Horizon-Free and Instance-Dependent Regret Bounds for Reinforcement Learning with General Function Approximation

Jiayi Huang[1]   **Han Zhong**[1]   **Liwei Wang**[1,2]   **Lin F. Yang**[3]

[1]Center for Data Science, Peking University
[2]National Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[3]University of California, Los Angeles

Abstract

To tackle long planning horizon problems in reinforcement learning with general function approximation, we propose the first algorithm, termed as UCRL-WVTR, that achieves both *horizon-free* and *instance-dependent*, since it eliminates the polynomial dependency on the planning horizon. The derived regret bound is deemed *sharp*, as it matches the minimax lower bound when specialized to linear mixture MDPs up to logarithmic factors. Furthermore, UCRL-WVTR is *computationally efficient* with access to a regression oracle. The achievement of such a horizon-free, instance-dependent, and sharp regret bound hinges upon (i) novel algorithm designs: weighted value-targeted regression and a high-order moment estimator in the context of general function approximation; and (ii) fine-grained analyses: a novel concentration bound of weighted non-linear least squares and a refined analysis which leads to the tight instance-dependent bound. We also conduct comprehensive experiments to corroborate our theoretical findings.

1 INTRODUCTION

Reinforcement Learning (RL) plays a pivotal role in solving complex decision-making problems, where an agent interacts with the environment to learn a policy that maximizes cumulative rewards. In the context of time-homogeneous episodic RL, where the total re-wards are bounded by 1, an open question arises concerning whether episodic RL is more difficult than bandits problems (Jiang and Agarwal, 2018). While numerous research endeavors have been dedicated to the development of algorithms with up to poly-logarithmic dependence on the planning horizon, thus are *horizon-free*, these efforts are restricted to either tabular (Wang et al., 2020a; Zhang et al., 2021; Li et al., 2022; Zhang et al., 2022) or linear mixture Markov Decision Processes (MDPs) (Zhang et al., 2021b; Kim et al., 2022; Zhou and Gu, 2022; Zhao et al., 2023). Consequently, this open question persists as a formidable challenge, necessitating further exploration and innovation to address a broader family of RL problems.

The difficulty of solving RL problems is intrinsically related to complexity of the underlying model. For instance, the performance of an RL algorithm typically improves when the stochasticity in unknown transition kernel is diminished. This observation motivated many works to design algorithms that achieve *instance-dependent* regret bounds (Zanette and Brunskill, 2019; Zhou et al., 2023; Zhao et al., 2023). These bounds depend on fine-grained, problem-specific quantities, thus providing tighter guarantees than worst-case regret bounds. Nevertheless, existing research for RL with general function approximation has primarily focused on minimizing regret in the worst-case scenario (Ayoub et al., 2020; Foster et al., 2023; Agarwal et al., 2023), thereby leaving an instance-dependent guarantee still challenging in this domain.

Thus, a natural question arises:

Is efficient, horizon-free and instance-dependent learning possible in RL with general function approximation?

In this work, we give an affirmative answer to this question by proposing an algorithm, termed as UCRL-WVTR, for RL with general function approximation. UCRL-WVTR enjoys an $\widetilde{O}(\sqrt{\dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}}} \cdot$

$\sqrt{\mathcal{Q}^*} + \dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}})$ regret, where $\dim_{\mathcal{F}}$ and $\log \mathcal{N}_{\mathcal{F}}$ respectively denote the generalized Eluder dimension and log-covering number of the function class $\mathcal{F}$, and $\mathcal{Q}^*$ is an instance-dependent quantity defined in Theorem 4.2. The generalized Eluder dimension is an extension of Eluder dimension (Russo and Van Roy, 2013) tailored to the context of weighted regression settings. The derived instance-dependent regret bound exhibits a log-polynomial dependence on the planning horizon H. Therefore, we simultaneously achieve the best of both worlds: horizon-free and instance-dependent. For more detailed insights and explanations, we refer readers to Section 4. The main theoretical contributions of this paper are summarized as follows:

- Our algorithm utilizes the Value-Targeted Regression (VTR) framework proposed by Ayoub et al. (2020), which was devised for estimating the unknown transition kernel. Our refinement of this framework includes a weighted design, which assigns higher importance to data points with low variance and uncertainty. Additionally, we adopt a high-order moment estimator to achieve more accurate estimations of variances. This improved algorithmic design plays a pivotal role in producing an accurate model estimate. Based on the estimated model, we further implement an efficient planning approach via a regression oracle for general function approximation, thus achieving computational efficiency.

- To theoretically characterize the deviation of the estimated model, we propose a novel Bernstein-style concentration bound for weighted non-linear regression and rigorously establish its tightness. This result provides insights for assigning weights to be both variance-aware and uncertainty-aware in our algorithm. Despite the inherent complexities posed by the non-linear function class, we have successfully conducted a refined analysis, resulting in a regret bound that simultaneously achieves horizon-free and instance-dependent.

- As a special case, our algorithm achieves a first-order regret scaling as $\widetilde{O}(d\sqrt{V_1^* K} + d^2)$ for linear mixture MDPs, where the transition kernel admits a linear combination of some basis transition models. This regret bound is sharp since it is reduced to $\widetilde{O}(d\sqrt{K} + d^2)$ in the worst case, matching the lower bound $\Omega(d\sqrt{K})$ (Zhou et al., 2021) when $K \geq d^2$ up to logarithmic factors. Considering this emerges as a direct consequence of a broader result, it shows that our novel algorithm designs and fine-grained analyses indeed handle the general RL problems in a sharp manner.

Road Map The rest of this paper is organized as follows. Section 2 provides works that are closely related to ours. Section 3 introduces the formal definition of RL with general function approximation. Section 4 states our main theoretical results. Section 5 presents the experimental results that corroborate our theoretical findings. We then make conclusions in Section 6. Additional experiments and detailed theoretical analyses are left in the Appendix.

Notations Let $[n] := \{1, 2, \ldots, n\}$. Let $\overline{[n]} := \{0, 1, \ldots, n\}$. For a collection of elements $\{X_t\}_{t \geq 1}$, let $X_{[t]} := \{X_1, \ldots, X_t\}$. Denote the ϵ-covering number of $\mathcal{F}$ w.r.t. ℓ_∞-norm as $\mathcal{N}_{\mathcal{F}}(\epsilon)$.

2 RELATED WORK

Horizon-Free Regret in RL Jiang and Agarwal (2018) raised an open question regarding the comparative difficulty of RL in comparison to bandit problems. Specifically, in the context of time-homogeneous tabular RL and assuming an upper bound of 1 on total rewards, they posited that any algorithm seeking to find an ϵ-optimal policy would need to exhibit a polynomial dependence on the planning horizon H in the sample complexity. This conjecture has been challenged by a line of works that have introduced algorithms capable of achieving horizon-free regret bounds (Wang et al., 2020a; Zhang et al., 2021a; Li et al., 2022; Zhang et al., 2022). Additionally, a series recent studies (Zhang et al., 2021b; Kim et al., 2022; Zhou and Gu, 2022; Zhao et al., 2023) further extended horizon-free learning to linear mixture MDPs. It is noteworthy that these aforementioned studies are based on the assumption of either a finite state space or the linear representation of transition probabilities. Regrettably, such assumptions are often less practical to align with real-world scenarios.

Instance-Dependent Regret in RL In recent years, there has been a substantial body of research dedicated to the developingment of algorithms with instance-dependent regret bounds, therefore providing tighter guarantee than traditional worst-case regret (Zanette and Brunskill, 2019; Zhou et al., 2023; Wagenmaker et al., 2022; Li and Sun, 2023; Huang et al., 2023; Zhao et al., 2023). Notably, Zhao et al. (2023) introduced a variance-adaptive algorithm with a horizon-free and variance-aware regret bound, but constrained to linear mixture MDPs. To our best knowledge, hardly few works have explored the concept of instance-dependent regret in RL with general function approximation. The exception to this is Wagenmaker and Foster (2023), which developed a non-asymptotic theory of instance-optimal RL with general

function approximation. However, despite theoretical advancements, their algorithm falls short of practicality since it deviates significantly from minimax optimality and computational efficiency.

RL with General Function Approximation We study model-based RL with general function approximation, where the agent learns and employs an explicit model of the environment for planning and decision-making. This area has witnessed a substantial surge of research (Osband and Van Roy, 2014; Sun et al., 2019; Ayoub et al., 2020; Wang et al., 2020b; Foster et al., 2021; Chen et al., 2022b,a; Zhong et al., 2022; Foster et al., 2023; Wagenmaker and Foster, 2023). Notably, Ayoub et al. (2020) stand as the closest precursor to our work. They adopted a novel VTR approach, which evaluates models based on their ability to predict values at the next states. This VTR framework was subsequently extended and refined, culminating in the attainment of horizon-free regrets when applied to linear mixture MDPs (Zhou and Gu, 2022). Concurrently, there exists a separate line of works concentrated on model-free RL (Jiang et al., 2017; Wang et al., 2020b; Du et al., 2021; Jin et al., 2021; Kong et al., 2021; Dann et al., 2021; Zhong et al., 2022; Liu et al., 2023; Agarwal et al., 2023), where the agent directly learns policies or value functions for decision-making based on their interactions with the environment. However, it is worth noting that the algorithms mentioned above either suffer from polynomial dependence on the planning horizon H or fail to provide a fine-grained instance-dependent regret guarantee. Our work achieves the best of both worlds for the first time in RL with general function approximation.

3 PRELIMINARIES

MDPs with General Function Approximation We study time-homogeneous episodic MDPs, which can be described by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \{r_h\}_{h \in [H]})$. Here, $\mathcal{S}$ and $\mathcal{A}$ are state space and action space, respectively, H is the length of planning horizon, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamics, $r_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the h-th step deterministic reward function known to the agents[1]. We consider the bounded reward setting that $\sum_{h=1}^{H} r_h(s_h, a_h) \leq 1$ for any trajectory $\{s_1, a_1, \ldots, s_H, a_H\}$. We consider deterministic policy throughout this paper. A deterministic policy $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h \in [H]}$ is a collection of H mappings from state space to action space. For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the action value function $Q_h^\pi(s, a)$ and (state) value function

$V_h^\pi(s)$ as follows:

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

where the expectation is taken with respect to the transition kernel $\mathbb{P}$ and the agent's policy π. Denote $V_h^*(s) := \sup_\pi V_h^\pi(s)$ and $Q_h^*(s, a) := \sup_\pi Q_h^\pi(s, a)$ as the optimal value functions. We introduce the following shorthands for simplicity. At the h-th step, for any value function $V : \mathcal{S} \to \mathbb{R}$, denote

$$[\mathbb{P}V](s, a) := \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, a)}[V(s')],$$
$$[\mathbb{V}V](s, a) := [\mathbb{P}V^2](s, a) - [\mathbb{P}V]^2(s, a)$$

as the conditional expectation and variance of V, respectively. We aim to design efficient algorithms for minimizing the K-episode regret defined as

$$\text{Regret}(K) := \sum_{k=1}^{K}[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)].$$

To tackle problems with large state spaces, we consider MDPs with general function approximation.

Assumption 3.1 (MDPs with general function approximation). Let $\mathcal{P}$ be a general function class composed of transition kernels mapping state-action pairs to measures over $\mathcal{S}$. We assume the transition model $\mathbb{P}$ of the MDP satisfies $\mathbb{P} \in \mathcal{P}$.

To characterize the complexity of the model class $\mathcal{P}$, we further introduce the function class

$$\mathcal{F} := \{f : \mathcal{S} \times \mathcal{A} \times \mathcal{V} \to \mathbb{R} \text{ such that } \exists \mathbb{P} \in \mathcal{P},$$
$$f(s, a, V) = [\mathbb{P}V](s, a) \text{ for any } V \in \mathcal{V}\},$$

where $\mathcal{V}$ encompasses all value functions V such that $V : \mathcal{S} \to [0, 1]$. As is demonstrated by Ayoub et al. (2020), a bijection denoted as $\phi : \mathcal{P} \to \mathcal{F}$ exists, establishing a direct correspondence between $\mathcal{P}$ and $\mathcal{F}$, such that for any $\mathbb{P} \in \mathcal{P}$, there exists a corresponding $f = \phi(\mathbb{P}) \in \mathcal{F}$ with $f(s, a, V) = [\mathbb{P}V](s, a)$. For brevity, we denote $f_\mathbb{P} := \phi(\mathbb{P})$. Additionally, we assume for any $(s, a, V) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V}$, $f_\mathbb{P}(s, a, V) = \sum_{s' \in \mathcal{S}} V(s')\mathbb{P}(s'|s, a)$ can be efficiently evaluated within $\mathcal{O}$ time. Such an assumption depends on the intrinsic property of the model class and it holds true for many MDP models. See Ayoub et al. (2020) for more discussions. We use the covering number and generalized Eluder dimension introduced later to gauge the complexity of $\mathcal{F}$.

Generalized Eluder Dimension Inspired by Agarwal et al. (2023), we use generalized Eluder dimension as a complexity measurement for the generalization property of the function class $\mathcal{F}$. Generalized

---

[1]Our result can be generalized to the unknown-reward cases since learning transition dynamics is more challenging than learning rewards.

Eluder dimension is an extension of the Eluder dimension defined in Russo and Van Roy (2013) to weighted regression settings. Intuitively, it can be thought of Eluder dimension of a scaling-enlarged function set.

Definition 3.2 (Generalized Eluder dimension)**.** Let $\lambda > 0$, a sequence of random elements $\boldsymbol{X} = \{X_t\}_{t \in [T]}$ and $\boldsymbol{\sigma} = \{\sigma_t\}_{t \in [T]}$ be given. The *generalized Eluder dimension* of a function class $\mathcal{F}$ consisting of function $f : \mathcal{X} \to \mathbb{R}$ is given by

$$\dim_{\mathcal{F}}(\sigma_{\min}, T)$$
$$:= \sup_{\boldsymbol{X}, \boldsymbol{\sigma}} \sum_{t=1}^{T} \min\left(1, \frac{1}{\sigma_t^2} D_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})\right),$$
$$\text{s.t. } |\boldsymbol{X}| = T, \boldsymbol{\sigma} \geq \sigma_{\min},$$

where uncertainty $D_{\mathcal{F}}$ is defined by

$$D_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})$$
$$:= \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \lambda}.$$

We also use the notation $\dim_{\mathcal{F}} := \dim_{\mathcal{F}}(\sigma_{\min}, T)$ when $\sigma_{\min}$ and T are clear from the context.

Remark 3.3. We remark that the notion of generalized Eluder dimension is initially introduced by Agarwal et al. (2023), with a primary focus on time-inhomogeneous MDPs under model-free settings. Furthermore, they require value closedness[2] (Wang et al., 2020b; Kong et al., 2021). In contrast, we study time-homogenous MDPs under model-based settings, where we do not need completeness-type assumptions. More importantly, our work endeavors to pursue horizon-free learning in the context of general functions, whereas their work only obtains a near-optimal regret bound with a polynomial dependency on the planning horizon H.

Linear Mixture MDPs We also consider linear mixture MDPs (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020), which is a special case where the transition kernel admits a linear representation.

Assumption 3.4 (Linear mixture MDPs)**.** There exists an unknown vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and a known feature $\boldsymbol{\phi}(\cdot | \cdot, \cdot) \in \mathbb{R}^d$ such that

$$\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$$

for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Meanwhile, we assume that $\|\boldsymbol{\theta}^*\|_2 \leq B$ and $\|\boldsymbol{\phi}_V(s, a)\|_2 \leq 1$ for any bounded value function $V : \mathcal{S} \to [0, 1]$ and any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, where

$$\boldsymbol{\phi}_V(s, a) := \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s, a) V(s').$$

---

[2]The Bellman projection of any value function lies in the target function class.

Algorithm 1 UCRL-WVTR

Require: λ, confidence radius $\{\widehat{\beta}_k\}_{k \geq 1}$, level M, variance parameters $\sigma_{\min}, \gamma$
1: For $m \in \overline{[M]}$, $\widehat{f}_{1,m} \leftarrow 0$, $D_{1,m}, D_{1,0,m} \leftarrow \varnothing$
2: **for** $k = 1, \ldots, K$ **do**
3:     $V_{k,H+1}(\cdot) = 0$
4:     **for** $h = H, \ldots, 1$ **do**
5:         $Q_{k,h}(\cdot, \cdot) \leftarrow \min\{r_h(\cdot, \cdot) + \widehat{f}_{k,0}(\cdot, \cdot, V_{k,h+1}) + \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(\cdot, \cdot, V_{k,h+1}; D_{k,0}), 1\}$
6:         $V_{k,h}(\cdot) = \max_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$,
        $\pi_h^k(\cdot) = \arg\max_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$
7:     **end for**
8:     Receive s_1^k
9:     **for** $h = 1, \ldots, H$ **do**
10:         Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive s_{h+1}^k
11:         For $m \in \overline{[M]}$, $z_{k,h,m} \leftarrow (s_h^k, a_h^k, V_{k,h+1}^{2^m})$,
        $y_{k,h,m} \leftarrow V_{k,h+1}^{2^m}(s_{h+1}^k)$
12:         $\{\bar{\sigma}_{k,h,m}\}_{m \in \overline{[M]}} \leftarrow \text{HOME}(\{z_{k,h,m}, \widehat{f}_{k,m}, D_{k,h-1,m}, D_{k,m}\}_{m \in \overline{[M]}}, \widehat{\beta}_k, \sigma_{\min}, \gamma)$
13:         For $m \in \overline{[M]}$, $D_{k,h,m} \leftarrow D_{k,h-1,m} \cup \{z_{k,h,m}, \bar{\sigma}_{k,h,m}\}$
14:     **end for**
15:     For $m \in \overline{[M]}$, $D_{k+1,m}, D_{k+1,0,m} \leftarrow D_{k,H,m}$
16:     For $m \in \overline{[M]}$, $\widehat{f}_{k+1,m} \leftarrow \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{k} \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{i,h}^2}(f(z_{i,h,m}) - y_{i,h,m})^2$
17: **end for**

The generalized Eluder dimension and log-covering number of linear mixture MDPs can be simplified.

Proposition 3.5. For d-dimensional, B-bounded linear mixture MDPs defined in Assumption 3.4, the generalized Eluder dimension and log-covering number satisfies

$$\dim_{\mathcal{F}}(\sigma_{\min}, T) = O(d \log(1 + TB/(d\lambda\sigma_{\min}^2))),$$
$$\log \mathcal{N}_{\mathcal{F}}(\epsilon) = \Theta(d \log(B/\epsilon)).$$

Proof. See Appendix D.1 for a detailed proof. $\square$

4 MAIN RESULTS

In this section, we propose a new algorithm UCRL-WVTR for MDPs with general function approximation, as detailed in Algorithm 1. We first give the high-level idea, then analyze the computational complexity and regret bound.

4.1 Algorithm Description

UCRL-WVTR capitalizes on the VTR methodology, which draws inspiration from UCRL-VTR presented

---

Algorithm 2 HOME

Require: $\{z_{k,h,m}, \widehat{f}_{k,m}, D_{k,h-1,m}, D_{k,m}\}_{m \in \overline{[M]}}$, $\quad \widehat{\beta}_k$,
$\quad \sigma_{\min}, \gamma$

Ensure: $\{\bar{\sigma}_{k,h,m}\}_{m \in \overline{[M]}}$

1: **for** $m = 0, \ldots, M-1$ **do**
2: $\quad [\bar{\mathbb{V}}_{k,m} V_{k,h+1}^{2^m}](s_h^k, a_h^k) \quad \leftarrow \quad \widehat{f}_{k,m+1}(z_{k,h,m+1}) - \widehat{f}_{k,m}^2(z_{k,h,m})$
3: $\quad E_{k,h,m} \leftarrow \min\{1, 2\widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m+1}; D_{k,m+1})\} + \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\}$
4: $\quad \bar{\sigma}_{k,h,m}^2 \quad \leftarrow \quad \max\{[\bar{\mathbb{V}}_{k,m} V_{k,h+1}^{2^m}](s_h^k, a_h^k) \quad + E_{k,h,m}, \sigma_{\min}^2, \gamma^2 \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,h-1,m})\}$
5: **end for**
6: $\bar{\sigma}_{k,h,M}^2 \leftarrow \max\{1, \sigma_{\min}^2, \gamma^2 \mathcal{D}_{\mathcal{F}}(z_{k,h,M}; D_{k,h-1,M})\}$

---

by Ayoub et al. (2020) on estimating transition dynamics. Furthermore, we employ *weighted* least squares regression to mitigate sub-optimality resulting from heterogeneous noise levels, specifically, differences in variances at each stage. Additionally, we utilize a high-order moment estimator for accurately estimating variances of the next-state value functions motivated by Zhang et al. (2021b); Zhou and Gu (2022). We highlight the primary enhancements of UCRL-WVTR in comparison to UCRL-VTR as follows:

Weighted VTR UCRL-WVTR utilizes a weighted value-targeted regression framework, where the weights are chosen to take advantage of a novel concentration inequality, as articulated in Theorem 4.4. Under the guidance of this inequality, at the end of the k-th episode, we estimate the ground truth model $f_* = f_{\mathbb{P}}$ as the solution of weighted least squares

$$\widehat{f}_{k+1,0} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^k \sum_{h=1}^H \frac{1}{\bar{\sigma}_{i,h,0}^2} (f(z_{i,h,0}) - V_{i,h+1}(s_{h+1}^i))^2.$$

Here, $z_{i,h,0}$ is defined in Line 11 of Algorithm 1. And $\bar{\sigma}_{i,h,0}^2$ is an upper bound for both the conditional variance of $V_{i,h+1}(s_{h+1}^i)$ and the uncertainty of $z_{i,h,0}$, defined as follows:

$$\bar{\sigma}_{k,h,0}^2 = \max\{[\bar{\mathbb{V}}_{k,0} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h,0}, \sigma_{\min}^2, \gamma^2 \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,h-1,0})\},$$

where $[\bar{\mathbb{V}}_{k,0} V_{k,h+1}](s_h^k, a_h^k)$ is the estimate of $[\mathbb{V} V_{k,h+1}](s_h^k, a_h^k)$, $E_{k,h,0}$ is the error bound specified in Line 3 of Algorithm 2 that satisfies

$$[\bar{\mathbb{V}}_{k,0} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h,0} \geq [\mathbb{V} V_{k,h+1}](s_h^k, a_h^k)$$

with high probability, $\sigma_{\min}$ is a small positive constant to avoid numerical instability, γ is a positive constant to be chosen, and $\mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,h-1,0})$ in Definition 3.2 is the uncertainty of $z_{k,h,0}$. Then we discuss how to estimate the conditional variance of $V_{k,h+1}(s_{h+1}^k)$. Since

$$[\mathbb{V} V_{k,h+1}](s_h^k, a_h^k) = [\mathbb{P} V_{k,h+1}^2](s_h^k, a_h^k) - [\mathbb{P} V_{k,h+1}]^2(s_h^k, a_h^k),$$

we estimate it by $[\bar{\mathbb{V}}_{k,0} V_{k,h+1}](s_h^k, a_h^k)$ defined as

$$[\bar{\mathbb{V}}_{k,0} V_{k,h+1}](s_h^k, a_h^k) = \widehat{f}_{k,1}(z_{k,h,1}) - \widehat{f}_{k,0}^2(z_{k,h,0}),$$

where $z_{i,h,1}$ is defined in Line 11 of Algorithm 1 and $\widehat{f}_{k,1}$ will be specified later.

Efficient Planning via Point-Wise Bonus Terms
Once obtaining the estimate $\widehat{f}_{k,0}$, we construct action value functions $\{Q_{k,h}\}_{h \in [H]}$ following the Upper Confidence Bound (UCB) scheme (Azar et al., 2017). More specifically, starting from step H and proceeding to step 1, UCRL-WVTR estimates the upper bound of the optimal value function at each step in a point-wise manner, then takes actions optimistically:

$$Q_{k,h}(s, a) = r_h(s, a) + \widehat{f}_{k,0}(s, a, V_{k,h+1}) + \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(s, a, V_{k,h+1}; D_{k,0}),$$

Here $\widehat{\beta}_k = \widetilde{O}(\sqrt{\log \mathcal{N}_{\mathcal{F}}})$ according to Theorem 4.4. However, it is worth noting that in the realm of non-linear functions, uncertainty $\mathcal{D}_{\mathcal{F}}$ generally lacks an analytical expression, impeding the computational efficiency of our algorithm. To address such a challenge, we employ a technique to efficiently compute uncertainty $\mathcal{D}_{\mathcal{F}}$ via a regression oracle, as explained later. In contrast, Ayoub et al. (2020) construct value functions through optimistic planning, which is generally computationally intractable.

High-Order Moment Estimator We utilize the high-order moment estimator in Algorithm 2 to achieve enhanced precision in estimating the ground truth model f_*. As previously discussed, the estimate $\widehat{f}_{k,0}$ necessitates weights $\bar{\sigma}_{k,h,0}$, which relies on $\widehat{f}_{k,1}$ for estimating variance of $V_{k,h+1}$. And $\widehat{f}_{k,1}$ can be derived by weighted least squares with predictors $z_{i,h,1}$, weights $\bar{\sigma}_{i,h,1}^2$ and targets $V_{i,h+1}^2(s_{h+1}^i)$. Here $\bar{\sigma}_{k,h,1}$ are in a similar manner as $\bar{\sigma}_{k,h,0}$, depending on the variance of $V_{k,h+1}^2$. Recursively, we estimate variance of $V_{k,h+1}^{2^m}$, which is the conditional 2^{m+1}-th central moment of $V_{k,h+1}$ until $m = M$, where M is chosen to meet the desired level of precision. Last, variance of $V_{k,h+1}^{2^M}$ is estimated with its upper bound 1. It is worth noting that this higher-order moment estimator has been previously utilized in research dedicated to horizon-free learning. However, these works focused on either tabular (Zhang et al., 2021a, 2022) or linear mixture MDPs (Zhang et al., 2021b; Zhou and Gu, 2022), while we study general function approximation.

4.2 Computational Complexity

Due to the non-linear nature of the function class $\mathcal{F}$, closed-form solutions of weighted least squares $\widehat{f}_{k,m}$

in Line 16 and uncertainty $\mathcal{D}_\mathcal{F}$ in Definition 3.2 are not readily attainable. To assess the computational efficiency of our proposed algorithm in such a context, we turn to the concept of oracle complexity (Wang et al., 2020b; Kong et al., 2021), which measures the number of calls to some optimization oracles.

Regression Oracle We introduce the regression oracle in Assumption 4.1 for solving the weighted nonlinear least squares regression. Inspired by Kong et al. (2021), we leverage this oracle to compute uncertainty $\mathcal{D}_\mathcal{F}$ through a binary search procedure. As detailed in Appendix B, we demonstrate that this quantity can be estimated efficiently with a mere $\widetilde{O}(1)$ number of calls to the regression oracle.

Assumption 4.1 (Regression oracle)**.** We assume access to a weighted least squares *regression oracle*, which takes a function class $\mathcal{G}$ consisting of function $g : \mathcal{X} \to \mathbb{R}$ and a t-sized of weighted examples $\{(X_s, v_s, Y_s)\}_{s \in [t]} \subset \mathcal{X} \times \mathbb{R}^+ \times \mathbb{R}$ as input, and outputs the solution of weighted least squares $\widehat{g}$ within $\mathcal{R}$ time, where $\widehat{g}$ is defined as

$$\widehat{g} = \underset{g \in \mathcal{G}}{\arg\min} \sum_{s=1}^{t} v_s (g(X_s) - Y_s)^2.$$

The availability of a regression oracle is a reasonably mild assumption, which appeared in many works concerning general function approximation (Krishnamurthy et al., 2017; Foster et al., 2018; Kong et al., 2021; Agarwal et al., 2023). It is noteworthy that this regression oracle admits an analytical solution under linear mixture models (Ayoub et al., 2020). In more general scenarios, when the function class $\mathcal{F}$ is characterized as a collection of differentiable functions, such as neural networks, the implementation of the regression oracle becomes feasible and computationally efficient through the utilization of gradient-based optimization algorithms (Bubeck et al., 2015).

Computational Complexity of UCRL-WVTR Recall that $\mathcal{O}$ denotes the evaluation time of any function $f \in \mathcal{F}$ and $\mathcal{R}$ represents the computational cost of the regression oracle. We consider the computation cost of the k-th episode and the h-th step. First, it takes $\widetilde{O}(\mathcal{O} + \mathcal{R})$ time to compute the action value function $Q_{k,h}$ in Line 5 for a given state-action pair (s, a), since it requires the evaluation of $\widehat{f}_{k,0}$ and the computation of $\mathcal{D}_\mathcal{F}$. Then, to take actions based on π_h^k, UCRL-WVTR needs to compute the action value functions $Q_{k,h}$ for $|\mathcal{A}|$ actions, with each to be computed within $\widetilde{O}(\mathcal{O} + \mathcal{R})$ time. Next, $\{\bar{\sigma}_{k,h,m}\}_{m \in \overline{[M]}}$ in Line 12 can be computed within $\widetilde{O}(M(\mathcal{O} + \mathcal{R}))$ time since they require the evaluation of $\widehat{f}_{k,m}$, $\widehat{f}_{k,m+1}$ and

the computation of $\mathcal{D}_\mathcal{F}$ for each $\bar{\sigma}_{k,h,m}$. Finally, it takes $M\mathcal{R}$ time to calculate $\{\widehat{f}_{k+1,m}\}_{m \in \overline{[M]}}$ in Line 16. Therefore, the total time cost of UCRL-WVTR is $\widetilde{O}(KHM\mathcal{O} + KHM\mathcal{R} + |\mathcal{A}|KH\mathcal{O} + |\mathcal{A}|KH\mathcal{R})$.

4.3 Regret Bound

Theorem 4.2 (Regret)**.** For the episodic MDPs with general function approximation defined in Section 3, we set parameters in Algorithm 1 as follows: $M = \lceil \log_2(3KH) \rceil, \lambda = \log \mathcal{N}_\mathcal{F}, \sigma_{\min}^2 = \sqrt{\dim_\mathcal{F} \log \mathcal{N}_\mathcal{F}}/(KH), \gamma^2 = \sqrt{\log \mathcal{N}_\mathcal{F}}, \epsilon = \log \mathcal{N}_\mathcal{F} \cdot \sigma_{\min}^2/(KH)$ and $\{\widehat{\beta}_k\}_{k \geq 1}$ as

$$\widehat{\beta}_k = 3\sqrt{\iota_k} + 2\frac{\iota_k}{\gamma^2} + \sqrt{\lambda} + \sqrt{6kH\epsilon/\sigma_{\min}^2}.$$

Then for any $\delta > 0$, with probability at least $1 - 4(M + 1)\delta$, the regret of UCRL-WVTR is bounded by

$$\widetilde{O}\left(\sqrt{\dim_\mathcal{F} \log \mathcal{N}_\mathcal{F}} \cdot \sqrt{\mathcal{Q}^*} + \dim_\mathcal{F} \log \mathcal{N}_\mathcal{F}\right), \quad (4.1)$$

where $\iota_k = \widetilde{O}(\log \mathcal{N}_\mathcal{F})$, $\dim_\mathcal{F} = \dim_\mathcal{F}(\sigma_{\min}, KH)$, $\mathcal{N}_\mathcal{F} = \mathcal{N}_\mathcal{F}(\epsilon)$, and $\mathcal{Q}^* = \min\{\mathcal{Q}_0^*, V_1^* K\}$ with

$$Q_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}(V_{h+1}^*)^{2^m}](s_h^k, a_h^k),$$

$$\mathcal{Q}_0^* = \max_{m \in \overline{[M]}} Q_m, \quad V_1^* = \frac{1}{K} \sum_{k=1}^{K} V_1^*(s_1^k).$$

Proof. The proof features a novel concentration bound and fine-grained analyses on the higher order expansions of MDPs with general function approximation. See Section 4.4 for a proof sketch and Appendix F for a detailed proof. □

Our instance-dependent result given by Theorem 4.2 successfully eliminates the dependence on H, with the exception of logarithmic factors. This accomplishment represents a groundbreaking advancement in RL with general function approximation, as it achieves both *horizon-free* and *instance-dependent* for the first time.

The Quantity $\mathcal{Q}^*$ Since $\mathcal{Q}^* = \min\{\mathcal{Q}_0^*, V_1^* K\} \leq V_1^* K$, our regret bound in Theorem 4.2 immediately implies a first-order regret

$$\widetilde{O}\left(\sqrt{\dim_\mathcal{F} \log \mathcal{N}_\mathcal{F}} \cdot \sqrt{V_1^* K} + \dim_\mathcal{F} \log \mathcal{N}_\mathcal{F}\right).$$

The quantity $\mathcal{Q}_0^*$ defined in Theorem 4.2 is the maximum of Q_m over $m \in \overline{[M]}$, with Q_m being the sum of the variance of 2^m-th order of the optimal value function along the k-th trajectory over $k \in [K]$. Thereby, $\mathcal{Q}_0^*$ quantifies the stochasticity of the MDP under the

optimal policy and it vanishes when the transition kernel is deterministic. In specific, our results also indicate that the regret of learning a deterministic MDP is bounded by $\widetilde{O}(\dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}})$. By Proposition 3.5, this further implies a $\widetilde{O}(d^2)$ regret bound for linear mixture MDPs, matching the result in Zhou and Gu (2022).

To illustrate our theory more, we present implied regret bounds for linear mixture MDPs.

Corollary 4.3 (Regret for Linear Mixture MDPs). For d-dimensional linear mixture MDPs defined in Assumption 3.4, we set parameters in Algorithm 1 according to Theorem 4.2. Then for any $\delta > 0$, with probability at least $1 - 4(M + 1)\delta$, the regret of UCRL-WVTR is bounded by

$$\widetilde{O}(d\sqrt{\mathcal{Q}^*} + d^2) \leq \widetilde{O}(d\sqrt{V_1^* K} + d^2) \leq \widetilde{O}(d\sqrt{K} + d^2).$$

Proof. We have $\dim_{\mathcal{F}}, \log \mathcal{N}_{\mathcal{F}} = \widetilde{O}(d)$ by Proposition 3.5, then the result follows from Theorem 4.2. $\square$

Corollary 4.3 demonstrates that we also achieve a first-order regret guarantee for linear mixture MDPs, which covers the state-of-the-art worst-case regret bound $\widetilde{O}(d\sqrt{K}+d^2)$ by Zhou and Gu (2022), therefore matching the lower bound $\Omega(d\sqrt{K})$ (Zhou et al., 2021) when $K \geq d^2$ up to logarithmic factors. The outcome presented in Zhou and Gu (2022) critically depends on the feature $\phi_V(s, a) \in \mathbb{R}^d$ defined in Assumption 3.4. This feature endows the value function with a simple linear structure, rendering it amenable to analyze. Consequently, their proposed algorithms and analytical methodologies encounter limitations when applied to broader function classes characterized by intricate non-linear structures. We refer readers to Appendix C for more explanations.

4.4 Proof Sketch

In this section, we provide a proof sketch of the regret bound in Theorem 4.2, which primarily relies on the following key lemmas.

Concentration of the Estimated Model Our first step is to establish the optimism of the estimated value function as in Line 5 of Algorithm 1, which hinges on a novel concentration inequality in Theorem 4.4. Notably, this theorem resembles a Bernstein-style bound that substantially expands the scope of weighted linear regression (Zhou et al., 2021; Zhou and Gu, 2022) to non-linear settings, thus accommodating a broader class of functions. We refer readers to Appendix C for more technical differences with Zhou and Gu (2022).

Theorem 4.4. Let $\{\mathcal{G}_t\}_{t \geq 1}$ be a filtration, and $\{X_t\}_{t \geq 1}, \{Y_t\}_{t \geq 1}$ be stochastic processes such that

$X_t \in \mathcal{X}$ is $\mathcal{G}_{t-1}$-measurable and $Y_t \in [0, L]$ is $\mathcal{G}_t$-measurable. Let $f_* \in \mathcal{F}$ with function class $\mathcal{F}$ consisting of functions $f : \mathcal{X} \to [0, L]$. Suppose $\mathbb{E}[Y_t|\mathcal{G}_{t-1}] = f_*(X_t)$. Let

$$\widehat{f}_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{t} w_s^2 (f(X_s) - Y_s)^2, \qquad (4.2)$$

where the $\mathcal{G}_{s-1}$-measurable random variable w_s satisfies for all $s \in [t]$, $|w_s| \leq W$, $w_s^2 \operatorname{Var}[Y_s|\mathcal{G}_{s-1}] \leq \sigma^2$. Then for any $\delta, \epsilon > 0$, with probability at least $1 - \delta$, we have for all $t \geq 1$,

$$\sum_{s=1}^{t} w_s^2 (\widehat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq \beta_{t+1}^2 \text{ with}$$

$$\beta_{t+1} = 3\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_{\mathcal{F}}(X_s; X_{[s-1]}, 1/w_{[s-1]})$$

$$+ \sqrt{\lambda} + \sqrt{6Lt\epsilon/\sigma_{\min}^2},$$

where $\mathcal{D}_{\mathcal{F}}$ is in Definition 3.2 and $\iota_t = \widetilde{O}(\log \mathcal{N}_{\mathcal{F}})$.

Proof. See Appendix E for a detailed proof. $\square$

Remark 4.5. Theorem 4.4 provides a variance and uncertainty-aware bound of the deviation of $\widehat{f}_{t+1}$. In contrast, the bound proposed by Ayoub et al. (2020) for unweighted regression gives $\beta'_{t+1} = \widetilde{O}(L\sqrt{\log \mathcal{N}_{\mathcal{F}}})$, which is Hoeffding-type, i.e., it scales with the range of noise. Therefore, UCRL-VTR deviates from achieving an instance-dependent regret bound.

Recall the estimated model $\widehat{f}_{k+1,m}$ in Line 16 of Algorithm 1 is the solution of weighted least squares regression. We apply Theorem 4.4 for all $m \in \overline{[M]}$ with $\{X_t, w_t, Y_t\}_{t \geq 1} = \{z_{i,h,m}, 1/\bar{\sigma}_{i,h,m}, y_{i,h,m}\}_{i,h \in [k] \times [H]}$, which gives the confidence radius

$$\widehat{\beta}_K = \widetilde{O}(\log \mathcal{N}_{\mathcal{F}}).$$

Then the optimism property of the estimated value function in Line 5 can be proved with high probability.

Higher Order Expansion The regret bound can be related to the summation of bonuses (see Appendix F.3):

$$\text{Regret}(K) \lesssim \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\} + C,$$

where $\lesssim$ hides constants and C represents lower order terms. Let R_m, S_m denote the summation of bonuses and moments with respect to m-th level as follows:

$$R_m := \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\},$$

$$S_m := \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}V_{k,h+1}^{2^m}](s_h^k, a_h^k).$$

Then we can further bound R_m with Lemma 4.6.

Lemma 4.6 (Informal)**.** We have for all $m \in \overline{[M-1]}$,

$$R_m \lesssim \widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{S_m + R_{m+1}} + \widehat{\beta}_K^2 \dim_{\mathcal{F}} + C,$$

Lemma 4.6 establishes relationships between R_m and S_m, the estimated variance to the higher moment. Such a structure recursively expands for $M = O(\log KH)$ times, resulting in a regret bound only log-polynomially dependent on H. We further conduct a fine-grained analysis, rendering these quantities instance-dependent:

Lemma 4.7 (Informal)**.** We have

$$R_0 \lesssim \widehat{\beta}_K \sqrt{\dim_{\mathcal{F}} R_0} + \widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q}^*} + \widehat{\beta}_K^2 \dim_{\mathcal{F}} + C.$$

The completion of the proof necessitates in-depth analyses of higher-order quantities to eliminate lower-order terms. We refer readers to Appendix F for the formal statements of these critical lemmas, as well as an exhaustive and comprehensive proof of Theorem 4.2.

5 EXPERIMENTS

In this section, we conduct numerical experiments to demonstrate the effectiveness of our algorithmic designs. We adopt the episodic RiverSwim environment (Strehl and Littman, 2008), which is also considered in Ayoub et al. (2020). Figure 1 shows a 5-state River-Swim environment with action space $\mathcal{A} = \{a_1, a_2\}$. We choose $H \in \{20, 100\}$, where $H = 20$ in consistency with Ayoub et al. (2020), and $H = 100$ to demonstrate the performance on long planning horizon problems. And we set the number of episodes $K = 5000$. We refer readers to Appendix A for implementation details and additional experimental results.
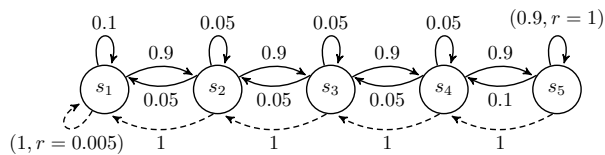

Figure 1: The transition probabilities and rewards of RiverSwim environment (Strehl and Littman, 2008). The solid arrow shows the transition kernel of action a_1. And the dashed arrow represents that of action a_2. The reward is denoted with the transition probabilities, if greater than 0.

We mainly consider three approaches to solve the episodic MDPs: (i) our proposed algorithm UCRL-WVTR; (ii) UCRL-WVTR without HOME; (iii) UCRL-VTR (Ayoub et al., 2020). We generate
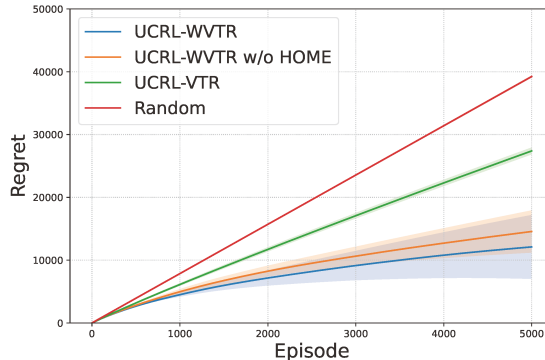
(a) A 5-state RiverSwim with $H = 20$.
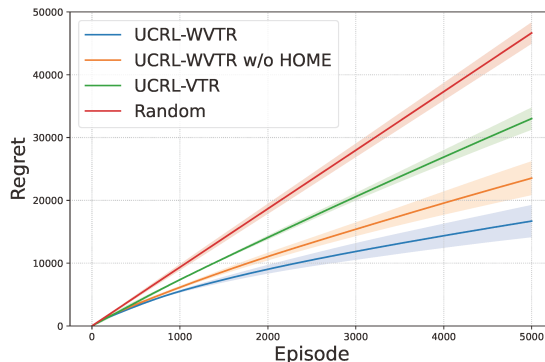
(b) A 5-state RiverSwim with $H = 100$.

Figure 2: Cumulative regret over episodes. `Random` represents the baseline where the agent acts uniformly.

10 independent paths for each setting. Their average regrets with error bars are shown in Figure 2.

As anticipated, UCRL-WVTR demonstrates superior performance compared to UCRL-VTR for both settings, underscoring the efficacy of weighted regression. It is noteworthy that the omission of the high-order moment estimator results in a slightly worse performance, thereby illustrating its practical utility. In summary, the experimental findings substantiate and reinforce our theoretical conclusions.

6 CONCLUSION

We propose a novel algorithm for model-based RL with general function approximation. We rigorously prove it achieves a horizon-free and instance-dependent regret bound. Moreover, it is computationally efficient given a regression oracle. Numerical experiments further validate the theoretical findings.

Acknowledgments

References

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

Agarwal, A., Jin, Y., and Zhang, T. (2023). VOQL: Towards optimal regret in model-free rl with non-linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR.

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

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Chen, F., Mei, S., and Bai, Y. (2022a). Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*.

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

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051.

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

Foster, D., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. (2018). Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR.

Foster, D. J., Golowich, N., and Han, Y. (2023). Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*.

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

Huang, J., Zhong, H., Wang, L., and Yang, L. (2023). Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. *Advances in Neural Information Processing Systems*, 36.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR.

Jiang, N. and Agarwal, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418.

Kim, Y., Yang, I., and Jun, K.-S. (2022). Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072.

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

Krishnamurthy, A., Agarwal, A., Huang, T.-K., Daumé III, H., and Langford, J. (2017). Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR.

Li, X. and Sun, Q. (2023). Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*.

Li, Y., Wang, R., and Yang, L. F. (2022). Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 965–976. IEEE.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. (2023). Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36.

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27.

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

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331.

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

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. (2022). First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR.

Wagenmaker, A. J. and Foster, D. J. (2023). Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR.

Wang, R., Du, S. S., Yang, L., and Kakade, S. (2020a). Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33:9075–9085.

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

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

Zhang, Z., Ji, X., and Du, S. (2021a). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR.

Zhang, Z., Ji, X., and Du, S. (2022). Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. (2021b). Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. (2023). Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4977–5020. PMLR.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. (2022). Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*.

Zhou, D. and Gu, Q. (2022). Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349.

Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR.

Zhou, R., Zihan, Z., and Du, S. S. (2023). Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pages 42878–42914. PMLR.

A ADDITIONAL EXPERIMENTS

Hyperparameters of Algorithms We provide more details about the three algorithms considered for the experiments in Section 5. These approaches share similar structures with Algorithm 1. (i) UCRL-WVTR: our proposed algorithm in Algorithm 1. (ii) UCRL-WVTR without HOME: it can be implemented by setting the level $M = 1$ in Algorithm 1. Intuitively, we simply use unweighted regression for solving $f_{k,h,1}$, which is then used for constructing the variance estimate $[\bar{\mathbb{V}}_{k,0}V_{k,h+1}](s_h^k, a_h^k) = \widehat{f}_{k,1}(z_{k,h,1}) - \widehat{f}_{k,0}^2(z_{k,h,0})$. (iii) UCRL-VTR: it can be implemented by setting the level $M = 0$, $\sigma_{\min} = 1$ and $\gamma = 0$ in Algorithm 1, since it simply uses unweighted regression, resulting the variance estimation $[\bar{\mathbb{V}}_{k,0}V_{k,h+1}](s_h^k, a_h^k) = 1$. Specifically, we list the hyperparameters in Table 1.

Table 1: Hyperparameters of Algorithms

Hyperparameters	UCRL-WVTR	UCRL-WVTR without HOME	UCRL-VTR
λ	0.001	0.001	0.001
$\sigma_{\min}$	0.01	0.01	1
γ	0.5	0.5	0
β	1	1	1
M	3	1	0

Environments Recall the RiverSwim environment consists of a series of n states organized in a linear sequence, as shown in Figure 3. The agent's initial position is on the far left, and it faces a decision at each state to either swim to the left or to the right. Notably, there exists a prevailing current that significantly facilitates leftward swimming while making rightward swimming more challenging. When swimming with the current, the agent is guaranteed to make progress to the left. However, swimming against the current predominantly leads to rightward movement but occasionally results in the agent moving left or remaining in the current state. The environment offers rewards only at the extreme ends: a reward of 0.05 on the far left and a reward of 1 on the far right. In essence, the agent is compelled to explore the environment thoroughly, ultimately weighing the decision of whether to brave the uncertain prospects of moving against the current for the potentially higher reward, or simply staying in the original position to secure a relatively smaller reward. Consequently, intelligent and strategic exploration becomes a fundamental requirement for acquiring an effective policy in this challenging environment.
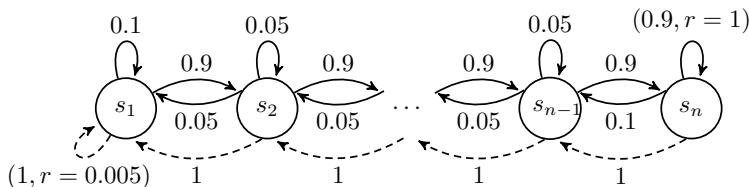

Figure 3: The transition probabilities and rewards of an n-state RiverSwim environment (Strehl and Littman, 2008). The solid arrow shows the transition kernel of action a_1. And the dashed arrow represents that of action a_2. The reward is denoted with the transition probabilities, if greater than 0.

Additional Results To demonstrate the effectiveness of our proposed algorithm more, we show additional experimental results. We set $H = 100$ to demonstrate the performance of these algorithms on long planning horizon problems. We will choose a series of gradually increasing the number of states of RiverSwim, i.e., $|S| \in \{6, 8, 10\}$, which will exponentially increase the difficulty of such an episodic RL problem.

As shown in Figure 4, the columns correspond to RiverSwim with $|S| = 6, 8, 10$, respectively. The first row shows the cumulative regret of each algorithm, and the second row represents average rewards within a single episode. We generate 10 independent paths for each algorithm, and show the average values plus or minus the standard deviation.

(a) Regret with $|S| = 6$ (b) Regret with $|S| = 8$ (c) Regret with $|S| = 10$

(d) Average Rewards with $|S| = 6$ (e) Average Rewards with $|S| = 8$ (f) Average Rewards with $|S| = 10$
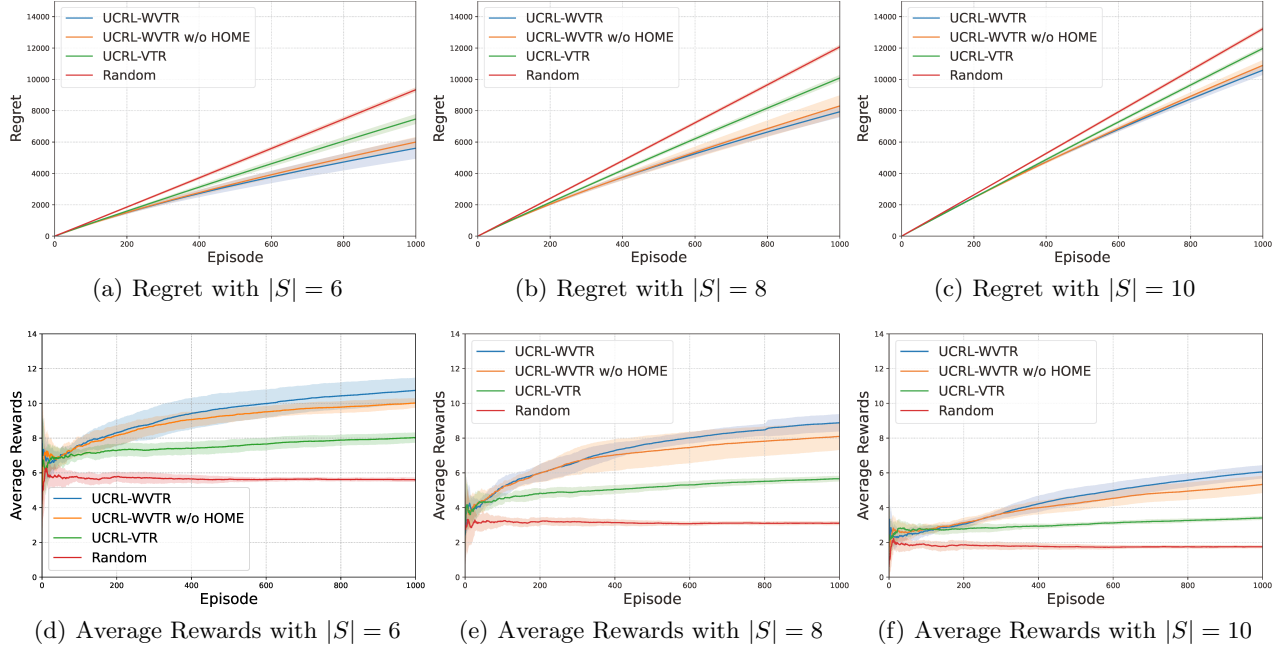
Figure 4: Regret and Average Rewards of 1000 Episodes

As expected, our proposed algorithm UCRL-WVTR outperforms UCRL-VTR, highlighting the effectiveness of weighted regression. It is worth noting that excluding the high-order moment estimator leads to a slightly inferior performance, thus demonstrating its practical value. While it introduces additional uncertainty in estimating higher-order moments, as clearly demonstrated by our results, the extra information from these variance estimations can leverage historical data more effectively, thereby accelerating the learning process. In summary, our experimental results provide strong support for and further reinforce our theoretical conclusions.

B COMPUTING UNCERTAINTY VIA THE REGRESSION ORACLE

In this section, we introduce how to efficiently compute uncertainty $\mathcal{D}_{\mathcal{F}}$ in Definition 3.2. The high-level idea is to decompose the task into solving a series of weighted least squares regression problems, whose solutions are directly available by the regression oracle defined in Assumption 4.1. Recall

$$D_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]}) = \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} (f_1(X_s) - f_2(X_s))^2 + \lambda}. \tag{B.1}$$

We denote $\mathcal{G} := \mathcal{F} - \mathcal{F}$. Assuming the optimal value of (B.1) is attained at f_1^*, f_2^*, let $g_* = f_1^* - f_2^* \in \mathcal{G}$, we claim that g_* is the solution to the following constrained optimization problem:

$$\max_{g \in \mathcal{G}} g(X_t) \quad \text{s.t.} \quad \sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g^2(X_s) \le \beta^*, \tag{B.2}$$

where $\beta^* = \sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_*^2(X_s)$ is unknown in prior. The claim can be easily proved by contradiction. If the solution to (B.2) is $\widetilde{g} \ne g_*$, then we have $\widetilde{g}(X_t) \ge g_*(X_t)$ while $\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} \widetilde{g}^2(X_s) \le \beta^* = \sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_*^2(X_s)$, which yields

$$\frac{\widetilde{g}^2(X_t)}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} \widetilde{g}^2(X_s) + \lambda} \ge \frac{g_*^2(X_t)}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_*^2(X_s) + \lambda},$$

a contradiction. Inspired by Foster et al. (2018), we then transform the problem (B.2) into solving the following weighted regression problem

$$\min_{g \in \mathcal{G}} \sum_{s=1}^{t} v_s(g(X_s) - Y_s)^2,$$

Algorithm 3 Computing Uncertainty

Require: Function class $\mathcal{F}$, dataset $\{X_{[t]}, \sigma_{[t-1]}\}$, λ, upper bound $\mathcal{B}$, precision ϵ and a regression oracle

Ensure: The estimated uncertainty $\widetilde{\mathcal{D}}$

1: Denote $R(g, v) := \sum_{s=1}^{t-1} v_s(g(X_s) - Y_s)^2 + v(g(X_t) - 1)^2$
2: Set $\mathcal{G} \leftarrow \mathcal{F} - \mathcal{F}, \bar{\beta} \leftarrow 2\lambda, \widetilde{\mathcal{D}} \leftarrow 0$
3: **while** $\bar{\beta} < 2\mathcal{B}$ **do**
4:     Set $\beta \leftarrow \bar{\beta} - \lambda$
5:     Set $v_L \leftarrow 0, v_H \leftarrow 2\beta/\epsilon$
6:     Set $g_L \leftarrow 0, g_H \leftarrow \mathrm{argmin}_{g \in \mathcal{G}} R(g, v_H)$
7:     Set $z_L \leftarrow 0, z_H \leftarrow g_H(X_t)$
8:     Set $\Delta \leftarrow \epsilon\beta/4$
9:     **while** $|z_H - z_L| > \epsilon$ and $|v_H - v_L| > \Delta$ **do**
10:         Set $\widetilde{v} \leftarrow (v_H + v_L)/2$
11:         Set $\widetilde{g} \leftarrow \mathrm{argmin}_{g \in \mathcal{G}} R(g, \widetilde{v})$
12:         Set $\widetilde{z} \leftarrow \widetilde{g}(X_t)$
13:         **if** $\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} \widetilde{g}^2(X_s) > \beta$ **then**
14:             $v_H \leftarrow \widetilde{v}, z_H \leftarrow \widetilde{g}(X_t)$
15:         **else**
16:             $v_L \leftarrow \widetilde{v}, z_L \leftarrow \widetilde{g}(X_t)$
17:         **end if**
18:     **end while**
19:     Set $\widetilde{\mathcal{D}} \leftarrow \max\left\{\widetilde{\mathcal{D}}, \widetilde{g}(X_t)/\sqrt{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} \widetilde{g}^2(X_s) + \lambda}\right\}$
20:     Set $\bar{\beta} = 2\bar{\beta}$
21: **end while**

---

where $v_s = 1/\sigma_s^2$ for $s \in [t-1]$, $v_t \in \mathbb{R}^+$ is some constant unknown to the agent, $Y_s = 0$ for $s \in [t-1]$ and $Y_t = 1$. The solution is available with access to a regression oracle. We leverage a doubling trick for β^* and a binary search technique for the proper value of v_t. See Algorithm 3 for a detailed description.

Proposition B.1. Consider the following optimization problem for solving the uncertainty

$$\max_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} (f_1(X_s) - f_2(X_s))^2 + \lambda}.$$

By definition, the optimal value is $\mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})$, and we assume its solution is $g^* = f_1^* - f_2^*$. For any $\epsilon > 0$, we run Algorithm 3 to solve the problem above with $\mathcal{B}$ satisfying $\mathcal{B} \geq \sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_*^2(X_s) + \lambda$. If the function class $\mathcal{F}$ is convex and closed under point-wise convergence, then Algorithm 3 terminates within $O(\log(\mathcal{B}/\lambda) \log(1/\epsilon))$ calls to the regression oracle and returns $\widetilde{\mathcal{D}}$ such that

$$\widetilde{\mathcal{D}} - \epsilon/\sqrt{\lambda} \leq D_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \leq \sqrt{2}\widetilde{\mathcal{D}} + \epsilon/\sqrt{\lambda}.$$

Proof. Note $\mathcal{G} = \mathcal{F} - \mathcal{F}$ is convex due to the convexity of $\mathcal{F}$. Denote g_α as the ϵ-precision solution of the following constrained optimization problem:

$$\max_{g \in \mathcal{G}} g(X_t) \quad \text{s.t.} \quad \sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g^2(X_s) \leq \lambda 2^\alpha - \lambda. \tag{B.3}$$

Then according to Theorem 1 of Foster et al. (2018), g_α is computed within $O(\log(1/\epsilon))$ time, and the estimated uncertainty $\widetilde{\mathcal{D}}$ returned by Algorithm 3 is actually the maximum of the following quantities over $\alpha \in \{0, 1, \ldots, \lceil \log_2(\mathcal{B}/\lambda) \rceil\}$, as shown in (B.4).

$$\widetilde{\mathcal{D}} = \max_\alpha \frac{g_\alpha(X_t)}{\sqrt{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_\alpha^2(X_s) + \lambda}}. \tag{B.4}$$

Due to the definition of $\mathcal{B}$, there exists $\alpha \in \{0, 1, \ldots, \lceil \log_2(\mathcal{B}/\lambda) \rceil - 1\}$, such that $\lambda 2^\alpha \le \beta^* + \lambda \le \lambda 2^{\alpha+1}$. On one hand,

$$
\mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]}) = \frac{g_*(X_t)}{\sqrt{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_*^2(X_s) + \lambda}} = \frac{g_*(X_t)}{\sqrt{\beta^* + \lambda}} \le \frac{g_*(X_t)}{\sqrt{\lambda 2^\alpha}} = \sqrt{2} \frac{g_*(X_t)}{\sqrt{\lambda 2^{\alpha+1}}}
$$

$$
\overset{(a)}{\le} \sqrt{2} \frac{g_{\alpha+1}(X_t)}{\sqrt{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2} g_{\alpha+1}^2(X_s) + \lambda}} + \frac{\epsilon}{\sqrt{\lambda 2^\alpha}} \overset{(b)}{\le} \sqrt{2}\widetilde{D} + \frac{\epsilon}{\sqrt{\lambda}},
$$

where (a) holds due to the definition of $g_{\alpha+1}$ in (B.3) and (b) holds due to (B.4). On the other hand, $\mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \ge \widetilde{D} - \epsilon/\sqrt{\lambda}$ due to the definition of $\mathcal{D}_{\mathcal{F}}$. The proof is completed. $\qquad\square$

Corollary B.2. To invoke Algorithm 3 in UCRL-WVTR, we set $\mathcal{B} = KH/\sigma_{\min}^2 + \lambda$, $\epsilon = \sqrt{\lambda}/(KH)$, then the estimated uncertainty $\widetilde{\mathcal{D}}$ can be derived within $\widetilde{O}(1)$ calls to the regression oracle such that $\widetilde{\mathcal{D}} - 1/(KH) \le \mathcal{D}_{\mathcal{F}} \le \sqrt{2}\widetilde{\mathcal{D}} + 1/(KH)$. The error in terms of the estimated uncertainty $\mathcal{D}_{\mathcal{F}}$ only enlarges the regret bound by at most constant terms.

Remark B.3. It is noteworthy that Kong et al. (2021) adopted a similar idea to compute the sensitivity score defined therein via the regression oracle. However, their approach is not directly applicable to ours, since (i) They consider the unweighted regression setting, which is a special case of ours by setting $\sigma_{[t]} = \mathbf{1}$. (ii) The sensitivity score in their work is used for constructing a dataset that approximates historical samples and for applying a low-switching updating scheme, while we use uncertainty to directly update the value function.

C COMPARISONS WITH PREVIOUS WORK

We make some comparisons with Zhou and Gu (2022), with emphasis on why their approaches for linear mixture MDPs defined in Assumption 3.4 can not be extended to RL with general function approximation.

First, at the core of their analysis is a Bernstein-style concentration bound (Theorem 4.3 in their work) for the deviation $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t}$, where $\boldsymbol{\mu}_t$ is the solution of least squares regression with (weighted) data $\{\boldsymbol{x}_s\}_{s\in[t]}$ and targets $\{y_s\}_{s\in[t]}$, $\mathbf{Z}_t = \sum_{s=1}^t \boldsymbol{x}_s \boldsymbol{x}_s^\top + \lambda \boldsymbol{I}$. Assuming a filtration $\{\mathcal{G}_t\}_{t\ge1}$, here $\boldsymbol{x}_t$ is $\mathcal{G}_{t-1}$-measurable, and $y_t = \langle \boldsymbol{\theta}^*, \boldsymbol{x}_t \rangle + \eta_t$ with noise $\eta_t \in \mathcal{G}_t$. Plugging in the analytical form of the estimate $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \sum_{s=1}^t \boldsymbol{x}_s y_s$, the following inequality holds

$$
\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le \left\| \sum_{s=1}^t \boldsymbol{x}_s \eta_s \right\|_{\mathbf{Z}_t^{-1}} + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2. \tag{C.1}
$$

Denote $\boldsymbol{d}_t = \sum_{s=1}^t \boldsymbol{x}_s \eta_s$, $Z_t = \|\boldsymbol{d}_t\|_{\mathbf{Z}_t^{-1}}$. They then decompose $\|\sum_{s=1}^t \boldsymbol{x}_s \eta_s\|_{\mathbf{Z}_t^{-1}}^2 = Z_t^2$ into two bounded martingale difference sequences as in (C.2) and apply Freedman's inequality for deriving high probability bounds of these two terms separately.

$$
Z_t^2 \le \sum_{s=1}^t \frac{2\eta_s \boldsymbol{x}_s^\top \mathbf{Z}_{s-1}^{-1} \boldsymbol{d}_{s-1}}{1 + w_s^2} + \sum_{s=1}^t \frac{\eta_s^2 w_s^2}{1 + w_s^2}, \tag{C.2}
$$

where $w_t = \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}$. Such a self-normalized expression in (C.1) holds thanks to the analytical solution $\boldsymbol{\mu}_t$ of weighted linear regression, and a decomposition in (C.2) heavily depends on the linear structure of the estimate $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \sum_{s=1}^t \boldsymbol{x}_s y_s$. Regrettably, these properties no longer hold in non-linear settings, necessitating a new approach for providing a tight concentration bound on the solution to least squares regression as shown in Theorem 4.4. We refer the readers to its proof in Appendix E for details.

Second, once deriving a tight estimate of model with $\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_k} \le \widehat{\beta}_k$, given any state-action pair (s, a) and the next-state value function V, one can naturally construct an upper bound of the next-state value function evaluated on (s, a) by explicitly adding a bonus of the following form

$$
|\widehat{f}_k(s, a, V) - f_*(s, a, V)| = \|\langle \boldsymbol{\phi}_V(s, a), \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_* \rangle\| \le \widehat{\beta}_k \|\boldsymbol{\phi}_V(s, a)\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}, \tag{C.3}
$$

where $\widehat{\boldsymbol{\Sigma}}_k = \sum_{i=1}^{k-1} \sum_{h=1}^H \frac{1}{\bar{\sigma}_{i,h}^2} \boldsymbol{\phi}_{i,h} \boldsymbol{\phi}_{i,h}^\top + \lambda \boldsymbol{I}$ with $\boldsymbol{\phi}_{k,h} = \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$. Then, adopting the policy that selecting action with the maximum value function, and using the standard technique, one can approximately bound the

K-episode regret as

$$\text{Regret}(K) \lesssim \widehat{\beta}_K \sum_{k=1}^{K} \sum_{h=1}^{H} \|\phi_{k,h}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}} \lesssim \widehat{\beta}_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\sigma}_{k,h}^2} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \left\| \frac{\phi_{k,h}}{\bar{\sigma}_{k,h}} \right\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}^2}. \tag{C.4}$$

Here $\sum_{k=1}^{K} \sum_{h=1}^{H} \left\| \frac{\phi_{k,h}}{\bar{\sigma}_{k,h}} \right\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}^2$ can be easily bounded by $\widetilde{O}(d)$ by elliptical potential lemma (Abbasi-Yadkori et al., 2011). However, in the realm of RL with general function approximation, the bonus term does not enjoy such a simple form as in (C.3), and the elliptical potential lemma is not available. To this end, we define the bonus as an optimization problem over the general function class, where an analytical solution is generally not attainable. To overcome this obstacle, we further introduce a computationally efficient algorithm for computing uncertainty for the bonus term via a regression oracle, thus achieving computational efficiency despite the non-linear structure. We then carefully bound the regret with respect to the generalized Eluder dimension. Please see our definition of the uncertainty and the generalized Eluder dimension in Definition 3.2, and computational complexity in Section 4.2.

Furthermore, they leverage a high-moment estimator for a tight upper bound of $\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\sigma}_{k,h}^2$ in (C.4). More specifically, they estimate the variances of the m-th moment of the next-state value functions up to $M = O(\log(KH))$ levels, and recursively establish the relationships between quantities of MDPs of these levels. Consequently, they obtain an upper bound of the variance term with only log-polynomial dependence on H, thus ultimately achieving a horizon-free regret bound. However, these lemmas with respect to the higher-order expansions of MDPs may not hold in the context of general function approximation. Despite the inherent complexities posed by the non-linear function class, we have extended these methodologies to achieve a horizon-free regret bound with log-polynomial dependence on the planning horizon H. Significantly, our refined analysis also establishes a connection with certain problem-specific quantities, consequently rendering the regret bound of UCRL-WVTR instance-dependent. Comprehensive elaboration on this achievement is provided in Section 4.3.

D   PROOFS FOR LINEAR MIXTURE MDPS

D.1   Proof of Proposition 3.5

Proof of Proposition 3.5. We first bound the covering number, then come to the generalized Eluder dimension.

Covering Number   For any $f \in \mathcal{F}$, we have for any (s, a, V),

$$f(s, a, V) = [\mathbb{P}V](s, a) = \sum_{s' \in \mathcal{S}} \langle \boldsymbol{\theta}^*, \phi(s'|s, a) \rangle V(s')$$

$$= \langle \boldsymbol{\theta}^*, \sum_{s' \in \mathcal{S}} \phi(s'|s, a) V(s') \rangle = \langle \boldsymbol{\theta}^*, \phi_V(s, a) \rangle, \tag{D.1}$$

where $\|\boldsymbol{\theta}^*\|_2 \leq B$. This implies f is a linear function of $\phi_V(s, a)$ and the ϵ-log-covering number of $\mathcal{F}$ follows from standard covering-number arguments.

Generalized Eluder Dimension   First, we show uncertainty $\mathcal{D}_{\mathcal{F}}$ has an analytic expression for linear mixture MDPs in the following lemma.

Lemma D.1. For d-dimensional, B-bounded linear mixture MDPs defined in Assumption 3.4, given any $\boldsymbol{X} = \{(s_t, a_t, V_t)\}_{t \in [T]}$, $\boldsymbol{\sigma} = \{\sigma_t\}_{t \in [T]}$, we have

$$\mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]}) = \|\phi_t\|_{\boldsymbol{\Sigma}_{t-1}},$$

where $\phi_t = \phi_{V_t}(s_t, a_t)$ and $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t} \frac{1}{\sigma_s^2} \phi_s \phi_s^\top + \lambda/B \cdot \boldsymbol{I}$ for all $t \in [T]$.

Proof. For any $f_1, f_2 \in \mathcal{F}$, let $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfies for any $X = (s, a, V)$,

$$f_1(X) - f_2(X) = \langle \boldsymbol{\theta}, \phi_V(s, a) \rangle,$$

where the linear representation of f_1 and f_2 follows from (D.1). Then we have

$$\frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \lambda} \overset{(a)}{\leq} \frac{\langle \boldsymbol{\theta}, \boldsymbol{\phi}_t \rangle^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2}\langle \boldsymbol{\theta}, \boldsymbol{\phi}_s \rangle^2 + \frac{\lambda}{B}\|\boldsymbol{\theta}\|_2^2} \overset{(b)}{=} \frac{\langle \boldsymbol{\theta}, \boldsymbol{\phi}_t \rangle^2}{\|\boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_{t-1}}^2} \overset{(c)}{\leq} \|\boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^2,$$

where (a) holds due to $\|\boldsymbol{\theta}\|_2 \leq B$, (b) holds due to the definition of $\boldsymbol{\Sigma}_{t-1}$ and (c) holds due to Cauchy-Schwarz inequality. Thus $\mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]}) = \|\boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}$ by Definition 3.2. $\qquad\square$

Next, we come to the generalized Eluder dimension. We further assume for all $t \in [T]$, $\sigma_t \geq \sigma_{\min}$, then it follows that

$$\sum_{t=1}^{T} \min\left\{1, \frac{1}{\sigma_t^2} \mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]})\right\} = \sum_{t=1}^{T} \min\left\{1, \frac{1}{\sigma_t^2}\|\boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^2\right\} \leq 2d\log\left(1 + \frac{TB}{d\lambda\sigma_{\min}^2}\right),$$

where the inequality holds due to Lemma H.3. Taking maximum of the LHS over $|\boldsymbol{X}| = T, \boldsymbol{\sigma} \geq \sigma_{\min}$ and the proof is completed by Definition 3.2. $\qquad\square$

E   PROOF OF THEOREM 4.4

Proof of Theorem 4.4. Recall the definition of $\widehat{f}_{t+1}$ in (4.2), which implies

$$\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq 2\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))(Y_s - f_*(X_s)).$$

For any fixed $f \in \mathcal{F}$, denote $E_s(f) = w_s^2(f(X_s) - f_*(X_s))(Y_s - f_*(X_s))$, which is a martingale difference sequence adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$. Inspired by Agarwal et al. (2023), we are to give a high-probability bound of $\sum_{s=1}^{t} E_s(f)$ in terms of $\sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2$. Note $|w_s| \leq W$ and $f(X_s), f_*(X_s), Y_s$ are bounded in $[0, L]$, thereby the expectation and summation of variances are upper bounded by

$$|E_s(f)| \leq W^2 L^2, \qquad \sum_{s=1}^{t} \mathbb{E}[E_s^2(f)|\mathcal{G}_{s-1}] \leq \sigma^2 W^2 L^2 t.$$

Simultaneously, the following problem-specific bounds hold:

$$\max_{s \in [t]} |E_s(f)| \overset{(a)}{\leq} L \max_{s \in [t]} w_s^2 \mathcal{D}_{\mathcal{F}}(X_s; X_{[s-1]}, 1/w_{[s-1]})\sqrt{\sum_{i=1}^{s-1} w_i^2(f(X_i) - f_*(X_i))^2 + \lambda}$$

$$\leq L \max_{s \in [t]} w_s^2 \mathcal{D}_{\mathcal{F}}(X_s; X_{[s-1]}, 1/w_{[s-1]}) \cdot \sqrt{\sum_{s=1}^{t-1} w_s^2(f(X_s) - f_*(X_s))^2 + \lambda},$$

$$\sum_{s=1}^{t} \mathbb{E}[E_s^2(f)|\mathcal{G}_{s-1}] \leq \sigma^2 \sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2,$$

where (a) holds due to the definition of $\mathcal{D}_{\mathcal{F}}$ in Definition 3.2. We denote $\mathcal{D}_s = \mathcal{D}_{\mathcal{F}}(X_s; X_{[s-1]}, 1/w_{[s-1]})$ for short. Let $\epsilon > 0$ be a constant and $\mathcal{V}$ be a ϵ-covering net of $\mathcal{F}$. Applying Lemma H.1 with $m = v = \sigma^2$, $\iota_t = 16\log\frac{2\mathcal{N}_{\mathcal{F}}(\epsilon)t^2(\log(\sigma^2 W^2 L^2 t) + 2)(\log(W^2 L^2) + 2)}{\delta}$, together with a union bound over $f \in \mathcal{V}$, with probability at

least $1 - \delta/(2t^2)$, we have for any fixed $t \geq 1$ and all $f \in \mathcal{V}$,

$$2\sum_{s=1}^{t} E_s(f) \leq \sqrt{\iota_t} \cdot \sqrt{\sigma^2 \sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2 + \sigma^4}$$

$$+ \iota_t \left(L \max_{s \in [t]} w_s^2 \mathcal{D}_s \cdot \sqrt{\sum_{s=1}^{t-1} w_s^2(f(X_s) - f_*(X_s))^2 + \lambda + \sigma^2} \right)$$

$$\overset{(a)}{\leq} \left(\sqrt{\iota_t}\sigma + \iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right) \sqrt{\sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2} + \sqrt{\lambda} \iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s + 2\iota_t \sigma^2$$

$$\overset{(b)}{\leq} \frac{1}{2}\sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2 + \frac{1}{2}\left(\sqrt{\iota_t}\sigma + \iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2$$

$$+ \frac{1}{2}\left(\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2 + \frac{1}{2}\lambda + 2\iota_t\sigma^2$$

$$\overset{(c)}{\leq} \frac{1}{2}\sum_{s=1}^{t} w_s^2(f(X_s) - f_*(X_s))^2 + \frac{1}{2}\left(\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2 + \frac{1}{2}\lambda + 2\iota_t\sigma^2,$$

where (a) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$ and $\iota_t \geq 1$, (b) holds due to $\sqrt{ab} \leq a/2 + b/2$ for any $a, b \geq 0$, and (c) holds due to $a^2 + b^2 \leq (a+b)^2$ for any $a, b \geq 0$. Let $g \in \mathcal{V}$ such that $\|g - \widehat{f}_{t+1}\|_\infty \leq \epsilon$, then

$$\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))^2$$

$$\leq 2\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))(Y_s - f_*(X_s))$$

$$\leq 2\sum_{s=1}^{t} w_s^2(g(X_s) - f_*(X_s))(Y_s - f_*(X_s)) + \frac{2Lt}{\sigma_{\min}^2}\epsilon$$

$$\leq \frac{1}{2}\sum_{s=1}^{t} w_s^2(g(X_s) - f_*(X_s))^2 + \frac{1}{2}\left(\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2 + \frac{1}{2}\lambda + 2\iota_t\sigma^2 + \frac{2Lt}{\sigma_{\min}^2}\epsilon$$

$$\leq \frac{1}{2}\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))^2 + \frac{1}{2}\left(\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2 + \frac{1}{2}\lambda + 2\iota_t\sigma^2 + \frac{3Lt}{\sigma_{\min}^2}\epsilon.$$

That is for any fixed $t > 0$, we have

$$\sum_{s=1}^{t} w_s^2(\widehat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq \left(\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s \right)^2 + \lambda + 4\iota_t\sigma^2 + \frac{6Lt}{\sigma_{\min}^2}\epsilon$$

$$\leq \left(3\sqrt{\iota_t}\sigma + 2\iota_t L \max_{s \in [t]} w_s^2 \mathcal{D}_s + \sqrt{\lambda} + \sqrt{\frac{6Lt\epsilon}{\sigma_{\min}^2}} \right)^2,$$

where the second inequality holds due to $2\sqrt{ab} \leq a + b$ and $a + b \leq (\sqrt{a} + \sqrt{b})^2$ for any $a, b \geq 0$. Finally, the result holds through a union bound over all $t \geq 1$ and the fact that $\sum_{t=1}^{\infty} \frac{1}{2t^2} \leq 1$. $\qquad\square$

F   PROOF OF THEOREM 4.2

We define filtration $\{\mathcal{G}_{k,h}\}_{k \in [K], h \in [H]}$ as follows. Let $\mathcal{G}_{k,h}$ be the σ-field generated by the state-action pairs up to h-th step and k-th episode. That is $\mathcal{G}_{k,h} = \sigma(\{s_h^i, a_h^i\}_{(i,h) \in [k-1] \times [H]} \cup \{s_j^k, a_j^k\}_{j \in [h]})$. For constants $\sigma_{\min}, \epsilon > 0$, let $\mathcal{N}_\mathcal{F} = \mathcal{N}_\mathcal{F}(\epsilon)$, and $\dim_\mathcal{F} := \dim_\mathcal{F}(\sigma_{\min}, KH)$ be the generalized Eluder dimension in Definition 3.2.

F.1 High Probability Events and Optimism

Let $\widehat{\mathcal{B}}_{k,m}$ denote the confidence region as follows:

$$\widehat{\mathcal{B}}_{k,m} := \left\{ f : \sum_{i=1}^{k-1} \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{i,h,m}^2} (\widehat{f}_{k,h,m}(z_{i,h,m}) - f(z_{i,h,m}))^2 \leq \widehat{\beta}_k^2 \right\}, \tag{F.1}$$

where

$$\widehat{\beta}_k = 3\sqrt{\iota_k} + 2\frac{\iota_k}{\gamma^2} + \sqrt{\lambda} + \sqrt{6kH\epsilon/\sigma_{\min}^2} \tag{F.2}$$

with $\iota_k = 16 \log \frac{2\mathcal{N}_{\mathcal{F}}(\epsilon)k^2 H^2 (\log(kH/\sigma_{\min}^2)+2)(\log(1/\sigma_{\min}^2)+2)}{\delta}$. We define event $\mathcal{E}$ as

$$\mathcal{E} := \{f_* \in \widehat{\mathcal{B}}_{k,m} \text{ for all } k \in [K], m \in \overline{[M]}\}.$$

The following lemmas hold.

Lemma F.1. For any $k \in [K], m \in [M]$, if $f_* \in \widehat{\mathcal{B}}_{k,m}$, we have for all $h \in [H]$,

$$|\widehat{f}_{k,m}(z_{k,h,m}) - f_*(z_{k,h,m})| \leq \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m}).$$

Furthermore,

$$V_{k,h}(s_h^k) - r_h(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \leq 2\min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\},$$
$$|[\bar{\mathbb{V}}_{k,h} V_{k,h+1}^{2^m}](s_h^k, a_h^k) - [\mathbb{V}_{k,h} V_{k,h+1}^{2^m}](s_h^k, a_h^k)| \leq E_{k,h,m}.$$

Proof. See Appendix G.1 for a detailed proof. □

Lemma F.2. Event $\mathcal{E}$ holds with probability at least $1 - (M+1)\delta$.

Proof. See Appendix G.2 for a detailed proof. □

Lemma F.3. On event $\mathcal{E}$, we have for all $k, h \in [K] \times [H]$, $Q_{k,h}(\cdot, \cdot) \geq Q_h^*(\cdot, \cdot)$, $V_{k,h}(\cdot) \geq V_h^*(\cdot)$.

Proof. See Appendix G.3 for a detailed proof. □

F.2 Higher Order Expansion of MDPs

Inspired by Zhang et al. (2021b); Zhou and Gu (2022); Zhao et al. (2023), we define the following quantities of MDPs. For all $k \in [K], h \in [H+1]$, We use I_h^k to denote the following events:

$$I_h^k := \mathbb{1}\left\{ \forall m \in \overline{[M]}, \sum_{j=1}^{h-1} \frac{1}{\bar{\sigma}_{k,j,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,j,m}; D_{k,j-1,m}) \leq 1 \right\}. \tag{F.3}$$

Note I_h^k is $\mathcal{G}_{k,h-1}$-measurable and monotonically decreasing. We define h_k for all $k \in [K]$ as the least h such that I_h^k vanishes.

$$h_k = \min\{h \in [H+1] | I_h^k = 0\}. \tag{F.4}$$

We use the quantity G to denote the number of episodes when the uncertainty quantity grows sharply:

$$G = \sum_{k=1}^{K} (1 - I_{H+1}^k). \tag{F.5}$$

We use $\check{V}_{k,h}(s)$ to denote the estimation error between the estimated value function and the optimal value function, and use $\widetilde{V}_{k,h}(s)$ to denote the sub-optimality gap of policy π^k at stage h:

$$\check{V}_{k,h}(s) = V_{k,h}(s) - V_h^*(s), \quad \forall s \in \mathcal{S}, (k,h) \in [K] \times [H], \tag{F.6}$$
$$\widetilde{V}_{k,h}(s) = V_h^*(s) - V_h^{\pi^k}(s), \quad \forall s \in \mathcal{S}, (k,h) \in [K] \times [H]. \tag{F.7}$$

In addition, we use $\check{S}_m, \widetilde{S}_m, S_m, Q_m$ to represent the total variance of 2^m-th order value functions $(\check{V}_{k,h+1}^{2^m}, \widetilde{V}_{k,h+1}^{2^m}, V_{k,h+1}^{2^m}, (V_{h+1}^*)^{2^m})$:

$$\check{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k [\mathbb{V}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k), \tag{F.8}$$

$$\widetilde{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k [\mathbb{V}\widetilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k), \tag{F.9}$$

$$S_m = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k [\mathbb{V}V_{k,h+1}^{2^m}](s_h^k, a_h^k), \tag{F.10}$$

$$Q_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}(V_{h+1}^*)^{2^m}](s_h^k, a_h^k). \tag{F.11}$$

We further use $\mathcal{Q}_0^*$ to denote the maximum of Q_m:

$$\mathcal{Q}_0^* = \max_{m \in [\overline{M}]} Q_m. \tag{F.12}$$

Then, for 2^m-th order value functions $(\check{V}_{k,h+1}^{2^m}, \widetilde{V}_{k,h+1}^{2^m}, V_{k,h+1}^{2^m})$, we use $\check{A}_m, \widetilde{A}_m, A_m$ to denote the summation of stochastic transition noise as follows:

$$\check{A}_m = \left| \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^m}(s_{h+1}^k) \right] \right|, \tag{F.13}$$

$$\widetilde{A}_m = \left| \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \widetilde{V}_{k,h+1}^{2^m}(s_{h+1}^k) \right] \right|, \tag{F.14}$$

$$A_m = \left| \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{k,h+1}^{2^m}](s_h^k, a_h^k) - V_{k,h+1}^{2^m}(s_{h+1}^k) \right] \right|. \tag{F.15}$$

Next, we use V_0 to represent the total rewards and V_1^* for the average optimal value functions over K episodes:

$$V_0 = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k r(s_h^k, a_h^k), \tag{F.16}$$

$$V_1^* = \frac{1}{K} \sum_{k=1}^{K} V_1^*(s_1^k). \tag{F.17}$$

Finally, we use the R_m to denote the summation of bonuses:

$$R_m = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\}. \tag{F.18}$$

Now, we introduce the following lemmas to build the connection between these quantities.

Lemma F.4. We have

$$G \leq (M+1) \dim_{\mathcal{F}}. \tag{F.19}$$

Proof. See Appendix G.4 for a detailed proof. □

Lemma F.5. On event $\mathcal{E}$, we have for all $m \in [\overline{M}]$,

$$\check{S}_m \leq \check{A}_{m+1} + G + 2^{m+1} \cdot (2R_0), \tag{F.20}$$

$$\widetilde{S}_m \leq \widetilde{A}_{m+1} + G + 2^{m+1} \cdot (2R_0 + \check{A}_0), \tag{F.21}$$

$$S_m \leq A_{m+1} + G + 2^{m+1} \cdot (V_0 + 2R_0). \tag{F.22}$$

Proof. See Appendix G.5 for a detailed proof. □

Lemma F.6. With probability at least $1 - 3(M+1)\delta$, we have for all $m \in \overline{[M]}$,

$$\check{A}_m \leq \sqrt{\zeta \check{S}_m} + \zeta, \tag{F.23}$$

$$\widetilde{A}_m \leq \sqrt{\zeta \check{S}_m} + \zeta, \tag{F.24}$$

$$A_m \leq \sqrt{\zeta S_m} + \zeta, \tag{F.25}$$

where $\zeta = 8\log(2(\log(KH) + 2)/\delta)$. We denote the corresponding event by $\mathcal{A}$.

Proof. See Appendix G.6 for a detailed proof. □

Lemma F.7. On event $\mathcal{E} \cup \mathcal{A}$, we have for all $m \in \overline{[M]}$,

$$\check{A}_m \leq \sqrt{\zeta} \cdot \sqrt{\check{A}_{m+1} + G + 2^{m+1} \cdot (2R_0)} + \zeta. \tag{F.26}$$

And

$$\check{A}_0 \leq 2\sqrt{2\zeta R_0} + 3\sqrt{\zeta G} + 7\zeta, \tag{F.27}$$

$$\check{A}_1 \leq 4\sqrt{\zeta R_0} + 3\sqrt{\zeta G} + 7\zeta. \tag{F.28}$$

Proof. See Appendix G.7 for a detailed proof. □

Lemma F.8. On event $\mathcal{E} \cup \mathcal{A}$, we have

$$\widetilde{A}_0 \leq 4\sqrt{2\zeta R_0} + 6\sqrt{\zeta G} + 15\zeta. \tag{F.29}$$

Proof. See Appendix G.8 for a detailed proof. □

Lemma F.9. We have

$$V_0 \leq V_1^* K + A_0 + \check{A}_0. \tag{F.30}$$

Proof. See Appendix G.9 for a detailed proof. □

Lemma F.10 (Formal version of Lemma 4.6). Let $\gamma^2 \leq \log \mathcal{N}_{\mathcal{F}}$. On event $\mathcal{E}$, we have for all $m \in \overline{[M-1]}$,

$$R_m \leq 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{S_m + 2R_{m+1} + KH\sigma_{\min}^2} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{F.31}$$

Proof. See Appendix G.10 for a detailed proof. □

Lemma F.11 (Formal version of Lemma 4.7). On event $\mathcal{E} \cup \mathcal{A}$, we have

$$R_0 + \check{A}_0 \leq 32\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}} R_0} + 15\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q}_0^* + G + KH\sigma_{\min}^2} + 587\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{F.32}$$

Proof. See Appendix G.11 for a detailed proof. □

Lemma F.12. On event $\mathcal{E} \cup \mathcal{A}$, we have

$$R_0 + A_0 \leq 32\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_1^* K + \check{A}_0} + 30\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 1174\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{F.33}$$

Proof. See Appendix G.12 for a detailed proof. □

F.3 Regret Analysis

Proof of Theorem 4.2. On event $\mathcal{E} \cup \mathcal{A}$, which holds with probability at least $1 - 4(M+1)\delta$ by Lemma F.2, F.6 and a union bound, we have all lemmas in this section hold. By the optimism implied by Lemma F.3, we have

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right] \leq \sum_{k=1}^{K} \left[V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k) \right]. \tag{F.34}$$

We further use Lemma F.13 to bound the regret with the higher-order quantities defined in Section F.2.

Lemma F.13. On event $\mathcal{E}$, we have

$$\sum_{k=1}^{K} \left[V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k) \right] \leq 2R_0 + \check{A}_0 + \widetilde{A}_0 + G. \tag{F.35}$$

Proof. See Appendix G.13 for a detailed proof. □

On one hand, we have

$$
\begin{aligned}
& 2R_0 + \check{A}_0 + \widetilde{A}_0 + G \\
& \overset{(a)}{\leq} 64\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}} R_0} + 30\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q}_0^* + G + KH\sigma_{\min}^2} + 1174\widehat{\beta}_K^2 \dim_{\mathcal{F}} \\
& \quad + 4\sqrt{2\zeta R_0} + 6\sqrt{\zeta G} + G \\
& \overset{(b)}{=} \widetilde{O}\left(\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q}_0^* + KH\sigma_{\min}^2} + \widehat{\beta}_K^2 \dim_{\mathcal{F}} \right),
\end{aligned}
\tag{F.36}
$$

where (a) holds due to (F.32) and (F.29), while we utilize $\zeta \leq \widehat{\beta}_K^2 \dim_{\mathcal{F}}$ (which is trivial since ζ only consists of logarithmic terms), $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ for any $x \geq 0$ and (F.19) for (b). On the other hand, we have

$$
\begin{aligned}
& 2R_0 + \check{A}_0 + \widetilde{A}_0 + G \\
& \overset{(a)}{\leq} 64\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_1^* K + \check{A}_0} + 60\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 2348\widehat{\beta}_K^2 \dim_{\mathcal{F}} \\
& \quad + 2\sqrt{2\zeta R_0} + 3\sqrt{\zeta G} + 7\zeta + 4\sqrt{2\zeta R_0} + 6\sqrt{\zeta G} + G \\
& \overset{(b)}{=} \widetilde{O}\left(\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_1^* K + KH\sigma_{\min}^2} + \widehat{\beta}_K^2 \dim_{\mathcal{F}} \right),
\end{aligned}
\tag{F.37}
$$

where (a) holds due to (F.33), (F.27) and (F.29), while we utilize $\zeta \leq \widehat{\beta}_K^2 \dim_{\mathcal{F}}$, $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ for any $x \geq 0$ and (F.19) in (b). Combining (F.34), (F.35), (F.36) and (F.37), we have

$$\text{Regret}(K) = \widetilde{O}\left(\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\min\{\mathcal{Q}_0^*, V_1^* K\} + KH\sigma_{\min}^2} + \widehat{\beta}_K^2 \dim_{\mathcal{F}} \right).$$

Recall that according to (F.2),

$$\widehat{\beta}_K = 3\sqrt{\iota_K} + 2\frac{\iota_K}{\gamma^2} + \sqrt{\lambda} + \sqrt{6kH\epsilon/\sigma_{\min}^2}$$

with $\iota_K = \widetilde{O}(\log \mathcal{N}_{\mathcal{F}})$. Moreover, setting $\lambda = \log \mathcal{N}_{\mathcal{F}}, \sigma_{\min}^2 = \sqrt{\dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}}}/(KH), \gamma^2 = \sqrt{\log \mathcal{N}_{\mathcal{F}}}, \epsilon = \log \mathcal{N}_{\mathcal{F}} \sigma_{\min}^2/(KH)$, we have $\widehat{\beta}_K = \widetilde{O}(\sqrt{\log \mathcal{N}_{\mathcal{F}}})$ and $KH\sigma_{\min}^2 = \sqrt{\dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}}}$, which yields a high-probability regret bound

$$\text{Regret}(K) = \widetilde{O}\left(\sqrt{\dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}}} \cdot \sqrt{\min\{\mathcal{Q}_0^*, V_1^* K\}} + \dim_{\mathcal{F}} \log \mathcal{N}_{\mathcal{F}} \right).$$

Notice $\mathcal{Q}^* = \min\{\mathcal{Q}_0^*, V_1^* K\}$, then the proof is completed. □

G   MISSING PROOFS IN SECTION F

G.1   Proof of Lemma F.1

Proof of Lemma F.1. Recall $D_{k,m} = \{z_{i,h,m}, \bar{\sigma}_{i,h,m}\}_{(i,h)\in[k-1]\times[H]}$ and $D_{k,h,m} := D_{k,m} \cup \{z_{k,j,m}, \bar{\sigma}_{k,j,m}\}_{j\in[h]}$. If $f_* \in \widehat{\mathcal{B}}_{k,m}$, it follows that

$$|\widehat{f}_{k,m}(z_{k,h,m}) - f_*(z_{k,h,m})|$$
$$\stackrel{(a)}{\leq} \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\sqrt{\sum_{i=1}^{k-1}\sum_{h=1}^{H}\frac{1}{\bar{\sigma}_{i,h,m}^2}(\widehat{f}_{k,h,m}(z_{i,h,m}) - f_*(z_{i,h,m}))^2 + \lambda}$$
$$\stackrel{(b)}{\leq} (\widehat{\beta}_k + \sqrt{\lambda})\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})$$
$$\stackrel{(c)}{\approx} \widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m}),$$

where (a) holds due to the definition of $\mathcal{D}_{\mathcal{F}}$ in Definition 3.2, (b) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, (c) holds due to $\sqrt{\lambda} = O(\widehat{\beta}_k)$, therefore we omit the term $\sqrt{\lambda}$ in the proof for simplicity. Furthermore, we have

$$V_{k,h}(s_h^k) - r_h(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)$$
$$= \min\{1, r_h(s_h^k, a_h^k) + \widehat{f}_{k,m}(z_{k,h,0}) + \widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\} - r_h(s_h^k, a_h^k) - f_*(z_{k,h,0})$$
$$\leq \min\{1, 2\widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\}.$$

And

$$|[\bar{\mathbb{V}}_{k,h}V_{k,h+1}^{2^m}](s_h^k, a_h^k) - [\mathbb{V}_{k,h}V_{k,h+1}^{2^m}](s_h^k, a_h^k)|$$
$$= |[f_{k,m+1}(z_{k,h,m+1}) - f_{k,m}^2(z_{k,h,m})] - [f_*(z_{k,h,m+1}) - f_*^2(z_{k,h,m})]|$$
$$\leq |f_{k,m+1}(z_{k,h,m+1}) - f_*(z_{k,h,m+1})| + |f_{k,m}^2(z_{k,h,m}) - f_*^2(z_{k,h,m})|$$
$$= |f_{k,m+1}(z_{k,h,m+1}) - f_*(z_{k,h,m+1})| + |f_{k,m}(z_{k,h,m}) + f_*(z_{k,h,m})| \cdot |f_{k,m}(z_{k,h,m}) - f_*(z_{k,h,m})|$$
$$\leq \min\{1, \widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,m+1}; D_{k,m+1})\} + \min\{1, 2\widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\}$$
$$= E_{k,h,m},$$

where the last inequality holds due to $f_{k,m}(z_{k,h,m}), f_*(z_{k,h,m}) \in [0,1]$. □

G.2   Proof of Lemma F.2

Proof of Lemma F.2. We will prove the statement by Theorem 4.4 and induction. For each $m \in \overline{[M-1]}$, denote $\xi_{k,h,m} := f_*(z_{k,h,m}) + \mathbb{1}\{f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1}\}[y_{k,h,m} - f_*(z_{k,h,m})]$, and $\xi_{k,h,M} := y_{k,h,M}$. Then we have for all $m \in \overline{[M]}$, $\xi_{k,h,m} \in \mathcal{G}_{k,h}$, $\mathbb{E}[\xi_{k,h,m}|\mathcal{G}_{k,h}] = f_*(z_{k,h,m})$. Recall the definition of $\bar{\sigma}_{k,h,m}$ in (4) of Algorithm 2, for $m \in \overline{[M-1]}$, we have

$$\bar{\sigma}_{k,h,m}^{-2}\text{Var}[\xi_{k,h,m}|\mathcal{G}_{k,h,m}] = \bar{\sigma}_{k,h,m}^{-2}\mathbb{1}\{f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1}\}\text{Var}[y_{k,h,m}|\mathcal{G}_{k,h,m}]$$
$$= \bar{\sigma}_{k,h,m}^{-2}\mathbb{1}\{f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1}\}[\mathbb{V}V_{k,h+1}^{2^m}](s_h^k, a_h^k)$$
$$= \bar{\sigma}_{k,h,m}^{-2}\mathbb{1}\{f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1}\}[[\mathbb{P}V_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2(s_h^k, a_h^k)]$$
$$= \bar{\sigma}_{k,h,m}^{-2}\mathbb{1}\{f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1}\}[f_*(z_{k,h,m+1}) - f_*^2(z_{k,h,m})]$$
$$\leq \bar{\sigma}_{k,h,m}^{-2}[f_{k,m+1}(z_{k,h,m+1}) - f_{k,m}^2(z_{k,h,m}) + \min\{1, \widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,m+1}; D_{k,m+1})$$
$$\qquad \min\{1, 2\widehat{\beta}_k\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\}\}]$$
$$= \bar{\sigma}_{k,h,m}^{-2}[[\bar{\mathbb{V}}_{k,m}V_{k,h+1}^{2^m}](s_h^k, a_h^k) + E_{k,h,m}]$$
$$\leq 1,$$

where the first inequality holds due to Lemma F.1 and $f_{k,m}(z_{k,h,m}) \in [0,1]$. And

$$\bar{\sigma}_{k,h,M}^{-2} \operatorname{Var}[\xi_{k,h,M}|\mathcal{G}_{k,h,m}] \leq \bar{\sigma}_{k,h,M}^{-2} \leq 1.$$

Furthermore, for all $m \in \overline{[M]}$, we have

$$\bar{\sigma}_{k,h,m}^{-2} \mathcal{D}_{\mathcal{F}}^2(z_{k,h,m}; D_{k,m}) \leq 1/\gamma^2.$$

For each $m \in \overline{[M]}$, we define

$$\widetilde{f}_{k+1,m} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{h=1}^{H} (f(z_{i,h,m}) - \xi_{i,h,m})^2. \tag{G.1}$$

Applying Theorem 4.4 using $X_t = z_{k,h,m}$, $Y_t = \xi_{k,h,m}$, $w_t = 1/\bar{\sigma}_{k,h,m}$, together with a union bound, with probability at least $1 - (M+1)\delta$, we have for all $m \in \overline{[M]}, k \in [K]$,

$$\sum_{i=1}^{k-1} \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{i,h,m}^2} (\widetilde{f}_{k,h,m}(z_{i,h,m}) - f_*(z_{i,h,m}))^2 \leq 3\sqrt{\iota_k} + 2\frac{\iota_k}{\gamma^2} + \sqrt{\lambda} + \sqrt{6kH\epsilon/\sigma_{\min}^2} = \widehat{\beta}_k^2, \tag{G.2}$$

where $\iota_k = 16 \log \frac{2\mathcal{N}_{\mathcal{F}}(\epsilon)k^2 H^2 (\log(kH/\sigma_{\min}^2)+2)(\log(1/\sigma_{\min}^2)+2)}{\delta}$.

We continue the proof by induction over $(k,m) \in [K] \times \overline{[M]}$. First, for $k = 1$, $m \in \overline{[M]}$, the result holds trivially.

Next, for $k \in [K], m = M$, we have $\xi_{k,h,M} = y_{k,h,M}$, therefore $\widetilde{f}_{k,M} = \widehat{f}_{k,M}$ by (G.1), which implies $\widehat{f}_{k,M} \in \widehat{\mathcal{B}}_{k,M}$ by (G.2).

Last, for all $k \in [k], m \in \overline{[M-1]}$, we have the following observations:

$$
\begin{aligned}
&f_* \in \widehat{\mathcal{B}}_{k,m} \cap \widehat{\mathcal{B}}_{k,m+1} \\
&\implies \xi_{k,h,m} = y_{k,h,m} \text{ for all } h \in [H] \\
&\overset{\text{(G.1)}}{\implies} \widetilde{f}_{k+1,m} = \widehat{f}_{k+1,m} \\
&\overset{\text{(G.2)}}{\implies} \widehat{f}_{k+1,m} \in \widehat{\mathcal{B}}_{k+1,m}.
\end{aligned} \tag{G.3}
$$

For any $k \in K$, we assume $\widehat{f}_{k',m} \in \widehat{\mathcal{B}}_{k',m}$ for all $k' \leq k, m \in \overline{[M]}$. Notice $\widehat{f}_{k+1,M} \in \widehat{\mathcal{B}}_{k+1,M}$, using (G.3), we have $\widehat{f}_{k+1,m} \in \widehat{\mathcal{B}}_{k+1,m}$ for all $m \in \overline{[M]}$. Then the proof is completed by induction. □

G.3 Proof of Lemma F.3

Proof of Lemma F.3. We prove the optimism by induction. When $h = H+1$, we have $V_{k,H+1}(\cdot) = V_{H+1}^*(\cdot) = 0$, and the result holds trivially. We assume the statement is true for all $h+1$, and prove the case of h. For any (s,a), if $Q_{k,h}(s,a) = 1$, then $Q_{k,h}(s,a) = 1 \geq Q_h^*(s,a)$. Otherwise, we have

$$
\begin{aligned}
Q_{k,h}(s,a) - Q_h^*(s,a) &= f_{k,0}(s,a,V_{k,h+1}) - f_*(s,a,V_{h+1}^*) + \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(s,a,V_{k,h+1}; D_{k,0}) \\
&\geq f_{k,0}(s,a,V_{k,h+1}) - f_*(s,a,V_{k,h+1}) + \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(s,a,V_{k,h+1}; D_{k,0}) \\
&\geq 0,
\end{aligned}
$$

where the first inequality holds due to $V_{k,h+1}(\cdot) \geq V_{h+1}^*(\cdot)$ and the second holds due to Lemma F.1. That is, we have $Q_{k,h}(\cdot,\cdot) \geq Q_h^*(\cdot,\cdot)$ and therefore $V_{k,h}(\cdot) \geq V_h^*(\cdot)$. Then the proof is completed by induction. □

G.4 Proof of Lemma F.4

Proof of Lemma F.4. Recall the definition of I_h^k, we have

$$1 - I_{H+1}^k = 1 \Leftrightarrow \exists m \in \overline{[M]}, \ \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{k,h,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,h,m}; D_{k,h-1,m}) > 1.$$

Let $\mathcal{D}_m$ denote the indices k such that

$$\mathcal{D}_m := \left\{ k \in [K] : \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{k,h,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,h,m}; D_{k,h-1,m}) > 1 \right\}.$$

Then we have $G \leq |\bigcup_{m=0}^{M} \mathcal{D}_m| \leq \sum_{m=0}^{M} |\mathcal{D}_m|$. For any $m \in \overline{[M]}$, we have

$$\begin{aligned}
|\mathcal{D}_m| &\leq \sum_{k=1}^{K} \min\left\{ 1, \sum_{h=1}^{H} \frac{1}{\bar{\sigma}_{k,h,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,h,m}; D_{k,h-1,m}) \right\} \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \min\left\{ 1, \frac{1}{\bar{\sigma}_{k,h,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,h,m}; D_{k,h-1,m}) \right\} \\
&\leq \dim_{\mathcal{F}}.
\end{aligned}$$

Taking the summation over $m \in \overline{[M]}$ gives the upper bound of G. $\qquad\square$

G.5   Proof of Lemma F.5

Proof of Lemma F.5. We are to bound $\check{S}_m, \widetilde{S}_m$ and S_m separately with similar arguments.

Bound $\check{S}_m$ Recall the definition of $\check{S}_m$ in (F.8), we have

$$\begin{aligned}
\check{S}_m &= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k [\mathbb{V} \check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) \\
&= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P} \check{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k)^2 \right] \\
&= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P} \check{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\
&\quad + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left(\check{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) - \check{V}_{k,h}^{2^{m+1}}(s_h^k) \right) \\
&\leq \check{A}_{m+1} + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] + \sum_{k=1}^{K} (1 - I_{h_k}^k) \check{V}_{k,h_k}^{2^{m+1}}(s_{h_k}^k) \\
&\leq \check{A}_{m+1} + G + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right], \qquad\qquad (G.4)
\end{aligned}$$

where h_k is defined in (F.4) and the last inequality holds since $|\check{V}_{k,h}(s_h^k)| \leq 1$ and I_h^k is monotonically decreasing. For the third term in (G.4), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]\\
&\overset{(a)}{\leq} \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}]^{2^{m+1}}(s_h^k, a_h^k) \right]\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\check{V}_{k,h}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) \right] \prod_{i=0}^{m} \left[\check{V}_{k,h}^{2^i}(s_h^k) + [\mathbb{P}\check{V}_{k,h+1}]^{2^i}(s_h^k, a_h^k) \right]\\
&\leq 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\left\{ \check{V}_{k,h}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k), 0 \right\}\\
&\overset{(b)}{\leq} 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\left\{ V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k), 0 \right\}\\
&\overset{(c)}{\leq} 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \cdot 2\min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\}\\
&= 2^{m+1}\cdot(2R_0),
\end{aligned}
\tag{G.5}
$$

where (a) holds due to $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, (b) holds due to the definition of $\check{V}_{k,h}$ and $V_h^*(s_h^k) \geq r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)$, while (c) is due to Lemma F.1. Substituting (G.5) into (G.4), we have

$$\check{S}_m \leq \check{A}_{m+1} + G + 2^{m+1}\cdot(2R_0).$$

Bound $\widetilde{S}_m$ Recall the definition of $\widetilde{S}_m$ in (F.9), we have

$$
\begin{aligned}
\widetilde{S}_m &= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k [\mathbb{V}\widetilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k)\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\widetilde{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k)^2 \right]\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\widetilde{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - \widetilde{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) \right] + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]\\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left(\widetilde{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) - \widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) \right)\\
&\leq \widetilde{A}_{m+1} + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] + \sum_{k=1}^{K}(1 - I_{h_k}^k)\widetilde{V}_{k,h_k}^{2^{m+1}}(s_{h_k}^k)\\
&\leq \widetilde{A}_{m+1} + G + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right],
\end{aligned}
\tag{G.6}
$$

where h_k is defined in (F.4) and the last inequality holds since $|\widetilde{V}_{k,h}(s_h^k)| \leq 1$ and I_h^k is monotonically decreasing. For the third term in (G.6), we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]$$

$$\overset{(a)}{\leq} \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}]^{2^{m+1}}(s_h^k, a_h^k) \right]$$

$$= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k) \right] \prod_{i=0}^{m} \left[\widetilde{V}_{k,h}^{2^i}(s_h^k) + [\mathbb{P}\widetilde{V}_{k,h+1}]^{2^i}(s_h^k, a_h^k) \right]$$

$$\leq 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \max \left\{ \widetilde{V}_{k,h}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k), 0 \right\}$$

$$\overset{(b)}{=} 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \max \left\{ V_h^*(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k), 0 \right\}$$

$$\overset{(c)}{\leq} 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \max \left\{ V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k), 0 \right\} + |[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k)|$$

$$\overset{(d)}{\leq} 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[2 \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\} + |[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k)| \right]$$

$$\leq 2^{m+1} \cdot (2R_0 + \check{A}_0), \tag{G.7}$$

where (a) holds due to $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, (b) holds due to the definition of $\widetilde{V}_{k,h}$ and $V_h^{\pi^k}(s_h^k) = r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k)$, (c) holds due to $V_h^*(s_h^k) \geq r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)$ and the definition of $\check{V}_{k,h}$, while (d) is due to Lemma F.1. Substituting (G.7) into (G.6), we have

$$\widetilde{S}_m \leq \widetilde{A}_{m+1} + G + 2^{m+1} \cdot (2R_0 + \check{A}_0).$$

Bound S_m Recall the definition of S_m in (F.10), we have

$$S_m = \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k [\mathbb{V}V_{k,h+1}^{2^m}](s_h^k, a_h^k)$$

$$= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]$$

$$= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - V_{k,h+1}^{2^{m+1}}(s_{h+1}^k) \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[V_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]$$

$$+ \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left(V_{k,h+1}^{2^{m+1}}(s_{h+1}^k) - V_{k,h}^{2^{m+1}}(s_h^k) \right)$$

$$\leq A_{m+1} + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[V_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] + \sum_{k=1}^{K} (1 - I_{h_k}^k) V_{k,h_k}^{2^{m+1}}(s_{h_k}^k)$$

$$\leq A_{m+1} + G + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[V_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right], \tag{G.8}$$

where h_k is defined in (F.4) and the last inequality holds since $|V_{k,h}(s_h^k)| \leq 1$ and I_h^k is monotonically decreasing. For the third term in (G.8), we have

$$\sum_{k=1}^K \sum_{h=1}^H I_h^k \left[V_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}]^2 (s_h^k, a_h^k) \right]$$

$$\overset{(a)}{\leq} \sum_{k=1}^K \sum_{h=1}^H I_h^k \left[V_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}V_{k,h+1}]^{2^{m+1}}(s_h^k, a_h^k) \right]$$

$$= \sum_{k=1}^K \sum_{h=1}^H I_h^k \left[V_{k,h}(s_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \right] \prod_{i=0}^m \left[V_{k,h}^{2^i}(s_h^k) + [\mathbb{P}V_{k,h+1}]^{2^i}(s_h^k, a_h^k) \right]$$

$$\leq 2^{m+1} \sum_{k=1}^K \sum_{h=1}^H I_h^k \max \left\{ V_{k,h}(s_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k), 0 \right\}$$

$$\overset{(b)}{\leq} 2^{m+1} \sum_{k=1}^K \sum_{h=1}^H I_h^k \left[r(s_h^k, a_h^k) + 2\min\{1, \widehat{\beta}_k \mathcal{D}_\mathcal{F}(z_{k,h,0}; D_{k,0})\} \right]$$

$$\leq 2^{m+1} \cdot (V_0 + 2R_0), \tag{G.9}$$

where (a) holds due to $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ and (b) holds due to Lemma F.1. Substituting (G.9) into (G.8), we have

$$S_m \leq A_{m+1} + G + 2^{m+1} \cdot (V_0 + 2R_0).$$

□

G.6 Proof of Lemma F.6

Proof of Lemma F.6. Let $X_{k,h} = I_h^k \left[[\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^m}(s_{h+1}^k) \right]$, then we have $\mathbb{E}[X_{k,h}|\mathcal{G}_{k,h}] = 0$, $|X_{k,h}| \leq 2$ and $\mathbb{E}[X_{k,h}^2|\mathcal{G}_{k,h}] = I_h^k[\mathbb{V}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k)$. Therefore, for any $m \in \overline{[M]}$, applying variance-aware Freedman's inequality in Lemma H.2, with probability at least $1 - \delta$, we have

$$\check{A}_m \leq \sqrt{\zeta \check{S}_m} + \zeta.$$

Thus, taking a union bound over $m \in \overline{[M]}$, with probability at least $1 - (M+1)\delta$, (F.23) holds. The proofs for (F.24) and (F.25) follow the same arguments as (F.23). □

G.7 Proof of Lemma F.7

Proof of Lemma F.7. On event $\mathcal{E} \cup \mathcal{A}$, we have (F.20) and (F.23) hold by Lemma F.5 and F.6. Substituting the bound of $\check{S}_m$ in (F.20) into (F.23), we have for all $m \in \overline{[M]}$,

$$\check{A}_m \leq \sqrt{\zeta} \cdot \sqrt{\check{A}_{m+1} + G + 2^{m+1} \cdot (2R_0)} + \zeta.$$

And we have for all $m \in \overline{[M]}$, $\check{A}_m \leq 2KH$. Then the result follows by Lemma H.6. □

G.8 Proof of Lemma F.8

Proof of Lemma F.8. On event $\mathcal{E} \cup \mathcal{A}$, we have (F.21) and (F.24) hold by Lemma F.5 and F.6. Substituting the bound of $\widetilde{S}_m$ in (F.21) into (F.24), we have for all $m \in \overline{[M]}$,

$$\widetilde{A}_m \leq \sqrt{\zeta} \cdot \sqrt{\widetilde{A}_{m+1} + G + 2^{m+1} \cdot (2R_0 + \check{A}_0)} + \zeta.$$

And we have for all $m \in \overline{[M]}$, $\widetilde{A}_m \leq 2KH$. Applying Lemma H.6, we have

$$
\begin{aligned}
\widetilde{A}_0 &\leq 2\sqrt{\zeta(2R_0 + \check{A}_0)} + 3\sqrt{\zeta G} + 7\zeta \\
&\overset{(a)}{\leq} 2\sqrt{2\zeta R_0} + 2\sqrt{\zeta \check{A}_0} + 3\sqrt{\zeta G} + 7\zeta \\
&\overset{(b)}{\leq} 2\sqrt{2\zeta R_0} + \zeta + \check{A}_0 + 3\sqrt{\zeta G} + 7\zeta \\
&\overset{(c)}{\leq} 4\sqrt{2\zeta R_0} + 6\sqrt{\zeta G} + 15\zeta,
\end{aligned}
$$

where (a) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, (b) holds due to $2\sqrt{ab} \leq a + b$ for $a, b \geq 0$ and (c) holds due to (F.27) in Lemma F.7. $\qquad\square$

G.9   Proof of Lemma F.9

Proof of Lemma F.9. Recall the definition of V_0 and V_1^* in (F.16) and (F.17), we have

$$
\begin{aligned}
V_0 - V_1^* K &= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k r(s_h^k, a_h^k) - \sum_{k=1}^{K} V_1^*(s_h^k) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[r(s_h^k, a_h^k) - V_h^*(s_h^k) + V_{h+1}^*(s_{h+1}^k) \right] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[V_{h+1}^*(s_{h+1}^k) - [\mathbb{P}_h V_{h+1}^*](s_h^k, a_h^k) \right] \\
&\leq A_0 + \check{A}_0,
\end{aligned}
$$

where the second inequality holds due to $V_h^*(s_h^k) \geq r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)$. $\qquad\square$

G.10   Proof of Lemma F.10

Proof of Lemma F.10. First, for all $m \in \overline{[M]}$, $R_m \leq KH$ holds trivially. For (k, h) where $I_h^k = 1$, we have $\sum_{j=1}^{h-1} \frac{1}{\bar{\sigma}_{k,j,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,j,m}; D_{k,j-1,m}) \leq 1$. By Lemma H.4, it follows that

$$
\begin{aligned}
&\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m}) \\
&\leq \exp\left\{ \frac{1}{2}\sum_{j=1}^{h-1} \frac{1}{\bar{\sigma}_{k,j,m}^2} \mathcal{D}_{\mathcal{F}}^2(z_{k,j,m}; D_{k,j-1,m}) \right\} \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,h-1,m}) \\
&\leq 2\mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,h-1,m}).
\end{aligned}
$$

Then we have

$$
R_m \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, I_h^k \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,h-1,m})\},
$$

which can be bounded by Lemma H.5, with $\beta_t = I_h^k \widehat{\beta}_k$, $\bar{\sigma}_t = \bar{\sigma}_{k,h,m}$ and $X_t = z_{k,h,m}$. We have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, I_h^k \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,h-1,m})\} \\
&\leq \dim_{\mathcal{F}} + \widehat{\beta}_K \gamma^2 \dim_{\mathcal{F}} + \widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[[\bar{\mathbb{V}}_{k,m} V_{k,h+1}^{2m}](s_h^k, a_h^k) + E_{k,h,m} \right] + KH\sigma_{\min}^2} \\
&\leq \dim_{\mathcal{F}} + \widehat{\beta}_K \gamma^2 \dim_{\mathcal{F}} + \widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[[\mathbb{V} V_{k,h+1}^{2m}](s_h^k, a_h^k) + 2E_{k,h,m} \right] + KH\sigma_{\min}^2},
\end{aligned}
$$

where the last inequality is due to Lemma F.1. Note that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k E_{k,h,m} = \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \min\left\{1, 2\widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m}; D_{k,m})\right\}$$
$$+ \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \min\left\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,m+1}; D_{k,m+1})\right\}$$
$$\le 2R_m + R_{m+1},$$

where the last inequality holds by the definition of R_m in (F.18). Thus, we have

$$R_m \le 2\dim_{\mathcal{F}} + 2\widehat{\beta}_K \gamma^2 \dim_{\mathcal{F}} + 2\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{S_m + 4R_m + 2R_{m+1} + KH\sigma_{\min}^2}$$
$$\overset{(a)}{\le} 16\widehat{\beta}_K^2 \dim_{\mathcal{F}} + 4\dim_{\mathcal{F}} + 4\widehat{\beta}_K\gamma^2 \dim_{\mathcal{F}} + 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{S_m + 2R_{m+1} + KH\sigma_{\min}^2}$$
$$\overset{(b)}{\le} 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{S_m + 2R_{m+1} + KH\sigma_{\min}^2} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}},$$

where (a) holds since $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$ and $x \le a\sqrt{x} + b$ implies $x \le a^2 + 2b$ for any $x \ge 0$, while (b) holds due to $\widehat{\beta}_K, \dim_{\mathcal{F}} \ge 1$ and $\gamma^2 \le \widehat{\beta}_K$. $\qquad\square$

G.11 Proof of Lemma F.11

Proof of Lemma F.11. On event $\mathcal{E} \cup \mathcal{A}$, we have (F.31) holds by Lemma F.10. And we have for all $m \in \overline{[M]}$,

$$S_m \le 2Q_m + 2\check{S}_m, \tag{G.10}$$

where the inequality holds due to $\mathrm{Var}[X+Y] \le 2\mathrm{Var}[X] + 2\mathrm{Var}[Y]$. Substituting the bound of S_m in (G.10) into (F.31), we have for all $m \in \overline{[M-1]}$,

$$R_m \le 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{2Q_m + 2\check{S}_m + 2R_{m+1} + KH\sigma_{\min}^2} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{G.11}$$

Next, on event $\mathcal{E} \cup \mathcal{A}$, we have (F.20) holds by Lemma F.5. Substituting the bound of $\check{S}_m$ in (F.20) into (G.11), we have

$$R_m \le 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{2R_{m+1} + 2\check{A}_{m+1} + 2Q_m + 2G + KH\sigma_{\min}^2 + 2^{m+1} \cdot (4R_0)} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}}$$
$$\le 4\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{R_{m+1} + \check{A}_{m+1} + 2^{m+1} \cdot (2R_0)}$$
$$+ 4\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q} + G + KH\sigma_{\min}^2} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}}, \tag{G.12}$$

where the last inequality holds due to the inequality that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$, and the definition of $\mathcal{Q}$ in (F.12). Then, on event $\mathcal{E} \cup \mathcal{A}$, we have (F.26) holds by Lemma F.7. Add (F.26) to (G.12), using $\sqrt{a} + \sqrt{b} \le \sqrt{2} \cdot \sqrt{a+b}$ for any $a, b \ge 0$ and assume $\zeta \le \widehat{\beta}_K^2 \dim_{\mathcal{F}}$ (which is trivial since ζ only consists of logarithmic terms), we have for all $m \in \overline{[M-1]}$,

$$R_m + \check{A}_m \le 8\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{R_{m+1} + 2\check{A}_{m+1} + 2^{m+1} \cdot (4R_0)}$$
$$+ 5\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q} + G + KH\sigma_{\min}^2} + 25\widehat{\beta}_K^2 \dim_{\mathcal{F}}$$
$$\le 8\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{R_{m+1} + \check{A}_{m+1} + 2^{m+1} \cdot (2R_0)}$$
$$+ 5\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{\mathcal{Q} + G + KH\sigma_{\min}^2} + 25\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{G.13}$$

And for all $m \in \overline{[M]}$, $R_m + \check{A}_m \le 3HK$. Then the result follows by Lemma H.6. $\qquad\square$

G.12   Proof of Lemma F.12

Proof of Lemma F.12. On event $\mathcal{E} \cup \mathcal{A}$, we have (F.22), (F.25) and (F.31) holds by Lemma F.5, F.6 and F.10. Substituting the bound of S_m in (F.22) into (F.25) and (F.31) respectively, we have for all $m \in \overline{[M-1]}$,

$$A_m \leq \sqrt{\zeta} \cdot \sqrt{A_{m+1} + G + 2^{m+1} \cdot (V_0 + 2R_0)} + \zeta, \tag{G.14}$$

$$\begin{aligned}
R_m &\leq 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{2R_{m+1} + A_{m+1} + G + KH\sigma_{\min}^2 + 2^{m+1} \cdot (V_0 + 2R_0)} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}} \\
&\leq 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{2R_{m+1} + A_{m+1} + 2^{m+1} \cdot (V_0 + 2R_0)} \\
&\quad + 4\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 24\widehat{\beta}_K^2 \dim_{\mathcal{F}}.
\end{aligned} \tag{G.15}$$

Add (G.14) to (G.15), using $\sqrt{a} + \sqrt{b} \leq \sqrt{2} \cdot \sqrt{a+b}$ for any $a, b \geq 0$ and assume $\zeta \leq \widehat{\beta}_K^2 \dim_{\mathcal{F}}$, we have for all $m \in \overline{[M-1]}$,

$$\begin{aligned}
R_m + A_m &\leq 4\widehat{\beta}_K \sqrt{2\dim_{\mathcal{F}}} \cdot \sqrt{2R_{m+1} + 2A_{m+1} + 2^{m+1} \cdot (2V_0 + 4R_0)} \\
&\quad + 5\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 25\widehat{\beta}_K^2 \dim_{\mathcal{F}} \\
&= 8\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{R_{m+1} + A_{m+1} + 2^{m+1} \cdot (V_0 + 2R_0)} \\
&\quad + 5\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 25\widehat{\beta}_K^2 \dim_{\mathcal{F}}.
\end{aligned} \tag{G.16}$$

And for all $m \in \overline{[M]}$, $R_m + \check{A}_m \leq 2HK$. Applying Lemma H.6, we have

$$R_0 + A_0 \leq 16\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_0 + 2R_0} + 15\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 331\widehat{\beta}_K^2 \dim_{\mathcal{F}}. \tag{G.17}$$

We have (F.30) holds by Lemma F.9. Substituting the bound of V_0 in (F.30) into (G.17), we have

$$\begin{aligned}
R_0 + A_0 &\leq 16\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_1^* K + A_0 + \check{A}_0 + 2R_0} + 15\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 331\widehat{\beta}_K^2 \dim_{\mathcal{F}} \\
&\leq 32\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{V_1^* K + \check{A}_0} + 30\widehat{\beta}_K \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{G + KH\sigma_{\min}^2} + 1174\widehat{\beta}_K^2 \dim_{\mathcal{F}},
\end{aligned}$$

where the last inequality holds since $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ for any $x \geq 0$. $\square$

G.13   Proof of Lemma F.13

Proof of Lemma F.13. First, we decompose $V_{k,1}(s_1^k)$ and $V_1^{\pi^k}(s_1^k)$ as follows

$$\begin{aligned}
V_{k,1}(s_1^k) &= \sum_{h=1}^{H} I_h^k [V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k)] + \sum_{h=1}^{H}(1 - I_h^k)[V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k)] \\
&= \sum_{h=1}^{H} I_h^k r(s_h^k, a_h^k) + \sum_{h=1}^{H} I_h^k \left[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \right] \\
&\quad + \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k) \right] + (1 - I_{h_k}^k)V_{k,h_k}(s_{h_k}^k), \\
V_1^{\pi^k}(s_1^k) &= \sum_{h=1}^{H} I_h^k [V_h^{\pi^k}(s_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] + \sum_{h=1}^{H}(1 - I_h^k)[V_h^{\pi^k}(s_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \\
&= \sum_{h=1}^{H} I_h^k r(s_h^k, a_h^k) + \sum_{h=1}^{H} I_h^k \left[V_h^{\pi^k}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) \right] \\
&\quad + \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \right] + (1 - I_{h_k}^k)V_{h_k}^{\pi^k}(s_{h_k}^k) \\
&= \sum_{h=1}^{H} I_h^k r(s_h^k, a_h^k) + \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \right] + (1 - I_{h_k}^k)V_{h_k}^{\pi^k}(s_{h_k}^k),
\end{aligned}$$

where h_k is defined in (F.4). Thus it follows that

$$
V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)
$$

$$
= \sum_{h=1}^{H} I_h^k \left[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \right] + (1 - I_{h_k}^k)V_{k,h_k}(s_{h_k}^k) - (1 - I_{h_k}^k)V_{h_k}^{\pi^k}(s_{h_k}^k)
$$

$$
+ \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k) \right] - \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \right]
$$

$$
= \sum_{h=1}^{H} I_h^k \left[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \right] + (1 - I_{h_k}^k)V_{k,h_k}(s_{h_k}^k) - (1 - I_{h_k}^k)V_{h_k}^{\pi^k}(s_{h_k}^k)
$$

$$
+ \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k) \right] + \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k) - \widetilde{V}_{k,h+1}(s_{h+1}^k) \right].
$$

Then we have

$$
\sum_{k=1}^{K} \left[V_{k,1}(s_1^k) - V^{\pi^k}(s_1^k) \right]
$$

$$
\overset{(a)}{\leq} 2 \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \min\{1, \widehat{\beta}_k \mathcal{D}_{\mathcal{F}}(z_{k,h,0}; D_{k,0})\} + \sum_{k=1}^{K} (1 - I_{h_k}^k)V_{k,h_k}(s_{h_k}^k)
$$

$$
+ \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k) \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[[\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k) - \widetilde{V}_{k,h+1}(s_{h+1}^k) \right]
$$

$$
\overset{(b)}{\leq} 2R_0 + \check{A}_0 + \widetilde{A}_0 + G,
$$

where (a) is due to Lemma F.1 and (b) holds since $|V_{k,h}(s_h^k)| \leq 1$ and I_h^k is monotonically decreasing. $\qquad\square$

H   AUXILLIARY LEMMAS

Lemma H.1 (Variance-aware and range-aware Freedman's inequality, Corollary 2 in Agarwal et al. (2023)). Let $M \geq m > 0, V \geq v > 0$ be fixed constants, and $\{X_s\}_{s \in [t]}$ be a stochastic process adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$, such that X_s is $\mathcal{G}_s$-measurable. Suppose $\mathbb{E}[X_s|\mathcal{G}_{s-1}] = 0, |X_s| \leq M$ and $\sum_{s=1}^{t} \mathbb{E}[X_s^2|\mathcal{G}_{s-1}] \leq V^2$ almost surely. Then for any $\delta > 0$, with probability at least $1 - (\log(V^2/v^2) + 2)(\log(M/m) + 2)\delta$, we have

$$
\sum_{s=1}^{t} X_s \leq \sqrt{2 \left(2 \sum_{s=1}^{t} \mathbb{E}[X_s^2|\mathcal{G}_{s-1}] + v^2 \right) \log \frac{1}{\delta}} + \frac{2}{3} \left(2 \max_{s \in [t]} |X_s| + m \right) \log \frac{1}{\delta}.
$$

Lemma H.2 (Variance-aware Freedman's inequality). Let $M > 0$ be fixed constants, and $\{X_s\}_{s \in [t]}$ be a stochastic process adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$, such that X_s is $\mathcal{G}_s$-measurable. Suppose $\mathbb{E}[X_s|\mathcal{G}_{s-1}] = 0$ and $|X_s| \leq M$ almost surely. Then for any $\delta > 0$, with probability at least $1 - 2(\log t + 2)\delta$, we have

$$
\sum_{s=1}^{t} X_s \leq 2 \sqrt{\sum_{s=1}^{t} \mathbb{E}[X_s^2|\mathcal{G}_{s-1}] \log \frac{1}{\delta}} + 4M \log \frac{1}{\delta}.
$$

Proof. The result follows by applying Lemma H.1 with $V^2 = M^2 t, m = v = M$. $\qquad\square$

Lemma H.3 (Elliptical Potential Lemma, Lemma 11 in Abbasi-Yadkori et al. (2011)). Let $\{\boldsymbol{x}_t\}_{t \in [T]} \subset \mathbb{R}^d$ and assume $\|\boldsymbol{x}_t\|_2 \leq L$ for all $t \in [T]$. Set $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t} \boldsymbol{x}_t \boldsymbol{x}_t^\top + \lambda \boldsymbol{I}$. Then it follows that

$$
\sum_{t=1}^{T} \min \left\{ 1, \|\boldsymbol{x}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^2 \right\} \leq 2d \log \left(1 + \frac{TL^2}{d\lambda} \right).
$$

Lemma H.4. Let $\mathcal{D}_{\mathcal{F}}$ in Definition 3.2, for any $t > t_0 \geq 1$, we have

$$\mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t_0]}, \sigma_{[t_0]}) \leq \exp\left\{ \sum_{s=t_0+1}^{t-1} \frac{1}{\sigma_s^2} \mathcal{D}_{\mathcal{F}}^2(X_s; X_{[s-1]}, \sigma_{[s-1]}) \right\} \mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]}).$$

Proof. Note

$$
\begin{aligned}
&\mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]}) \\
=\ & \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \lambda} \\
\geq\ & \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{(1 + \frac{1}{\sigma_{t-1}^2}\mathcal{D}_{\mathcal{F}}^2(X_{t-1}; X_{[t-2]}, \sigma_{[t-2]})) \sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \lambda} \\
\geq\ & \frac{1}{1 + \frac{1}{\sigma_{t-1}^2}\mathcal{D}_{\mathcal{F}}^2(X_{t-1}; X_{[t-2]}, \sigma_{[t-2]})} \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(X_t) - f_2(X_t))^2}{\sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \lambda} \\
=\ & \frac{1}{1 + \frac{1}{\sigma_{t-1}^2}\mathcal{D}_{\mathcal{F}}^2(X_{t-1}; X_{[t-2]}, \sigma_{[t-2]})} \mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-2]}, \sigma_{[t-2]}) \geq \dots \\
\geq\ & \frac{1}{\prod_{s=t_0+1}^{t-1}(1 + \frac{1}{\sigma_s^2}\mathcal{D}_{\mathcal{F}}^2(X_s; X_{[s-1]}, \sigma_{[s-1]}))} \mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t_0]}, \sigma_{[t_0]}),
\end{aligned}
$$

where the first inequality holds due to

$$
\begin{aligned}
&\sum_{s=1}^{t-1} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 \\
=\ & \sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \frac{1}{\sigma_{t-1}^2}(f_1(X_{t-1}) - f_2(X_{t-1}))^2 \\
\leq\ & \sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 + \frac{1}{\sigma_{t-1}^2}\mathcal{D}_{\mathcal{F}}^2(X_{t-1}; X_{[t-2]}, \sigma_{[t-2]}) \sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2 \\
=\ & \left(1 + \frac{1}{\sigma_{t-1}^2}\mathcal{D}_{\mathcal{F}}^2(X_{t-1}; X_{[t-2]}, \sigma_{[t-2]})\right) \sum_{s=1}^{t-2} \frac{1}{\sigma_s^2}(f_1(X_s) - f_2(X_s))^2.
\end{aligned}
$$

Thus, we have

$$\mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t_0]}, \sigma_{[t_0]}) \leq \prod_{s=t_0+1}^{t-1} \left(1 + \frac{1}{\sigma_s^2}\mathcal{D}_{\mathcal{F}}^2(X_s; X_{[s-1]}, \sigma_{[s-1]})\right) \mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})$$

$$\leq \exp\left\{ \sum_{s=t_0+1}^{t-1} \frac{1}{\sigma_s^2} \mathcal{D}_{\mathcal{F}}^2(X_s; X_{[s-1]}, \sigma_{[s-1]}) \right\} \mathcal{D}_{\mathcal{F}}^2(X_t; X_{[t-1]}, \sigma_{[t-1]}),$$

where the second inequality holds due to the inequality $1 + x \leq \exp\{x\}$. $\qquad\square$

Lemma H.5. Let $\{\sigma_t, \beta_t\}_{t \geq 1}$ be a sequence of non-negative numbers, $\sigma_{\min}, \gamma, \lambda > 0$, $\{X_t\}_{t \geq 1} \subset \mathcal{X}$ and $\{\bar{\sigma}_k\}_{k \geq 1}$ be recursively defined:

$$\bar{\sigma}_t^2 = \max\{\sigma_t^2, \sigma_{\min}^2, \gamma^2 \mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]})\}.$$

Then we have

$$\sum_{t=1}^{T} \min\{1, \beta_t \mathcal{D}_{\mathcal{F}}(X_t; X_{[t-1]}, \sigma_{[t-1]})\} \leq \dim_{\mathcal{F}} + \max_{t \in [T]} \beta_t \cdot \gamma^2 \dim_{\mathcal{F}} + \sqrt{\dim_{\mathcal{F}}} \cdot \sqrt{\sum_{t=1}^{T} \beta_t^2(\sigma_t^2 + \sigma_{\min}^2)},$$

where $\mathcal{D}_{\mathcal{F}}$ and $\dim_{\mathcal{F}} = \dim_{\mathcal{F}}(\sigma_{\min}, T)$ are in Definition 3.2.

Proof. We decompose $[T]$ as the union of three disjoint sets $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$:

$$\mathcal{J}_1 = \left\{t \in [T] \Big| \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]}) > 1\right\},$$

$$\mathcal{J}_2 = \left\{t \in [T] \Big| \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \leq 1, \bar{\sigma}_t \in \{\sigma_t, \sigma_{\min}\}\right\},$$

$$\mathcal{J}_3 = \left\{t \in [T] \Big| \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \leq 1, \bar{\sigma}_t = \gamma \sqrt{\mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})}\right\}.$$

For the summation over $\mathcal{J}_1$, we have

$$\sum_{t \in \mathcal{J}_1} \min\{1, \beta_t \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})\} \leq |\mathcal{J}_1| \leq \sum_{t \in \mathcal{J}_1} \min\left\{1, \frac{1}{\bar{\sigma}_t^2} \mathcal{D}_\mathcal{F}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})\right\} \leq \dim_\mathcal{F}.$$

Next, for the summation over $\mathcal{J}_2$, we have

$$\sum_{t \in \mathcal{J}_2} \min\{1, \beta_t \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})\}$$

$$\leq \sum_{t \in \mathcal{J}_2} \beta_t \bar{\sigma}_t \cdot \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})\}$$

$$\leq \sum_{t=1}^T \beta_t \max\{\sigma_t, \sigma_{\min}\} \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})\}$$

$$\overset{(a)}{\leq} \sqrt{\sum_{t=1}^T \beta_t^2(\sigma_t^2 + \sigma_{\min}^2)} \cdot \sqrt{\sum_{t=1}^T \min\left\{1, \frac{1}{\bar{\sigma}_t^2} \mathcal{D}_\mathcal{F}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})\right\}}$$

$$\leq \sqrt{\dim_\mathcal{F}} \cdot \sqrt{\sum_{t=1}^T \beta_t^2(\sigma_t^2 + \sigma_{\min}^2)},$$

where (a) holds due to Cauchy-Schwartz inequality and $\frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \leq 1$. Then, for the summation over $\mathcal{J}_3$, we have

$$\sum_{t \in \mathcal{J}_3} \min\{1, \beta_t \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})\}$$

$$\leq \sum_{t \in \mathcal{J}_3} \beta_t \bar{\sigma}_t \cdot \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})$$

$$\overset{(a)}{\leq} \max_{t \in [T]} \beta_t \cdot \gamma^2 \sum_{t=1}^T \min\left\{1, \frac{1}{\bar{\sigma}_t^2} \mathcal{D}_\mathcal{F}^2(X_t; X_{[t-1]}, \sigma_{[t-1]})\right\}$$

$$\leq \max_{t \in [T]} \beta_t \cdot \gamma^2 \dim_\mathcal{F},$$

where (a) holds due to $\bar{\sigma}_t = \gamma^2 \frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]})$ and $\frac{1}{\bar{\sigma}_t} \mathcal{D}_\mathcal{F}(X_t; X_{[t-1]}, \sigma_{[t-1]}) \leq 1$. Finally, putting pieces together finishes the proof. □

Lemma H.6 (Modified from Lemma 2 in Zhang et al. 2021a). *Let $\lambda_1, \lambda_2, \lambda_4 > 0$, $\lambda_3 \geq 1$ and $i' = \lceil \log_2 \lambda_1 \rceil$. Let $a_0, a_1, a_2, \ldots, a_{i'}$ be non-negative reals such that $a_i \leq \lambda_1$ for any $0 \leq i \leq i'$, and $a_i \leq \lambda_2 \sqrt{a_{i+1} + 2^{i+1} \cdot \lambda_3} + \lambda_4$ for any $0 \leq i < i'$. Then we have*

$$a_0 \leq \max\left\{\left(\lambda_2 + \sqrt{\lambda_2^2 + \lambda_4}\right)^2, \lambda_2\sqrt{4\lambda_3} + \lambda_4\right\} \leq \lambda_2\sqrt{4\lambda_3} + 4\lambda_2^2 + 3\lambda_4,$$

$$a_1 \leq \max\left\{\left(\lambda_2 + \sqrt{\lambda_2^2 + \lambda_4}\right)^2, \lambda_2\sqrt{8\lambda_3} + \lambda_4\right\} \leq \lambda_2\sqrt{8\lambda_3} + 4\lambda_2^2 + 3\lambda_4.$$