Non-stationary Reinforcement Learning under General Function Approximation

Songtao Feng [1] **Ming Yin** [2] **Ruiquan Huang** [3] **Yu-Xiang Wang** [2] **Jing Yang** [3] **Yingbin Liang** [1]

Abstract

General function approximation is a powerful tool to handle large state and action spaces in a broad range of reinforcement learning (RL) scenarios. However, theoretical understanding of non-stationary MDPs with general function approximation is still limited. In this paper, we make the first such an attempt. We first propose a new complexity metric called dynamic Bellman Eluder (DBE) dimension for non-stationary MDPs, which subsumes majority of existing tractable RL problems in static MDPs as well as non-stationary MDPs. Based on the proposed complexity metric, we propose a novel confidence-set based model-free algorithm called SW-OPEA, which features a sliding window mechanism and a new confidence set design for non-stationary MDPs. We then establish an upper bound on the dynamic regret for the proposed algorithm, and show that SW-OPEA is provably efficient as long as the variation budget is not significantly large. We further demonstrate via examples of non-stationary linear and tabular MDPs that our algorithm performs better in small variation budget scenario than the existing UCB-type algorithms. To the best of our knowledge, this is the first dynamic regret analysis in non-stationary MDPs with general function approximation.

1. Introduction

Reinforcement learning (RL) commonly refers to the sequential decision making framework modeled by a Markov Decision Process (MDP), where agent aims to maximize its cumulative reward in an unknown environment (Sutton & Barto, 2018). RL has achieved great success in a variety of practical applications, including games (Silver et al., 2016; 2017; 2018; Vinyals et al., 2019), robotics (Kober et al., 2013; Gu et al., 2017), and autonomous driving (Yurtsever et al., 2019). So far, most existing RL works have focused on a *static* MDP model, in which both the reward and the transition kernel are time-invariant. However, *non-stationarity*[1] naturally occurs in many sequential decision problems such as online advertisement auctions (Cai et al., 2017; Lu et al., 2019), traffic management (Chen et al., 2020), health-care operations (Shortreed et al., 2010), and inventory control (Agrawal & Jia, 2019).

Compared to static RL, non-stationary RL can be significantly challenging. Under the general non-stationary environment, designing algorithm that achieve sublinear regret might not be possible due to the worst scenario where rewards and transitions change drastically over time. A line of extensive studies have focused on tabular non-stationary MDPs (Auer et al., 2008; Gajane et al., 2018; Even-Dar et al., 2009; Yu & Mannor, 2009; Yu et al., 2009; Neu et al., 2010; 2012; Zimin & Neu, 2013; Dekel & Hazan, 2013; Rosenberg & Mansour, 2019; Jin et al., 2020; Cheung et al., 2020; Fei et al., 2020; Mao et al., 2021). However, the performance of these algorithms suffers from large number of states in non-stationary MDPs, which precludes its usage in exponentially large or even continuous state spaces. Therefore, function approximation has become a prominent tool and several works proposed algorithms for non-stationary MDPs with structural assumptions, such as state-action forming a metric space (Domingues et al., 2020), linear MDPs (Zhou et al., 2022; Touati & Vincent, 2020), linear mixture MDPs (Zhong et al., 2021). However, the structural function approximation of (such as linear) MDPs typically restrict the designed algorithms to perform well only under limited classes of MDPs, and may not be applicable generally. This naturally leads to the following open question:

Can we design an algorithm that achieves a desired regret performance for non-stationary MDPs under general function approximation?

To this end, there are a few challenges. (a) We need to identify an appropriate complexity metric for non-stationary MDPs that covers many existing problems of interest;

[1]The Ohio State Univsersity [2]The University of California, Santa Barbara [3]The Pennsylvania State University. Correspondence to: Yingbin Liang <liang.889@osu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We emphasize non-stationarity is different from time-inhomogeneity (*e.g.* (Yin et al., 2021)). The latter allows transition P_t to be different for $t \in [H]$, but P_t's are fixed across episodes.

(b) We need to design an algorithm that can handle non-stationary without function structures on transition kernels and rewards to exploit; and (c) Establishing a dynamic regret bound that potentially improves those for non-stationary simpler MDPs such as linear and tabular cases is non-trivial. In this paper, we give an affirmative answer to the above question by addressing the aforementioned challenges.

We summarize our contributions as follows.

Complexity metric: We propose a new complexity metric named the Dynamic Bellman Eluder (DBE) dimension for non-stationary MDPs, which generalizes the Bellman Eluder (BE) dimension designed for stationary MDPs (Jin et al., 2021), and subsumes a broad class of RL problems including low BE dimension problems in stationary RL and linear MDPs in non-stationary RL. Moreover, when the non-stationarity is relatively small compared to a universal gap (which still allows a certain non-stationarity), we show that the DBE dimension is the same as the BE dimension of one MDP instance of the non-stationary MDPs.

Algorithm: We then design a new confidence-set based algorithm SW-OPEA for non-stationary MDP, by greedily selecting the candidate value function in the confidence region. This is in contrast to the UCB-type algorithms adopted by all previous studies of non-stationary MDPs. In fact, a UCB-type algorithm is not easily applicable to non-stationary MDPs with general function approximation due to the difficulty of finding an appropriate bonus term. Our main design novelty lies in the construction of the confidence region, which features the sliding window mechanism, and incorporates local variation budget in order to exactly capture the distribution mismatch between 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.

Theory: We theoretically characterize the dynamic regret of SW-OPEA in Theorem 5.2. To demonstrate the advantage of SW-OPEA , we compare our regret bound of SW-OPEA to that of previously proposed UCB-type algorithms (Zhou et al., 2022) for non-stationary 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 the minimax lower bound given in Zhou et al. (2022). Our bound is slightly worse in the average variation budget, which suggests that our algorithm is advantageous over UCB-type algorithms in the small variation scenario.

Analysis: Technically, 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.

1.1. Related Work

Non-stationary tabular MDPs: Most works on non-stationary tabular MDPs considered static regret (Auer et al., 2008; Gajane et al., 2018; Even-Dar et al., 2009; Yu & Mannor, 2009; Yu et al., 2009; Neu et al., 2010; 2012; Zimin & Neu, 2013; Dekel & Hazan, 2013; Rosenberg & Mansour, 2019; Jin et al., 2020). A few recent studies (Cheung et al., 2020; Fei et al., 2020; Mao et al., 2021) focused on dynamic regret for non-stationary tabular MDPs. Specifically, assuming time-varying transitions and rewards, Cheung et al. (2020) proposed a sliding window approach, and Mao et al. (2021) used restart mechanism to handle non-stationarity. While the first two works adopted value-based algorithms, Fei et al. (2020) applied a policy optimization algorithm for full-information feedback of rewards and time-invariant transitions.

Non-stationary MDPs with function approximation: Under non-stationary MDPs with continuous environment where the state-action forms a metric space, Domingues et al. (2020) proposed a kernel-based algorithm. Two concurrent works Zhou et al. (2022) and Touati & Vincent (2020) considered non-stationary RL under linear MDPs, where Zhou et al. (2022) considered dynamic regret and Touati & Vincent (2020) studied static regret. To handle non-stationarity, Zhou et al. (2022) adopted a scheme of restarting the base LSVI-UCB algorithm while Touati & Vincent (2020) used weighted least squares value iteration with exponential weights on past data. Under the non-stationary MDPs with linear mixture function approximation of both transitions and rewards, Zhong et al. (2021) considered bandit feedback rewards and dynamic regret. Moreover, Wei & Luo (2021) proposed black-box reduction approach that converts algorithm with optimal regret in stationary MDPs into another algorithm for non-stationary MDPs.

Recently, Foster et al. (2022) generalized the decision-estimation coefficient (DEC) framework to non-stationary RL setting with the goal of minimizing the static regret. Their framework can potentially cover majority problems but the connection between their result and the existing results is still not well understood. We also remark that the performance under the DEC framework is often worse than the best-known result when restricted to special cases. Further, their work focused on the static regret, whereas our work potentially maintains the sharp performance when restricting to special cases, and our performance metric of dynamic regret is more general.

Static MDP with general function approximation: Broadly speaking, the line of research on designing sample-

efficient RL algorithms with general function approximations in the past has been mainly focused on the static RL setting. Russo & Van Roy (2013); Osband & Roy (2014) initiated the study on the minimal structural assumptions that render sample-efficient learning by proposing a structural condition called Eluder dimension, and Wang et al. (2020) then extended LSVI-UCB for general function approximation with small Eluder dimension. Another well-studied direction is the low-rank conditions, including Bellman rank (Dong et al., 2019; Jiang et al., 2017) for model-free setting and witness rank (Sun et al., 2018) for model-based setting. Jin et al. (2021) proposed a complexity named Bellman Eluder (BE) dimension for model-free setting, which subsumes low Bellman rank and low Eluder dimension as special cases. Du et al. (2021) proposed Bilinear class, which unifies both model-based and model-free RL for a broad class of loss estimators including Bellman error. Sharing the same spirit of unifying model-free and model-based RL, Foster et al. (2021) proposed DEC, which is a necessary and sufficient condition for sample-efficient learning, and then they extended it to an adversarial decision making problem with static regret in Foster et al. (2022). While the sample complexity of Foster et al. (2021); Du et al. (2021) is generally worse than the best-known result when restricted to special cases, Chen et al. (2022) recently extended BE dimension and proposed an Admissible Bellman Characterization (ABC) framework to include both model-free and model-based RL while maintaining sharp sample efficiency. Very recently, Yin et al. (2023); Zhang et al. (2022) consider parametric differentiable function approximation in offline RL, but there is no study in the online regime.

Non-stationary bandits: Broadly speaking, our work is also related to a line of research on non-stationary bandits. Methods have been proposed to handle non-stationarity for various non-stationary multi-armed bandit (MAB) settings, including decaying memorey and sliding windows (Garivier & Moulines, 2011; Keskin & Zeevi, 2017) and restart mechanism (Auer et al., 2002; Besbes et al., 2014b;a), which are widely employed in non-stationary RL. More recently, several works developed methods for unknown variation budget (Karnin & Anava, 2016; Cheung et al., 2022), and abrupt changes (Auer et al., 2019). Another line of works focused on Markovian bandits (Ma, 2018), non-stationary contextual bandits (Luo et al., 2017; Chen et al., 2019), linear bandits (Cheung et al., 2019; Zhao et al., 2020), and bandits with slowly changing rewards (Besbes et al., 2019).

2. Preliminaries

2.1. Non-stationary MDPs

We consider an episodic MDP with time-varying transitions and rewards $(\mathcal{S}, \mathcal{A}, H, P, r, x_1)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of each episode, $P = \{P_h^k\}_{(k,h)\in[K]\times[H-1]}$ is the collection of non-stationary transition kernels with $P_h^k : \mathcal{S}\times\mathcal{A} \mapsto \triangle(\mathcal{S})$, $r = \{r_h^k\}_{(k,h)\in[K]\times[H]}$ is the collection of adversarial deterministic reward functions with $r_h^k : \mathcal{S}\times\mathcal{A} \mapsto [0,1]$, and x_1 is the fixed initial state.

Suppose an agent sequentially interacts with the non-stationary MDP $(\mathcal{S}, \mathcal{A}, H, P, r, x_1)$. At the beginning of the k-th episode, the reward $\{r_h^k\}_{h\in[H]}$ are adversarially chosen by the environment, which possibly depends on the $(k-1)$ historical trajectories. Meanwhile, the agent determines a policy $\pi^k = \{\pi_h^k\}_{h\in[H]}$ where $\pi_h^k : \mathcal{S} \mapsto \triangle(\mathcal{A})$. At the h-th step, the agent observes the state x_h^k, takes an action following $a_h^k \sim \pi_h^k(\cdot|x_h^k)$, obtains the reward function r_h^k which determines the received reward $r_h^k(x_h^k, a_h^k)$, and the MDP evolves into the next state $x_{h+1}^k \sim P_h^k(\cdot|x_h^k, a_h^k)$. The k-th episode ends after receiving the last reward $r_H^k(x_H^k, a_H^k)$. For convenience, let x_{H+1} be a dummy state and $P_H^k(x_{H+1}|x_H^k, a_H^k) = 1$ for any $(x_H^k, a_H^k) \in \mathcal{S}\times\mathcal{A}$. Define the state and state-action value functions of policy $\pi = \{\pi_h\}_{h\in[H]}$ recursively via the following equation

$$Q_{h;(*,k)}^{\pi}(x,a) = r_h^k(x,a) + (\mathbb{P}_h V_{h+1;(*,k)}^{\pi})(x,a),$$
$$V_{h;(*,k)}^{\pi}(x) = \langle Q_{h;(*,k)}^{\pi}(x,\cdot), \pi_h(\cdot|x)\rangle_{\mathcal{A}}, \;\; V_{H+1;(*,k)} = 0,$$

where $\mathbb{P}_h$ is the operator defined as $(\mathbb{P}_h f)(x,a) := \mathbb{E}\left[f(x')|x' \sim P_h(x'|x,a)\right]$ 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 performance is measured by the dynamic regret, which quantifies the performance difference between the learning policy and the benchmark policy $\{\pi^{(*,k)}\}_{k\in[K]}$ where $\pi^{(*,k)} = \arg\max_{\pi} V_{1;(*,k)}^{\pi}(x_1)$. Specifically, the dynamic regret for K episodes is defined as

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

2.2. Function Approximation

Consider a function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where $\mathcal{F}_h \subseteq (\mathcal{S}\times\mathcal{A} \mapsto [0, H-h+1])$ offers a collection of candidate functions to approximate $Q_{h;(*,k)}^{\pi^{(*,k)}}$, denoted as $Q_{h;(*,k)}^{*}$. Since each episode ends in H steps, we set $f_{H+1} = 0$. We make the following standard assumptions on the function class $\mathcal{F}$.

Assumption 2.1 (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., $(Q_{1;(*,k)}^{*}, Q_{2;(*,k)}^{*}, \cdots, Q_{H;(*,k)}^{*}) \in \mathcal{F}$ for $k \in [K]$.

Given functions $f = (f_1, f_2, \cdots, f_H)$ where $f_h \in (\mathcal{S}\times$

$\mathcal{A} \mapsto [0, H-h+1]$), define

$$(\mathcal{T}_h^k f_{h+1})(x,a) := r_h^k(x,a) + (\mathbb{P}_h^k f_{h+1})(x,a),$$
$$(\mathbb{P}_h^k f_{h+1})(x,a) = \mathbb{E}_{x' \sim P_h^k(\cdot|x,a)}[\max_{a' \in \mathcal{A}} f_{h+1}(x', a')],$$

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

Assumption 2.2 (Completeness)**.** $\mathcal{T}_h^k \mathcal{F}_{h+1} \subseteq \mathcal{F}_h$ for all $(k,h) \in [K] \times [H]$.

Completeness assumption requires the function class $\mathcal{F}$ is closed under Bellman operators of any episode.

3. Dynamic Eluder Dimension

In this section, we introduce a new complexity measure for non-stationary MDPs. We start with the following ϵ-independence between distributions and the distributional Eluder dimension.

Definition 3.1 (ϵ-independence between distributions (Jin et al., 2021))**.** Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \mu_2, \cdots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ϵ-independent of $\{\mu_1, \mu_2. \cdots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists $g \in \mathcal{G}$ such that $\sum_{i=1}^n (\mathbb{E}_{x \sim \mu_i}[g(x)])^2 \leq \epsilon^2$, but $|\mathbb{E}_{x \sim \nu}[g(x)]| > \epsilon$.

Definition 3.2 (Distributional Eluder (DE) dimension (Jin et al., 2021))**.** Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and Π be a family of probability measures over $\mathcal{X}$. The distributinoal Eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \rho_2, \cdots, \rho_n\} \subseteq \Pi$ such that there exists $\epsilon' \geq \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \rho_2, \cdots, \rho_{i-1}\}$ for all $i \in [n]$.

The next definition of Bellman Eluder dimension is first introduced in Jin et al. (2021) for stationary MDPs.

Definition 3.3 (Bellman Eluder dimension (BE))**.** Let $(I - \mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}, k \in [K]\}$ 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

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

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

Definition 3.4 (Dynamic Bellman Eluder (DBE) dimension)**.** Let $(I - \bar{\mathcal{T}}_h)\mathcal{F} := \{f_h - \mathcal{T}_h^k f_{h+1} : f \in \mathcal{F}, k \in [K]\}$ 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 dynamic Bellman Eluder dimension of $\mathcal{F}$ with respect to Π is defined as

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

We focus on the following choice of distribution family $\mathcal{D}_\Delta = \{\mathcal{D}_{\Delta,h}\}_{h \in [H]}$ where $\mathcal{D}_{\Delta,h} = \{\delta_{(s,a)} : s \in \mathcal{S}, a \in \mathcal{A}\}$. However, our result can be adapted to $\mathcal{D}_\mathcal{F} = \{\mathcal{D}_{\mathcal{F},h}\}_{h \in [H]}$ where $\mathcal{D}_{\mathcal{F},h}$ denotes the collection of all probability measures over $\mathcal{S} \times \mathcal{A}$ at h-th step, generated by executing the greedy policy π_f induced by any $f \in \mathcal{F}$.

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 non-stationary 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 $\widetilde{\delta}_\epsilon^u$ defined below, then the DBE dimension equals to the BE dimension with respect to one MDP instance of the non-stationary MDPs.

Definition 3.5 (Universal gap)**.** If ν is ϵ-independent of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ with respect to $g \in \mathcal{G}$, we define gap $\widetilde{\delta}_{g,\epsilon;\boldsymbol{\mu},\nu} = |\mathbb{E}_{x \sim \nu}[g(x)]| - \epsilon$. The universal gap with respect to a function class $\mathcal{G}$ is $\widetilde{\delta}_\epsilon^u = \inf_{g \in \mathcal{G}, \epsilon' \geq \epsilon, \boldsymbol{\mu}_g} \widetilde{\delta}_{g,\epsilon';\boldsymbol{\mu}_g}$ where $\boldsymbol{\mu}_g$ is any ϵ'-independent sequence with respect to g.

Proposition 3.6 (Informal)**.** *If the variations in transitions and rewards are relatively small compared to the universal gap $\widetilde{\delta}_\epsilon^u$ with respect to $(I - \mathcal{T}_h^k)\mathcal{F}$ for $k \in [2:K]$, then*

$$\dim_{\mathrm{DBE}}(\mathcal{F}, \Pi, \epsilon) = \max_{h \in [H]} \dim_{\mathrm{DE}}((I - \mathcal{T}_h^1)\mathcal{F}, \Pi, \epsilon),$$

where the latter is exactly the BE dimension of the first MDP instance of the non-stationary MDPs.

The formal statement of the proposition (see Proposition A.4) and its proof is provided in Section A. The intuition is if the variations in transitions and rewards are small (but does not necessarily vanish), then the set of functions $(I - \mathcal{T}_h^k)\mathcal{F}$ for $k \in [2:K]$ is relatively close to $(I - \mathcal{T}_h^1)\mathcal{F}$. Therefore their union $(I - \bar{\mathcal{T}}_h)\mathcal{F}$, constructed for the DBE dimension, remains close to $(I - \mathcal{T}_h^1)\mathcal{F}$.

Under static MDPs, the DBE dimension naturally reduces to BE dimension, and therefore it subsumes a majority tractable problem classes in stationary RL. Moreover, the DBE framework further includes more tractable problem

classes in non-stationary RL. Below we show that our DBE dimension covers non-stationary linear MDPs.

Definition 3.7 (Non-stationary Linear MDPs (Zhou et al., 2022))**.** For linear MDP with feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, there exists an unknown measure μ_h^k on $\mathcal{S}$ and a vector $\theta_h^k \in \mathbb{R}^d$ satisfying $P_h^k(s'|s,a) = \phi(s,a)^\top \mu_h^k(s')$ and $r_h^k(s,a) = \phi(s,a)^\top \theta_h^k$, where $\|\phi(s,a)\| \le 1$ and $\max\{\|\mu_h^k\|, \|\theta_h^k\|\} \le \sqrt{d}$ for all $(h,k) \in [H] \times [K]$.

The next proposition shows that the DBE dimension of non-stationary linear MDPs scales with the linear feature dimension $\widetilde{\mathcal{O}}(d)$. The proof is shown in Appendix B.

Proposition 3.8. *The DBE dimension of non-stationary linear MDPs with the feature dimension d satisfies*

$$\dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \le \mathcal{O}\left(1 + d \log\left(16H^2 d/\epsilon^2 + 1\right)\right).$$

4. Algorithm

In this section, we propose our algorithm SW-OPEA for non-stationary MDPs with general function approximation.

At high level, SW-OPEA differentiates from the GOLF algorithm (Jin et al., 2021) for static MDPs with general function approximation in its novel designs to handle the non-stationarity of transition kernels and rewards. Specifically, SW-OPEA features the sliding window mechanism and incorporates local variation budget in order to exactly capture the distribution mismatch between 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.

Further in the context of the previous studies of non-stationary MDPs, SW-OPEA is the first confidence-set based algorithm, to the best of our knowledge. In fact, a UCB-type algorithm is not easily applicable to non-stationary MDPs with general function approximation due to the difficulty of finding an appropriate bonus term. As we will show in Section 5 by an example of non-stationary linear MDPs, SW-OPEA performs better than the best known UCB-type algorithms in small variation budget scenarios.

The pseudocode of SW-OPEA is presented in Algorithm 4. 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.

Algorithm 1 SW-OPEA (Sliding Window Optimistic-based Exploration and Approximation under non-stationary MDPs)

1: **Input:** $\mathcal{D}_1, \cdots, \mathcal{D}_H \leftarrow \varnothing$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
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 $(x_1^k, a_1^k, \cdots, x_H^k, a_H^k, x_{H+1}^k)$ by following π^k and reward function $\{r_h^k\}_{h\in[H]}$.
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 $\mathcal{B}^k = \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \inf_{g\in\mathcal{G}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1}) + \beta + 2H^2\Delta_P^w(k,h), \forall h \in [H]\}$.
7: **end for**

Sliding window squared Bellman error is defined as

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

Note that in episode k, we use the latest reward information r_h^k over the entire window, rather than r_h^t, to form the sliding window squared Bellman error. Such construction exploits the most recent information of the reward function r_h^k to maximally reduce the non-stationarity of rewards. Therefore, $\mathcal{L}_{\mathcal{D}_h}$ tends to be small as long as the transition kernel difference between episode k and t is small. Furthermore, we adopt the sliding window in the squared loss (1), which is based on the "forgetting principle" (Garivier & Moulines, 2011) where the squared loss estimated at episode k relies on the observed history during episode $1 \vee (k-w)$ to k instead of all prior observations. The rationale is that under non-stationarity 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 $\mathcal{D}_h$

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

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

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

Since the transition kernel varies across episodes, we include an additional term of the local variation budget $\Delta_P^w(k,h)$ in the definition of $\mathcal{B}_k$. Intuitively, the local variation budget $\Delta_P^w(k,h)$ captures the cumulative transition kernel differences between current episode and all previous episode in the sliding window. Therefore, by compensating a term involving $\Delta_P^w(k,h)$ 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 (see Lemma C.2).

We remark that the assumption on the local variation budget involving transition functions are unknown could be relaxed. Inspired by the standard technique to handle unknown variation budget as in linear nonstationary MDPs (Zhou et al., 2022), we propose the following modification of the algorithm. We remove the local variation budget in the bonus term in the algorithm, and instead, design a strategy to adapt the window size to the variation budget (without knowing its value) as in the EXP3-P algorithm (Bubeck & Cesa-Bianchi, 2012). It has been shown that as long as the window sizes are picked to densely cover the entire value range of the window size, such a scheme will result in a performance close enough to the case as if the window size is picked in an optimal way when the variation budget is known. We expect that such a scheme will achieve the same regret (in terms of scaling) as the case with the knowledge of the variation budget. We will investigate the feasibility of the proposed strategy and leave the detailed mathematical analysis in the future work.

5. Theoretical Guarantees

In this section, we first provide our main theoretical result for SW-OPEA, and then present a proof sketch that highlights our novel developments in the analysis.

5.1. Main Results

In this section, we provide our characterization of the dynamic regret for SW-OPEA.

We first state the following generalized completeness assumption (Antos et al., 2008; Chen & Jiang, 2019; Jin et al., 2021). 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 5.1 (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 5.1 is equivalent to the standard completeness assumption (Assumption 2.2). Without loss of generality, we assume $\mathcal{F} \subseteq \mathcal{G}$ and $\mathcal{G} = \mathcal{F} \cup \mathcal{G}$.

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

$$\Delta_R(K) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sup_{x \in \mathcal{S}, a \in \mathcal{A}} |(r_h^k - r_h^{k-1})(x,a)|, \tag{3}$$

$$\Delta_P(K) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sup_{x \in \mathcal{S}, a \in \mathcal{A}} \left\|(P_h^k - P_h^{k-1})(\cdot|x,a)\right\|_1, \tag{4}$$

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 5.2 (Dynamic regret of SW-OPEA). *Under Assumption 2.1 and Assumption 5.1, 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 $k \geq \min\{w+1, \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})\}$ we have*

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

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 $\Delta_P^w(t,h)$ 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

$$L = \max_{h \in [H], t < k} \frac{\sum_{s=t}^{k-1} \sup_{x,a} \|(P_h^{s+1} - P_h^s)(\cdot|x,a)\|_1}{k-t}. \tag{5}$$

Clearly, we have $L \leq 1$ and $\Delta_P^w(k,h) \leq Lw^2$, and L can be viewed as the the greatest average variation of transition kernels 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.

Corollary 5.3. *Under the condition of Theorem 5.2 and $|\mathcal{G}| > 10$, with probability at least $1-\delta$, the following argument holds: if $\sqrt{L} > \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|} - \frac{1}{H\sqrt{d}}\right)$, select $w = \lceil \frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L} + \frac{1}{HK\sqrt{d}}} \rceil$ and the dynamic regret is bounded by*

$$\widetilde{\mathcal{O}}\Big(H^{\frac{3}{2}} K^{\frac{1}{2}} d^{\frac{1}{4}} (\log|\mathcal{G}|)^{\frac{1}{4}} + H^2 K d^{\frac{1}{2}} L^{\frac{1}{4}} (\log|\mathcal{G}|)^{\frac{1}{4}} + \Delta_R + H\Delta_P\Big); \tag{6}$$

otherwise, select $w = K$ and the dynamic regret is bounded by $\widetilde{O}\left(H^2 K^{\frac{1}{2}} d^{\frac{1}{2}} (\log|\mathcal{G}|)^{\frac{1}{2}}\right)$, 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 5.2 and Corollary 5.3, 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 (6) corresponds to the regret of the static MDP while the remaining term arises due to the non-stationarity. As a result, when transitions and rewards remain the same over time, our result reduces to $\widetilde{\mathcal{O}}\left(H^2K^{\frac{1}{2}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{2}}\right)$, which matches the static regret of GOLF in Jin et al. (2021)[2].

Advantage of SW-OPEA: To understand the advantage of SW-OPEA over the UCB-based algorithms, we take non-stationary linear MDPs as an example. When specializing to non-stationary linear and tabular MDPs, our result becomes $\widetilde{\mathcal{O}}\left(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}+HT\widetilde{d}^{\frac{3}{4}}L^{\frac{1}{4}}+TL_\theta\right)$ where $T=HK$, $\widetilde{d}$ is the feature dimension for linear MDPs and $\widetilde{d}=|\mathcal{S}||\mathcal{A}|$ for tabular MDPs, and L_θ is the average variation budget in rewards. For non-stationary linear MDPs, the result in Zhou et al. (2022) 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[3] to be for tabular MDPs, which gives $\widetilde{\mathcal{O}}\left(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}^{\frac{3}{2}}+H^{\frac{4}{3}}\widetilde{d}^{\frac{3}{2}}T\widetilde{L}^{\frac{1}{3}}+H^{\frac{4}{3}}\widetilde{d}^{\frac{4}{3}}TL_\theta^{\frac{1}{3}}\right)$. 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 non-stationarity 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 (note that $L\leq 1$). For the last term caused by the non-stationary of rewards, our result performs better in the variation budget of rewards, horizon H as well as the feature dimension $\widetilde{d}$.

It also interesting to compare our result with the minimax dynamic regret lower bound $\Omega\left(H^{\frac{1}{2}}T^{\frac{1}{2}}\widetilde{d}+H^{\frac{1}{3}}T\widetilde{d}^{\frac{2}{3}}\widetilde{L}^{\frac{1}{3}}\right)$ developed in Zhou et al. (2022) for linear MDPs with non-stationary transitions. For such a case, our result becomes $\widetilde{\mathcal{O}}\left(H^{\frac{3}{2}}T^{\frac{1}{2}}\widetilde{d}+HT\widetilde{d}^{\frac{3}{4}}L^{\frac{1}{4}}\right)$. The first term is the regret under stationary MDPs and the second term arises due to the non-stationarity of transitions. We can see that our first term corresponding to static MDPs matches the lower bound both in terms of T and $\widetilde{d}$, whereas the upper bound in Zhou et al. (2022) matches the lower bound only in T. For the non-stationarity term, our dependency on H and $\widetilde{d}$ is closer to the lower bound than that in Zhou et al. (2022), whereas our dependency on the variation budget is close but does not match the lower bound. Overall, these comparisons suggest that our confidence-set based algorithm performs better than UCB-type algorithms in small variation budget scenario under non-stationary linear MDPs.

When the state-action set forms a metric space, Domingues et al. (2020) proposed a kernel-based approach in nonstationary RL. Ignoring term regarding static MDPs, their result renders $\widetilde{O}\left(SA^{\frac{1}{2}}H^{\frac{4}{3}}TL^{\frac{1}{3}}+SA^{\frac{1}{2}}H^{\frac{4}{3}}TL_\theta^{\frac{1}{3}}\right)$ regret bound in the tabular case while our result becomes $\widetilde{O}\left((SA)^{\frac{3}{4}}HTL^{\frac{1}{4}}+TL_\theta\right)$. For the first term caused by the non-stationarity of transition kernels, our result has better dependency on step H, but is worse in the average variation budget of transitions. For the second term caused by the non-stationarity of rewards, the dependency on the variation budget of rewards, horizon H as well as the cardinality of state and action spaces is improved. The comparison suggests our confidence-set based algorithm is advantageous over the kernel-based algorithm in small variation budget and small action space scenario under non-stationary MDPs.

5.2. Proof Sketch of Theorem 5.2

In this section, we provide a sketch of the proof for Theorem 5.2 and defer all the details to Appendix C.

The preliminary step is to decompose the dynamic regret of SW-OPEA into three terms as follows:

$$\begin{aligned}
&\mathrm{D-Regret}(k)\leq\\
&H+\underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(r_h^{t-1}-r_h^t)(x_h,a_h)]}_{\text{(I)}}\\
&+\underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H}\left[\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}-\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t))}\right][r_h^t(x_h,a_h)]}_{\text{(II)}}\\
&+\underbrace{\sum_{t=1}^{k}\left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}}-V_{1;(*,t-1)}^{\pi^t}\right)(x_1)}_{\text{(III)}}. \qquad (7)
\end{aligned}$$

Term (I) can be bounded by $\Delta_R(k)$ by the definition of the variation budget of rewards (3). 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 $\widetilde{V}_{h';,(*,k)}^{\pi}$. Then,

[2]The additional H here is due to the definition of $r_h\in[0,1]$, whereas Jin et al. (2021) assumes $\sum_h r_h\leq 1$.

[3]They consider bandit feedback. By adapting their algorithm and analysis, it turns out that the dynamic regret does not benefit from full information feedback in non-stationary linear MDPs.

we show in Lemma C.1 that

$$\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k))}\right)[r_h^k(x_h,a_h)] = \left[\widetilde{V}^{\pi^k}_{1;(*,k-1)} - \widetilde{V}^{\pi^k}_{1;(*,k)}\right](x_1) \le \sum_{i=1}^{h-1}\sup_{x,a}\left\|(P_h^k - P_h^{k-1})(\cdot|x,a)\right\|_1.$$

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

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

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 non-stationary MDPs.

Lemma 5.4 (Probability distribution shift lemma)**.** *Suppose P and Q are two probability distributions of a random variable x and define $f_m = \sup_x |f(x)|$. Then we have*

$$\left|\left(\mathop{\mathbb{E}}_{x\sim P} f(x) - C\right)^2 - \left(\mathop{\mathbb{E}}_{x\sim Q} f(x) - C\right)^2\right| \le (2f_m + 2|C|)f_m \cdot \mathrm{TV}(P,Q).$$

The proof can be found in Appendix C.6.

Next, we show in Lemma C.2 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

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

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 5.4, 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 $\Delta_P^w(k,h)$. 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 1) guarantees $V^*_{1;(*,k-1)}(x_1) \le \sup_a f_1^k(x_1,a)$ for every episode $k\in[K]$. Combining the optimism and the generalized policy loss decomposition (see Lemma C.8), we have

$$\text{(III)} \le \sum_{t=1}^{k}\left(\max_{a\in\mathcal{A}} f_1^t(x_1,a) - V^{\pi^t}_{1;(*,t-1)}(x_1)\right) \le \sum_{h=1}^{H}\sum_{t=1}^{k}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1}f_{h+1}^t)(x_h,a_h)]. \tag{8}$$

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 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 Lemma C.3 that with high probability, for all $(k,h)\in[K]\times[H]$,

$$\sum_{t=1\vee(k-w-1)}^{k-1}\Big[f_h^k(s_h^t,a_h^t) - r_h^{k-1}(s_h^t,a_h^t) - \mathop{\mathbb{E}}_{x'\sim P_h^{k-1}(x_h^t,a_h^t)}\max_{a'\in\mathcal{A}} f_{h+1}^k(s',a')\Big]^2 \le 6H^2\Delta_P^w(k-1,h) + \mathcal{O}(\beta). \tag{9}$$

Technically, we define the following helpful random variable

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

to form an appropriate martingale and obtain the martingale concentration result. Then, applying our probability distribution shift lemma (Lemma 5.4), the definition of $\mathcal{B}^k$ and the completeness assumption gives (9).

Step IV: We establish the relationship between (8) and (9). Specifically, we aim to upper bound (8) given (9) 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 $i\in[1\vee(k-w):k-1]$. It turns out that the DBE dimension plays an important role in connecting these two terms, as summarized in the following lemma.

Lemma 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 $\omega>0$,*

$$\sum_{t=1\vee(k-w)}^{k}|\mathbb{E}_{x\sim\mu_t}[\phi_t(x)]| \le \mathcal{O}\Big(\sqrt{\dim_{\mathrm{DE}}(\Phi,\Pi,\theta)\beta[k\wedge(w+1)]} + \min\{w+1,k,\dim_{\mathrm{DE}}(\Phi,\Pi,\theta)\}C + [k\wedge(w+1)]\theta\Big).$$

The proof is adapted from the proof of Lemma 41 in (Jin et al., 2021) and provided in Appendix C.5.

Based on the DBE dimension and Lemma 5.5, we are ready to bound (III) via term (8) . 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

$$\sum_{h=1}^{H}\sum_{t=1}^{k}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1}f_{h+1}^t)(x_h,a_h)]$$

$$\leq \mathcal{O}\Big(H\sqrt{w} + \frac{H^2k}{\sqrt{w}}\sqrt{\dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_\Delta, \sqrt{1/K})\log\frac{KH|\mathcal{F}|}{\delta}}$$
$$+ \frac{Hk}{\sqrt{w}}\sqrt{\dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_\Delta, \sqrt{1/K})}\sum_{h=1}^{H}\sqrt{\sup_{k\in[K]}\Delta_P^w(k,h)}\Big).$$

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

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

where we suppress the first term H in (7) since it is dominated by the fourth term herein.

5.3. Bandit Feedback

In this section, we extend our algorithm to bandit feedback scenario. We defer all the details to Appendix D.

In bandit feedback scenario, the reward function $r_h^k(\cdot,\cdot)$ is no long available, and the agent can only get access to the reward obtained from the trajectory. Therefore, we need to capture the non-stationarity of rewards in the construction of the sliding window Bellman error and the confidence set. Specifically, we replace the sliding window squared Bellman error (1) with

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

where r_h^t is the reward obtained at step h in episode t. Moreover, the local regression constraint for the confidence set is

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

where β is a confidence parameter, Δ_P^w is the local variation budget in transitions defined in (2) and Δ_R^w is the local variation budget in rewards defined as

$$\Delta_P^w(k,h) = \textstyle\sum_{t=1\vee(k-w)}^{k}\sup_{x\in\mathcal{S},a\in\mathcal{A}}|(r_h^k - r_h^t)(x,a)|.$$

Our main theoretical result for the bandit feedback scenario is provided in the next theorem.

Theorem 5.6. *Under Assumption 2.1 and Assumption 5.1, 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 $k \geq \min\{w+1, \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})\}$ we have*

$$\mathrm{D-Regret}(k) = \Delta_R(k) + H\Delta_P(k) + \mathcal{O}\Big(H\sqrt{w}$$
$$+ \frac{H^2k}{\sqrt{w}}\sqrt{d\log[KH|\mathcal{G}|/\delta]} + \frac{H^2k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_P^w(t,h)}$$
$$+ \frac{H^{3/2}k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_R^w(t,h)}\Big),$$

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

Besides the average variation budget L in transitions defined in (5), we define the average variation budget L_θ in rewards

$$L_\theta = \max_{h\in[H],t<k}\frac{\sum_{s=t}^{k-1}\sup_{x,a}|(r_h^{s+1} - r_h^s)(x,a)|}{k-t}. \quad (10)$$

By optimizing the window size w, we have the following corollary.

Corollary 5.7. *Under the condition of Theorem 5.6 and $|\mathcal{G}| > 10$, with probability at least $1-\delta$, the following argument holds: if $\sqrt{L} + \frac{\sqrt{L_\theta}}{\sqrt{H}} > \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|} - \frac{1}{H\sqrt{d}}\right)$, select $w = \lceil\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L} + \frac{\sqrt{L_\theta}}{\sqrt{H}} + \frac{1}{HK\sqrt{d}}}\rceil$, the dynamic regret is upper-bounded by*

$$\widetilde{\mathcal{O}}\Big(H^{\frac{3}{2}}K^{\frac{1}{2}}d^{\frac{1}{4}}(\log|\mathcal{G}|)^{\frac{1}{4}} + H^2KL^{\frac{1}{4}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{4}}$$
$$+ H^{\frac{7}{4}}KL_\theta^{\frac{1}{4}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{4}} + \Delta_R + H\Delta_P\Big);$$

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

6. Conclusion and Future Work

In this paper, we proposed a new complexity metric named Dynamic Bellman Eluder (DBE) dimension for non-stationary MDPs, which extends the Bellman Eluder (BE) dimension for static MDPs. When the variations in transition kernels and rewards are relatively small compared to a universal gap, we show that the DBE dimension is exactly the BE dimension of one MDP instance in the non-stationary MDPs. We then incorporated the sliding window mechanism and a novel design for the confidence set into our confidence-set based algorithm SW-OPEA, and provided its theoretical upper bound on the dynamic regret. We further demonstrate the advantage of our algorithm by comparing our dynamic regret bound to that of previously proposed algorithms for non-stationary linear and tabular MDPs. One interesting future direction is to further improve the dependency of the dynamic regret on the average variation L.

Acknowledgements

The work of S. Feng was supported in part by the startup fund of the Ohio State University. The work of Y. Liang was supported in part by the U.S. National Science Foundation under the grants DMS-2134145 and RINGS-2148253. The work of R. Huang and J. Yang was supported by the U.S. National Science Foundation under the grants CNS-1956276 and CNS-2003131. M. Yin and Y. Wang were partially supported by National Science Foundation grants #2007117 and #2003257.

References

Agrawal, S. and Jia, R. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.*, 71(1):89–129, apr 2008.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2008.

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, 2014a.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, 2014b.

Besbes, O., Gur, Y., and Zeevi, A. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012.

Cai, H., Ren, K., Zhang, W., Malialis, K., Wang, J., Yu, Y., and Guo, D. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., Xu, K., and Zhenhui. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *AAAI Conference on Artificial Intelligence*, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. A general framework for sample-efficient function approximation in reinforcement learning. *ArXiv*, abs/2209.15634, 2022.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning for non-stationary Markov decision processes: The blessing of (More) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, 68(3):1696–1713, 2022.

Dekel, O. and Hazan, E. Better rates for any adversarial deterministic MDP. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Domingues, O. D., M'enard, P., Pirotta, M., Kaufmann, E., and Valko, M. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Dong, K., Peng, J., Wang, Y., and Zhou, Y. $\sqrt{n}$-regret for learning in markov decision processes with function approximation and low bellman rank. *ArXiv*, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, 2021.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems*, 2020.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *ArXiv*, 2021.

Foster, D. J., Rakhlin, A., Sekhari, A., and Sridharan, K. On the Complexity of Adversarial Decision Making. *ArXiv*, 2022.

Gajane, P., Ortner, R., and Auer, P. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *ArXiv*, 2018.

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, 2011.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, 2017.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

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

Karnin, Z. S. and Anava, O. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems*, 2016.

Keskin, N. B. and Zeevi, A. Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307, 2017.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Lu, J., Yang, C., Gao, X., Wang, L., Li, C., and Chen, G. Reinforcement learning with sequential information clustering in real-time bidding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

Luo, H., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. In *Annual Conference Computational Learning Theory*, 2017.

Ma, W. Improvements and generalizations of stochastic knapsack and markovian bandits approximation algorithms. *Mathematics of Operations Research*, 43(3):789–812, 2018.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Neu, G., György, A., and Szepesvari, C. The online loop-free stochastic shortest-path problem. In *Annual Conference Computational Learning Theory*, 2010.

Neu, G., Gyorgy, A., and Szepesvari, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

Osband, I. and Roy, B. V. Model-based reinforcement learning and the eluder dimension. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, 2019.

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

Shortreed, S. M., Laber, E. B., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84:109–136, 2010.

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

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *ArXiv*, 2017.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Annual Conference Computational Learning Theory*, 2018.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Touati, A. and Vincent, P. Efficient learning in non-stationary linear markov decision processes. *ArXiv*, 2020.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature*, pp. 1–5, 2019.

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

Wei, C.-Y. and Luo, H. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *ArXiv*, 2021.

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021.

Yin, M., Wang, M., and Wang, Y.-X. Offline reinforcement learning with differentiable function approximation is provably efficient. *International Conference on Learning Representations*, 2023.

Yu, J. Y. and Mannor, S. Arbitrarily modulated markov decision processes. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, 2009.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2019.

Zhang, R., Zhang, X., Ni, C., and Wang, M. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *International Conference on Machine Learning*, 2022.

Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. A simple approach for non-stationary linear bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

Zhong, H., Yang, Z., Wang, Z., and Szepesvári, C. Optimistic policy optimization is provably efficient in non-stationary MDPs. *ArXiv*, 2021.

Zhou, H., Chen, J., Varshney, L. R., and Jagmohan, A. Nonstationary reinforcement learning with linear function approximation. *Transactions on Machine Learning Research*, 2022.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, 2013.

A. Proof of Proposition 3.6

In this section, we extend Bellman Eluder (BE) dimension to dynamic Bellman Eluder dimension (DBE) under the setting of small variations in transitions and rewards in the following steps.

The **First Step** is to generalize the class of Bellman residues considered in Bellman Eluder dimension. We restate the definition of Bellman Eluder dimension (Jin et al., 2021).

Definition A.1 (Bellman Eluder dimension (BE))**.** Let $(I-\mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}\}$ 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 defines as

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

For ease of presentation, we use $(f,\mathcal{T}_h)$ to denote the element $f_h - \mathcal{T}_h f_{h+1}$ in the set $(I-\mathcal{T}_h)\mathcal{F}$. For any $(f,\mathcal{T}_h)$ pair, we introduce the complement of $(f,\mathcal{T}_h)$, denoted by $(-f,-\mathcal{T}_h)$, where $-\mathcal{T}_h f' = -r_h + P_h f'$ for any f'. Let $-(I-\mathcal{T}_h)\mathcal{F}$ be the set of all complements of $(f,\mathcal{T}_h)\in(I-\mathcal{T}_h)\mathcal{F}$. Then, we define the extended class of Bellman residuals

$$\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon) := \max_{h\in[H]} \dim_{\mathrm{DE}}\left((I-\widetilde{\mathcal{T}}_h)\widetilde{\mathcal{F}},\Pi_h,\epsilon\right),$$

where $(I-\widetilde{\mathcal{T}}_h)\widetilde{\mathcal{F}} = ((I-\mathcal{T}_h)\mathcal{F})\cup(-(I-\mathcal{T}_h)\mathcal{F})$.

We first show that the BE dimension of the extended class of Bellman residuals equals to that of the original class of Bellman residuals, as formalized in the following lemma.

Lemma A.2. *Let $\tilde{\mathcal{F}}$ be defined in the above context, then we have* $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon) = \dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$.

Proof. Since $(I-\mathcal{T}_h)\mathcal{F}\subseteq(I-\widetilde{\mathcal{T}}_h)\widetilde{\mathcal{F}}$, it is obvious that $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon)\geq\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$. Next, we show $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon)\leq\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$. Let μ be independent of $\rho_1,\ldots,\rho_m$ with respect to $(I-\widetilde{\mathcal{T}}_h)\widetilde{\mathcal{F}}$. We aim to show μ is also independent of $\rho_1,\ldots,\rho_m$ with respect to $(I-\mathcal{T}_h)\mathcal{F}$.

By the definition of ϵ-independence between distributions, there exists a function g from either $(I-\mathcal{T}_h)\mathcal{F}$ or $-(I-\mathcal{T}_h)\mathcal{F}$ such that the there exists $\epsilon'\geq\epsilon$ such that $\sqrt{\sum_{i=1}^m \mathbb{E}_{\rho_i}[g]}\leq\epsilon$ and $|\mathbb{E}_\mu[g]|>\epsilon$. If g is from $(I-\mathcal{T}_h)\mathcal{F}$, then μ is obviously independent of $\rho_1,\ldots,\rho_m$ with respect to $(I-\mathcal{T}_h)\mathcal{F}$. If g is from $-(I-\mathcal{T}_h)\mathcal{F}$, i.e., g has form $g=-f_h-(-\mathcal{T}_h)(-f_{h+1})$ for some f and $\mathcal{T}_h$, we have

$$\sum_{t=1}^m(\mathbb{E}_{x\sim\rho_i}[-f_h-(-\mathcal{T}_h)(-f_{h+1})])^2 = \sum_{t=1}^m(\mathbb{E}_{x\sim\rho_i}[-f_h+r_h+P_hf_{h+1}])^2\leq\epsilon^2,$$
$$|\mathbb{E}_{x\sim\mu}[-f_h-(-\mathcal{T}_h)(-f_{h+1})]| = |\mathbb{E}_{x\sim\mu}[-f_h+r_h+P_hf_{h+1}]|>\epsilon,$$

which again implies μ is independent of $\rho_1,\ldots,\rho_m$ with respect to $\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$.

We have shown that if μ be independent of $\rho_1,\ldots,\rho_m$ with respect to $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon)$, then μ is also independent of $\rho_1,\ldots,\rho_m$ with respect to $\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$. Therefore, the length of the longest independent sequence in $\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$ must be equal or longer than that in $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon)$, i.e., $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}},\Pi,\epsilon)\leq\dim_{\mathrm{BE}}(\mathcal{F},\Pi,\epsilon)$. □

The **Second step** is to investigate the difference between two BE dimensions for different Bellman operators. Before we proceed, we define the gap in the definition of ϵ-independence between distributions.

It turns out that if the variation of the transitions and rewards are smaller than the gap $\widetilde{\delta}^u_\epsilon$, which will be defined later, then two BE dimensions induced by difference Bellman operators are comparable, as summarized in the following theorem.

Lemma A.3. *Suppose there are two MDP instances with Bellman operator $\mathcal{T}_h^1$ and $\mathcal{T}_h^2$, where $h\in[H]$. Let $\widetilde{\delta}^u_\epsilon$ be the universal gap with respect to $(I-\widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}$ (see Definition 3.5). Then, if the variation of two instances is relatively small compared to the universal gap $\widetilde{\delta}$ satisfying*

$$\max_h\sqrt{6m_{\max}H\left(\sup_{x,a}|r_h^1-r_h^2|+H\cdot\mathrm{TV}(P_h^1,P_h^2)\right)}+\left(\sup_{x,a}|r_h^1-r_h^2|+H\cdot\mathrm{TV}(P_h^1,P_h^2)\right)\leq\widetilde{\delta}^u_\epsilon,$$

where $m_{\max} = \dim_{\mathrm{BE}}((I - \mathcal{T}_h^2)\mathcal{F}, \Pi, \epsilon)$*, then*

$$\dim_{\mathrm{DE}}((I - \mathcal{T}_h^2)\mathcal{F}, \Pi, \epsilon) \le \dim_{\mathrm{DE}}((I - \mathcal{T}_h^1)\mathcal{F}, \Pi, \epsilon),$$

and

$$\dim_{\mathrm{DE}}(((I - \mathcal{T}_h^2)\mathcal{F}) \cup ((I - \mathcal{T}_h^1)\mathcal{F}), \Pi, \epsilon) = \dim_{\mathrm{DE}}((I - \mathcal{T}_h^1)\mathcal{F}, \Pi, \epsilon),$$

Proof. Fix $h \in [H]$. Let $\mu_1, \ldots, \mu_m$ be independent sequence with respect to $(I - \widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}$. By the definition of BE dimension, $m \le \dim_{\mathrm{BE}}((I - \widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}, \Pi_h, \epsilon)$. If we can show $\mu_1, \ldots, \mu_m$ is also an independent sequence with respect to $(I - \widetilde{\mathcal{T}}_h^1)\widetilde{\mathcal{F}}$, then the longest independent sequence with respect to $(I - \widetilde{\mathcal{T}}_h^1)\widetilde{\mathcal{F}}$ must be equal or longer than that with respect to $(I - \widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}$ and the proof is complete. In the following, we will focus on proving this argument.

By the condition, there exists $\epsilon' \ge \epsilon$ such that for all $i \in [m]$ we have

$$\sum_{t=1}^{i-1} (\mathbb{E}_{\mu_t}[f_h^i - \mathcal{T}_h^2 f_{h+1}^i])^2 \le \epsilon'^2,$$
$$|\mathbb{E}_{\mu_i}[f_h^i - \mathcal{T}_h^2 f_{h+1}^i]| \ge \epsilon' + \widetilde{\delta}_{i;\mu_1,\ldots,\mu_i}.$$

Here, with a little abuse of notation, the subscript i of $\widetilde{\delta}_{i;\mu_1,\ldots,\mu_i}$ represents the function $f_h^i - \mathcal{T}_h^2 f_{h+1}^i$.

By Lemma A.5, we have

$$\begin{aligned}\sum_{t=1}^{i-1} (\mathbb{E}_{x\sim\mu_t}[f_h^i - \mathcal{T}_h^1 f_h^i])^2 &\le \epsilon'^2 + 6mH\left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right) \\ &\le \left(\epsilon' + \sqrt{6mH\left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right)}\right)^2. \end{aligned} \tag{11}$$

We point it out that both the $(f^i, \mathcal{T}_h^1)$ from $(I - \mathcal{T}_h)\mathcal{F}$ and $(-f^i, -\mathcal{T}_h^i)$ from $-(I - \mathcal{T}_h)\mathcal{F}$ satisfy the above inequality.

Next, consider

$$\begin{aligned}\min\Bigg\{&\Big|\big|\mathbb{E}_{x\sim P}[f_h - r_h^2 - P_h^2 f_{h+1}]\big| - \big|\mathbb{E}_{x\sim P}[f_h - r_h^1 - P_h^1 f_{h+1}]\big|\Big|, \\ &\Big|\big|\mathbb{E}_{x\sim P}[-f_h - (-r_h^2) - P_h^2(-f_{h+1})]\big| - \big|\mathbb{E}_{x\sim P}[f_h - r_h^1 - P_h^1 f_{h+1}]\big|\Big|\Bigg\}.\end{aligned}$$

The first argument in the $\min$ function corresponds to the difference between pair $(f, \mathcal{T}_h^1)$ and $(f, \mathcal{T}_h^2)$ while the second one is the difference between pair $(f, \mathcal{T}_h^1)$ and $(-f, -\mathcal{T}_h^2)$.

If $(\mathbb{E}_{x\sim P}[f_h - r_h^2 - P_h^2 f_{h+1}])(\mathbb{E}_{x\sim P}[f_h - r_h^1 - P_h^1 f_{h+1}]) \ge 0$, then by Lemma A.6, the first argument in the $\min$ function is upper bounded by $\sup_{x,a}|r_h^1 - r_h^2| + \mathrm{TV}(P_h^1, P_h^2)$.

If $(\mathbb{E}_{x\sim P}[f_h - r_h^2 - P_h^2 f_{h+1}])(\mathbb{E}_{x\sim P}[f_h - r_h^1 - P_h^1 f_{h+1}]) < 0$, then by Lemma A.6, the second argument is upper bounded by $\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)$.

Therefore, the quantity we considered is upper bounded by $\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)$. By triangle inequality, either

$$\begin{aligned}\big|\mathbb{E}_{x\sim\mu_1}[f_h^i - \mathcal{T}_h^1 f_h^i]\big| &\ge \big|\mathbb{E}_{x\sim\mu_1}[f_h^i - \mathcal{T}_h^2 f_h^i]\big| - \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right) \\ &\ge \epsilon' + \widetilde{\delta}_{i;\mu_1,\ldots,\mu_i} - \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right) \end{aligned} \tag{12}$$

holds or

$$\big|\mathbb{E}_{x\sim\mu_1}[-f_h^i - (-\mathcal{T}_h^1)(-f_h^i)]\big| \ge \big|\mathbb{E}_{x\sim\mu_1}[f_h^i - \mathcal{T}_h^2 f_h^i]\big| - \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right)$$

$$\geq \epsilon' + \widetilde{\delta}_{i;\mu_1,\ldots,\mu_i} - \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right) \tag{13}$$

holds.

Recall $\widetilde{\delta}_\epsilon^u$ is the universal gap with respect to $(I - \widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}$ (see Definition 3.5), and $m_{\max} = \dim_{\mathrm{BE}}((I - \widetilde{\mathcal{T}}_h^2)\widetilde{\mathcal{F}}, \Pi_h, \epsilon)$. If it holds that

$$\max_h \sqrt{6m_{\max}H\left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right)} + \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right) \leq \widetilde{\delta}_\epsilon^u,$$

which implies

$$\begin{aligned}\gamma &= \left(\epsilon' + \sqrt{6mH\left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right)}\right) - \left(\epsilon' + \widetilde{\delta}_{i;\mu_1,\ldots,\mu_i} - \left(\sup_{x,a}|r_h^1 - r_h^2| + H \cdot \mathrm{TV}(P_h^1, P_h^2)\right)\right) \\ &\geq 0.\end{aligned}$$

The above inequality together with (11)-(13) shows that for the sequence $\mu_1, \ldots, \mu_i$, there exists a function g from $(I - \widetilde{\mathcal{T}}_h^1)\widetilde{\mathcal{F}}$, and a $\widetilde{\epsilon} \in [\epsilon', \epsilon' + \gamma)$ satisfying $\widetilde{\epsilon} \geq \epsilon$ such that

$$\begin{aligned}&\sum_{t=1}^{i-1}(\mathbb{E}_{\mu_t}[g])^2 \leq \widetilde{\epsilon}^2, \\ &|\mathbb{E}_{\mu_i}[g]| > \widetilde{\epsilon}.\end{aligned}$$

The above argument holds for all $i \in [m]$ and we conclude that $\mu_1, \ldots, \mu_m$ is again an independent sequence with respect to $(I - \widetilde{\mathcal{T}}_h^1)\widetilde{\mathcal{F}}$. The proof for the first inequality is complete by noting that $\dim_{\mathrm{BE}}(\widetilde{\mathcal{F}}, \Pi, \epsilon) = \dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon)$ by Lemma A.2.

For the second inequality, we are left to show $\dim_{\mathrm{DE}}(((I - \mathcal{T}_h^2)\mathcal{F}) \cup ((I - \mathcal{T}_h^1)\mathcal{F}), \Pi, \epsilon) \leq \dim_{\mathrm{DE}}((I - \mathcal{T}_h^1)\mathcal{F}, \Pi, \epsilon)$. The proof is by showcasing every independence sequence with respect to $((I - \mathcal{T}_h^2)\mathcal{F}) \cup ((I - \mathcal{T}_h^1)\mathcal{F})$ must also be independent with respect to $(I - \mathcal{T}_h^1)\mathcal{F}$, which follows exactly the same argument as above and is omitted here. □

The **Step three** is to build connection between BE dimension to DBE dimension when the variations in transitions and rewards are small. In general, DBE dimension could be substantially larger than BE dimension of one MDP instance in the non-stationary MDPs. However, if the variation of all instances are small enough compared to the universal gap $\widetilde{\delta}_{k;\epsilon}^u$ with respect to $(I - \mathcal{T}_h^k)\mathcal{F}$ for all $k \in [2 : K]$, DBE dimension is indeed equal to BE dimension. The following proposition is an immediate result from Lemma A.3.

Proposition A.4. *If it holds that for all k,*

$$\max_h \sqrt{6m_k H\left(\sup_{x,a}|r_h^1 - r_h^k| + H \cdot \mathrm{TV}(P_h^1, P_h^k)\right)} + \left(\sup_{x,a}|r_h^1 - r_h^k| + H \cdot \mathrm{TV}(P_h^1, P_h^k)\right) \leq \widetilde{\delta}_{k;\epsilon}^u,$$

where $m_k = \dim_{\mathrm{BE}}((I - \mathcal{T}_h^k)\mathcal{F}, \Pi, \epsilon)$, and $\widetilde{\delta}_{k;\epsilon}^u$ is the universal gap with respect to function class $(I - \mathcal{T}_h^k)\mathcal{F}$. Then

$$\dim_{\mathrm{DBE}}(\mathcal{F}, \Pi, \epsilon) = \dim_{\mathrm{DE}}((I - \mathcal{T}_h^1)\mathcal{F}, \Pi, \epsilon),$$

where the latter is exactly the BE dimension for the first MDP instance.

A.1. Supporting Lemmas

Lemma A.5. *Suppose $f_h \leq H$ for all h, and $r, r' \leq 1$, we have*

$$\left|(\mathbb{E}_{x\sim P}[f_h - r - Pf_{h+1}])^2 - (\mathbb{E}_{x\sim P}[f_h - r' - P'f_{h+1}])^2\right| \leq 6H\left(\sup_{x,a}(r - r') + \mathrm{TV}(P, P')\right).$$

Proof. Note that

$$\begin{aligned}
&\left|\left(\mathbb{E}_{x\sim P}[f_h - r - Pf_{h+1}]\right)^2 - \left(\mathbb{E}_{x\sim P}[f_h - r' - P'f_{h+1}]\right)^2\right| \\
&\leq 6H\left|\mathbb{E}_{x\sim P}[r - r'] + \mathbb{E}_{x\sim P}[(P-P')f_{h+1}]\right| \\
&\leq 6H\left(\left|\mathbb{E}_{x\sim P}[r - r']\right| + H\left|\mathbb{E}_{x\sim P}[(P-P')f_{h+1}]\right|\right) \\
&\leq 6H\left(\sup_{x,a}(r-r') + H\cdot \mathrm{TV}(P,P')\right).
\end{aligned}$$

□

Lemma A.6. *Suppose $f_h \leq H$ for all h, and $r, r' \leq 1$, we have*

$$\left|\mathbb{E}_{x\sim P}[f_h - r - Pf_{h+1}] - \mathbb{E}_{x\sim P}[f_h - r' - P'f_{h+1}]\right| \leq \sup_{x,a}(r-r') + H\cdot \mathrm{TV}(P,P').$$

Proof. Note that

$$\begin{aligned}
&\left|\mathbb{E}_{x\sim P}[f_h - r - Pf_{h+1}] - \mathbb{E}_{x\sim P}[f_h - r' - P'f_{h+1}]\right| \\
&\leq \left|\mathbb{E}_{x\sim P}[r - r'] + \mathbb{E}_{x\sim P}[(P-P')f_{h+1}]\right| \\
&\leq \left|\mathbb{E}_{x\sim P}[r - r']\right| + \left|\mathbb{E}_{x\sim P}[(P-P')f_{h+1}]\right| \\
&\leq \sup_{x,a}(r-r') + H\cdot \mathrm{TV}(P,P').
\end{aligned}$$

□

B. Proof of Propostion 3.8

In this section, we show the DBE dimension of non-stationary linear MDP is $\widetilde{\mathcal{O}}(d)$ where d is the feature dimension.

Define $m = \dim_{\mathrm{DBE}}((I-\mathcal{T}_h)\mathcal{F}, \mathcal{D}_{\mathcal{F},h}, \epsilon)$ and let $h = \arg\max_{h\in[H]} \dim_{\mathrm{DBE}}((I-\mathcal{T}_h)\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon)$.

Let $\mu_1, \ldots, \mu_m$ be an independent sequence with respect to $(I-\mathcal{T}_h)\mathcal{F}$. By definition, there exists $(f^1, \mathcal{T}^1), \ldots, (f^i, \mathcal{T}^i)$ such that for all $i \in [m]$, we have

$$\begin{aligned}
&\sum_{t=1}^{i-1}\left(\mathop{\mathbb{E}}_{(x,a)\sim\mu_t}\left[(f_h^i - \mathcal{T}_h^{i-1}f_{h+1}^i)(x,a)\right]\right)^2 \leq \epsilon^2, \quad \text{and} \\
&\left|\mathop{\mathbb{E}}_{(x,a)\sim\mu_i}\left[(f_h^i - \mathcal{T}_h^{i-1}f_{h+1}^i)(x,a)\right]\right| > \epsilon.
\end{aligned}$$

We aim to show $m = \widetilde{\mathcal{O}}(d)$.

For linear MDP, a natural function class $\mathcal{F}_h$ is

$$\{f \in ((\mathcal{S}\times\mathcal{A}) \mapsto [0, H-h+1]) : \phi(x,a)^\top w_h, \|\phi(x,a)\| \leq 1, \forall (x,a) \text{ and } \|w_h\| \in 2(H-h+1)\sqrt{d}\}.$$

Note that

$$(f_h^i - \mathcal{T}_h^{i-1}f_{h+1}^i)(x,a) = \phi(x,a)^\top(w_{h_i} - \widetilde{w}_{h,i}),$$

where $\widetilde{w}_{h,i} = \theta_{h,i-1} + \int_{x'}\mu_{h,i-1}(x')\max_a f_{h+1}^i(x',a)$ and we have $\max\{\|w_{h,i}\|, \|\widetilde{w}_{h,i}\|\} \leq 2H\sqrt{d}$ for all $h \in [H]$.

Therefore, for all $i \in [m]$

$$\sum_{t=1}^{i-1}\left(\mathop{\mathbb{E}}_{(x,a)\sim\mu_t}\left[\phi(x,a)^\top(\widetilde{w}_h^i - w_h^{i-1})\right]\right)^2 \leq \epsilon^2, \quad \text{and}$$

$$\left| \mathop{\mathbb{E}}_{(x,a)\sim\mu_i} \left[\phi(x,a)^\top (\widetilde{w}_h^i - w_h^{i-1})\right] \right| > \epsilon.$$

For ease of exposition, we set

$$\mathbf{x}_i = \widetilde{w}_h^i - w_h^{i-1}, \quad \mathbf{z}_i = \mathop{\mathbb{E}}_{(x,a)\sim\mu_i} [\phi(x,a)], \quad \mathbf{V}_i = \sum_{t=1}^{i-1} \mathbf{z_t} \mathbf{z_t}^\top + \frac{\epsilon^2}{\zeta} \cdot I,$$

where $\zeta = 4H\sqrt{d}$.

The previous argument implies that for all $i \in [m]$,

$$\|\mathbf{x}_i\|_{\mathbf{V}_i} \leq \sqrt{2}\epsilon,$$
$$\|\mathbf{x}_i\|_{\mathbf{V}_i} \cdot \|\mathbf{z}_i\|_{\mathbf{V}_i^{-1}} > \epsilon.$$

Therefore, we have $\|\mathbf{z}_i\|_{\mathbf{V}_i^{-1}} \geq \frac{1}{\sqrt{2}}$.

By matrix determinant lemma,

$$\det[\mathbf{V}_m] = \det[\mathbf{V}_{m-1}] \left(1 + \|\mathbf{z}_m\|^2_{\mathbf{V}_m^{-1}}\right) \geq \cdots \geq (\frac{3}{2})^{m-1} (\frac{\epsilon^2}{\zeta})^d.$$

Moreover,

$$\det[\mathbf{V}_m] \leq \left(\frac{\mathrm{tr}[\mathbf{V}_m]}{d}\right)^d \leq \left(\frac{\zeta(m-1)}{d} + \frac{\epsilon^2}{\zeta}\right)^d.$$

Therefore,

$$(\frac{3}{2})^{m-1} \leq \left(\frac{\zeta^2(m-1)}{d\epsilon^2} + 1\right)^d.$$

Taking logarithm on both sides gives

$$m \leq 4\left[1 + d\log\left(\frac{\zeta^2(m-1)}{d\epsilon^2} + 1\right)\right],$$

which implies

$$m \leq \mathcal{O}\left(1 + d\log\left(\frac{\zeta^2}{\epsilon^2} + 1\right)\right).$$

C. Proofs of SW-OPEA

In this section, we provide the formal Proof of Theorem 5.2.

C.1. Proof of Theorem 5.2

We decompose the dynamic regret in the following way

$$\mathrm{D-Regret}(k) = \sum_{t=1}^{k} \left(V_{1;(*,t)}^{\pi^{(*,t)}} - V_{1;(*,t)}^{\pi^t}\right)(x_1)$$
$$= \sum_{t=1}^{k} \left(V_{1;(*,t)}^{\pi^{(*,t)}} - V_{1;(*,t-1)}^{\pi^{(*,t-1)}} + V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t} + V_{1;(*,t-1)}^{\pi^t} - V_{1;(*,t)}^{\pi^t}\right)(x_1)$$

$$
\begin{aligned}
&= \left(V_{1;(*,k)}^{\pi^{(*,k)}} - V_{1;(*,0)}^{\pi^{(*,0)}}\right)(x_1) + \sum_{t=1}^{k}\left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t}\right)(x_1) \\
&\qquad + \sum_{t=1}^{k}\sum_{h=1}^{H}\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[r_h^{t-1}(x_h,a_h)] - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t))}[r_h^t(x_h,a_h)]\right) \\
&\le H + \underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H}\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(r_h^{t-1} - r_h^t)(x_h,a_h)]\right)}_{\text{(I)}} \\
&\quad + \underbrace{\sum_{t=1}^{k}\sum_{h=1}^{H}\left(\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t))}\right)[r_h^t(x_h,a_h)]\right)}_{\text{(II)}} \\
&\quad + \underbrace{\sum_{t=1}^{k}\left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t}\right)(x_1)}_{\text{(III)}}.
\end{aligned}
$$

By the definition of variation in rewards (3), we have $\text{(I)} \le \Delta_R(k)$.

We bound (II) using the following lemma.

Lemma C.1. *Fix $(k,h) \in [K]\times[H]$, we have*

$$
\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k))}\right)[r_h^k(x_h,a_h)] \le \sum_{i=1}^{h-1}\left\|\mathbb{P}_i^{k-1} - \mathbb{P}_i^k\right\|_\infty.
$$

Moreover,

$$
\sum_{t=1}^{k}\sum_{h=1}^{H}\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t))}\right)[r_h^t(x_h,a_h)] \le H\Delta_P(k).
$$

The proof of C.1 is provided in Appendix C.3.

Therefore,

$$
\mathrm{D-Regret}(k) = H + \Delta_R(k) + H\Delta_P(k) + \underbrace{\sum_{t=1}^{k}\left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t}\right)(x_1)}_{\text{(III)}}.
$$

Before we proceed, we present the next two lemmas.

Lemma C.2. *If $\beta = cH^2\log\frac{KH|\mathcal{G}|}{\delta}$, then with probability at least $1-\delta$, we have $Q^*_{(*,k)} \in \mathcal{B}^k$ for all $k \in [K]$.*

Lemma C.3. *If $\beta = cH^2\log\frac{KH|\mathcal{G}|}{\delta}$, then with probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, we have*

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

The proofs of Lemma C.2 and C.3 are based on martingale concentration and provided in Appendix C.4.

By Lemma C.2, with probability at least $1-\delta$, we have

$$
\text{(III)} = \sum_{t=1}^{k}\left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t}\right)(x_1)
$$

$$\leq \sum_{t=1}^{k} \left(\max_{a \in \mathcal{A}} f_1^t(x_1, a) - V_{1;(*,t-1)}^{\pi^t}(x_1) \right)$$
$$\leq \sum_{h=1}^{H} \sum_{t=1}^{k} \mathop{\mathbb{E}}_{(x_h,a_h) \sim (\pi^t,(*,t-1))} [(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h, a_h)],$$

where the first inequality follows from Lemma C.2 and the optimistic planning step (line 3) in Algorithm 4 which guarantees that $V_{1;(*,k-1)}^* \leq \sup_a f_1^k(x_1, a)$ for every episode k, the last inequality follows from generalized policy loss decomposition (Lemma C.8) and the fact that $\pi^k = \pi_{f^k}$ (line 3 in Algorithm 4).

The next lemma is adapted from ((Jin et al., 2021)) and the proof can be found in Appendix C.5.

Lemma C.4. *Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \leq C$ 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 $k \in [K]$, $\sum_{t=1}^{k-1} (\mathbb{E}_{x\sim\mu_t}[\phi_k(x)])^2 \leq \beta$. Then for all $k \in [K]$ and $\omega > 0$,*

$$\sum_{t=1\vee(k-w)}^{k} |\mathbb{E}_{x\sim\mu_t}[\phi_t(x)]|$$
$$\leq \mathcal{O}\left(\sqrt{\dim_{\mathrm{DE}}(\Phi, \Pi, \theta)\beta[k \wedge (w+1)]} + \min\{w+1, k, \dim_{\mathrm{DE}}(\Phi, \Pi, \theta)\}C + [k \wedge (w+1)]\theta \right).$$

We invoke Lemma 5.5 and Lemma C.3 with

$$\theta = \sqrt{\frac{1}{w}}, C = H,$$
$$\mathcal{X} = \mathcal{S} \times \mathcal{A}, \Phi = (I - \mathcal{T}_h)\mathcal{F}, \text{and } \Pi = \mathcal{D}_{\Delta,h},$$
$$\phi_k = f_h^k - \mathcal{T}_h^{k-1} f_{h+1}^k, \mu_k = \mathbf{1}\{\cdot = (x_h^k, a_h^k)\}$$

and obtain

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

where the second inequality follows from Azuma-Hoeffding inequality, and in the last inequality, we use $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any positive $a, b \geq 0$ and we define $d = \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})$.

Summing over step $h \in [H]$ gives

$$\sum_{h=1}^{H} \sum_{t=1}^{k} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))} [(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h, a_h)]$$
$$\leq \mathcal{O}\left(\frac{H^2 k}{\sqrt{w}} \sqrt{d \log[KH|\mathcal{G}|/\delta]} + \frac{H^2 k}{\sqrt{w}} \sqrt{d \sup_{t\in[k]} \Delta_P^w(t, h)} + H\sqrt{w} \right),$$

which completes the proof.

C.2. Proof of Corollary 5.3

For ease of exposition, let $d = \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})$. We adopt average variation L defined in (5). Then we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{K} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)] \\
&\le \widetilde{\mathcal{O}}\left(\frac{H^2K}{\sqrt{w}}\sqrt{d}\sqrt{\log|\mathcal{G}|} + \frac{H^2K}{\sqrt{w}}\sqrt{dLw^2} + H\sqrt{w}\right) \\
&\le \widetilde{\mathcal{O}}\left(H^2K\sqrt{d}\left(\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{w}} + (\sqrt{L} + \frac{1}{HK\sqrt{d}})\sqrt{w}\right)\right).
\end{aligned}$$

Note first that $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{1}{HK\sqrt{d}}} > 1$ when $|\mathcal{G}| > 10$.

If $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{1}{HK\sqrt{d}}} \ge K$, i.e., $\sqrt{L} \le \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|} - \frac{1}{H\sqrt{d}}\right)$, we select $w = K$ and we have

$$\sum_{h=1}^{H}\sum_{t=1}^{K} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)] \le \widetilde{O}\left(H^2K^{\frac{1}{2}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{2}}\right).$$

If $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{1}{HK\sqrt{d}}} < K$, i.e., $\sqrt{L} > \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|} - \frac{1}{H\sqrt{d}}\right)$, we select $w = \lceil \frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{1}{HK\sqrt{d}}} \rceil$ and we have

$$\sum_{h=1}^{H}\sum_{t=1}^{K} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)] \le \widetilde{\mathcal{O}}\left(H^2KL^{\frac{1}{4}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{4}} + H^{\frac{3}{2}}K^{\frac{1}{2}}d^{\frac{1}{4}}(\log|\mathcal{G}|)^{\frac{1}{4}}\right).$$

C.3. Proof of Lemma C.1

Proof. Fix $(k,h) \in [K]\times[H]$, define reward function $\widetilde{r}_{h'} = r_h^k(x,a)\mathbf{1}\{h'=h\}$ for all $h' \in [H]$. For an episodic MDP $(\mathcal{S},\mathcal{A},H,P^k,\widetilde{r},x_1)$ where $\{P_{h'}^k\}_{h'\in[H]}$ and $\{\widetilde{r}_{h'}\}_{h'\in[H]}$, the state value function and state-action value function of policy $\{\pi_{h'}\}_{h'\in[H]}$ are $\widetilde{V}_{h';,(*,k)}^{\pi}$ and $\widetilde{Q}_{h';,(*,k)}^{\pi}$. Clearly, we have

$$\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k))}\right)[r_h^k(x_h,a_h)] = \left(\widetilde{V}_{1;(*,k-1)}^{\pi^k} - \widetilde{V}_{1;(*,k)}^{\pi^k}\right)(x_1).$$

For any function $f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and any $(k,h,x) \in [K]\times[H]\times\mathcal{S}$, define the following operator

$$(\mathbb{J}_{k,h}f)(x) = \langle f(x,\cdot), \pi_h^k(\cdot|x)\rangle.$$

Note that

$$\begin{aligned}
&\widetilde{V}_{1;(*,k-1)}^{\pi^k} - \widetilde{V}_{1;(*,k)}^{\pi^k} \\
&= \mathbb{J}_{k,1}\left(\widetilde{Q}_{1;(*,k-1)}^{\pi^k} - \widetilde{Q}_{1;(*,k)}^{\pi^k}\right) \\
&= \mathbb{J}_{k,1}\left(\mathbb{P}_1^{k-1}\widetilde{V}_{2;(*,k-1)}^{\pi^k} - \mathbb{P}_1^{k}\widetilde{V}_{2;(*,k)}^{\pi^k}\right) \\
&= \mathbb{J}_{k,1}\mathbb{P}_1^{k-1}\left(\widetilde{V}_{2;(*,k-1)}^{\pi^k} - \widetilde{V}_{2;(*,k)}^{\pi^k}\right) + \mathbb{J}_{k,1}\left(\mathbb{P}_1^{k-1} - \mathbb{P}_1^{k}\right)\widetilde{V}_{2;(*,k)}^{\pi^k} \\
&= \prod_{i=1}^{h}\left(\mathbb{J}_{k,i}\mathbb{P}_i^{k-1}\right)\underbrace{\left(\widetilde{V}_{h+1;(*,k-1)}^{\pi^k} - \widetilde{V}_{h+1;(*,k)}^{\pi^k}\right)}_{=0} + \sum_{i=1}^{h}\prod_{\ell=1}^{i-1}\left(\mathbb{J}_{k,\ell}\mathbb{P}_\ell^{k-1}\right)\mathbb{J}_{k,i}\left(\mathbb{P}_i^{k-1} - \mathbb{P}_i^{k}\right)\widetilde{V}_{i+1,(*,k)}^{\pi^k} \\
&= \sum_{i=1}^{h-1}\prod_{\ell=1}^{i-1}\left(\mathbb{J}_{k,\ell}\mathbb{P}_\ell^{k-1}\right)\mathbb{J}_{k,i}\left(\mathbb{P}_i^{k-1} - \mathbb{P}_i^{k}\right)\widetilde{V}_{i+1,(*,k)}^{\pi^k}.
\end{aligned}$$

where in the second equality we use the fact that reward $\widetilde{r}$ is identical. I.e.,

$$\begin{aligned}
&\left(\widetilde{V}^{\pi^k}_{1;(*,k-1)} - \widetilde{V}^{\pi^k}_{1;(*,k)}\right)(x_1) \\
&= \sum_{i=1}^{h-1} \mathop{\mathbb{E}}_{(x_i,a_i)\sim(\pi^k,(*,k-1))} \left[\left(\left(\mathbb{P}_i^{k-1} - \mathbb{P}_i^k\right)\widetilde{V}^{\pi^k}_{i+1,(*,k)}\right)(x_i,a_i)\right] \\
&\leq \sum_{i=1}^{h-1} \sup_{x,a} \left\|P_i^{k-1}(\cdot|x,a) - P_i^k(\cdot|x,a)\right\|_1
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{k=1}^{k'}\sum_{h=1}^{H}\left(\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k-1))} - \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^k,(*,k))}\right)[r_h^k(x_h,a_h)] \\
&\leq \sum_{k=1}^{k'}\sum_{h=1}^{H}\sum_{i=1}^{h-1} \sup_{x,a} \left\|P_i^{k-1}(\cdot|x,a) - P_i^k(\cdot|x,a)\right\|_1 \\
&\leq \sum_{k=1}^{k'}\sum_{h=1}^{H}\sum_{i=1}^{H} \sup_{x,a} \left\|P_i^{k-1}(\cdot|x,a) - P_i^k(\cdot|x,a)\right\|_1 \\
&\leq \sum_{h=1}^{H}\left(\sum_{k=1}^{k'}\sum_{i=1}^{H} \sup_{x,a} \left\|P_i^{k-1}(\cdot|x,a) - P_i^k(\cdot|x,a)\right\|_1\right) \\
&\leq H\Delta_P(k').
\end{aligned}$$

□

C.4. Proofs of concentration lemmas

The Freedman's inquaulity controls the sum of martingale difference by the sum of their variance.

Lemma C.5 (Freedman's inequality ((Jin et al., 2021))). *Let $\{Z_t\}_{t\in[T]}$ be a real-valued martingale difference sequence adapted to filtration $\mathcal{F}_t$, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. If $|Z_t| \leq R$ almost surely, then for any $\eta \in (0, R)$, it holds that with probability at least $1-\delta$,*

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

C.4.1. Proof of Lemma C.2

Proof. Define

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

Fix a tuple $(k,h,g) \in [K]\times[H]\times\mathcal{G}$. Let

$$\begin{aligned}
W_t(h,f) :&= \left[g_h(x_h^t,a_h^t) - r_h^k - \max_{a'\in\mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t,a'))\right]^2 - \left[\#_{k,h}(x_h^t,a_h^t) - r_h^k - \max_{a'\in\mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t,a'))\right]^2 \\
&= \left[g_h(x_h^t,a_h^t) - \#_{k,h}(x_h^t,a_h^t)\right]\left[g_h(x_h^t,a_h^t) + \#_{k,h}(x_h^t,a_h^t) - 2\left(r_h^k + \max_{a'\in\mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t,a')\right)\right]
\end{aligned}$$

and $\mathcal{F}_{t,h}$ be the filtration induced by $\{x_1^i, a_1^i, \cdots, x_H^i\}_{i\in[t-1]} \cup \{x_1^t, a_1^t, \cdots, x_h^t, a_h^t\} \cup \{r_h^i\}_{h\in[H]}^{i\in[t-1]}$. We have

$$\mathbb{E}[W_t(h,g)|\mathcal{F}_{t,h}] = \left[\left(g_h - \#_{k,h}\right)(x_h^t,a_h^t)\right]^2,$$

$$\mathrm{Var}[W_t(h,g)|\mathcal{F}_{t,h}] \le 36H^2\mathbb{E}[W_t(h,g)|\mathcal{F}_{t,h}].$$

By Freedman's inequality, with probability at least $1-\delta$,

$$\left|\sum_{t=1\vee(k-w)}^{k} W_t(h,g) - \sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2\right| \le \mathcal{O}\left(H\sqrt{\log(1/\delta)\sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2}+\log(1/\delta)\right).$$

Taking union bound over $[K]\times[H]\times\mathcal{G}$,

$$\left|\sum_{t=1\vee(k-w)}^{k} W_t(h,g) - \sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2\right| \le \mathcal{O}\left(H\sqrt{\iota\sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2}+\iota\right),$$

where $\iota=\log(HK|\mathcal{G}|/\delta)$. We have

$$\begin{aligned}
&-\sum_{t=1\vee(k-w)}^{k} W_t(h,g)\\
&\le -\sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2+\mathcal{O}\left(H\sqrt{\iota\sum_{t=1\vee(k-w)}^{k}\left[\left(g_h(x_h^t,a_h^t)-\#_{k,h}\right)(x_h^t,a_h^t)\right]^2}+\iota\right)\\
&\le \mathcal{O}(H^2\iota).
\end{aligned}$$

I.e.,

$$\begin{aligned}
&\sum_{t=1\vee(k-w)}^{k}\left[\#_{k,h}(x_h^t,a_h^t)-r_h^k-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2\\
&\qquad\le \sum_{t=1\vee(k-w)}^{k}\left[g_h(x_h^t,a_h^t)-r_h^k-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2+\mathcal{O}(H^2\iota).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{t=1\vee(k-w)}^{k}\left[Q_{h;(*,k)}(x_h^t,a_h^t)-r_h^k-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2\\
&\le \sum_{t=1\vee(k-w)}^{k}\left[\#_{k,h}(x_h^t,a_h^t)-r_h^k-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2+2H^2\Delta_P^w(k,h)\\
&\le \sum_{t=1\vee(k-w)}^{k}\left[g_h(x_h^t,a_h^t)-r_h^k-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2+2H^2\Delta_P^w(k,h)+\mathcal{O}(H^2\iota),
\end{aligned}$$

where the first inequality follows from Lemma C.7 and Eqn. (2). By the definition of $\mathcal{B}^k$ and $\beta=cH^2\log\frac{KH|\mathcal{G}|}{\delta}$ with some large absolute constant c, we conclude that with probability at least $1-\delta$, $Q_{(*,k)}\in\mathcal{B}^k$ for all $k\in[K]$. □

C.4.2. PROOF OF LEMMA C.3

Proof. Define

$$\#^f_{k,h}(x^t_h, a^t_h) = r^k_h(s^t_h, a^t_h) + \mathop{\mathbb{E}}_{x' \sim P^t_h(x^t_h, a^t_h)} \max_{a' \in \mathcal{A}} f_{h+1}(s', a').$$

Fix a tuple $(k, h, f) \in [K] \times [H] \times \mathcal{G}$. Let

$$\begin{aligned} W_t(h, f) :&= \left[f_h(x^t_h, a^t_h) - r^k_h - \max_{a' \in \mathcal{A}} f_{h+1}(x^t_{h+1}, a') \right]^2 - \left[\#^f_{k,h}(x^t_h, a^t_h) - r^k_h - \max_{a' \in \mathcal{A}} f_{h+1}(x^t_{h+1}, a') \right]^2 \\ &= [f_h(x^t_h, a^t_h) - \#^f_{k,h}(x^t_h, a^t_h)] \left[f_h(x^t_h, a^t_h) + \#^f_{k,h}(x^t_h, a^t_h) - 2 \left(r^k_h + \max_{a' \in \mathcal{A}} f_{h+1}(x^t_{h+1}, a') \right) \right] \end{aligned}$$

and $\mathcal{F}_{t,h}$ be the filtration induced by $\{x^i_1, a^i_1, \cdots, x^i_H\}_{i \in [t-1]} \cup \{x^t_1, a^t_1, \cdots, x^t_h, a^t_h\} \cup \{r^i_h\}^{i \in [t-1]}_{h \in [H]}$. We have

$$\begin{aligned} \mathbb{E}[W_t(h, f) | \mathcal{F}_{t,h}] &= \left[\left(f_h - \#^f_{k,h} \right)(x^t_h, a^t_h) \right]^2, \\ \mathrm{Var}[W_t(h, f) | \mathcal{F}_{t,h}] &\le 36 H^2 \mathbb{E}[W_t(h, g) | \mathcal{F}_{t,h}]. \end{aligned}$$

By Freedman's inequality, we have

$$\begin{aligned} &\left| \sum_{t=1 \vee (k-w)}^{k} W_t(h, f) - \sum_{t=1 \vee (k-w)}^{k} \left[(f_h - \#^f_{k,h})(x^t_h, a^t_h) \right]^2 \right| \\ &\qquad \le \mathcal{O}\left(H \sqrt{\log(1/\delta) \sum_{t=1 \vee (k-w)}^{k} \left[(f_h - \#^f_{k,h})(x^t_h, a^t_h) \right]^2} + \log(1/\delta) \right). \end{aligned}$$

Taking union bound over $[K] \times [H] \times \mathcal{G}$, we have

$$\left| \sum_{t=1 \vee (k-w)}^{k} W_t(h, g) - \sum_{t=1}^{k} \left[(f_h - \#^f_{k,h})(x^t_h, a^t_h) \right]^2 \right| \le \mathcal{O}\left(H \sqrt{\iota \sum_{t=1 \vee (k-w)}^{k} \left[(f_h - \#^f_{k,h})(x^t_h, a^t_h) \right]^2} + \iota \right),$$

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

Note that

$$\begin{aligned} &\sum_{t=1 \vee (k-w-1)}^{k-1} W_t(h, f^k) \\ &= \sum_{t=1 \vee (k-w-1)}^{k-1} \left[f^k_h(x^t_h, a^t_h) - r^{k-1}_h - \max_{a' \in \mathcal{A}} f^k_{h+1}(x^t_{h+1}, a') \right]^2 \\ &\qquad - \sum_{t=1 \vee (k-w-1)}^{k-1} \left[\#^{f^k}_{k-1,h}(x^t_h, a^t_h) - r^{k-1}_h - \max_{a' \in \mathcal{A}} f^k_{h+1}(x^t_{h+1}, a') \right]^2 \\ &\le \sum_{t=1 \vee (k-w-1)}^{k-1} \left[f^k_h(x^t_h, a^t_h) - r^{k-1}_h - \max_{a' \in \mathcal{A}} f^k_{h+1}(x^t_{h+1}, a') \right]^2 \\ &\qquad - \sum_{t=1 \vee (k-w-1)}^{k-1} \left[\mathcal{T}^{k-1}_h f^k_{h+1}(x^t_h, a^t_h) - r^{k-1}_h - \max_{a' \in \mathcal{A}} f^k_{h+1}(x^t_{h+1}, a') \right]^2 + 2H^2 \Delta^w_P(k-1, h) \\ &\le \sum_{t=1 \vee (k-w-1)}^{k-1} \left[f^k_h(x^t_h, a^t_h) - r^{k-1}_h - \max_{a' \in \mathcal{A}} f^k_{h+1}(x^t_{h+1}, a') \right]^2 \end{aligned}$$

$$- \inf_{g\in\mathcal{G}} \sum_{t=1\vee(k-w-1)}^{k-1} \left[g_h(x_h^t, a_h^t) - r_h^{k-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 + 2H^2\Delta_P^w(k-1,h)$$
$$\leq \beta + 4H^2\Delta_P^w(k-1,h),$$

where the first inequality follows from Lemma C.7 and Eqn. (2), the second inequality follows from Assumption 5.1, and the last inequality follows from the definition of $\mathcal{B}^{k-1}$.

Therefore,

$$\sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \#_{k-1,h}^{f^k})(x_h^t, a_h^t) \right]^2 \leq \beta + 4H^2\Delta_P^w(k-1,h) + \mathcal{O}\left(H^2\iota\right).$$

Finally, we use Lemma C.7 once more and obtain

$$\sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \mathcal{T}_h^{k-1} f_{h+1}^k)(x_h^t, a_h^t) \right]^2$$
$$\leq \sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \#_{k-1,h}^{f^k})(x_h^t, a_h^t) \right]^2 + 2H^2\Delta_P^w(k-1,h)$$
$$\leq 6H^2\Delta_P^w(k-1,h) + \mathcal{O}(\beta).$$

□

C.5. Proof of Lemma 5.5

The proof in the subsection essentially follows the same arguments as in (Jin et al., 2021), and we adapt it to the sliding window scenario.

Lemma C.6. *Given a function class Φ defined on $\mathcal{X}\times\mathcal{Y}$, and a family of probability measures Π over $\mathcal{X}$. Suppose sequence $\{\phi_k\}_{k\in[K]} \subseteq \Phi$ and $\{\mu_k\}_{k\in[K]} \subseteq \Pi$ satisfy that for all $k\in[K]$, $\sum_{t=1\vee(k-w-1)}^{k-1}(\mathbb{E}_{x\sim\mu_t}[\phi_k(x)])^2 \leq \beta$. Then for all $k\in[K]$,*

$$\sum_{t=1\vee(k-w)}^{k} \mathbf{1}\{|\mathbb{E}_{x\sim\mu_t}[\phi_t(x)]| > \epsilon\} \leq (\frac{\beta}{\epsilon^2}+1)\dim_{\mathrm{DE}}(\Phi,\Pi,\epsilon)$$

Proof. First, suppose for all $k\in[\kappa]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{x\sim\mu_t}[\phi_k(x)])^2 \leq \beta$, we show that if for some $k\in[\kappa]$ we have $|\mathbb{E}_{x\sim\mu_k}[\phi_k(x)]| > \epsilon$, then μ_k is ϵ-dependent on at most $\lceil\beta/\epsilon^2\rceil - 1$ disjoint subsequences in $\{\mu_1,\ldots,\mu_{k-1}\}$. By definition of GDE, if $|\mathbb{E}_{x\sim\mu_k}[\phi_k(x)]| > \epsilon$ and μ_k is ϵ-dependent on a subsequence $\{\nu_1,\ldots,\nu_\ell\}$ of $\{\mu_1,\ldots,\mu_{k-1}\}$, then we should have $\sum_{t=1}^{\ell}(\mathbb{E}_{x\sim\nu_t}[\phi_k(x)])^2 > \epsilon^2$. It implies that if μ_k is ϵ-dependent on L disjoint subsequences in $\{\mu_1,\ldots,\mu_{k-1}\}$, we have

$$\beta \geq \sum_{t=1\vee(k-w-1)}^{k-1} (\mathbb{E}_{x\sim\mu_t}[\phi_k(x)])^2 > L\epsilon^2,$$

which implies $L \leq \lceil\beta/\epsilon^2\rceil - 1$.

Second, we show that for any sequence $\{\nu_1,\ldots,\nu_\kappa\}\subseteq\Pi$, there exists $j\in[\kappa]$ such that ν_j is ϵ-dependent on at least $L = \lceil(\kappa-1)/\dim_{\mathrm{DE}}(\Phi,\Pi,\epsilon)\rceil$ disjoint subsequences in $\{\nu_1,\ldots,\nu_{j-1}\}$. We prove the argument by the following artificial procedure: we start with singleton sequences $B_1=\{\nu_1\}$, $B_2=\{\nu_2\}$, …, $B_L=\{\nu_L\}$ and $j=L+1$. For each j, if ν_j is ϵ-dependent on $B_1,B_2,\ldots,B_L$, we already achieved the goal and we stop; otherwise, we pick an $i\in[L]$ such that ν_j is ϵ-independent of B_i and update $B_i\cup\{\nu_j\}$. Then we increment j by 1 continue this process. By the definition of GDE dimension, the size of each $B_1,B_2,\ldots,B_L$ cannot get bigger than $\dim_{\mathrm{DE}}(\Phi,\Pi,\epsilon)$ at any point in this process. Therefore, the process stops before or on $j = L\dim_{\mathrm{DE}}(\Phi,\Pi,\epsilon)+1 \leq \kappa$.

Now fix $k \in [K]$ and let $\{\nu_1, \ldots, \nu_\kappa\}$ be the subsequence of $\{\mu_{1\vee(k-w)}, \ldots, \mu_k\}$, consisting of elements for which $|\mathbb{E}_{x\sim\mu_t}[|\phi_t(x)|]| > \epsilon$ and the corresponding bijective function is $\theta : [\kappa] \mapsto [1 \vee (k-w) : k]$. Note that for all $\ell \in [\kappa]$, we have $|\mathbb{E}_{x\sim\nu_\ell}[|\phi_{\theta(\ell)}(x)|]| > \epsilon$ and

$$\sum_{t=1}^{\ell-1}(\mathbb{E}_{x\sim\nu_t}[\phi_{\theta(\ell)}(x)]) \le \sum_{t=1\vee\theta(\ell)-w-1}^{\theta(\ell)-1}(\mathbb{E}_{x\sim\mu_t}[\phi_{\theta(\ell)}(x)]) \le \beta.$$

Using the first claim, we know that each ν_j is ϵ-dependent on at most $L < \lceil\beta/\epsilon^2\rceil - 1$ disjoint subsequences of $\{\nu_1, \nu_2, \ldots, \nu_{j-1}\}$. Using the second claim, we know that there exists $j \in [\kappa]$ such that ν_j is ϵ-dependent on at least $\lceil(\kappa-1)/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\rceil$ disjoint subsequences of $\{\nu_1, \nu_2, \ldots, \nu_{j-1}\}$. Therefore, we have $\lceil(\kappa-1)/\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)\rceil \le \lceil\beta/\epsilon^2\rceil - 1$, which implies

$$\kappa < (\frac{\beta}{\epsilon^2} + 1)\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon).$$

□

Proof of Lemma 5.5. Fix $k \in [K]$ and let $d = \dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon)$. Sort the sequence

$$\big\{|\mathbb{E}_{x\sim\mu_{1\vee(k-w)}}[\phi_{1\vee(k-w)}(x)]|, \ldots, |\mathbb{E}_{x\sim\mu_k}[\phi_k(x)]|\big\}$$

in decreasing order and denote it by $\{e_1, e_2, \ldots, e_{k\wedge(w+1)}\}$ ($e_1 \ge e_2 \ge \cdots \ge e_{k\wedge(w+1)}$). Note that

$$\begin{aligned}\sum_{t=1\vee(k-w)}^{k}|\mathbb{E}_{x\sim\mu_t}[\phi_t(x,y)]| = \sum_{t=1}^{k\wedge(w+1)} e_t = \sum_{t=1}^{k\wedge(w+1)} e_t\mathbf{1}\{e_t \le \theta\} + \sum_{t=1}^{k\wedge(w+1)} e_t\mathbf{1}\{e_t > \theta\}\\ \le [k\wedge(w+1)]\theta + \sum_{t=1}^{k\wedge(w+1)} e_t\mathbf{1}\{e_t > \theta\}\end{aligned}$$

For $t \in [k]$, we show that if $e_t > \theta$, then we have $e_t \le \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Assume $t \in [k]$ satisfies $e_t > \theta$. Then there exists an α such that $e_t > \alpha \ge \theta$. By Lemma C.6, we have

$$t \le \sum_{i=1}^{k\wedge(w+1)}\mathbf{1}\{e_i > \alpha\} \le (\frac{\beta}{\alpha^2} + 1)\dim_{\mathrm{DE}}(\Phi, \Pi, \alpha) \le (\frac{\beta}{\alpha^2} + 1)\dim_{\mathrm{DE}}(\Phi, \Pi, \omega),$$

which implies $\alpha \le \sqrt{\frac{d\beta}{t-d}}$. Letting $\alpha \to e_t$, we have $e_t \le \sqrt{\frac{d\beta}{t-d}}$. Besides, recall $e_t \le C$, so we have $e_t \le \min\{\sqrt{\frac{d\beta}{t-d}}, C\}$. Finally, we have

$$\begin{aligned}\sum_{t=1}^{k\wedge(w+1)} e_t\mathbf{1}\{e_t > \omega\} &\le \min\{d, k, w+1\}C + \sum_{t=d+1}^{k\wedge(w+1)}\sqrt{\frac{d\beta}{t-d}}\\ &\le \min\{d, k, w+1\}C + \sqrt{d\beta}\int_0^{k\wedge(w+1)}\frac{1}{\sqrt{t}}\,\mathrm{d}t\\ &= \min\{d, k, w+1\}C + 2\sqrt{d\beta[k\wedge(w+1)]}.\end{aligned}$$

C.6. Auxiliary Lemmas

Lemma C.7. *Suppose P and Q are two probability distributions of a random variable x, then*

$$\left|\Big(\mathop{\mathbb{E}}_{x\sim P} f(x) - C\Big)^2 - \Big(\mathop{\mathbb{E}}_{x\sim Q} f(x) - C\Big)^2\right| \le (2f_m + 2|C|)f_m \cdot \mathrm{TV}(P, Q),$$

where $f_m = \sup_x |f(x)|$.

Proof. Note that

$$\begin{aligned}
&\left|\left(\mathop{\mathbb{E}}_{x\sim P} f(x) - C\right)^2 - \left(\mathop{\mathbb{E}}_{x\sim Q} f(x) - C\right)^2\right| \\
&= \left|\left(\mathop{\mathbb{E}}_{x\sim P} f(x) + \mathop{\mathbb{E}}_{x\sim Q} f(x) - 2C\right)\left(\mathop{\mathbb{E}}_{x\sim P} f(x) - \mathop{\mathbb{E}}_{x\sim Q} f(x)\right)\right| \\
&\le (2f_m + 2|C|)\left|\int_x f(x)(\mathrm{d}P - \mathrm{d}Q)\right| \\
&\le (2f_m + 2|C|) f_m \cdot \mathrm{TV}(P, Q).
\end{aligned}$$

□

Lemma C.8 (Generalized policy loss decomposition)**.** *For any t, k, we have*

$$f_1^t(x_1, \pi_1^t(x_1)) - V_{1;(*,k)}^{\pi^t}(x_1) = \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,k))} \left[(f_h^t - r_{h;(*,k)} - \mathbb{P}_h^k f_{h+1}^t)(x_h, a_h)\right],$$

where $\pi_t := \pi_{f^t}$, the greedy policy under function approximation f^t.

Proof. Note that

$$\begin{aligned}
&\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,k))} \left[(f_h^t - r_{h;(*,k)} - \mathbb{P}_h^k f_{h+1}^t)(x_h, a_h)\right] \\
&= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(x_h,a_h,x_{h+1})\sim(\pi^t,(*,k))} \left[f_h^t(x_h, a_h) - r_{h;(*,k)}(x_h, a_h) - \max_{a\in\mathcal{A}} f_{h+1}^t(x_{h+1}, a)\right] \\
&= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,k))} \left[f_h^t(x_h, a_h) - r_{h;(*,k)}(x_h, a_h) - \mathop{\mathbb{E}}_{(x_{h+1},a_{h+1})\sim(\pi^t,(*,k))}[f_{h+1}^t(x_{h+1}, a_{h+1})]\right] \\
&= \mathop{\mathbb{E}}_{(x_{1:H},a_{1:H})\sim(\pi^t,(*,k))} \left[\sum_{h=1}^{H} \left(f_h^t(x_h, a_h) - f_{h+1}^t(x_{h+1}, a_{h+1})\right)\right] - \mathop{\mathbb{E}}_{(x_{1:H},a_{1:H})\sim(\pi^t,(*,k))} \left[\sum_{h=1}^{H} r_{h;(*,k)}(x_h, a_h)\right] \\
&= f_1^t(x_1, \pi_1^t(x_1)) - V_{1;(*,k)}^{\pi^t}(x_1),
\end{aligned}$$

where the second equality follows from $\pi_t = \pi_{f^t}$. □

D. Bandit Feedback

We extend our algorithm to bandit feedback scenario, and the pseudocode is presented in Algorithm 2. In bandit feedback scenario, the reward function $r_h^k(\cdot,\cdot)$ is no long available, and the agent can only get access to the reward obtained from the trajectory. Therefore, the non-stationarity of rewards plays an important role in the construction of the sliding window Bellman error and the confidence set. Specifically, we replace the sliding window squared Bellman error (1) with

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

where r_h^t is the reward obtained at step h in episode t. Moreover, the local regression constraint is

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

where β is a confidence parameter, Δ_P^w is the local variation budget defined in (2) and Δ_R^w is defined as

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

Algorithm 2 SW-OPEA (bandit feedback)

Input: $\mathcal{D}_1, \cdots, \mathcal{D}_H \leftarrow \varnothing$, $\mathcal{B}^0 \leftarrow \mathcal{F}$.
for episode k from 1 to K **do**
  Choose $\pi^k = \pi_{f^k}$,
  where $f^k = \arg\max_{f \in \mathcal{B}^{k-1}} f_1(x_1, \pi_f(x_1))$.
  Collect a trajectory $(x_1^k, a_1^k, \cdots, x_H^k, a_H^k, x_{H+1}^k)$ by following π^k and reward function $\{r_h^k\}_{h\in[H]}$.
  Augment $\mathcal{D}_h = \mathcal{D}_h \cup \{(x_h^k, a_h^k, x_{h+1}^k)\}$, $\forall h \in [H]$.
  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 $\mathcal{L}_{\mathcal{D}_h}(\xi_h, \zeta_{h+1}) = \sum_{t=1\vee(k-w)}^k \left(\xi_h(x_h^t, a_h^t) - r_h^t - \max_{a'\in\mathcal{A}} \zeta_{h+1}(x_{h+1}^t, a')\right)^2$
end for

D.1. Algorithm and Theorem

Theorem D.1. *Under Assumption 2.1 and Assumption 5.1, 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 $k \ge \min\{w+1, \dim_{\mathrm{DBE}}(\mathcal{F}, \mathcal{D}_{\Delta,h}, \sqrt{1/w})\}$ we have*

$$\begin{aligned}\mathrm{D-Regret}(k) =& \Delta_R(k) + H\Delta_P(k) \\ &+ \mathcal{O}\left(H\sqrt{w} + \frac{H^2k}{\sqrt{w}}\sqrt{d\log[KH|\mathcal{G}|/\delta]} + \frac{H^2k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_P^w(t,h)} + \frac{H^{3/2}k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_R^w(t,h)}\right).\end{aligned}$$

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

D.2. Proof of Theorem 5.6

Following the same argument in Appendix C gives

$$\mathrm{D-Regret}(k) = H + \Delta_R(k) + H\Delta_P(k) + \underbrace{\sum_{t=1}^k \left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t}\right)(x_1)}_{\text{(I)}}.$$

In the sequel, we strive to bound term (I). We first introduce a different probability distribution shift lemma. Compared to Lemma 5.4, the new lemma is more general and can handle the bandit feedback scenario.

Lemma D.2. *Suppose P and Q are two probability distributions of a random variable x, then*

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

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

Proof. Note that

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

□

Thanks to Lemma D.2, we are able to obtain the following two lemmas.

Lemma D.3. *If $\beta = cH^2 \log \frac{KH|\mathcal{G}|}{\delta}$, then with probability at least $1-\delta$, we have $Q^*_{(*,k)} \in \mathcal{B}^k$ for all $k \in [K]$.*

Proof. Define

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

Fix a tuple $(k, h, g) \in [K] \times [H] \times \mathcal{G}$. Let

$$\begin{aligned} W_t(h, f) :&= \left[g_h(x_h^t, a_h^t) - r_h^t - \max_{a' \in \mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t, a')) \right]^2 - \left[\#_{k,h}(x_h^t, a_h^t) - r_h^t - \max_{a' \in \mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t, a')) \right]^2 \\ &= [g_h(x_h^t, a_h^t) - \#_{k,h}(x_h^t, a_h^t)] \left[g_h(x_h^t, a_h^t) + \#_{k,h}(x_h^t, a_h^t) - 2 \left(r_h^t + \max_{a' \in \mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t, a') \right) \right] \end{aligned}$$

and $\mathcal{F}_{t,h}$ be the filtration induced by $\{x_1^i, a_1^i, \cdots, x_H^i\}_{i \in [t-1]} \cup \{x_1^t, a_1^t, \cdots, x_h^t, a_h^t\} \cup \{r_h^i\}_{h \in [H]}^{i \in [t-1]}$. We have

$$\begin{aligned} &\mathbb{E}[W_t(h, g) | \mathcal{F}_{t,h}] = \left[(g_h - \#_{k,h})(x_h^t, a_h^t) \right]^2, \\ &\mathrm{Var}[W_t(h, g) | \mathcal{F}_{t,h}] \le 36 H^2 \mathbb{E}[W_t(h, g) | \mathcal{F}_{t,h}]. \end{aligned}$$

By Freedman's inequality, with probability at least $1-\delta$,

$$\begin{aligned} &\left| \sum_{t=1 \vee (k-w)}^{k} W_t(h, g) - \sum_{t=1 \vee (k-w)}^{k} \left[\left(g_h(x_h^t, a_h^t) - \#_{k,h} \right)(x_h^t, a_h^t) \right]^2 \right| \\ &\qquad \le \mathcal{O}\left(H \sqrt{\log(1/\delta) \sum_{t=1 \vee (k-w)}^{k} \left[(g_h(x_h^t, a_h^t) - \#_{k,h})(x_h^t, a_h^t) \right]^2} + \log(1/\delta) \right). \end{aligned}$$

Taking union bound over $[K] \times [H] \times \mathcal{G}$,

$$\begin{aligned} &\left| \sum_{t=1 \vee (k-w)}^{k} W_t(h, g) - \sum_{t=1 \vee (k-w)}^{k} \left[\left(g_h(x_h^t, a_h^t) - \#_{k,h} \right)(x_h^t, a_h^t) \right]^2 \right| \\ &\qquad \le \mathcal{O}\left(H \sqrt{\iota \sum_{t=1 \vee (k-w)}^{k} \left[(g_h(x_h^t, a_h^t) - \#_{k,h})(x_h^t, a_h^t) \right]^2} + \iota \right), \end{aligned}$$

where $\iota = \log(HK|\mathcal{G}|/\delta)$. We have

$$\begin{aligned} &- \sum_{t=1 \vee (k-w)}^{k} W_t(h, g) \\ &\le - \sum_{t=1 \vee (k-w)}^{k} \left[\left(g_h(x_h^t, a_h^t) - \#_{k,h} \right)(x_h^t, a_h^t) \right]^2 + \mathcal{O}\left(H \sqrt{\iota \sum_{t=1 \vee (k-w)}^{k} \left[(g_h(x_h^t, a_h^t) - \#_{k,h})(x_h^t, a_h^t) \right]^2} + \iota \right) \\ &\le \mathcal{O}(H^2 \iota). \end{aligned}$$

I.e.,

$$\begin{aligned} &\sum_{t=1 \vee (k-w)}^{k} \left[\#_{k,h}(x_h^t, a_h^t) - r_h^t - \max_{a' \in \mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t, a') \right]^2 \\ &\qquad \le \sum_{t=1 \vee (k-w)}^{k} \left[g_h(x_h^t, a_h^t) - r_h^t - \max_{a' \in \mathcal{A}} Q_{h+1;(*,k)}(x_{h+1}^t, a') \right]^2 + \mathcal{O}(H^2 \iota). \end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{t=1\vee(k-w)}^{k}\left[Q_{h;(*,k)}(x_h^t,a_h^t)-r_h^t-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2\\
&\le\sum_{t=1\vee(k-w)}^{k}\left[\#_{k,h}(x_h^t,a_h^t)-r_h^t-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2+2H^2\Delta_P^w(k,h)+2H\Delta_R^w(k,h)\\
&\le\sum_{t=1\vee(k-w)}^{k}\left[g_h(x_h^t,a_h^t)-r_h^t-\max_{a'\in\mathcal{A}}Q_{h+1;(*,k)}(x_{h+1}^t,a')\right]^2+2H^2\Delta_P^w(k,h)+2H\Delta_R^w(k,h)+\mathcal{O}(H^2\iota),
\end{aligned}$$

where the first inequality follows from Lemma D.2 and the definition of Δ_P^w and Δ_P^w. By the definition of $\mathcal{B}^k$ and $\beta=cH^2\log\frac{KH|\mathcal{G}|}{\delta}$ with some large absolute constant c, we conclude that with probability at least $1-\delta$, $Q_{(*,k)}\in\mathcal{B}^k$ for all $k\in[K]$. □

Lemma D.4. *If $\beta=cH^2\log\frac{KH|\mathcal{G}|}{\delta}$, then with probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, we have*

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

Proof. Define

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

Fix a tuple $(k,h,f)\in[K]\times[H]\times\mathcal{G}$. Let

$$\begin{aligned}
W_t(h,f):&=\left[f_h(x_h^t,a_h^t)-r_h^t-\max_{a'\in\mathcal{A}}f_{h+1}(x_{h+1}^t,a')\right]^2-\left[\#_{k,h}^f(x_h^t,a_h^t)-r_h^t-\max_{a'\in\mathcal{A}}f_{h+1}(x_{h+1}^t,a')\right]^2\\
&=[f_h(x_h^t,a_h^t)-\#_{k,h}^f(x_h^t,a_h^t)]\left[f_h(x_h^t,a_h^t)+\#_{k,h}^f(x_h^t,a_h^t)-2\left(r_h^t+\max_{a'\in\mathcal{A}}f_{h+1}(x_{h+1}^t,a')\right)\right]
\end{aligned}$$

and $\mathcal{F}_{t,h}$ be the filtration induced by $\{x_1^i,a_1^i,\cdots,x_H^i\}_{i\in[t-1]}\cup\{x_1^t,a_1^t,\cdots,x_h^t,a_h^t\}\cup\{r_h^i\}_{h\in[H]}^{i\in[t-1]}$. We have

$$\begin{aligned}
&\mathbb{E}[W_t(h,f)|\mathcal{F}_{t,h}]=\left[\left(f_h-\#_{k,h}^f\right)(x_h^t,a_h^t)\right]^2,\\
&\mathrm{Var}[W_t(h,f)|\mathcal{F}_{t,h}]\le 36H^2\mathbb{E}[W_t(h,g)|\mathcal{F}_{t,h}].
\end{aligned}$$

By Freedman's inequality, we have

$$\begin{aligned}
&\left|\sum_{t=1\vee(k-w)}^{k}W_t(h,f)-\sum_{t=1\vee(k-w)}^{k}\left[(f_h-\#_{k,h}^f)(x_h^t,a_h^t)\right]^2\right|\\
&\qquad\le\mathcal{O}\left(H\sqrt{\log(1/\delta)\sum_{t=1\vee(k-w)}^{k}\left[(f_h-\#_{k,h}^f)(x_h^t,a_h^t)\right]^2}+\log(1/\delta)\right).
\end{aligned}$$

Taking union bound over $[K]\times[H]\times\mathcal{G}$, we have

$$\left|\sum_{t=1\vee(k-w)}^{k}W_t(h,g)-\sum_{t=1}^{k}\left[(f_h-\#_{k,h}^f)(x_h^t,a_h^t)\right]^2\right|\le\mathcal{O}\left(H\sqrt{\iota\sum_{t=1\vee(k-w)}^{k}\left[(f_h-\#_{k,h}^f)(x_h^t,a_h^t)\right]^2}+\iota\right),$$

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

Note that

$$
\begin{aligned}
&\sum_{t=1\vee(k-w-1)}^{k-1} W_t(h, f^k) \\
&= \sum_{t=1\vee(k-w-1)}^{k-1} \left[f_h^k(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 \\
&\qquad - \sum_{t=1\vee(k-w-1)}^{k-1} \left[\#_{k-1,h}^{f^k}(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 \\
&\le \sum_{t=1\vee(k-w-1)}^{k-1} \left[f_h^k(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 \\
&\quad - \sum_{t=1\vee(k-w-1)}^{k-1} \left[\mathcal{T}_h^{k-1} f_{h+1}^k(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 + 2H^2\Delta_P^w(k-1,h) + 2H\Delta_R^w(k-1,w) \\
&\le \sum_{t=1\vee(k-w-1)}^{k-1} \left[f_h^k(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 \\
&\qquad - \inf_{g\in\mathcal{G}} \sum_{t=1\vee(k-w-1)}^{k-1} \left[g_h(x_h^t, a_h^t) - r_h^{t-1} - \max_{a'\in\mathcal{A}} f_{h+1}^k(x_{h+1}^t, a') \right]^2 + 2H^2\Delta_P^w(k-1,h) + 2H\Delta_R^w(k-1,w) \\
&\le \beta + 4H^2\Delta_P^w(k-1,h) + 4H\Delta_R^w(k-1,w),
\end{aligned}
$$

where the first inequality follows from Lemma D.2 and the definition of Δ_P^w and Δ_P^w, the second inequality follows from Assumption 5.1, and the last inequality follows from the definition of $\mathcal{B}^{k-1}$.

Therefore,

$$
\sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \#_{k-1,h}^{f^k})(x_h^t, a_h^t) \right]^2 \le \beta + 4H^2\Delta_P^w(k-1,h) + 4H\Delta_R^w(k-1,w) + \mathcal{O}\left(H^2\iota\right).
$$

Finally, we use Lemma D.2 again and obtain

$$
\begin{aligned}
&\sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \mathcal{T}_h^{k-1} f_{h+1}^k)(x_h^t, a_h^t) \right]^2 \\
&\le \sum_{t=1\vee(k-w-1)}^{k-1} \left[(f_h^k - \#_{k-1,h}^{f^k})(x_h^t, a_h^t) \right]^2 + 2H^2\Delta_P^w(k-1,h) + 2H\Delta_R^w(k-1,w) \\
&\le 6H^2\Delta_P^w(k-1,h) + 6H\Delta_R^w(k-1,w) + \mathcal{O}(\beta).
\end{aligned}
$$

□

By Lemma D.3, with probability at least $1-\delta$, we have

$$
\begin{aligned}
\text{(I)} &= \sum_{t=1}^{k} \left(V_{1;(*,t-1)}^{\pi^{(*,t-1)}} - V_{1;(*,t-1)}^{\pi^t} \right)(x_1) \\
&\le \sum_{t=1}^{k} \left(\max_{a\in\mathcal{A}} f_1^t(x_1, a) - V_{1;(*,t-1)}^{\pi^t}(x_1) \right)
\end{aligned}
$$

$$\leq \sum_{h=1}^{H}\sum_{t=1}^{k} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)],$$

where the first inequality follows from Lemma D.3 and the optimistic planning step (line 3) in Algorithm 2 which guarantees that $V^*_{1;(*,k-1)} \leq \sup_a f_1^k(x_1,a)$ for every episode k, the last inequality follows from generalized policy loss decomposition (Lemma C.8) and the fact that $\pi^k = \pi_{f^k}$ (line 3 in Algorithm 2).

Now we invoke Lemma 5.5 and Lemma D.4 with

$$\begin{aligned}
&\theta = \sqrt{\frac{1}{w}}, C = H,\\
&\mathcal{X} = \mathcal{S}\times\mathcal{A}, \Phi = (I-\mathcal{T}_h)\mathcal{F}, \text{and } \Pi = \mathcal{D}_{\Delta,h},\\
&\phi_k = f_h^k - \mathcal{T}_h^{k-1} f_{h+1}^k, \mu_k = \mathbf{1}\{\cdot = (x_h^k, a_h^k)\}
\end{aligned}$$

and obtain

$$\begin{aligned}
&\sum_{t=1}^{k} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)]\\
&\leq \sum_{t=1}^{k}(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h^t,a_h^t) + \mathcal{O}\left(\sqrt{k\log(k)}\right)\\
&\leq \mathcal{O}\left(\frac{k}{w}\sqrt{w\cdot \dim_{\mathrm{DBE}}(\mathcal{F},\mathcal{D}_{\Delta,h},\sqrt{1/w})\left(H^2\log[KH|\mathcal{G}|/\delta] + H^2\sup_{t\in[k]}\Delta_P^w(t,h) + H\sup_{t\in[k]}\Delta_R^w(t,h)\right)} + \sqrt{w}\right)\\
&\leq \mathcal{O}\left(\frac{Hk}{\sqrt{w}}\sqrt{d\log[kH|\mathcal{G}|/\delta]} + \frac{Hk}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_P^w(t,h)} + \frac{\sqrt{H}k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_R^w(t,h)} + \sqrt{w}\right),
\end{aligned}$$

where the second inequality follows from Azuma-Hoeffding inequality, and in the last inequality, we use $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for any positive $a,b\geq 0$ and we define $d = \dim_{\mathrm{DBE}}(\mathcal{F},\mathcal{D}_{\Delta,h},\sqrt{1/w})$.

Summing over step $h\in[H]$ gives

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{k} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)]\\
&\leq \mathcal{O}\left(\frac{H^2k}{\sqrt{w}}\sqrt{d\log[KH|\mathcal{G}|/\delta]} + \frac{H^2k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_P^w(t,h)} + \frac{H^{3/2}k}{\sqrt{w}}\sqrt{d\sup_{t\in[k]}\Delta_R^w(t,h)} + H\sqrt{w}\right),
\end{aligned}$$

which completes the proof.

D.3. Proof of Corollary 5.7

For ease of exposition, let $d = \dim_{\mathrm{DBE}}(\mathcal{F},\mathcal{D}_{\Delta,h},\sqrt{1/w})$. We adopt average variation L defined in (5) and average variation L in rewards defined in (10). Then we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{K} \mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t - \mathcal{T}_h^{t-1} f_{h+1}^t)(x_h,a_h)]\\
&\leq \widetilde{\mathcal{O}}\left(\frac{H^2K}{\sqrt{w}}\sqrt{d}\sqrt{\log|\mathcal{G}|} + \frac{H^2K}{\sqrt{w}}\sqrt{dLw^2} + \frac{H^{\frac{3}{2}}K}{\sqrt{w}}\sqrt{dL_\theta w^2} + H\sqrt{w}\right)\\
&\leq \widetilde{\mathcal{O}}\left(H^2K\sqrt{d}\left(\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{w}} + (\sqrt{L} + \frac{\sqrt{L_\theta}}{\sqrt{H}} + \frac{1}{HK\sqrt{d}})\sqrt{w}\right)\right).
\end{aligned}$$

Note first that $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}}+\frac{1}{HK\sqrt{d}}} > 1$ when $|\mathcal{G}| > 10$.

If $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}}+\frac{1}{HK\sqrt{d}}} \geq K$, i.e., $\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}} \leq \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|}-\frac{1}{H\sqrt{d}}\right)$, we select $w = K$ and we have

$$\sum_{h=1}^{H}\sum_{t=1}^{K}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t-\mathcal{T}_h^{t-1}f_{h+1}^t)(x_h,a_h)] \leq \widetilde{O}\left(H^2K^{\frac{1}{2}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{2}}\right).$$

If $\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}}+\frac{1}{HK\sqrt{d}}} < K$, i.e., $\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}} > \frac{1}{K}\left(\sqrt{\log|\mathcal{G}|}-\frac{1}{H\sqrt{d}}\right)$, we select $w = \lceil\frac{\sqrt{\log|\mathcal{G}|}}{\sqrt{L}+\frac{\sqrt{L_\theta}}{\sqrt{H}}+\frac{1}{HK\sqrt{d}}}\rceil$ and we have

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{t=1}^{K}\mathop{\mathbb{E}}_{(x_h,a_h)\sim(\pi^t,(*,t-1))}[(f_h^t-\mathcal{T}_h^{t-1}f_{h+1}^t)(x_h,a_h)]\\&\leq \widetilde{\mathcal{O}}\left(H^2KL^{\frac{1}{4}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{4}}+H^{\frac{7}{4}}KL_\theta^{\frac{1}{4}}d^{\frac{1}{2}}(\log|\mathcal{G}|)^{\frac{1}{4}}+H^{\frac{3}{2}}K^{\frac{1}{2}}d^{\frac{1}{4}}(\log|\mathcal{G}|)^{\frac{1}{4}}\right).\end{aligned}$$