Toward General Function Approximation in Nonstationary Reinforcement Learning

Songtao Feng, Ming Yin, Ruiquan Huang, Yu-Xiang Wang, *Member, IEEE*, Jing Yang, *Senior Member, IEEE*, and Yingbin Liang, *Fellow, IEEE*

***Abstract*—Function approximation has experienced significant success in the field of reinforcement learning (RL). Despite a handful of progress on developing theory for nonstationary RL with function approximation under structural assumptions, existing work for nonstationary RL with general function approximation is still limited. In this work, we investigate two different approaches for nonstationary RL with general function approximation: confidence-set based algorithm and UCB-type algorithm. For the first approach, we introduce a new complexity measure called dynamic Bellman Eluder (DBE) for nonstationary MDPs, and then propose a confidence-set based algorithm SW-OPEA based on the complexity metric. SW-OPEA features the sliding window mechanism and a novel confidence set design for nonstationary MDPs. For the second approach, we propose a UCB-type algorithm LSVI-Nonstationary following the popular least-square-value-iteration (LSVI) framework, and mitigate the computational efficiency challenge of the confidence-set based approach. LSVI-Nonstationary features the restart mechanism and a new design of the bonus term to handle nonstationarity. The two proposed algorithms outperform the existing algorithms for nonstationary linear and tabular MDPs in the small variation budget setting. To the best of our knowledge, the two approaches are the first confidence-set based algorithm and UCB-type algorithm in the context of nonstationary MDPs.**

***Index Terms*—Nonstationary MDPs, general function approximation, Eluder dimension, LSVI.**

Manuscript received 29 October 2023; revised 6 February 2024; accepted 19 March 2024. Date of publication 29 March 2024; date of current version 3 May 2024. The work of Songtao Feng and Yingbin Liang was supported in part by the U.S. National Science Foundation under Grant RINGS-2148253, Grant DMS-2134145, and Grant CNS-2112471. The work of Ming Yin and Yu-Xiang Wang was supported in part by the National Science Foundation under Grant 2007117 and Grant 2003257. The work of Ruiquan Huang and Jing Yang was supported in part by the U.S. National Science Foundation under Grant CNS-1956276, Grant CNS-2003131, and Grant CNS-2030026. This work was presented in part at the Proceedings of the 40th International Conference on Machine Learning, 2023 [1]. *(Corresponding author: Songtao Feng.)*

Songtao Feng is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32603 USA (e-mail: sfeng1@ufl.edu).

Ming Yin is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: my0049@princeton.edu).

Ruiquan Huang and Jing Yang are with the School of Electrical Engineering and Computer Science, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: rzh5514@psu.edu; yangjing@psu.edu).

Yu-Xiang Wang is with the Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: yuxiangw@cs.ucsb.edu).

Yingbin Liang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.889@osu.edu).

Digital Object Identifier 10.1109/JSAIT.2024.3381818

I. Introduction

REINFORCEMENT learning (RL) focuses on the problem of maximizing the cumulative reward through interactions with an unknown environment. RL has witnessed a great success in practical applications, including robotics [2], [3], games [4], [5], [6], [7], and autonomous driving [8]. The unknown environment in RL is commonly modeled as a Markov decision process (MDP), where the set of states $\mathcal{S}$ describes all possible status of the environment. At a state $s \in \mathcal{S}$, an agent takes an action a from an action set $\mathcal{A}$ to interact with the environment, after which the environment transits to the next state $s' \in \mathcal{S}$ drawn from some unknown transition distributions, and then the agent receives an immediate reward. The interaction between the agent and the environment takes place episodically, where each episode consists of H steps. The notion called regret has been typically employed to measure the performance of RL algorithms, which measures how much worse an agent performs following its current policy comparison to the optimal policy in hindsight. The goal of the agent is to strategically interact with the environment to balance the exploration and exploitation tradeoff to minimize the regret.

Most existing RL studies adopt a static MDP model, in which both the reward and the transition kernel are time-invariant across episodes. However, stationary environment is insufficient to model enormous sequential decision problems such as online advertisement auctions [9], [10], traffic management [11], health care operations [12], and inventory control [13]. In contrast, nonstationary RL takes variations in rewards and transitions into consideration and is able to characterize larger classes of problems of interest [14]. In general, it is impossible to design algorithms that achieve sublinear regret for MDPs with drastically changing rewards and transitions in the worst case [15]. Therefore, one fundamental issue in the theoretical study of nonstationary RL is to investigate the maximum nonstationarity an agent can tolerate to adapt to the nonstationary dynamics of an MDP in order to achieve sublinear regret.

Without additional assumptions on the structure of the MDP, there is a line of extensive studies on nonstationary tabular MDPs [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. However, the performance of nonstationary tabular MDPs suffers from large state and action spaces, which limits its applicability in scenarios with exponentially large or continuous state spaces. Therefore,

function approximation has become a prominent tool to cope with this challenge. Several works have developed RL algorithms for nonstationary MDPs under structural assumptions, such as state-action set forming a metric space [29], linear MDPs [30], [31], linear mixture MDPs [32]. Although the developed algorithms are much more efficient than the algorithms designed for tabular setting, these algorithms require strong structural assumptions on the function approximation (such as a well-designed feature extractor in linear MDPs), which severely restricts the range of situations where these approaches can be employed. This naturally leads to the following open question:

Can we design an algorithm that achieves a "desired" regret performance[1] *for nonstationary MDPs under general function approximation?*

In this paper, we give an affirmative answer to the above question by investigating two different approaches and address the following challenges: First, we need to identify an appropriate complexity metric for nonstationary MDPs that covers many existing problems of interest. Second, We need to design an algorithm that can handle nonstationarity without additional structural assumptions on transition kernels and rewards. Third, it is non-trivial to establish a dynamic regret bound of the proposed algorithm that potentially improves those for nonstationary tabular and linear MDPs. The contributions of our work is summarized based on two different approaches as follows.

Confidence-set based algorithm. We propose a new complexity metric named the Dynamic Bellman Eluder (DBE) dimension for nonstationary MDPs, which generalizes the Bellman Eluder (BE) dimension designed for static MDPs [33], and subsumes a broad class of RL problems including low BE dimension problems in static RL and nonstationary tabular and linear MDPs in nonstationary RL. We then design a new confidence-set based algorithm SW-OPEA for nonstationary MDPs, by greedily selecting the candidate value function in the confidence region. Our design novelty lies in the construction of the confidence region, which features the sliding window mechanism, and incorporates local variation budget in order to accurately capture the distribution mismatch between the current episode and all episodes in the sliding window. Such a design ensures the optimal state-action value function in current episode to lie within the confidence region, and hence the optimism principle remains valid.

We theoretically characterize the dynamic regret of SW-OPEA. To demonstrate the advantage of SW-OPEA, we compare our regret bound of SW-OPEA to that of previously proposed UCB-type algorithms [30] for nonstationary linear and tabular MDPs. The comparison shows that our confidence-set based algorithm performs better in terms of the linear feature dimension $\tilde{d}$ and the horizon H, where the dependency on H also matches with the minimax lower bound given in [30], while performs slightly worse in the average variation budget. Therefore, the comparison suggests that our algorithm outperforms their algorithm in the small variation scenario. Our analysis features a few new developments. (a) We develop a distribution shift lemma to handle transition kernel variations over time. (b) We come up with new auxiliary random variables to form appropriate martingale differences and obtain the concentration results. (c) We use an auxiliary MDP to help bound the difference of two expectations under different underlying models.

UCB-type algorithm. To mitigate the computational inefficiency of the confidence-set based algorithm, we propose a UCB-type algorithm LSVI-Nonstationary for nonstationary MDPs with general function approximation, which adopts LSVI with upper confidence bound to handle the exploration and exploitation tradeoff. In order to handle nonstationarity, our algorithm features the restart mechanism, and incorporating the local variation budget in the design of the bonus term to ensure the optimism of the learned state-action value function.

We use the Eluder dimension to measure the complexity of the state-value function class $\mathcal{F}$ for nonstationary MDPs. We then theoretically characterize the dynamic regret of the proposed UCB-type algorithm, which depends on the Eluder dimension of function class $\mathcal{F}$. Our newly proposed UCB-type algorithm matches with the performance of SW-OPEA in terms of horizon H, average variation budget in transitions L_P and average variation budget in rewards L_r, while performing slightly worse in the number of states and actions $|\mathcal{S}|, |\mathcal{A}|$ under tabular MDPs and the same in the linear feature $\tilde{d}$ in linear MDPs. Our result suggests the benefit of UCB-type algorithm over confidence-set based algorithm.

Our main technical development for this approach lies in the single step optimization error for the least-square optimization in our UCB-type algorithm. We do not take the distribution drift in transitions and rewards into consideration, which may lead to non-trivial estimation error. In our analysis, we explicitly capture a non-trivial term due to the nonstationarity of the environment. We show that by compensating such a term involving local variation budget into the standard term due to concentration, the difference between the least-square predictor and the one-step backup estimate $r_h^k + P_h^k V_{h+1}^k$ is still bounded.

A. Related Work

Static Regret of Nonstationary MDPs: Static regret in nonstationary MDPs have been considered extensively in the past [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. Static regret has also been studied for nonstationary MDPs with function approximation. In particular, [31] characterizes the static regret for the weighted least squares value iteration method. Reference [34] studies the nonstationary RL setting with general function approximation, where the static regret is captured through a more general notion called decision-estimation coefficient (DEC).

Dynamic Regret of Nonstationary MDPs: Many studies in the past have been focused on the metric of dynamic regret, which quantifies the performance difference between the learning policy and the optimal policy at each step. For nonstationary tabular MDPs, value-based approaches have

[1]The performance of the algorithm relies on the variation budget of rewards and transitions. Mildly changed rewards and transitions results in a sublinear regret while drastically changed rewards and transitions leads to linear regret.

been proposed in [26], [28], where they respectively propose a sliding window strategy and a restart mechanism to handle nonstationarity. Further, [27] adopted a different method based on policy optimization. For nonstationary MDPs with function approximation, [30] and [32] focus on linear function approximation and linear-mixture function approximation, respectively, and [29] consider a kernel-based approach for nonstationary MDPs when state-action set forms a metric space. Further, [35] propose a unified approach to nonstationary MDPs that relies on an oracle algorithm with optimal regret for stationary MDPs to develop a useful algorithm for nonstationary MDPs.

Static MDPs with General Function Approximation: MDPs with general function approximation have been well studied in the static setting, where the transition kernel and reward function do not change over time. References [36] and [37] first introduce the notion of Eluder dimension to characterize the complexity of the function class, and study the performance based on such a metric. Later on, the notion Eluder dimension has been extended to Bellman Eluder dimension [33], and other notions have also been proposed, including Admissible Bellman Characterization (ABC) [38] and decision-estimation coefficient (DEC) [39]. Another line of research is based on low-rank conditions, including Bellman rank [40], [41], witness rank [42], and bilinear class [43]. Closest to our work here are the studies by [33] and [44]. For the confidence-set based algorithm, we generalize the Bellman Eluder dimension [33] for static MDPs to dynamic Eluder dimension for nonstationary MDPs, while for the UCB-type algorithm, we extend the study of UCB-type of approach in static MDPs [44] to nonstationary MDPs. Both of our approaches feature new elements in algorithm design and analysis tailored to nonstationary MDPs.

B. Relationship Between Trustworthy RL and Nonstationary RL

The goal of trustworthy reinforcement learning is to design algorithms competent in solving challenging real-world problems, including robustly handling perturbations, satisfying safety constraint, and generalizing to unseen environments. Nonstationary RL studied in this work is closely related to those three aspects. First, nonstationarity naturally occurs in robust MDPs. In classical robust RL setting, we aim to find a policy that maximizes the worst-case performance against uncertainty variable U, where uncertainty U could be either state s, action a, reward r, or transition P. When environment discrepancies are considered, i.e., uncertain variable follows $U = (P, r)$, and they satisfy the variation budget constraint, our algorithms provide candidate policies for robust MDPs with performance guarantee. Second, nonstationary MDPs can be viewed as a special case of safe RL problems. The nonstationarity, characterized by the variation budgets, serves as the constraint on the total variations of rewards and transitions, and our algorithms provide safe policies (satisfying variation budget constraints) with good performances. Third, nonstationary MDPs can help understand generalization in RL. Consider the scenario where testing environments are drawn from time-variant nonstationary distributions, and the agents are expected to learn how to leverage past experience and identify new environment. The nonstationary RL could serve as a general framework to study such a problem, and help understand generalization in RL.

Notation: For a set $\mathcal{S}$, $\Delta(\mathcal{S})$ represents a collection of distributions over $\mathcal{S}$. For a positive integer N, we use $[N]$ to denote the set of positive integers $\{1, 2, \ldots, N\}$. For positive integers m, n, define $\{\cdot\}_{[m:n]} = \varnothing$ if $m > n$. Let f, h be a function of n, $f(n) = O(h(n))$ is equivalently to that there exists positive k such that $f(n) = O(h(n)\log^k(h(n)))$. Given a dataset $\mathcal{D} = \{(x_i, a_i, q_i)\}_{i=1}^{|\mathcal{D}|} \subseteq \mathcal{S} \times \mathcal{A} \times \mathbb{R}$, for a function $f : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, define $\|f\|_{\mathcal{D}} = (\sum_{i=1}^{|\mathcal{D}|}(f(x_i, a_i) - q_i)^2)^{\frac{1}{2}}$. For a set of state-action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$, for a function $f : \mathcal{S} \times \mathcal{A}$, define $\|f\|_{\mathcal{Z}} = (\sum_{(x,a)\in\mathcal{Z}}(f(x_i, a_i))^2)^{\frac{1}{2}}$. For a set of functions $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}\}$, we define the width function of the state-action pair as $w(\mathcal{F}; x, a) = \sup_{f,f'\in\mathcal{F}}(f - f')(x, a)$.

II. Preliminaries

A. Nonstationary MDPs

Our setting can be formulated as a nonstationary finite-horizon episodic Markov decision process, captured by a tuple $(\mathcal{S}, \mathcal{A}, H, K, P, r, x_1)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of each episode, K is the total number of episodes, $P = \{P_h^k\}_{(k,h)\in[K]\times[H-1]}$ where $P_h^k : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the transition kernel at step h in the k-th episode, $r = \{r_h^k\}_{(k,h)\in[K]\times[H]}$ where $r_h^k : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the mean reward function at step h in the k-th episode, and x_1 is the fixed initial state.

The agent interacts with the nonstationary MDP sequentially. At the beginning of k-th episode, the agent chooses a policy $\pi^k = \{\pi_h^k\}_{h\in[H]}$ where $\pi_h^k : \mathcal{S} \mapsto \Delta(\mathcal{A})$. At step h, the agent observes the state x_h^k, takes an action following $a_h^k \sim \pi_h^k(\cdot|x_h^k)$, obtains a reward $\widetilde{r}_h^k$ (we also use r_h^k if there is no ambiguity) with mean $r_h^k(x_h^k, a_h^k)$, and the MDP evolves into the next state $x_{h+1}^k \sim P_h^k(x_h^k, a_h^k)$. The process ends after receiving the last reward r_H^k. We define the state and state-action value functions of policy $\pi = \{\pi_h\}_{h\in[H]}$ recursively via the following equation

$$Q^{\pi}_{h;(*,k)}(x, a) = r_h^k(x, a) + \left(P_h^k V^{\pi}_{h+1;(*,k)}\right)(x, a),$$
$$V^{\pi}_{h;(*,k)}(x) = \langle Q^{\pi}_{h;(*,k)}(x, \cdot), \pi_h^k(\cdot|x)\rangle_{\mathcal{A}}, \quad V_{H+1;(*,k)} = 0,$$

where $(*, k)$ represents the true model in the k-th episode, P_h^k is the operator defined as $(\mathbb{P}_h^k f)(x, a) := \mathbb{E}[f(x')|x' \sim P_h^k(x'|x, a)]$ for any function $f : \mathcal{S} \mapsto \mathbb{R}$. Here $\langle\cdot, \cdot\rangle_{\mathcal{A}}$ denotes the inner product over action space $\mathcal{A}$ and the subscript $\mathcal{A}$ is omitted when appropriate.

The learning objective is to find the optimal policy via interactions with the environment to minimize the dynamic regret

$$\mathrm{D-Regret}(K) := \sum_{k=1}^{K}\left(V^{\pi^{(*,k)}}_{1;(*,k)} - V^{\pi^k}_{1;(*,k)}\right)(x_1),$$

which quantifies the performance difference between the learning policy and the benchmark policy $\{\pi^{(*,k)}\}_{k\in[K]}$ where $\pi^{(*,k)} = \arg\max_{\pi} V^{\pi}_{1;(*,k)}(x_1)$.

B. Function Approximation

Consider a function class $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_H$, where $\mathcal{F}_h \subseteq \{f : \mathcal{S} \times \mathcal{A} \mapsto [0, H-h+1]\}$ is the candidate function class to approximate $Q^{\pi_{(*,k)}}_{h;(*,k)}$. For convenience, we set $f_{H+1} = 0$, and therefore $\mathcal{F}_{H+1} = \{f(s,a) = 0 : (s,a) \in \mathcal{S} \times \mathcal{A}\}$.

Assumption 1 (Realizability): $Q^*_{h;(*,k)} \in \mathcal{F}_h$ for all $(k,h) \in [K] \times [H]$.

Realizability assumption requires that the optimal state-action value function in each episode is contained in the function class $\mathcal{F}$ with no approximation error, i.e., $(Q^*_{1;(*,k)}, \ldots, Q^*_{H;(*,k)}) \in \mathcal{F}$ for $k \in [K]$.

Given functions $f = (f_1, f_2, \ldots, f_H)$ where $f_h \in (\mathcal{S} \times \mathcal{A} \mapsto [0, H-h+1])$, define

$$\left(\mathcal{T}^k_h f_{h+1}\right)(x,a) \coloneqq r^k_h(x,a) + \left(P^k_h f_{h+1}\right)(x,a),$$

$$\left(P^k_h f_{h+1}\right)(x,a) = \mathbb{E}_{x' \sim P^k_h(\cdot|x,a)}\left[\max_{a' \in \mathcal{A}} f_{h+1}(x', a')\right],$$

where $\mathcal{T}^k_h$ is the Bellman operator at step h in episode k. Note that the optimal state-action value function satisfies $Q^*_{h;(*,k)}(x,a) = (\mathcal{T}^k_h Q^*_{h+1;(*,k)})(x,a)$ for all valid x, a, h. Moreover, we define $\mathcal{T}^k_h \mathcal{F}_{h+1} = \{\mathcal{T}^k_h f_{h+1} : f_{h+1} \in \mathcal{F}_{h+1}\}$.

Assumption 2 (Completeness): $\mathcal{T}^k_h \mathcal{F}_{h+1} \subseteq \mathcal{F}_h$ for all $(k,h) \in [K] \times [H]$.

For the completeness assumption, we require that after applying the Bellman operator $\mathcal{T}^k_h$ of any episode k to a function f_{h+1} in the function class $\mathcal{F}_{h+1}$ at step $h+1$, the resulting function lies in the function class $\mathcal{F}_h$ at previous step h.

C. Complexity Measures

In this section, we introduce two complexity measures for a class of functions. One is Eluder dimension and the other one is distributional Eluder dimension.

The definition of Eluder dimension was first proposed in [44], and is based on the ϵ-independence of points, as illustrated in the following definition.

Definition 1 (Eluder Dimension): Let $\epsilon \geq 0$ and $\mathcal{Z} = \{(x_i, a_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A}$ be a sequence of state-action pairs.

- A state-action pair $(x,a) \in \mathcal{S} \times \mathcal{A}$ is ϵ-dependent on $\mathcal{Z}$ with respect to $\mathcal{F}$ if any $f, f' \in \mathcal{F}$ satisfying[2] $\|f - f'\|_{\mathcal{Z}} \leq \epsilon$ also satisfies $|f(x,a) - f'(x,a)| \leq \epsilon$.
- An (x,a) is ϵ-independent on $\mathcal{Z}$ with respect to $\mathcal{F}$ if (x,a) is not ϵ-dependent on $\mathcal{Z}$.
- The Eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ of a function class $\mathcal{F}$ is the length of the longest sequence of elements in $\mathcal{S} \times \mathcal{A}$ such that, for some $\epsilon' \geq \epsilon$, every element is ϵ'-independent of its predecessors.

It has been shown in [36] that $\dim_E(\mathcal{F}, \epsilon) \leq |\mathcal{S}||\mathcal{A}|$ for tabular MDPs, and $\dim_E(\mathcal{F}, \varepsilon) \leq \widetilde{O}(\widetilde{d})$ for linear MDPs where $\widetilde{d}$ is the feature dimension.[3]

We extend the notion of ϵ-independence of points to ϵ-independence of distributions, and obtain the definition of distributional Eluder dimension [33].

[2] $\|\cdot\|_{\mathcal{Z}}$ is formally defined in Section I Notation.

[3] The proofs for the nonstationary setting are essentially the same as the proof for the stationary setting therein, and we do not differentiate the two settings.

Definition 2 (Distributional Eluder Dimension): Let $\epsilon \geq 0$ and $\{\nu_i\}_{i=1}^n \subseteq \Delta(\mathcal{S} \times \mathcal{A})$ be a sequence of probability distributions.

- A distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is ϵ-dependent on $\{\nu_1, \ldots, \nu_n\}$ with respect to $\mathcal{F}$ if any $f \in \mathcal{F}$ satisfying $\sqrt{\sum_i (\mathbb{E}_{\nu_i} f)^2} \leq \epsilon$ also satisfies $|\mathbb{E}_\mu f| \leq \epsilon$.
- A μ is ϵ-independent on $\{\nu_1, \ldots, \nu_n\}$ with respect to $\mathcal{F}$ if μ is not ϵ-dependent on $\{\nu_1, \ldots, \nu_n\}$.
- The distributional Eluder dimension $\dim_E(\mathcal{F}, \Pi, \epsilon)$ of a function class $\mathcal{F}$ and distribution class Π is the length of the longest sequence of elements in Π such that, for some $\epsilon' \geq \epsilon$, every element is ϵ'-independent of its predecessors.

III. Confidence-Set Based Algorithm

To the best of our knowledge, the proposed SW-OPEA is the first confidence-set based algorithm in the context of nonstationary MDPs. At high level, confidence-set based algorithm consists of three key steps: optimistic planning, data collection and confidence set updating. Compared to static MDPs, we adopt sliding window mechanism and incorporate local variation budgets in transitions and rewards to compensate for the distribution mismatch between the current episode and all episodes in the sliding window to handle nonstationarity. Despite of the new technical developments for the analysis, our algorithm for nonstationary MDPs remains concise and simple.

A. Dynamic Eluder Dimension

In this section, we first review the Bellman Eluder (BE) dimension for static MDPs, and propose a new complexity metric Dynamic Eluder (DBE) dimension for nonstationary MDPs. Both BE dimension and DBE dimension are based on the distributional Eluder dimension (see Definition 2). However, compared to Bellman Eluder dimension, the new Dynamic Eluder Dimension nicely captures the nonstationarity of the problem.

The definition of Bellman Eluder dimension was first introduced in [33] for static MDPs.

Definition 3 (Bellman Eluder dimension (BE)): Let $(I - \mathcal{T}_h)\mathcal{F} \coloneqq \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}, k \in [K]\}$ be the set of Bellman residuals in all episodes induced by $\mathcal{F}$ at step h, and $\Pi = \{\Pi_h\}_{h \in [H]}$ be a collection of H probability measure families over $\mathcal{S} \times \mathcal{A}$. The ϵ-Bellman Eluder dimension of $\mathcal{F}$ with respect to Π is defined as

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

For nonstationary MDPs, the Bellman operators $\mathcal{T}_h$ varies over episodes, and hence we introduce our new complexity measure called dynamic Bellman Eluder dimension for nonstationary MDPs.

Definition 4 (Dynamic (Bellman) Eluder (DBE) dimension): Let $(I - \bar{\mathcal{T}}_h)\mathcal{F} \coloneqq \{f_h - \mathcal{T}^k_h f_{h+1} : f \in \mathcal{F}, k \in [K]\}$ be the set of Bellman residuals in all episodes induced by $\mathcal{F}$ at step h, and $\Pi = \{\Pi_h\}_{h \in [H]}$ be a collection of H probability measure

Algorithm 1 GOLF (Sketch)

1: **Input:** $\mathcal{D}_1, \ldots, \mathcal{D}_H \leftarrow \varnothing$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
2: **for episode** k from 1 to K **do**
3: **Choose** $\pi^k = \pi_{f^k}$, where $f^k = \arg\max_{f \in \mathcal{B}^{k-1}} f_1(x_1, \pi_f(x_1))$.
4: **Collect** a trajectory $(x_1, a_1, r_1, \ldots, x_H, a_H, r_H, x_{H+1})$ by following π^k.
5: **Augment** $\mathcal{D}_h = \mathcal{D}_h \cup \{(x_h, a_h, r_h, x_{h+1})\}$, $\forall h \in [H]$.
6: Update $\mathcal{B}^k = \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta, \ \forall h \in [H]\}$, where $\mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{(s,a,r,s') \in \mathcal{D}_h} \big(\xi_h(x_h^t, a_h^t) - r - \max_{a' \in \mathcal{A}} \zeta_{h+1}(x_{h+1}^t, a')\big)^2$
7: **end for**

families over $\mathcal{S} \times \mathcal{A}$. The dynamic Bellman Eluder dimension of $\mathcal{F}$ with respect to Π is defined as

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

We focus on the following choice of distribution family $\mathcal{D}_\Delta = \{\mathcal{D}_{\Delta,h}\}_{h \in [H]}$ where $\mathcal{D}_{\Delta,h} = \{\delta_{(s,a)} : s \in \mathcal{S}, a \in \mathcal{A}\}$, i.e., the collections of probability measures that put measure 1 on as single state-action pair.

The DBE dimension is the distributional Eluder dimension on the function class $(I - \bar{\mathcal{T}}_h)\mathcal{F}$ in all episodes, maximizing over step $h \in [H]$, which can be viewed as an extension of BE dimension to nonstationary MDPs. The main difference between DBE dimension and BE dimension is that the Bellman operator $\mathcal{T}_h^k$ is time-varying, and we include all the Bellman residues induced by $\mathcal{T}_h^k$ for $k \in [K]$ in the function class. In general, the DBE dimension could be substantially larger than the BE dimension due the fact that the class of functions can be significantly larger. However, we can show that, if the variations in both transitions and rewards are relatively small compared to a universal gap, then the DBE dimension equals to the BE dimension with respect to one MDP instance of the nonstationary MDP [1]. Moreover, the DBE dimension of nonstationary linear MDPs scales linearly with the linear feature dimension $\widetilde{\mathcal{O}}(\widetilde{d})$ [1].

B. Algorithm SW-OPEA

In this section, we propose our confidence-set based algorithm SW-OPEA for nonstationary MDPs with general function approximation.

Overview of GOLF [33]: We first give a brief introduction of GOLF in Algorithm 1 for static MDPs with general function approximation. There are three key components: Optimistic planning (line 3), data collection (line 4), and updating confidence set B^k (line 6). The key step is to construct the confidence set $\mathcal{B}^k$, and GOLF maintains a local regression constraint using collected data $\mathcal{D}_h$ at this step $\mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta$, where β is a confidence parameter, and $\mathcal{L}_{\mathcal{D}_h}$ is the squared loss proxy to the squared Bellman error at step h. It was shown that the regret of GOLF is $\widetilde{O}(H\sqrt{dK})$, where $d = \dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, 1/\sqrt{K})$ is the BE dimension.

Algorithm 2 Sliding Window Optimistic Exploration and Approximation (SW-OPEA)

1: **Input:** $\mathcal{D}_1, \ldots, \mathcal{D}_H \leftarrow \varnothing$, $\mathcal{B}^0 \leftarrow \mathcal{F}$, local variation budgets $\Delta_P^w(k,h)$, $\Delta_R^w(k,h)$.
2: **for episode** k from 1 to K **do**
3: **Choose** $\pi^k = \pi_{f^k}$, where $f^k = \arg\max_{f \in \mathcal{B}^{k-1}} f_1(x_1, \pi_f(x_1))$.
4: **Collect** a trajectory $(x_1^k, a_1^k, r_1^k, \ldots, x_H^k, a_H^k, r_H^k, x_{H+1}^k)$ by following π^k.
5: **Augment** $\mathcal{D}_h = \mathcal{D}_h \cup \{(x_h^k, a_h^k, x_{h+1}^k)\}$, $\forall h \in [H]$.
6: Update $\mathcal{B}^k = \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta + 2H^2\Delta_P^w(k,h) + 2H\Delta_R^w(k,h), \ \forall h \in [H]\}$, where $\mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1})$ is defined in (1).
7: **end for**

At a high level, SW-OPEA differentiates from the GOLF algorithm [33] for static MDPs with general function approximation in its novel designs to handle the nonstationarity of transition kernels and rewards. Specifically, SW-OPEA features the sliding window mechanism and incorporates local variation budget in order to accurately capture the distribution mismatch between the current episode and all episodes in the sliding window. Such a design ensures the optimal state-action value function in the current episode to lie within the confidence region, and hence the optimism principle remains valid.

The pseudocode of SW-OPEA is presented in Algorithm 2. SW-OPEA initializes the dataset $\{\mathcal{D}_h\}_{h \in [H]}$ to be empty sets, and confidence set $\mathcal{B}^0$ to be $\mathcal{F}$. Then, in each episode, SW-OPEA performs the following two steps:

Optimistic planning step (Line 3) greedily selects the most optimistic state-action value function f^k from the confidence set $\mathcal{B}^{k-1}$ constructed in the last episode, and chooses the corresponding greedy policy π_k associated with f^k.

Sliding window squared Bellman error is defined as

$$\mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{t=1 \vee (k-w)}^{k} \Big(\xi_h(x_h^t, a_h^t) - r_h^t - \max_{a' \in \mathcal{A}} \zeta_{h+1}(x_{h+1}^t, a')\Big)^2. \quad (1)$$

Note that in episode k, we use bandit reward r_h^t in the construction of the sliding window squared Bellman error, and $\mathcal{L}_{\mathcal{D}_h}$ tends to be small as long as the transition kernel difference between episode k and t is small. Furthermore, based on the "forgetting principle" [45], we adopt the **sliding window** in the squared loss (1), where the data used to estimate the squared loss at episode k relies on the latest $w+1$ observations (when iteration number is sufficiently large) during episode $1 \vee (k-w)$ to k instead of all prior observations. The rationale is that under nonstationarity setting, the historical observations far in the past are obsolete, and they are not as informative for the evaluation of the squared loss.

Confidence set updating step (Line 4–6) first executes policy π^k and collects data for the current episode, and then updates the confidence set based on the new data.

The key novel ingredient of SW-OPEA lies in the construction of the confidence set $\mathcal{B}^k$. For each $h \in [H]$, SW-OPEA maintains a local regression constraint using the collected data $\mathcal{D}_h$

$$\mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta + 2H^2 \Delta_P^w(k,h) + 2H \Delta_R^w(k,h),$$

where β is a confidence parameter, and Δ_P^w, Δ_R^w are the local variation budgets defined by

$$\Delta_P^w(k,h) = \sum_{t=1\vee(k-w)}^{k} \sup_{x\in\mathcal{S}, a\in\mathcal{A}} \left\| \left(P_h^k - P_h^t\right)(\cdot|x,a) \right\|_1. \quad (2)$$

$$\Delta_R^w(k,h) = \sum_{t=1\vee(k-w)}^{k} \sup_{x\in\mathcal{S}, a\in\mathcal{A}} \left| \left(r_h^k - r_h^t\right)(x,a) \right|. \quad (3)$$

Since the transition kernel varies across episodes, we include an additional term of the local variation budget $\Delta_P^w(k,h)$ and $\Delta_R^w(k,h)$ in the definition of $\mathcal{B}_k$. Intuitively, the local variation budget $\Delta_P^w(k,h)$ and $\Delta_R^w(k,h)$ captures the cumulative transition kernel and reward differences between current episode and all previous episode in the sliding window. Therefore, by compensating a term involving $\Delta_P^w(k,h)$ and $\Delta_R^w(k,h)$ in the confidence set $\mathcal{B}_k$, we ensure that the optimal state-action value function in the k-th episode $Q^*_{h;(*,k)}$ still lies in the confidence set $\mathcal{B}^k$ with high probability.

C. Theoretical Guarantees

In this section, we provide our main theoretical result for SW-OPEA, and defer the proof sketch that highlights our novel developments in the analysis to Appendix A.

We first state the following generalized completeness assumption [33], [46], [47]. Let $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_H$ be an auxiliary function class provided to the learner where $\mathcal{G}_h \subseteq (\mathcal{S} \times \mathcal{A} \mapsto [0, H-h+1])$.

Assumption 3 (Generalized Completeness): $\mathcal{T}_h^k \mathcal{F}_{h+1} \subseteq \mathcal{G}_h$ for all $(k,h) \in [K] \times [H]$.

If we choose $\mathcal{G} = \mathcal{F}$, then Assumption 3 is equivalent to the standard completeness assumption (see Assumption 2). Without loss of generality, we assume $\mathcal{F} \subseteq \mathcal{G}$, which implies $\mathcal{G} = \mathcal{F} \cup \mathcal{G}$.

Moreover, to quantify the nonstationarity, we define the variation in rewards of adjacent episodes and the variation in transition kernels of adjacent episodes as

$$\Delta_R(K) = \sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{x\in\mathcal{S}, a\in\mathcal{A}} \left| \left(r_h^k - r_h^{k-1}\right)(x,a) \right|, \quad (4)$$

$$\Delta_P(K) = \sum_{k=1}^{K}\sum_{h=1}^{H} \sup_{x\in\mathcal{S}, a\in\mathcal{A}} \left\| \left(P_h^k - P_h^{k-1}\right)(\cdot|x,a) \right\|_1, \quad (5)$$

where we define $P_h^0 = P_h^1$ and $r_h^0 = r_h^1$ for all $h \in [H]$.

The dynamic regret of our algorithm SW-OPEA is characterized in the following theorem.

Theorem 1: Under Assumption 1 and Assumption 3, there exists an absolute constant c such that for any $\delta \in (0,1]$, $K \in \mathbb{N}$, if we choose $\beta = cH^2 \log \frac{KH|\mathcal{G}|}{\delta}$ in SW-OPEA, then with probability at least $1-\delta$, for all $k \in [K]$, when $k \ge \min\{w+1, \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})\}$, $\mathrm{D-Regret}(k)$ equals

$$\Delta_R(k) + H\Delta_P(k) + \mathcal{O}lH\sqrt{w} + \frac{H^2 k}{\sqrt{w}}\sqrt{d\log\left[KH|\mathcal{G}|/\delta\right]} + \frac{H^2 k}{\sqrt{w}}\sqrt{d \sup_{t\in[k]} \Delta_P^w(t,h)} + \frac{H^{3/2}k}{\sqrt{w}}\sqrt{d \sup_{t\in[k]} \Delta_R^w(t,h)}\, l,$$

where $d = \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})$.

Note that the last term depends on the sliding window size w, and we can further optimize w if an upper bound of the local variation budget $\Delta_P^w(t,h)$ and $\Delta_R^w(t,h)$ is given. Below we give an example for optimizing sliding window size w.

Before we proceed, we first define the average variation budget L as

$$L_P = \max_{h\in[H], t<k} \frac{\sum_{s=t}^{k-1} \sup_{x,a} \left\| \left(P_h^{s+1} - P_h^s\right)(\cdot|x,a) \right\|_1}{k-t}, \quad (6)$$

$$L_r = \max_{h\in[H], t<k} \frac{\sum_{s=t}^{k-1} \sup_{x,a} \left| \left(r_h^{s+1} - r_h^s\right)(x,a) \right|}{k-t}. \quad (7)$$

Clearly, we have $L_P, L_r \le 1$ and $\Delta_P^w(k,h) \le L_P w^2$, $\Delta_R^w(k,h) \le L_r w^2$. L_P, L_r can be viewed as the greatest average variation of transition kernels and rewards across adjacent episodes over any period of episodes maximized over step $h \in [H]$. Then the following corollary characterizes the dynamic regret by optimizing the window size w based on L_P and L_r.

Corollary 1: Under the conditions of Theorem 1 and $|\mathcal{G}| > 10$, with probability at least $1-\delta$, the following argument holds: if $\sqrt{L_P} + \frac{\sqrt{L_r}}{\sqrt{H}} > \frac{1}{K}(\sqrt{\log|\mathcal{G}|} - \frac{1}{H\sqrt{d}})$, select $w = \lceil \frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L_P} + \frac{\sqrt{L_r}}{\sqrt{H}} + \frac{1}{HK\sqrt{d}}} \rceil$, the dynamic regret is upper-bounded by

$$\widetilde{\mathcal{O}}\Big(H^{\frac{3}{2}} K^{\frac{1}{2}} d^{\frac{1}{4}} (\log|\mathcal{G}|)^{\frac{1}{4}} + H^2 K L_P^{\frac{1}{4}} d^{\frac{1}{2}} (\log|\mathcal{G}|)^{\frac{1}{4}} + H^{\frac{7}{4}} K L_r^{\frac{1}{4}} d^{\frac{1}{2}} (\log|\mathcal{G}|)^{\frac{1}{4}} + \Delta_R + H\Delta_P\Big); \quad (8)$$

otherwise, select $w = K$ and the dynamic regret is upper-bounded by $\widetilde{O}(H^2 K^{\frac{1}{2}} d^{\frac{1}{2}} (\log|\mathcal{G}|)^{\frac{1}{2}})$, where $d = \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})$.

We remark that $|\mathcal{G}|$ appearing in the log term can be replaced by its ϵ-covering number $\mathcal{N}_{\mathcal{G}}(\epsilon)$ to handle the classes with infinite cardinality. In both Theorem 1 and Corollary 2, we do not omit $\log|\mathcal{G}|$ in $\widetilde{\mathcal{O}}$ since for many function classes, $\log|\mathcal{G}|$ (or $\log\mathcal{N}_{\mathcal{G}}(\epsilon)$) can contribute to a polynomial factor. For example, for $\widetilde{d}$ dimensional linear function class, $\log\mathcal{N}_{\mathcal{G}}(\epsilon) = \widetilde{\mathcal{O}}(\widetilde{d})$ where $\widetilde{d}$ is the linear feature dimension.

Our first term in (8) corresponds to the regret of the static MDP while the remaining term arises due to the nonstationarity. As a result, when transitions and rewards remain the same over time, our result reduces to $\widetilde{\mathcal{O}}(H^2 K^{\frac{1}{2}} d^{\frac{1}{2}} (\log|\mathcal{G}|)^{\frac{1}{2}})$, which matches with the static regret of GOLF in [33].[4]

Advantage of SW-OPEA: To understand the advantage of SW-OPEA over existing algorithms for nonstationary MDPs, we take nonstationary linear MDPs as an example. When

[4]The additional H here is due to the definition of $r_h \in [0,1]$, whereas [33] assumes $\sum_h r_h \le 1$.

specializing to nonstationary linear and tabular MDPs, our result becomes $\widetilde{\mathcal{O}}(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d} + HT\widetilde{d}^{\frac{3}{4}}L_P^{\frac{1}{4}} + H^{\frac{3}{4}}T\widetilde{d}^{\frac{3}{4}}L_r^{\frac{1}{4}})$ where $T = HK$, $\widetilde{d}$ is the feature dimension for linear MDPs and $\widetilde{d}$ equals $|\mathcal{S}||\mathcal{A}|$ for tabular MDPs, and L_r is the average variation budget in rewards. For nonstationary linear MDPs, the result in [30] is not comparable to ours due to the different definitions of the variation budget of transition kernels. To make a fair comparison, we convert their bound on the dynamic regret to be that for tabular MDPs, which gives $\widetilde{\mathcal{O}}(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}^{\frac{3}{2}} + H^{\frac{4}{3}}\widetilde{d}^{\frac{3}{2}}T\widetilde{L}_P^{\frac{1}{3}} + H^{\frac{4}{3}}\widetilde{d}^{\frac{4}{3}}TL_r^{\frac{1}{3}})$. The first term corresponds to the regret of static linear MDPs and our result has better dependency on the feature dimension $\widetilde{d}$. For the second term due to the nonstationarity of transition kernels, our bound is better in terms of the horizon H and feature dimension $\widetilde{d}$ while worse in terms of the average variation budget of transitions L_P (note that $L_P \leq 1$).[5] Similarly for the last term caused by the nonstationary of rewards, our result performs better in terms of the horizon H and feature dimension $\widetilde{d}$ while worse in terms of the average variation budget of rewards L_r (note that $L_r \leq 1$).

It is also interesting to compare our result with the minimax dynamic regret lower bound $\Omega(H^{\frac{1}{2}}T^{\frac{1}{2}}\widetilde{d} + H^{\frac{1}{3}}T\widetilde{d}^{\frac{2}{3}}\widetilde{L}_P^{\frac{1}{3}})$ developed in [30] for linear MDPs with nonstationary transitions. For such a case, our result becomes $\widetilde{\mathcal{O}}(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d} + HT\widetilde{d}^{\frac{3}{4}}L_P^{\frac{1}{4}})$. The first term is the regret under stationary MDPs and the second term arises due to the nonstationarity of transitions. We can see that our first term corresponding to static MDPs matches with the lower bound both in terms of T and $\tilde{d}$, whereas the upper bound in [30] matches with the lower bound only in T. For the nonstationarity term, our dependency on H and $\tilde{d}$ is closer to the lower bound than that in [30], whereas our dependency on the variation budget is close but does not match with the lower bound.

IV. UCB-Type Algorithm

In this section, we investigate a new UCB-type algorithm LSVI-Nonstationary. Our proposed algorithm falls into the popular LSVI framework, which uses LSVI with upper confidence bound to handle exploration and exploitation tradeoff. While designing the bonus term is simple in static tabular and linear MDPs, it becomes difficult in nonstationary MDPs with general function approximation. Our algorithm features the restart mechanism and incorporate the local variation budget in the design of the bonus term to handle nonstationarity. Moreover, it alleviates the computational inefficiency in the confidence-set based approach to select the optimistic state-action value functions in each step altogether.

We begin with the bounded complexity assumption [44] on the function class $\mathcal{F}$ and the state-action pairs $\mathcal{S} \times \mathcal{A}$.

Assumption 4: For any $\varepsilon > 0$, the following statements hold:

- There exists an ε-cover $\mathcal{C}(\mathcal{F}, \varepsilon) \subseteq \mathcal{F}$ with size $|\mathcal{C}(\mathcal{F}, \varepsilon)| \leq \mathcal{N}(\mathcal{F}, \varepsilon)$, such that for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{C}(\mathcal{F}, \varepsilon)$ with $\|f - f'\|_\infty \leq \varepsilon$;
- There exists an ε-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ with size $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon) \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon)$, such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$, there exists $(x', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ with $\sup_{f \in \mathcal{F}} |f(x, a) - f(x', a')| \leq \varepsilon$.

This assumption essentially requires both the function class and the state-action pairs have bounded covering numbers. It is acceptable for the covers to have exponential size since the regret bound scales logarithmically on both $\mathcal{N}(\mathcal{F}, \cdot)$ and $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \cdot)$. For the tabular case when $\mathcal{S}, \mathcal{A}$ are finite, $\log \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(|\mathcal{S}||\mathcal{A}|)$ and $\log \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \log(|\mathcal{S}||\mathcal{A}|)$. For the linear MDPs with feature dimension $\widetilde{d}$, $\log \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(\widetilde{d})$ and $\log \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \log(\widetilde{d})$.

Algorithm 3 $\mathcal{F}$-LSVI (Sketch)

1: **Input:** failure probability $\delta \in (0, 1)$, number of episodes K.
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3: Receive initial state $s_1^k \sim \mu$.
4: $Q_{H+1}^k(\cdot, \cdot) \leftarrow 0$ and $V_{H+1}^k(\cdot) \leftarrow 0$.
5: $\mathcal{Z}_h^k = \{(x_h^\ell, a_h^\ell)\}_{\ell \in [1:k-1]}$.
6: **for** $h = H, H-1, \ldots, 1$ **do**
7: $\mathcal{D}_h^k = \{(x_h^\ell, a_h^\ell, \widetilde{r}_h^\ell + V_{h+1}^k(x_{h+1}^\ell))\}_{\ell \in [1,k-1]}$.
8: $f_h^k \leftarrow \arg\min_{f \in \mathcal{F}_h} \|f\|_{\mathcal{D}_h^k}^2$.
9: $b_h^k \leftarrow \text{bonus}(\mathcal{F}_h, f_h^k, \mathcal{Z}_h^k, \delta)$.
10: $Q_h^k(\cdot, \cdot) \leftarrow \min\{f_h^k(\cdot, \cdot) + b_h^k(\cdot, \cdot), H\}$ and $V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
11: $\pi_h^k(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
12: **end for**
13: **for** $h = 1, 2, \ldots, H$ **do**
14: Take action $a_h^k \leftarrow \widetilde{\pi}_h^k(x_h^k)$ and observe $x_{h+1}^k \sim P_h^k(\cdot|x_h^k, a_h^k)$ and $\widetilde{r}_h^k \sim r_h^k(x_h^k, a_h^k)$.
15: **end for**
16: **end for**

A. Algorithm LSVI-Nonstationary

In this section, we describe our proposed UCB-type algorithm LSVI-Nonstationary for nonstationary MDPs with general function approximation.

Overview of $\mathcal{F}$-LSVI [44]: We begin with UCB-type algorithm $\mathcal{F}$-LSVI in Algorithm 3 for static MDPs[6] with general function approximation. At the beginning of each episode k, we set $Q_{H+1}^k = 0$, and calculate $Q_H^k, Q_{H-1}^k, \ldots, Q_1^k$ iteratively (line 8–10). Then, the greedy policy with respect to Q_h^k to collect data for the k-th episode. The procedure is repeated until all K episodes are finished. The key ingredient is the design of the bonus term b_h^k in line 9 based on sensitivity sampling to tightly characterize the estimation error, so that the optimistic Q function estimated (line 10) always upper bounds the optimal action-value function. The regret of $\mathcal{F}$-LSVI was shown to be $\widetilde{O}(\sqrt{\widetilde{d}^3H^3K})$ in tabular MDPs where $\widetilde{d} = |\mathcal{S}||\mathcal{A}|$, and $\widetilde{O}(\sqrt{\widetilde{d}^4H^3K})$ in linear MDPs where $\widetilde{d}$ is the feature dimension.

[5] Strictly speaking, their average variation budget $\widetilde{L}_P$ is not comparable to L_P, and the argument holds approximately.

[6] The algorithm presented here is slightly different from its original version [44] in using function class $\mathcal{F}_h$ instead of $\mathcal{F}$ for estimating value function at step h.

Algorithm 4 LSVI-Nonstationary

1: **Input:** failure probability $\delta \in (0, 1)$, number of episodes K and epoch size W.
2: **for** $j = 1, 2, \ldots, \lceil \frac{K}{W} \rceil$ **do**
3: $\tau \leftarrow (j-1)W + 1$.
4: **for** episode $k = \tau, \ldots, \min\{\tau + W, K\}$ **do**
5: Receive initial state $s_1^k \sim \mu$.
6: $Q_{H+1}^k(\cdot, \cdot) \leftarrow 0$ and $V_{H+1}^k(\cdot) \leftarrow 0$.
7: $\mathcal{Z}_h^k = \{(x_h^\ell, a_h^\ell)\}_{\ell \in [\tau:k-1]}$.
8: **for** $h = H, H-1, \ldots, 1$ **do**
9: $\mathcal{D}_h^k = \{(x_h^\ell, a_h^\ell, \widetilde{r}_h^\ell + V_{h+1}^k(x_{h+1}^\ell))\}_{\ell \in [\tau, k-1]}$.
10: $f_h^k \leftarrow \arg\min_{f \in \mathcal{F}_h} \|f\|_{\mathcal{D}_h^k}^2$.
11: $b_h^k \leftarrow \text{bonus}(\mathcal{F}_h, f_h^k, \mathcal{Z}_h^k, \delta, j)$ (Algorithm 6).
12: $Q_h^k(\cdot, \cdot) \leftarrow \min\{f_h^k(\cdot, \cdot) + b_h^k(\cdot, \cdot), H\}$ and $V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
13: $\pi_h^k(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
14: **end for**
15: **for** $h = 1, 2, \ldots, H$ **do**
16: Take action $a_h^k \leftarrow \widetilde{\pi}_h^k(x_h^k)$ and observe $x_{h+1}^k \sim P_h^k(\cdot|x_h^k, a_h^k)$ and $\widetilde{r}_h^k \sim r_h^k(x_h^k, a_h^k)$.
17: **end for**
18: **end for**
19: **end for**

From a high level point of view, our algorithm features two key ingredients: least-square value iteration (LSVI) and a restart mechanism. Our algorithm uses LSVI with upper confidence bound to handle the exploration and exploitation tradeoff, where we incorporate the local variation budget in the design of bonus term to ensures the optimism of the learned state-action value function. Moreover, we use the epoch restart mechanism to adapt to the nonstationarity of the environment. Those ingredients make our design significantly different from the $\mathcal{F}$-LSVI algorithm [44] for static MDPs.

The pseudocode of LSVI-Nonstationary is presented in Algorithm 4. Our algorithm runs in epochs with length W. Within each episode, we follow these steps: Firstly, we estimate the state-action value function through a least-square problem using historical data from the current epoch. Next, we create an upper confidence bound for the state-action value function and select the policy that maximizes this upper confidence bound. A new trajectory is then collected by following the greedy policy. Finally, we periodically restart the algorithm to handle the nonstationarity of the environment.

Least-square value iteration: At the beginning of each episode k, we maintain a replay buffer of existing samples $\{x_h^\ell, a_h^\ell, r_h^\ell\}_{\ell \in [\tau:k-1]}$, where τ is the first episode of the epoch containing episode k. Let $Q_{H+1}^k = 0$, and we set $Q_H^k, Q_{H-1}^k, \ldots, Q_1^k$ iteratively as follows (line 10–12). For $h = H, H-1, \ldots, 1$,

$$f_h^k(\cdot, \cdot) = \arg\min_{f \in \mathcal{F}_h} \sum_{\ell=\tau}^{k-1} \left(f\left(x_h^\ell, a_h^\ell\right) - r_h^\ell - \max_a Q_{h+1}^k\left(x_{h+1}^\tau, a\right) \right)^2,$$

$$Q_h^k(\cdot, \cdot) = \min\{f_h^k(\cdot, \cdot) + b_h^k(\cdot, \cdot), H\},$$

where $b_h^k(\cdot, \cdot)$ is the bonus function to be defined shortly. After obtaining $Q_h^k(\cdot, \cdot)$, we then use the greedy policy with respect to Q_h^k to collect data (line 13) for the kth episode. Note that the least-square problem does not take into consideration the distribution drift in transitions and rewards, which may potentially result in significant estimation errors. However, our analysis shows that these estimation errors can adapt to the nonstationarity. Specifically, we incorporate such estimation errors into the design of the bonus term to ensure the state-action value estimate is an optimistic upper bound of the optimal state-action value function.

Stable upper-confidence bonus function: With more collected data, the least-square solution is expected to provide a better approximation to the optimal state-action value function. To encourage exploration, we add additional bonus function $b_h^k(\cdot, \cdot)$ to guarantee that with high probability, $Q_{h+1}^k(\cdot, \cdot)$ is an optimistic upper bound of the optimal state-action value function. The design of bonus term b_h^k has two features: First, we leverage the importance sampling technique [44] to prioritize important data in the replay buffer so that the bonus function b_h^k is stable even when the replay buffer has large cardinality. Second, the distribution drift of the transitions and the rewards is characterized in the design of bonus term b_h^k in order to obtain the optimistic upper bound of the optimal state-action value function.

We define bonus function to be the width function $b_h^k(\cdot, \cdot) = w(\mathcal{F}_h^k; \cdot, \cdot)$, where $\mathcal{F}_h^k$ is defined as the confidence set so that the estimation error of the one-step backup $(r_h^k + P_h^k V_{h+1}^k)(\cdot, \cdot)$ lies within $\mathcal{F}_h^k$ with high probability. By the definition of width function, $b_h^k(\cdot, \cdot)$ provides an upper bound on the confidence interval of the state-action value estimate, since the width function maximizes the difference between all pairs of state-action value functions within the confidence set. Specifically, we define the confidence set as $\mathcal{F}_h^k = \{f \in \mathcal{F}_h : \|f - f_h^k\|_{\mathcal{Z}_h^k}^2 \le \beta + H\Delta_h^k\}$ where β is the confidence parameter properly selected so that $(r_h^k + P_h^k V_{h+1}^k)(\cdot, \cdot) \in \mathcal{F}_h^k$ with high probability, $\mathcal{Z}_h^k$ consists of the collected samples (line 5), and Δ_h^k is the local variation budget defined by $\Delta_h^k =$

$$\sum_{\ell=\tau}^{k-1} \sup_{x,a} |\left(r_h^k - r_h^\ell\right)(x, a)| + H \sum_{\ell=\tau}^{k-1} \sup_{x,a} \left\| \left(P_h^k - P_h^\ell\right)(\cdot|x, a) \right\|_1.$$

Note that the complexity of a bonus function could be high as it is defined by the dataset $\mathcal{Z}_h^k$ whose size can be as large as W. We adopt importance sampling technique in [44] to reduce the size of the dataset. Moreover, the data samples in $\mathcal{Z}_h^k$ are collected from nonstationary environment, we include an additional term of local variation budget Δ_h^k in the definition of $\mathcal{F}_h^k$. Intuitively, the local variation budget Δ_h^k captures the discrepancy between current episode k and previous episodes in the current epoch. By incorporating a term involving Δ_h^k in the design of $\mathcal{F}_h^k$, we ensure the true action-value function of the kth episode lies within the confidence set $\mathcal{F}_h^k$ with high probability. The formal definition of bonus term b_h^k and the selection of β is deferred to Appendix B.

Restart mechanism: We use epoch restart mechanism to handle the nonstationary environment. Specifically, we

restart every W episodes as illustrated in the outer loop of Algorithm 4 (line 4), and the estimate of the state-action value function are calculated only by the samples collected in the current epoch, independent of all previous epochs. Note that while in general the epoch length W can vary for different epochs, we consider a fixed length and the corresponding dynamic regret upper bound in this work.

Compared to the confidence-set based algorithm SW-OPEA, which relies on an computational inefficient oracle to select the optimistic state-action value function within the confidence set. Instead, our algorithm is based on the popular UCB-based approach, which simplified the algorithm design and can be potentially implemented computationally efficiently [44].

B. Theoretical Guarantees

In this section, we provide the theoretical guarantee for Algorithm 4, and defer proofs to Appendix C.

For clarity, assume K/W is an integer throughout this section. The variation budget of an epoch $w \in [1{:}K/W]$ is defined as

$$\Delta_h^{(w)} = \sum_{\ell=w(W-1)+1}^{wW} \sup_{x,a} |\left(r_h^k - r_h^\ell\right)(x,a)| + H \sum_{\ell=w(W-1)+1}^{wW} \sup_{x,a} \left\| \left(P_h^k - P_h^\ell\right)(\cdot|x,a) \right\|_1.$$

The dynamic regret of LSVI-Nonstationary is characterized in the following theorem.

Theorem 2 (Dynamic Regret of LSVI-Nonstationary): Under Assumption 1, Assumption 2 and Assumption 4, with probability at least $1-\delta$, LSVI-Nonstationary achieves a dynamic regret bound of $\mathrm{D-Regret}(K) =$

$$\widetilde{O}\left(\frac{4H^2Kd_m}{W} + \frac{KH^2}{\sqrt{W}}\sqrt{\iota} + \sqrt{d_m HW}\sum_{h=1}^{H}\sum_{w=1}^{K/W}\sqrt{\Delta_h^{(w)}}\right)$$

where $d_m = \sup_h \dim_E(\mathcal{F}_h, 1/W)$ and

$$\iota \le c \sup_h \log^3(T/\delta)\dim_E^2\left(\mathcal{F}_h, \delta/16W^2\right) \ln(\mathcal{N}(\mathcal{F}_h, \delta/576W)/\delta)\ln\left(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \frac{1}{16\sqrt{W/\delta}})W/\delta\right).$$

for some absolute constant $c > 0$.

Note that the last term depends on the length of the restart epoch W, and the dynamic regret upper bound can be further optimized by setting appropriate W. We adopt the same definition of the average variation budget in transitions L_P and rewards L_r defined in (6) and (7).

The following corollary characterize the dynamic regret by optimizing the restart epoch length W based on the average variation budget L for both nonstationary tabular and nonstationary linear MDPs.

Corollary 2: Consider the same condition as in Theorem 2. For tabular MDPs with $\widetilde{d} = |\mathcal{S}||\mathcal{A}|$, let $\mathcal{F}_h$ be the entire function space of $\mathcal{S}\times\mathcal{A} \mapsto [0, H-h+1]$ for $h \in [H]$. Since $\mathcal{S}, \mathcal{A}$ are finite, for $\varepsilon > 0$, we have $\dim_E(\mathcal{F}_h, \varepsilon) \le \widetilde{d}$, $\log(\mathcal{N}(\mathcal{F}, \varepsilon)) = \widetilde{O}(\widetilde{d})$, and $\log(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \varepsilon)) = O(\log(\widetilde{d}))$, and the dynamic regret is bounded by

$$\widetilde{O}\left(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}^{\frac{3}{2}} + HT\widetilde{d}L_P^{\frac{1}{4}} + H^{\frac{3}{4}}T\widetilde{d}L_r^{\frac{1}{4}}\right).$$

For linear MDPs with feature dimension $\widetilde{d}$, $\dim_E(\mathcal{F}_h, \varepsilon) \le \widetilde{O}(\widetilde{d})$, $\log(\mathcal{N}(\mathcal{F}, \varepsilon)) = \widetilde{O}(\widetilde{d})$, and $\log(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \varepsilon)) = \widetilde{O}(\widetilde{d})$, and the dynamic regret is bounded by

$$\widetilde{O}\left(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}^{\frac{3}{2}} + HT\widetilde{d}^{\frac{5}{4}}L_P^{\frac{1}{4}} + H^{\frac{3}{4}}T\widetilde{d}^{\frac{5}{4}}L_r^{\frac{1}{4}}\right).$$

Compare to SW-OPEA: Under nonstationary MDPs with general function approximation, we compare the dynamic regret upper bound of our UCB-type algorithm to the dynamic regret bound of the confidence-set based algorithm SW-OPEA. For both nonstationary tabular and linear MDPs, SW-OPEA gives $\widetilde{O}(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d} + HT\widetilde{d}^{\frac{3}{4}}L_P^{\frac{1}{4}} + H^{\frac{3}{4}}T\widetilde{d}^{\frac{3}{4}}L_r^{\frac{1}{4}})$, where $\widetilde{d}$ equals $|\mathcal{S}||\mathcal{A}|$ for tabular MDPs and equals the feature dimension for linear MDPs. We see that the dynamic regret bound of UCB-type algorithm matches that of confidence-set based algorithm in horizon H as well as average variation budgets L_P and L_r while perform slightly worse in terms of $\widetilde{d}$. Similarly to the static MDPs [44], a more refined analysis specialized to the tabular and linear setting can potentially improve the dynamic regret bound. We would like to point out that our algorithm and analysis handles the nonstationary MDPs with general function approximation, which is much harder than and contains the nonstationary tabular and linear MDPs.

V. Conclusion

In this paper, we investigate two approaches for the nonstationary MDPs with general function approximation: confidence-set based algorithm and UCB-type algorithm. Based on the notion of dynamic Eluder dimension, the confidence-set based algorithm SW-OPEA incorporates the sliding window mechanism, and a novel design for the confidence set. The dynamic regret of SW-OPEA is shown to outperform the existing algorithms in nonstationary linear and tabular MDPs in the small variation budget regime. To alleviate the computational inefficiency challenge for the oracle used to select the optimistic state-action value function within an confidence set in the confidence-set based algorithm, we further propose a different UCB-type algorithm, which follows the popular LSVI framework. To handle nonstationarity, the UCB-type algorithm LSVI-Nonstationary features the restart mechanism, and the novel design of the bonus term to ensure the optimism of the learned state-action value function. LSVI-Nonstationary performs no worse than the confidence-set based algorithm SW-OPEA, while considerably simplifies the algorithm design. Our future directions include studying the unknown variation budget scenario and establishing lower bound for nonstationary MDPs with general function approximation.

Appendix A

We provide a sketch of the proof of Theorem 1, and the details of the proof of Theorem 1 and the proof of Corollary 1

can be found in [1]. The preliminary step is to decompose the dynamic regret of SW-OPEA into three terms as follows:

$$\begin{aligned}
&\mathrm{D-Regret}(k) \le H \\
&+ \underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}\left[\left(r_h^{t-1}-r_h^t\right)(x_h,a_h)\right]}_{\text{(I)}} \\
&+ \underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H}\left[\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t))}\right]\left[r_h^t(x_h,a_h)\right]}_{\text{(II)}} \\
&+ \underbrace{\sum_{t=1}^{k}\left(V^{\pi^{(*,t-1)}}_{1;(*,t-1)} - V^{\pi^t}_{1;(*,t-1)}\right)(x_1)}_{\text{(III)}}. \qquad (9)
\end{aligned}$$

Term (I) can be bounded by $\Delta_R(k)$ by the definition of the variation budget of rewards (4). In the sequel, we aim to bound (II) in step II and bound (III) in the remaining steps.

Step I: We introduce a novel auxiliary MDP to help bound term (II). For a fixed tuple $(k,h) \in [K]\times[H]$, we design an episodic MDP $(\mathcal{S},\mathcal{A},H,P^k,\widetilde{r},x_1)$ with reward $\widetilde{r}_{h'} = r_h^k(x,a)\mathbf{1}\{h'=h\}$ and the corresponding state value function of policy $\{\pi_{h'}\}_{h'\in[H]}$ is defined as $\widetilde{V}^{\pi}_{h';,(*,k)}$. Then, by [1, Lemma C.1], we have

$$\begin{aligned}
\text{(II)} &\le \left[\widetilde{V}^{\pi^k}_{1;(*,k-1)} - \widetilde{V}^{\pi^k}_{1;(*,k)}\right](x_1) \\
&\le \sum_{i=1}^{h-1}\sup_{x,a}\left\|\left(P_h^k - P_h^{k-1}\right)(\cdot|x,a)\right\|_1.
\end{aligned}$$

Replacing k by t, and summing over $t\in[k]$, $h\in[H]$ gives

$$\begin{aligned}
\text{(II)} &\le \sum_{t=1}^{k}\sum_{h=1}^{H}\sup_{x,a}\sum_{i=1}^{h-1}\left\|\left(P_i^{t-1}-P_i^t\right)(\cdot|x,a)\right\|_1 \\
&\le \sum_{h=1}^{H}\left(\sum_{t=1}^{k}\sum_{i=1}^{H}\sup_{x,a}\left\|\left(P_i^{t-1}-P_i^t\right)(\cdot|x,a)\right\|_1\right) \le H\Delta_P(k).
\end{aligned}$$

Step II: This step together with the next step establishes important properties to bound term (III) in step IV.

First, we develop the following crucial probability distribution shift lemma, which will handle the transition kernel variation in nonstationary MDPs.

Lemma 1 [1, Lemma D.2]: Suppose P and Q are two probability distributions of a random variable x, then

$$\begin{aligned}
&\left|\left(\mathop{\mathbb{E}}_{x\sim P} f(x)+\mathbb{E}g_1(y)-C\right)^2 - \left(\mathop{\mathbb{E}}_{x\sim Q} f(x)+\mathbb{E}g_2(y)-C\right)^2\right| \\
&\quad\le (2f_m+2g_m+2|C|)f_m \mathrm{TV}(P,Q),
\end{aligned}$$

where $f_m = \sup_x|f(x)|$, $g_m = \max_{i=1,2}\sup_y g_i(y)$.

Next, we show that $Q^*_{(*,k)}$, the optimal state-action value function at step h, lies in the confidence set $\mathcal{B}^k$ for all $k\in[K]$ with high probability. The argument is proved by the martingale concentration and the confidence set we design. Technically, we define

$$\#_{k,h}(x_h^t,a_h^t) = r_h^k(x_h^t,a_h^t) + \mathop{\mathbb{E}}_{x'\sim P_h^t(\cdot|x_h^t,a_h^t)}\max_{a'\in\mathcal{A}} Q_{h+1;(*,k)}(x',a'),$$

to form an appropriate martingale difference, which is similar to the h-th step Bellman update of the state-action value function in episode k except that the expectation is taken with respect to P_h^t instead of P_h^k. By Lemma 1, the cumulative mismatch during the sliding window between $\#_{k,h}(x_h^t,a_h^t)$ and the h-step Bellman update of state-action value function in episode k is captured by the local path length $\Delta_P^w(k,h)$ and $\Delta_R^w(k,h)$. Finally, by the design of confidence set $\mathcal{B}^k$, we can show that $Q^*_{(*,k)}\in\mathcal{B}^k$.

Given $Q^*_{(*,k)}\in\mathcal{B}^k$ for all $k\in[K]$, the optimistic planning step (Line 3) guarantees that $V^*_{1;(*,k-1)}(x_1) \le \sup_a f_1^k(x_1,a)$ for every episode $k\in[K]$. Combining the optimism and the generalized policy loss decomposition [1, Lemma C.8], we have

$$\begin{aligned}
\text{(III)} &\le \sum_{t=1}^{k}\left(\max_{a\in\mathcal{A}} f_1^t(x_1,a) - V^{\pi^t}_{1;(*,t-1)}(x_1)\right) \\
&\le \sum_{h=1}^{H}\sum_{t=1}^{k}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}\left[\left(f_h^t - \mathcal{T}_h^{t-1}f_{h+1}^t\right)(x_h,a_h)\right]. \qquad (10)
\end{aligned}$$

Step III: We will show the sharpness of our confidence set $\mathcal{B}^k$. Under the construction of $\mathcal{B}^k$, f^k selected from $\mathcal{B}^{k-1}$ is guaranteed to have small loss $\mathcal{L}_{\mathcal{D}_h}(f_h^k, f_h^{h+1})$. Note that the data used in episode k are collected by executing π^i for one episode for all $i\in[1\vee(k-w),k]$, by the concentration and the completeness assumption. We can show in [1, Lemma D.4] that with high probability, for all $(k,h)\in[K]\times[H]$,

$$\begin{aligned}
&\sum_{t=1\vee(k-w-1)}^{k-1}\left[f_h^k(x_h^t,a_h^t) - r_h^t - \mathop{\mathbb{E}}_{x'\sim P_h^{k-1}(x_h^t,a_h^t)}\max_{a'\in\mathcal{A}} f_{h+1}^k(x',a')\right]^2 \\
&\quad\le 6H^2\Delta_P^w(k-1,h) + 6H\Delta_R^w(k-1,h) + \mathcal{O}(\beta). \qquad (11)
\end{aligned}$$

Technically, we define the following helpful random variable

$$\#^f_{k,h}(x_h^t,a_h^t) = r_h^k(x_h^t,a_h^t) + \mathop{\mathbb{E}}_{x'\sim P_h^t(x_h^t,a_h^t)}\max_{a'\in\mathcal{A}} f_{h+1}(x',a')$$

to form an appropriate martingale and obtain the martingale concentration result. Then, applying our probability distribution shift lemma (Lemma 1), the definition of $\mathcal{B}^k$ and the completeness assumption gives (11).

Step IV: We establish the relationship between (10) and (11). Specifically, we aim to upper bound (10) given (11) holds. Note that their forms are similar except that the latter is the squared Bellman error, and the data (s_t,a_t) is taken under policy π^i for $i=[1\vee(k-w):k-1]$. It turns out that the DBE dimension plays an important role in connecting these two terms, as summarized in the following lemma.

Lemma 2 [1, Lemma 5.5]: Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)|\le C$ for all $(g,x)\in\Phi\times\mathcal{X}$, and a family of probability measures Π over $\mathcal{X}$. Suppose $\{\phi_k\}_{k\in[K]}\subseteq\Phi$ and $\{\mu_k\}_{k\in[K]}\subseteq\Pi$ satisfy that for all $k\in[K]$, $\sum_{t=1\vee(k-w-1)}^{k-1}(\mathbb{E}_{x\sim\mu_t}[\phi_k(x)])^2\le\beta$. Then for all $k\in[K]$ and $\omega>0$, $\sum_{t=1\vee(k-w)}^{k}|\mathbb{E}_{x\sim\mu_t}[\phi_t(x)]|$ is upper bounded by

$$\begin{aligned}
&\mathcal{O}\Big(\sqrt{\dim_{\mathrm{DE}}(\Phi,\Pi,\theta)\beta[k\wedge(w+1)]} \\
&\qquad + \min\{w+1,k,\dim_{\mathrm{DE}}(\Phi,\Pi,\theta)\}C + [k\wedge(w+1)]\theta\Big).
\end{aligned}$$

Based on the DBE dimension and Lemma 2, we are ready to bound (III) via term (10). By choosing Φ to be the function class of Bellman residuals, and μ_k to be the distribution under policy π^k, term (III) is upper bounded by

$$\sum_{h=1}^{H}\sum_{t=1}^{k}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}\Big[\Big(f_h^t-\mathcal{T}_h^{t-1}f_{h+1}^t\Big)(x_h,a_h)\Big]$$
$$\le \mathcal{O}\Big(H\sqrt{w}+\frac{H^2k}{\sqrt{w}}\sqrt{\dim_{\mathrm{DBE}}\Big(\mathcal{F},\mathcal{D}_\Delta,\sqrt{1/K}\Big)\log\frac{KH|\mathcal{F}|}{\delta}}$$
$$+\frac{Hk}{\sqrt{w}}\sqrt{\dim_{\mathrm{DBE}}\Big(\mathcal{F},\mathcal{D}_\Delta,\sqrt{1/K}\Big)\sum_{h=1}^{H}\sqrt{\sup_{k\in[K]}\Delta_P^w(k,h)}}\Big).$$

Combining all the steps, the dynamic regret of our algorithm SW-OPEA is

$$\mathrm{D-Regret}(k)\le\Delta_R(k)+H\Delta_P(k)+\mathcal{O}\Big(H\sqrt{w}$$
$$\frac{H^2k}{\sqrt{w}}\sqrt{d\log\big[KH|\mathcal{G}|/\delta\big]}+\frac{H^2k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_P^w(t,h)}\Big)$$

where we suppress the first term H in (9) since it is dominated by the fourth term herein.

Appendix B
The Stable Bonus Function via Importance Sampling

The framework of subsampling a given dataset in RL was first established by [44], which builds upon the sensitivity sampling technique [48], [49], [50]. For sake of completeness, we provide the formal definition of sensitivity and important results to be used in our analysis, and the proofs are omitted as they are similar to those in [44].

We begin with the definition of sensitivity function.

Definition 5 [44]: For a given set of state-action pairs $\mathcal{Z}\subseteq\mathcal{S}\times\mathcal{A}$ and a function class $\mathcal{F}$, for each $z\in\mathcal{Z}$, define the λ-sensitivity of (s,a) with respect to $\mathcal{Z}$ and $\mathcal{F}$ as

$$\mathrm{sensitivity}_{\mathcal{Z},\mathcal{F},\lambda}(x,a)=\sup_{f,f'\in\mathcal{F},\|f-f'\|_{\mathcal{Z}}^2\ge\lambda}\frac{\big(f(x,a)-f'(x,a)\big)^2}{\|f-f'\|_{\mathcal{Z}}^2}.$$

λ-sensitivity measures the importance of data points in $\mathcal{Z}$ which contributes the most to $\|f-f'\|_{\mathcal{Z}}^2$ for $f,f'\in\mathcal{F}$ whenever $\|f-f'\|_{\mathcal{Z}}^2\ge\lambda$. The algorithm to subsample the dataset is provided in Algorithm 5, where the sampling probability for each state-action pair is proportional to the sensitivity.

The next lemma shows the relations between the subsampled dataset and the original dataset.

Lemma 3 [44, Proposition 1]: With probability at least $1-\delta$, the size of $\mathcal{Z}'$ returned by Algorithm 6 satisfies $|\mathcal{Z}'|\le 4|\mathcal{Z}|/\delta$, the number of distinct elements in $\mathcal{Z}$ is at most

$$1728\dim_E(\mathcal{F},\lambda/|\mathcal{Z}|)\log\Big((H+1)^2|\mathcal{Z}|/\lambda\Big)\ln(|\mathcal{Z}|)$$
$$\ln\Big(4\mathcal{N}(\mathcal{F},\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|})/\delta\Big)/\varepsilon^2,$$

and for any $f,f'\in\mathcal{F}$,

$$(1-\varepsilon)\|f-f'\|_{\mathcal{Z}}^2-2\lambda\le\|f-f'\|_{\mathcal{Z}'}^2$$
$$\le(1+\varepsilon)\|f-f'\|_{\mathcal{Z}}^2+8|\mathcal{Z}|\lambda/\delta.$$

Algorithm 5 Sensitivity-Sampling($\mathcal{F},\mathcal{Z},\lambda,\varepsilon,\delta$)

Input: function class $\mathcal{F}$, reference function $\bar{f}\in\mathcal{F}$, set of state-action pairs $\mathcal{Z}\subseteq\mathcal{S}\times\mathcal{A}$, and failure probability $\delta\in(0,1)$.
Initialize $\mathcal{Z}'\leftarrow\varnothing$.
For each $z\in\mathcal{Z}$, let p_z to be the smallest real number such that $1/p_z$ is an integer and

$$p_z\ge\min\{1,\mathrm{sensitivity}_{\mathcal{Z},\mathcal{F},\lambda}(z)\cdot 72\ln\Big(4\mathcal{N}(\mathcal{F},\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}/\delta)\Big)/\varepsilon^2\}.$$

For each $z\in\mathcal{Z}$, independently add $1/p_z$ copies of z into $\mathcal{Z}'$ with probability p_z.
return $\mathcal{Z}'$.

Algorithm 6 Bonus($\mathcal{F},\bar{f},\mathcal{Z},\delta,j$)

Input: function class $\mathcal{F}$, set of state-action pairs $\mathcal{Z}\subseteq\mathcal{S}\times\mathcal{A}$, accuracy parameters $\lambda,\varepsilon>0$ and failure probability $\delta\in(0,1)$.
$\mathcal{Z}'\leftarrow$ Sensitivity-sampling($\mathcal{F},\mathcal{Z},\delta/(16W),1/2,\delta$).
$\mathcal{Z}'\leftarrow\varnothing$ if $|\mathcal{Z}'|\ge 4T/\delta$ or the number of distinct elements in $\mathcal{Z}'$ exceeds

$$6912\dim_E\Big(\mathcal{F},\delta/(16W^2)\Big)\log\Big(64H^2W^2/\delta\Big)\ln W\ln T\cdot\ln(\mathcal{N}(\mathcal{F},\delta/576W)/\delta)$$

Let $\hat{f}\in\mathcal{C}(\mathcal{F},1/(8\sqrt{4W/\delta}))$ be such that $\big\|\bar{f}-\hat{f}\big\|_\infty\le 1/(8\sqrt{4W/\delta})$.
$\hat{\mathcal{Z}}\leftarrow\varnothing$.
for $z\in\mathcal{Z}'$ **do**
  Let $\hat{z}\in\mathcal{C}(\mathcal{S}\times\mathcal{A},1/(8\sqrt{4W/\delta}))$ be such that $\sup_{f,f'\in\mathcal{F}}|f(z)-f'(z)|\le 1/(8\sqrt{4W/\delta})$.
  $\hat{\mathcal{Z}}\leftarrow\hat{\mathcal{Z}}\cup\{\hat{z}\}$.
end for
return $\hat{w}(\cdot,\cdot):=w(\hat{\mathcal{F}};\cdot,\cdot)$, where $\hat{F}=\{f\in\mathcal{F}:\big\|f-\hat{f}\big\|_{\hat{\mathcal{Z}}}^2\le 3\beta(\mathcal{F},\delta)+2\}$ where

$$\beta(\mathcal{F},\delta)=c'\Big(H\sqrt{\log(T/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)+\log|\mathcal{W}_h|}+\sqrt{H\Delta_h^{(j)}}\Big)^2$$

for some absolute constant $c'>0$.

Equipped with the subsampling procedure, we are able to formally define the stable bonus function in Algorithm 6. In line 10, the variation budget $\Delta_h^{(w)}$ is defined as

$$\Delta_h^{(w)}=\sum_{\ell=w(W-1)+1}^{wW}\sup_{x,a}\big|\Big(r_h^k-r_h^\ell\Big)(x,a)\big|$$
$$+H\sum_{\ell=w(W-1)+1}^{wW}\sup_{x,a}\Big\|\Big(P_h^k-P_h^\ell\Big)(\cdot|x,a)\Big\|_1.$$

At a high level, we first subsample the given dataset $\mathcal{Z}$, and define the confidence set based on the newly subsampled dataset and the reference function. Note that the subsampled dataset will be discarded if its size is too large, which is guaranteed to happen with low probability.

Based on Lemma 3, we have the following lemma, which is adapted from [44, Proposition 2] for nonstationary MDPs with restart epoch W.

Lemma 4: For Algorithm 6, suppose $|\mathcal{Z}| \le W$, the following statements hold:

- With probability at least $1-\delta/(16T)$,
$$w\big(\underline{\mathcal{F}}; x, a\big) \le \widehat{w}(x,a) \le w\big(\overline{\mathcal{F}}; x, a\big),$$
where $\underline{\mathcal{F}} = \{f \in \mathcal{F} : \|f-\bar{f}\|_{\mathcal{Z}}^2 \le \beta(\mathcal{F},\delta)\}$, and $\overline{\mathcal{F}} = \{f \in \mathcal{F} : \|f-\bar{f}\|_{\mathcal{Z}}^2 \le 9\beta(\mathcal{F},\delta)+12\}$.
- $\widehat{w}(\cdot,\cdot) \le \mathcal{W}$ for a function set $\mathcal{W}$ with
$$\begin{aligned}\log|\mathcal{W}| &\le 6912\dim_E\Big(\mathcal{F},\delta/(16W^2)\Big)\log\Big(16(H+1)^2W^2/\delta\Big)\\ &\quad\cdot\ln W\ln(64T\mathcal{N}(\mathcal{F},\delta/(576W)/\delta)\\ &\quad\cdot\log\Big(\mathcal{N}\Big(\mathcal{S}\times\mathcal{A},1/(8\sqrt{4W/\delta}\cdot 4W/\delta)\Big)\Big)\\ &\quad+\log\Big(\mathcal{N}\Big(\mathcal{S}\times\mathcal{A},1/(8\sqrt{4W/\delta})\Big)\Big)\\ &\le C\dim_E\Big(\mathcal{F},\delta/(16W^2)\Big)\cdot\log\Big(H^2W^2/\delta\Big)\cdot\ln W\\ &\quad\cdot\ln T\ln(\mathcal{N}(\mathcal{F},\delta/576W)/\delta)\\ &\quad\cdot\log(\mathcal{N}\Big(\mathcal{S}\times\mathcal{A},1/(8\sqrt{4W/\delta})\cdot 4W/\delta\Big)\end{aligned}$$
for some absolute constant $C>0$ when T is sufficiently large.

Appendix C
Proof of Theorem 2

Step I: We analyze the complexity of the stable bonus function. The framework of subsampling a given dataset in RL was first established by [44]. We adapt the analysis therein to our setting for a given epoch of length W. The main result is presented in Lemma 4.

Step II: This step shows that the state-action value function estimate $Q_h^k(\cdot,\cdot)$ in Algorithm 4 is an optimistic upper bound for the optimal state-action value function. Our new development lies in developing the single step optimization error for nonstationary MDPs, and the construction of the confidence set.

We first establish the single step optimization error bound in the following lemma.

Lemma 5 (Single Step Optimization Error): Consider fixed $(k,h) \in [K]\times[H]$. Denote τ as the first episode of an epoch containing episode k. Let
$$\mathcal{Z}_h^k = \Big\{\Big(x_h^\ell, a_h^\ell\Big)\Big\}_{\ell\in[\tau:k-1]}$$
as defined in Line 7 of Algorithm 4. For any $V:\mathcal{S}\mapsto[0,H-h]$, define
$$\mathcal{D}_{V;h}^k = \Big\{\Big(x_h^\ell, a_h^\ell, \widetilde{r}_h^\ell + V(x_{h+1}^\ell)\Big)\Big\}_{\ell\in[\tau,k-1]}$$
and
$$\widehat{f}_{V;h} = \arg\min_{f\in\mathcal{F}}\|f\|_{\mathcal{D}_{V;h}^k}^2.$$
For any $h\in[H]$, $V:\mathcal{S}\mapsto[0,H-h]$ and $\delta\in(0,1)$, there is an event $\mathcal{I}_{h,V,\delta}$ which holds with probability at least $1-\delta$, such that conditioned on $\mathcal{I}_{h,V,\delta}$, for any $V':\mathcal{S}\mapsto[0,H-h]$ with $\|V'-V\|_\infty \le 1/W$, we have
$$\begin{aligned}&\Big\|\widehat{f}_{V'}(\cdot,\cdot) - r_h^k(\cdot,\cdot) - \sum_{s'\in\mathcal{S}}P_h^k(s'|\cdot,\cdot)V'(s')\Big\|_{\mathcal{Z}_h^k}\\ &\quad\le c\Big(H\cdot\sqrt{\log(1/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}+\sqrt{H\Delta_k}\Big)\end{aligned}$$
for some absolute constant $c>0$, where
$$\Delta_h^k = \sum_{\ell=\tau}^{k-1}\sup_{x,a}|\Big(r_h^k - r_h^\ell\Big)(x,a)| + H\sum_{\ell=\tau}^{k-1}\sup_{x,a}|\Big(P_h^k - P_h^\ell\Big)(x,a)|.$$

Proof: Consider a fixed $V:\mathcal{S}\mapsto[0,H-h]$. For any $f\in\mathcal{F}_h$, consider $\sum_{\ell=\tau}^{k-1}\xi_h^\ell(f)$ where
$$\begin{aligned}\xi_h^\ell(f) &= 2\Big(f - r_h^k - \mathbb{P}_h^kV\Big)\Big(x_h^\ell,a_h^\ell\Big)\cdot\\ &\quad+\Big(r_h^\ell\Big(x_h^\ell,a_h^\ell\Big)\Big(\mathbb{P}_h^\ell V\Big)\Big(x_h^\ell,a_h^\ell\Big) - \widetilde{r}_h^\ell - V\Big(x_{h+1}^\ell\Big)\Big).\end{aligned}$$
For any $(\ell,h)\in[k-1]\times[H]$, define $\mathbb{F}_h^\ell$ as the filtration induced by the sequence
$$\Big\{\Big(x_{h'}^t,a_{h'}^t\Big)\}_{(t,h')\in[\ell-1]\times[H]}\cup\{\Big(x_1^\ell,a_1^\ell\Big),\ldots,\Big(x_{h-1}^\ell,a_{h-1}^\ell\Big)\Big\}.$$
Then, $\mathbb{E}[\xi_h^\ell(f)|\mathbb{F}_h^\ell]=0$ and
$$|\xi_h^\ell(f)| \le 2(H-h+1)\Big|\Big(f-r_h^k-\mathbb{P}_h^kV\Big)\Big(x_h^\ell,a_h^\ell\Big)\Big|.$$
By Azuma-Hoeffding's inequality, we have
$$\mathbb{P}\left[\left|\sum_{\ell=\tau}^{k-1}\xi_h^\ell(f)\right|\ge\varepsilon\right] \le 2e^{\left(-\frac{\varepsilon^2}{8(H-h+1)\|f-r_h^k-\mathbb{P}_h^kV\|_{\mathcal{Z}_h^k}^2}\right)}.$$
Let
$$\begin{aligned}\varepsilon &= \left(8(H-h+1)^2\log\Big(\frac{2\mathcal{N}(\mathcal{F}_h,1/W)}{\delta}\Big)\Big\|f-r_h^k-\mathbb{P}_h^kV\Big\|_{\mathcal{Z}_h^k}^2\right)^{1/2}\\ &\le 4(H-h+1)\Big\|f-r_h^k-\mathbb{P}_h^kV\Big\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}.\end{aligned}$$
Then, with probability at least $1-\delta$, for all $f\in\mathcal{C}(\mathcal{F}_h,1/W)$,
$$\begin{aligned}\left|\sum_{\ell=\tau}^{k-1}\xi_h^\ell(f)\right| &\le 4(H-h+1)\Big\|f-r_h^k-\mathbb{P}_h^kV\Big\|_{\mathcal{Z}_h^k}\\ &\quad\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}.\end{aligned}$$
Define the above event to be $\mathcal{I}_{h,V,\delta}$, and we condition on this event for the rest of the proof.

For all $f\in\mathcal{F}_h$, there exists $g\in\mathcal{C}(\mathcal{F}_h,1/W)$, such that $\|f-g\|_\infty\le 1/W$, and we have
$$\begin{aligned}\left|\sum_{\ell=\tau}^{k-1}\xi_h^\ell(f)\right| &\le \left|\sum_{\ell=\tau}^{k-1}\xi_h^\ell(g)\right| + 2(H-h+1)\\ &\le 4(H-h+1)\Big\|g-r_h^k-\mathbb{P}_h^kV\Big\|_{\mathcal{Z}_h^k}\\ &\quad\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}+2(H-h+1)\\ &\le 4(H-h+1)\left(\Big\|f-r_h^k-\mathbb{P}_h^kV\Big\|_{\mathcal{Z}_h^k}+1\right)\end{aligned}$$

$$\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}+2(H-h+1).$$

Consider $V':\mathcal{S}\mapsto[0,H-h]$ with $\|V'-V\|\le 1/W$. We have

$$\left\|r_h^k+\mathbb{P}_h^kV'-r_h^k-\mathbb{P}_h^kV\right\|_\infty\le\|V'-V\|_\infty\le 1/W.$$

Note first that for any $f,g\in\mathcal{F}$, we have

$$\begin{aligned}\|f\|_{\mathcal{D}^k_{V';h}}-\|g\|_{\mathcal{D}^k_{V';h}}=\|f-g\|_{\mathcal{Z}_h^k}&\\+2\sum_{\ell=\tau}^{k-1}\Big(f(x_h^\ell,a_h^\ell)-g(x_h^\ell,a_h^\ell)\Big)&\\\cdot\Big(g(x_h^\ell,a_h^\ell)-\widetilde{r}_h^\ell-V'(x_{h+1}^\ell)\Big).&\end{aligned}$$

Replacing g with $r_h^k+\mathbb{P}_h^kV'$ gives

$$\begin{aligned}\|f\|^2_{\mathcal{D}^k_{V';h}}-\left\|r_h^k+\mathbb{P}_h^kV'\right\|^2_{\mathcal{D}^k_{V';h}}&=\left\|f-r_h^k-\mathbb{P}_h^kV'\right\|^2_{\mathcal{Z}_h^k}\\&+2\sum_{\ell=\tau}^{k-1}\Big(f-r_h^k-\mathbb{P}_h^kV'\Big)\Big(x_h^\ell,a_h^\ell\Big)\\&\cdot\Big(r_h^\ell(x_h^\ell,a_h^\ell)+(\mathbb{P}_h^\ell V')(x_h^\ell,a_h^\ell)-\widetilde{r}_h^\ell-V'(x_{h+1}^\ell)\Big)\qquad(I_2)\\&+2\sum_{\ell=\tau}^{k-1}\Big(f-r_h^k-\mathbb{P}_h^kV'\Big)\Big(x_h^\ell,a_h^\ell\Big)\cdot(r_h^k(x_h^\ell,a_h^\ell)\\&+(\mathbb{P}_h^kV')(x_h^\ell,a_h^\ell)-r_h^\ell(x_h^\ell,a_h^\ell)-(\mathbb{P}_h^\ell V')(x_h^\ell,a_h^\ell))\qquad(I_3)\end{aligned}$$

For the second term I_2, we have

$$\begin{aligned}I_2&\ge 2\sum_{\ell=\tau}^{k-1}\Big(f-r_h^k-\mathbb{P}_h^kV\Big)\Big(x_h^\ell,a_h^\ell\Big)(r_h^k(x_h^\ell,a_h^\ell)\\&\quad+(\mathbb{P}_h^kV)(x_h^\ell,a_h^\ell)-\widetilde{r}_h^\ell-V(x_{h+1}^\ell))-4(H-h+1)\\&=\sum_{\ell=\tau}^{k-1}\xi_h^\ell(f)-4(H-h+1)\\&\ge-4(H-h+1)\left(\left\|f-r_h^k-\mathbb{P}_h^kV'\right\|_{\mathcal{Z}_h^k}+2\right)\\&\quad\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}-6(H-h+1).\end{aligned}$$

For the third term I_3, we have

$$\begin{aligned}I_3&\ge-2(H-h+1)\left(\sum_{\ell=\tau}^{k-1}\sup_{x,a}\left|\Big(r_h^k-r_h^\ell\Big)(x,a)\right|\right.\\&\qquad\left.+H\sum_{\ell=\tau}^{k-1}\sup_{x,a}\left|\Big(P_h^k-P_h^\ell\Big)(x,a)\right|\right)\\&=-2(H-h+1)\Delta_h^k.\end{aligned}$$

Since $\widehat{f}_{V;h}=\arg\min_{f\in\mathcal{F}}\|f\|^2_{\mathcal{D}^k_{V;h}}$, we have

$$\begin{aligned}0&\ge\left\|\widehat{f}_{V;h}\right\|^2_{\mathcal{D}^2_{V';h}}-\left\|r_h^k+\mathbb{P}_h^kV'\right\|^2_{\mathcal{D}^2_{V';h}}\\&\ge\left\|\widehat{f}_{V;h}-r_h^k-\mathbb{P}_h^kV'\right\|^2_{\mathcal{Z}_h^k}\\&\quad-4(H-h+1)\left(\left\|\widehat{f}_{V;h}-r_h^k-\mathbb{P}_h^kV'\right\|_{\mathcal{Z}_h^k}+2\right)\\&\quad\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}\\&\quad-6(H-h+1)-2(H-h+1)\Delta_h^k.\end{aligned}$$

Solving the above inequality, we have

$$\begin{aligned}&\left\|\widehat{f}_{V;h}-r_h^k-\mathbb{P}_h^kV'\right\|_{\mathcal{Z}_h^k}\\&\le c'\left(H\cdot\sqrt{\log(1/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)}+\sqrt{H\Delta_h^k}\right)\end{aligned}$$

for some absolute constant $c'>0$. ■

Based on the single step optimization error of nonstationary MDPs, we devise the confidence set $\mathcal{F}_h^k$ which contains both the least square solution f_h^k and the one-step backup $r_h^k+P_h^kV_{h+1}^k$, as summarized in the following lemma.

Lemma 6 (Confidence Set): Define

$$\mathcal{F}_h^k=\left\{f\in\mathcal{F}_h:\left\|f-f_h^k\right\|_{\mathcal{Z}_h^k}\le\beta(\mathcal{F}_h,\delta)\right\},$$

where $\beta(\mathcal{F}_h,\delta)=c'(H\sqrt{\log(T/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)+\log|\mathcal{W}_h|}+\sqrt{H\Delta_h^W})^2$ for some absolute constant $c'>0$, and $\mathcal{W}_h$ is given in Lemma 4 with $\mathcal{F}$ replaced by $\mathcal{F}_h$. Then with probability at least $1-\delta/8$, for all $k,h\in[K]\times[H]$, we have

$$r_h^k+\mathbb{P}_h^kV_{h+1}^k\in\mathcal{F}_h^k.$$

Proof: For all $(k,h)\in[K]\times[H]$ the bonus function $b_h^k(\cdot,\cdot)=w(\mathcal{F}_h;\cdot,\cdot)\in\mathcal{W}$. Note that

$$\mathcal{Q}=\{\min\{f(\cdot,\cdot)+w(\cdot,\cdot),H\}:w\in\mathcal{W},f\in\mathcal{C}(\mathcal{F}_h,1/W)\}\cup\{0\}$$

is a $(1/W)$-cover of

$$Q_{h+1}^k(\cdot,\cdot)=\begin{cases}\min\{f_{h+1}^k(\cdot,\cdot)+b_{h+1}^k,H\}, & h<H,\\ 0, & h=H.\end{cases}$$

In other words, there exists $q\in\mathcal{Q}$ such that $\|q-Q_{h+1}^k\|_\infty\le 1/W$, which implies

$$\mathcal{V}=\{\max_{a\in\mathcal{A}}q(\cdot,a):q\in\mathcal{Q}\}$$

is a $(1/W)$-cover of V_{h+1}^k with $\log|\mathcal{V}|\le\log|\mathcal{W}|+\log\mathcal{N}(\mathcal{F}_h,1/W)+1$. For each $V\in\mathcal{V}$, let $\mathcal{I}_{h,V,\delta/(8|\mathcal{V}|T)}$ be the event defined in Lemma 5. By Lemma 5, we have $\mathbb{P}[\cap_{V\in\mathcal{V}}\mathcal{I}_{h,\mathcal{V},\delta/(8|\mathcal{V}|T)}]\ge 1-\delta/(8T)$. We condition on $\cap_{V\in\mathcal{V}}\mathcal{I}_{h,\mathcal{V},\delta/(8|\mathcal{V}|T)}$ in the rest of the proof.

Since f_h^k is the solution of the optimization problem in Line 10 of Algorithm 1, i.e., $f_h^k=\min_{f\in\mathcal{F}}\|f\|^2_{\mathcal{D}_h^k}$. Let $V\in\mathcal{V}$ such that $\|V-V_{h+1}^k\|_\infty\le 1/W$. Thus, by Lemma 5, we have

$$\begin{aligned}&\left\|f_h^k-r_h^k-P_h^kV_{h+1}^k\right\|_{\mathcal{Z}_h^k}\le\\&\quad c'\left(H\cdot\sqrt{\log(T/\delta)+\log\mathcal{N}(\mathcal{F}_h,1/W)+\log|\mathcal{W}_h|}+\sqrt{H\Delta_h^k}\right)\end{aligned}$$

for some absolute constant c'. Therefore, by a union bound, for all $(k,h)\in[K]\times[H]$, we have $r_h^k+\mathbb{P}_h^kV_{h+1}^k\in\mathcal{F}_h^k$ with probability at least $1-\delta/8$. ■

Since the bonus term b_h^k is defined to be the width of confidence set $\mathcal{F}_h^k$, we conclude that $Q_h^k(\cdot,\cdot)$ defined by $\min\{H,(f_h^k+b_h^k(\cdot,\cdot))\}$ is an optimistic upper bound for $(r_h^k+P_h^kV_{h+1}^k)(\cdot,\cdot)$.

Lemma 7: With probability at least $1-\delta/4$, for all $(k,h) \in [K]\times[H]$, for all $(x,a)\in\mathcal{S}\times\mathcal{A}$,

$$Q_h^*(x,a) \le Q_h^k(x,a) \le r_h(x,a) + \left(P_h^k V_{h+1}^k\right)(x,a) + 2b_h^k(x,a).$$

Proof: For each $(k,h)\times[K]\times[H]$, define $\mathcal{F}_h^k = \{f \in \mathcal{F}_h : \|f - f_h^k\|_{\mathcal{Z}_h^k} \le \beta(\mathcal{F}_h,\delta)\}$. By Lemma 6, the event that for all $(k,h)\in[K]\times[H]$, $r_h^k + \mathbb{P}_h^k V_{h+1}^k \in \mathcal{F}_h^k$ holds with probability at least $1-\delta/8$. Moreover, by Lemma 4, the event $b_h^k(x,a) > w(\mathcal{F}_h^k;x,a)$ holds with probability at least $1-\delta/8$. We condition on those two events in the rest of the proof.

Note that

$$\max_{f\in\mathcal{F}_h^k} |f(x,a) - f_h^k(x,a)| \le w\left(\mathcal{F}_h^k;x,a\right) \le b_h^k(x,a).$$

Since $r_h^k + \mathbb{P}_h^k V_{h+1}^k \in \mathcal{F}_h^k$, we have

$$\left|r_h^k(x,a) + \left(\mathbb{P}_h^k V_{h+1}^k\right)(x,a) - f_h^k(x,a)\right| \le b_h^k(x,a).$$

Therefore,

$$\begin{aligned} Q_h^k(x,a) &\le f_h^k(x,a) + b_h^k(x,a) \\ &\le r_h^k(x,a) + \left(\mathbb{P}_h^k V_{h+1}^k\right)(x,a) + 2b_h^k(x,a). \end{aligned}$$

Next we show $Q_h^*(x,a) \le Q_h^k(x,a)$ by induction on h, When $h = H+1$, the desired inequality clearly holds. Suppose $Q_h^*(\cdot,\cdot) \le Q_{h+1}^k(\cdot,\cdot)$ for some h. Clearly, $V_{h+1}^*(\cdot) \le V_{h+1}^k(\cdot)$. Therefore, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} Q_h^*(x,a) &= r_h^k(x,a) + \left(\mathbb{P}_h^k V_{h+1}^*\right)(x,a) \\ &\le \min\left\{H, r_h^k(x,a) + \left(\mathbb{P}_h^k V_{h+1}^k\right)(x,a)\right\} \\ &\le \min\left\{H, f_h^k(x,a) + b_h^k(x,a)\right\} = Q_h^k(x,a). \end{aligned}$$

■

Step III: We decompose the dynamic regret and further bound it via Eluder dimension.

By standard regret decomposition for UCB-type algorithms, the dynamic regret is upper bounded by the summation of the bonus function, as shown in the following lemma.

Lemma 8: With probability at least $1-\delta/2$,

$$\mathrm{D-Regret}(K) \le 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k\left(x_h^k,a_h^k\right) + 4H\sqrt{KH\log(8/\delta)}.$$

Proof: Define $\xi_h^k = P_h^k(V_{h+1}^k - V_{h+1}^{\pi_k})(x_h^k,a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi_k})(x_{h+1}^k)$ and $\mathbb{F}_h^k$ as the filtration induced by $\{(x_{h'}^{k'},a_{h'}^{k'})\}_{(h',k')\in[H]\times[k-1]} \cup \{(x_1^k,a_1^k),\ldots,(x_h^k,a_h^k)\}$. Then

$$\mathbb{E}\left[\xi_h^k|\mathbb{F}_h^k\right] = 0 \text{and} |\xi_h^k| \le 2H.$$

By Azuma-Hoeffding'e inequality, with probability at least $1-\delta/4$,

$$\sum_{k=1}^{H}\sum_{h=1}^{H-1}\xi_h^k \le 4H\sqrt{KH\log(8/\delta)}.$$

We condition on the above event, and the event defined in Lemma 7 which holds with probability $1-\delta/4$. We have

$$\begin{aligned} \mathrm{D-Regret}(K) &= \sum_{k=1}^{K}\left(V_1^*\left(x_1^k\right) - V_1^{\pi_k}\left(x_1^k\right)\right) \\ &\le \sum_{k=1}^{K}\left(V_1^k\left(x_1^k\right) - V_1^{\pi_k}\left(x_1^k\right)\right) \\ &\le \sum_{k=1}^{K}\Big(r_1^k\left(x_1^k,a_1^k\right) + \left(P_1^k V_2^k\right)\left(x_1^k,a_1^k\right) + 2b_1^k\left(x_1^k,a_1^k\right) \\ &\quad - r_1^k\left(x_1^k,a_1^k\right) - \left(P_1^k V_2^{\pi_k}\right)\left(x_1^k,a_1^k\right)\Big) \\ &= \sum_{k=1}^{K}\left(P_1^k\left(V_2^k - V_2^{\pi_k}\right)\left(x_1^k,a_1^k\right) + 2b_1^k\left(x_1^k,a_1^k\right)\right) \\ &= \sum_{k=1}^{K}\left(\xi_1^k + \left(V_2^k - V_2^{\pi_k}\right)\left(x_2^k\right) + 2b_1^k\left(x_1^k,a_1^k\right)\right) \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H-1}\xi_h^k + 2\sum_{k=1}^{K}\sum_{h=1}^{H-1} b_h^k\left(x_h^k,a_h^k\right) \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H-1} b_h^k\left(x_h^k,a_h^k\right) + 4H\sqrt{KH\log(8/\delta)}. \end{aligned}$$

To bound the summation of the bonus function, we use a similar argument in [44] to show that the summation of bonus term can be upper bounded by the Eluder dimension of the function class $\mathcal{F}_h$. ■

Lemma 9: With probability at least $1-\delta/4$, for any $\varepsilon > 0$,

$$\sum_{k=1}^{K}\mathbf{1}\left\{b_h^k\left(x_h^k,a_h^k\right) > \varepsilon\right\} \le \left(\frac{c\beta(\mathcal{F}_h,\delta)}{\varepsilon^2} + 1\right)\dim_E(\mathcal{F}_h,\varepsilon),$$

for some absolute constant $c > 0$.

Proof: Let $\overline{\mathcal{F}}_h^k = \{f\in\mathcal{F}_h : \|f - f_h^k\|_{\mathcal{Z}_h^k}^2 \le 9\beta(\mathcal{F}_h,\delta) + 12\}$. By Lemma 4, the event that for all $(k,h)\in[K]\times[H]$, $b_h^k(\cdot,\cdot) \le w(\overline{\mathcal{F}}_h^k,\cdot,\cdot)$ holds with probability at least $1-\delta/4$. We condition on such event in the rest of the proof.

Let $\mathcal{L} = \{(x_{h'}^k,a_{h'}^k) : h' = h, b_h^k(x_h^k,a_h^k) > \varepsilon\}$ with $|\mathcal{L}| = L$. We show that there exists $(x_h^k,a_h^k)\in\mathcal{L}$ such that (x_h^k,a_h^k) is ε-dependent on at least $L/\dim_E(\mathcal{F}_h,\varepsilon)-1$ disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}$ if K is sufficiently large. Consider the following procedure: Let $\mathcal{L}_1,\ldots,\mathcal{L}_{L/\dim_E(\mathcal{F}_h,\varepsilon)-1}$ be $L/\dim_E(\mathcal{F}_h,\varepsilon)-1$ disjoint subsequences of $\mathcal{L}$ which are initially empty. Consider $(x_h^k,a_h^k)\cap\mathcal{L}$ for each $k\in[K]$ sequentially. Find a j such that (x_h^k,a_h^k) is ε-independent of $\mathcal{L}_j$ and then add (x_h^k,a_h^k) into $\mathcal{L}_j$. If such j does not exist, then the process terminates. By the definition of ε-dependence, $|\mathcal{L}_j| \le \dim_E(\mathcal{F}_h,\varepsilon)$ for all j. Therefore, (x_h^k,a_h^k) must be ε-dependent on at least $\lfloor L/\dim_E(\mathcal{F}_h,\varepsilon)\rfloor$ disjoint sequences in $\mathcal{Z}_h^k\cap\mathcal{L}$.

Note that since $(x_h^k,a_h^k)\in\mathcal{L}$, i.e., $b_h^k(x_h^k,a_h^k) > \varepsilon$, which implies there exists $f,f'\in\mathcal{F}_h$ with $\|f - f_h^k\|_{\mathcal{Z}_h^k}^2 \le 9\beta(\mathcal{F}_h,\delta) + 12$ and $\|f' - f_h^k\|_{\mathcal{Z}_h^k}^2 \le 9\beta(\mathcal{F}_h,\delta) + 12$ such that $\|f - f'\|_\infty > \epsilon$. By triangle inequality, $\|f - f'\|_{\mathcal{Z}_h^k}^2 \le 36\beta(\mathcal{F}_h,\delta) + 48$. Therefore

$$\lfloor L/\dim_E(\mathcal{F}_h,\varepsilon)\rfloor\varepsilon^2 \le \|f - f'\|_{\mathcal{Z}_h^k}^2 \le 36\beta(\mathcal{F}_h,\delta) + 48,$$

which gives $L \le (\frac{36\beta(\mathcal{F}_h,\delta)}{\varepsilon}+1)\dim_E(\mathcal{F}_h,\varepsilon)$.

Lemma 10: With probability at least $1-\delta/4$,

$$\sum_{k=1}^{W} b_h^k\Big(x_h^k, a_h^k\Big) \le 4H\dim_E(\mathcal{F}_h, 1/W) + c'\sqrt{\dim_E(\mathcal{F}_h,1/W)\cdot W\cdot\beta(\mathcal{F}_h,\delta)},$$

for some absolute constant c'.

Proof: Let $w_1 \ge \ldots \ge w_W$ be a permutation of $\{b_h^k(x_h^k,a_h^k)\}_{k\in[W]}$. By Lemma 9, for any $w_t \ge 1/W$, we have

$$t \le \left(\frac{c\beta(\mathcal{F}_h,\delta)}{w_t^2}+1\right)\dim_E(\mathcal{F}_h,w_t) \le \left(\frac{c\beta(\mathcal{F}_h,\delta)}{w_t^2}+1\right)\dim_E(\mathcal{F}_h,1/W),$$

which implies

$$w_t \le \left(\frac{t}{\dim_E(\mathcal{F}_h,1/W)}-1\right)^{-\frac{1}{2}}\cdot\sqrt{c\beta(\mathcal{F}_h,\delta)}.$$

Moreover, we have $w_t \le 4H$. Therefore

$$\sum_{t=1}^{W} w_t \le 4H\cdot\dim_E(\mathcal{F}_h,1/W) + \sum_{\dim_E(\mathcal{F}_h,1/W)\le t\le W}\left(\frac{t}{\dim_E(\mathcal{F}_h,1/W)}-1\right)^{-\frac{1}{2}}\sqrt{c\beta(\mathcal{F}_h,\delta)} \le 4H\dim_E(\mathcal{F}_h,1/W) + 2\sqrt{c\dim_E(\mathcal{F}_h,1/W)W\beta(\mathcal{F}_h,\delta)}.$$

■

Proof of Theorem 2. Combining Lemma 8 and Lemma 10 and the value of $\beta(\mathcal{F}_h,\delta)$, and summing over all epochs $w\in[1,K/W]$, we obtain the dynamic regret upper bound for our proposed algorithm LSVI-Nonstationary

$$\mathrm{D-Regret}(K) \le \sum_{h=1}^{H}\sum_{w=1}^{\lceil K/W\rceil}\sum_{t=w(W-1)+1}^{\min\{wW,K\}} b_h^t\big(x_h^t,a_h^t\big) \le \frac{4H^2Kd_m}{W} + \widetilde{O}\left(\frac{KH^2}{\sqrt{W}}\sqrt{\iota}+\sqrt{d_mHW}\sum_{h=1}^{H}\sum_{w=1}^{\lceil K/W\rceil}\sqrt{\Delta_h^{(w)}}\right),$$

where $d_m = \sup_h \dim_E(\mathcal{F}_h,1/W)$, and we use $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ in the second inequality.

A. Proof of Corollary 2

Recall the definition of average variation budget L_P and L_r. By Theorem 2, we have

$$\begin{aligned}&\mathrm{D-Regret}(K)\\ &= \widetilde{O}\left(\frac{4H^2Kd_m}{W}+\frac{KH^2}{\sqrt{W}}\sqrt{\iota}+\sqrt{d_mHW}\sum_{h=1}^{H}\sum_{w=1}^{K/W}\sqrt{\Delta_h^{(w)}}\right)\\ &\le \widetilde{O}\left(\frac{KH^2}{\sqrt{W}}\sqrt{\iota}+\sqrt{d_mHW}\sum_{h=1}^{H}\sum_{w=1}^{K/W}\sqrt{LW^2}\right)\\ &= \widetilde{O}\left(KH^2\iota^{\frac{1}{2}}W^{-\frac{1}{2}}+d_m^{\frac{1}{2}}KH^{\frac{3}{2}}\left(H^{\frac{1}{2}}L_P^{\frac{1}{2}}+L_r^{\frac{1}{2}}\right)W^{\frac{1}{2}}\right),\end{aligned}$$

■

where $d_m = \sup_h \dim_E(\mathcal{F}_h,1/W)$ and

$$\iota \le c'\sup_h \log^3(T/\delta)\dim_E^2\Big(\mathcal{F}_h,\delta/16W^2\Big)\ln(\mathcal{N}(\mathcal{F}_h,\delta/576W)/\delta)\cdot\ln\left(\mathcal{N}(\mathcal{S}\times\mathcal{A},\frac{1}{16\sqrt{W/\delta}})\cdot W/\delta\right).$$

If $\frac{\iota^{\frac{1}{2}}H^{\frac{1}{2}}}{d_m^{\frac{1}{2}}(\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}})} \ge K$, i.e., $\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}} \le \frac{\sqrt{H\iota}}{K\sqrt{d_m}}$, we select $W=K$ and we have

$$\mathrm{D-Regret}(K) \le \widetilde{O}\Big(H^2K^{\frac{1}{2}}\iota^{\frac{1}{2}}\Big).$$

If $\frac{\iota^{\frac{1}{2}}H^{\frac{1}{2}}}{d_m^{\frac{1}{2}}(\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}})} < K$, i.e., $\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}} > \frac{\sqrt{H\iota}}{K\sqrt{d_m}}$, select $W = \lceil\frac{\iota^{\frac{1}{2}}H^{\frac{1}{2}}}{d_m^{\frac{1}{2}}(\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}})}\rceil$ and we have

$$\mathrm{D-Regret}(K) \le \widetilde{O}\left(KH^2\iota^{\frac{1}{4}}d_m^{\frac{1}{4}}L_P^{\frac{1}{4}}+KH^{\frac{7}{4}}\iota^{\frac{1}{4}}d_m^{\frac{1}{4}}L_P^{\frac{1}{4}}\right).$$

Consider tabular MDPs with $\widetilde{d}=|\mathcal{S}||\mathcal{A}|$. Let $\mathcal{F}_h$ be the entire function space of $\mathcal{S}\times\mathcal{A}\mapsto[0,H-h+1]$ for $h\in[H]$. Since $\mathcal{S},\mathcal{A}$ are finite, for $\varepsilon>0$, we have $\dim_E(\mathcal{F}_h,\varepsilon)\le\widetilde{d}$, $\log(\mathcal{N}(\mathcal{F},\varepsilon))=\widetilde{O}(\widetilde{d})$, and $\log(\mathcal{N}(\mathcal{S}\times\mathcal{A},\varepsilon))=O(\log(\widetilde{d}))$, we have $\iota=\widetilde{O}(\widetilde{d}^3)$ and $d_m=\widetilde{O}(\widetilde{d})$. Therefore, when $\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}} > \frac{\sqrt{H\widetilde{d}}}{K}$, we have

$$\mathrm{D-Regret}(K) \le \widetilde{O}\left(KH^2\widetilde{d}L_P^{\frac{1}{4}}+KH^{\frac{7}{4}}\widetilde{d}L_P^{\frac{1}{4}}\right).$$

For linear MDPs with feature dimension $\widetilde{d}$, $\dim_E(\mathcal{F}_h,\varepsilon)\le\widetilde{O}(\widetilde{d})$, $\log(\mathcal{N}(\mathcal{F},\varepsilon))=\widetilde{O}(\widetilde{d})$, and $\log(\mathcal{N}(\mathcal{S}\times\mathcal{A},\varepsilon))=\widetilde{O}(\widetilde{d})$, we have $\iota=\widetilde{O}(\widetilde{d}^4)$ and $d_m=\widetilde{O}(\widetilde{d})$. Therefore, when $\sqrt{L_P}+\frac{\sqrt{L_r}}{\sqrt{H}} > \frac{\sqrt{H}\widetilde{d}^{\frac{3}{2}}}{K}$, we have

$$\mathrm{D-Regret}(K) \le \widetilde{O}\left(KH^2\widetilde{d}^{\frac{5}{4}}L_P^{\frac{1}{4}}+KH^{\frac{7}{4}}\widetilde{d}^{\frac{5}{4}}L_P^{\frac{1}{4}}\right).$$

REFERENCES

[1] S. Feng, M. Yin, R. Huang, Y.-X. Wang, J. Yang, and Y. Liang, "Non-stationary reinforcement learning under general function approximation," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 9976–10007.

[2] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[3] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, 2017, pp. 3389–3396.

[4] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[5] D. Silver et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," 2017, *arXiv:1712.01815*.

[6] D. Silver et al., "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

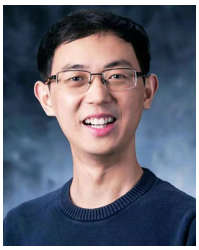
[7] O. Vinyals et al., "Grandmaster level in starcraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[8] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[9] H. Cai et al., "Real-time bidding by reinforcement learning in display advertising," in *Proc. 10th ACM Int. Conf. Web Search Data Min. (WSDM)*, 2017, pp. 661–670.

[10] J. Lu, C. Yang, X. Gao, L. Wang, C. Li, and G. Chen, "Reinforcement learning with sequential information clustering in real-time bidding," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, 2019, pp. 1633–1641.
[11] C. Chen et al., "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3414–3421.
[12] S. M. Shortreed, E. B. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: An empirical study," *Mach. Learn.*, vol. 84, pp. 109–136, Jul. 2011.
[13] S. Agrawal and R. Jia, "Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management," in *Proc. ACM Conf. Econ. Comput.*, 2019, pp. 743–744.
[14] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
[15] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 737–757, 2009.
[16] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1–8.
[17] P. Gajane, R. Ortner, and P. Auer, "A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions," 2018, *arXiv:1805.10066*.
[18] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online Markov decision processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 726–736, 2009.
[19] J. Y. Yu and S. Mannor, "Arbitrarily modulated Markov decision processes," in *Proc. 48h IEEE Conf. Decis. Control (CDC) Held Jointly 28th Chin. Control Conf.*, 2009, pp. 2946–2953.
[20] G. Neu, A. György, and C. Szepesvari, "The online loop-free stochastic shortest-path problem," in *Proc. Annu. Conf. Comput. Learn. Theory*, 2010, pp. 231–243.
[21] G. Neu, A. Gyorgy, and C. Szepesvari, "The adversarial stochastic shortest path problem with unknown transition probabilities," in *Proc. 15th Int. Conf. Artif. Intell. Stat.*, 2012, pp. 805–813.
[22] A. Zimin and G. Neu, "Online learning in episodic Markovian decision processes by relative entropy policy search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–9.
[23] O. Dekel and E. Hazan, "Better rates for any adversarial deterministic MDP," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 675–683.
[24] A. Rosenberg and Y. Mansour, "Online convex optimization in adversarial Markov decision processes," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5478–5486.
[25] C. Jin, T. Jin, H. Luo, S. Sra, and T. Yu, "Learning adversarial Markov decision processes with bandit feedback and unknown transition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4860–4869.
[26] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Reinforcement learning for non-stationary Markov decision processes: The blessing of (More) optimism," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1843–1854.
[27] Y. Fei, Z. Yang, Z. Wang, and Q. Xie, "Dynamic regret of policy optimization in non-stationary environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6743–6754.
[28] W. Mao, K. Zhang, R. Zhu, D. Simchi-Levi, and T. Basar, "Near-optimal model-free reinforcement learning in non-stationary episodic MDPs," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 7447–7458.
[29] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann, and M. Valko, "A kernel-based approach to non-stationary reinforcement learning in metric spaces," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 3538–3546.
[30] H. Zhou, J. Chen, L. R. Varshney, and A. Jagmohan, "Nonstationary reinforcement learning with linear function approximation," 2020, *arXiv:2010.04244*.
[31] A. Touati and P. Vincent, "Efficient learning in non-stationary linear Markov decision processes," 2021, *arXiv:2010.12870*.
[32] H. Zhong, Z. Yang, Z. Wang, and C. Szepesvári, "Optimistic policy optimization is provably efficient in non-stationary MDPs," 2022, *arXiv:2110.08984*.
[33] C. Jin, Q. Liu, and S. Miryoosefi, "Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 13406–13418.
[34] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan, "On the complexity of adversarial decision making," 2022, *arXiv:2206.13063*.
[35] C.-Y. Wei and H. Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," 2021, *arXiv:2102.05406*.
[36] D. Russo and B. Van Roy, "Eluder dimension and the sample complexity of optimistic exploration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–9.
[37] I. Osband and B. V. Roy, "Model-based reinforcement learning and the eluder dimension," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1–9.
[38] Z. Chen, C. J. Li, A. Yuan, Q. Gu, and M. I. Jordan, "A general framework for sample-efficient function approximation in reinforcement learning," 2022, *arXiv:2209.15634*.
[39] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, "The statistical complexity of interactive decision making," 2023, *arXiv:2112.13487*.
[40] K. Dong, J. Peng, Y. Wang, and Y. Zhou, "$\sqrt{n}$-regret for learning in Markov decision processes with function approximation and low bellman rank," 2020, *arXiv:1909.02506*.
[41] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, "Contextual decision processes with low Bellman rank are PAC-learnable," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1704–1713.
[42] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford, "Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches," in *Proc. 32nd Annu. Conf. Comput. Learn. Theory*, 2018, pp. 2898–2933.
[43] S. S. Du et al., "Bilinear classes: A structural framework for provable generalization in RL," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2826–2836.
[44] R. Wang, R. R. Salakhutdinov, and L. Yang, "Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6123–6135.
[45] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. 22nd Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 174–188.
[46] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path," *Mach. Learn.*, vol. 71, no. 1, p. 89–129, Apr. 2008.
[47] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1042–1051.
[48] M. Langberg and L. J. Schulman, "Universal ε-approximators for integrals," in *Proc. 21st Annu. ACM-SIAM Symp. Discrete Algorithms*, 2010, pp. 598–607.
[49] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, 2011, pp. 569–578.
[50] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering," in *Proc. 24th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2013, pp. 434–1453.

Songtao Feng received the Ph.D. degree in electrical engineering from The Pennsylvania State University, University Park. He is currently a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Florida. His research interests include reinforcement learning, optimization, and wireless communications and networking.



Ming Yin received the B.S. degree in applied mathematics from the University of Science and Technology of China, and the Ph.D. degree from the University of California at Santa Barbara. He is a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, Princeton University.

Ruiquan Huang received the B.S. degree in applied mathematics from the University of Science and Technology of China, and the M.S. degree in applied mathematics from Columbia University. He is currently pursuing the Ph.D. degree in electrical engineering with Penn State University. His research interests lie in multiarmed bandits and reinforcement learning, federated learning, and learning theory.

Yu-Xiang Wang (Member, IEEE) received the Ph.D. degree from Carnegie Mellon University in 2017. He is an Associate Professor of Computer Science with the University of California at Santa Barbara. His research interests revolve around the intersection of machine learning, statistics and optimization with special focus on statistical theory and methodology, differential privacy, large-scale machine learning, reinforcement learning, and theory of deep learning. He served as an Area Chair and a Senior Program Committee Member at ICML, NeurIPS, ICLR, COLT, and AISTATS and an Action Editor for *Transactions on Machine Learning Research* and IEEE TRANSACTIONS ON INFORMATION THEORY.

Jing Yang (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park. She is an Associate Professor of Electrical Engineering with The Pennsylvania State University. Her research interests lie in multiarmed bandits and reinforcement learning, federated learning, and wireless communications and networking. She received the National Science Foundation CAREER Award in 2015 and the IEEE WICE Early Achievement Award in 2020, and was selected as one of the 2020 N2Women: Stars in Computer Networking and Communications. She served as a Symposium/Track/Workshop Co-Chair for Asilomar 2023, ICC 2021, INFOCOM 2021-AoI Workshop, WCSP 2019, CTW 2015, and PIMRC 2014, a TPC member of several conferences, and an Editor for IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING from 2017 to 2020. She is currently an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING.

Yingbin Liang (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign in 2005, and served on the faculty of the University of Hawaii and Syracuse University before she joined The Ohio State University (OSU). She is currently a Professor with the Department of Electrical and Computer Engineering, OSU, and a core faculty of Ohio State Translational Data Analytics Institute. She is also currently serving as the Deputy Director of the AI-Edge Institute, OSU. Her research interests include machine learning, optimization, information theory, and statistical signal processing. She received the National Science Foundation CAREER Award and the State of Hawaii Governor Innovation Award in 2009. She also received the EURASIP Best Paper Award in 2014.