*(x_1) - V_1^{\pi_{f^t}}(x_1) &= \sum_{t=1}^T [V_{1,f^t} - V_1^{\pi_{f^t}}] + \sum_{t=1}^T [V_1^* - V_{1,f^t}] \\
&\leq \underbrace{\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{(I)}} - \underbrace{\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{(II)}}.
\end{aligned} \tag{9}$$

Proof. The value decomposition lemma has been widely used in the literature (Jiang et al., 2017a; Cai et al., 2020). We present the proof in the appendix for completeness. □

In comparison, term (II) is hard to control in the online setting because we have no knowledge about π^* and cannot design algorithms in terms of it. The idea is to adopt the general principle, "Optimism in the Face of Uncertainty" (OFU, Auer et al. (2002); Dani et al. (2008); Li et al. (2010); Jiang et al. (2017b); Jin et al. (2021a); Du et al. (2021)) so that $V_{1,f^t} \geq V_1^*$ and the second term is eliminated. In this case, our target becomes the *out-of-sample* Bellman error:

$$\sum_{t=1}^{T}[V_{1,f^t} - V_1^{\pi_{f^t}}] = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right].$$

We mention in passing that in the offline setting, however, we only have a pre-determined dataset $\mathcal{D}$ and $T = 1$. In this case, controlling term (I) may require a good coverage over all $f \in \mathcal{H}$ (because we need to evaluate all $f \in \mathcal{H}$ with this pre-determined $\mathcal{D}$), which is too strong to hold in practice. On the other hand, if we adopt the pessimism principle such that $V_{1,f} \leq V_1^{\pi_f}$, we only need a good coverage over π^* (single-policy coverage). We refer interested readers to Jin et al. (2021b) for a detailed illustration.

4 Eluder Coefficient

We now extend the idea from the linear class to a more general scenario with the prediction error identified in the last section.

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

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

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

where $B^\dagger > 0$ is some problem-dependent constant for regularization. In general, we may use a different notion of training error, and define the generalized eluder coefficient as follows:

$$\sum_{t=1}^{T} V_{1,f^t} - V_1^{\pi_{f^t}} \lesssim \Bigg[\underbrace{d(\epsilon)}_{\textit{cost of generalization}} \sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\sum_{s=1}^{t-1}\ell_h^s(f^t)}_{\textit{training error}}\Bigg]^{1/2},$$

where $\ell_h^s(\cdot) : \mathcal{H} \to R^+$ and $\ell_h^s(f^*) = 0$ holds for any $(s, h) \in [T] \times [H]$. We omit the burn-in cost for a clearer presentation.

Ignoring the burn-in cost, which is typically non-dominating, the eluder coefficient suggests that the prediction error can be upper bounded by the cumulative training error on average, although the training error is amplified by the eluder coefficient. Therefore, the eluder coefficient can be used to measure the hardness of such a generalization, thus further serving to measure the hardness of learning the MDP. We make several remarks before continuing.

Remark 2. *The first definition presented here is from Dann et al. (2021) (up to a Cauchy-Schwarz inequality), except that the expectation is inside the square in the training error. This is a very subtle difference and Xie et al. (2022) thoroughly studies it from a statistical viewpoint. We choose this formulation because this can be handled with only realizability. Otherwise, one need Bellman completeness condition to achieve sample-efficient learning.*

Remark 3. *There is an expectation in the notion of loss, instead of evaluating the loss at a specific point as in the Lemma 1. Therefore, the implicit structure assumption is now imposed on the interplay between the MDP and the hypothesis class, which allows this formulation to capture more problems.*

4.1 EXAMPLE: LINEAR MDP

We now illustrate the idea in terms of the linear MDP Jin et al. (2020).

Example 3 (Linear MDP). *MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a (known) feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown signed measures $\mu_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$, we have $\mathbb{P}_h(\cdot \mid x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle$, $r_h(x, a) = \langle \phi(x, a), \theta_h \rangle$. Without loss of generality, we assume that $||\phi(x, a)|| \le 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{||\mu_h(\mathcal{S})||, ||\theta_h||\} \le \sqrt{d}$ for all $h \in [H]$.*

For linear MDP, both the transition kernel and the reward function are linear in a known feature, which further implies that the Bellman residual of any $f \in \mathcal{H}$ is linear, thus reducing to the case of Theorem 1.

Lemma 3 (Linear MDP has a low eluder coefficient). *For the linear MDP defined in Example 3, if we take $\mathcal{H}_h = \{Q_{h,f}(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \theta_{h,f} : \|\theta_{h,f}\| \le \sqrt{d}H\}$, then it has an eluder coefficient of $d(\epsilon) = \mathcal{O}\big(Hd \log(1 + \frac{T}{\epsilon})\big)$.*

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

$$\begin{aligned} \mathcal{T}_h V(x, a) &= r_h(x, a) + (\mathbb{P}_h V)(x, a) = \phi(x, a)^\top \theta_h + \int_{\mathcal{S}} V(x_{h+1}) \left\langle \phi(x, a), d\mu_h(x_{h+1}) \right\rangle \\ &= \left\langle \phi(x, a), \theta_h + \int_{\mathcal{S}} V(x_{h+1}) d\mu_h(x_{h+1}) \right\rangle := \langle \phi(x, a), w_h \rangle . \end{aligned} \tag{10}$$

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

Lemma 4. *For linear MDP with hypothesis class $\mathcal{H}_h = \{Q_{h,f}(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \theta_{h,f} : \|\theta_{h,f}\| \le \sqrt{d}H\}$, for any $f \in \mathcal{H}$ and $h \in [H]$, there exists a $X_h(f) \in \mathbb{R}^d$ such that $\mathbb{E}_{\pi_f} \mathcal{E}_h(f, x_h, a_h) = \langle X_h(f), \theta_{h,f} - w_{h,f} \rangle$, where $Q_{h,f}(x, a) = \phi(x, a)^\top \theta_{h,f}$ and $\mathcal{T}_h V_{h+1,f}(x, a) = \phi(x, a)^\top w_{h,f}$. Moreover, by definition of linear MDP, it holds*

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

Proof. By equation 10, we know that for any $V_{h+1,f}$ associated with $f \in \mathcal{H}$, we can assume that there exists a $w_{h,f} \in \mathbb{R}^d$ such that $\mathcal{T}_h V_{h+1,f}(\cdot, \cdot) = \phi(\cdot, \cdot)^\top w_{h,f}$. Meanwhile, for any $Q_{h,f}$, it can be represented by $Q_{h,f}(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \theta_{h,f}$ for some $\theta_{h,f} \in \mathbb{R}^d$. As a result, for any $f \in \mathcal{H}$, the Bellman residual is also linear:

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

Therefore, we know that $X_h(f) = \mathbb{E}_{\pi_f} \phi(x_h, a_h)$ satisfies the condition. □

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

$$\begin{aligned} &\sum_{t=1}^{T} V_{1,f^t}(x_1) - V_1^{\pi^t}(x_1) = \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \left(\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} + \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} > 1\}\right) \\ &\le H \cdot \sum_{t=1}^{T} \sum_{h=1}^{H} \min\left\{ |\langle X_h(f^t), \frac{\theta_{h,f^t} - w_{h,f^t}}{H} \rangle|, 1 \right\} \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} \\ &\quad + H \cdot \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} > 1\} \\ &\le H \cdot \sum_{t=1}^{T} \sum_{h=1}^{H} \min\left\{ |\langle X_h(f^t), \frac{\theta_{h,f^t} - w_{h,f^t}}{H} \rangle|, 1 \right\} \mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}} \le 1\} + \min\{H\widetilde{d}, H^2 T\}, \end{aligned} \tag{11}$$

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

$$\begin{aligned}
&\min\left\{|\left\langle X_h(f^t), \frac{\theta_{h,f^t}-w_{h,f^t}}{H}\right\rangle|, 1\right\}\mathbf{1}\{\|X_h(f^t)\|_{\Sigma_{t;h}^{-1}}\le 1\}\\
&\le ||\frac{\theta_{h,f^t}-w_{h,f^t}}{H}||_{\Sigma_{t;h}}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}\\
&=\frac{1}{H}\Big[\lambda||\theta_{h,f^t}-w_{h,f^t}||^2+\sum_{s=1}^{t-1}|\left\langle X_h(f^s),\theta_{h,f^t}-w_{h,f^t}\right\rangle|^2\Big]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}\\
&\le\Big[\lambda d+\frac{1}{H^2}\sum_{s=1}^{t-1}\left(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h)\right)^2\Big]^{1/2}\cdot\min\{||X_h(f^t)||_{\Sigma_{t;h}^{-1}},1\}
\end{aligned}\tag{12}$$

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

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

Plugging equation 12 into equation 11, we obtain that

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

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

As a motivating example, we first introduce the following linear mixture MDP Ayoub et al. (2020); Cai et al. (2020); Modi et al. (2020).

Example 4. *We say an MDP is a liner mixture model if there exists (known) feature $\phi:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}^d$ and $\psi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$; and (unknown) $\theta^*\in\mathbb{R}^d$, such that for all $h\in[H]$ and $(x,a,x')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, we have*

$$\mathbb{P}_h(x'\mid x,a)=\langle\theta_h^*,\phi(x,a,x')\rangle,\qquad r_h(x,a)=\langle\theta_h^*,\psi(x,a)\rangle.\tag{13}$$

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

Unfortunately, one cannot obtain a low eluder coefficient for linear mixture MDP by taking the loss function as $(\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f^t,x_h,a_h))^2$. However, it achieves a low eluder coefficient by considering a different loss function.

Lemma 5 (Linear mixture MDP has a low generalized eluder coefficient)**.** *We consider the hypothesis space $\mathcal{H}=\{f=(\theta_{1,f},\cdots,\theta_{H,f}):\forall h\in[H],||\theta_{h,f}||\le B\}$ and adopt the following loss function for linear mixture MDP:*

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

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

There is also a line of work assuming different bilinear structures on the induced Bellman residual space, including Bellman rank (Jiang et al., 2017b), witness rank (Sun et al., 2019), and bilinear rank (Du et al., 2021), whose GEC can be bound similarly.

4.2 Example: Eluder Dimension

The analysis presented in the last subsection heavily relies on the linear structure of the problem. In the literature, there is another general complexity measure, i.e., the eluder dimension Russo & Van Roy (2013); Wang et al. (2020); Jin et al. (2021a), that can cover a wide range of RL problems. We first introduce the definition of ϵ-independence.

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

Intuitively, independence means that we can find a $g \in \mathcal{G}$ such that even though it is consistent with the historical data (small in-sample error on $\{\mu_1, \mu_2, \cdots, \mu_n\}$), we can still suffer from a large error on the new distribution ν.

Definition 3 (Distributional eluder (DE) dimension)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{Z}$, and Π be a family of probability measures over $\mathcal{Z}$. The distributional eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \cdots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \ge \epsilon$ with ρ_i being ϵ'-independent of $\{\rho_1, \cdots, \rho_{i-1}\}$ for all $i \in [n]$.*

The distributional eluder dimension is simply the length of the longest sequence such that these independent things can happen consecutively in sequence. To be more specific, in what follows, we consider $\mathcal{G}_h = \{\mathcal{E}_h(f, \cdot, \cdot) : f \in \mathcal{H}\}$ as the induced Bellman residual class (in this case, it is referred to as the Bellman elduer dimension (Jin et al., 2021a)), but we note that the reduction technique can be applied to general function class. The following lemma (Dann et al., 2021) shows that a problem with a low distributional eluder dimension also has a low eluder coefficient.

Lemma 6. *Suppose that a problem has a distributional eluder dimension of $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ and suppose that $|g|$ is bounded by H for all $g \in \mathcal{G}$. Then, the eluder coefficient satisfies $d(\epsilon) \le \mathcal{O}(\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon) \log T)$ in the following sense: for arbitrary sequence of $\{(d_t, g_t) \in \Pi \times \mathcal{G}\}_{t=1}^T$, we have*

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

Remark 4. *We note that Russo & Van Roy (2013) developed a technique that could also achieve the goal of reducing the out-of-sample error into the historical one (see proposition 3 and lemma 2 of Russo & Van Roy (2013) for details). However, the result and subsequent adaptations (e.g. Jin et al. (2021a)) require a non-decreasing sequence $\{\beta_t\}_{t=1}^T$ such that $\sum_{s=1}^{t-1} (\mathbb{E}_{d_s} g_t)^2 \le \beta_t$ for all $t \in [T]$. This is suitable for the algorithms based on confidence set, but cannot be applied to the posterior sampling and also the algorithms we will present in Algorithm 5.*

In contrast, we note that the problems with a low eluder coefficient can have a large eluder dimension, as shown by the following lemma adapted from Xie et al. (2022).

Lemma 7. *Fix the time horizon $T > 0$ and $H = 2$. Let $\Pi = \Pi_1 \times \cdots \Pi_H$ and Π_h be a collection of probability measure families over $\mathcal{S} \times \mathcal{A}$ induced by following π_f. Then, there exists a class of MDPs such that the distributional eluder dimension $\max_{h \in [2]} \dim_{\mathrm{DE}}(\mathcal{G}_h, \Pi_h, 1/T^{1/3})$ is lower bounded by $\Omega(T^{1/3})$, while the eluder coefficient $d(\epsilon)$ is upper bounded by $\mathcal{O}(\log T)$, regardless of the ϵ, in the sense of: for an arbitrary sequence of $\{f^t \in \mathcal{H}\}_{t=1}^T$, we have*

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

To summarize, the bilinear-type low rank structure assumption and the eluder dimension assumption can be unified in the presented framework.

5 Algorithmic Design

5.1 Loss Estimator: Supervised Guarantee

One issue is that $\ell_h^s(\cdot)$ may not be directly available from the dataset because it involves an expectation. We introduce the following loss estimator based on the collected samples.

Definition 4 (Loss estimator with batch sampling)**.** *We consider a general sampling strategy where we sample m i.i.d. trajectories $\{\zeta_{i,h}^k\}_{i=1}^m$ [1] at each stage k and assume that T is divisible by m without loss of generality. With $K := T/m$, for each iteration $k \in [K]$, we suppose that we have access to a loss estimator $L_h^{1:k-1}(\cdot) : \mathcal{H} \to \mathbb{R}$, which only depends on the history:*

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

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

$$\sum_{s=1}^{k-1} \ell_h^s(f) \le L_h^{1:k-1}(f) + \Delta_h^k, \tag{14}$$

and

$$L_h^{1:k-1}(f^*) \le \Delta_h^k. \tag{15}$$

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

We first focus on the value-based case, where the loss function is $\ell_h^s(f) = \left(\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. As a motivating example, if we want to estimate $\sum_{s=1}^{k-1} (\mathbb{E} X_s)^2$ with a collection of samples $X_1, \cdots, X_{k-1}$ from the underlying distribution sequence, we cannot directly use the loss estimator $\sum_{s=1}^{k-1} X_s^2$ because

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

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

Lemma 8 (In-sample error estimation with trajectory average Jiang et al. (2017b); Du et al. (2021))**.** *Suppose that $\ell_h^s(f) = \left(\mathbb{E}_{x_h \sim \pi_{f^s}, a_h \sim \pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. We can independently sample m trajectories $\{\zeta_{i,h}^k\}_{i=1}^m$ by following π_{f^k} for each $k \in [K]$ and take*

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

where it satisfies equation 14 and equation 15 with $\Delta_h^k = \frac{4(k-1)H^2\iota}{m}$, and $\iota = O(\log(KH|\mathcal{H}|/\delta))$.

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

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

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

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

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

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

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

To prove equation 15, we note $\mathcal{E}_h(f^*, x_h, a_h) = 0$ for any (x_h, a_h). □

The guarantee provided by Lemma 8 can be sub-optimal since we use m samples for one hypothesis choice f^t. We should view the Lemma 8 as

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

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

Assumption 2 (Bellman Completeness). *We consider a value-based hypothesis $\mathcal{H}_h = \mathcal{F}_h \subset \{f_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. The hypothesis class is said to be Bellman complete, if for each $h \in [H]$, $\mathcal{T}_h^* \mathcal{F}_{h+1} \subset \mathcal{F}_h$, where $\mathcal{T}_h^* \mathcal{F}_{h+1} = \{\mathcal{T}_h f_{h+1} : f_{h+1} \in \mathcal{F}_{h+1}\}$ and*

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

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

Lemma 9 (In-sample error estimation with minimax formulation Antos et al. (2008); Jinglin & Jiang (2019); Jin et al. (2021a); Dann et al. (2021)). *Suppose that $\ell_h^s(f) = \left(\mathbb{E}_{x_h \sim \pi_{f^s}, a_h \sim \pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. We set $m = 1$ so $T = K$ in this case. For each $t \in [T]$, we can collect the trajectory ζ^t by following π_{f^t} and take $L_h^{1:t}(f) := \sum_{s=1}^t L_h^s(f)$ where*

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

where it satisfies equation 14 and equation 15 with $\Delta_h^t = O(H^2\iota)$ and $\iota = \log(H|\mathcal{H}|T/\delta)$.

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

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

where the conditional expectation of the second term is exactly the sampling variance, thus canceling the variance term. The reason is that we can approximate the second term (which is not available to the actually executed algorithm) by $\inf_{f_h' \in \mathcal{H}_h} \left(Q_{h,f'}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\right)^2$ and the Bellman completeness condition. The loss estimator in Lemma 9 is referred to as minimax formulation Antos et al. (2008) because it can be also written as

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

Remark 5. *Both Lemma 8 and Lemma 9 can ensure a small average loss $\sum_{s=1}^{t-1} \left(\mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f, x_h, a_h)\right)^2$. However, the minimax formulation further ensures a small $\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left(\mathcal{E}_h(f, x_h, a_h)\right)^2$, while the trajectory average cannot. We mention in passing that Chen et al. (2022b) further study such a formulation with the general loss function beyond Bellman residual.*

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

Lemma 10 (In-sample error estimation of model-based method)**.** *Suppose that* $\ell_h^s(f) = \mathbb{E}_{\pi_{f^s}} D_{\mathrm{H}}^2\big(\mathbb{P}_{h,f}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h)\big)$*. We set* $m = 1$ *so* $T = K$ *in this case. For each* $h \in [H]$*, we can take* $L_h^{1:t}(f) := \sum_{s=1}^t L_h^s(f) := \frac{1}{2}\sum_{s=1}^t -\log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s)$*. However, implementation with* $L_h^s(\cdot)$ *does not lead to a condition as required in Definition 4 for* $L_h^s(\cdot)$ *itself. The key observation is that adopting* $L_h^s(\cdot)$ *is equivalent to implementing Algorithm 1 with* $\widetilde{L}_h^s(f) = -\frac{1}{2}\log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s) + \frac{1}{2}\log \mathbb{P}_{h,f^*}(x_{h+1}^s \mid x_h^s, a_h^s)$ *because we subtract the same amount of loss for all hypotheses. Therefore, we can use it in our theoretical analysis, and it satisfies equation 14 and equation 15 with* $\Delta_h^t = \log(H|\mathcal{H}|/\delta)$*.*

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

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

$$f^k := \operatorname*{argmin}_{f \in \mathcal{H}} L_h^{1:k-1}(f). \tag{16}$$

The model-free version is referred to as the fitted Q-iteration (FQI) in the literature Szepesvári (2010). We may call equation 16 as fitted ℓ-iteration (FLI) since the target here is the loss function ℓ in the definition of GEC. FLI picks the hypothesis that best approximates the ground truth. Unfortunately, this does not ensure a low regret bound or a mild sample complexity. This is because a low cumulative prediction error does not necessarily lead to a low regret because of the difference between $V_{1,f}$ and V_1^*.

5.2 Optimistic FLI

The idea is to modify the FLI by adding another "feel-good" term to the objective:

$$f^t = \operatorname*{argmax}_{f \in \mathcal{H}} \Big[V_{1,f}(x_1) - \eta \sum_{s=1}^{t-1}\sum_{h=1}^{H} L_h^s(f)\Big], \tag{17}$$

where $\eta > 0$ is a hyper-parameter to control the relative importance of the feel-good term. The technical consideration of this modification is inspired by Zhang (2022). The complete algorithm is presented in Algorithm 1.

Algorithm 1 Optimistic FLI

1: **Input:** Hypothesis space $\mathcal{H}$, $\eta > 0$, batch size $m > 0$.
2: **for** $k = 1, 2, \cdots, K := T/m$ **do**
3: Select f^k by solving

$$f^k = \operatorname*{argmax}_{f \in \mathcal{H}} \Big[V_{1,f}(x_1) - \eta \sum_{h=1}^{H} L_h^{1:k-1}(f)\Big]. \tag{18}$$

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

Comparison to existing OFU-based algorithms. The algorithm can be viewed as a unification of the BiLin-UCB Du et al. (2021), GOLF Jin et al. (2021a), and OMLE Liu et al. (2022), where the generality is mainly from the flexible choice of the loss function. However, the main difference is that we do not explicitly maintain a confidence set and perform a constraint optimization subroutine on it:

$$\begin{aligned} &\max_{f \in \mathcal{F}} V_{1,f} \\ &\text{subject to } \sum_{h=1}^{H} L_h^{1:k-1}(f) \le \beta^k, \end{aligned} \tag{19}$$

where $\beta^k > 0$ is the confidence level. To the best of our knowledge, such an implicit formulation is new in the literature of online RL and we believe that it helps to better illustrate the role of optimism in the algorithmic design. We mention in passing that Xie et al. (2021) has a similar algorithmic idea in the pessimism-based offline setting. Combining this with the in-sample error estimation bound, we can establish the main result of this paper.

Theorem 2. *Under Assumption 1, we consider the problem with a low GEC $d(\epsilon)$ and the loss estimator given in Definition 4 with estimation interval Δ_h^k. Then, with $\eta = \sqrt{d(\epsilon)/(2\sum_{h=1}^H \sum_{k=1}^K \Delta_h^k)} > 0$, Algorithm 1 satisfies that with probability at least $1-\delta$,*

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

where $T = mK$.

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

$$\begin{aligned}
&\sum_{k=1}^K V_1^* - V_1^{\pi^k} := \sum_{k=1}^K [V_{1,f^k} - V_1^{\pi^k} + V_1^* - V_{1,f^k}] \\
&\le \sum_{k=1}^K [V_1^* - V_{1,f^k}] + \eta \sum_{h=1}^H\sum_{k=1}^K \Big(\sum_{s=1}^{k-1} \ell_h^s(f^k)\Big) + \underbrace{\frac{1}{\eta}\cdot d + 2\min\{Hd, H^2K\} + \epsilon B^\dagger K}_{\Xi} \qquad \text{//GEC} \\
&\le \sum_{k=1}^K V_1^* + \sum_{k=1}^K \big[-V_{1,f^k} + \eta \sum_{h=1}^H L_h^{1:k-1}(f^k)\big] + \eta\sum_{h=1}^H\sum_{k=1}^K \Delta_h^k + \Xi \qquad \text{//Loss estimator.} \\
&\le \sum_{k=1}^K V_1^* + \sum_{k=1}^K \big[-V_1^* + \eta \sum_{h=1}^H L_h^{1:k-1}(f^*)\big] + \eta\sum_{h=1}^H\sum_{k=1}^K \Delta_h^k + \Xi \qquad \text{//Update rule.} \\
&\le 2\eta\sum_{h=1}^H\sum_{k=1}^K \Delta_h^k + \frac{1}{\eta}\cdot d + 2\min\{Hd, H^2K\} + \epsilon B^\dagger K.
\end{aligned}$$

Now we optimize the last equation with respect to $\eta > 0$ and take $\eta = \sqrt{d/(2\sum_{h=1}^H \sum_{k=1}^K \Delta_h^k)} > 0$:

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

Since for each iteration, we sample m trajectories in total, the final regret is given by:

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

□

5.3 Posterior Sampling

While the optimization-based framework Algorithm 1 provides a satisfactory statistical guarantee, it is known that posterior sampling usually outperforms this type of algorithms empirically, including Chapelle & Li (2011) for bandits, and Osband et al. (2016) for RL. We can extend the idea to

the posterior sampling, which is presented in Algorithm 2. For instance, the Conditional Posterior Sampling presented in Dann et al. (2021) adopt the following posterior:

$$L_h^{1:t}(f) = -\eta \sum_{s=1}^{t} [Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 \\ - \log \mathbb{E}_{\widetilde{f}_h \sim p_h^0(\cdot)} \left[\exp\left(-\eta \sum_{s=1}^{t} [Q_{h,\widetilde{f}}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 \right) \right], \tag{20}$$

where $\eta > 0$ is a hyper-parameter and the highlighted term can be viewed as the soft version in the minimax formulation and is used to cancel the sampling variance. Another model-based version (Zhong et al., 2022) takes

$$L_h^{1:t}(f) := \eta \sum_{s=1}^{t} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s). \tag{21}$$

The analyses of the posterior sampling are developed in Zhang (2022); Dann et al. (2021); Zhong et al. (2022) and lead to matching upper bounds as compared to Algorithm 1, and we refer the interested readers to these works for details.

In comparison, the sampling-based framework requires knowledge of the environment by assuming that we have access to a good prior distribution that approximately satisfies the Assumption 1 and the Assumption 2. In contrast, the optimization-based framework presented in this paper does not require knowledge of the environment. However, as we mentioned before, while two types of algorithms achieve similar theoretical guarantees, empirical studies show that sampling-based algorithms are usually superior in practice. We have some preliminary experimental results for the optimization-based framework in Liu et al. (2023) and leave the posterior sampling for future work.

Algorithm 2 Optimistic Posterior Sampling

1: **Input:** Hypothesis space $\mathcal{H}$, $\gamma > 0$, batch size $m > 0$, and a prior distribution p_0.
2: **for** $k = 1, 2, \cdots, K := T/m$ **do**
3: Sample f^k from

$$p^k(f) \propto \underbrace{p^0(f) \exp(\gamma V_{1,f})}_{\text{Feel-good prior}} \cdot \exp(L_h^{1:k-1}(f)). \tag{22}$$

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

6 Related work

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

Tabular MDP. For a tabular MDP, we assume that the state space $\mathcal{S}$ and action space $\mathcal{A}$ are small. But we do not impose any structural assumption across states and actions. The goal in the tabular case is to design algorithms that achieve a regret depending polynomially on S, A and also the horizon H. The tabular MDP has been extensively studied in the literature Auer et al. (2008); Azar et al. (2017); Dann et al. (2017); Jin et al. (2018); Agrawal & Jia (2017); Zanette & Brunskill (2019); Zhang et al. (2020; 2021); Ménard et al. (2021); Li et al. (2021); Wu et al. (2022); Zhang et al. (2022). Among them, Azar et al. (2017) designs a model-based algorithm UCB-VI that explicitly models the transition matrix of the MDP, and attains the minimax-optimal regret bound $\mathcal{O}(\sqrt{H^2SAT})$. After this, Jin et al. (2018) proposes an optimistic variant of Q-learning UCB-B, which is model-free because it directly learns the optimal Q-value instead of the model dynamic, and attains a regret of $\mathcal{O}(\sqrt{H^3SAT})$ with Bernstein-type bonus. This is later improved by Zhang et al. (2020) to close the gap to the lower bound by leveraging the idea of variance reduction by a reference function Johnson & Zhang (2013). In comparison, the model-free algorithms typically require less

time space and storage space as compared to the model-based counterparts. Since both the model-based and model-free algorithms attain the minimax-optimal regret bound, the tabular settings are well-studied. However, since the lower bound depends on the $\sqrt{SA}$, we cannot handle modern RL problems with large state space without further structural assumptions.

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

Linear function approximation and learnability. Linear function approximation is arguably the most fundamental one Wang et al. (2019); Yang & Wang (2019); Cai et al. (2020); Jin et al. (2020); Zanette et al. (2020); Ayoub et al. (2020); Modi et al. (2020); Zhou et al. (2021); Zhong & Zhang (2023); Agarwal et al. (2022); He et al. (2022) in function approximation. Typically, we will assume that we have access to a d-dimensional feature map of the state-action pair $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. A natural idea is to assume that the optimal Q-value Q^* is linear in this feature, where we refer it as the *linear Q^* condition*, in the sense that there exists a $\theta_h \in \mathbb{R}^d$ and $||\theta_h^*|| \le B$:

$$Q_h^*(x,a) := \langle \phi(x,a), \theta_h^* \rangle, \qquad \forall h \in [H]. \tag{23}$$

The main technical consideration is that due to the extra linear structure, we can generalize from the visited states to the unseen states by the standard analysis of linear regression (in comparison, the tabular MDP does not impose any structure across states and actions so such a generalization is impossible). Then, we can approximate the Q^* by $\mathcal{H} := \{Q_h(x,a) = \langle \phi(x,a), \theta_h \rangle : ||\theta_h|| \le B, h \in [H]\}$, which satisfies the realizability assumption (Assumption 1). Unfortunately, there exists a negative result Wang et al. (2021), saying that this is not sufficient for sample-efficient learning.

Proposition 2 (Linear-realizability is not sufficient Wang et al. (2021))**.** *There exists an MDP with feature map ϕ that satisfies equation 23 but any algorithms must have*

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

To bypass this hardness result, Jin et al. (2020) imposes a stronger assumption that both the transition kernel and the reward function are linear in the feature, where we refer to this condition as the linear MDP (Example 3). Jin et al. (2020) designs an optimistic variant of LSVI (Least Squares Value Iteration), referred to as the LSVI-UCB, that achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{d^3H^4T})$. This result is recently improved by Agarwal et al. (2022); He et al. (2022) to $\widetilde{\mathcal{O}}(\sqrt{d^2H^3T})$, which matches the minimax lower bound in Zhou et al. (2021). Another line of work Ayoub et al. (2020); Modi et al. (2020); Cai et al. (2020); Zhou et al. (2021) study the linear mixture MDP, where the transition kernel is a linear mixture of a number of basis kernels and also has designed algorithms based on linear regression to achieve the minimax-optimal regret bound. As we have seen in the main text, the linear MDP and the linear mixture MDP essentially limit the generalization from the historical data samples to the newly arrived trajectory by limiting the freedom of the transition kernels and reward functions. Such a generalization ability is $\widetilde{O}(d)$ in the complexity measure proposed in this paper. We note that both the linear MDP and linear mixture MDP are rather strong assumptions. There are also some works that consider linear realizability different from equation 23. For instance, Du et al. (2019) assumes that the Q-value of any policy is linear:

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

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

Sample-efficient RL with general function approximation. While the linear MDP assumption permits sample-efficient learning, it is rather limited in practice. A long line of work extends to

the general *non-linear* function approximation and designs algorithms to effectively solve these RL problems. Generally speaking, these works can be largely grouped into two categories (i) impose certain low-rank structures so that some results in the linear MDP can be generalized; (ii) limit the sequence length of effective state (and/or action) distributions with respect to the hypothesis space. We now review them as follows.

A line of work imposes a low-rank structure on certain parts of the RL problems. The seminal work Jiang et al. (2017b) considers the model-free approach with value-based hypothesis $\mathcal{F}$ and proposes the *Bellman rank*, which is imposed on the class of induced Bellman residual: $\{\mathbb{E}_{\pi_f}\mathcal{E}_h(g, x_h, a_h) : f, g \in \mathcal{F}\}$[2]. If we view this set as a matrix of $\mathbb{R}^{|\Pi|\times|\mathcal{F}|}$ with $\Pi := \{\pi_f : f \in \mathcal{F}\}$, the Bellman rank d is the rank of the matrix, maximized over $h \in [H]$. In this case, the linearly independent rows of the matrix one can find are at most d, thus limiting the generalization of the problems. More generally, one can extend the finite-dimensional case by considering a bilinear structure. Specifically, we assume that we have two unknown embeddings $W_h : \mathcal{F} \to \mathcal{V}$ and $X_h : \mathcal{F} \to \mathcal{V}$ where $\mathcal{V}$ is a Hilbert space such that

$$\mathbb{E}_{\pi_f}\mathcal{E}_h(g, x_h, a_h) = \langle W_h(g) - W_h(f^*), X_h(f)\rangle\,, \tag{24}$$

where the Bellman rank can be described in terms of the information gain Srinivas et al. (2009). Jiang et al. (2017b) also proposes OLIVE, based on the OFU principle and hypothesis elimination, to solve the problems with a low Bellman rank. After Bellman rank and OLIVE, Sun et al. (2019) shows that there exists an exponential separation between the model-free approach and model-based approach, in the sense that the Bellman rank of factored MDPs Kearns & Koller (1999) can be exponentially large. Alternatively, Sun et al. (2019) extends the idea of Bellman rank to the model-based setting, and proposes the *witness rank*, to capture the factored MDPs. Du et al. (2021) generalizes them by proposing the bilinear class, which assumes that the average Bellman error and a *discrepancy loss* have a special bilinear structure. Moreover, the bilinear class allows a flexible choice of "discrepancy function" to capture both the model-free and model-based problems. Du et al. (2021) also proposes BiLin-UCB, which is more similar to the optimism-based algorithms that have been analyzed in contextual bandit Dani et al. (2008); Li et al. (2010); Abbasi-Yadkori et al. (2011). Specifically, BiLin-UCB maintains a confidence set $\mathcal{H}^t$ at each iteration where $f^* \in \mathcal{H}^t$ with high probability. Then, the agent chooses the estimator with the highest value estimations (that is why we say it is optimistic) such that the estimation is higher than that of the ground truth to encourage exploration:

$$f^t = \operatorname*{argmax}_{f\in\mathcal{H}^t} V_{1,f}(x_1). \tag{25}$$

Another line of work focuses on explicitly limiting the length of the longest sequence of effective distributions with respect to the hypothesis space. Russo & Van Roy (2013) proposes the notion of the eluder dimension, which generalizes the notion of linear independence in $\mathbb{R}^d$. The eluder dimension is later leveraged to RL by Wang et al. (2020), which includes the linear MDP Jin et al. (2020) as a special example. However, Wang et al. (2020) only characterizes the eluder dimension of the hypothesis space $\mathcal{H}$, and the covered problems are rather limited. Jin et al. (2021a) further considers the distributional eluder dimension on the induced Bellman residual space (referred to as the Bellman eluder dimension) so that the eluder dimension is imposed on the interplay between the function class and the underlying MDP, and captures more RL problems. Jin et al. (2020) also proposes a model-free OFU-based algorithm, GOLF, which is also based on confidence sets and optimism equation 25. The main difference between BiLin-UCB and GOLF is that GOLF additionally assumes the Bellman completeness condition (see Assumption 2) thus leveraging a minimax formulation Antos et al. (2008) to give a more efficient estimation of the Bellman errors (see Chapter 5 for a detailed interpretation of these algorithmic choices). Consequently, GOLF achieves a $\sqrt{T}$-regret, while BiLin-UCB only achieves a $T^{2/3}$-regret (by online-to-bach conversion). We also note that Chen et al. (2022b) generalizes the Bellman completeness assumption to the more general discrepancy function and applies the minimax formulation to achieve a better regret bound.

It is known that neither the Bellman eluder dimension nor bilinear class captures each other (see comments on page 6 of Du et al. (2021)). Attempts have been made since then to unify these two

[2]The example presented here is referred to as the Q-type Bellman rank, which is different from the V-type one considered in Jiang et al. (2017b). We choose the Q-type one for a clear presentation.

rich tackable RL problems. Dann et al. (2021) introduces the notion of the *eluder coefficient*[3] and studies the Q-type model-free problems. The eluder coefficient $d(\mu)$ explicitly relates the *out-of-sample* average Bellman residual to the *in-sample* average (squared) Bellman error:

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\pi_{f^t}}\mathcal{E}_h(f^t,x_h,a_h)\le\mu\sum_{h=1}^{H}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}[\mathcal{E}_h(f^t,x_h,a_h)^2]+\frac{d(\mu)}{4\mu}. \tag{26}$$

Intuitively, the eluder coefficient quantifies the rate at which the prediction error can grow in comparison to the cumulative training error on average in an online manner, thus explicitly limiting the generalization from the visited state-action distributions to the unseen part. Technically, we note that the eluder coefficient serves to reduce the RL problems into an in-sample error estimation problem over the hypothesis space $\mathcal{H}$ in a supervised learning manner, which is relatively well studied in the literature. Our complexity measure is mostly related to this eluder coefficient, but with several important extensions so that it can unify most of the existing complexity measures. Dann et al. (2021) also proposes conditional posterior sampling with an optimistic modification in the prior, and develops new analysis techniques for sampling-based algorithms.

There is also another line of work with distinct technical considerations compared to the above-mentioned works. Foster et al. (2021; 2022) propose the decision estimation coefficient (DEC) to unify the complexity measures in interactive decision-making, which takes the MDPs as a special example. Given a model class $\mathcal{M}$ and a reference model $\widehat{M}$, the DEC is given by

$$\mathrm{dec}_\gamma(\mathcal{M},\widehat{M})=\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}[\underbrace{V_{1,M}(x_1)-V_{1,M}^{\pi}(x_1)}_{\text{regret of decision}}-\gamma\cdot\underbrace{D^2_{\mathrm{H}}(M(\pi),\widehat{M}(\pi))}_{\text{Easy to control}}], \tag{27}$$

where $V_{1,M}^{\pi}(x_1)$ is the V-value of policy π when M is the underlying model and $M(\pi)$ denotes the trajectory distribution jointly determined by the model M and the executed policy π. The technical consideration is to convert the RL problems into an online learning problem, by reducing the out-of-sample regret to another out-of-sample divergence $D^2_{\mathrm{H}}(M(\pi),\widehat{M}(\pi))$. We note such a technical treatment also arises in the decoupling coefficient proposed in Zhang (2022); Agarwal & Zhang (2022a;b). The idea of such a conversion may date back to the *information ratio* Russo & Van Roy (2014). DEC is a more general complexity measure that captures both the bilinear class and Bellman eluder dimension. The most appealing part of DEC is the matching lower bound in terms of DEC in some decision-making problems, which suggests that a low DEC is necessary for sample-efficient learning. However, the vanilla DEC equation 27 cannot be applied in a model-free manner. To address this issue, Foster et al. (2022) adopt an optimistic modification as in Zhang (2022), and extend it to the model-free scenario. However, the regret bound of the model-free E2D obtained in Foster et al. (2022) is inferior compared to that of Du et al. (2021) under only realizability. We mention in passing that Chen et al. (2022a) also independently studies the optimistic variant of DEC. A notable feature of DEC is that it involves a minimax operation, which accounts for the minimax subroutine in their algorithms, Estimation-to-Decisions (E2D). Such a black box minimax subroutine may lack efficient computation guidelines in practice and the DEC may not be applied to analyze the classic OFU-based or sampling-based algorithms. In comparison, the decoupling coefficient Zhang (2022); Agarwal & Zhang (2022a) does not require solving a minimax problem, and can be applied to analyze the optimistic algorithms based on posterior sampling or Maximum Likelihood Estimation (MLE).

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

REFERENCES

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

[3]The eluder coefficient is referred to as the decoupling coefficient in Dann et al. (2017). However, we note that the decoupling coefficient proposed in Zhang (2022) is distinctly different from the eluder coefficient in terms of intuition and technical consideration so we use the name eluder coefficient here.

Alekh Agarwal and Tong Zhang. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022a.

Alekh Agarwal and Tong Zhang. Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling. *arXiv preprint arXiv:2203.08248*, 2022b.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. *arXiv preprint arXiv:2212.06069*, 2022.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

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

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

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

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

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022a.

Zixiang Chen, Chris Junchi Li, Angela Yuan, Quanquan Gu, and Michael I Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022b.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

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

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

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

Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*, 2022.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022.

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. 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*, pp. 1704–1713. PMLR, 06–11 Aug 2017b.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 18–24 Jul 2021b.

ChenM Jinglin and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pp. 740–747, 1999.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022.

Zhihan Liu, Miao Lu, Wei Xiong, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, and Zhaoran Wang. One objective to rule them all: A maximization objective fusing estimation and planning for exploration. *arXiv preprint arXiv:2305.18258*, 2023.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pp. 7609–7618. PMLR, 2021.

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

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

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

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.

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

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

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

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021.

Tianhao Wu, Yunchang Yang, Han Zhong, Liwei Wang, Simon Du, and Jiantao Jiao. Nearly optimal policy optimization with stable at any time guarantee. In *International Conference on Machine Learning*, pp. 24243–24265. PMLR, 2022.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

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

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

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

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

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pp. 3858–3904. PMLR, 2022.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841*, 2023.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

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