Uniform-PAC Guarantees for Model-Based RL with Bounded Eluder Dimension

Yue Wu[1] **Jiafan He**[1] **Quanquan Gu**[1]

[1]Computer Science Department , University of California, Los Angeles, California, USA

Abstract

Recently, there has been remarkable progress in reinforcement learning (RL) with general function approximation. However, all these works only provide regret or sample complexity guarantees. It is still an open question if one can achieve stronger performance guarantees, i.e., the uniform probably approximate correctness (Uniform-PAC) guarantee that can imply both a sub-linear regret bound and a polynomial sample complexity for any target learning accuracy. We study this problem by proposing algorithms for both nonlinear bandits and model-based episodic RL using the general function class with a bounded eluder dimension. The key idea of the proposed algorithms is to assign each action to different levels according to its width with respect to the confidence set. The achieved Uniform-PAC sample complexity is tight in the sense that it matches the state-of-the-art regret bounds or sample complexity guarantees when reduced to the linear case. To the best of our knowledge, this is the first work for Uniform-PAC guarantees on bandit and RL that goes beyond linear cases.

1 INTRODUCTION

Designing efficient algorithms to learn and plan in the sequential decision-making environment modeled by a Markov decision process (MDP) is one of the main tasks in reinforcement learning (RL). However, traditional tabular RL algorithms suffer from the curse-of-dimensionality due to the large size of the state and action spaces in practice. To enable learning in high-dimensional state and action spaces, using a predefined function class to approximate the underlying transition dynamic or the value function is a common approach. Most existing works for RL with function approximation focus on simple linear function classes such as the linear mixture MDP [Modi et al., 2020, Ayoub et al., 2020, Zhou et al., 2021b], which can replace the size of the state and action spaces with the dimension of the linear function class. However, these assumptions are often too restrictive to hold in practice. Recently, a line of works [Russo and Van Roy, 2013, Du et al., 2021, Jin et al., 2021] emerged that studies RL with general function approximation, introducing new complexity measures for the general function class and proposing new algorithms with regret bounds or PAC guarantees in terms of the complexity of the general function class. All existing results of RL with a general function class are limited to either regret bounds or PAC sample complexity, both of which cannot ensure convergence to the optimal policy up to arbitrary accuracy. To address this, Dann et al. [2017] proposed the strongest possible performance measure, the Uniform-PAC guarantee, which provides upper bounds on the number of ϵ-suboptimal episodes for any accuracy parameter $\epsilon > 0$ uniformly. They also proposed the UBEV algorithm with $\widetilde{O}(SAH^4/\epsilon^2)$ Uniform-PAC sample complexity, which guarantees convergence to the optimal policy for any $\epsilon > 0$. Recently, He et al. [2021] proposed the FLUTE algorithm for RL with linear function approximation, which has an $\widetilde{O}(d^3H^5/\epsilon^2)$ Uniform-PAC sample complexity, where d is the dimension of the linear function class.

In this paper, we propose new algorithms for both the bandit problem and model-based RL problem with a general function class, focusing on a model-based RL problem called linear mixture MDPs [Ayoub et al., 2020]. Our approach uses the eluder dimension [Russo and Van Roy, 2013] as a complexity measure for the general function class $\mathcal{F}$, which generalizes the linear independence relation in the standard vector space to capture the nonlinear independence in the function space approximately. The core of our algorithms is to assign each action to different levels based on its width with respect to the confidence set. For function classes with a bounded eluder dimension, we prove that both algorithms have Uniform-PAC guarantees and near-optimal sample

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

complexity bounds. Our key contributions are highlighted below.

- In the nonlinear bandit problem, where the reward function f is from a known class $\mathcal{F}$ with bounded eluder dimension, we propose the $\mathcal{F}$-UPAC-OFUL algorithm that achieves an $\widetilde{O}(d_K d_E/\epsilon^2)$ Uniform-PAC guarantee. Here d_K relates to the metric entropy (log-covering number) of $\mathcal{F}$, d_E is its eluder dimension, and ϵ is the accuracy parameter. Our result implies a $\widetilde{O}(\sqrt{d_K d_E K})$ regret guarantee in the first T rounds, matching the result in Russo and Van Roy [2013] up to a logarithmic factor. This is the first Uniform-PAC guarantee for nonlinear bandits.
- We also consider a model-based episodic RL problem where the transition probability P^* belongs to a known family $\mathcal{P}$. We propose the $\mathcal{F}$-UPAC-VTR algorithm and obtain a $\widetilde{O}(H^3 d_K d_E/\epsilon^2)$ Uniform-PAC sample complexity, where H is the horizon length, d_K relates to the metric entropy and d_E is the eluder dimension for the value function class $\mathcal{F}$ induced by the transition probability family $\mathcal{P}$. This result also implies a $\widetilde{O}(\sqrt{d_K d_E H^3 K})$ regret bound in the first K episodes, matching the result of the UCRL-VTR algorithm in [Ayoub et al., 2020] up to a logarithmic factor.

For the ease of comparison, we list the results of our algorithms and the most related algorithms in Tables 1 and 2.

2 RELATED WORK

2.1 RL WITH LINEAR FUNCTION APPROXIMATION

There is a body of research on learning Markov Decision Processes (MDPs) under the assumption of linear function approximation, which can be divided into model-free and model-based approaches. For model-free algorithms, Jin et al. [2019] studied the linear MDP model, where the reward function $r(s,a)$ and transition probability function $\mathbb{P}(s'|s,a)$ are linear with respect to a given feature map $\boldsymbol{\phi}(s,a)$. The LSVI-UCB algorithm proposed by Jin et al. [2019] achieved an $O(\sqrt{d^3 H^3 K})$ regret bound. Zanette et al. [2020] extended the linear MDP assumption to the low inherent Bellman error assumption, where the Bellman backup can be approximated by a linear function class.

For model-based algorithms, Jia et al. [2020] investigated linear mixture MDPs [Modi et al., 2020], where the transition probability function $\mathbb{P}(s'|s,a)$ is linear with respect to a given feature mapping $\boldsymbol{\phi}(s'|s,a)$. They proposed the UCRL-VTR algorithm, which has a regret guarantee of $\widetilde{O}(d\sqrt{H^3 K})$. Zhou et al. [2021a] improved the regret guarantee to $\widetilde{O}(dH\sqrt{K})$ by introducing a Bernstein-type bonus in the linear mixture model.

Linear bandit problems, as a special case of RL with linear function approximation, have also been extensively studied. For infinite-arm linear bandit problems, Dani et al. [2008] obtained an $O(d\sqrt{K \log^3 K})$ regret guarantee with the Confidence Ball algorithm. Abbasi-Yadkori et al. [2011] proposed the OFUL algorithm, which improved the result to $O(d\sqrt{K \log^2 K})$.

2.2 RL WITH GENERAL FUNCTION APPROXIMATION

Recently, a line of research has emerged in RL with general function approximation to relax the linear function assumption [Russo and Van Roy, 2013, Jiang et al., 2017, Sun et al., 2019, Dong et al., 2020, Yang et al., 2020, Wang et al., 2020, Ayoub et al., 2020, Jin et al., 2021, Du et al., 2021]. Jiang et al. [2017] introduced the Bellman rank, a complexity measure that depends on the function class $\mathcal{F}$ and the roll-in policy, and proposed the OLIVE algorithm for models with low Bellman rank, which has a polynomial PAC-bound guarantee. The AVE algorithm by Dong et al. [2020] was proposed for learning low-Bellman-rank MDPs and obtained the first $O(\sqrt{K})$-regret guarantee. Sun et al. [2019] extended the Bellman rank to the model-based setting and proposed a complexity measure called "witness rank", which is always no larger than the Bellman rank and obtained a polynomial PAC bound in terms of the "witness rank".

In contrast, the eluder dimension [Russo and Van Roy, 2013] measures the complexity of a function class $\mathcal{F}$ from a different perspective, indicating how effectively the underlying function $f \in \mathcal{F}$ can be deduced from the current information. UCB-type and Thompson sampling-type algorithms for bandit problems were proposed by Russo and Van Roy [2013], both obtaining a $\sqrt{K}$-regret guarantee for bandit models with low eluder dimension. Osband and Van Roy [2014] extended this notion to reinforcement learning models and proposed a post-sampling algorithm, while Wang et al. [2020] and Ayoub et al. [2020] extended it to more general model-free and model-based reinforcement learning problems. Under the assumption that the value function or the transition probability belongs to a function class $\mathcal{F}$ with low eluder dimension, both the model-free algorithm $\mathcal{F}$-LSVI [Wang et al., 2020] and the model-based algorithm UCRL-VTR achieve $O(\sqrt{K})$-regret guarantees.

Recently, Jin et al. [2021] extended the eluder dimension to the Bellman eluder dimension, which considers the general function class and possible MDP structures simultaneously. The low Bellman eluder dimension covers both the low eluder dimension and the low Bellman rank, and the GOLF algorithm by Jin et al. [2021] achieves both $O(\sqrt{K})$-regret and polynomial PAC-bound guarantees. However, the Bellman eluder dimension does not cover the model-based RL problem. Concurrently, Du et al. [2021] proposed the bilinear class with bounded effective dimension, which contains

Table 1: Comparison of algorithms for bandits with linear or general reward function in terms of sample complexity and regret. Note that $d = \widetilde{\Theta}(d_K) = \widetilde{\Theta}(d_E)$ in the linear case. "–" means the corresponding Uniform-PAC sample complexity is not available or not applicable. Our results are tight given the nearly matching upper and lower bounds.

Algorithm	Sample Complexity	Regret	Uniform-PAC	Reward Function
OFUL [Abbasi-Yadkori et al., 2011]	–	$\widetilde{O}(d\sqrt{K})$	✘	Linear
UPAC-OFUL [He et al., 2021]	$\widetilde{O}(d^2/\epsilon^2)$	$\widetilde{O}(d\sqrt{K})$	✓	Linear
Linear UCB [Russo and Van Roy, 2013]	–	$\widetilde{O}(\sqrt{d_K d_E K})$	✘	General
$\mathcal{F}$-UPAC-OFUL Our work	$\widetilde{O}(d_K d_E/\epsilon^2)$	$\widetilde{O}(\sqrt{d_K d_E K})$	✓	General
Lower bound [Lattimore and Szepesvári, 2018]	–	$\widetilde{\Omega}(d\sqrt{K})$	–	Linear

Table 2: Comparison of algorithms for model-based RL with linear or general function approximation in terms of sample complexity and regret. Note that $d = \widetilde{\Theta}(d_K) = \widetilde{\Theta}(d_E)$ in the linear case.

Algorithm	Sample Complexity	Regret	Uniform-PAC	Function Approximation
UCRL-VTR+ [Zhou et al., 2021a]	–	$\widetilde{O}(dH\sqrt{K})$	✘	Linear
UCRL-VTR [Ayoub et al., 2020]	–	$\widetilde{O}(\sqrt{H^3 d_K d_E K})$	✘	General
$\mathcal{F}$-UCRL-VTR Our work	$\widetilde{O}(H^3 d_K d_E/\epsilon^2)$	$\widetilde{O}(\sqrt{H^3 d_K d_E K})$	✓	General
Lower bound [Zhou et al., 2021a]	–	$\widetilde{\Omega}(dH\sqrt{K})$	–	Linear

many MDP models such as low Bellman rank and low witness rank, and the BiLin-UCB algorithm with polynomial PAC bound. Notably, neither the bilinear class nor the low Bellman eluder dimension class contains each other.

All these works with general function approximation concern either a PAC sample complexity bound or a regret bound. This motivates us to provide a stronger guarantee on general function approximation, and we begin with the low eluder dimension as a first attempt.

3 PRELIMINARIES

Notation We use lower case letters to denote scalars, lower and upper case bold letters to denote vectors and matrices. We use $\|\cdot\|_\infty$ to indicate the infinity norm. For a probability measure $P(s)$ and a function $V(s)$ on the space $\mathcal{S}$, we use $\langle P, V\rangle$ or $\langle P(\cdot), V(\cdot)\rangle$ to denote the integral $\int_{\mathcal{S}} V(s)dP(s)$. We also use the standard O and Ω notations. We say $a_n = O(b_n)$ if and only if $\exists C > 0, N > 0, \forall n > N, a_n \leqslant Cb_n$; $a_n = \Omega(b_n)$ if $a_n \geqslant Cb_n$. The notation $\widetilde{O}$ is used to hide logarithmic factors.

3.1 EPISODIC MARKOV DECISION PROCESS

In this work, we consider the homogeneous, episodic Markov Decision Process (MDP). Each episodic MDP is denoted by a tuple $M\big(\mathcal{S}, \mathcal{A}, H, r(\cdot,\cdot), P^*(\cdot|\cdot,\cdot)\big)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the finite action space, H is the horizon length (i.e., length of each episode), $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function and $P^*(s'|s,a) : \mathcal{S}\times\mathcal{A} \to \Delta^{|\mathcal{S}|}$ is the transition probability function that denotes the probability for state s to transit to the next state s' given action a. A policy $\pi_h(\cdot|s) : \mathcal{S}\times[H] \to \Delta^{\mathcal{A}}$ is a function which maps a state s and the current step h to a distribution over the action space $\mathcal{A}$. In addition, for any policy π and step $h \in [H]$, we define the action-value function $Q_h^\pi(s,a)$ and value function $V_h^\pi(s)$ as the expected total reward from step h by following the policy π:

$$V_h^\pi(s) := \mathbb{E}_\pi\big[\textstyle\sum_{h'=h}^H r(s_{h'}, a_{h'})|s_h = s\big],$$
$$Q_h^\pi(s,a) := r(s,a) + \mathbb{E}_\pi\big[\textstyle\sum_{h'=h+1}^H r(s_{h'}, a_{h'})\big|s_h = s, a_h = a\big],$$

where $s_{h'+1} \sim P^*(\cdot|s_{h'}, a_{h'})$ and $a_{h'} \sim \pi_h(s_{h'})$. With this definition, the value function $V_h^\pi(s)$ and $Q_h^\pi(s,a)$ are bounded in $[0, H]$.

We define the optimal value function V_h^* and Q_h^* as $V_h^*(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$. For each step $h \in [H]$ and policy π, we have the following Bellman equation and Bellman optimality equation:

$$\begin{aligned} Q_h^\pi(s,a) &= r(s,a) + \langle P^*(\cdot|s,a), V_{h+1}^\pi(\cdot)\rangle, \\ Q_h^*(s,a) &= r(s,a) + \langle P^*(\cdot|s,a), V_{h+1}^*(\cdot)\rangle, \end{aligned} \tag{3.1}$$

where $V_{H+1}^\pi(s') = V_{H+1}^*(s') = 0$.

We study the online RL problem where the learning agent is given with $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$ and reward r but does not know the transition probability $P^* \in \mathcal{P}$, where $\mathcal{P}$ is a class of possible transition probabilities.

Generally speaking, the goal of an RL agent is to maximize the expected total reward over all K episodes. When P is known, the optimal policy is also known and computable via dynamic programming. Denote π_k as the policy the agent follows at episode k, the suboptimality gap incurred at episode k is defined as the difference between the optimal value function and value function for policy π_k: $\Delta_k := V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,1})$. With this notation, the pseudo-regret in first K episodes is

$$\text{Regret}(K) = \sum_{k=1}^K \Delta_k = \sum_{k=1}^K V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,1}).$$

Most works in the literature focus on establishing an upper bound on $\text{Regret}(K)$. In the next subsection, we introduce the notion of a stronger guarantee.

3.2 THE UNIFORM-PAC GUARANTEE

We say an algorithm is (ϵ,δ)-PAC, if for any $\epsilon, \delta \in (0,1)$, there exists a function $N(\epsilon,\delta)$ that is polynomial in ϵ^{-1} and $\log(\delta^{-1})$, such that

$$\mathbb{P}\big(\textstyle\sum_{k=1}^\infty \mathbb{1}\{\Delta_k > \epsilon\} \leqslant N(\epsilon,\delta)\big) \geqslant 1-\delta.$$

Here $N(\epsilon,\delta)$ is the sample complexity function. However, both the regret guarantee and the PAC guarantee have their limitations. For example, algorithms with sub-linear regret $o(K)$ in the first K episodes may suffer ϵ-suboptimality infinite times and fail to learn the optimal policy. For the PAC bound, it only controls the number of times that $\Delta_k > \epsilon$. And the algorithm may incur a smaller ϵ'-suboptimality gap for infinite times for $\epsilon' < \epsilon$.

To overcome these limitations, Dann et al. [2017] introduced a stronger notion of guarantee called Uniform-PAC, which provides PAC-guarantees for all accuracy parameter ϵ uniformly. More specifically, we say an algorithm is Uniform-PAC for some $\delta \in (0,1)$, if there exists a function $N(\epsilon,\delta)$ polynomial in ϵ^{-1} and $\log(\delta^{-1})$, such that

$$\mathbb{P}\big(\forall \epsilon > 0,\ \textstyle\sum_{k=1}^\infty \mathbb{1}\{\Delta_k > \epsilon\} \leqslant N(\epsilon,\delta)\big) \geqslant 1-\delta.$$

Dann et al. [2017] showed that the Uniform-PAC guarantee is *strictly stronger* than both the regret bound and the PAC-guarantee:

Theorem 3.1 (Theorem 3, Dann et al. [2017])**.** If an algorithm `Alg` is $N(\epsilon,\delta)$-Uniform-PAC with sample complexity $N(\epsilon,\delta) = \widetilde{O}(C_1/\epsilon + C_2/\epsilon^2)$, where C_1, C_2 are constant parameter for the algorithm `Alg` and only depend on $\text{poly}\big(S, A, H, \log(1/\delta)\big)$, then, the algorithm `Alg` has the following results:

- 1: `Alg` will converge to optimal policies with high probability at least $1-\delta$: $\mathbb{P}\big(\lim_{k\to+\infty}\Delta_k = 0\big) \geqslant 1-\delta$
- 2: Algorithm `Alg` is also (ϵ,δ)-PAC with the same sample complexity $\widetilde{O}(C_1/\epsilon + C_2/\epsilon^2)$ for all $\epsilon > 0$.
- 3: With probability at least $1-\delta$, for each $K \in \mathbb{N}$, the regret for `Alg` in the first K episodes is upper bounded by $\widetilde{O}\big(\sqrt{C_2 K} + \max\{C_1, C_2\}\big)$.

Due to the strong implication of the Uniform-PAC guarantee, one may wonder if Uniform-PAC is also achievable under the setting of general function approximation.

3.3 COMPLEXITY MEASURE OF A FUNCTION CLASS

To deal with general function class $\mathcal{F}$, we will use two complexity measures. The first one is covering number, which is formally defined as follow.

Definition 3.2. Suppose $\|\cdot\|$ is a norm on $\mathcal{F}$ and $\alpha > 0$. A α-covering with respect to $\|\cdot\|$ is a subset $\mathcal{G} \subseteq \mathcal{F}$, such that $\forall f \in \mathcal{F}, \exists g \in \mathcal{G}$, s.t. $\|f-g\| \leqslant \alpha$. The covering number $\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|)$ is the minimal cardinality of any α-covering of $\mathcal{F}$ with respect to $\|\cdot\|$.

For many function classes, the log-covering number (a.k.a. metric entropy) $\log\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|)$ is linear in the dimension of $\mathcal{F}$ and only logarithmic in α^{-1}. One such function class $\{f_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ has been discussed in Russo and Van Roy [2013], where $f_{\boldsymbol{\theta}}(\mathbf{x})$ is L-Lipschitz with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Theta} = [0,1]^d$. It can be shown that $\log\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty) \leqslant d\log(1+L/\alpha)$. Therefore, we make the following mild assumption.

Assumption 3.3. The metric entropy $\log\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty)$ of the function class $\mathcal{F}$ is bounded linearly by $\log(\alpha^{-1})$, i.e., $\log\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty) \leqslant d_K \log(\alpha^{-1})$.

We can view d_K as an upper bound on the Kolmogorov dimension of the function class $\mathcal{F}$ (see e.g., Osband and Van Roy [2014]for more details).

The other complexity measure is the eluder dimension, which is first proposed by Russo and Van Roy [2013] based on the concept of ϵ-independence:

Definition 3.4. An input $\mathbf{x} \in \mathcal{X}$ is ϵ-dependent on inputs $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$ with respect to $\mathcal{F}$ if any pair of function $f_1, f_2 \in \mathcal{F}$ satisfying $\sum_{i=1}^n \big(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)\big)^2 \leqslant$

ϵ^2 also satisfies $f_1(\mathbf{x}) - f_2(\mathbf{x}) \leqslant \epsilon$. $\mathbf{x}$ is ϵ-independent of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$ with respect to $\mathcal{F}$ if it is not ϵ-dependent on $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$.

Then the eluder dimension is formally defined as follows.

Definition 3.5. The *ϵ-eluder dimension* $\dim_E(\mathcal{F}, \epsilon)$ is the length d of the longest sequence of elements in $\mathcal{X}$ such that, for some $\epsilon' > \epsilon$, every element is ϵ'-independent of its predecessors.

4 UNIFORM-PAC BOUNDS FOR NONLINEAR BANDITS

4.1 THE BANDIT PROBLEM WITH GENERAL REWARD FUNCTIONS

In the nonlinear bandit problem, at each round $k \in \mathbb{N}$, the agent selects an action $\mathbf{x}_k$ from the action set $\mathcal{A}_k$, and then receives the reward $R_k = f_{\boldsymbol{\theta}*}(\mathbf{x}_k) + \eta_k$, where the true reward function $f_{\boldsymbol{\theta}*}$ is assumed to lie in a set of bounded, real-valued functions $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \mathcal{X} \to [0,1] | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ which is indexed by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. η_k is a conditionally unbiased 1-sub-Gaussian noise:

$$\forall k \in \mathbb{N}, \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda \eta_k} | \mathbf{x}_1, \eta_1, \mathbf{x}_2, \ldots, \mathbf{x}_k] \leqslant e^{\lambda^2/2}.$$

Formally speaking, an agent selecting $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ achieves Uniform-PAC with complexity $N(\epsilon, \delta)$ if and only if

$$\mathbb{P}\big(\forall \epsilon > 0,\ \textstyle\sum_{k=1}^{\infty} \mathbb{1}\{\Delta_k > \epsilon\} \leqslant N(\epsilon, \delta)\big) \geqslant 1 - \delta,$$

where $\Delta_k := \max_{\mathbf{x} \in \mathcal{A}_k} f_{\boldsymbol{\theta}*}(\mathbf{x}) - f_{\boldsymbol{\theta}*}(\mathbf{x}_k)$ denotes the sub-optimality gap at round k.

4.2 ALGORITHM

We first present an $\mathcal{F}$-UPAC-OFUL algorithm in Algorithm 1. The high-level idea of Algorithm 1 is to split the rounds into several disjoint sets $\mathcal{C}^l$, and apply the approach of optimistic exploration within each set.

More specifically, in each set $\mathcal{C}^l$, we use the information from the rounds k such that $k \in \mathcal{C}^l$ to construct the confidence set $\mathcal{F}^l \subseteq \mathcal{F}$ (Line 4), and try to find the most optimistic action (Lines 6-9), namely the action with largest upper confidence bound (UCB). Line 4 defines the confidence set $\mathcal{F}^l$ with:

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}^l}(f, \widehat{f}) &:= \textstyle\sum_{k \in \mathcal{C}^l} \big(f(\mathbf{x}_k) - \widehat{f}(\mathbf{x}_k)\big)^2, \\
\widehat{f}^l &:= \arg\min_{f \in \mathcal{F}} \textstyle\sum_{k \in \mathcal{C}^l} \big(f(\mathbf{x}_k) - R_k\big)^2, \\
\beta_t^l &:= 8 \log\big(\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta\big) \\
&\quad + 2\alpha t\big(8 + \sqrt{8 \log(4t^2/2^{-l}\delta)}\big), \qquad (4.1)
\end{aligned}$$

where the metric entropy can be replaced by its upper bound. When the action $\mathbf{x}_k$ is chosen, the algorithm needs to decide which level set the index k should be assigned to (Lines 10-14). The decision is made by the scale of the width $w_{\mathcal{F}^l}(\mathbf{x}_k)$ at Line 11:

$$w_{\mathcal{F}^l}(\mathbf{x}_k) := \sup_{f \in \mathcal{F}^l} f(\mathbf{x}_k) - \inf_{f \in \mathcal{F}^l} f(\mathbf{x}_k).$$

Once the index k is assigned to a particular set, the algorithm updates the total level S as the number of non-empty sets.

It can be shown that when $\mathcal{F}$ is a linear function class, i.e., $f_{\boldsymbol{\theta}*}(\mathbf{x}_k) = \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle$, the confidence set for level l becomes an ellipsoid $\{\boldsymbol{\theta} | \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^l\|_{\mathbf{V}^l} \leqslant \beta_t^l\}$, where $\mathbf{V}^l = \sum_{k \in \mathcal{C}^l} \mathbf{x}_k \mathbf{x}_k^\top$ is the covariance matrix of the contexts at level l. In this case, the width $w_{\mathcal{F}^l}(\mathbf{x})$ has a closed form $w_{\mathcal{F}^l}(\mathbf{x}) = 2\beta_T^l \|\mathbf{x}\|_{\mathbf{V}^l}$. This is exactly the bonus term used in LinUCB/OFUL, and He et al. [2021] uses this term as a criterion to assign the contexts to the appropriate level in order to achieve uniform PAC guarantee for UPAC-OFUL.

Algorithm 1 $\mathcal{F}$-UPAC-OFUL

1: Set $\mathcal{C}^l \leftarrow \varnothing, l \in \mathbb{N}$ and the total level $S = 1$
2: **for** round $k = 1, 2, ..$ **do**
3: **for** all level $l \in [S]$ **do**
4: Denote $\mathcal{F}^l = \{f | \mathcal{L}_{\mathcal{C}^l}(f, \widehat{f}^l) \leqslant \beta_{|\mathcal{C}^l|}^l\}$
5: **end for**
6: Receive the action set $\mathcal{A}_k$
7: Choose action
8: $\mathbf{x}_k \leftarrow \operatorname{argmax}_{\mathbf{x} \in \mathcal{A}_k} \sup_{f \in \bigcap_{l \in [S]} \mathcal{F}^l} f(\mathbf{x})$
9: Receive the reward R_k
10: Set level $l = 1$
11: **while** $w_{\mathcal{F}^l}(\mathbf{x}_k) \leqslant 2^{-l}$ and $l \leqslant S$ **do**
12: $l \leftarrow l + 1$
13: **end while**
14: Add the new element k to the set $\mathcal{C}^l$ and update $\mathcal{F}^l$ accordingly
15: Set the total level $S = \max_{l : |\mathcal{C}^l| > 0} l$
16: **end for**

4.3 MAIN RESULTS

Before presenting the main result for Algorithm 1, we explain how β_t^l in (4.1) is chosen. For each $l \in \mathbb{N}$, we set β_t^l as $\alpha = U_l^{-1}$, where U_l satisfies $U_l = 64 d_K d_E 4^l \log U_l/\delta$. Later in the proof, we will see that U_l serves as an upper bound on the cardinality of $\mathcal{C}^l$.

Theorem 4.1. Suppose $\mathcal{F}$ satisfies Assumption 3.3, and denote $\Delta_k := \max_{\mathbf{x} \in \mathcal{A}_k} f_{\boldsymbol{\theta}*}(\mathbf{x}) - f_{\boldsymbol{\theta}*}(\mathbf{x}_k)$ and $d_E := \dim_E(\mathcal{F}, \epsilon/2)$. Then there exists a constant c such that with probability $1 - 2\delta$, for all $\epsilon > 0$, Algorithm 1 satisfies

$$\textstyle\sum_{k=1}^{\infty} \mathbb{1}\{\Delta_k > \epsilon\} \leqslant c \cdot \frac{d_K d_E}{\epsilon^2} \log(\frac{d_K d_E}{\epsilon \delta}).$$

Optimality of the Result According to Theorem 3 in Dann et al. [2017], our result in Theorem 4.1 can be converted to the same regret bound as Russo and Van Roy [2013], i.e., $\widetilde{O}(\sqrt{d_K d_E K})$. Under the linear bandit setting, this result becomes $\widetilde{O}(d\sqrt{T})$ because $d_K = d_E = d$ and cannot be improved without additional assumptions. Another evidence is that Wagenmaker et al. [2022] provided a lower bound on fixed-epsilon PAC for linear bandits of $\Omega(d^2/\epsilon^2)$ (Theorem 2). Once again, Theorem 3 in Dann et al. [2017] can convert our result into theirs since Uniform-PAC covers both regret and PAC. Therefore, this suggests that our result is tight (See Table 1 for details).

Computational Efficiency Generally speaking, the most computationally expensive step is to compute $\arg\sup_{f\in\mathcal{F}'} f(\mathbf{x})$ and $w_{\mathcal{F}^l}(\mathbf{x})$, that is to find the optimum function on some given input $\mathbf{x}$ within the confidence set $\mathcal{F}'$ (e.g., Line 8 or Line 11). While these optimization problems can be solved efficiently (or even analytically) for the linear function class, for general function class, the computational efficiency will be more subtly related to the structure of the function class, as well as the optimization algorithm for finding the maximizer.

5 UNIFORM-PAC BOUNDS FOR EPISODIC MDPS

5.1 ALGORITHM

As described in Section 3, we study the homogeneous episodic MDP, where the unknown, true transition probability P^* lies in a known family $\mathcal{P}$.

Following Ayoub et al. [2020], our results depend on the complexity of a function class $\mathcal{F}$ associated with $\mathcal{P}$. Let $\mathcal{V}$ be the set of optimal value functions under some transition probability in $\mathcal{P}$, that is $\mathcal{V} = \{V_h^{*,P}(\cdot)|h \in [H], P \in \mathcal{P}\}$. Note that any $V \in \mathcal{V}$ is positive and bounded by H. Let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$, we can see that any triplet $X_{k,h} = (s_{k,h}, a_{k,h}, V_{k,h+1}) \in \mathcal{X}$. The function class $\mathcal{F}$ is the collection of functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$\mathcal{F} := \big\{f_P(s,a,V) = \big\langle P(\cdot|s,a), V(\cdot)\big\rangle \big| P \in \mathcal{P}\big\}.$$

Note that any $f \in \mathcal{F}$ is positive and bounded by H because $P(\cdot|s,a)$ is a probability measure and $V(\cdot) \in [0,H]$. We also assume the metric entropy is linearly dependent on d_K as in Assumption 3.3.

We present an $\mathcal{F}$-UPAC-VTR algorithm in Algorithm 2. Similar to Algorithm 1, Algorithm 2 will maintain several disjoint sets $\mathcal{C}^l$, and construct the confidence set (Line 2) within each set. In particular, we adapt the algorithm design from Ayoub et al. [2020], and the confidence set $\mathcal{B}^l$ in Line 2 is defined using:

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}_k^l}(P,\widehat{P}) &:= \sum_{(k,h)\in\mathcal{C}_k^l} \big(\big\langle P(\cdot|s_{k,h},a_{k,h}) \\
&\qquad - \widehat{P}(\cdot|s_{k,h},a_{k,h}), V_{k,h+1}(\cdot)\big\rangle\big)^2, \\
\widehat{P}_k^l &:= \arg\min_{P\in\mathcal{P}} \sum_{(k,h)\in\mathcal{C}_k^l} \big(\big\langle P(\cdot|s_{k,h},a_{k,h}), V_{k,h+1}\big\rangle \\
&\qquad - V_{k,h+1}(s_{k,h+1})\big)^2, \\
\beta_t^l &:= 2H^2 \log\big(\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty)/\delta\big) \\
&\quad + 2\alpha t\big(8H + \sqrt{2H^2\log(4t^2/2^{-l}\delta)}\big).
\end{aligned} \tag{5.1}$$

Based on the confidence sets, Algorithm 2 performs the optimistic planning (Line 4-11) at the beginning of each episode k. In particular, Algorithm 2 first chooses an optimistic model P_k that belongs to the confidence sets of all levels (Line 5). As long as the true transition probability P^* belongs to the confidence sets, the optimistic optimal value function $V_1^{*,P_k}(s_{k,1})$ under P_k will serve as an upper bound on the true value function $V_1^*(s_{k,1})$.

Then, Algorithm 2 computes the optimal value function under the optimistic model via dynamic programming (Line 6):

$$\begin{aligned}
V_{k,H+1}(s) &= 0, \\
Q_{k,h}(s,a) &= r(s,a) + \big\langle P_k(\cdot|s,a), V_{k,h+1}(\cdot)\big\rangle, \\
V_{k,h}(s) &= \max_{a\in\mathcal{A}} Q_{k,h}(s,a).
\end{aligned} \tag{5.2}$$

Based on the value function, the policy π_k is implicitly defined as $\pi_{k,h}(s_{k,h}) = \mathrm{argmax}_{a\in\mathcal{A}} Q_{k,h}(s_{k,h},a)$. The algorithm will follow this policy (Line 8) and receive the corresponding reward and the next state. We denote the triplet of state, action, value function at step h in episode k as $X_{k,h} := (s_{k,h}, a_{k,h}, V_{k,h+1})$.

After the planning phase, Algorithm 2 enters the assigning phase (Line 12-18). At Line 14, we utilize the width $w_{\mathcal{B}^l}(X_{k,h})$ to determine which level l the index (k,h) should be assigned to. The diameter of a function class $\mathcal{B}$ on the triplet $X_{k,h}$ is defined as

$$\begin{aligned}
w_{\mathcal{B}}(X_k) := \sup_{P\in\mathcal{B}} &\big\langle P(\cdot|s_{k,h},a_{k,h}), V_{k,h+1}(\cdot)\big\rangle \\
&- \inf_{P\in\mathcal{B}} \big\langle P(\cdot|s_{k,h},a_{k,h}), V_{k,h+1}(\cdot)\big\rangle.
\end{aligned}$$

Once all indices are assigned and the confidence sets are updated (Line 17), the algorithm updates the maximum level S and repeats the next episode (Line 18).

5.2 MAIN RESULTS

For each $l \in \mathbb{N}$, we set β_t^l as $\alpha = U_l^{-1}$, where U_l satisfies $U_l = 64H^2 d_K d_E 4^l \log U_l/\delta$. In Lemma A.1 in Appendix, we will see that U_l serves as an upper bound on the cardinality of $\mathcal{C}^l$.

Algorithm 2 $\mathcal{F}$-UPAC-VTR

Require: Confidence radius $\beta_t^l(l, t \in \mathbb{N})$
Set $\mathcal{C}^l \leftarrow \varnothing, l \in \mathbb{N}$ and the total level $S = 1$
Denote $\mathcal{B}^l = \{P \in \mathcal{P} | \mathcal{L}_{\mathcal{C}^l}(P, \widehat{P}^l) \leqslant \beta^l_{|\mathcal{C}^l|}\}$
for episode $k = 1, 2, ..$ **do**
  Receive the initial state $s_{k,1}$
  Choose the optimistic model $P_k \leftarrow \operatorname{argmax}_{P \in \bigcap_{l \in [S]} \mathcal{B}^l} V_1^{*,P}(s_{k,1})$
  Compute value functions $Q_{k,h}$ and $V_{k,h}, h \in [H]$ for P_k according to Equation (5.2).
  for $h = 1, 2, \ldots, H$ **do**
    Choose the current action $a_{k,h} \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_{k,h}(s_{k,h}, a)$
    Receive the reward and the next state $s_{k,h+1}$
    Denote $X_{k,h} = (s_{k,h}, a_{k,h}, V_{k,h+1})$ for $h < H$
  end for
  for $h = 1, 2, \ldots, H-1$ **do**
    Set level $l = 1$
    while $w_{\mathcal{B}^l}(X_{k,h}) \leqslant H2^{-l}$ and $l \leqslant S$ **do**
      $l \leftarrow l + 1$
    end while
    Add the new element (k, h) to the set $\mathcal{C}^l$ (and update $\mathcal{B}^l$ accordingly)
    Set $S = \max_{l: |\mathcal{C}^l| > 0} l$
  end for
end for

Theorem 5.1. Suppose $\mathcal{F}$ satisfies Assumption 3.3, and denote $d_E := \dim_E(\mathcal{F}, \epsilon/8H)$. Then there exists a constant c such that with probability $1-3\delta$, for all $\epsilon > 0$, Algorithm 2 satisfies

$$\sum_{k=1}^{\infty} \mathbb{1}\left\{V_1^*(s_{k,1}) - V_{k,1}^{\pi_k}(s_{k,1}) > \epsilon\right\} \leqslant c \cdot \frac{H^3 d_K d_E \log\left(H^2 d_K d_E/(\epsilon\delta)\right)}{\epsilon^2}.$$

Relation to the Regret Bound in Ayoub et al. [2020] Our result $\widetilde{O}(H^3 d_K d_E \epsilon^{-2})$ can be converted into a regret bound of order $\widetilde{O}(\sqrt{H^3 d_K d_E K})$, where K is the total number of episodes. This matches the regret bound $\widetilde{O}(\sqrt{H^3 d_K d_E K})$ from Theorem 1 in Ayoub et al. [2020] up to logarithmic factors.
Additionally, Zhou et al. [2021a] shows that the regret bound of Ayoub et al. [2020] is nearly tight, off by a factor of $\sqrt{H}$ from optimal, in the setting of linear mixture MDPs. This suggests that our results under Uniform-PAC guarantee can no longer be improved in terms of d_K, d_E and K (see Table 2 for details).

5.3 POTENTIAL EXTENSIONS TO MORE GENERAL RL CLASSES

One may ask whether our results can be extended to more general RL classes such as Bellman eluder (BE) dimension [Jin et al., 2021] and bilinear classes [Du et al., 2021]. Here, we discuss the possibility of establishing Uniform-PAC guarantees for these RL classes.

First of all, we would like to clarify that our model-based setting cannot be covered by the BE dimension, because the BE dimension cannot cover linear mixture MDPs (see e.g., Figure 1 by Chen et al. [2022] for a detailed classification), let alone the nonlinear generalization of linear mixture MDPs considered in our work.

For low BE dimension Jin et al. [2021], the original paper considers two different distribution families, $\mathcal{D}_{\mathcal{F}}$ and $\mathcal{D}_{\Delta}$. $\mathcal{D}_{\mathcal{F}}$ generalizes the notion of Bellman rank, while $\mathcal{D}_{\Delta}$ generalizes the eluder dimension. The "low BE dimension" actually means one of the distribution families has a low BE dimension. The multi-level partition scheme can be applied to the GOLF algorithm proposed by Jin et al. [2021], in the sense that we can use the instance-wise Bellman error $|[f_h^k - \mathcal{T}_h f_h^{k+1}](s_h^k, a_h^k)| \in (2^{-l}, 2^{-l+1}]$ (here h stands for the h-th step and k stands for the k-th episode) as the criterion to assign step (k, h) to level l. By doing so, we can establish the Uniform-PAC guarantee in terms of the BE dimension of the single state-action pair distribution family $\mathcal{D}_{\Delta}$. On the other hand, it is hard to establish the Uniform-PAC guarantee for the family $\mathcal{D}_{\mathcal{F}}$, because we cannot access the expected Bellman error from a single sample. The difference between the expected Bellman error and the instance-wise Bellman error will cause error in the level assignment, which needs to be carefully controlled in order to achieve ϵ^{-2} sample complexity.

For bilinear class, Du et al. [2021] assumed the expected Bellman error for some hypothesis f has a bilinear form of $|\langle W_h(f) - W_h(f^*), X_h(f)\rangle|$. Intuitively, at episode k and step h, the quantity $\|X_h(f_k)\|_{\Sigma_{k;h}^{-1}}$ can be used as a criterion

for level assignment. However, in most cases, $X_h(f_k)$ is defined as an expectation over the stationary distribution (see e.g., Section 4.3 by Du et al. [2021]), which cannot be computed exactly and has to be estimated using a single sample. Therefore, we again face the challenge of controlling the estimation error of the criterion, which may need a more delicate technique to achieve ϵ^{-2} PAC bound.

We leave both extensions as future work.

6 PROOF OVERVIEW FOR NONLINEAR BANDITS

In this section, we show the proof of Theorem 4.1, which will illustrate the key idea of achieving Uniform-PAC guarantee for bandits and model-based RL with small eluder dimension.

We will use $\mathcal{C}_k^l$ to denote the level set $\mathcal{C}^l$ before insert the index k into any level set. $\mathcal{B}_k^l$ and $\widehat{P}_k^l$ are defined based on $\mathcal{C}_k^l$ accordingly. l_k is the level that the index k is assigned to.

The proof relies on the following two lemmas. The first lemma states that each level set only contains bounded number of indices.

Lemma 6.1. Suppose we set $\alpha = U_l^{-1}$ for $\beta_t^l = 8\log(\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha t(8 + \sqrt{8\log(4t^2/2^{-l}\delta)})$, where U_l is defined via the equality $U_l = 64 d_K d_E 4^l \log U_l/\delta$. Then for each level $l > 0$ and each round $k > 0$, the total number of actions in the set $\mathcal{C}_k^l$ is bounded as $|\mathcal{C}_k^l| < U_l$, which also means

$$|\mathcal{C}_k^l| < 128 d_K d_E 4^l \log(64 d_K d_E 4^l/\delta).$$

Here, $d_E = \dim_E(2^{-l})$.

The second lemma states the designed confidence set contains the true parameter with high probability.

Lemma 6.2. Suppose $\beta_t^l = 8\log(\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha t(8 + \sqrt{8\log(4t^2/2^{-l}\delta)})$, and $\mathcal{F}_k^l = \{f \in \mathcal{F} | \mathcal{L}_{\mathcal{C}_k^l}(f, \widehat{f}_k^l) \leqslant \beta_{|\mathcal{C}_k^l|}^l\}$ With probability $1 - 2\delta$, we have for all $k > 0$ and $l \in [S_k]$, $f^* \in \mathcal{F}_k^l$.

Now we proceed to prove the main result.

Proof of Theorem 4.1. Denote $\mathbf{x}_k^* = \operatorname{argmax}_{\mathbf{x}\in\mathcal{A}_k} f^*(\mathbf{x})$. Under the event of Lemma 6.2, we have

$$\begin{aligned}
\max_{\mathbf{x}\in\mathcal{A}_k} f^*(\mathbf{x}) - f^*(\mathbf{x}_k) &= f^*(\mathbf{x}_k^*) - f^*(\mathbf{x}_k) \\
&\leqslant \max_{f\in\bigcap_{l\in[S]}\mathcal{F}_k^l} f(\mathbf{x}_k^*) - \min_{f\in\mathcal{F}_k^{l_k}} f(\mathbf{x}_k) \\
&\leqslant \max_{f\in\mathcal{F}_k^{l_k}} f(\mathbf{x}_k) - \min_{f\in\mathcal{F}_k^{l_k}} f(\mathbf{x}_k) \\
&= w_{\mathcal{F}_k^{l_k}}(\mathbf{x}_k),
\end{aligned}$$

where the first inequality is by Lemma 6.2; the second inequality is by the definition of $\mathbf{x}_k$ that is the optimistic action. The last equality is by the definition of $w_{\mathcal{F}_k^{l_k}}(\cdot)$.

Therefore, by choosing the level l_0 such that $2^{-l_0} < \epsilon \leqslant 2^{-(l_0-1)}$, we have

$$\begin{aligned}
&\textstyle\sum_{k=1}^\infty \mathbb{1}\left\{\max_{a\in\mathcal{A}_k} f_{\boldsymbol{\theta}^*}(\mathbf{x}) - f_{\boldsymbol{\theta}^*}(\mathbf{x}_k) > \epsilon\right\} \\
&\leqslant \textstyle\sum_{k=1}^\infty \mathbb{1}\left\{w_{\mathcal{F}_k^{l_k}}(\mathbf{x}_k) > \epsilon\right\} \\
&\leqslant \textstyle\sum_{k=1}^\infty \mathbb{1}\left\{w_{\mathcal{F}_k^{l_k}}(\mathbf{x}_k) > 2^{-l_0}\right\} \\
&\leqslant \textstyle\sum_{k=1}^\infty \sum_{l=1}^{l_0} \mathbb{1}\left\{l_k = l\right\} \leqslant \sum_{l=1}^{l_0} U_l,
\end{aligned}$$

where the first line is by the inequality we established above; the second one is by $\epsilon > 2^{-l_0}$; the third inequality is by the design of the level set $\mathcal{C}^l$ and the last inequality is due to Lemma 6.1.

From Lemma 6.1, we can further bound $\sum_{l=1}^{l_0} U_l$ by

$$\begin{aligned}
&\sum_{l=1}^{l_0} 128 d_K \dim_E(2^{-l}) 4^l \log(64 d_K \dim_E(2^{-l}) 4^l/\delta) \\
&\leqslant 128 d_K \dim_E(\epsilon/2) \sum_{l=1}^{l_0} 4^l (\log(64 d_K \dim_E(\epsilon/2)/\delta) + 2l) \\
&\leqslant 128 d_K d_E (4^{l_0+1}\log(64 d_K d_E/\delta) + 4^{l_0+1} l_0) \\
&\leqslant c \cdot \frac{d_K d_E}{\epsilon^2} \log(d_K d_E/\epsilon\delta),
\end{aligned}$$

where the first line is due to Lemma 6.1;the second line holds because $2^{-l} \geqslant 2^{-l_0} \geqslant \epsilon/2$ (recall that $d_E = \dim_E(\epsilon/2)$); the third line relies on $\sum_{l=1}^{l_0} 4^l < 4^{l_0+1}$ and $\sum_{l=1}^{l_0} l 4^l < l_0 4^{l_0+1}$; the last line holds because $2^{-l_0} < \epsilon \leqslant 2^{-(l_0-1)}$, thus $2^{l_0} < 2\epsilon^{-1}$. c in the last line is some constant. □

7 CONCLUSION AND DISCUSSION

In this work, we consider online decision-making with general function approximations and propose two new algorithms for nonlinear bandits and episodic MDPs. With the help of the multi-level partition scheme, our $\mathcal{F}$-UPAC-OFUL algorithm and $\mathcal{F}$-UPAC-VTR algorithm obtain the Uniform-PAC guarantee to find the near-optimal policy with the state-of-the-art sample complexity. To the best of our knowledge, these results for the first time show that it is possible to achieve a Uniform-PAC guarantee in bandits and RL problems with general function approximation.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. YW, JH and QG are supported in part by the National

Science Foundation CAREER Award 1906169 and research fund from UCLA-Amazon Science Hub. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

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*, 2022.

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. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-PAC bounds for reinforcement learning with linear function approximation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

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

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*, pages 1704–1713. JMLR. org, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

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

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

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264. Citeseer, 2013.

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

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. *arXiv preprint arXiv:2201.11206*, 2022.

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, 2020.

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

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

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

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR, 2021b.