Proceedings of Machine Learning Research vol 195:1–77, 2023 36th Annual Conference on Learning Theory

VOQL: Towards Optimal Regret in Model-free RL with Nonlinear Function Approximation

Alekh Agarwal ALEKHAGARWAL@GOOGLE.COM
Google Research

Yujia Jin YUJIAJIN@STANFORD.EDU
Stanford University

Tong Zhang TOZHANG@GOOGLE.COM
Google Research

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study time-inhomogeneous episodic reinforcement learning (RL) under general function approximation and sparse rewards. We design a new algorithm, Variance-weighted Optimistic Q-Learning (VOQL), based on Q-learning and bound its regret assuming closure under Bellman backups, and bounded Eluder dimension for the regression function class. As a special case, VOQL achieves $\widetilde{O}(d\sqrt{TH} + d^6H^5)$ regret over T episodes for a horizon H MDP under (d-dimensional) linear function approximation, which is asymptotically optimal. Our algorithm incorporates weighted regression-based upper and lower bounds on the optimal value function to obtain this improved regret. The algorithm is computationally efficient given a regression oracle over the function class, making this the first computationally tractable and statistically optimal approach for linear MDPs.

Keywords: Reinforcement learning, nonlinear function approximation, model-free algorithms, eluder dimension.

1. Introduction

Optimally trading off exploration and exploitation to achieve a low regret is a fundamental question in Reinforcement Learning (RL) research. The last few years have seen some significant advances on this front, with the development of algorithms that achieve optimal regret guarantees when the underlying state space is finite (Azar et al., 2017; Zanette and Brunskill, 2019; Zhang and Ji, 2019; Simchowitz and Jamieson, 2019; Zhang et al., 2020b; He et al., 2020). Remarkably, a line of work has shown that when the overall reward of each trajectory is bounded independent of the horizon, then the regret has no explicit polynomial horizon dependence (Zanette and Brunskill, 2019; Zhang et al., 2021a, 2020a; Ren et al., 2021; Tarbouriech et al., 2021). However, these fall some way short of being applicable to real-world RL settings with large state spaces, and where we rely on the use of function approximation to generalize across related states. Some recent works (Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022) do generalize ideas from the tabular setting for a special class of RL problems with linear function approximation, called linear mixture MDPs, using a model-based approach that is more amenable to ideas from the tabular setting. Motivated by this landscape our paper asks if we can develop model-free techniques that attain optimal regret guarantees with general function approximation.

In particular, we consider function approximation in the so-called Q-type setting (Jin et al., 2021), where we explore using pointwise notions of optimism, and which generally avoids some

Setting	Method	Regret
linear MDPs	LSVI-UCB (Jin et al., 2020)	$d^{\frac{3}{2}} H \sqrt{T}$
	ELEANOR (Zanette et al., 2020)	$dH\sqrt{T}$
	VOQL (**our work, Theorem 8**)	$d\sqrt{HT}$
	Lower bound (Zhou et al., 2021)	$d\sqrt{HT}$
general function class	$\mathcal{F}$-LSVI (Wang et al., 2020)	$\dim(\mathcal{F})\sqrt{\log \mathcal{N} \log \mathcal{N}'} \cdot H\sqrt{T}$
	GOLF (Jin et al., 2021)	$\sqrt{\dim(\mathcal{F}) \log \mathcal{N}} \cdot H\sqrt{T}$
	VOQL (**our work, Theorem 9**)	$\sqrt{\dim(\mathcal{F}) \log \mathcal{N} \cdot HT}$

Table 1: **Comparison of regret:** In-homogeneous episodic RL with horizon H, with T trajectories, and sparse rewards $\sum_{h\in[H]} r^h \leq 1$. We only state the leading $O(\sqrt{T})$ term and hide poly-logarithmic factors in T, H, ϵ and δ. For linear MDPs, concurrent to our work, He et al. (2022) also obtained a regret bound of $\widetilde{O}(d\sqrt{HT})$. For general function class, $\dim(\mathcal{F})$ and $\log \mathcal{N}$ refer to properties of the function class $\mathcal{F}$, i.e. the generalized Eluder dimension (see Definition 2) and $\mathcal{N} = |\mathcal{F}|$.[2] In Wang et al. (2020), $\mathcal{N}' = \mathcal{N}(\mathcal{X} \times \mathcal{A})$ refers to the covering number of the state-action space.

additional horizon factors which arise in more general V-type settings. We adapt Eluder dimension based techniques (Russo and Van Roy, 2013; Wang et al., 2020) to design an exploration bonus, motivated by the empirical success of such bonus-based approaches (Feng et al., 2021; Burda et al., 2018; Henaff et al., 2022), and establish horizon-free guarantees in terms of a generalized Eluder dimension.

Model and Our Results. We study time-inhomogenous finite horizon MDPs with a horizon h, meaning that the transition dynamics and rewards at each step $h = 1, 2, \ldots, H$ can be different. We focus on model-free and value-based approaches, where the goal is to learn the optimal value function $Q_\star$ by searching over some function class $\mathcal{F}$. When $\mathcal{F}$ satisfies standard realizability and completeness assumptions, and the cumulative reward over each trajectory is at most 1, we show that the regret of our algorithm Variance-weighted Optimistic Q-Learning (VOQL)[3] scales as $\widetilde{O}\left(\sqrt{TH\dim(\mathcal{F})\log \mathcal{N}}\right)$. Here $\dim(\mathcal{F})$ is a weighted generalization of the standard Eluder dimension, which still captures (generalized) linear models. For the special case of $\mathcal{F}$ being linear, such as in linear MDPs, VOQL attains a sample complexity of $\widetilde{O}(d\sqrt{HT})$. As shown in Table 1, our guarantee is minimax optimal. Unlike the ELEANOR algorithm of Zanette et al. (2020), our approach is both model-free and computationally efficient, and indeed the first optimal model-free result for linear MDPs.

2. Formally, we allow $\mathcal{F}^h$ to vary as a function of h, in which case the precise result can be found in Theorem 6. Here we discuss the setting of $\mathcal{F}$ being shared across h for simplicity.

3. pronounced vocal

Overview of Techniques Our algorithm VOQL is based on optimistic Q-learning in a finite horizon setting, where we add an exploration bonus to the rewards for learning an optimistic estimate of the optimal value function $Q_\star$. There are two main challenges which result in sub-optimal horizon or dimension dependence in prior works. The bonus added across rounds typically grows as $O(H)$, even when the rewards add up to 1. The bonus is also data dependent, and a direct uniform concentration argument (Jin et al., 2020) yields a sub-optimal scaling with dimension. Taking a cue from the insights in tabular results, we use *weighted regression* to estimate $Q_\star$, such that the variance of our estimator is bounded independent of the horizon. We additionally establish a key *monotonicity property of our optimistic estimates*, which has been a significant challenge in prior works (Hu et al., 2021). Taken together, these techniques result in our optimal horizon scaling. For dimension dependence, we generalize the idea in Wagenmaker et al. (2022) which avoids the additional leading-order factors due to uniform convergence in the PAC setting. In particular, we *decouple the uniform convergence argument* into a higher order $O(d)$ term, and a lower order $O(d^2)$ term, which yields optimal scaling in the dominant term of the regret. Each of these techniques is potentially of independent interest in future works.

We note that the weighted regression technique is inpired by the linear mixture MDP work of Zhou et al. (2021), but the adaptation to linear MDPs has significant challenges. The use of over optimistic $Q_\star$ estimates that we use was introduced in the original version of Hu et al. (2022) posted prior to this work, but their original result had an error in the monotonicity analysis as explained in Appendix B. Concurrently with our work, an independent work He et al. (2022) also obtain a similar result for linear MDPs, using a different rare policy-switching argument, combined with weighted regression.

2. Preliminaries

We consider the following time-inhomogeneous episodic Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P} := \{P^h\}_{h\in[H]}, \mathcal{R} := \{r^h\}_{h\in[H]})$ with horizon length $H \in \mathbb{Z}_{>0}$, where $[H]$ is a shorthand for the set $\{1, 2, \ldots, H\}$. Here we let $P^h : \mathcal{X} \times \mathcal{A} \to \Delta^{\mathcal{X}}$ and $r^h : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ characterize the transition kernel and instantaneous reward at a given level $h \in [H]$ respectively. We consider a sparse reward setting where $\sum_{h\in[H]} r^h \in [0, 1]$ under the realization of any policy. We use $(x^h, a^h) \in \mathcal{X} \times \mathcal{A}$ to denote an arbitrary state-action pair at level h (omitting h when clear from context), and write $z = (x, a)$ as shorthand. A policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from state space to action space.[4] Since the optimal policy is non-stationary in an episodic MDP, we use π to refer to the H-tuple $\{\pi^h\}_{h\in[H]}$

Given an episodic MDP $\mathcal{M}$ and some policy $\pi := \{\pi^h\}_{h\in[H]}$, the V-value and Q-value functions are defined as the expected cumulative rewards, starting at level h from state-action pair $z^h = (x^h, a^h)$ or state x^h, when following the policy π, i.e.

$$Q^h_\pi(x^h, a^h) = \sum_{h' \geq h} \mathbb{E}\big[r^{h'}(x^{h'}, a^{h'}) \mid x^h, a^h, a^{h'} = \pi^{h'}(x^{h'})\big], \; V^h_\pi(x^h) = Q^h_\pi(x^h, \pi^h(x^h)). \tag{1}$$

Given some initial distribution $q \in \Delta^{\mathcal{X}}$, the optimal policy $\pi_\star \in \arg\max_{\pi^h, h\in[H]} \mathbb{E}_{x^1\sim q} V^1_\pi(x^1)$. For simplicity we write $Q^h_\star = Q^h_{\pi_\star} = \sup_\pi Q^h_\pi$ and $V^h_\star = V^h_{\pi_\star} = \sup_\pi V^h_\pi$ when clear from context.

4. Many works also consider randomized policies $\pi : \mathcal{X} \to \Delta^{\mathcal{A}}$ in reinforcement learning. In this paper it suffices to constrain to the class of deterministic policies.

We define the Bellman operator $\mathcal{T}$ on functions $f : \mathcal{X} \to \mathbb{R}$ so that $(\mathcal{T}f)(x^h, a^h) = \mathbb{E}_{r^h, x^{h+1}}[r^h + f(x^{h+1})|x^h, a^h]$. We often use the shorthand $f(x) = \max_{a \in \mathcal{A}} f(x, a)$. The definition of value functions ensures the validity of the *Bellman equation*, i.e. $Q^h_\star(x^h, a^h) = (\mathcal{T}V^{h+1}_\star)(x^h, a^h)$. We also define the Bellman operator for second moment as $\mathcal{T}_2 f(x^h, a^h) = \mathbb{E}_{r^h, x^{h+1}} \left[\left(r^h + f(x^{h+1}) \right)^2 |x^h, a^h \right]$.

We consider a class of episodic MDPs such that the value functions satisfy the (approximate) realizability assumption under a general function class $\mathcal{F} := \{\mathcal{F}^h\}_{h \in [H]}$. More concretely, we introduce the following assumption:

Assumption 1 (ϵ-realizability under Bellman backups) *Given $\{\mathcal{F}^h\}_{h \in [H]}$ where each set $\mathcal{F}^h$ is composed of functions $f^h : \mathcal{X} \times \mathcal{A} \to [0, L]$. We assume for each $h \in [H]$, and any $V : \mathcal{X} \to [0, 1]$ there exists $f^h \in \mathcal{F}^h$ such that $\max_{x,a \in \mathcal{X} \times \mathcal{A}} \left| f^h(x, a) - \mathcal{T}V(x, a) \right| \le \epsilon$, and $\max_{x,a \in \mathcal{X} \times \mathcal{A}} \left| f^h(x, a) - \mathcal{T}_2 V(x, a) \right| \le \epsilon$. Also we assume there exists some $f^h_\star \in \mathcal{F}^h$ such that $\left\| f^h_\star - Q^h_\star \right\|_\infty \le \epsilon$, for all $h \in [H]$. We assume $L = O(1)$ and use $\mathcal{N}$ to denote the maximal size of function class $\max_{h \in [H]} |\mathcal{F}^h|$ throughout the paper.*

When $\epsilon = 0$, the assumption states that the function class is complete and well-specified under Bellman backups for any function V. Such an assumption is stronger than the classical completeness assumption $\mathcal{T}\mathcal{F}^{h+1} \subseteq \mathcal{F}^h$ (see e.g. Chen and Jiang, 2019). This realizability assumption is standard for analyzing value-based methods relying on regression (see e.g. Wang et al., 2020; Jin et al., 2021) under general function approximation, due to the non-linearity in the bonus terms. While this assumption can be avoided in an information-theoretic sense using ideas developed in Jiang et al. (2017) and follow-ups, avoiding it will introduce computational overhead as in Zanette et al. (2020) and thus we make this assumption in the interest of obtaining sharper guarantees. The assumption naturally holds for tabular and linear MDPs. More generally, ϵ allows us to capture a bounded misspecification. When we instantiate the function class as a cover of a larger infinite class, the covering might also induce a non-zero ϵ in Assumption 1.

Since we use linear MDPs as a running example to illustrate our key definitions and assumptions, we define them formally next.

Definition 1 (Linear MDPs (Yang and Wang, 2020; Jin et al., 2020)) *An MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ is a linear MDP if there exists a known feature mapping $\phi^h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ for every $h \in [H]$, such that for any $h \in [H]$, and any $(x^h, a^h) \in \mathcal{X} \times \mathcal{A}$, we have $P^h(\cdot|x^h, a^h) = \langle \phi^h(x^h, a^h), \mu^h(\cdot) \rangle$ and $\mathbb{E}\left[r^h | x^h, a^h \right] = \langle \phi^h(x^h, a^h), \theta^h \rangle$, for some unknown measures $\boldsymbol{\mu}^h = \{\mu^h(x)\}_{x \in \mathcal{X}}$ where each $\mu^h(x) \in \mathbb{R}^d$ and $\theta^h \in \mathbb{R}^d$. We assume that $\sup_{x,a \in \mathcal{X} \times \mathcal{A}} \|\phi(x, a)\|_2 \le 1$, $\left\| \sum_{x \in \mathcal{X}} \mu^h(x) \right\|_2 + \left\| \theta^h \right\|_2 \le B^h$ (possibly scaling with d), and $\sum_{h \in [H]} r^h \in [0, 1]$.*

Jin et al. (2020) show that linear MDPs satisfy the realizability assumption under the linear function class $\mathcal{F}^h_{\mathsf{lin}}$ defined as

$$\mathcal{F}^h_{\mathsf{lin}} := \{ \langle w^h, \phi^h(\cdot, \cdot) \rangle : w^h \in \mathbb{R}^d, \|w^h\|_2 \le B^h \}, \text{ for any } h \in [H]. \tag{2}$$

We also define $\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})$ be an ϵ_{c}-cover of $\mathcal{F}^h_{\mathsf{lin}}$ under the ℓ_∞ norm, so that $\log |\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})| = O(d \log \frac{B^h}{\epsilon_{\mathrm{c}}})$.

While the realizability completeness assumption allows us to control the error of our regression solution to $Q_\star$ under the data distribution used in regression, it does not control the complexity of exploration in the MDP, when the learner uses the classes $\{\mathcal{F}^h\}_{h \in [H]}$. To capture this complexity, we now define an additional quantity which we call a *generalized Eluder dimension*, which extends the original definition of Russo and Van Roy (2013) to weighted regression settings, based on recent work of Gentile et al. (2022) (also see Zhang (2023)).

Definition 2 (Generalized Eluder dimension) *Let $\lambda > 0$, a sequence of state-action pairs $Z = \{z_i\}_{i\in[T]}$ and $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in[T]}$ be given. The* generalized Eluder dimension *of a function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, L]$ is given by $\dim_{\alpha,T}(\mathcal{F}) := \sup_{Z,\boldsymbol{\sigma}:|Z|=T,\boldsymbol{\sigma}\geq\alpha} \dim(\mathcal{F}, Z, \boldsymbol{\sigma})$, where*

$$\dim(\mathcal{F}, Z, \boldsymbol{\sigma}) := \sum_{i=1}^{T} \sigma_i^{-2} D_{\mathcal{F}}^2(z_i; z_{[i-1]}, \sigma_{[i-1]}) \vee 1,$$

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

We also use $d_\alpha := \frac{1}{H}\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)$ when function class $\{\mathcal{F}^h\}_{h\in[H]}$ is clear from context.

For linear MDPs, the definition of generalized Eluder dimension for the relevant function class $\mathcal{F}_{\mathsf{lin}}^h$ can be simplified as follows:

Lemma 3 *For the class $\mathcal{F}_{\mathsf{lin}}^h$ defined in (2), letting $\mathcal{F}_{\mathsf{lin}}^h(\epsilon_{\mathrm{c}})$ be the ϵ_{c}-cover of $\mathcal{F}_{\mathsf{lin}}^h$ for some $\epsilon_{\mathrm{c}} > 0$, we have $\dim_{\alpha,T}(\mathcal{F}_{\mathsf{lin}}^h(\epsilon_{\mathrm{c}})) \leq \dim_{\alpha,T}(\mathcal{F}_{\mathsf{lin}}^h) = O\big(d \log\big(1 + \frac{(B^h)^2 T}{\alpha^2 d\lambda}\big)\big) = \widetilde{O}(d)$.*

Remark 4 (Relation to standard Eluder dimension) *When $\boldsymbol{\sigma} \equiv 1$, $\max_{Z:|Z|=T} \dim(\mathcal{F}, Z, \mathbf{1}) \leq \dim_E(\mathcal{F}, \sqrt{\lambda/T}) + 1$, where $\dim_E(\mathcal{F}, \varepsilon)$ is the standard Eluder dimension of $\mathcal{F}$ as defined in Russo and Van Roy (2013). The unweighted version of our definition has also appeared in Gentile et al. (2022). The generalized definition we give takes supremum over any $\boldsymbol{\sigma} \geq \alpha$, and thus is incomparable with the standard Eluder dimension even when $\alpha = 1$.*

We use the learning protocol in episodic reinforcement learning where at every episode $t \in [T]$ and horizon level $h \in [H]$, the learner explores the trajectory based on some exploration rule that only depends on the historical data. The learner then generates new data $\{x_t^h, a_t^h, r_t^h\}_{h\in[H]}$ based on her data exploration rule, chooses action a_t^h, transitions to the next state $x_t^{h+1} \sim P^h(x_t^h, a_t^h)$ and receives reward $r_t^h = r^h(x_t^h, a_t^h, x_t^{h+1})$. The goal of learner is to optimize her *regret* while interacting with the environment in T episodes, where the initial distribution $x^1 \sim \mu$ generates from some given fixed initial distribution. Formally, we define the regret as:

$$\mathrm{Regret}(T) = \sum_{t\in[T]} \mathbb{E}_{x^1\sim\mu}\left[V_\star^1\left(x^1\right) - V_t^1\right], \quad \text{where } V_t^1 := \mathbb{E}\left[\sum_{h\in[H]} r_t^h | x^1, f_{t,1}^{[h]}, f_{t,2}^{[h]}\right]. \tag{3}$$

Here V_t^1 denotes the expected cumulative reward in the t_{th} trajectory, where the exploration policy may depend on some functions $f_{t,1}^{[H]}$, $f_{t,2}^{[H]}$ based on the history (see exploration rule (9)).

In general for episodic reinforcement learning, the optimal policy as some Markovian non-stationary policy (that only depends on the horizon level h and current state x) always exists.

3. Algorithm

We now discuss our algorithm, Variance-weighted Optimistic Q-Learning (VOQL) in detail. The pseudocode for the algorithm is presented in Algorithm 1. At a high-level, the algorithm performs optimistic Q-learning style updates (Line 9 to Line 12), where we repeatedly apply the empirical Bellman optimality operator at each level h to an optimistic value function for $h+1$, and add an additional bonus to the resulting function to account for the regression uncertainty in the empirical Bellman operator. We additionally maintain over-optimistic and over-pessimistic estimates (Line 13 to Line 17), which are combined to form both a variance estimate in reweighting our regression objective, as well as in defining the data collection policy in Equation (9) (used in Line 21). We first describe some of the key elements of the algorithm, before discussing how they fit together.

For brevity, we only specify the parameters of the algorithm somewhat informally in the following discussion, focusing on the dependency on the key parmeters. Precise parameter settings can be found in Table 3 in the Appendix.

Regression and weighted regression. In episodic reinforcement learning, many online algorithms iteratively solve the following (weighted) regression problem to fit the past dataset: At episode t, given a target function $f_t^{h+1} \approx V_\star^{h+1}$ learnt from the past data, in order to fit $\mathcal{T} f_t^{h+1} = \mathbb{E}[r^h + f_t^{h+1}(x^{h+1})]$ we define

$$\text{least square estimator } \hat{f}_t^h = \arg\min_{f^h \in \mathcal{F}^h} \sum_{s \in [t-1]} (\bar{\sigma}_s^h)^{-2} \left(f^h\left(x_s^h, a_s^h\right) - r_s^h - f_t^{h+1}(x_s^{h+1}) \right)^2,$$
$$\text{version space } \mathcal{F}_t^h := \left\{ f^h \in \mathcal{F}^h : \sum_{s \in [t-1]} (\bar{\sigma}_s^h)^{-2} \left(f^h(x_s^h, a_s^h) - \hat{f}_t^h(x_s^h, a_s^h) \right)^2 \le \left(\beta_t^h\right)^2 \right\}. \tag{5}$$

Standard approaches in Q-learning solve least squares problems like Equation (5) by solving an unweighted regression with $\bar{\sigma}^h \equiv 1$ for all $h \in [H]$, where $f_t^{h+1} \approx Q_\star^{h+1}$. Note that we take a backup of the iterate f_t^{h+1} instead of f_{t-1}^{h+1} as it is natural to do bottom-up approximate dynamic programming in a finite horizon setting. Optimistic variants of Q-learning (Jin et al., 2020; Wang et al., 2020) use some optimistic estimate f_t^{h+1} of $Q_\star^{h+1}$ instead. More recently, a line of work has studied the benefits of using weights informed by the variance to satisfy $\mathbb{V}\left[r^h + f_t^{h+1}(x^{h+1}) | z_s^h\right] \le \left(\bar{\sigma}_s^h\right)^2 \le \widetilde{O}(1)$, for obtaining stronger guarantees in terms of their horizon dependence in linear bandits and linear mixture MDP settings (Zhou et al., 2021; Zhou and Gu, 2022), as well as in linear MDPs (Hu et al., 2022). Our formulation of weighted regression here is motivated by these works, extending such techniques to a non-linear and model-free setting.

The radius of confidence interval β_t^h is properly chosen to ensure that $\mathcal{T} f_t^{h+1} \in \mathcal{F}_t^h$ with high probability (up to small additive error element-wise due to Assumption 1). The solution of this regression problem also admits pointwise confidence bounds on the error to $Q_\star^h$ under the *bounded Eluder dimension* condition (see Definition 2).

Optimistic value estimation and bonus oracle. We now concretely describe how we use weighted regression to construct an optimistic estimate of $Q_\star^h$. Since $Q_\star^h(x,a) = \mathbb{E}[r^h + V_\star^{h+1}(x')|x,a]$, there are two source of uncertainty which need to be upper bounded. First is from the error in our estimates of $V_\star^{h+1}$, that is addressed by using a regression target f_t^{h+1} which is optimistic for $V_\star^{h+1}$. The second source is the estimation error in the conditional expectation using samples at step h. In

Algorithm 1: Variance-weighted Optimistic Q-Learning (VOQL)

Input: function class $\{\mathcal{F}^h\}_{h\in[H]}$, a consistent bonus oracle $\mathcal{B}$, $\epsilon > 0$
Parameters: $\{u_t\}_{t\in[T]}$, λ, bonus error ϵ_{b}, α, δ, $\{\beta^h_{t,1}, \beta^h_{t,2}, \bar{\beta}^h_t\}^{h\in[H]}_{t\in[T]}$
Initialize $\mathcal{D}^h_{[0]} = \emptyset$ for all $h \in [H]$
for *episode* $t = 1, 2, \cdots, T$ **do**
  Initialize last step $f^{H+1}_{t,j}(\cdot) \leftarrow 0$, for all $j = 1, 2, -2$
  if $t > 1$ **then**
    for $h = H, H-1, \cdots, 1$ **do**
      When $t > 1$, define $\bar{\sigma}^h_{t-1}$ as in Equation (8)
      Solve $\hat{f}^h_{t,1} = \arg\min_{f^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}^h_s)^2} \left(f^h(x^h_s, a^h_s) - r^h_s - f^{h+1}_{t,1}(x^{h+1}_s)\right)^2$
      Set $b^h_{t,1} \leftarrow \mathcal{B}\left(\{\bar{\sigma}^h_s\}_{s\in[t-1]}, \mathcal{D}^h_{[t-1]}, \mathcal{F}^h, \hat{f}^h_{t,1}, \beta^h_{t,1}, \lambda, \epsilon_{\mathrm{b}}\right)$ (see Definition 5)
      Update $f^h_{t,1}(\cdot) = \min\left(\hat{f}^h_{t,1}(\cdot) + b^h_{t,1}(\cdot) + \epsilon, 1\right)$
      Update optimistic V-value $f^h_{t,1}(x) = \max_a f^h_{t,1}(x, a)$ for all $x \in \mathcal{X}$
      Solve $\hat{f}^h_{t,j} = \arg\min_{f^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \left(f^h(x^h_s, a^h_s) - r^h_s - f^{h+1}_{t,j}(x^{h+1}_s)\right)^2$, $j = \pm 2$
      Set $b^h_{t,2} \leftarrow \mathcal{B}\left(\mathbf{1}_{[t-1]}, \mathcal{D}^h_{[t-1]}, \mathcal{F}^h, \hat{f}^h_{t,2}, \beta^h_{t,2}, \lambda, \epsilon_{\mathrm{b}}\right)$ (see Definition 5)
      Update $f^h_{t,2}(\cdot) = \min\left(\hat{f}^h_{t,2}(\cdot) + 2b^h_{t,1}(\cdot) + b^h_{t,2}(\cdot) + 3\epsilon, 2\right)$
      Update $f^h_{t,-2}(\cdot) = \max\left(\hat{f}^h_{t,-2}(\cdot) - b^h_{t,2}(\cdot) - \epsilon, 0\right)$
      Update $f^h_{t,\pm 2}(x) = \max_a f^h_{t,\pm 2}(x, a)$ for all $x \in \mathcal{X}$
      Solve $\hat{g}^h_t = \arg\min_{g^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \left(g^h(x^h_s, a^h_s) - \left(r^h_s + f^{h+1}_{t,2}(x^{h+1}_s)\right)^2\right)^2$
  Receive initial state $x^1_t \sim \mu$
  for $h = 1, 2, \cdots, H$ **do**
    Generate $\mathcal{D}^{[H]}_t$ from x^1_t according to u_t and data collection policy (Equation (9))
    if $t = 1$ **then** $\left(\sigma^h_t\right)^2 = 4$
    else

$$\left(\sigma^h_t\right)^2 = \min\left(4, \hat{g}^h_t(z^h_t) - \left(\hat{f}^h_{t,-2}(z^h_t)\right)^2 + D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}^h_{[t-1]}) \cdot \left(\sqrt{\left(\bar{\beta}^h_t\right)^2 + \lambda} + 2L\sqrt{\left(\beta^h_{t,2}\right)^2 + \lambda}\right) + 2(1+L)\epsilon\right) \tag{4}$$

simple settings where the space $\mathcal{X}$ is discrete or when the class $\mathcal{F}$ is linear, this error is quantified as an optimistic bonus using either the number of samples for x, a or the standard elliptical bonus (see e.g. Abbasi-Yadkori et al., 2011; Jin et al., 2020). An optimistic function at time h is then defined as $f^h_{t,1} = \hat{f}^h_t + b^h_t$, where b^h_t is the optimistic bonus and $\hat{f}^h_t$ is as defined in (5) with an optimistic target $f^{h+1}_{t,1}$. The reason for denoting the optimistic function as $f^h_{t,1}$ instead of f^h_t will be shortly clarified when we define an additional overly optimistic sequence.

For a general function class, we use $b_t^h(z^h) = \max_{f^h \in \mathcal{F}_t^h} f^h(z^h) - \min_{f^h \in \mathcal{F}_t^h} f^h(z^h)$ to capture the regression uncertainty for all $z^h \in \mathcal{X} \times \mathcal{A}$ (Feng et al., 2021). However, this uncertainty bonus has a high complexity in that the maximizing and minimizing functions can differ arbitrarily for each z^h. Consequently, the target function $f_{t,1}^{h+1}$ defined using $\hat{f}_t^{h+1} + b_t^{h+1}$ is very complex. Since f_t^{h+1} is random in the regression objective (5), this high complexity induces a potentially poor confidence bound for the solution of regression. To circumvent this issue, we assume the existence of a low complexity bonus oracle which roughly dominates the value obtained by the pointwise maximization over $\mathcal{F}_t^h$ for now. We subsequently provide concrete instantiations of this bonus oracle for linear MDPs and general settings with a low Eluder dimension in Section 4.

Definition 5 (Oracle $\mathcal{B}$ for bonus function b_t^h) *Given $h \in [H]$, $t \in [T]$, sequence of $\{\bar{\sigma}_s^h\}_{s\in[t-1]}$ and dataset $\mathcal{D}_{[t-1]}^h = \{(x_s^h, a_s^h, r_s^h, x_s^{h+1})\}_{s\in[t-1]}$, function class $\mathcal{F}^h$ with $\hat{f}^h \in \mathcal{F}^h$, $\beta^h, \lambda \geq 0$, error parameter $\epsilon_{\mathrm{b}} \geq 0$* [5]*, the* bonus oracle *$\mathcal{B}(\{\bar{\sigma}_s^h\}_{s\in[t-1]}, \mathcal{D}_{[t-1]}^h, \mathcal{F}^h, \hat{f}^h, \beta^h, \lambda, \epsilon_{\mathrm{b}})$ outputs a bonus function $b^h(\cdot)$ such that:*

(i) $b^h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ belongs to a function class $\mathcal{W}$ and we use $\mathcal{N}_b$ to denote the size of bonus function class $|\mathcal{W}|$;

(ii) $b^h(z^h) \geq \max\left\{|f^h(z^h) - \hat{f}^h(z^h)|,\ f^h \in \mathcal{F}^h : \sum_{s\in[t-1]} \frac{\left(f^h(z_s^h) - \hat{f}_t^h(z_s^h)\right)^2}{\left(\bar{\sigma}_s^h\right)^2} \leq \left(\beta^h\right)^2\right\}$, for any $z^h \in \mathcal{X} \times \mathcal{A}$;

(iii) $b^h(z^h) \leq C \cdot \left(D_{\mathcal{F}^h}(z^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \cdot \sqrt{\left(\beta^h\right)^2 + \lambda} + \epsilon_{\mathrm{b}} \cdot \beta^h\right)$ for all $z^h \in \mathcal{X} \times \mathcal{A}$ for some $C > 0$.

Further we say the oracle $\mathcal{B}$ is consistent *if for any $t < t'$ with $\{\bar{\sigma}_s^h\}_{s\in[t-1]} \subseteq \{\bar{\sigma}_s^h\}_{s\in[t'-1]}$, $\mathcal{D}_{[t-1]}^h \subseteq \mathcal{D}_{[t'-1]}^h$, β_t^h non-decreasing in t for each $h \in [H]$ and $\mathcal{F}_t^h, \hat{f}_t^h$ as defined in (5), we have $\mathcal{B}(\{\bar{\sigma}_s^h\}_{s\in[t-1]}, \mathcal{D}_{[t-1]}^h, \mathcal{F}_t^h, \hat{f}_t^h, \beta_t^h, \lambda, \epsilon_{\mathrm{b}}) \geq \mathcal{B}(\{\bar{\sigma}_s^h\}_{s\in[t'-1]}, \mathcal{D}_{[t'-1]}^h, \mathcal{F}_{t'}^h, \hat{f}_{t'}^h, \beta_{t'}^h, \lambda, \epsilon_{\mathrm{b}})$ element-wise.*

With such an oracle $\mathcal{B}$, we define the optimistic sequence $f_{t,1}^h \approx \hat{f}_{t,1}^h + b_{t,1}^h$ (approximation due to additive ϵ term and truncation), where $b_{t,1}^h = \mathcal{B}(\{\bar{\sigma}_s^h\}_{s\in[t-1]}, \mathcal{D}_{[t-1]}^h, \mathcal{F}_t^h, \hat{f}_t^h, \beta_{t,1}^h, \lambda, \epsilon_{\mathrm{b}})$, $\beta_{t,1}^h \approx \sqrt{\log \mathcal{N}}$ and $\hat{f}_{t,1}^h$ is a solution to (5) with $f_{t,1}^{h+1}$ as the target function.

Overly optimistic and overly pessimistic value estimates. A sharp analysis of the convergence of the optimistic estimates $f_{t,1}^h$ requires appropriately estimated variances for weighting the regression examples, which satisfy $\mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_s^h\right] \leq \left(\bar{\sigma}_s^h\right)^2$ as mentioned before. In order to produce such variance estimates, we first define an auxiliary set of value function estimates, which are then used in variance estimation. Specifically, we define an *overly optimistic* estimate $f_{t,2}^h$, as well as an *overly pessimistic* estimate $f_{t,-2}^h$. Roughly, these functions are designed to ensure that

$$\underbrace{f_{s,-2}^h(z^h)}_{\text{overly-pessimistic}} \leq \underbrace{f_\star^h(z^h)}_{:= Q_\star^h,\text{ true optimal}} \leq \underbrace{f_{t,1}^h(z^h)}_{\text{optimistic}} \leq \underbrace{f_{s,2}^h(z^h)}_{\text{overly optimistic}} \quad \text{for any } z^h \in \mathcal{X} \times \mathcal{A} \text{ and } s \leq t. \tag{6}$$

Concretely, the functions $f_{t,\pm 2}^h$ are defined by finding an unweighted regression solution $\hat{f}_{t,\pm 2}^h$ to (5) with target function $f_{t,\pm 2}^{h+1}$, and defining $f_{t,\pm 2}^h \approx \hat{f}_{t,\pm 2}^h \pm 2b_{t,2}^h \pm b_{t,1}^h$ (approximation due to ϵ

5. Here ϵ_{b} here characterizes the error we are able to tolerate in the bonus oracle, and may differ from ϵ and ϵ_{c}.

term and truncation), for some bonus $b^h_{t,2}$ defined using our bonus oracle. A key difference, however, is that the weight sequence $\bar{\sigma} \equiv 1$ in defining $\hat{f}^h_{t,\pm 2}$, so that we use standard unweighted regression in (5). We note that the idea of having an overly optimistic sequence for variance estimation is first introduced in Hu et al. (2022). However, they do not perform unweighted regression like us, which leads to some technical problems in their analysis as described in Appendix B. The bonuses $b^h_{t,2}$ are given by $\mathcal{B}(\mathbf{1}_{[t-1]}, \mathcal{D}^h_{[t-1]}, \mathcal{F}^h_t, \hat{f}^h_t, \beta^h_{t,2}, \lambda, \epsilon_{\mathrm{b}})$ and $\beta^h_{t,2} \approx \sqrt{\log \mathcal{N}\mathcal{N}_b}$. We note that the bonus multiplier $\beta_{t,2}$ also incorporates the complexity of the bonus oracle class $\mathcal{W}$. For intuition, $\beta_{t,1} = \widetilde{O}(\sqrt{d})$ in a linear MDP, while $\beta_{t,2} = \widetilde{O}(d)$, and making this distinction is crucial to obtaining the asymptotically optimal d dependence in our bounds.

Estimating variance. Next we discuss how to construct an appropriate variance upper bound $(\bar{\sigma}^h_s)^2$. Unlike the convenient condition of $(\bar{\sigma}^h_s)^2 \geq \mathbb{V}_{r^h, x^{h+1}}[r^h + f^{h+1}_{t,1}(x^{h+1})|x^h_s, a^h_s]$, which necessitates reasoning about a changing target, we perform a more careful analysis and leverage properties of the overly optimistic function $f^h_{s,2}$ to show that it suffices to ensure that $(\bar{\sigma}^h_s)^2 \geq \mathbb{V}[r^h + f^{h+1}_\star(x^{h+1})|x^h_s, a^h_s]$ at all rounds $s \leq t$. This change, which crucially relies on (6), fixes the target function to be $f^h_\star$, and enables the creation of a valid variance estimate.

Since variance involves second moment and squared expectation, we estimate the two separately. We estimate the second moment directly using an *unweighted regression* described below.

$$\forall\, t \in [T], h \in [H], \quad \hat{g}^h_t = \arg\min_{g^h \in \mathcal{F}^h} \sum_{s \in [t-1]} \left(g^h\left(x^h_s, a^h_s\right) - \left(r^h_s + f^{h+1}_{t,2}\left(x^{h+1}_s\right)\right)^2 \right)^2, \tag{7}$$

and choose $\bar{\beta}^h_t \approx \sqrt{\log \mathcal{N}\mathcal{N}_b}$.

A natural variance upper bound at step t can then be obtained as $\hat{g}^h_t - (f^h_{t,-2})^2$, but we need additional terms to account for the estimation errors and obtain a valid upper bound. This is achieved through the sequence $\bar{\sigma}^h_t$ defined as (informally here, see Equation (12) for the precise setting)

$$\sigma^h_t \text{ is as defined in Equation (4),}$$
$$\bar{\sigma}^h_t = \max\left\{ \sigma^h_t, \alpha, \sqrt{\widetilde{\Theta} \cdot \left(f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)\right)}, \sqrt{\widetilde{\Theta} \cdot D_{\mathcal{F}^h}\left(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}\right)} \right\}. \tag{8}$$

Here α is a lower bound for the variance estimate which we set inverse-polynomially in T, H for stability of the algorithm. We also let $\widetilde{\Theta} = \widetilde{\Theta}(\log \mathcal{N}\mathcal{N}_b)$, which also contains other logarithmic factors in $T, H, L, 1/\alpha, 1/\delta$ that are precisely defined in Equation (12).

Design of exploration policy. A natural exploration policy, given the optimistic sequence $f^h_{t,1}$ is to be greedy with respect to it. However, it is beneficial to sometimes act greedily with respect to the overly optimistic function $f^h_{t,2}$ if the two sequences begin to differ by a lot. At iteration t we choose actions as per the following rule:

$$a^h_t = \begin{cases} \operatorname{argmax}_{a \in \mathcal{A}} f^h_{t,1}(x^h_t, a) & \text{if } f^h_{t,1}(x^{h'}_t) \geq f^h_{t,2}(x^{h'}_t) - u_t \text{ for all } h' \leq h, \\ \operatorname{argmax}_{a \in \mathcal{A}} f^h_{t,2}(x^h_t, a) & \text{otherwise,} \end{cases} \tag{9}$$

where u_t is an appropriately chosen threshold. Note that the action sequence defined this way is not a Markovian policy since it depends on the entire prefix trajectory at each step. However, it does constitute a valid exploration scheme, which suffices for regret minimization. In comparison, Hu et al. (2022) also use overly optimistic function in exploration, but there the agent only acts greedily with respect to the overly optimistic function.

4. Main Result and Applications

We now give the main guarantee for Algorithm 1, which is the main result of the paper.

Theorem 6 (Regret bound for VOQL, complete version in Theorem 15 and Theorem 45) *Suppose we are given function classes $\{\mathcal{F}^h\}_{h\in[H]}$ satisfy Assumption 1 with $\epsilon \in [0,1]$, have generalized Elude dimension $\dim_{\alpha,T}(\mathcal{F}^h)$, $h \in [H]$ as defined as in Definition 2 with $\lambda = 1$, and access to a consistent bonus oracle $\mathcal{B}$ satisfying Definition 5 with $TH\epsilon_{\mathrm{b}} = O(1)$. Let $d_\alpha = \frac{1}{H} \cdot \sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)$ with $\alpha = \sqrt{1/TH}$, and set $u_t = O(H^2\epsilon + H\delta + \sqrt{\log(\mathcal{N}TH/\delta) + T^2H\epsilon} \cdot (\log(\mathcal{N}\mathcal{N}_b TH/\delta)\sqrt{H^5 d_\alpha})/\sqrt{t})$. For any $\delta < 1/(T+H^2+11)$ the regret of VOQL satisfies $\mathbb{E}R_T = O\Big(\sqrt{\log(\mathcal{N}TH/\delta) + T^2H\epsilon}\cdot\sqrt{THd_\alpha} + \big(\log(\mathcal{N}TH/\delta)+T^2H\epsilon\big)\cdot\log^2(\mathcal{N}\mathcal{N}_bTH/\delta)H^5 d_\alpha\Big)$. The regret bound also holds with probability at least $1-\delta$.*

We remark that we don't pay attention to optimize the low-order $\mathrm{poly}(H)$ terms, which may be easily improved through more careful analysis. When $\epsilon = 0$ in Assumption 1 and $T = \widetilde{\Omega}\left(d_\alpha H^9 \log^4(\mathcal{N}\mathcal{N}_b)\log\mathcal{N}\right)$, we see that the regret scales as $\widetilde{O}(\sqrt{THd_\alpha \log \mathcal{N}})$. The dependence on T, $\log\mathcal{N}$ and the generalized Eluder dimension d_α is standard and unimprovable, as we will see in the special case of a linear MDP shortly. Also, we note that the regret scales as $\sqrt{H}$. This might appear sub-optimal since we assume that the trajectory level rewards, and hence values, are normalized in $[0,1]$. This scaling, however, captures the model complexity of an inhomogeneous process and is unavoidable due to a matching lower bound in linear MDPs as we discuss shortly.

We give a high-level proof sketch of the theorem in Section 5, with details deferred to Appendices F and G. First we present some specific consequences of the general result using concrete instantiations of the bonus oracle $\mathcal{B}$ for both linear and nonlinear function approximation. Due to the efficient implementation of such bonus oracles, our algorithm is computationally tractable for both linear and general function classes, modulo the tractability of regression in the general case.

4.1. Linear Function Approximation

We apply the theorem to the specific setting of linear MDPs (see Definition 1). Recall that in this case, the finite function class $\mathcal{F}^h$ is defined as an ϵ_{c}-cover $\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})$ at each level h for the linear function class $\{\langle w^h, \phi^h(\cdot,\cdot)\rangle : w^h \in \mathbb{R}^d, \|w^h\|_2 \le B^h\}$. We define $B = \max_{h\in[H]} B^h$ as a constant (same effect as L), so that $\log|\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})| = O\big(d\log(1+B/\epsilon_{\mathrm{c}})\big)$. Furthermore, we know from Lemma 3 that in this case $d_\alpha = O\big(d\log\big(1+\frac{B^2T}{\alpha^2 d\lambda}\big)\big)$.

The bonus oracle for linear MDPs is easily instantiated using the standard elliptical bonus, and satisfies all our properties as we show below. We refer readers to Appendix D for a complete proof.

Lemma 7 (Bonus oracle $\mathcal{B}$ for linear MDPs) *Given $T, H \in \mathcal{Z}_+$, suppose all $\beta^h_t \le \beta$ and β^h_t is non-decreasing in $t \in [T]$ for each $h \in [H]$. For any $t \ge 1$, $h \in [H]$, variances $\{\bar\sigma^h_s\}_{s\in[t-1]}$ satisfying $\bar\sigma^h_s \ge \alpha$ for some $\alpha > 0$, dataset $\mathcal{D}^h_{[t-1]} = \{(\phi^h(z^h_s), a^h_s, r^h_s, \phi(z^{h+1}_s))\}_{s\in[t-1]}$, function class $\mathcal{F}^h_t$ and $\hat f^h_t \in \mathcal{F}^h_t$ defined via weighted regression (5), and parameters $\lambda, \epsilon_{\mathrm{c}} > 0$, let $\mathcal{B}(\{\bar\sigma^h_s\}_{s\in[t-1]}, \mathcal{D}^h_{[t-1]}, \mathcal{F}^h_t, \hat f^h_t, \beta^h_t, \lambda, \epsilon_{\mathrm{c}}) = \|\phi^h(x,a)\|_{(\Sigma^h_t)^{-1}}\sqrt{(\beta^h_t)^2+\lambda}$, where $\Sigma^h_t = \frac{\lambda}{4(B^h)^2}I + \sum_{s\in[t-1]}\frac{1}{(\bar\sigma^h_s)^2}\phi^h(z^h_s)\phi^h(z^h_s)^\top$. For any choice of covering radius $\epsilon_{\mathrm{c}} \le \alpha\sqrt{\lambda/8T}$, the oracle satisfies all the properties of Definition 5 with $\log\mathcal{N}_b = \log|\mathcal{W}| = O\big(d^2\log\big(1+B^2\sqrt{d}\beta/(\lambda\epsilon_{\mathrm{c}}^2)\big)\big)$.*

Combined with Theorem 6, we obtain the following result when applying VOQL to linear MDPs with aforementioned function class $\mathcal{F}_{\mathsf{lin}}^h(\epsilon_\mathrm{c}), h \in [H]$ and bonus oracle $\mathcal{B}$.

Theorem 8 (Regret of VOQL for linear MDPs) *Under conditions of Theorem 6, suppose that the underlying MDP is linear, so that the original function class $\mathcal{F}_{\mathsf{lin}}^h$ satisfies Assumption 1 with $\epsilon = 0$ and the ϵ_c-cover $\mathcal{F}_{\mathsf{lin}}^h(\epsilon_\mathrm{c})$ satisfies Assumption 1 with $\epsilon = \epsilon_\mathrm{c}$. Choosing $\lambda = 1$, $u_t = \widetilde{\Theta}(d^3 H^{5/2}/\sqrt{t})$, $\alpha = \sqrt{1/HT}$, $\epsilon_\mathrm{b} = \epsilon_\mathrm{c} \leq 1/8HT$ and $\delta < 1/(T + H^2 + 11)$, VOQL with the bonus oracle defined in Lemma 7, achieves a total regret of $\mathbb{E}R_T = \widetilde{O}\left(d\sqrt{HT} + d^6 H^5\right)$. The regret bound also holds with probability at least $1 - \delta$.*

The regret bound is *asymptotically optimal* in the leading order term by adapting the construction of Zhou et al. (2021) to our setting. Specifically, we take the construction for linear MDP mentioned in Remark 5.8 of their paper, and rescale the rewards to be $1/H$ in the absorbing state. Their proof is based on embedding H independent linear bandit instances in a horizon H MDP, where the learner needs to solve $\Omega(H)$ of the bandit instances. It can be checked that Lemma C.8 of their analysis, which specifies the regret incurred in each bandit instance, now simply scales to be $\Omega(1/H)$, and the rest of the argument remains unchanged, leading to an overall lower bound of $\Omega(d\sqrt{HT})$ after T episodes for a horizon H time-inhomogenous linear MDP.

We note that all prior bonus-based methods suffer from a sub-optimal horizon and dimension scaling for linear MDPs. Given that this comes out of a consequence of a more general result here shows that our algorithm and analysis handle the uncertainty in our predictions in a sharp manner.

We also note that the Theorem 8 does not strictly require a linear MDP assumption, since we only require Assumption 1 to hold for $\mathcal{F}_{\mathsf{lin}}^h$, where we allow the error term ϵ to handle a small model misspecification in this definition. As a minor point, the instantiation of our general result from Theorem 6 requires us to instantiate $\mathcal{F}_{\mathsf{lin}}^h$ to be a finite ϵ_c-cover of the linear function class, and perform regression over this cover. It is also possible to directly analyze both the general and this special case in terms of a covering argument, which allows us to run our regressions directly on the original function class, which is preferable in practice.

4.2. Nonlinear Function Approximation

In the general setting of non-linear $\mathcal{F}$ with a bound on the generalized Eluder dimension, the ideas from the linear case can be extended by leveraging the techniques of Wang et al. (2020). They define a bonus function roughly as $b_t^h(z) = \max_{f \in \mathcal{F}_t^h} f(z) - \hat{f}_t^h(z)$. To avoid the high complexity arising with this definition, as remarked in Section 3 they instead consider an approximation to the class $\mathcal{F}_t^h$ in the maximization, defined using a subsampled set of the data. By using standard online subsampling arguments Kong et al. (2021), we can argue that the predictive differences between functions are preserved up to constant factors, while the amount of data required is significantly smaller. This procedure allows us to implement $\mathcal{B}$ as required in Definition 5 with $\log|\mathcal{W}| = O\left(\max_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h) \cdot \log\frac{T\mathcal{N}}{\delta} \log\frac{T|\mathcal{X}\times\mathcal{A}|}{\delta}\right)$. For linear MDPs, this gives an alternative method to build $\mathcal{W}$ so that $\log\mathcal{N}_b = \log|\mathcal{W}| = \widetilde{O}(d^2 \cdot \log|\mathcal{X}\times\mathcal{A}|)$. We refer the readers to Corollary 14 and Algorithm 2 in Appendix E for additional details on implementing the oracle and only state the main regret guarantee here.

Theorem 9 (Regret of VOQL using subsampling based bonus oracle $\mathcal{B}$) *Under conditions in Theorem 6, suppose the original function class $\mathcal{F}^h$ satisfies Assumption 1 with $T^2 H\epsilon = O(1)$, choosing*

$\lambda = 1$, $\alpha = \sqrt{1/TH}$, $\delta < 1/(T + H^2 + 11)$, $\epsilon_{\mathrm{b}} = 0$, *VOQL with the subsampling based bonus oracle* $\mathcal{B}$ *in Algorithm 2, and* $u_t = \widetilde{\Theta}\big(\log^{1.5} \mathcal{N} \cdot \log |\mathcal{X} \times \mathcal{A}| \sqrt{d_\alpha} \cdot \max_{h \in [H]} \dim_{\alpha,T}(\mathcal{F}^h) H^{5/2}/\sqrt{t}\big)$, *achieves a total regret of* $\mathbb{E}R_T = \widetilde{O}\big(\sqrt{\log \mathcal{N} d_\alpha HT} + \log^3 \mathcal{N} \cdot \log^2 |\mathcal{X} \times \mathcal{A}| d_\alpha \cdot \big(\max_{h \in [H]} \dim_{\alpha,T}(\mathcal{F}^h)\big)^2 H^5\big)$. *The regret bound also holds with probability at least* $1 - \delta$.

Apart from improving upon the dependence in H and d_α relative to the earlier result of Wang et al. (2020), as highlighted in Table 1, a key improvement in Theorem 9 is that the log-covering number of the state-action space only appears in the lower order term. This is due to the bonus parameter $\beta_{t,1}^h$ for the optimistic value function being independent of the size of the bonus class, and a key insight in our analysis.

5. Proof Sketch

In this section, we provide a proof sketch of Theorem 6. For simplicity we focus on proving an informal version of expected regret: we assume $\epsilon = 0$ in Assumption 1, $\epsilon_{\mathrm{b}} = 0$ in Definition 5, sufficiently small δ, and use $\widetilde{\cdot}$ to hide logarithmic factors in $T, H, 1/\alpha, 1/\delta$. We refer readers to Appendices F and G for the complete analysis. The proof works in the following three steps.

Step 1. Confidence interval and the good event.

Proposition 10 (Full in Proposition 33) *For our chosen* $\beta_{t,1}^h \approx \sqrt{\log \mathcal{N}}$, *in Algorithm 1 using a consistent bonus oracle as defined in Definition 5, with probability* $1 - \Theta(\delta)$ *we have* $\mathcal{T} f_{t,1}^{h+1}$ *satisfies* $\sum_{s \in [t-1]} \frac{1}{(\bar{\sigma}_s^h)^2} \Big((\mathcal{T} f_{t,1}^{h+1})(x_s^h, a_s^h) - \hat{f}_{t,1}^h(x_s^h, a_s^h) \Big)^2 \leq \big(\beta_{t,1}^h\big)^2$.

Showing $\mathcal{T} f_{t,1}^{h+1}$ satisfies this property, and thus lying in the defined confidence interval, requires a careful argument. We first ensure our chosen variance estimator remains a valid upper bound for the changing target, i.e. $\big(\bar{\sigma}_s^h\big)^2 \geq \mathbb{V}_{r^h, x^{h+1}}[r^h + f_{t,1}^{h+1}(x^{h+1})]$ for any $t \geq s$, which underpins the reason why we need to introduce the safe upper bounds and lower bounds $f_{\cdot,\pm 2}$ to define $\bar{\sigma}_s^h$. Further, we also separate the variance in two additive terms: one depending on $r + f_\star^{h+1}$, and the other on $f_{t,1}^{h+1} - f_\star^{h+1}$. The former has a lower complexity as dealing with a fixed target $f_\star^h$, while the second is lower-order by carefully choosing $\big(\bar{\sigma}_s^h\big)^2 \geq \widetilde{\Omega}(\log(\mathcal{N}\mathcal{N}_b)) \cdot (f_{s,2}^h(z_s^h) - f_{s,-2}^h(z_s^h)) \geq \widetilde{\Omega}(\log(\mathcal{N}\mathcal{N}_b)) \cdot (f_{t,2}^h(z_s^h) - f_{t,-2}^h(z_s^h)) \geq \widetilde{\Omega}(\log(\mathcal{N}\mathcal{N}_b)) \cdot \mathbb{V}_{,x^{h+1}}[f_{t,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})]$ due to monotonicity (see (6), or formally Lemma 31). This allows us to set $\beta_{t,1}^h \approx \sqrt{\log \mathcal{N}}$ instead of $\sqrt{\log \mathcal{N}\mathcal{N}_b}$, which lies central to the optimal dependence on d_α in the leading $\sqrt{T}$-order term.

Using a more standard martingale concentration analysis, we can also show with probability $1 - \Theta(\delta)$, $\mathcal{T} f_{t,\pm 2}^h$ and $\mathcal{T}_2 f_{t,1}^h$ all lie in the desired confidence intervals given $\beta_{t,2}^h, \bar{\beta}_t^h \approx \sqrt{\log \mathcal{N}\mathcal{N}_b}$. Under these conditions, the regret can be bounded by a usual sum of bonus terms $b_{t,1}^h$ when we are greedy wrt $f_{t,1}^h$ (see step 3). But on the remaining rounds, the bonus can be large and we separately bound the number of such rounds in the next step.

Step 2. Exploration rounds using $f_{t,2}^h$. In our exploration strategy Equation (9), we also sometimes need to use $f_{t,2}^h$ to ensure the variance upper bound $\bar{\sigma}_t^h$ estimated through $f_{t,2}^h$ is not too pessimistic while upper bounding $\mathbb{V}_{r^h, x^{h+1}}[r^h + f_{t,1}^{h+1}(x^{h+1})]$. In the second step, we show the agent only uses $f_{t,2}^h$ in the exploration *occasionally*: we denote the subset of all such iterations as $\mathcal{T}_{\mathrm{oo}} \subseteq [T]$.

Proposition 11 (Full in Lemma 41) *Given* $\alpha \le 1$*,* $u_t \ge \widetilde{\Omega}\left(\sqrt{\log \mathcal{N}} \cdot \log \mathcal{N}\mathcal{N}_b \cdot H^{5/2} \cdot \sqrt{d_\alpha}/\sqrt{t}\right)$*, it holds that* $\mathbb{E}\left[|\mathcal{T}_{\rm oo}||\mathcal{E}_{\le T}\right] \le \widetilde{O}\left(T/(\log(\mathcal{N}\mathcal{N}_b)\cdot H^3)\right)$*.*

Intuitively, when u_t is large enough, $f_{t,2}^h(x_t^h) \ge f_{t,1}^h(x_t^h) + u_t \ge V_t^h + u_t$ cannot happen too often, given the upper bound between $f_{t,2}^h - V_t^h$ shown in Lemma 36 in Appendix F.4. The particular threshold of u_t in Proposition 11 also serves the purpose that $\sum_{t\in[T],h\in[H]} u_t = \widetilde{O}\left(\mathrm{poly}(H, d_\alpha)\cdot\sqrt{T}\right)$, which we need to control when bounding the summation of bonus in the next step.

Step 3. Bounding the summation of bonus terms.

Proposition 12 (Full in Lemma 44) *Given* $\lambda = 1$*,* $\alpha = 1/\sqrt{TH}$ *and a bonus oracle* $\mathcal{B}$ *as in Definition 5, conditioning on good event in step 1, choosing* u_t *in step 2, bonus terms* $b_{t,1}^h$ *in Line 10 satisfy* $\mathbb{E}\sum_{t\in[T],h\in[H]} \min(1+L, b_{t,1}^h(z_t^h)) = \widetilde{O}\left(\sqrt{\log\mathcal{N}\cdot d_\alpha}\cdot\sqrt{HT} + \log\mathcal{N}\cdot\log^{1.5}\mathcal{N}\mathcal{N}_b\cdot d_\alpha\cdot H^5\right)$*.*

On the high level, we use the definition of $\beta_{t,1}^h$, Cauchy-Schwartz inequality, law of total variance (see Corollary 40) and the converging property of $\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)$ (see Lemma 43).

Combining the steps together we bound total regret via bounding $\sum_{t\in[T]}(f_{t,1}^1(x_t^1) - V_t^1)$, which boils down to $\sum_h \min(1+L, b_{t,1}^h(z_t^h))$ for most iterations, and larger bonus $\sum_h \min\left(1+L, b_{t,2}^h(z_t^h)\right)$ occasionally for $t \in \mathcal{T}_{\rm oo}$ (see Appendix F.5). Putting them together proves the main theorem.

6. Future Directions

In this work, we design VOQL, a new algorithm for time-inhomogeneous episodic reinforcement learning which achieves asymptotically-optimal regret for linear MDPs, and improved regret for general function class. Here we list a few directions as important open problems for future research following this work.

First-order regret bounds and tight low-order terms. In our analysis we didn't focus on analyzing the low-order factors in the tightest possible way: getting better low-order dependence could be possible using the proposed algorithm in the paper. Beyond, in order to get tight low-order terms, or to achieve a first-order (instance-dependent) regret bound as in Zanette and Brunskill (2019); Wagenmaker et al. (2022), one might need to introduce new component such as recursive high-order moment estimation (Zhang et al., 2021b) in the current algorithmic framework. Exploring alternative and potentially simpler algorithm which leads to better regret bounds are important directions for future research.

Horizon-free regret bounds for homogenous reinforcement learning. It is natural to ask if the horizon dependence can be removed if the episodic reinforcement learning process is homogenous, for both linear MDPs and the more general setting. We suspect that this is not feasible with the current approach as the regression functions across the levels do not share any parameters, and hence fail to leverage the homogenous structure of the problem.

Improved regret bounds for V-type function approximation. Another important and interesting open question is to design algorithms which enjoy better horizon dependence in the regret bounds for V-type function approximation Jin et al. (2021); Agarwal and Zhang (2022), a notion more general than Q-type approximation. This requires alternative algorithmic frameworks different from the optimistic Q-learning, since pointwise bonuses do not easily combine with V-type assumptions. Obtaining horizon-free guarantees in this setting largely remains open.

Acknowledgments

We thank anonymous reviewers for suggestions to improve the presentation and clarity of the paper. YJ was supported in part by the Danzig-Lieberman Graduate Fellowship. Part of the work was done while YJ was a research intern at Google Research, NY.

References

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

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

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

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

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

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

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

Xiequan Fan, Ion Grama, and Quansheng Liu. Martingale inequalities of type dzhaparidze and van zanten. *Statistics*, 51(6):1200–1213, 2017.

Fei Feng, Wotao Yin, Alekh Agarwal, and Lin Yang. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pages 3263–3273. PMLR, 2021.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

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

Claudio Gentile, Zhilei Wang, and Tong Zhang. Fast rates in pool-based batch active learning. 2022. URL https://arxiv.org/abs/2202.05448. short version appeared in ICML 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Minimax optimal reinforcement learning for discounted mdps. *arXiv preprint arXiv:2010.00587*, 2020.

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

Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via elliptical episodic bonuses. *arXiv preprint arXiv:2210.05805*, 2022.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR, 2022.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*, 2021.

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

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

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 2021.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *arXiv preprint arXiv:2111.03289*, 2021.

Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533, 2020.

Yuanzhi Li, Ruosong Wang, and Lin F Yang. 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, 2022.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634, 2021.

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

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34:6843–6855, 2021.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022.

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

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

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

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. URL http://tongzhang-ml.org/lt-book.html. in press, also available as http://tongzhang-ml.org/lt-book.html.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020a.

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

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

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021b.

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

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *arXiv preprint arXiv:2205.11507*, 2022.

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

Appendix A. Related Work

Improving regret bounds in RL problems. Azar et al. (2013, 2017) first studied the Bellman property of variance in tabular RL problems and used it for improving the horizon dependence in the sample complexity and regret bounds. In the tabular setting, Zanette and Brunskill (2019) later improved the result to further achieve problem-dependent regret bounds which match the classical regret (Azar et al., 2017) in the worst-case, but obtain horizon-free guarantees in the sparse reward case when the cumulative reward is at most 1 in any trajectory. Foster and Krishnamurthy (2021) also studied problem-dependent regret bounds for tabular MDPs. Additionally, there are works on obtaining fine-grained regret bounds in other RL settings, e.g. the data-dependent regret bounds for adversarial bandits and MDPs (Lee et al., 2020) and problem-dependent regret bounds for RL under linear function approximation (Wagenmaker et al., 2022).

Horizon-free bounds for the sparse reward setting (when total reward of each trajectory is independent of horizon length H) has received increased attention recently, starting from the COLT open problem posed in Jiang and Agarwal (2018). In the tabular setting, a line of work (Zanette and Brunskill, 2019; Zhang et al., 2021a, 2020a; Ren et al., 2021; Tarbouriech et al., 2021) designs algorithms that incur a poly-logarithmic in H regret, using tighter concentration bounds in their analysis. Another line of work further studies how to obtain completely horizon-free methods, at the cost of paying extra exponential (Li et al., 2022) or polynomial factors Zhang et al. (2022) in other problem parameters like state or action size. These ideas have been further generalized to linear mixture MDPs (Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022). However, the model-based approach they rely on is challenging to extend to model-free settings with function approximation. In particular, the challenge of designing a montonic variance upper bound in the face of a changing regression target does not arise in the model-based setting.

Linear function approximation. Linear MDPs have become a popular simple model for understanding function approximation beyond the tabular setting. Many works, such as Jin et al. (2020); Yang and Wang (2020); Zanette et al. (2020) obtain $O(\sqrt{T})$ regret bounds, and Zanette et al. (2020) obtain a $O(dH\sqrt{T})$ bound, which is optimal in the scaling with d, but sub-optimal in the scaling with H by a $\sqrt{H}$ factor. Table 1 provides more detailed comparisons. In terms of techniques, our approach is closest to that of Jin et al. (2020) in using a bonus based approach in a model-free setting, but incorporates weighted regression and other analysis improvements to obtain optimal guarantees. More closely related is the very recent result of Hu et al. (2022), who provide near-optimal regret bound for linear MDPs. Unfortunately, their analysis suffers from a technical issue that we explain in Appendix B, but their algorithm nevertheless contains many important design elements that we also incorporate. A key tool is the use of an *over-optimistic* and *over-pessimistic* value function estimates, in addition to a standard optimistic estimate, to help with bounding the variance of our regression targets. Like them, we also use the greedy policy of the over-optimistic function on certain rounds, as opposed to that of the optimistic function. Overall, these over-optimistic and over-pessimistic values, which we learn using unweighted regression, provide a safety net for our estimates and allow us to trade-off the amount of exploration with some extra regret by acting greedily with respect to the over-optimistic function.

Nonlinear function approximation. RL under nonlinear function approximation has gained increasing emphasis to model complex function spaces like neural networks, which are routinely used in empirical works. Several works have developed rank-based measures to capture the hardness of RL in such settings, in frameworks such as Bellman rank (Jiang et al., 2017), Bellman-Eluder dimension (Jin et al., 2021), Bilinear classes (Du et al., 2021) and DEC (Foster et al., 2021). Many of the upper bounds in these frameworks, however, yield sub-optimal guarantees when specialized to linear or tabular MDPs owing to their generality. An alternative approach builds on the Eluder dimension framework of Russo and Van Roy (2013), which has been extended to model-free RL in Wang et al. (2020). A related class of problems is that of small Q-type Bellman-Eluder dimension studied in Jin et al. (2021). Among these, the work of Wang et al. (2020) is closest to ours. Like their work, we also use model-free regression to estimate value functions, use an Eluder dimension style argument to control the exploration complexity, and use their sensitivity sampling argument to create a bonus oracle. However, we use weighted regression for function fitting, and correspondingly use a generalized Eluder dimension to handle such weighted objectives. Our generalization of the Eluder dimension is based on recent ideas from active learning in Gentile et al. (2022), although their work does not consider weighted settings. A related analysis of RL with non-linear function approximation with a similar definition of the Eluder dimension is also carried out in Zhang (2023), but they do not consider weighted regression and the guarantees are sub-optimal in d and H factors. Compared with Wang et al. (2020), we also avoid paying a state-action covering number in the leading order term in the regret in Table 1.

Appendix B. A Technical Issue in Hu et al. (2022)

The original version of Hu et al. (2022) designed an algorithm for linear MDPs and the main result in their paper states the method achieves a minimax-optimal regret of $\widetilde{O}\left(d\sqrt{HT}\right)$. Unfortunately, their analysis suffered from a technical mistake which we explain in detail here.

Their algorithm crucially relied on the assumption the over-optimistic values $\dot{\widehat{V}}_{i,h}(\cdot)$ upper bound the optimistic values $\widehat{V}_{i,h}(\cdot)$ point-wise with high probability. Specifically, Lemma D.2 and Equation (38) stated that $\dot{\widehat{V}}_{i,h}(s) \geq \widehat{V}_{j,h}(s)$, $\forall i \leq j$ and $s \in \mathcal{X}$, where we use their notations with s denoting state instead of x as in this paper. This was a critical condition needed in the proof.

However, the proof of Lemma D.2 in the original version of Hu et al. (2022) was incorrect. In particular, the authors used induction to prove Lemma D.2: Assuming the condition holds for $h+1$, they argued

$$\dot{\widehat{Q}}_{i,h}(s,a) - \widehat{Q}_{j,h}(s,a) = \mathcal{T}\widehat{V}_{i,h+1}(s,a) - \mathcal{T}\widehat{V}_{j,h+1}(s,a) \geq 0,$$

where we recall $\mathcal{T}\widehat{V}_{i,h+1}(s,a) = \mathbb{E}_{r^h,s^{h+1}}[r^h + \widehat{V}_{i,h+1}(s^{h+1})|s,a]$. Yet the authors did not explicitly prove the last inequality: the proof stated that the last inequality is directly from the induction hypothesis: $\dot{\widehat{V}}_{i,h+1}(\cdot) \geq \widehat{V}_{j,h+1}(\cdot)$, which was incorrect. In fact the inductive argument fails with their algorithm, which cannot be resolved by modifying the analysis. In this paper, we modify the overly optimistic value estimation procedure using unweighted regression (together with a different exploration policy) to ensure that the estimation made at any time i remains valid for all time $j \geq i$.

We note that after the submission of our work, Hu et al. (2022) updated their paper with the rare switching technique from the recent work of He et al. (2022) to resolve this technical issue for linear MDPs. With rare switching, the use of over optimism to ensure monotonicity can be replaced with an explicit maximization over all value functions from the history, as these functions only change rarely. This is identical to the arguments of He et al. (2020), which are developed independent of and concurrent with our work.

Appendix C. Proof for Generalized Eluder Dimension

Here we prove the remark comparing the generalized Eluder dimension defined in this paper (see Definition 2) with the standard Eluder dimension defined in literature (see, e.g. Russo and Van Roy (2013)). We first restate the remark for completeness.

Remark 4 (Relation to standard Eluder dimension) *When* $\boldsymbol{\sigma} \equiv 1$, $\max_{Z:|Z|=T} \dim(\mathcal{F}, Z, \mathbf{1}) \leq \dim_E(\mathcal{F}, \sqrt{\lambda/T})+1$, *where* $\dim_E(\mathcal{F}, \varepsilon)$ *is the standard Eluder dimension of* $\mathcal{F}$ *as defined in Russo and Van Roy (2013). The unweighted version of our definition has also appeared in Gentile et al. (2022). The generalized definition we give takes supremum over any* $\boldsymbol{\sigma} \geq \alpha$, *and thus is incomparable with the standard Eluder dimension even when* $\alpha = 1$.

Proof [Proof of Remark 4] Suppose $\dim_E(\mathcal{F}, \varepsilon) = n$. By definition of Eluder dimension, for any length T sequence Z, there are at most n distinct (sorted) indices $t_i \in [T]$, $i \in [n]$ such that for z_{t_i}, $D^2_{\mathcal{F}}(z_{t_i}, 1; z_{[t_i-1]}, \mathbf{1}_{[t_i-1]}) \geq \frac{\varepsilon^2}{\lambda}$. We then bound

$$\max_{Z:|Z|=T} \dim(\mathcal{F}, Z, \mathbf{1}) \leq \max_{Z:|Z|=T} \left(n + \frac{\varepsilon^2}{\lambda} T\right) = n + \frac{\varepsilon^2}{\lambda} T \leq \dim_E(\mathcal{F}, \varepsilon) + \varepsilon^2 \cdot \frac{T}{\lambda}.$$

Here for the first inequality we use the definition of $\dim(\mathcal{F}, Z, \mathbf{1})$ and the fact that only n terms in the summation of $\min(1, D^2_{\mathcal{F}})$ are upper bounded by 1 instead of $\frac{\varepsilon^2}{\varepsilon^2+\lambda}$ as argued above. Plugging in the choice of $\varepsilon = \sqrt{\lambda/T}$ concludes the proof. ∎

Appendix D. Proofs for Linear Function Approximation

Here we provide the full proofs of several properties of linear function class as stated in the main paper.

The first property is about the Eluder dimension for linear MDPs Russo and Van Roy (2013).

Lemma 3 *For the class* $\mathcal{F}^h_{\mathsf{lin}}$ *defined in (2), letting* $\mathcal{F}^h_{\mathsf{lin}}(\epsilon_c)$ *be the* ϵ_c*-cover of* $\mathcal{F}^h_{\mathsf{lin}}$ *for some* $\epsilon_c > 0$, *we have* $\dim_{\alpha,T}(\mathcal{F}^h_{\mathsf{lin}}(\epsilon_c)) \leq \dim_{\alpha,T}(\mathcal{F}^h_{\mathsf{lin}}) = O\big(d \log\big(1 + \frac{(B^h)^2 T}{\alpha^2 d\lambda}\big)\big) = \widetilde{O}(d)$.

Proof [Proof of Lemma 3] Fix $h \in [H]$. Recalling the definitions of linear function classes $\mathcal{F}^h_{\mathsf{lin}}$, we can simplify the definition of generalized Eluder dimension to be the follows:

$$\left(\bar{\sigma}^h\right)^{-2} D^2_{\mathcal{F}^h_{\mathsf{lin}}}(z^h; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}) \leq \left(\bar{\sigma}^h\right)^{-2} D^2_{\mathcal{F}^h_{\mathsf{lin}}}(z^h; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}) = \left\|\frac{1}{\bar{\sigma}^h}\phi^h(z^h)\right\|^2_{\left(\Sigma^h_t\right)^{-1}},$$

$$\text{where} \quad \Sigma^h_t := \frac{\lambda}{\left(B^h\right)^2} I + \sum_{s\in[t-1]} \frac{1}{\left(\bar{\sigma}^h_s\right)^2}\phi^h(z^h_s)\left(\phi^h(z^h_s)\right)^\top.$$

Consequently, in this case we can bound for any $\boldsymbol{\sigma} \geq \alpha$ that

$$\begin{aligned}\dim(\mathcal{F}_{\mathsf{lin}}^h, \mathcal{Z}, \sigma) &\overset{(i)}{\leq} \sum_{t\in[T]} \frac{2\left\|\frac{1}{\bar{\sigma}^h}\phi^h(z_t^h)\right\|^2_{(\Sigma_t^h)^{-1}}}{1+\left\|\frac{1}{\bar{\sigma}^h}\phi^h(z_t^h)\right\|^2_{(\Sigma_t^h)^{-1}}} \overset{(ii)}{=} \sum_{t\in[T]} 2\left\|\frac{1}{\bar{\sigma}^h}\phi^h(z_t^h)\right\|^2_{(\Sigma_{t+1}^h)^{-1}} \\ &\overset{(iii)}{=} 2\sum_{t\in[T]}\left(\log\det(\Sigma_{t+1}^h) - \log\det(\Sigma_t^h)\right) = O\left(\log\left|\frac{\left(B^h\right)^2}{\lambda}\Sigma_T^h\right|\right) \\ &= O\left(d\log\left(1+\frac{\left(B^h\right)^2 T}{\alpha^2 d\lambda}\right)\right).\end{aligned}$$

Here we use (i) the inequality that $\min(1,x) \leq \frac{2x}{1+x}$ for any $x \geq 0$, (ii) the Sherman-Morrison formula, and (iii) writing $\left\|\frac{1}{\bar{\sigma}^h}\phi^h(z_t^h)\right\|^2_{(\Sigma_{t+1}^h)^{-1}} = \mathrm{trace}\left(\left(\Sigma_{t+1}^h\right)^{-1}(\Sigma_{t+1}^h - \Sigma_t^h)\right) = \log\det(\Sigma_{t+1}^h) - \log\det(\Sigma_t^h)$. This is the classical bound of summation of Elliptical bonuses, see e.g. Lemma 11 of Abbasi-Yadkori et al. (2011). The above inequality shows that

$$\dim_{\alpha,T}(\mathcal{F}_{\mathsf{lin}}^h) = O\left(d\log\left(1+\frac{\left(B^h\right)^2 T}{\alpha^2 d\lambda}\right)\right) = \widetilde{O}(d).$$

Since by definition any ϵ_{c}-cover of $\mathcal{F}_{\mathsf{lin}}^h(\epsilon_{\mathrm{c}})$ is just a subset of $\mathcal{F}_{\mathsf{lin}}^h$, we have $\dim_{\alpha,T}(\mathcal{F}_{\mathsf{lin}}^h(\epsilon_{\mathrm{c}})) = \widetilde{O}(d)$ by definition of generalized Eluder dimension (Definition 2). ■

The next lemma shows that standard elliptical bonus functions satisfy the definition of bonus oracle $\mathcal{B}$.

Lemma 7 (Bonus oracle $\mathcal{B}$ for linear MDPs) *Given $T, H \in \mathcal{Z}_+$, suppose all $\beta_t^h \leq \beta$ and β_t^h is non-decreasing in $t \in [T]$ for each $h \in [H]$. For any $t \geq 1$, $h \in [H]$, variances $\{\bar{\sigma}_s^h\}_{s\in[t-1]}$ satisfying $\bar{\sigma}_s^h \geq \alpha$ for some $\alpha > 0$, dataset $\mathcal{D}_{[t-1]}^h = \{(\phi^h(z_s^h), a_s^h, r_s^h, \phi(z_s^{h+1}))\}_{s\in[t-1]}$, function class $\mathcal{F}_t^h$ and $\hat{f}_t^h \in \mathcal{F}_t^h$ defined via weighted regression (5), and parameters $\lambda, \epsilon_{\mathrm{c}} > 0$, let $\mathcal{B}(\{\bar{\sigma}_s^h\}_{s\in[t-1]}, \mathcal{D}_{[t-1]}^h, \mathcal{F}_t^h, \hat{f}_t^h, \beta_t^h, \lambda, \epsilon_{\mathrm{c}}) = \|\phi^h(x,a)\|_{(\Sigma_t^h)^{-1}}\sqrt{(\beta_t^h)^2+\lambda}$, where $\Sigma_t^h = \frac{\lambda}{4(B^h)^2}I + \sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\phi^h(z_s^h)\phi^h(z_s^h)^\top$. For any choice of covering radius $\epsilon_{\mathrm{c}} \leq \alpha\sqrt{\lambda/8T}$, the oracle satisfies all the properties of Definition 5 with $\log\mathcal{N}_b = \log|\mathcal{W}| = O\big(d^2\log\big(1+B^2\sqrt{d}\beta/(\lambda\epsilon_{\mathrm{c}}^2)\big)\big)$.*

Proof [Proof of Lemma 7] By definition of the class $\mathcal{F}_{\mathsf{lin}}^h(\epsilon_{\mathrm{c}})$, we note that for any z^h:

$$|f^h(z^h) - \hat{f}_t^h(z^h)| \leq \|w-\hat{w}_t\|_{\Sigma_t^h}\|\phi^h(z^h)\|_{(\Sigma_t^h)^{-1}},$$

where w and $\hat{w}_t$ are the weight parameters underlying f^h and $\hat{f}_t^h$ respectively. By the definition (5) of the class $\mathcal{F}_t^h$, we have that for any $f \in \mathcal{F}_t^h$, defining the bonus as $\|\phi^h(x,a)\|_{(\Sigma_t^h)^{-1}}\sqrt{(\beta_t^h)^2+\lambda}$ verifies the second property of Definition 5.

For the third property, we note there must exist $\Delta w_\star$ satisfying $\|\Delta w_\star\| = 2B^h$ and that

$$\|\phi^h(z^h)\|^2_{(\Sigma_t^h)^{-1}} = \frac{\Delta w_\star^\top\phi^h(z^h)\phi^h(z^h)^\top\Delta w_\star}{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\Delta w_\star^\top\phi^h(z_s^h)\phi^h(z_s^h)^\top\Delta w_\star + \lambda}.$$

Thus, by assumption of $\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})$ being an ϵ_{c}-cover we can find $\tilde{\Delta} w_\star = w - w'$ for $w, w' \in \mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})$ such that $\langle \phi^h(z^h), \tilde{\Delta} w_\star - \Delta w_\star \rangle \le 2\epsilon_{\mathrm{c}}$ for all z^h and thus

$$\begin{aligned}
&\frac{\Delta w_\star^\top \phi^h(z^h)\phi^h(z^h)^\top \Delta w_\star}{\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}^h_s)^2} \Delta w_\star^\top \phi^h(z^h_s)\phi^h(z^h_s)^\top \Delta w_\star + \lambda} \\
&\quad \le \frac{2\tilde{\Delta} w_\star^\top \phi^h(z^h)\phi^h(z^h)^\top \tilde{\Delta} w_\star + 2\cdot(2\epsilon_{\mathrm{c}})^2}{\frac{1}{2}\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}^h_s)^2} \tilde{\Delta} w_\star^\top \phi^h(z^h_s)\phi^h(z^h_s)^\top \tilde{\Delta} w_\star - (\frac{2\epsilon_{\mathrm{c}}}{\alpha})^2(t-1) + \lambda} \\
&\quad \overset{(\star)}{\le} 4\frac{\tilde{\Delta} w_\star^\top \phi^h(z^h)\phi^h(z^h)^\top \tilde{\Delta} w_\star}{\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}^h_s)^2} \tilde{\Delta} w_\star^\top \phi^h(z^h_s)\phi^h(z^h_s)^\top \tilde{\Delta} w_\star + \lambda} + 16\epsilon_{\mathrm{c}}^2/\lambda \\
&\quad \le 4 \sup_{w,w'\in\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})} \frac{(w-w')^\top \phi^h(z^h)\phi^h(z^h)^\top (w-w')}{\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}^h_s)^2} (w-w')^\top \phi^h(z^h_s)\phi^h(z^h_s)^\top (w-w') + \lambda} + 16\epsilon_{\mathrm{c}}^2/\lambda,
\end{aligned}$$

where for inequality $(\star)$ we use the choice of ϵ_{c} so that $\epsilon_{\mathrm{c}}^2 \le \frac{\alpha^2\lambda}{8T}$. Thus by taking square root on both sides, we have

$$\|\phi^h(z^h)\|_{(\Sigma^h_t)^{-1}} \le 2\bar{\sigma}^h_t \cdot D_{\mathcal{F}^h_{\mathsf{lin}}(\epsilon_{\mathrm{c}})}(z^h, \bar{\sigma}^h_t; z_{[t-1]}, \bar{\sigma}^h_{[t-1]}) + 4\epsilon_{\mathrm{c}}/\sqrt{\lambda},$$

which by multiplying over $\sqrt{(\beta^h_t)^2 + \lambda}$ on both sides proves the third property.

Since the matrix Σ^h_t is data dependent, the standard way to specify the bonus class is to parameterize it by all possible choices of the matrix in the Mahalanobis norm (Jin et al., 2020) , which means that the class $\mathcal{W}$ consists of all bonus functions of the form

$$b^h(z^h) \in \left\{ \left\|\phi^h(z^h)\right\|_A | \text{where } A \in \mathcal{C}_A \right\}, \text{ for any } h \in [H],$$
$$\text{where } \mathcal{C}_A \text{ is an } \epsilon_{\mathrm{c}}^2\text{-cover of } \left\{ A \in \mathbb{R}^{d\times d}, \ \|A\|_{\mathsf{F}} \le \frac{4(B^h)^2\sqrt{d}}{\lambda} \cdot \sqrt{\beta^2+\lambda} \right\}.$$

By standard argument we have the size of $\log|\mathcal{C}_A|$ is bounded by $O\left(d^2 \log\left(1 + \frac{(B^h)^2\sqrt{d}\cdot\beta}{\lambda\epsilon_{\mathrm{c}}^2}\right)\right)$. Thus, by definition of $\mathcal{W}$ we also have $\log|\mathcal{W}| = O\left(d^2 \log\left(1 + \frac{B^2\sqrt{d}\cdot\beta}{\lambda\epsilon_{\mathrm{c}}^2}\right)\right)$.

The consistency of the oracle follows from the fact that β^h_t is non-decreasing in t element-wise for each $h \in [H]$, thus completing the proof. ∎

Appendix E. Implementing Bonus Oracle $\mathcal{B}$ using Online-subsampling

The guarantees of Algorithm 1 hold assuming a consistent bonus oracle $\mathcal{B}$ satisfying Definition 5. To implement such an oracle, we follow the online sensitivity sub-sampling approach described in Kong et al. (2021), which is a follow-up of the original sensitivity sub-sampling idea proposed in Wang et al. (2020).

For completeness here we restate this sub-sampling procedure in Algorithm 2 and its guarantees in Proposition 13. [6]

6. Different from the original result, we don't consider a cover class since we are already working with a finite function class.

We first define the weighted dataset $\mathcal{Z}$ so that each element in it is $(z, \bar{\sigma}(z))$ and define the *weighted sensitivity score* as

$$\text{sensitivity}_{\mathcal{Z},\mathcal{F},\beta,\alpha}(z) = \min\left\{ \sup_{f,f'\in\mathcal{F}} \frac{\frac{1}{\bar{\sigma}^2(z)}\left(f(z)-f'(z)\right)^2}{\min\left\{\sum_{z'\in\mathcal{Z}} \frac{1}{\bar{\sigma}^2(z')}\left(f(z')-f'(z')\right)^2, \frac{T(H+1)^2}{\alpha^2}\right\}+\beta^2}, 1\right\}$$

For the weighted dataset $\mathcal{Z}^h_{[t-1]} = \{(x^h_s, a^h_s), \bar{\sigma}^h_s\}_{s\in[t-1]}$, we define $\|f\|^2_{\mathcal{Z}} = \sum_{z\in\mathcal{Z}} \frac{1}{\bar{\sigma}^2(z)} f^2(z)$, i.e. weighted sum of ℓ_2-norm square. Now we introduce the sub-sampling procedure.

Algorithm 2: Online Sensitivity Sub-sampling with Weights

Input function class $\mathcal{F}$, current sub-sampled dataset $\hat{\mathcal{Z}} \subseteq \mathcal{X}\times\mathcal{A}$, new state-action pair z, parameter β, threshold $\alpha > 0$, failure probability δ

Parameter $1 \le C < \infty$

Let p_z be the smallest real number such that

$$1/p_z \text{ is an integer and } p_z \ge \min\left(1, C\cdot \text{sensitivity}_{\mathcal{Z},\mathcal{F},\beta,\alpha}(z)\cdot\log(T\mathcal{N}/\delta)\right)$$

Independently add $1/p_z$ copies of $(z, \bar{\sigma}(z))$ into $\hat{\mathcal{Z}}_+$ with probability p_z

Return $\hat{\mathcal{Z}}_+$.

This algorithm can be called at every step t, h with $\hat{\mathcal{Z}}^h_{[t-1]}$ and the new data z^h_t to obtain the next $\hat{\mathcal{Z}}^h_{[t]}$. Below we will refer to the original dataset as $\mathcal{Z}^h_{[t-1]} = \{(x^h_s, a^h_s), \bar{\sigma}^h_s\}_{s\in[t-1]}$, and $\hat{\mathcal{Z}}^h_{[t-1]}$ as the dataset subsampled from $\mathcal{Z}^h_{[t-1]}$.

Proposition 13 (Guarantees in weighted case, generalizing Proposition 1 and 2 in Kong et al. (2021)) *When $\bar{\sigma}(z) \ge \alpha$ always holds for any z, with probability $1-\delta$, it holds that*

$$\begin{aligned}\sup_{f_1,f_2:\|f_1-f_2\|^2_{\mathcal{Z}^h_t}\le\beta^2} |f_1(z)-f_2(z)| &\le \sup_{f_1,f_2:\|f_1-f_2\|^2_{\hat{\mathcal{Z}}^h_t}\le 10^2\beta^2} |f_1(z)-f_2(z)| \\ &\le \sup_{f_1,f_2:\|f_1-f_2\|^2_{\mathcal{Z}^h_t}\le 10^4\beta^2} |f_1(z)-f_2(z)|.\end{aligned}$$

Further, for each $h\in[H]$, the number of different elements in sub-sampled dataset $\hat{\mathcal{Z}}^h_t$ ($t = 1, 2, \cdots, T$) is always bounded by $S_{\max} = O\left(\log\frac{T\mathcal{N}}{\delta}\cdot\max_{h\in[H]}\dim_{\alpha,T}(\mathcal{F}^h)\right)$ and the total size (counting repetitions) is bounded by $O(T^3/\delta)$.

This leads to the formal guarantee of implementing $\mathcal{B}$ in Definition 5 for general function class.

Corollary 14 (Implementing $\mathcal{B}$ using online-subsampling) *There exists an algorithm (see Algorithm 2) that with probability $1-\delta$ implements a consistent bonus oracle $\mathcal{B}$ with $\epsilon_{\mathrm{b}} = 0$ for all iterations $t\in[T]$, $h\in[H]$ where $\log|\mathcal{W}| \le O\left(\max_{h\in[H]}\dim_{\alpha,T}(\mathcal{F}^h)\cdot\log\frac{T\mathcal{N}}{\delta}\log\frac{T|\mathcal{X}\times\mathcal{A}|}{\delta}\right)$.*

We remark that here we state a slightly generalized version adapted to weighted regression, which includes unweighted regression stated in Kong et al. (2021) as a special case when we set $\bar{\sigma}(z)\equiv 1$. The result follows a straightforward generalization from Kong et al. (2021) taking weights into consideration.

Appendix F. Full Analysis of Theorem 6

This section provides a full proof for the bound on the expected regret in Algorithm 1. This is stated in Theorem 6 in the main paper and here we first state a more formal version of that theorem.

Theorem 15 (Bound on expected regret) *Suppose function class* $\{\mathcal{F}^h\}_{h\in[H]}$ *satisfies Assumption 1 with* $\epsilon \in [0,1]$ *and Definition 2 with* $\lambda = 1$*, and given consistent bonus oracle* $\mathcal{B}$ *(output function in class* $\mathcal{W}$*) satisfying Definition 5, VOQL with* $\alpha = \sqrt{1/TH}$*,* $\delta < 1/(T+10)$*,* $\epsilon \le 1$ *and*

$$u_t = C \cdot \left(\frac{\sqrt{\log \frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\log \frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H^{5/2}\sqrt{d_\alpha} + \sqrt{t}H\epsilon_{\mathrm{b}}\right)}{\sqrt{t}} + H^2\epsilon + H\delta \right)$$

for sufficiently large constant $C < \infty$*, achieves a total regret* $\mathbb{E}R_T = O\Bigg(\sqrt{\log \frac{\mathcal{N}TH}{\delta} + T^2H\epsilon} \cdot \sqrt{THd_\alpha} + \left(\log \frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right) \cdot \left(\log^2 \frac{\mathcal{N}\mathcal{N}_b TH}{\delta} \cdot H^5 d_\alpha + T^2\epsilon_{\mathrm{b}}^2\right)\Bigg)$.

The section is organized as follows: We first introduce some general notations, definitions, and helper lemmas that will be used throughout the proof in this section in Appendix F.1. In Appendix F.2, we prove the properties of constructed confidence intervals $\mathcal{F}^h_{t,j}$, $j = 1, \pm 2$ and $\mathcal{G}^h_t$. In Appendix F.3 we prove the key point-wise monotonicity property and also some properties of our chosen variance estimator $\left(\sigma^h_t\right)^2$ defined in Equation (4). In Appendix F.4, we show the approximation error between our constructed values $f^h_{t,j}$, $j = 1, \pm 2$ with respect to the true expected reward V^h_t. In Appendix F.5 we provide the formal proofs for bounding the regret. We refer readers to Appendix G for the complete theorem statement and full proof for the *high-probability* regret bound.

In Table 2 we summarize the main notations used in the paper. We also provide the concrete choices of parameters of Algorithm 1 for obtaining the claimed regret bounds in Table 3.

F.1. Notations and Preliminaries

Here we briefly give self-contained notations and definitions used throughout the proof. We summarize the main notations used in Table 2 and the specific choice of parameters in Table 3 for easy reference.

Iterates and functions. In general, we use $z = (x, a)$, $z^h = (x^h, a^h)$ and $z^h_t = (z^h_t, a^h_t)$ interchangeably. We also use $f^h_\star \in \mathcal{F}^h$ to denote either the optimal Q-value function $Q^h_\star$ or the optimal V-value function $V^h_\star$ when clear from context.

As in the main paper, we will use the notation $\mathcal{T}f(z^h) = \mathbb{E}_{r^h,x^{h+1}}[r^h + f(x^{h+1})|z^h]$ and also $\mathcal{T}_2 f(z^h) = \mathbb{E}_{r^h,x^{h+1}}\left[\left(r^h + f(x^{h+1})\right)^2|z^h\right]$ to be the conditional expectation of future values and their second moment under any function f at level h and state-action pair z^h.

Notation	Meaning	Remark
$\mathcal{X}, \mathcal{A}$	state space, action space	
t, h	$t \in [T]$ trajectory/step, $h \in [H+1]$ level	
r_t^h, x_t^h, a_t^h	reward, state and action at step t and level h	
r^h, x^h, a^h	random reward, state and action at step t and level h	
z	shorthand for state-action pair (x, a)	
$\mathcal{D}_{[t-1]}^h$	data set $\{(x_s^h, a_s^h, r_s^h, x_s^{h+1})\}_{s\in[t-1]}$	
$\mathcal{F}^h$	function class for $h \in [H]$	Ass. 1
$\mathcal{F}_{\mathsf{lin}}^h$	general linear function class	Eqn. (2)
$\mathcal{F}_{\mathsf{lin}}^h(\epsilon_c)$	ϵ_c-cover of general linear function class $\mathcal{F}_{\mathsf{lin}}^h$	Eqn. (2)
$\mathcal{W}$	bonus function class defined for $\mathcal{B}$	Def. 5
ϵ_b	error paremeter for bonus oracle	Eqn. (2)
$\mathcal{N}$	maximal size of function class $\max_{h\in[H]} \lvert\mathcal{F}^h\rvert$	
$\mathcal{N}_b$	size of bonus function class $\lvert\mathcal{W}\rvert$	Def. 5
$D_{\mathcal{F}}(z; z_{[t-1]}, \sigma_{[t-1]})$	$:= \sqrt{\sup_{f_1,f_2\in\mathcal{F}} \frac{(f_1(z)-f_2(z))^2}{\sum_{s\in[t-1]} \frac{1}{\sigma_s^2}(f_1(z_s)-f_2(z_s))^2+\lambda}}$	λ param.
$\dim_{\alpha,T}(\mathcal{F})$	generalized Eluder dimension defined in Definition 2	α param.
d_α	shorthand for $\frac{1}{H}\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)$ (Definition 2)	α param.
$f_{t,1}^h$	optimistic value function at step t, h	
$\hat{f}_{t,1}^h$	solution of fitting weighted regression at step t, h	Eqn. (11)
$\bar{f}_{t,1}^h$	in $\mathcal{F}^h$ and $\max_{z^h} \left\lvert \bar{f}_{t,1}^h(z^h) - \mathbb{E}\left[r^h + f_{t,1}^{h+1}(x^{h+1})\vert z^h\right]\right\rvert \le \epsilon$	Ass. 1
$\mathcal{F}_{t,1}^h$	version space of optimistic value functions at step t, h	Eqn. (11)
$f_{t,2}^h$	overly optimistic value function at step t, h	
$f_{t,-2}^h$	overly pessimistic value function at step t, h	
$\hat{f}_{t,\pm 2}^h$	solution of fitting unweighted regression at step t, h	Eqn. (16)
$\bar{f}_{t,\pm 2}^h$	in $\mathcal{F}^h$ and $\max_{z^h} \left\lvert \bar{f}_{t,j}^h(z^h) - \mathbb{E}\left[r^h + f_{t,j}^{h+1}(x^{h+1})\vert z^h\right]\right\rvert \le \epsilon$	Ass. 1
$\mathcal{F}_{t,\pm 2}^h$	version space of overly optimistic(pessimistic) value functions at t, h	Eqn. (16)
$\hat{g}_t^h$	solution of fitting second-moment regression at step t, h	Eqn. (19)
ψ_t^h	in $\mathcal{F}^h$ and $\max_{z^h} \left\lvert \psi_t^h(z^h) - \mathbb{E}\left[\left(r^h + f_{t,2}^{h+1}(x^{h+1})\right)^2 \vert z^h\right]\right\rvert \le \epsilon$	Ass. 1
$\mathcal{G}_t^h$	version space of second-moment estimates at t, h	Eqn. (19)
$\mathcal{E}_t^h$	event that $\{\bar{f}_{t,j}^h \in \mathcal{F}_{t,j}^h$ for $j = 1, \pm 2$ and $\psi_t^h \in \mathcal{G}_t^h\}$	
$\mathcal{E}_t, \mathcal{E}_{\le t}$	joint event that $\cap_{h\in[H]}\mathcal{E}_t^h$ or $\cap_{s\in[t]}\cap_{h\in[H]}\mathcal{E}_s^h$	
$h_t \in [H+1]$	random h when starting to take greedy w.r.t $f_{t,2}^h$ at step t	
$\mathcal{T}_o, \mathcal{T}_{oo}$	disjoint subsets of $[T]$ when $h_t = H+1$ or $h_t \in [H]$	
V_t^h	expected reward during exploration at time t from step h onwards	
$V_\star^h, Q_\star^h$	optimal V-value or Q-value function at level h	
$f_\star^h$	equivalent to $Q_\star^h$ or $V_\star^h$ depending on the context	
$\xi_{t,j}^h$	$:= r_t^h + f_{t,j}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h,x^{h+1}}\left[r^h + f_{t,j}^{h+1}(x^{h+1})\vert z_t^h\right]$ for $j = 1, \pm 2$	
ξ_t^h	$:= r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1}\vert z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]$	
$b_{t,j}^h$	bonus term obtained in Lines 10 and 14 using $\mathcal{B}$, for $j = 1, 2$	Def. 5
$\mathcal{T}$	Bellman operator $\mathcal{T}f(z^h) = \mathbb{E}[r^h + f(x^{h+1})\vert z^h]$	
$\mathcal{T}_2$	second-moment operator $\mathcal{T}_2 f(z^h) = \mathbb{E}\left[\left(r^h + f(x^{h+1})\right)^2 \vert z^h\right]$	

Table 2: **Summary of notations.** Here we use $\mathbb{E}[\cdot\vert z^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot\vert z^h]$ for brevity.

Parameter	Choice	Remark
δ	$\delta < 1/(T+10)$ in Theorem 15, $\delta < 1/(H^2+11)$ in Theorem 45	
$\delta_{t,h}$	$\delta/(T+1)(H+1)$	Eqn. (14)
ϵ	$\epsilon \in [0,1]$, model class misspecification error	Ass. 1
ϵ_{c}	error due to taking covering of function class	
α	$\sqrt{1/TH}$	Def. 2
λ	1	Def. 2
$\upsilon(\delta)$	$\sqrt{\log \frac{\mathcal{N}^2(2\log(4LT/\alpha)+2)(\log(8L/\alpha^2)+2)}{\delta}}$	Eqn. (14)
$\iota(\delta)$	$3\sqrt{\log \frac{\mathcal{N}\mathcal{N}_b(2\log(4LT/\alpha)+2)(\log(8L/\alpha^2)+2)}{\delta}}$	Eqn. (15)
$\beta_{t,1}^h$	$\sqrt{\left(6\sqrt{\lambda}+156\right)} \cdot \sqrt{\log \frac{\mathcal{N}^2(T+1)(H+1)\left(2\log\frac{4LT}{\alpha}+2\right)\left(\log\frac{8L}{\alpha^2}+2\right)}{\delta}} + \sqrt{\frac{8tL}{\alpha^2}\cdot\epsilon}$	Eqn. (13)
$i(\delta)$	$\sqrt{2\log \frac{\mathcal{N}\mathcal{N}_b(2\log(18LT)+2)(\log(18L)+2)}{\delta}}$	Eqn. (18)
$\beta_{t,2}^h$	$\sqrt{2\,(24L+21)\,i^2(\delta_{t,h}) + 20tL\epsilon}$	Eqn. (17)
$\iota'(\delta)$	$\sqrt{2\log \frac{\mathcal{N}\mathcal{N}_b(2\log(32LT)+2)(\log(32L)+2)}{\delta}}$	Eqn. (21)
$\bar{\beta}_t^h$	$\sqrt{8(11+9L)\,(\iota'(\delta_{t,h}))^2 + 32tL\epsilon}$	Eqn. (20)
$\left(\sigma_t^h\right)^2$	$\min\left(4, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h)\cdot\left(\sqrt{\left(\bar{\beta}_t^h\right)^2+\lambda} + 2L\sqrt{\left(\beta_{t,2}^h\right)^2+\lambda}\right) + \hat{g}_t^h(z_t^h) - \left(\hat{f}_{t,-2}^h(z_t^h)\right)^2 + 2(1+L)\epsilon\right)$ for $t \geq 2$	Eqn. (4)
$\bar{\sigma}_t^h$	$\max\Big\{\sigma_t^h, \alpha, \sqrt{2}\iota(\delta_{t,h})\sqrt{f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)}, 2\left(\sqrt{\upsilon(\delta_{t,h})}+\iota(\delta_{t,h})\right)\cdot\sqrt{D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}\Big\}$	Eqn. (12)

Table 3: **Summary of parameter choices.**

We recall the definition of $f_{t,j}$ for $j = 1, \pm 2$ in Algorithm 1 that

$$
\begin{aligned}
f_{t,1}^h(\cdot) &:= \min\left(\hat{f}_{t,1}^h(\cdot) + b_{t,1}^h(\cdot) + \epsilon, 1\right), \\
f_{t,2}^h(\cdot) &:= \min\left(\hat{f}_{t,2}^h(\cdot) + 2b_{t,1}^h(\cdot) + b_{t,2}^h(\cdot) + 3\epsilon, 2\right), \\
f_{t,-2}^h(\cdot) &:= \max\left(\hat{f}_{t,-2}^h(\cdot) - b_{t,2}^h(\cdot) - \epsilon, 0\right),
\end{aligned} \tag{10}
$$

where $\hat{f}_{t,j}^h$ is the center of the constructed confidence interval (see next paragraph for concrete definitions) respectively. For each $f_{t,j}^h$ with $j = 1, \pm 2$, we will let $\bar{f}_{t,j}^h(\cdot) \in \mathcal{F}^h$ to approximate the conditional expectation with target $f_{t,j}^{h+1}$, i.e. $\max_{z^h}\left|\bar{f}_{t,j}^h(z^h) - \mathcal{T} f_{t,j}^{h+1}(z^h)\right| \leq \epsilon$ (such $\bar{f}$ exists due to Assumption 1). Similarly, we will let $\psi_t^h(\cdot) \in \mathcal{F}^h$ to approximate the conditional expectation of second moment with target $f_{t,2}^{h+1}$, i.e. $\max_{z^h} \left|\psi_t^h(z^h) - \mathcal{T}_2 f_{t,1}^{h+1}(z^h)\right| \le \epsilon$ (existence is guaranteed similarly).

Regression and confidence intervals. We define the following (weighted) regression problems and their induced confidence intervals at each step $t \in [T]$, $h \in [H]$ when the dataset $\mathcal{D}_{[t-1]}^h := \{(x_s^h, a_s^h, r_s^h, x_s^{h+1})\}_{s\in[t-1]}$ is given. Throughout we use $\mathcal{N} := \max_{h\in[H]} |\mathcal{F}^h|$ and $\mathcal{N}_b := |\mathcal{W}|$ to denote the sizes of the function class $\mathcal{F}^h, h \in [H]$ and the bonus function class $\mathcal{W}$.

The weighted regression problem for fitting optimistic value functions $f_{t,1}^h$ is:

$$\begin{aligned} \hat{f}_{t,1}^h &= \arg\min_{f^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \frac{\left(f^h\left(x_s^h, a_s^h\right) - r_s^h - f_{t,1}^{h+1}\left(x_s^{h+1}\right)\right)^2}{\left(\bar{\sigma}_s^h\right)^2}, \\ \text{and let } \mathcal{F}_{t,1}^h &:= \left\{ f^h \in \mathcal{F}^h : \sum_{s\in[t-1]} \frac{1}{\left(\bar{\sigma}_s^h\right)^2} \left(f^h(x_s^h, a_s^h) - \hat{f}_t^h(x_s^h, a_s^h)\right)^2 \le \left(\beta_{t,1}^h\right)^2 \right\}. \end{aligned} \tag{11}$$

The parameters are as follows (for $t \ge 2$):

$$\begin{aligned} \bar{\sigma}_t^h := \max\Big\{&\sigma_t^h, \alpha, \sqrt{2}\iota(\delta_{t,h})\sqrt{f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)}, \\ &2\left(\sqrt{\upsilon(\delta_{t,h})} + \iota(\delta_{t,h})\right)\cdot\sqrt{D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}\Big\}, \end{aligned} \tag{12}$$

$$\beta_{t,1}^h := \sqrt{\left(6\sqrt{\lambda} + 156\right)} \cdot \sqrt{\log \frac{\mathcal{N}^2(T+1)(H+1)\left(2\log\frac{4LT}{\alpha} + 2\right)\left(\log\frac{8L}{\alpha^2} + 2\right)}{\delta}} + \sqrt{\frac{8tL}{\alpha^2} \cdot \epsilon}, \tag{13}$$

$$\delta_{t,h} := \frac{\delta}{(T+1)(H+1)} \quad , \quad \upsilon(\delta_{t,h}) := \sqrt{\log\frac{\mathcal{N}^2\left(2\log(4LT/\alpha) + 2\right)\left(\log(8L/\alpha^2) + 2\right)}{\delta_{t,h}}}, \tag{14}$$

$$\text{and} \quad \iota(\delta_{t,h}) := 3\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(4LT/\alpha) + 2\right)\left(\log(8L/\alpha^2) + 2\right)}{\delta_{t,h}}}. \tag{15}$$

The unweighted regression for fitting overly optimistic and overly pessimistic value functions is as follows:

$$\begin{aligned} &\forall\, t \in [T], h \in [H], \quad \hat{f}_{t,\pm2}^h = \arg\min_{f^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \left(f^h\left(x_s^h, a_s^h\right) - r_s^h - f_{t,\pm2}^{h+1}\left(x_s^{h+1}\right)\right)^2, \\ &\text{and we let } \mathcal{F}_{t,\pm2}^h := \left\{ f^h \in \mathcal{F}^h : \sum_{s\in[t-1]} \left(f^h(x_s^h, a_s^h) - \hat{f}_{t,\pm2}^h(x_s^h, a_s^h)\right)^2 \le \left(\beta_{t,2}^h\right)^2 \right\}. \end{aligned} \tag{16}$$

We choose the parameters as follows:

$$\text{Note} \quad \max_{h\in[H]} \left|\mathcal{F}^h + 2\mathcal{W} + \mathcal{W}\right| \leq \mathcal{N}\mathcal{N}_b^2,$$

$$\beta_{t,2}^h := \sqrt{2\left(24L+21\right) \imath^2(\delta_{t,h}) + 20tL\epsilon}, \tag{17}$$

$$\delta_{t,h} := \frac{\delta}{(T+1)(H+1)} \text{ and } \imath(\delta) = \sqrt{2\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(18LT)+2\right)\left(\log(18L)+2\right)}{\delta}}. \tag{18}$$

The unweighted regression for fitting second-moment function values is as follows:

$$\begin{aligned} \forall\, t\in[T], h\in[H], \quad \hat{g}_t^h &= \arg\min_{g^h\in\mathcal{F}^h} \sum_{s\in[t-1]} \left(g^h\left(x_s^h, a_s^h\right) - \left(r_s^h + f_{t,1}^{h+1}\left(x_s^{h+1}\right)\right)^2\right)^2, \\ \text{and similarly let } \mathcal{G}_t^h &:= \left\{g^h\in\mathcal{F}^h : \sum_{s\in[t-1]} \left(g^h(x_s^h,a_s^h) - \hat{g}_t^h(x_s^h,a_s^h)\right)^2 \leq \left(\bar{\beta}_t^h\right)^2\right\}. \end{aligned} \tag{19}$$

Here we choose parameters:

$$\bar{\beta}_t^h := \sqrt{8(11L+9)\left(\iota'(\delta_{t,h})\right)^2 + 32tL\epsilon}, \tag{20}$$

$$\delta_{t,h} := \frac{\delta}{(T+1)(H+1)} \quad \text{and} \quad \iota'(\delta) := \sqrt{2\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(32LT)+2\right)\left(\log(32L)+2\right)}{\delta}}. \tag{21}$$

In all the definitions above, we note that $\delta_{t,h}$ is independent of t and h. We also observe $\sum_{t\in[T],h\in[H]} \delta_{t,h} \leq \delta$. We will define the following good (probabilistic) event: $\mathcal{E}_{t,j}^h := \{\bar{f}_{t,j}^h \in \mathcal{F}_{t,j}^h\}$ with $j=1,\pm 2$. Further, we let $\bar{\mathcal{E}}_t := \{\psi_t^h \in \mathcal{G}_t^h\}$. Further, we let $\mathcal{E}_t^h = \mathcal{E}_{t,1}^h \cap \mathcal{E}_{t,2}^h \cap \mathcal{E}_{t,-2}^h \cap \bar{\mathcal{E}}_t^h$, and use $\mathcal{E}_t = \cap_{h\in[H]}\mathcal{E}_t^h$ and $\mathcal{E}_{\leq t} = \cap_{t'\leq t}\mathcal{E}_{t'}$ as shorthand for joint events.

Design of exploration policy. We restate the exploration policy for generating new data trajectory as stated in the main paper, i.e. Equation (9). At each iteration t, the algorithm collects data using both optimistic sequence $f_{t,1}^h$ and overly optimistic $f_{t,2}^h$. Given a sequence of pre-specified $\{u_t\}_{t\in[T]}$, at each iteration t, with function $f_{t,1}^h(\cdot)$ and $f_{t,2}^h(\cdot)$ at hand, we choose actions based on the following rule:

$$a_t^h = \begin{cases} \mathrm{argmax}_{a\in\mathcal{A}}\, f_{t,1}^h(x_t^h,a) & \text{if } f_{t,1}^h(x_t^{h'}) \geq f_{t,2}^h(x_t^{h'}) - u_t \text{ for all } h' \leq h, \\ \mathrm{argmax}_{a\in\mathcal{A}}\, f_{t,2}^h(x_t^h,a) & \text{otherwise,} \end{cases} \tag{22}$$

We also use $h_t \in [H+1]$ to denote the (random) threshold at which we first start taking greedy action based on overly optimistic sequence $f_{t,2}$. We divide the iteration set into the disjoint subsets $[T] = \mathcal{T}_{\mathrm{o}} \cup \mathcal{T}_{\mathrm{oo}}$ so that

$$\mathcal{T}_{\mathrm{o}} := \{t\in[T] : h_t = H+1\} \text{ and } \mathcal{T}_{\mathrm{oo}} := \{t\in[T] : h_t \in [H]\}. \tag{23}$$

Al step t, we note the policy induced by our defined exploration rule (22) at step t,h given state x_t^h depends on $\mathcal{D}_{[t-1]}^h$ and also the new generated trajectory $\{x_1^h, x_2^h, \cdots, x_{[t-1]}^h, x_t^h\}$. For each

$h \in [H]$, we use V_t to denote the expected V-value only at the trajectory under the exploration rule at step t, formally as follows:

$$V_t^h := \mathbb{E}\left[\sum_{h' \geq h} r^{h'} | x_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]. \tag{24}$$

The expectation is taken with respect to both model transition and exploration policy.

Other notations. We define the martingale difference sequence so that

$$\xi_{t,j}^h := r_t^h + f_{t,j}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h, x^{h+1}}\left[r^h + f_{t,j}^{h+1}(x^{h+1}) | z_t^h\right], \text{ for } j = 1, -2, 2,$$
$$\xi_t^h := r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1} | z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right].$$

We use the following shorthand of summation of bonus terms.

$$I := \sum_{t \in [T]} \sum_{h \in [H]} \min\left(1 + L, b_{t,1}^h(z_t^h)\right),$$
$$II := \sum_{t \in [T]} \sum_{h \in [H]} \min\left(1 + L, b_{t,2}^h(z_t^h)\right).$$

In the high-probability regret proof in Appendix G, we also define probabilistic event $\mathcal{E}_{\xi_{\text{dif}}}$, $\mathcal{E}_{\mathbb{V}}$, $\mathcal{E}_{\xi}$, $\mathcal{E}_{\xi_1}$, $\mathcal{E}_{\xi_2}$, $\mathcal{E}_{\xi_{-2}}$, we refer readers to Lemmas 46 and 47 and corollary 40 for their concrete meanings.

Helper lemmas. We now include a few helper lemmas that will be used in multiple parts of the analysis of Algorithm 1. The first few lemmas characterize the concentration behavior of martingale difference sequence, and are used heavily in Appendix F.2, and for proving Lemmas 46 and 47.

Lemma 16 (Freedman's inequality, cf. Theorem of Beygelzimer et al. (2011)) *Let $M, v > 0$ be constants, and $\{x_i\}_{i \in [t]}$ be stochastic process adapted to a filtration $\{\mathcal{H}_i\}_{i \in [t]}$. Suppose $\mathbb{E}[x_i | \mathcal{H}_{i-1}] = 0$, $|x_i| \leq M$ and $\sum_{i \in [t]} \mathbb{E}[x_i^2 | \mathcal{H}_{i-1}] \leq V^2$. Then for any $\delta > 0$, with probability at least $1 - \delta$ we have*

$$\sum_{i \in [t]} x_i \leq 2V\sqrt{\log(1/\delta)} + M \log(1/\delta).$$

Lemma 17 (Freedman's inequality variant, cf. Dzhaparidze and Van Zanten (2001); Fan et al. (2017)) *Let $\{x_i\}_{i \in [t]}$ be adapted to filtration $\{\mathcal{H}_i\}_{i \in [t]}$. Suppose $\mathbb{E}[x_i | \mathcal{H}_{i-1}] = 0$ and $\mathbb{E}[x_i^2 | \mathcal{H}_{i-1}] < \infty$ almosy surely. Then for any a, v, $y > 0$ we have*

$$\mathbb{P}\left(\sum_{i \in [t]} x_i > a, \sum_{i \in [t]} \left(\mathbb{E}[x_i^2 | \mathcal{H}_{i-1}] + x_i^2 \cdot \mathbf{1}_{\{|x_i| > y\}}\right) < v^2\right) \leq \exp\left(\frac{-a^2}{2(v^2 + ay/3)}\right).$$

Corollary 18 *Let $M > m > 0, V > v > 0$ be constants, and $\{x_i\}_{i \in [t]}$ be stochastic process adapted to a filtration $\{\mathcal{H}_i\}_{i \in [t]}$. Suppose $\mathbb{E}[x_i | \mathcal{H}_{i-1}] = 0$, $|x_i| \leq M$ and $\sum_{i \in [t]} \mathbb{E}[x_i^2 | \mathcal{H}_{i-1}] \leq V^2$ almost surely. Then for any $\delta, \epsilon > 0$, let $\iota = \sqrt{\log \frac{(2\log(V/v)+2)\cdot(\log(M/m)+2)}{\delta}}$ we have*

$$\mathbb{P}\left(\sum_{i \in [t]} x_i > \iota \sqrt{2\left(2\sum_{i \in [t]} \mathbb{E}[x_i^2 | \mathcal{H}_{i-1}] + v^2\right) + \frac{2}{3}\iota^2 \left(2\max_{i \in [t]} |x_i| + m\right)}\right) \leq \delta.$$

Next, we state some helper lemmas on law of total variance, which are used in proving Corollary 40.

Proposition 19 (Law of total variance, LTV) *Suppose at step t, we use policy π_t and have value function following such policy as $\{V_{\pi_t}^h\}_{h\in[H]}$, then by law of total variance we have*

$$\mathbb{V}\left[\sum_{h\in[H]} r_t^h\right] = \mathbb{E}\left[\sum_{h\in[H]} \mathbb{V}\left[r_t^h + V_{\pi_t}^{h+1}|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]\right] \leq 1.$$

Proof We use $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$ for simplicity. By conditional expectation and law of total variance we can show that

$$\begin{aligned}
\mathbb{V}\left[\sum_{h\in[H]} r_t^h\right] &= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{V}[r^h + V_{\pi_t}^{h+1}(x^{h+1})|z_t^h] \mid f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \\
&= \mathbb{E}\left[\left(\sum_{h=1}^{H}\left(r_t^h + V_{\pi_t}^{h+1}(x_t^{h+1}) - \mathbb{E}[r^h + V_{\pi_t}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \\
&\overset{(\star)}{=} \mathbb{E}\left[\left(\sum_{h=1}^{H} r_t^h - V_{\pi_1}^1(x_t^1)\right)^2 \mid f_{t,1}^{[H]}, f_{t,2}^{[H]}\right],
\end{aligned}$$

here we use $(\star)$ the definition of π_t and V_{π_t}. ■

We now provide an adapted version of LTV applying to $f_{t,1}$, which generates our greedy policy only when $t \in \mathcal{T}_\text{o}$ based on the exploration rule as in Equation (22).

Proposition 20 (Adapted version using LTV) *Suppose at step t, the agent explores based on rule (22). Then conditioning on the past*

$$\mathcal{H}_{t-1}^H = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots; \cdots; x_{t-1}^1, r_{t-1}^1, \cdots, r_{t-1}^H, x_{t-1}^{H+1}),$$

we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=1}^{H} \mathbb{V}_{r^h,x^{h+1}}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h] \mid \mathcal{H}_{t-1}^H\right] &\leq 2\,\mathbb{E}\left[\left(\sum_{h=1}^{H} r^h - f_{t,1}^1(x_t^1)\right)^2 \mid \mathcal{H}_{t-1}^H\right] \\
&+ 2\,\mathbb{E}\left[\left(\mathbf{1}(t\in\mathcal{T}_\text{o})\sum_{h=1}^{H}\left(f_{t,1}^h(x_t^h, a_t^h) - \mathop{\mathbb{E}}_{r^h,x^{h+1}}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid \mathcal{H}_{t-1}^H\right] \\
&+ 2\,\mathbb{E}\left[\left(\mathbf{1}(t\in\mathcal{T}_\text{oo})\sum_{h=1}^{H}\left(f_{t,1}^h(x_t^h) - \mathop{\mathbb{E}}_{r^h,x^{h+1}}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid \mathcal{H}_{t-1}^H\right].
\end{aligned}$$

Proof We use $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$ for simplicity. By conditional expectation and law of total variance we can show that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{V}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h] \mid \mathcal{H}_{t-1}^H\right] \\
&\quad = \mathbb{E}\left[\left(\sum_{h=1}^{H}\left(r_t^h + f_{t,1}^{h+1}(x_t^{h+1}) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid \mathcal{H}_{t-1}^H\right] \\
&\quad \overset{(i)}{=} \mathbb{E}\Bigg[\Bigg(\sum_{h=1}^{H} r_t^h - f_{t,1}^1(x_t^1) + \sum_{h=1}^{H}\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{o}}\}}\left(f_{t,1}^h(x_t^h,a_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right) \\
&\qquad\qquad + \sum_{h=1}^{H}\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{oo}}\}}\left(f_{t,1}^h(x_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}]\right)\Bigg)^2 \mid \mathcal{H}_{t-1}^H\Bigg] \\
&\quad \overset{(ii)}{\leq} 2\,\mathbb{E}\left[\left(\sum_{h=1}^{H} r_t^h - f_{t,1}^1(x_t^1)\right)^2 \mid \mathcal{H}_{t-1}^H\right] \\
&\qquad\qquad + 2\,\mathbb{E}\left[\left(1_{\{t\in\mathcal{T}_{\mathrm{o}}\}}\sum_{h=1}^{H}\left(f_{t,1}^h(x_t^h,a_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid \mathcal{H}_{t-1}^H\right] \\
&\qquad\qquad + 2\,\mathbb{E}\left[\left(1_{\{t\in\mathcal{T}_{\mathrm{oo}}\}}\sum_{h=1}^{H}\left(f_{t,1}^h(x_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\right)\right)^2 \mid \mathcal{H}_{t-1}^H\right],
\end{aligned}$$

here we use (i) the fact that whenever $t \in \mathcal{T}_{\mathrm{o}}$, we have $f_{t,1}^h(x_t^h, a_t^h) = f_{t,1}^h(x_t^h)$, and (ii) the factor 2 comes from the AM-GM inequality and by noting that summation terms of $t \in \mathcal{T}_{\mathrm{o}}$ and $t \in \mathcal{T}_{\mathrm{oo}}$ are mutually exclusive. ■

F.2. Confidence Intervals' Properties

In this subsection, we justify the choices of parameters $\beta_{t,j}^h$ and $\bar{\beta}_t^h$ in constructing our confidence intervals $\mathcal{F}_{t,j}^h$ for $j = 1, \pm 2$ and $\mathcal{G}_t^h$. We show with high probability under moderate assumptions, it holds that $\bar{f}_{t,j} \in \mathcal{F}_{t,j}^h$ for $j = 1, \pm 2$ and $\psi_t^h \in \mathcal{G}_t^h$ using martingale concentration. We write $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot|z_t^h]$ where the randomness is taken with respect to only r^h and x^{h+1} due to model transition by abusing notation, when the meaning is clear from context.

Confidence interval of optimistic sequence. This paragraph proves the property of the optimistic confidence interval we construct for $Q_\star^h$.

Lemma 21 *At step $t \in [T]$ and horizon $h \in [H]$, suppose*

$$(\sigma_s^h)^2 \geq \mathbb{V}\left[r^h + f_\star^{h+1}(x^{h+1})|z_s^h\right], \qquad \forall\, s \in [t-1], \tag{25}$$

and $$\mathbb{E}\left[|f_{t,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})| \mid z_s^h\right] \leq f_{s,2}^h(z_s^h) - f_{s,-2}^h(z_s^h), \qquad \forall\, s \in [t-1], z_s^h \in \mathcal{X} \times \mathcal{A}. \tag{26}$$

Recalling that $\bar{f}_{t,1}^h(x^h, a^h) \in \mathcal{F}^h$ as some function such that $|\bar{f}_{t,1}^h(z^h) - \mathcal{T} f_{t,1}^{h+1}(z^h)| \leq \epsilon$ for all $z^h = (x^h, a^h)$, we have with probability $1 - 2\delta_{t,h}$, it holds that $\bar{f}_{t,1}^h \in \mathcal{F}_{t,1}^h$ for the constructed $\mathcal{F}_{t,1}^h$ based on the definition of confidence interval in (11) and $\beta_{t,1}^h$ in (13).

We will prove that the assumptions of Lemma 21 hold true in the next section following a recursive argument (see Proposition 33). Now we first work under these assumptions and provide a complete proof of Lemma 21. In order to prove this lemma, we first show the concentration properties of two martingale difference sequences (MDSs), building on Corollary 18.

Lemma 22 *Under the same setting and condition (25) as in Lemma 21, consider filtration defined as $\mathcal{H}_s^h = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots, r_2^H, x_2^{H+1}; \cdots, x_s^1, r_s^1, \cdots, r_s^h, x_s^{h+1})$, we consider for any fixed $f, \tilde{f} \in [0, L]$, define*

$$\eta_s^h := r_s^h + f_\star^{h+1}\left(x_s^{h+1}\right) - \mathbb{E}\left[r^h + f_\star^{h+1}\left(x^{h+1}\right)|z_s^h\right],$$

$$\textit{and MDS} \quad \mathsf{D}_s^h[f, \tilde{f}] := 2\frac{\eta_s^h}{(\bar{\sigma}_s^h)^2} \cdot \left(f\left(z_s^h\right) - \tilde{f}\left(z_s^h\right)\right),$$

then we have with probability $1 - \delta_{t,h}/\mathcal{N}^2$,

$$\sum_{s\in[t-1]} \mathsf{D}_s^h[f, \tilde{f}]$$

$$\leq \frac{4}{3}\upsilon(\delta_{t,h})\sqrt{\lambda} + \frac{2}{3}\upsilon^2(\delta_{t,h}) + \sqrt{2}\upsilon(\delta_{t,h}) + \frac{16^2\upsilon^2(\delta_{t,h})}{3^2} + \frac{\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h) - \tilde{f}(z_s^h)\right)^2}{4}, \tag{27}$$

where we recall $\delta_{t,h} = \frac{\delta}{(T+1)(H+1)}$ and $\upsilon(\delta_{t,h}) = \sqrt{\log \frac{\mathcal{N}^2(2\log(4LT/\alpha)+2)(\log(8L/\alpha^2)+2)}{\delta_{t,h}}}$ as in (14).

Proof By definition of $\mathsf{D}_s^h[f, \tilde{f}] = 2\frac{\eta_s^h}{(\bar{\sigma}_s^h)^2} \cdot \left(f\left(z_s^h\right) - \tilde{f}\left(z_s^h\right)\right)$, it is a well-defined martingale difference sequence adapted to $\mathcal{H}_s^h$. In order to apply Corollary 18, we first give the almost-surely bounds on its maximum scale M and sum of second moment V. It holds that

$$\left|\mathsf{D}_s^h\left[f, \tilde{f}\right]\right| \leq \frac{2|\eta_s^h| \cdot \max_{z_s^h} |f(z_s^h) - \tilde{f}(z_s^h)|}{\alpha^2} \leq \frac{8L}{\alpha^2} =: M \text{ w.p. } 1,, \tag{28}$$

$$\begin{aligned}\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}_s^h\left[f, \tilde{f}\right]\right)^2 |z_s^h\right] &= \sum_{s\in[t-1]} 4\frac{\mathbb{E}\left[\left(\eta_s^h\right)^2 |z_s^h\right]}{(\bar{\sigma}_s^h)^4}\left(f(z_s^h) - \tilde{f}(z_s^h)\right)^2 \\ &\leq \frac{16L^2}{\alpha^2}(t-1) \leq \left(\frac{4LT}{\alpha}\right)^2 =: V^2.\end{aligned} \tag{29}$$

Here we only use the size bound on $f, \tilde{f} \in [0, L]$ by Assumption 1, $\bar{\sigma}_s^h \geq \alpha$ by definition (12), and $|\eta_s^h| \leq 2$ since $r_s^h, f_\star^{h+1} \in [0,1]$ always hold true.

Additionally, we also have the following realization-dependent bounds on the maximum scale and sum of second moment for the MDS sequence.

$$\begin{aligned}\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}_s^h\left[f,\tilde{f}\right]\right)^2 | z_s^h\right] &= \sum_{s\in[t-1]} 4\frac{\mathbb{E}\left[\left(\eta_s^h\right)^2 | z_s^h\right]}{\left(\bar{\sigma}_s^h\right)^4}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2\\ &\overset{(i)}{\leq} 4\sum_{s\in[t-1]}\frac{1}{\left(\bar{\sigma}_s^h\right)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2 \end{aligned} \tag{30}$$

Here we use (i) the assumption in (25) such that $\mathbb{E}\left[\left(\eta_s^h\right)^2 | z_s^h\right] = \mathbb{V}\left[r^h + f_\star^{h+1}(x^{h+1}) | z_s^h\right] \leq \left(\sigma_s^h\right)^2 \leq \left(\bar{\sigma}_s^h\right)^2$.

$$\begin{aligned}\max_{s\in[t-1]}\left|D_s^h\left[f,\tilde{f}\right]\right| &= \max_{s\in[t-1]} 2\left|\frac{\eta_s^h}{\left(\bar{\sigma}_s^h\right)^2}\right|\cdot\left|f(z_s^h)-\tilde{f}(z_s^h)\right|\\ &\overset{(ii)}{\leq} \max_{s\in[t-1]}\frac{4}{\left(\bar{\sigma}_s^h\right)^2}\sqrt{D^2_{\mathcal{F}^h}(z_s^h; z^h_{[s-1]}, \bar{\sigma}^h_{[s-1]})\left(\sum_{i\in[s-1]}\frac{1}{\left(\bar{\sigma}_i^h\right)^2}\left(f(z_i^h)-\tilde{f}(z_i^h)\right)^2+\lambda\right)}\\ &\overset{(iii)}{\leq} \frac{1}{\upsilon(\delta_{t,h})}\sqrt{\sum_{s\in[t-2]}\frac{1}{\left(\bar{\sigma}_s^h\right)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2+\lambda}, \end{aligned} \tag{31}$$

Here we use (ii) the size bound that $|\eta_s^h| \leq 2$ since $r^h, f_\star^{h+1} \in [0,1]$ together with the definition of $D_{\mathcal{F}^h}$, and (iii) the choice of $\bar{\sigma}_s^h \geq 2\sqrt{D_{\mathcal{F}^h}(z_s^h; z^h_{[s-1]}, \bar{\sigma}^h_{[s-1]})\cdot\upsilon(\delta_{t,h})}$ for all $s \in [t-1]$ and then taking the $\max_{s\in[t-1]}$ inside the summation of $i \in [s-1]$.

Applying Corollary 18 with $M = 8L/\alpha^2$ (28), $V = 4LT/\alpha$ (29), $v = m = 1$, using (30) and (31) we conclude that with probability at least $1 - \delta_{t,h}/\mathcal{N}^2$, for

$$\upsilon(\delta_{t,h}) = \sqrt{\log\frac{\mathcal{N}^2\left(2\log(4LT/\alpha)+2\right)\left(\log(8L/\alpha^2)+2\right)}{\delta_{t,h}}},$$

we have

$$\sum_{s\in[t-1]} 2\frac{\eta_s^h}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right) \le \upsilon(\delta_{t,h})\sqrt{16\left(\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2\right)+2}$$

$$+\frac{2}{3}\upsilon^2(\delta_{t,h})\cdot\left(\frac{2}{\upsilon(\delta_{t,h})}\sqrt{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2+\lambda}+1\right)$$

$$\le \frac{4}{3}\upsilon(\delta_{t,h})\sqrt{\lambda}+\frac{2}{3}\upsilon^2(\delta_{t,h})+\sqrt{2}\upsilon(\delta_{t,h})+\frac{16}{3}\upsilon(\delta_{t,h})\sqrt{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}$$

$$\le \frac{4}{3}\upsilon(\delta_{t,h})\sqrt{\lambda}+\frac{2}{3}\upsilon^2(\delta_{t,h})+\sqrt{2}\upsilon(\delta_{t,h})+\frac{2\cdot 16^2\upsilon^2(\delta_{t,h})}{2\cdot 3^2}+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}{2\cdot 2}.$$

The last inequality above uses AM-GM inequality and this concludes the proof. ■

Lemma 23 *Under the same setting and condition (26) as in Lemma 21, consider filtration defined as* $\mathcal{H}_s^h = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots, r_2^H, x_2^{H+1}; \cdots, x_s^1, r_s^1, \cdots, r_s^h, x_s^{h+1})$, *we consider for any fixed* $f, \tilde{f} \in [0, L]$, *and* $f' = f_{t,1}^{h+1}$, *define*

$$\xi_s^h[f'] = f'\left(x_s^{h+1}\right) - f_\star^{h+1}\left(x_s^{h+1}\right) - \mathbb{E}\left[f'\left(x^{h+1}\right) - f_\star^{h+1}\left(x^{h+1}\right)|z_s^h\right],$$

and MDS $\Delta_s^h\left[f,\tilde{f},f'\right] = 2\frac{\xi_s^h[f']}{(\bar{\sigma}_s^h)^2}\cdot\left(f\left(z_s^h\right)-\tilde{f}\left(z_s^h\right)\right)$,

then we have with probability $1-\delta_{t,h}/\mathcal{N}^3\mathcal{N}_b$,

$$\sum_{s\in[t-1]}\Delta_s^h\left[f,\tilde{f},f'\right]$$

$$\le \frac{4}{3}\sqrt{\lambda}+\frac{2}{3}\cdot\frac{\iota^2(\delta_{t,h})}{\log\mathcal{N}_b}+\sqrt{2}\cdot\frac{\iota(\delta_{t,h})}{\sqrt{\log\mathcal{N}_b}}+\frac{16^2}{3^2}+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}{4}, \quad (32)$$

where we recall $\iota(\delta_{t,h}) = 3\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(4LT/\alpha)+2\right)\left(\log(8L/\alpha^2)+2\right)}{\delta_{t,h}}}$ *as in (15).*

Proof Note that the martingale difference $\Delta_s^h\left[f,\tilde{f},f'\right]$ is adapted to $\mathcal{H}_s^h$. In order to apply Corollary 18, we again first give the almost-surely bounds on its maximum scale M and sum of second moment V. It holds that

$$\left|\Delta_s^h\left[f,\tilde{f},f'\right]\right| \le \frac{2|\xi_s^h[f']|\cdot\max_{z_s^h}|f(z_s^h)-\tilde{f}(z_s^h)|}{\alpha^2} \le \frac{8L}{\alpha^2} =: M, \quad (33)$$

$$\sum_{s\in[t-1]}\mathbb{E}\left[\left(\Delta_s^h\left[f,\tilde{f},f'\right]\right)^2|z_s^h\right] = \sum_{s\in[t-1]}4\frac{\mathbb{E}\left[\left(\xi_s^h[f']\right)^2|z_s^h\right]}{(\bar{\sigma}_s^h)^4}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2 \le \left(\frac{4L}{\alpha}T\right)^2 =: V^2. \quad (34)$$

Here we again use $f, \tilde{f} \in [0, L]$ and $\bar{\sigma}_s^h \geq \alpha$ and $|\xi_s^h| \leq 2$ since $f' = f_{t,1}^{h+1} \in [0,1]$.

Additionally, we also have the following realization-dependent bound on the maximum magnitude.

$$\sum_{s\in[t-1]} \mathbb{E}\left[\left(\Delta_s^h\left[f,\tilde{f},f'\right]\right)^2 |z_s^h\right] = \sum_{s\in[t-1]} 4\frac{\mathbb{E}\left[\left(\xi_s^h[f']\right)^2|z_s^h\right]}{\left(\bar{\sigma}_s^h\right)^4}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2$$
$$\overset{(i)}{\leq} \frac{4}{\iota^2(\delta_{t,h})}\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2, \tag{35}$$

Here we use (i) given assumption (26) and that $\delta_{t,h}$ doesn't depend on t by definition so that $\mathbb{E}\left[\left(\xi_s^h[f']\right)^2|z_s^h\right] \leq \mathbb{E}\left[\left(f_{t,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})\right)^2|z_s^h\right] \leq 2\mathbb{E}\left[|f_{t,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})||z_s^h\right]$ $\leq 2(f_{s,2}^h(z_s^h) - f_{s,-2}^h(z_s^h)) \leq \iota^{-2}(\delta_{t,h})\left(\bar{\sigma}_s^h\right)^2$.

We also have the following bound on the sum of second moment.

$$\max_{s\in[t-1]}\left|\Delta_s^h\left[f,\tilde{f},f'\right]\right| = \max_{s\in[t-1]} 2\left|\frac{\xi_s^h[f']}{(\bar{\sigma}_s^h)^2}\right|\cdot\left|f(z_s^h)-\tilde{f}(z_s^h)\right|$$
$$\overset{(ii)}{\leq} \max_{s\in[t-1]}\frac{4}{(\bar{\sigma}_s^h)^2}\sqrt{D^2_{\mathcal{F}^h}(z_s^h; z_{[s-1]}^h, \bar{\sigma}_{[s-1]}^h)\left(\sum_{i\in[s-1]}\frac{1}{(\bar{\sigma}_i^h)^2}\left(f(z_i^h)-\tilde{f}(z_i^h)\right)^2+\lambda\right)}$$
$$\overset{(iii)}{\leq} \frac{1}{\iota^2(\delta_{t,h})}\sqrt{\sum_{s\in[t-2]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2+\lambda}. \tag{36}$$

For (ii) we use the size bound that $|\xi_s^h[f']| \leq 2$ since $r^h \in [0,1]$, and $f' \in [0,1]$ together with the definition of $D_{\mathcal{F}^h}$ and (iii) the choice of $\bar{\sigma}_s^h \geq 2\sqrt{D_{\mathcal{F}^h}(z_s^h; z_{[s-1]}^h, \bar{\sigma}_{[s-1]}^h)\cdot\iota^2(\delta_{t,h})}$ for all $s \in [t-1]$ and taking the $\max_{s\in[t-1]}$ inside the summation of $i \in [s-1]$.

Applying Corollary 18 with $M = 8L/\alpha^2$ (33), $V = 4LT/\alpha$ (34), $v = 1/\sqrt{\log \mathcal{N}_b}$, $m = 1/\log \mathcal{N}_b$, using (35) and (36) we conclude that with probability at least $1 - \delta_{t,h}/(\mathcal{N}^3\mathcal{N}_b)$, it holds that

$$\sum_{s\in[t-1]} 2\frac{\xi_s^h[f']}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right) \leq \iota(\delta_{t,h})\sqrt{\frac{16}{\iota^2(\delta_{t,h})}\left(\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2\right)+2\cdot\frac{1}{\log\mathcal{N}_b}}$$
$$+\frac{2}{3}\iota^2(\delta_{t,h})\cdot\left(\frac{2}{\iota^2(\delta_{t,h})}\sqrt{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2+\lambda}+\frac{1}{\log\mathcal{N}_b}\right)$$
$$\leq \frac{4}{3}\sqrt{\lambda}+\frac{2}{3}\cdot\frac{\iota^2(\delta_{t,h})}{\log\mathcal{N}_b}+\sqrt{2}\cdot\frac{\iota(\delta_{t,h})}{\sqrt{\log\mathcal{N}_b}}+\frac{16}{3}\sqrt{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}$$
$$\leq \frac{4}{3}\sqrt{\lambda}+\frac{2}{3}\cdot\frac{\iota^2(\delta_{t,h})}{\log\mathcal{N}_b}+\sqrt{2}\cdot\frac{\iota(\delta_{t,h})}{\sqrt{\log\mathcal{N}_b}}+\frac{2\cdot 16^2}{2\cdot 3^2}+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}{2\cdot 2},$$

for the choice of

$$\iota(\delta_{t,h}) = 3\sqrt{\log \frac{\mathcal{N}\mathcal{N}_b \left(2\log(4LT/\alpha)+2\right)\left(\log(8L/\alpha^2)+2\right)}{\delta_{t,h}}}$$
$$\geq \sqrt{\log \frac{\mathcal{N}^3\mathcal{N}_b \left(2\log\left(\frac{4LT\sqrt{\log \mathcal{N}_b}}{\alpha}\right)+2\right)\left(\log \frac{8L\log\mathcal{N}_b}{\alpha^2}+2\right)}{\delta_{t,h}}}.$$

Here for the last inequality we use $\log\log\mathcal{N}_b \leq \mathcal{N}_b$. ■

Making use of these two helper lemmas, we provide the complete proof for Lemma 21.

Proof [Proof of Lemma 21] At step $t \in [T], h \in [H]$, we locally denote the probability event $\mathcal{E}_{t,h}$ as follows so that $\{\bar{f}_{t,1}^h \notin \mathcal{F}_{t,1}^h\} \subseteq \mathcal{E}_{t,h}$:

$$\mathcal{E}_{t,h} := \left\{ \sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}_s^h)^2} \left(\bar{f}_{t,1}^h\left(x_s^h, a_s^h\right) - \hat{f}_{t,1}^h\left(x_s^h, a_s^h\right)\right)^2 > \left(\beta_{t,1}^h\right)^2 \right\}.$$

Note by definition of $\hat{f}_{t,1}^h$, we know that with probability 1 it holds that

$$\sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right) - \bar{f}_{t,1}^h\left(z_s^h\right)\right)^2$$
$$\leq 2\sum_{s\in[t-1]} \frac{\left(r_s^h + f_{t,1}^{h+1}\left(x_s^{h+1}\right) - \bar{f}_{t,1}^h(z_s^h)\right)}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right) - \bar{f}_{t,1}^h\left(z_s^h\right)\right).$$

Thus this event can be equivalently expressed as

$$\left\{ \begin{array}{l} \sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right) - \bar{f}_{t,1}^h\left(z_s^h\right)\right)^2 \leq 2\sum_{s\in[t-1]} \frac{\left(r_s^h + f_{t,1}^{h+1}\left(x_s^{h+1}\right) - \bar{f}_{t,1}^h(z_s^h)\right)}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right) - \bar{f}_{t,1}^h\left(z_s^h\right)\right) \\ \sum_{s\in[t-1]} \frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right) - \bar{f}_{t,1}^h\left(z_s^h\right)\right)^2 > \left(\beta_{t,1}^h\right)^2 \end{array} \right\}.$$

Now let $\hat{f}_{t,1}^h = f$, $\bar{f}_{t,1}^h = \tilde{f}$ and $f_{t,1}^{h+1} = f'$ so that $f_{t,1}^{h+1} = \min\left(f'' + \epsilon, 1\right)$ for some $f'' \in \mathcal{F}^{h+1} + \mathcal{W}$. Now we apply Lemmas 22 and 23 with these choices, along with union bounds. Then, it holds that with probability at least $1 - \delta_{t,h}$ that

$$2\sum_{s\in[t-1]} \frac{\eta_s^h}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h) - \bar{f}_{t,1}^h(z_s^h)\right)$$
$$\leq \frac{4}{3}\upsilon(\delta_{t,h})\sqrt{\lambda} + \frac{2}{3}\upsilon^2(\delta_{t,h}) + \sqrt{2}\upsilon(\delta_{t,h}) + \frac{16^2\upsilon^2(\delta_{t,h})}{3^2} + \frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h) - \bar{f}_{t,1}^h(z_s^h)\right)^2}{4}, \tag{37}$$

and also

$$
\begin{aligned}
&2\sum_{s\in[t-1]}\frac{\xi_s^h[f_{t,1}^{h+1}]}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)\\
&\quad\le\frac{4}{3}\sqrt{\lambda}+\frac{2}{3}\cdot\frac{\iota^2(\delta_{t,h})}{\log\mathcal{N}_b}+\sqrt{2}\cdot\frac{\iota(\delta_{t,h})}{\sqrt{\log\mathcal{N}_b}}+\frac{16^2}{3^2}+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(f(z_s^h)-\tilde{f}(z_s^h)\right)^2}{4}\\
&\quad\le\frac{4}{3}\sqrt{\lambda}+6\left(1+\upsilon^2(\delta_{t,h})\right)+3\sqrt{2}\sqrt{1+\upsilon^2(\delta_{t,h})}+\frac{16^2}{3^2}+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)^2}{4}.
\end{aligned}
\tag{38}
$$

Above for the last inequality we also use by definition of $\upsilon(\delta_{t,h})$ and $\iota(\delta_{t,h})$ that $\iota^2(\delta_{t,h})/\log\mathcal{N}_b\le 9(1+\upsilon^2(\delta_{t,h}))$.

Combining Equations (37) and (38) and using $\upsilon^2(\delta_{t,h})\ge 1$ for upper bounding the coefficients, we have with probability $1-2\delta_{t,h}$,

$$
\begin{aligned}
&2\sum_{s\in[t-1]}\frac{\left(r_s^h+f_{t,1}^{h+1}\left(x_s^{h+1}\right)-\bar{f}_{t,1}^h(z_s^h)\right)}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h\left(z_s^h\right)-\bar{f}_{t,1}^h\left(z_s^h\right)\right)\\
&\quad\le 4t\frac{\epsilon}{\alpha^2}L+2\sum_{s\in[t-1]}\frac{\eta_s^h+\xi_s^h[f_{t,1}^{h+1}]}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right) && \text{(Assumption 1)}\\
&\quad\le\frac{1}{2}\cdot\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)^2+(3\sqrt{\lambda}+78)\upsilon^2(\delta_{t,h})+\frac{4tL}{\alpha^2}\epsilon.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_{t,h})&\overset{(i)}{\le}\mathbb{P}\left(\sum_{s\in[t-1]}2\frac{\left(r_s^h+f_{t,1}^{h+1}\left(x_s^{h+1}\right)-\bar{f}_{t,1}^h(z_s^h)\right)}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)>\frac{1}{2}\left(\beta_{t,1}^h\right)^2\right.\\
&\qquad\left.+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_t^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)^2}{2}\right)\\
&\overset{(ii)}{\le}\mathbb{P}\left(\sum_{s\in[t-1]}2\frac{\left(r_s^h+f_{t,1}^{h+1}\left(x_s^{h+1}\right)-\bar{f}_{t,1}^h(z_s^h)\right)}{(\bar{\sigma}_s^h)^2}\left(\hat{f}_{t,1}^h(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)>\left(3\sqrt{\lambda}+8\right)\upsilon(\delta_{t,h})\right.\\
&\qquad\left.+70\upsilon^2(\delta_{t,h})+\frac{4tL}{\alpha^2}\epsilon+\frac{\sum_{s\in[t-1]}\frac{1}{(\bar{\sigma}_t^h)^2}\left(f_n(z_s^h)-\bar{f}_{t,1}^h(z_s^h)\right)^2}{2}\right)\\
&\le 2\delta_{t,h},
\end{aligned}
$$

where we use (i) the definition of $\mathcal{E}_{t,h}$ and (ii) the choice of $\beta_{t,1}^h$ as in Equation (13). Consequently,

$$
\mathbb{P}\left(\bar{f}_{t,1}^h\notin\mathcal{F}_{t,1}^h\right)\le\mathbb{P}\left(\mathcal{E}_{t,h}\right)\le 2\delta_{t,h},
$$

which implies with probability $1-2\delta_{t,h}$, $\bar{f}^h_{t,1} \in \mathcal{F}^h_{t,1}$ for any fixed given $t \in [T-1]$, $h \in [H]$ (the case $t=0$ holds with probability 1 by definition). ■

Confidence interval of overly optimistic and overly pessimistic sequence. Here we prove properties of the overly optimistic and overly pessimistic confidence interval we construct for $Q^h_\star$.

Lemma 24 *At step $t \in [T]$ and horizon $h \in [H]$, recall $\bar{f}^h_{t,2}(x^h, a^h) \in \mathcal{F}^h$ is some function such that $|\bar{f}^h_{t,2}(z^h) - \mathcal{T} f^{h+1}_{t,2}(z^h)| \le \epsilon$ for all $z^h = (x^h, a^h)$, then we have with probability $1-\delta_{t,h}$, it holds that $\bar{f}^h_{t,2} \in \mathcal{F}^h_{t,2}$ for the constructed $\mathcal{F}^h_{t,2}$ based on the definition of confidence interval and $\beta_{t,2}$ in (13).*

Similar to proving the confidence interval of optimistic sequence, we first provide the following lemma.

Lemma 25 *Under the same setting as in Lemma 24, consider filtration $\mathcal{H}^h_s$ and any fixed pair functions $f \in [0, L]$ and $f' \in [0, 2]$ we define random variables*

$$\eta^h_s[f'] := r^h_s + f'\left(x^{h+1}_s\right) - \mathbb{E}\left[r^h + f'\left(x^{h+1}\right)|z^h_s\right],$$

$$\textit{and MDS}\ \ \mathsf{D}^h_s[f, f'] := 2\eta^h_s[f'] \cdot \left(f\left(z^h_s\right) - \mathcal{T} f'\left(z^h_s\right)\right),$$

then we have with probability $1-\delta_{t,h}/(\mathcal{N}^2\mathcal{N}_b^2)$,

$$\sum_{s\in[t-1]} \mathsf{D}^h_s[f, f'] \le (24L+21)\iota^2(\delta_{t,h}) + \frac{\sum_{s\in[t-1]}\left(f(z^h_s) - \mathcal{T} f'(z^h_s)\right)^2}{2}, \tag{39}$$

$$\textit{where we recall}\ \ \iota(\delta_{t,h}) = \sqrt{2\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(18LT)+2\right)\left(\log(18L)+2\right)}{\delta_{t,h}}}\ \textit{as in (18)}.$$

Proof Similar to the proof of Lemma 22, we apply Corollary 18 on the defined MDS sequence D^h_s. We first bound the quantities of interest:

$$|\mathsf{D}^h_s[f, f']| \le 2|\eta^h_s[f']| \max_{z^h_s} |f(z^h_s) - \mathcal{T} f'(z^h_s)| \overset{(i)}{\le} 18L =: M,$$

$$\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}^h_s[f, f']\right)^2 |z^h_s\right] = \sum_{s\in[t-1]} 4\mathbb{E}\left(\eta^h_s[f']\right)^2 \left(f(z^h_s) - \mathcal{T} f'(z^h_s)\right)^2 \overset{(i)}{\le} (18LT)^2 =: V^2;$$

$$\begin{aligned}\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}^h_s[f, f']\right)^2 |z^h_s\right] &= \sum_{s\in[t-1]} 4\mathbb{E}\left[\left(\eta^h_s[f']\right)^2 |z^h_s\right] \left(f(z^h_s) - \mathcal{T} f'(z^h_s)\right)^2 \\ &\overset{(i)}{\le} 36 \sum_{s\in[t-1]} \left(f(z^h_s) - \mathcal{T} f'(z^h_s)\right)^2,\end{aligned}$$

where we use (i) the size bound that $|\eta^h_s| \le 3$ and $\max_{z^h_s} |f(z^h_s) - \mathcal{T} f'(z^h_s)| \le 3L$ (using $L \ge 1$).

Thus, applying Corollary 18 with $M = 18L$, $V = 18LT$, $v = m = 1$ to bound its summation we can conclude that with probability at least $1 - \delta_{t,h}/(\mathcal{N}^2\mathcal{N}_b^2)$,

$$\begin{aligned}
\sum_{s\in[t-1]} 2\eta_s^h[f']\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right)^2 &\le \iota(\delta_{t,h})\sqrt{36\left(\sum_{s\in[t-1]}(f(z_s^h) - \mathcal{T}f'(z_s^h))^2\right) + 2v^2} \\
&\quad + \frac{4}{3}\iota^2(\delta_{t,h})\cdot 18L + \frac{2}{3}\iota^2(\delta_{t,h}) \\
&\le \sqrt{2}\iota(\delta_{t,h}) + \left(24L + \frac{2}{3}\right)\iota^2(\delta_{t,h}) + 6\iota(\delta_{t,h})\sqrt{\sum_{s\in[t-1]}(f(z_s^h) - \mathcal{T}f'(z_s^h))^2} \\
&\le (24L+21)\iota^2(\delta_{t,h}) + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right)^2}{2}.
\end{aligned}$$

The last inequality again uses AM-GM inequality and the fact that $\iota(\delta_{t,h}) \ge 1$ by definition. ■

This lemma helps us prove Lemma 24 as follows.

Proof [Proof of Lemma 24] At step $t \in [T], h \in [H]$, we locally define the probability event

$$\mathcal{E}_{t,h} := \left\{ \sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h\left(x_s^h, a_s^h\right) - \bar{f}_{t,2}^h\left(x_s^h, a_s^h\right)\right)^2 > \left(\beta_{t,2}^h\right)^2 \right\}$$

so that $\{\bar{f}_{t,2}^h \notin \mathcal{F}_{t,2}^h\} \subseteq \mathcal{E}_{t,h}$.

Now by definition of $\hat{f}_{t,2}^h$, we know that with probability 1 it holds that

$$\sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h\left(z_s^h\right) - \bar{f}_{t,2}^h\left(z_s^h\right)\right)^2 \le 2\sum_{s\in[t-1]}\left(r_s^h + f_{t,2}^{h+1}\left(x_s^{h+1}\right) - \bar{f}_{t,2}^h(z_s^h)\right)\left(\hat{f}_{t,2}^h\left(z_s^h\right) - \bar{f}_{t,2}^h\left(z_s^h\right)\right).$$

This event can be equivalently expressed as

$$\left\{\begin{array}{l}\sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h\left(z_s^h\right) - \bar{f}_{t,2}^h\left(z_s^h\right)\right)^2 \le 2\sum_{s\in[t-1]}\left(r_s^h + f_{t,2}^{h+1}\left(x_s^{h+1}\right) - \bar{f}_{t,2}^h(z_s^h)\right)\left(\hat{f}_{t,2}^h\left(z_s^h\right) - \bar{f}_{t,2}^h\left(z_s^h\right)\right) \\ \sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h\left(z_s^h\right) - \bar{f}_{t,2}^h\left(z_s^h\right)\right)^2 > \left(\beta_{t,2}^h\right)^2\end{array}\right\}.$$

Now for each particular pair of $f \in \mathcal{F}^h$ where $\hat{f}_{t,2}^h = f$ and $f' = \min(1, f'' + 3\epsilon) = f_{t,2}^{h+1}$ where $f'' \in \mathcal{F}^{h+1} + 2\mathcal{W} + \mathcal{W}$, we define the random variables

$$\begin{aligned}
\eta_s^h[f'] &:= r_s^h + f'\left(x_s^{h+1}\right) - \mathbb{E}\left[r^h + f'\left(x^{h+1}\right)|z_s^h\right], \\
\text{and MDS}\quad \mathsf{D}_s^h[f,f'] &= 2\eta_s^h[f']\cdot\left(f\left(z_s^h\right) - \mathcal{T}f'\left(z_s^h\right)\right).
\end{aligned}$$

Following (39) we have with probability at least $1 - \delta_{t,h}/\left(\mathcal{N}^2\mathcal{N}_b^2\right)$,

$$\sum_{s\in[t-1]} 2\eta_s^h[f']\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right) \le (16L+21)\iota^2(\delta_{t,h}) + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right)^2}{2}.$$

This implies that for any function $\bar{f}[f']$ satisfying $\|\bar{f}[f'] - \mathcal{T}f'\|_\infty \le \epsilon$, it holds that with probability at least $1 - \delta_{t,h}/\left(\mathcal{N}^2\mathcal{N}_b^2\right)$,

$$\begin{aligned}
2\sum_{s\in[t-1]} & \left(\eta_s^h[f'] + \mathcal{T}f'\left(z_s^h\right) - \bar{f}[f'](z_s^h)\right)\left(f(z_s^h) - \bar{f}[f'](z_s^h)\right) \\
& \le \sum_{s\in[t-1]} 2\eta_s^h[f']\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right) + 4tL\epsilon + 4t\epsilon \\
& \le (24L+21)i^2(\delta_{t,h}) + 8tL\epsilon + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}f'(z_s^h)\right)^2}{2} \\
& \le (24L+21)i^2(\delta_{t,h}) + 10tL\epsilon + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \bar{f}[f'](z_s^h)\right)^2}{2}.
\end{aligned}$$

Note the size of $\mathcal{F}^h$ and $\mathcal{F}^{h+1} + 2\mathcal{W} + \mathcal{W}$ are bounded by $\mathcal{N}$ and $\mathcal{N}\mathcal{N}_b^2$, we thus take a union bound over all choices of $f \in \mathcal{F}^h$ and $f'' \in \mathcal{F}^{h+1} + 2\mathcal{W} + \mathcal{W}$ so that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{t,h}) & \overset{(i)}{\le} \mathbb{P}\left(\sum_{s\in[t-1]} 2\left(\eta_s^h + \mathbb{E}[r^h + f_{t,2}^{h+1}(x^{h+1})|z_s^h] - \bar{f}_{t,2}^h(z_s^h)\right)\left(\hat{f}_{t,2}^h(z_s^h) - \bar{f}_{t,2}^h(z_s^h)\right) > \right. \\
& \qquad \left. \frac{1}{2}\left(\beta_{t,2}^h\right)^2 + \frac{\sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h(z_s^h) - \bar{f}_{t,2}^h(z_s^h)\right)^2}{2}\right) \\
& \overset{(ii)}{\le} \mathbb{P}\left(\sum_{s\in[t-1]} 2\left(\eta_s^h + \mathbb{E}[r^h + f_{t,2}^{h+1}(x^{h+1})|z_s^h] - \bar{f}_{t,2}^h(z_s^h)\right)\left(\hat{f}_{t,2}^h(z_s^h) - \bar{f}_{t,2}^h(z_s^h)\right) > \right. \\
& \qquad \left. (24L+21)i^2(\delta_{t,h}) + 10tL\epsilon + \frac{\sum_{s\in[t-1]}\left(\hat{f}_{t,2}^h(z_s^h) - \bar{f}_{t,2}^h(z_s^h)\right)^2}{2}\right) \\
& \le \delta_{t,h},
\end{aligned}$$

where we use (i) the definition of $\mathcal{E}_{t,h}$ and (ii) the choice of $\beta_{t,2}^h$. Consequently,

$$\mathbb{P}\left(\bar{f}_{t,2}^h \notin \mathcal{F}_{t,2}^h\right) \le \delta_{t,h},$$

which implies with probability $1 - \delta_{t,h}$, $\bar{f}_{t,2}^h \in \mathcal{F}_{t,2}^h$ for any fixed given $t \in [T-1]$, $h \in [H]$ (the case $t = 0$ holds with probability 1 by definition). ■

Similarly, we have for overly pessimistic values $f_{t,-2}^h$ and $\bar{f}_{t,-2}^h$, the following lemma:

Lemma 26 *At step $t \in [T]$ and horizon $h \in [H]$, recall $\bar{f}_{t,-2}^h(x^h, a^h) \in \mathcal{F}^h$ is some function such that $|\bar{f}_{t,-2}^h(z^h) - \mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1})|z^h\right] \le \epsilon$ for all $z^h = (x^h, a^h)$, then we have with probability $1 - \delta_{t,h}$, it holds that $\bar{f}_{t,-2}^h \in \mathcal{F}_{t,-2}^h$ for the constructed $\mathcal{F}_{t,-2}^h$ based on the definition of confidence interval in Algorithm 1 and $\beta_{t,2}$ in (17).*

Proof The proof is symmetric as that of Lemma 24. ■

We also give the following consequence of Lemma 26 together with the definition of generalized Eluder dimension, which will be useful to justify our definition of σ_t^h in Equation (4).

Lemma 27 *Conditioning on the good event $\mathcal{E}_{t,-2}^h$, we have*

$$\left|\left[\bar{f}_{t,-2}^h(z_t^h)\right]^2 - \left[\hat{f}_{t,-2}^h(z_t^h)\right]^2\right| \le 2L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda} \cdot D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h).$$

Proof To see this, we note that conditioning on the good event $\bar{f}_{t,-2}^h(\cdot) \in \mathcal{F}_{t,-2}^h$, we have for any z,

$$\begin{aligned}\left[\bar{f}_{t,-2}^h(z)\right]^2 - \left[\hat{f}_{t,-2}^h(z)\right]^2 &\le 2L\left|\bar{f}_{t,-2}^h(z) - \hat{f}_{t,-2}^h(z)\right| \\ &\le 2L \cdot D_{\mathcal{F}^h}(z; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \sqrt{\sum_{s\in[t-1]}\left(\bar{f}_{t,-2}^h(z_s^h) - \hat{f}_{t,-2}^h(z_s^h)\right)^2 + \lambda} \\ &\le 2L \cdot D_{\mathcal{F}^h}(z; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h)\sqrt{(\beta_{t,2}^h)^2 + \lambda}.\end{aligned}$$

Plugging the particular choice of $z = z_t^h$ concludes the proof. ■

Confidence interval of second-moment sequence. Here we prove the property of the optimistic confidence interval we construct for the second-moment sequence.

Lemma 28 *At step $t \in [T]$ and horizon $h \in [H]$, recall $\psi_t^h(x^h, a^h) \in \mathcal{F}^h$ satisfies $|\psi_t^h(z^h) - \mathcal{T}_2 f_{t,1}^{h+1}(z^h)| \le \epsilon$ for any $z^h = (x^h, a^h)$, then we have with probability $1 - \delta_{t,h}$, it holds that $\psi_t^h \in \mathcal{G}_t^h$ for the constructed $\mathcal{G}_t^h$ based on the definition of confidence interval in (19) and $\bar{\beta}$ in (20).*

Similar to proving the confidence intervals above we first provide the following lemma.

Lemma 29 *Under the same setting as in Lemma 28, consider filtration $\mathcal{H}_s^h$ and any fixed pair functions $f \in [0, L]$, $f' \in [0, 1]$ we define random variables*

$$\eta_s^h[f'] := \left(r_s^h + f'\left(x_s^{h+1}\right)\right)^2 - \mathbb{E}\left[\left(r^h + f'\left(x^{h+1}\right)\right)^2 | z_s^h\right],$$

$$\textit{and MDS}\quad \mathsf{D}_s^h[f, f'] := 2\eta_s^h[f'] \cdot \left(f\left(z_s^h\right) - \mathcal{T}_2 f'\left(z_s^h\right)\right),$$

then we have with probability $1 - \delta_{t,h}/(\mathcal{N}^2\mathcal{N}_b)$,

$$\sum_{s\in[t-1]} \mathsf{D}_s^h[f, f'] \le 4(11L + 9)\left(\iota'(\delta_{t,h})\right)^2 + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2}{2}, \tag{40}$$

$$\textit{where we recall}\quad \iota'(\delta_{t,h}) = \sqrt{2\log\frac{\mathcal{N}\mathcal{N}_b\left(2\log(32LT) + 2\right)\left(\log(32L) + 2\right)}{\delta_{t,h}}}\quad \textit{as in (21).}$$

Proof Recall the definition of $\mathcal{T}_2 f(z^h) = \mathbb{E}\left[\left(r^h + f(z^{h+1})\right)^2 | z^h\right]$. We note the difference sequence D_s^h as defined is adapted to $\mathcal{H}_s^h$ and satisfies

$$|\mathsf{D}_s^h[f,f']| \le 2|\eta_s^h| \max_{z_s^h} |f(z_s^h) - \mathcal{T}_2 f'(z_s^h)| \overset{(i)}{\le} 32L =: M,$$

$$\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}_s^h[f,f']\right)^2 | z_s^h\right] = \sum_{s\in[t-1]} 4\mathbb{E}\left(\eta_s^h[f']\right)^2 \left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2 \overset{(i)}{\le} (32LT)^2 =: V^2$$

$$\sum_{s\in[t-1]} \mathbb{E}\left[\left(\mathsf{D}_s^h[f,f']\right)^2 | z_s^h\right] = \sum_{s\in[t-1]} 4\mathbb{E}\left[\left(\eta_s^h[f']\right)^2 | z_s^h\right] \left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2$$

$$\overset{(i)}{\le} 64 \sum_{s\in[t-1]} \left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2,$$

where we use (i) the size bound that $|\eta_s^h[f']| \le 4$ and $\max_{z_s^h} |f(z_s^h) - \mathcal{T}_2 f'(z_s^h)| \le 4L$.

Applying Corollary 18 with $M = 32L$, $V = 32LT$, $v = m = 1$ to bound its summation we can conclude that with probability at least $1 - \delta_{t,h}/\mathcal{N}^2\mathcal{N}_b^2$,

$$\begin{aligned}
\sum_{s\in[t-1]} 2\eta_s^h[f']\left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right) &\le \iota'(\delta_{t,h})\sqrt{64\left(\sum_{s\in[t-1]} (f(z_s^h) - \mathcal{T}_2 f'(z_s^h))^2\right) + 2v^2} \\
&\quad + \frac{4}{3}\left(\iota'(\delta_{t,h})\right)^2 \cdot 32L + \frac{2}{3}\left(\iota'(\delta_{t,h})\right)^2 \\
&\le \left(\frac{2+128L}{3}\right)\left(\iota'(\delta_{t,h})\right)^2 + \sqrt{2}\iota'(\delta_{t,h}) + 8\iota'(\delta_{t,h})\sqrt{\sum_{s\in[t-1]} (f(z_s^h) - \mathcal{T}_2 f'(z_s^h))^2} \\
&\le 4(11L+9)\left(\iota'(\delta_{t,h})\right)^2 + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2}{2}.
\end{aligned}$$

For the last inequality we use AM-GM inequality and by definition the fact that $\iota'(\delta_{t,h}) \ge 1$. ■

This lemma again helps us prove Lemma 28 as follows.

Proof [Proof of Lemma 28] At step $t \in [T], h \in [H]$, we locally define the probability event

$$\mathcal{E}_{t,h} := \left\{ \sum_{s\in[t-1]} \left(\hat{g}_t^h\left(x_s^h, a_s^h\right) - \psi_t^h\left(x_s^h, a_s^h\right)\right)^2 > \left(\bar{\beta}_t^h\right)^2 \right\}.$$

so that $\{\psi_t^h \notin \mathcal{G}_t^h\} \subseteq \mathcal{E}_{t,h}$

Now by definition of $\hat{g}_t^h$, we know that with probability 1 it holds that $\sum_{s\in[t-1]} \left(\hat{g}_t^h\left(z_s^h\right) - \psi_t^h\left(z_s^h\right)\right)^2 \le 2\sum_{s\in[t-1]} \left(\left(r_s^h + f_{t,1}^{h+1}\left(x_s^{h+1}\right)\right)^2 - \psi_t^h(z_s^h)\right)\left(\hat{g}_t^h\left(z_s^h\right) - \psi_t^h\left(z_s^h\right)\right)$. This event can be equivalently expressed as This event can be equivalently expressed as

$$\left\{\begin{array}{l}
\sum_{s\in[t-1]} \left(\hat{g}_t^h\left(z_s^h\right) - \psi_t^h\left(z_s^h\right)\right)^2 \le 2\sum_{s\in[t-1]} \left(\left(r_s^h + f_{t,1}^{h+1}\left(x_s^{h+1}\right)\right)^2 - \psi_t^h(z_s^h)\right)\left(\hat{g}_t^h\left(z_s^h\right) - \psi_t^h\left(z_s^h\right)\right) \\
\sum_{s\in[t-1]} \left(\hat{g}_t^h\left(z_s^h\right) - \psi_t^h\left(z_s^h\right)\right)^2 > \left(\bar{\beta}_t^h\right)^2
\end{array}\right\}.$$

Now for any given pair of $f \in \mathcal{F}^h$ and f' such that $f' = \min(f'' + \epsilon, 1)$ for some $f'' \in \mathcal{F}^{h+1} + \mathcal{W}$, when $\hat{g}_t^h = f$ and $f_{t,1}^{h+1} = f'$ we define random variable $\eta_s^h[f'] = \left(r_s^h + f'\left(x_s^{h+1}\right)\right)^2 - \mathbb{E}\left[\left(r^h + f'\left(x^{h+1}\right)\right)^2 | z_s^h\right]$ and the martingale difference sequence $\mathsf{D}_s^h[f, f'] = 2\eta_s^h[f'] \cdot \left(f\left(z_s^h\right) - \mathcal{T}_2 f'\left(z_s^h\right)\right)$.

Equation (40) of Lemma 29 implies that for each particular pair of (f, f') where $\hat{g}_t^h = f$ and $f_{t,1}^{h+1} = f'$, for any function $\bar{f}[f'] \in \mathcal{F}^h$ satisfying $\|\bar{f}[f'] - \mathcal{T}_2 f'\|_\infty \leq \epsilon$, it holds that with probability at least $1 - \delta_{t,h}/\mathcal{N}^2\mathcal{N}_b$, we have

$$\begin{aligned}
&2\sum_{s\in[t-1]}\left(\eta_s^h[f'] + \mathcal{T}_2 f'\left(z_s^h\right) - \bar{f}[f'](z_s^h)\right)\left(f(z_s^h) - \bar{f}[f'](z_s^h)\right)\\
&\quad\leq \sum_{s\in[t-1]} 2\eta_s^h[f']\left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right) + 4tL\epsilon + 8L\epsilon\\
&\quad\leq 4(11L+9)\left(\iota'(\delta_{t,h})\right)^2 + 12tL\epsilon + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \mathcal{T}_2 f'(z_s^h)\right)^2}{2}\\
&\quad\leq 4(11L+9)\left(\iota'(\delta_{t,h})\right)^2 + 16tL\epsilon + \frac{\sum_{s\in[t-1]}\left(f(z_s^h) - \bar{f}[f'](z_s^h)\right)^2}{2}.
\end{aligned}$$

Consequently, by union bound over all choices of $f \in \mathcal{F}^h$ and $f' = \min(1, f'' + 3\epsilon) = f_{t,2}^{h+1}$ where $f'' \in \mathcal{F}^{h+1} + \mathcal{W}$, similar to Lemma 24 we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{t,h}) \leq \mathbb{P}\Bigg(&\sum_{s\in[t-1]} 2\left(\eta_s^h + \mathbb{E}[(r^h + f_{t,2}^{h+1}(x^{h+1}))^2|z_s^h] - \psi_t^h(z_s^h)\right)\left(\hat{g}_t^h(z_s^h) - \psi_t^h(z_s^h)\right) >\\
&4(11L+9)\left(\iota'(\delta_{t,h})\right)^2 + 16tL\epsilon + \frac{\sum_{s\in[t-1]}\left(\hat{g}_t^h(z_s^h) - \psi_t^h(z_s^h)\right)^2}{2}\Bigg) \leq \delta_{t,h},
\end{aligned}$$

and consequently,

$$\mathbb{P}\left(\psi_t^h \notin \mathcal{G}_t^h\right) \leq \delta_{t,h},$$

which implies with probability $1 - \delta_{t,h}$, $\psi_t^h \in \mathcal{G}_t^h$ for any fixed given $t \in [T-1]$, $h \in [H]$ (the case $t = 0$ holds with probability 1 by definition). ■

We also give the following consequence of Lemma 28 together with the definition of generalized Eluder dimension, which will be useful to justify our definition of σ_t^h in Equation (4).

Lemma 30 *Conditioning on the good event $\bar{\mathcal{E}}_t^h$, we have*

$$|\psi_t^h(z_t^h) - \hat{g}_t^h(z_t^h)| \leq D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h)\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda}.$$

Proof The proof is similar to that of Lemma 27 so we omit here for brevity. ■

F.3. Validity of Variance Estimator

In this section, we show that our variance over-estimate σ_s^h at iteration s is valid for all iterations afterwards $t \geq s$, and bound its difference with the true variance. We recall the definition of $f_{t,j}^h$ in Algorithm 1 such that $f_{t,1}^h(\cdot) := \min\left(\hat{f}_{t,1}^h(\cdot) + b_{t,1}^h(\cdot) + \epsilon, 1\right)$ and $f_{t,2}^h(\cdot) := \min\left(\hat{f}_{t,2}^h(\cdot) + 2b_{t,1}^h(\cdot) + b_{t,2}^h(\cdot) + 3\epsilon, 2\right)$, $f_{t,-2}^h(\cdot) := \max\left(\hat{f}_{t,-2}^h(\cdot) - b_{t,2}^h(\cdot) - \epsilon, 0\right)$, and $\bar{f}_{t,j}^h$ as the conditional expectations. Abusing notation again, we use $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot|z_t^h]$ where the randomness is taken with respect to r^h and x^{h+1} conditioning on z_t^h when the meaning is clear from context. We also recall the definition of event $\mathcal{E}_t^h = \{\bar{f}_{t,j}^h \in \mathcal{F}_{t,j}^h$ for all $j = 1, \pm 2$ and $\psi_t^h \in \mathcal{G}_t^h\}$ and $\mathcal{E}_{\leq t} = \cap_{s \leq t, h \in [H]} \mathcal{E}_s^h$.

First, we show that $f_{t,j}^h$ satisfies a pointwise monotonic relation conditioning on previous events. This is an important property that we need to satisfy for fulfilling the assumptions (25) and (26) required in Lemma 21, and a reason that we design overly optimistic sequence using *unweighted regression*.

We first state Lemma 31 and then provide its full proof right after.

Lemma 31 (Pointwise monotonicity) *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. For any fixed $t \in [T]$, and $h \in [H]$, conditioning on events $\mathcal{E}_{\leq t-1} \cap \left(\cap_{h'=h}^H \mathcal{E}_t^{h'}\right)$, we have for all $z^h, z^{h-1} \in \mathcal{X} \times \mathcal{A}$,*

1. $f_\star^h(z^h) \leq f_{t,1}^h(z^h)$;
2. $f_{t,-2}^h(z^h) \leq f_\star^h(z^h)$;
3. $f_{s,2}^h(z^h) \geq \max\left(\mathcal{T} f_{t,1}^{h+1}(z^h), f_{t,1}^h(z^h)\right)$ *for all $s \in [t]$.*

Proof We use induction to prove each inequality. Note that under the conditioning $\mathcal{E}_{\leq t-1} \cap \left(\cap_{h'=h}^H \mathcal{E}_t^{h'}\right)$, we have that $\bar{f}_{t,j}^{h'} \in \mathcal{F}_{t,j}^{h'}$ for $j = 1, \pm 2$ and all $h' \geq h$, by definition of the events.

For the first inequality, note that at step t this holds trivially for $h' = H + 1$. Now suppose this holds for some $h + 1 \leq h' + 1 \leq H + 1$, i.e. we have $f_\star^{h'+1}(z^{h'+1}) \leq f_{t,1}^{h'+1}(z^{h'+1})$ for any $z^{h'+1}$ and thus $f_\star^{h'+1}(x^{h'+1}) \leq f_{t,1}^{h'+1}(x^{h'+1})$ for any $x^{h'+1}$. Then for level h', we have conditioning on $\mathcal{E}_t^{h'}$, for any $z^{h'}$, it holds that

$$\begin{aligned}\hat{f}_{t,1}^{h'}(z^{h'}) + b_{t,1}^{h'}(z^{h'}) + \epsilon &\geq \bar{f}_{t,1}^{h'}(z^{h'}) + \epsilon \\ &\geq \mathbb{E}\left[r^{h'} + f_{t,1}^{h'+1}(x^{h'+1})|z^{h'}\right] \geq \mathbb{E}\left[r^{h'} + f_\star^{h'+1}(x^{h'+1})|z^{h'}\right] = f_\star^{h'}(z^{h'}),\end{aligned}$$

where the first inequality is due to the definition of bonus term $b_{t,1}^{h'}$ as in Definition 5 and conditioning event $\mathcal{E}_t^{h'}$ so that $\bar{f}_{t,1}^{h'}(z^{h'}) \in \mathcal{F}_t^{h'}$, the second inequality is due to definition of $\bar{f}_{t,1}^{h'}$ and the third inequality is due to induction.

Recall the definition of $f_{t,1}^{h'}(\cdot) := \min\left(\hat{f}_{t,1}^{h'}(\cdot) + b_{t,1}^{h'}(\cdot) + \epsilon, 1\right)$ in Line 11 of Algorithm 1, together with the upper bound of 1 for $f_\star$ by the sparse reward assumption, we have consequently $f_{t,1}^{h'}(z^{h'}) \geq f_\star^{h'}(z^{h'})$ for any $z^{h'} \in \mathcal{X} \times \mathcal{A}$ and any $h' \geq h$, which proves the inequality when $h' = h$.

For the second inequality, note it also holds trivially for $h' = H+1$. Now suppose this holds for some $h+1 \le h'+1 \le H+1$, i.e. we have $f_{t,-2}^{h'+1}(\cdot) \le f_\star^{h'+1}(\cdot)$. Then for level h', conditioning on $\mathcal{E}_t^{h'}$ we have for any $z^{h'}$,

$$\begin{aligned}\hat{f}_{t,-2}^{h'}(z^{h'}) - b_{t,2}^{h'}(z^{h'}) - 2\epsilon &\overset{(i)}{\le} \bar{f}_{t,-2}^{h'}(z^{h'}) - 2\epsilon \overset{(ii)}{\le} \mathbb{E}[r^{h'} + f_{t,-2}^{h'+1}(x^{h'+1})|z^{h'}] - \epsilon \\ &\overset{(iii)}{\le} \mathbb{E}[r^{h'} + f_\star^{h'+1}(x^{h'+1})|z^{h'}] - \epsilon \le f_\star^{h'}(z^{h'}).\end{aligned}$$

Here we use (i) the definition of $b_{t,2}^{h'}$ and conditioning event of $\mathcal{E}_t^{h'}$, (ii) the definition of $\bar{f}_{t,-2}^{h'}$, and (iii) the induction assumption. Recall the definition of $f_{t,-2}^{h'}(\cdot) := \max\left(\hat{f}_{t,-2}^{h'}(\cdot) - b_{t,2}^{h'}(\cdot) - \epsilon, 0\right)$ in Line 16 of Algorithm 1, together with the fact that in above display RHS ≥ 0 always holds true by definition, we thus conclude by taking max with 0 in above inequality that $f_{t,-2}^{h'}(z^{h'}) \le f_\star^{h'}(z^{h'})$ for all $h' \ge h$ and specifically for $h' = h$.

For the third inequality given any fixed $s \le t$, note it also holds trivially for $h = H+1$. Now suppose this holds for some $h+1 \le h'+1 \le H+1$, i.e. we have $f_{s,2}^{h'+1}(z^{h'+1}) \ge f_{t,1}^{h'+1}(z^{h'+1})$ for all $z^{h'+1}$, and consequently $f_{s,2}^{h'+1}(x^{h'+1}) \ge f_{t,1}^{h'+1}(x^{h'+1})$ for all $x^{h'+1}$. Then for level h', conditioning on $\mathcal{E}_s^{h'}$ and $\mathcal{E}_t^{h'}$ we have for any $z^{h'}$, it holds that

$$\begin{aligned}\hat{f}_{s,2}^{h'}(z^{h}) + 2b_{s,1}^{h'}(z^{h'}) + b_{s,2}^{h'}(z^{h'}) + 3\epsilon &\overset{(i)}{\ge} \mathbb{E}[r^{h'} + f_{s,2}^{h'+1}(x^{h'+1})|z^{h'}] + 2b_{s,1}^{h'}(z^{h'}) + 2\epsilon \\ &\overset{(ii)}{\ge} \mathbb{E}[r^{h'} + f_{t,1}^{h'+1}(x^{h'+1})|z^{h'}] + 2b_{t,1}^{h'}(z^{h'}) + 2\epsilon.\end{aligned} \tag{41}$$

Here we use (i) the definition of $b_{s,2}$ so that $\hat{f}_{s,2}^{h'}(z^{h'}) + b_{s,2}^{h'}(z^{h'}) \ge \bar{f}_{s,2}^{h'}(z^{h'})$ conditioning on $\mathcal{E}_s^{h'}$ and definition of $\bar{f}_{s,2}^{h'}(z^{h'}) + \epsilon \ge \mathbb{E}[r^{h'} + f_{s,2}^{h'+1}(z^{h'+1})|z^{h'}]$, (ii) the induction assumption together with the consistency condition on bonus thus that $b_{s,1}^{h'}(z^{h'}) \ge b_{t,1}^{h'}(z^{h'})$. Recall definition $f_{s,2}^{h'}(\cdot) := \min\left(\hat{f}_{s,2}^{h'}(\cdot) + 2b_{s,1}^{h'}(\cdot) + b_{s,2}^{h'}(\cdot) + 3\epsilon, 2\right)$ in Line 15 of Algorithm 1, by taking min with 2 on both sides of (41) and use non-negativity of $b_{t,1}$, we have $f_{s,2}^{h'}(z^{h'}) \ge \mathcal{T} f_{t,1}^{h'+1}(z^{h'})$.

Additionally, we also have

$$\mathbb{E}[r^{h'} + f_{t,1}^{h'+1}(x^{h'+1})|z^{h'}] + 2b_{t,1}^{h'}(z^{h'}) + 2\epsilon \ge \hat{f}_{t,1}^{h'}(z^{h'}) + b_{t,1}^{h'}(z^{h'}) + \epsilon$$

using $\hat{f}_{t,1}^{h'}(z^{h'}) \le \mathbb{E}[r^{h'} + f_{t,1}^{h'+1}(x^{h'+1})|z^{h'}] + b_{t,1}^{h'}(z^{h'}) + \epsilon$ due to definition of $b_{t,1}$ conditioning on $\mathcal{E}_t^{h'}$. Now taking min with 2 on both sides we also obtain $f_{s,2}^{h'}(z^{h'}) \ge f_{t,1}^{h'}(z^{h'})$ for all $h' \ge h$ to make the inductive argument. And thus the third inequality also holds when $h' = h$.

The inequalities for all x^h is an immediate consequence of taking maximum over $a^h \in \mathcal{A}$ for each inequality. ■

Such point-wise monotonicity also allows us to prove the upper bound on constructed variance estimator σ_t^h for each iteration $t \in [T]$ and level $h \in [H]$, formally stated in the next lemma.

Lemma 32 (Lower bound of variance estimator) *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. At step $t \ge 2$, conditioning on the good event $\mathcal{E}_{\le t}$, the variance estimate σ_t^h satisfies $\left(\sigma_t^h\right)^2 \ge \mathbb{V}_{r^h, x^{h+1}}\left[r^h + f_\star^{h+1}(x^{h+1})|z_t^h\right]$ for all $h \in [H]$.*

Proof Fix any $h \in [H]$, we first consider proving the stated inequality for

$$\left(\tilde{\sigma}_t^h\right)^2 := \hat{g}_t^h(z_t^h) - \left(\hat{f}_{t,-2}^h(z_t^h)\right)^2 + D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \left(\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 2L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right) + 2(1+L)\epsilon.$$

Conditioning on the good event $\mathcal{E}_t$, by Lemma 27 and Lemma 30, we know that

$$(\tilde{\sigma}_t^h)^2 \geq \psi_t^h(z_t^h) - \left(\bar{f}_{t,-2}^h(z_t^h)\right)^2 + (1+2L)\epsilon. \tag{42}$$

Plugging $\left|\psi_t^h(z_t^h) - \mathbb{E}\left[(r^h + f_{t,1}^{h+1}(x^{h+1}))^2|z_t^h\right]\right| \leq \epsilon$, $\left|\bar{f}_{t,-2}^h(z_t^h) - \mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1})|z_t^h\right]\right| \leq \epsilon$ and $\left(\bar{f}_{t,-2}^h(z_t^h) + \mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1})|z_t^h\right]\right) \leq (1+2L)$ into Equation (42), we further have

$$(\tilde{\sigma}_t^h)^2 \geq \mathbb{E}\left[(r^h + f_{t,1}^{h+1}(x^{h+1}))^2|z_t^h\right] - \left(\mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1})\right]\right)^2.$$

Now using the monotonic property $f_{t,1}^{h+1}(\cdot) \geq f_\star^{h+1}(\cdot) \geq f_{t,-2}^{h+1}(\cdot) \geq 0$ conditioning on $\mathcal{E}_{\leq t}$, we have $\left(\tilde{\sigma}_t^h\right)^2 \geq \mathbb{V}\left[r^h + f_\star^{h+1}(x^{h+1})|x_t^h, a_t^h\right]$.

So far we have proven the stated inequality holds for $\left(\tilde{\sigma}_t^h\right)^2$. To show the inequality also holds true for $\left(\sigma_t^h\right)^2 = \min\left(4, \left(\tilde{\sigma}_t^h\right)^2\right)$, we take minimum with 4 and note $\mathbb{V}\left[r^h + f_\star^{h+1}(x^{h+1})|x_t^h, a_t^h\right] \leq 4$ always holds true. ■

The previous two lemmas on point-wise monotonicity and variance lower bound of $\left(\sigma_t^h\right)^2$ immediately imply that the good event $\mathcal{E}_{\leq T}$ happens with high probability, following from an inductive argument.

Proposition 33 *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. With probability $1-5\delta$, the good event $\mathcal{E}_{\leq T}$ happens, that is, $\bar{f}_{t,1}^h \in \mathcal{F}_{t,1}^h$, $\bar{f}_{t,\pm 2} \in \mathcal{F}_{t,\pm 2}^h$ and $\psi_t^h \in \mathcal{F}_t^h$ for all $t \in [T]$ and $h \in [H]$.*

Proof For any $t \geq 1$, conditioning on $\mathcal{E}_{\leq t} \cap \left(\cap_{h'=h+1}^H \mathcal{E}_{t+1}^{h'}\right)$, we first show the assumptions needed for step $t+1, h$ in Lemma 21 holds with probability 1.

The first assumption $\left(\sigma_s^h\right)^2 \geq \mathbb{V}\left[r^h + f_\star^{h+1}(x^{h+1})|z_s^h\right]$ for all $s \in [t]$ in (25) holds due to Lemma 32.

For second assumption in (26), it holds naively when $h = H$ since $f_{s,-2}^H \leq f_{s,2}^H$ point-wise for all $s \in [t]$ using Lemma 31. When $h < H$ we have $f_\star^{h+1}(x^{h+1}) \leq f_{t+1,1}^{h+1}(x^{h+1})$ for all x^{h+1} conditioning on the event $\mathcal{E}_{\leq t} \cap \left(\cap_{h'=h+1}^H \mathcal{E}_{t+1}^{h'}\right)$ due to the first inequality of Lemma 31. Consequently,

$$\begin{aligned}\mathbb{E}\left[|f_{t+1,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})| \mid z_s^h\right] &= \mathbb{E}\left[f_{t+1,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1}) \mid z_s^h\right] \\ &= \mathcal{T} f_{t+1,1}^{h+1}(z_s^h) - f_\star^h(z_s^h).\end{aligned}$$

Now since we condition on $\mathcal{E}_{\leq t} \cap \left(\cap_{h'=h+1}^{H} \mathcal{E}_{t+1}^{h'}\right)$, the second inequality of Lemma 31 implies that $f_{s,-2}^h(z_s^h) \leq f_\star^h(z_s^h)$ and the third inequality of Lemma 31 implies that $\mathcal{T} f_{t+1,1}^{h+1}(z_s^h) \leq f_{s,2}^h$ for all $s \in [t]$. Plugging these inequalities back we have $\mathbb{E}\left[|f_{t+1,1}^{h+1}(x^{h+1}) - f_\star^{h+1}(x^{h+1})| \mid z_s^h\right] \leq f_{s,2}^h(z_s^h) - f_{s,-2}^h(z_s^h)$ for all $s \in [t]$ and $z_s^h \in \mathcal{X} \times \mathcal{A}$. This also shows that the second assumption required in (26) holds.

Thus we have shown conditioning on $\mathcal{E}_{\leq t} \cap \left(\cap_{h'=h+1}^{H} \mathcal{E}_{t+1}^{h'}\right)$, the event $\mathcal{E}_{t+1}^h$ happens with probability $1 - 5\delta_{t+1,h}$ due to Lemmas 21, 24, 26 and 28. Taking a union bound and note $\delta_{t,h} = \delta/(T+1)(H+1)$ we thus conclude that with probability $1 - 5\delta$ the good event $\mathcal{E}_{\leq T}$ happens. ■

Next, we also provide an upper bound on the variance estimator $\left(\sigma_t^h\right)^2$. It shows the estimator is not much bigger than the variance when taking greedy policy induced by optimistic function $f_{t,1}$.

Lemma 34 (Upper bound of variance estimator) *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. For any step $t \geq 2$ conditioning on the good event $\mathcal{E}_{\leq t}$, the variance we estimate σ_t^h satisfies*

$$\begin{aligned}\left(\sigma_t^h\right)^2 &\leq \mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | x_t^h, a_t^h\right] + 4\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) \\ &\quad + 4\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \cdot \left(2\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 4L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right)\right) + 4(2+L)\epsilon.\end{aligned}$$

Proof

Condioning on $\mathcal{E}_{\leq t}$, we have $\bar{f}_{t,-2}^h \in \mathcal{F}_{t,2}^h$ and $\psi_t^h \in \mathcal{G}_t^h$ due to Lemma 27 and Lemma 30. Thus, by definition of bonus oracle and definition of σ_t^h we have

$$\begin{aligned}(\sigma_t^h)^2 &\leq \psi_t^h(z_t^h) - \left(\bar{f}_{t,-2}^h(z_t^h)\right)^2 \\ &\quad + D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \left(2\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 4L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right) + 2(1+L)\epsilon. \quad (43)\end{aligned}$$

Recall $\left|\psi_t^h(z_t^h) - \mathbb{E}\left[(r^h + f_{t,1}^{h+1}(x^{h+1}))^2 | z_t^h\right]\right| \leq \epsilon$ and $\left|\bar{f}_{t,-2}^h(z_t^h) - \mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1}) | z_t^h\right]\right| \leq \epsilon$, Equation (43) implies

$$\begin{aligned}\left(\sigma_t^h\right)^2 \leq &\mathbb{E}\left[(r^h + f_{t,1}^{h+1}(x^{h+1}))^2 | z_t^h\right] - \left(\mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1}) | z_t^h\right]\right)^2 \\ &+ D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \left(2\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 4L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right) + 4(1+L)\epsilon.\end{aligned} \quad (44)$$

Note we have

$$\begin{aligned}&\left(\mathbb{E}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right]\right)^2 - \left(\mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1}) | z_t^h\right]\right)^2 \\ &\overset{(i)}{\leq} 4\mathbb{E}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) - \left(r^h + f_{t,-2}^{h+1}(x^{h+1})\right) | z_t^h\right] \\ &\overset{(ii)}{\leq} 4\left(\mathbb{E}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right] - f_{t,-2}^h(z_t^h) + \epsilon\right) \\ &\overset{(iii)}{\leq} 4\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h) + \epsilon\right).\end{aligned} \quad (45)$$

Here we use (i) the size bounds that $r^h, f_{t,1}^{h+1}, f_{t,-2}^{h+1} \in [0,1]$, (ii) $f_{t,-2}^h(z_t^h) \le \bar{f}_{t,-2}^h(z_t^h) \le \mathcal{T} f_{t,-2}^{h+1} + \epsilon$ due to the definition of $b_{t,2}$ and $\bar{f}_{t,-2}^h \in \mathcal{F}_{t,-2}^h$ conditioning on $\mathcal{E}_{\le t}$, and (iii) the inequality that $f_{t,2}^h(z_t^h) \ge \mathcal{T} f_{t,1}^{h+1}(z_t^h)$ conditioning on $\mathcal{E}_{\le t}$ due to the third inequality in Lemma 31.

Plugging (45) back to (44), we have

$$\begin{aligned}\left(\sigma_t^h\right)^2 \le & \mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right] + 4\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h) + \epsilon\right) \\ & + D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \left(2\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 4L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right) + 4(1+L)\epsilon \\ \le & \mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right] + 4\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) \\ & + D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h) \cdot \left(2\sqrt{\left(\bar{\beta}_t^h\right)^2 + \lambda} + 4L\sqrt{\left(\beta_{t,2}^h\right)^2 + \lambda}\right) + 4(2+L)\epsilon.\end{aligned}$$

■

F.4. Approximation Error of Optimistic, Overly Optimistic(Pessimistic) Values

In this section, we will provide a few inequalities for bounding the optimistic values, overly optimistic values, and overly pessimistic values sequence. We hope to show they will not deviate much from the expected value V_t under exploration rule as defined in Equation (24), and thus not deviate much from $V_\star$ as well.

We again recall the definition of $f_{t,j}^h$ for $j = 1, \pm 2$ from (10) and also the use of $\mathcal{T}_{\mathrm{o}}$ and $\mathcal{T}_{\mathrm{oo}}$ for disjoint set of iterations such that $[T] = \mathcal{T}_{\mathrm{o}} \cup \mathcal{T}_{\mathrm{oo}}$ depending on whether $h_t \in [H]$ or not, as in (23).

We also define the martingale difference sequence so that for $j = 1, -2, 2$, $\xi_{t,j}^h := r_t^h + f_{t,j}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h, x^{h+1}}\left[r^h + f_{t,j}^{h+1}(x^{h+1}) | z_t^h\right]$, $\xi_t^h := r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1} | z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]$. Abusing notation again, we use $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h, x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h, x^{h+1}}[\cdot|z_t^h]$ where the randomness is taken with respect to r^h and x^{h+1} conditioning on z_t^h when the meaning is clear from context.

For the overly pessimistic sequence $f_{t,-2}^h$, we first have the following guarantee on its lower bound.

Lemma 35 *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. Conditioning on good event $\mathcal{E}_{\le t}$, recall $\xi_{t,-2}^h = r_t^h + f_{t,-2}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h, x^{h+1}}\left[r^h + f_{t,-2}^{h+1}(x^{h+1}) | z_t^h\right]$, and $\xi_t^h = r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1} | z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]$. Then we have for any $t \in \mathcal{T}$ and any $h \in [H]$, it holds that,*

$$f_{t,-2}^h(z_t^h) - V_t^h \ge -2 \sum_{h \le h' \le H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h \le h' \le H} \left(\xi_t^{h'} - \xi_{t,-2}^{h'}\right) - 2(H-h+1)\epsilon.$$

Proof We recall the definition of ξ_t^h and $\xi_{t,-2}^h$. Similar to the previous lemma, for the base case we have $f_{t,-2}^H(z_t^H) \ge \bar{f}_{t,-2}^H(z_t^H) - 2b_{t,2}^H(z_t^H) - \epsilon \ge V_t^H - 2b_{t,2}^H(z_t^H) - 2\epsilon$ using definition of $f_{t,-2}^H$ as in (10). Now suppose the condition holds true for step $h+1$ where $2 \le h+1 \le H$. That is, it holds $f_{t,-2}^{h+1}(z_t^{h+1}) \geq V_t^{h+1} - 2\sum_{h+1\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h+1\leq h'\leq H}\left(\xi_t^{h'} - \xi_{t,-2}^{h'}\right) - 2(H-h)\epsilon$. By definition, this implies that $f_{t,-2}^{h+1}(x_t^{h+1}) \geq f_{t,-2}^{h+1}(z_t^{h+1}) \geq V_t^{h+1} - 2\sum_{h+1\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h+1\leq h'\leq H}\left(\xi_t^{h'} - \xi_{t,-2}^{h'}\right) - 2(H-h)\epsilon$. Then for $z = z_t^h$ at level h, we have

$$\begin{aligned}
f_{t,-2}^{H}(z_t^h) - V_t^h &= f_{t,-2}^{h}(z_t^h) - \bar{f}_{t,-2}^{h}(z_t^h) + \bar{f}_{t,-2}^{h}(z_t^h) - V_t^h \\
&\overset{(i)}{\geq} -2b_{t,2}^h(z_t^h) - 2\epsilon + \mathbb{E}\left[r^h + f_{t,-2}^{h+1}(x^{h+1}) - \left(r^h + V_t^{h+1}\right)|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \\
&\overset{(ii)}{=} -2b_{t,2}^h(z_t^h) - 2\epsilon + \xi_t^h - \xi_{t,-2}^h + f_{t,-2}^{h+1}(x_t^{h+1}) - V_t^{h+1} \\
&\overset{(iii)}{\geq} -2b_{t,2}^h(z_t^h) + \xi_t^h - \xi_{t,-2}^h \\
&\quad + \left(-2\sum_{h+1\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h+1\leq h'\leq H}\xi_t^{h'} - \sum_{h+1\leq h'\leq H}\xi_{t,-2}^{h'} - 2(H-h)\epsilon\right) - 2\epsilon \\
&= -2\sum_{h\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h\leq h'\leq H}\left(\xi_t^{h'} - \xi_{t,-2}^{h'}\right) - 2(H-h+1)\epsilon.
\end{aligned}$$

Here we use (i) the fact that $\bar{f}_{t,-2}^h \in \mathcal{F}_{t,-2}^h$ by assumption and definition of $b_{t,2}^h$, (ii) definition of $\xi_t^h, \xi_{t,-2}^h$, and (iii) the recursion together with the definition that $f_{t,-2}^{h+1}(x_t^{h+1}) \geq f_{t,-2}^{h+1}(z_t^{h+1})$. ■

Next, we bound the optimistic sequence $f_{t,1}^h$ and the overly optimistic sequence $f_{t,2}^h$, depending on whether $t \in \mathcal{T}_{\mathrm{o}}, h_t = H+1$ or $t \in \mathcal{T}_{\mathrm{oo}}, h_t \in [H]$.

Lemma 36 *Suppose Algorithm 1 uses a consistent bonus oracle satisfying Definition 5. Conditioning on good event $\mathcal{E}_{\leq t}$, recall $\xi_{t,2}^h = r_t^h + f_{t,2}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h,x^{h+1}}\left[r^h + f_{t,2}^{h+1}(x^{h+1})|z_t^h\right]$, $\xi_t^h = r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1}|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]$. Then we have for any $h \geq h_t$,*

$$f_{t,2}^h(x_t^h) - V_t^h \leq 2\sum_{h\leq h'\leq H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h\leq h'\leq H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) + 4(H-h+1)\epsilon.$$

Further, recall $\xi_{t,1}^h = r_t^h + f_{t,1}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h,x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]$, for any $h \leq h_t$ we have

$$\begin{aligned}
f_{t,1}(x_t^h) - V_t^h \leq & 2\sum_{h\leq h'\leq H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h_t\leq h'\leq H} b_{t,2}^{h'}(z_t^{h'}) \\
& + \sum_{h_t\leq h'\leq H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) + \sum_{h\leq h'<h_t}\left(\xi_t^{h'} - \xi_{t,1}^{h'}\right) + 4(H-h+1)\epsilon.
\end{aligned}$$

Proof Conditioning on good event $\mathcal{E}_t$, we prove this by math induction on $s = 1, 2, .., t$ and $h = H, \cdots, 1$.

We first recall the definition of ξ_t^h and $\xi_{t,2}^h$. It is obvious that for $h = H$, we have $\hat{f}_{t,2}^H(z) + 2b_{t,1}^H(z) + b_{t,2}^H(z) \leq \bar{f}_{t,2}^H(z) + 2b_{t,1}^H(z) + 2b_{t,2}^H(z) \leq \mathbb{E}[r^H|z] + 2b_{t,1}^H(z) + 2b_{t,2}^H(z) + \epsilon$ for any $z \in \mathcal{X}\times\mathcal{A}$. Using definition of $f_{t,2}^H$ as in (10), this means in particular we have $f_{t,2}^H(x_t^H) = f_{t,2}^H(z_t^H) \leq V_t^H + 2b_{t,1}^H(z_t^H) + 2b_{t,2}^H(z_t^H) + 4\epsilon$. Now suppose the condition holds true for step

$h+1$ where $h_t+1 \le h+1 \le H$. That is, it holds that $f_{t,2}^{h+1}(x_t^{h+1}) = f_{t,2}^{h+1}(z_t^{h+1}) \le V_t^{h+1} + 2\sum_{h+1\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h+1\le h'\le H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h+1\le h'\le H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right)$. Then for $z = z_t^h$ at level $h \ge h_t$, we have

$$
\begin{aligned}
f_{t,2}^h(z_t^h) - V_t^h &= f_{t,2}^h(z_t^h) - \bar{f}_{t,2}^h(z_t^h) + \bar{f}_{t,2}^h(z_t^h) - V_t^h \\
&\overset{(i)}{\le} 2b_{t,1}^h(z_t^h) + 2b_{t,2}^h(z_t^h) + 4\epsilon + \mathbb{E}\left[r^h + f_{t,2}^{h+1}(x^{h+1}) - \left(r^h + V_t^{h+1}\right)|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \\
&\overset{(ii)}{=} 2b_{t,1}^h(z_t^h) + 2b_{t,2}^h(z_t^h) + 4\epsilon + \xi_t^h - \xi_{t,2}^h + f_{t,2}^{h+1}(x_t^{h+1}) - V_t^{h+1} \\
&\overset{(iii)}{\le} 2b_{t,1}^h(z_t^h) + 2b_{t,2}^h(z_t^h) + \xi_t^h - \xi_{t,2}^h + \left(2\sum_{h+1\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h+1\le h'\le H} b_{t,2}^{h'}(z_t^{h'})\right. \\
&\qquad \left. + \sum_{h+1\le h'\le H}\xi_t^{h'} - \sum_{h+1\le h'\le H}\xi_{t,2}^{h'}\right) + 4(H-h+1)\epsilon \\
&= 2\sum_{h\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h\le h'\le H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h\le h'\le H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) + 4(H-h+1)\epsilon.
\end{aligned}
$$

Here we use (i) the fact that $\bar{f}_{t,2}^{h+1} \in \mathcal{F}_{t,2}^h$ by assumption and definition of b_t^h, (ii) definition of ξ_t^h, $\xi_{t,2}^h$, and (iii) the recursion. By noting due to choice of greedy policy $f_{t,2}^h(z_t^h) = \max_{a^h} f_{t,2}^h(x_t^h, a^h) = f_{t,2}^h(x_t^h)$ for all $h \ge h_t$ concludes the final inequality.

For the second inequality, we note that by Lemma 31 it holds that $f_{t,1}^h(\cdot) \le f_{t,2}^h(\cdot)$ point-wise and consequently $f_{t,1}^h(x_t^h) \le f_{t,2}^h(x_t^h)$ for $h = h_t \in [H]$, which implies when $h = h_t \in [H]$,

$$
f_{t,1}^h(x_t^h) - V_t^h \le 2\sum_{h\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h\le h'\le H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h\le h'\le H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) + 4(H-h+1)\epsilon.
$$

The above case also holds true when $h_t = H+1$. Thus, this shows the base case holds true when $h = h_t \in [H+1]$. Now suppose the inequality we want to show for $f_{t,1}^h$ holds for $h+1$ where $1 \le h+1 \le h_t$, then for $z = z_t^h$ at level h, we have

$$
\begin{aligned}
f_{t,1}^h(z_t^h) - V_t^h &= f_{t,1}^h(z_t^h) - \bar{f}_{t,1}^h(z_t^h) + \bar{f}_{t,1}^h(z_t^h) - V_t^h \\
&\overset{(i)}{\le} 2b_{t,1}^h(z_t^h) + 2\epsilon + \mathbb{E}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) - \left(r^h + V_t^{h+1}\right)|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \\
&\overset{(ii)}{=} 2b_{t,1}^h(z_t^h) + 2\epsilon + \xi_t^h - \xi_{t,1}^h + f_{t,1}^{h+1}(x_t^{h+1}) - V_t^{h+1} \\
&\overset{(iii)}{\le} 2b_{t,1}^h(z_t^h) + \xi_t^h - \xi_{t,1}^h + 2\sum_{h+1\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h_t\le h'\le H} b_{t,2}^{h'}(z_t^{h'}) \\
&\qquad + \sum_{h_t\le h'\le H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) + \sum_{h+1\le h'<h_t}\left(\xi_t^{h'} - \xi_{t,1}^{h'}\right) + 4(H-h+1)\epsilon \\
&\le 2\sum_{h\le h'\le H} b_{t,1}^{h'}(z_t^{h'}) + 2\sum_{h_t\le h'\le H} b_{t,2}^{h'}(z_t^{h'}) + \sum_{h_t\le h'\le H}\left(\xi_t^{h'} - \xi_{t,2}^{h'}\right) \\
&\qquad + \sum_{h\le h'<h_t}\left(\xi_t^{h'} - \xi_{t,1}^{h'}\right) + 4(H-h+1)\epsilon.
\end{aligned}
$$

Here we use (i) the fact that $\bar{f}^h_{t,1} \in \mathcal{F}^h_{t,1}$ by assumption and definition of b^h_t, (ii) definition of $\xi^h_t, \xi^h_{t,1}$, and (iii) the recursion. By noting due to choice of greedy policy $f^h_{t,1}(z^h_t) = \max_{a^h} f^h_{t,1}(x^h_t, a^h) = f^h_{t,1}(x^h_t)$ concludes the final inequality. ■

F.5. Bounding the Regret in Expectation

From now on we will denote the good event $\mathcal{E}_{\leq T}$, which by Proposition 33 happens with probability $1 - 5\delta$.

When $\mathcal{E}_{\leq T}$ happens, following Lemma 36 the regret can be expressed as

$$
\begin{aligned}
R_T &= \sum_{t\in[T]} \left(f^1_\star(x^1_t) - V^1_t\right) \leq O(1) + \sum_{2\leq t\leq T} \left(f^1_{t,1}(x^1_t) - V^1_t\right) \\
&\leq O(1 + TH\epsilon) + 2\sum_{t\in\mathcal{T}_\mathrm{o}} \min\left(1 + L, \sum_{h\in[H]} b^h_{t,1}(z^h_t)\right) + \sum_{t\in\mathcal{T}_\mathrm{o}} \sum_{h\in[H]} \left(\xi^h_t - \xi^h_{t,1}\right) \\
&\quad + 2\sum_{t\in\mathcal{T}_\mathrm{oo}} \min\left(1 + L, \sum_{h\in[H]} b^h_{t,1}(z^h_t)\right) + 2\sum_{t\in\mathcal{T}_\mathrm{oo}} \min\left(1 + L, \sum_{h_t\leq h\leq H} b^h_{t,2}(z^h_t)\right) \\
&\quad + \sum_{t\in\mathcal{T}_\mathrm{oo}} \left(\sum_{1\leq h<h_t} \left(\xi^h_t - \xi^h_{t,1}\right) + \sum_{h_t\leq h\leq H} \left(\xi^h_t - \xi^h_{t,2}\right)\right) \\
&\leq O(1 + TH\epsilon) + 2\underbrace{\sum_{t\in[T]} \sum_{h\in[H]} \min\left(1 + L, b^h_{t,1}(z^h_t)\right)}_{I} + 2\underbrace{\sum_{t\in\mathcal{T}_\mathrm{oo}} \sum_{h\in[H]} \min\left(1 + L, b^h_{t,2}(z^h_t)\right)}_{III} \\
&\quad + \left[\sum_{t\in[T],h\in[H]} \xi^h_t - \sum_{t\in[T]} \sum_{h\in[H]} \xi_{t,1} + \sum_{t\in\mathcal{T}_\mathrm{oo}} \left(\sum_{h_t\leq h\leq H} \xi^h_{t,1} - \sum_{h_t\leq h\leq H} \xi^h_{t,2}\right)\right].
\end{aligned}
\tag{46}
$$

We again recall a few notations that we heavily use throughout this section.

$$
\begin{aligned}
I &:= \sum_{t\in[T]} \sum_{h\in[H]} \min\left(1 + L, b^h_{t,1}(z^h_t)\right), \\
II &:= \sum_{t\in[T]} \sum_{h\in[H]} \min\left(1 + L, b^h_{t,2}(z^h_t)\right), \\
III &:= \sum_{t\in\mathcal{T}_\mathrm{oo}} \sum_{h\in[H]} \min\left(1 + L, b^h_{t,2}(z^h_t)\right).
\end{aligned}
$$

We also recall the following notations

$$\mathcal{T}_{\text{oo}} := \{t \in [T] : \text{there exists some } h \in [H] \text{ that exploration is guided by } f_{t,2}^h\},$$

$$\sum_{t\in[T],h\in[H]} \xi_t^h := \left(\sum_{t\in[T],h\in[H]} r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1}|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]\right),$$

$$\sum_{t\in[T],h\in[H]} \xi_{t,1}^h := \sum_{t\in[T],h\in[H]} \left(r_t^h + f_{t,1}^{h+1}(x_t^{h+1}) - \mathop{\mathbb{E}}_{r^h,x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]\right),$$

$$\sum_{t\in[T],h\in[H]} \xi_{t,\pm 2}^h := \sum_{t\in[T],h\in[H]} \left(r_t^h + f_{t,\pm 2}^{h+1}(x_t^{h+1}) - \mathop{\mathbb{E}}_{r^h,x^{h+1}}\left[r^h + f_{t,\pm 2}^{h+1}(x^{h+1})|z_t^h\right]\right),$$

$$d_\alpha := H^{-1}\left(\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)\right).$$

The last equation defining d_α will be used mainly for notational simplicity.

The next is a simple fact about the defined ξ_t^h and $\xi_{t,j}^h$, which will be useful multiple times in the later analysis. The fact builds on the observation that a_t^h is determined solely by the filtration of $\mathcal{H}_t^{h-1} = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots, r_2^H, x_2^{H+1}; \cdots, x_t^1, r_t^1, \cdots, r_t^{h-1}, x_t^h)$, due to policy exploration rule (22).

Lemma 37 *By definition ξ_t^h and $\xi_{t,j}^h$ for $j = 1, \pm 2$ are all adapted to filtration $\mathcal{H}_t^{h-1}$. They are martingale difference sequence (MDS) satisfying $\mathbb{E}[\xi_t^h|\mathcal{H}_t^{h-1}] = 0$ and $\mathbb{E}[\xi_{t,j}^h|\mathcal{H}_t^{h-1}] = 0$. As an immediate consequence, we have for all $j = 1, \pm 2$,*

$$\mathbb{E}[\sum_{t\in[T]}\sum_{h\in[H]} \xi_t^h] = \mathbb{E}[\sum_{t\in[T]}\sum_{h\in[H]} \xi_{t,j}^h] = 0,$$

$$\textit{and } \mathbb{E}[\sum_{t\in\mathcal{T}_{\text{oo}}}\sum_{h_t\le h\le H} \xi_t^h] = \mathbb{E}[\sum_{t\in\mathcal{T}_{\text{oo}}}\sum_{h_t\le h\le H} \xi_{t,j}^h] = 0.$$

Proof We only prove the last equation for ξ_t^h. Note we can write

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\text{oo}}}\sum_{h_t\le h\le H} \xi_t^h\right] = \mathbb{E}\left[\sum_{t\in[T]}\sum_{1\le h\le H} \mathbb{E}[\mathbf{1}_{\{h_t\ge h\}}\xi_t^h|\mathcal{H}_t^{h-1}]\right]$$
$$\stackrel{(\star)}{=} \mathbb{E}\left[\sum_{t\in[T]}\sum_{1\le h\le H} \mathbf{1}_{\{h_t\ge h\}}\mathbb{E}[\xi_t^h|\mathcal{H}_t^{h-1}]\right] = 0.$$

Here for equation $(\star)$ we use the fact that random variable $\mathbf{1}_{\{h_t\ge h\}}$ is adapted to $\mathcal{H}_t^{h-1}$. Similar for the proof of $\xi_{t,j}^h$. ■

Now first of all, in the analysis we bound term II. To do so we rely on the definition of bonus oracle and the assumption of bounded generalized Eluder dimension.

Lemma 38 (Crude bound on II) *Given $b_{t,2}(\cdot) \le C \cdot \left(D_{\mathcal{F}^h}(\cdot; z^h_{[t-1]}, \mathbf{1}^h_{[t-1]})\sqrt{\left(\beta^h_{t,2}\right)^2 + \lambda} + \epsilon_{\mathrm{b}} \cdot \beta^h_{t,2}\right)$, when $\lambda = \Theta(1)$, $\alpha \le 1$, we have for a subset $\mathcal{T} \subseteq [T]$ the following inequality holds true:*

$$\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1+L, b^h_{t,2}(z^h_t)\right) = O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon} \cdot \left(H \cdot d_\alpha + H\sqrt{|\mathcal{T}| \cdot d_\alpha} + |\mathcal{T}|H\epsilon_{\mathrm{b}}\right)\right),$$

This immediately implies that

$$II := \sum_{t\in[T]}\sum_{h\in[H]} \min\left(1+L, b^h_{t,2}(z^h_t)\right) = O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon} \cdot \left(H \cdot d_\alpha + H\sqrt{T \cdot d_\alpha} + TH\epsilon_{\mathrm{b}}\right)\right).$$

Proof We first note that by assumption of $b_{t,2}$,

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1+L, b^h_{t,2}(z^h_t)\right)\\
&\overset{(i)}{=} O\left(\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]}) \cdot \sqrt{\left(\beta^h_{t,2}\right)^2 + \lambda}\right) + |\mathcal{T}|H\epsilon_{\mathrm{b}} \cdot \max_{t\in[T],h\in[H]} \beta^h_{t,2}\right)\\
&\overset{(ii)}{=} O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon} \cdot \left(\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]})\right) + |\mathcal{T}|H\epsilon_{\mathrm{b}}\right)\right).
\end{aligned}$$

Here we use (i) the assumption on $b^h_{t,2}$ and (ii) the definition of $\beta^h_{t,2}$ as in Equation (17).

Now we divide the indices of $(t,h) \in \mathcal{T} \times [H]$ in two cases:

$$\begin{aligned}
\mathcal{I}_1 &= \{(t,h) \in \mathcal{T} \times [H] \mid D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]}) \ge 1\},\\
\mathcal{I}_2 &= \{(t,h) \in \mathcal{T} \times [H] \mid D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]}) < 1\}.
\end{aligned}$$

We then consider the summation of terms respectively, note

$$\sum_{(t,h)\in\mathcal{I}_1} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]})\right) \le \sum_{(t,h)\in\mathcal{I}_1} D^2_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]}) \le H \cdot d_\alpha,$$

where the last inequality holds for any $\alpha \le 1$.

Also using Cauchy-Schwarz inequality,

$$\begin{aligned}
\sum_{(t,h)\in\mathcal{I}_2} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]})\right) &\le \sqrt{\sum_{(t,h)\in\mathcal{I}_2} 1^2} \cdot \sqrt{\sum_{(t,h)\in\mathcal{I}_2} D^2_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]})}\\
&\le H\sqrt{|\mathcal{T}| \cdot d_\alpha},
\end{aligned}$$

where the last inequality holds again for any $\alpha \le 1$.

Combining two terms together,

$$\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \mathbf{1}_{[t-1]})) \le O\left(H \cdot d_\alpha + H\sqrt{|\mathcal{T}| \cdot d_\alpha}\right). \tag{47}$$

■

Using the same idea we could get a similar crude bound for term I, formally as follows:

Lemma 39 (Crude bound on I**)** *Given* $b_{t,1}(\cdot) \le C\left(\cdot D_{\mathcal{F}^h}(\cdot; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})\sqrt{\left(\beta^h_{t,1}\right)^2 + \lambda} + \epsilon_{\mathrm{b}} \cdot \beta^h_{t,1}\right)$*, when* $\lambda = \Theta(1)$*,* $\alpha \le 1$*, we have for a subset* $\mathcal{T} \subseteq [T]$ *the following inequality holds true:*

$$\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1+L, b^h_{t,1}(z^h_t)\right) =$$
$$O\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}} \cdot H\sqrt{|\mathcal{T}| \cdot d_\alpha} + \log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta} \cdot Hd_\alpha + |\mathcal{T}|H\epsilon_{\mathrm{b}}\right)\right).$$

Proof We first note that by assumption of $b_{t,1}$,

$$\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1+L, b^h_{t,1}(z^h_t)\right)$$
$$\stackrel{(i)}{=} O\left(\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}) \cdot \sqrt{\left(\beta^h_{t,1}\right)^2 + \lambda}\right) + |\mathcal{T}|H\epsilon_{\mathrm{b}} \cdot \max_{t\in\mathcal{T}, h\in[H]} \beta^h_{t,1}\right)$$
$$\stackrel{(ii)}{=} O\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\sum_{t\in\mathcal{T}}\sum_{h\in[H]} \min\left(1, \bar{\sigma}^h_t \cdot \left(\bar{\sigma}^h_t\right)^{-1} D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})\right) + |\mathcal{T}|H\epsilon_{\mathrm{b}}\right)\right).$$

Here we use (i) the assumption on $b^h_{t,1}$ and (ii) the definition of $\beta^h_{t,1}$ as in Equation (13).

Now we divide the indices of $(t,h) \in \mathcal{T} \times [H]$ in the following cases similarly to the previous proof:

$$\mathcal{I}_1 = \{(t,h) \in \mathcal{T} \times [H] \mid \left(\bar{\sigma}^h_t\right)^{-1} D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}) \ge 1\},$$
$$\mathcal{I}_2 = \{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}^h_t = \alpha\},$$
$$\mathcal{I}_3 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}^h_t = 2\left(\sqrt{\upsilon(\delta_{t,h})} + \iota(\delta_{t,h})\right) \cdot \sqrt{D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})}\right\},$$
$$\mathcal{I}_4 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}^h_t = \sigma^h_t\right\},$$
$$\mathcal{I}_5 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}^h_t = \sqrt{2}\iota(\delta_{t,h})\sqrt{f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)}\right\}.$$

We then consider the summation of terms respectively, for $\mathcal{I}_1$ we have

$$\begin{aligned}\sum_{(t,h)\in\mathcal{I}_1} \min\left(1, \bar{\sigma}^h_t \cdot \left(\bar{\sigma}^h_t\right)^{-1} D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})\right) &\le \sum_{(t,h)\in\mathcal{I}_1} \left(\bar{\sigma}^h_t\right)^{-2} D^2_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]}) \\ &\le \sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h).\end{aligned} \tag{48}$$

For $\mathcal{I}_2$ we use Cauchy-Schwarz inequality to get

$$\begin{aligned}
&\sum_{(t,h)\in\mathcal{I}_2} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \\
&\quad \le \sqrt{\alpha^2 TH} \cdot \sqrt{\sum_{(t,h)\in\mathcal{I}_2} \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)} \\
&\quad \le \sqrt{\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)}.
\end{aligned} \tag{49}$$

For $\mathcal{I}_3$ we have

$$\begin{aligned}
&\sum_{(t,h)\in\mathcal{I}_3} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \\
&\quad \overset{(i)}{\le} \sum_{(t,h)\in\mathcal{I}_3} \left(8\upsilon(\delta_{t,h}) + \iota^2(\delta_{t,h})\right) \cdot \min\left(1, \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \\
&\quad \le O\left(\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}} + \log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}\right) \cdot \sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)\right).
\end{aligned} \tag{50}$$

Here for inequality (i) we use the choice that $\bar{\sigma}_t^h = 2\left(\sqrt{\upsilon(\delta_{t,h})} + \iota(\delta_{t,h})\right)\sqrt{D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}$, dividing both by $\sqrt{\bar{\sigma}_t^h}$ and rearranging gives $\bar{\sigma}_t^h \le 8\left(\upsilon(\delta_{t,h}) + \iota^2(\delta_{t,h})\right)\left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)$, and also the property that $\left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \le 1$ when $(t,h)\in\mathcal{I}_3$ due to definition of $\mathcal{I}_3$.

For $\mathcal{I}_4$ we use Cauchy-Schwarz inequality and upper bound $\bar{\sigma}_t^h = \sigma_t^h = O(1)$ to get

$$\begin{aligned}
&\sum_{(t,h)\in\mathcal{I}_4} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \\
&\quad \le \sqrt{\sum_{(t,h)\in\mathcal{I}_4} \left(\bar{\sigma}_t^h\right)^2} \cdot \sqrt{\sum_{(t,h)\in\mathcal{I}_4} \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)} \\
&\quad \le \sqrt{|\mathcal{T}|H \cdot \left(\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)\right)}.
\end{aligned}$$

For $\mathcal{I}_5$ we use Cauchy-Schwarz inequality to get

$$\begin{aligned}
&\sum_{(t,h)\in\mathcal{I}_5} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \\
&\qquad \le O\left(\sqrt{\sum_{(t,h)\in\mathcal{I}_5} \iota^2(\delta_{t,h})} \cdot \sqrt{\sum_{(t,h)\in\mathcal{I}_5} \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}\right) \\
&\qquad \le O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta}} \cdot \sqrt{|\mathcal{T}|H \cdot \left(\sum_{h\in[H]} \dim_{\alpha,T}(\mathcal{F}^h)\right)}\right),
\end{aligned}$$

where we use $f_{t,2}^h - f_{t,-2}^h = O(1)$ for the first inequality and the definition of $\iota(\delta_{t,h})$ for the second inequality.

Summing all terms above together we have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} \sum_{h\in[H]} \min(1, \bar{\sigma}_t^h \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)) \\
&\qquad \le O\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H \cdot d_\alpha + \sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta}} \cdot H\sqrt{|\mathcal{T}|d_\alpha}\right).
\end{aligned}$$

■

Corollary 40 (Corollary from adapted version using LTV) *Recall the filtration definition*

$$\mathcal{H}_{t-1}^H = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots, r_2^H, x_2^{H+1}; \cdots, x_{t-1}^1, r_{t-1}^1, \cdots, r_{t-1}^H, x_{t-1}^{H+1}).$$

Also we use $\mathbb{E}[\cdot|z_t^h] = \mathbb{E}_{r^h,x^{h+1}}[\cdot|z_t^h]$ and $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$ where the expectation is only taken over r^h and x^{h+1} due to model transition for shorthand. When $L = O(1)$ we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^H \mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right] \mid \mathcal{H}_{t-1}^H\right] \\
&\quad \le O\left(1 + H^2\delta + H^2 \cdot \mathbb{E}\left[\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{oo}}\}} \mid \mathcal{H}_{t-1}^H\right] + H \cdot \mathbb{E}\left[\sum_{h\in[H]} \left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) \mid \mathcal{H}_{t-1}^H\right]\right).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]} \sum_{h\in[H]} \mathbb{V}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]\right] \\
&\qquad \le O\left(T + TH^2\delta + H^2\,\mathbb{E}\left[|\mathcal{T}_{\mathrm{oo}}|\right] + H \cdot \mathbb{E}\sum_{t\in[T]} \sum_{h\in[H]} \left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)\right).
\end{aligned}$$

Proof For the first inequality, applying Proposition 20 we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{V}\Big[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\Big] \mid \mathcal{H}_{t-1}^H\right]\\
&\le 2\,\mathbb{E}\left[\left(\sum_{h=1}^{H} r_t^h - f_{t,1}^1(x_t^1)\right)^2 \mid \mathcal{H}_{t-1}^H\right]\\
&\qquad + 2\,\mathbb{E}\left[\left(\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{o}}\}}\sum_{h=1}^{H}\Big(f_{t,1}^h(z_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\Big)\right)^2 \mid \mathcal{H}_{t-1}^H\right]\\
&\qquad + 2\,\mathbb{E}\left[\left(\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{oo}}\}}\sum_{h=1}^{H}\Big(f_{t,1}^h(x_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\Big)\right)^2 \mid \mathcal{H}_{t-1}^H\right]\\
&\le O(1) + O\left(H\right)\cdot\mathbb{E}\left[\left|\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{o}}\}}\sum_{h\in[H]}\Big(f_{t,1}^h(z_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\Big)\right| \mid \mathcal{H}_{t-1}^H\right]\\
&\qquad + O(H^2)\cdot\mathbb{E}\left[\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{oo}}\}} \mid \mathcal{H}_{t-1}^H\right].
\end{aligned}$$

Here for the last inequality we note $|f_{t,1}^h|\le 1$, so that $\sum_{h=1}^H \Big(f_{t,1}^h(x_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h]\Big) \le O(H)$.

Note we can bound $f_{t,1}^h(z_t^h) - \mathbb{E}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h] \le f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)$ conditioning on $\mathcal{E}_{\le T}$ due to Lemma 31, plugging this back into the above inequality we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{V}[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h] \mid \mathcal{H}_{t-1}^H\right]\\
&\le O\left(1 + H^2\delta + H^2\cdot\mathbb{E}\left[\mathbf{1}_{\{t\in\mathcal{T}_{\mathrm{oo}}\}}\mid\mathcal{H}_{t-1}^H\right] + H\cdot\mathbb{E}\left[\sum_{h\in[H]}\Big(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\Big) \mid \mathcal{H}_{t-1}^H\right]\right).
\end{aligned}$$

The second inequality is an immediate consequence of this corollary together with definition of $\mathcal{T}_{\mathrm{oo}}$. ■

Following the previous expression of regret in (46) and using bound in Lemma 37, we have when $\delta \le 1/6$,

$$\begin{aligned}
\mathbb{E}R_T &= \mathbb{E}\left[\mathbf{1}(\mathcal{E}_{\le T})\mathbb{E}[R_T|\mathcal{E}_{\le T}] + \mathbf{1}(\text{not }\mathcal{E}_{\le T})\mathbb{E}[R_T|\text{ not }\mathcal{E}_{\le T}]\right]\\
&\le O(TH\delta) + (1-5\delta)\mathbb{E}\left[O\left(1+TH\epsilon\right) + 2\cdot I + 2\cdot III \mid \mathcal{E}_{\le T}\right]\\
&\le O\left(TH\delta + TH\epsilon + 1\right) + 2\mathbb{E}[I|\mathcal{E}_{\le T}] + 2\mathbb{E}[III|\mathcal{E}_{\le T}].
\end{aligned} \tag{51}$$

Lemma 41 (Bounding size of $\mathcal{T}_{\mathrm{oo}}$) *Suppose $\alpha \le 1$, we set*

$$u_t \ge C\cdot\left(\frac{\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon}\cdot\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}\cdot H^{5/2}\sqrt{d_\alpha} + \sqrt{t}H\epsilon_{\mathrm{b}}\right)}{\sqrt{t}} + H^2\epsilon + H\delta\right),$$

for some large enough constant $C < \infty$ and $\epsilon \le 1$, then we have the following facts about $\mathcal{T}_{\rm oo}$ holds true:

$$\mathbb{E}\left[|\mathcal{T}_{\rm oo}||\mathcal{E}_{\le T}\right] \le O\left(\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta}\cdot H^3}\right).$$

Proof We will condition on $\mathcal{E}_{\le T}$ throughout the arguments. Now we prove by contradiction, recall the definition of h_t, since for each $t \in \mathcal{T}_{\rm oo}$ we have $f_{t,2}^{h_t}(x_t^{h_t}) \ge f_{t,1}^{h_t}(x_t^{h_t}) + u_t$, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{\rm oo}}\left(f_{t,2}^{h_t}(x_t^{h_t}) - f_{t,1}^{h_t}(x_t^{h_t})\right)\\
&\quad\ge \frac{C}{4}\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}}\cdot\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}}\cdot H^{5/2}\sqrt{d_\alpha}\frac{|\mathcal{T}_{\rm oo}|}{\sqrt{T}}+|\mathcal{T}_{\rm oo}|H\epsilon_{\rm b}\right)\right)\\
&\qquad+\frac{C}{4}\left(|\mathcal{T}_{\rm oo}|H^2\epsilon+|\mathcal{T}_{\rm oo}|H\delta\right).
\end{aligned}$$

Note we also have conditioning on $\mathcal{E}_{\le T}$, since $f_{t,1}^h \ge f_\star^h \ge V_t^h$, it holds that

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{\rm oo}}\left(f_{t,2}^{h_t}(x_t^{h_t}) - f_{t,1}^{h_t}(x_t^{h_t})\right) \le \sum_{t\in\mathcal{T}_{\rm oo}}\left(f_{t,2}^{h_t}(x_t^{h_t}) - V_t^{h_t}\right)\\
&\overset{(i)}{\le} 2\sum_{t\in\mathcal{T}_{\rm oo},h\in[H]}\min\left(4, b_{t,1}^h(z_t^h)\right)+2\sum_{t\in\mathcal{T}_{\rm oo},h\in[H]}\min\left(4, b_{t,2}^h(z_t^h)\right)\\
&\qquad+\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}\left(\xi_t^h-\xi_{t,2}^h\right)+O(|\mathcal{T}_{\rm oo}|H^2\epsilon)\\
&\overset{(ii)}{\le} O\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}}H\sqrt{|\mathcal{T}_{\rm oo}|\cdot d_\alpha}+\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}H\cdot d_\alpha+|\mathcal{T}_{\rm oo}|H\epsilon_{\rm b}\right)\right)\\
&\qquad\quad+\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}\left(\xi_t^h-\xi_{t,2}^h\right)+O(|\mathcal{T}_{\rm oo}|H^2\epsilon),
\end{aligned}$$

where for (i) we use Lemma 36, and for (ii) we use Lemma 38 and Lemma 39 with $\mathcal{T} = \mathcal{T}_{\rm oo}$. Thus, conditioning on $\mathcal{E}_{\le T}$, taking expectation and note

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}(\xi_t^h-\xi_{t,2}^h)|\mathcal{E}_{\le T}\right] &\le O\Big(\mathbb{E}[|\mathcal{T}_{\rm oo}||\mathcal{E}_{\le T}]H\delta+\mathbb{E}\Big[\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}(\xi_t^h-\xi_{t,2}^h)\Big]\Big)\\
&= O\Big(\mathbb{E}[|\mathcal{T}_{\rm oo}||\mathcal{E}_{\le T}]H\delta\Big)
\end{aligned}$$

due to Lemma 37. Thus, in order for the two inequalities hold true simultaneously it must hold that $\mathbb{E}[|\mathcal{T}_{\rm oo}||\mathcal{E}_{\le T}] \le O\left(T/(H^3\cdot\log(\mathcal{N}\mathcal{N}_bTH/\alpha\delta))\right)$. ■

Building on this bound of $|\mathcal{T}_{\rm oo}|$, we show the next corollary on a tighter bound for the summation terms in III.

Corollary 42 (Fine-grained bound on III) *Given* $b_{t,2}(\cdot) \le C \cdot \left(D_{\mathcal{F}^h}(\cdot, z^h_{[t-1]}, \mathbf{1}^h_{[t-1]}) \sqrt{\left(\beta^h_{t,2}\right)^2 + \lambda} + \epsilon_{\mathrm{b}} \cdot \beta^h_{t,2} \right)$ *and using the particular choice of* u_t *as in Lemma 41, when* $\lambda = \Theta(1)$, $\alpha \le 1$, *we have the following inequality holds true:*

$$\begin{aligned}\mathbb{E}[III \mid \mathcal{E}_{\le T}] &:= \mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\sum_{h\in[H]} \min\left(1+L, b^h_{t,2}(z^h_t)\right) \mid \mathcal{E}_{\le T}\right] \\ &= O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T\epsilon}\cdot\sqrt{T\cdot d_\alpha} + \sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon}\cdot(H\cdot d_\alpha + T\epsilon_{\mathrm{b}})\right).\end{aligned}$$

Proof This is an immediate corollary by combining Lemma 38 and Lemma 41. ∎

Next, we proceed to bound $\mathbb{E}[I|\mathcal{E}_{\le T}]$ properly. To do so, we will provide an additional helper lemma before bounding I.

Lemma 43 *When* $\lambda = \Theta(1)$, $\alpha \le 1$, $\epsilon \le 1$, $\delta \le 1/10$, *we have*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in[T]}\sum_{h\in[H]}\left[f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)\right] \mid \mathcal{E}_{\le T}\right] \\ &\qquad \le O\left(H\cdot\mathbb{E}[I|\mathcal{E}_{\le T}] + H\cdot\mathbb{E}[II|\mathcal{E}_{\le T}] + H^2\cdot\mathbb{E}\left[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}\right] + TH^2\epsilon + TH^2\delta\right) + H\cdot\sum_{t\in\mathcal{T}_{\mathrm{o}}} u_t.\end{aligned}$$

Proof For $t \in \mathcal{T}_{\mathrm{oo}}$, it holds that

$$\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\sum_{h\in[H]}\left[f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)\right] = O(|\mathcal{T}_{\mathrm{oo}}|H),$$

and consequently $\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\sum_{h\in[H]}\left[f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)\right] \mid \mathcal{E}_{\le T}\right] = O(\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}| \mid \mathcal{E}_{\le T}]\cdot H)$

Otherwise, for iterations $t \in \mathcal{T}_{\mathrm{o}}$, we know that it always holds true that $f^h_{t,2}(x^h_t) \le f^h_{t,1}(x^h_t) + u_t$, which implies

$$\begin{aligned}f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t) &\overset{(i)}{\le} f^h_{t,1}(z^h_t) - f^h_{t,-2}(z^h_t) + u_t \\ &\overset{(ii)}{\le} u_t + O\left(\sum_{h\le h'\le H} b^{h'}_{t,1}(z^{h'}_t) + \sum_{h\le h'\le H} b^{h'}_{t,2}(z^{h'}_t) + \sum_{h\le h'\le H}\left(-\xi^{h'}_{t,1} + \xi^{h'}_{t,-2}\right) + H\epsilon\right),\end{aligned}$$

where we use (i) the fact that $f^h_{t,2}(z^h_t) \le f^h_{t,1}(z^h_t)$ by Lemma 31 and (ii) the upper bound on $f^h_{t,1}$ as in Lemma 36 when $h_t = H+1$ and the lower bound on $f^h_{t,-1}$ as in Lemma 35 conditioning on $\mathcal{E}_{\le T}$. Also, we note $f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t) \le O(1)$.

Now conditioning on $\mathcal{E}_{\le T}$, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right]\\
&\quad\le O\left(\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,1}^{h'}(z_t^{h'})\right)+\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,2}^{h'}(z_t^{h'})\right)\right)\\
&\qquad+\sum_{t\in\mathcal{T}_{\mathrm{o}}}H\cdot u_t+O\left(\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}-\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}+TH^2\epsilon\right)\\
&\quad\le O\left(\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,1}^{h'}(z_t^{h'})\right)+\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,2}^{h'}(z_t^{h'})\right)\right)\\
&\qquad+\sum_{t\in\mathcal{T}_{\mathrm{o}}}H\cdot u_t+O\left(\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}-\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}+|\mathcal{T}_{\mathrm{oo}}|H^2+TH^2\epsilon\right).
\end{aligned}$$

This implies that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right]\mid\mathcal{E}_{\le T}\right]\\
&\overset{(i)}{\le}O\left(\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,1}^{h'}(z_t^{h'})\right)+\sum_{t\in\mathcal{T}_{\mathrm{o}}}\sum_{h\in[H]}\min\left(1,\sum_{h\le h'\le H}b_{t,2}^{h'}(z_t^{h'})\right)\mid\mathcal{E}_{\le T}\right]\right)\\
&\qquad+\sum_{t\in\mathcal{T}_{\mathrm{o}}}H\cdot u_t+O(H^2\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}]+TH^2\epsilon+TH^2\delta)\\
&\overset{(ii)}{\le}O\left(H\cdot\mathbb{E}[I\mid\mathcal{E}_{\le T}]+H\cdot\mathbb{E}[II\mid\mathcal{E}_{\le T}]+H\sum_{t\in\mathcal{T}_{\mathrm{o}}}u_t+H^2\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}]+TH^2\epsilon+TH^2\delta\right).
\end{aligned}$$

Here we use (i) since $\mathcal{E}_{\le T}$ happens with probability $1-5\delta\ge 1/2$, and $|\xi_{t,j}^h|\le O(1)$ for any t,h and any $j=-2,1$, so that we have $\mathbb{E}\big[\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}\mid\mathcal{E}_{\le T}\big]=O(TH^2\delta)$ and $\mathbb{E}\big[\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}\mid\mathcal{E}_{\le T}\big]=O(TH^2\delta)$ using Lemma 37. For (ii) we simply use the definition of I and II and the fact that all bonus terms are non-negative.

Thus summing the two cases gives the claimed bound. ■

Lemma 44 (Fine-grained bound on I) *Recall the definition of $b_{t,1}$, $b_{t,2}$ as in Lemma 39, Lemma 38. When $\lambda=1$, $\alpha=1/\sqrt{TH}$, $\epsilon\le 1$ and $\delta\le 1/10$, conditioning on the event $\mathcal{E}_{\le T}$, we have the fol-*

lowing inequality holds true:

$$
\begin{aligned}
&\mathbb{E}[I|\mathcal{E}_{\le T}] := \mathbb{E}\left[\sum_{t\in[T]}\sum_{h\in[H]} \min\left(1+L, b_{t,1}^h(z_t^h)\right)|\mathcal{E}_{\le T}\right] \\
&= O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\sqrt{T}\cdot\sqrt{Hd_\alpha}\right) \\
&+ O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{TH^3(\epsilon+\delta)+H^3\mathbb{E}[|\mathcal{T}_{\text{oo}}||\mathcal{E}_{\le T}]+H^2\sum_{t\in\mathcal{T}_{\text{o}}}u_t}\cdot\sqrt{Hd_\alpha}\right) \\
&+ O\left(\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\log^{1.5}\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^{7/2}d_\alpha+\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot TH\epsilon_{\text{b}}\right).
\end{aligned}
$$

Proof We first note that by assumption and definition,

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{h\in[H]} \min\left(1+L, b_{t,1}^h(z_t^h)\right) \\
&\quad = O\left(\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\cdot\sqrt{\left(\beta_{t,1}^h\right)^2+\lambda}\right)+TH\epsilon_{\text{b}}\cdot\max_{t,h}\beta_{t,1}^h\right) \\
&\quad = O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\left(\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)+TH\epsilon_{\text{b}}\right)\right).
\end{aligned} \tag{52}
$$

Treating $L = O(1)$ as defined (see Assumption 1), we now bound the summation terms

$$
\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) = \sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, \bar{\sigma}_t^h\cdot\left(\bar{\sigma}_t^h\right)^{-1}D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)
$$

by dividing into cases.

We consider separating the index set $\{(t,h) : t\in[T], h\in[H]\}$ as follows, same as the cases we consider in Lemma 39.

$$
\begin{aligned}
&\mathcal{I}_1 = \{(t,h)\in\mathcal{T}\times[H] \mid \left(\bar{\sigma}_t^h\right)^{-1}D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\ge 1\}, \\
&\mathcal{I}_2 = \{(t,h)\in\mathcal{T}\times[H] \mid (t,h)\notin\mathcal{I}_1,\ \bar{\sigma}_t^h = \alpha\}, \\
&\mathcal{I}_3 = \left\{(t,h)\in\mathcal{T}\times[H] \mid (t,h)\notin\mathcal{I}_1,\ \bar{\sigma}_t^h = 2\left(\sqrt{\upsilon(\delta_{t,h})}+\iota(\delta_{t,h})\right)\cdot\sqrt{D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}\right\}, \\
&\mathcal{I}_4 = \left\{(t,h)\in\mathcal{T}\times[H] \mid (t,h)\notin\mathcal{I}_1,\ \bar{\sigma}_t^h = \sigma_t^h\right\}, \\
&\mathcal{I}_5 = \left\{(t,h)\in\mathcal{T}\times[H] \mid (t,h)\notin\mathcal{I}_1,\ \bar{\sigma}_t^h = \sqrt{2}\iota(\delta_{t,h})\sqrt{f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)}\right\}.
\end{aligned}
$$

For the terms restricting on $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ we recall the bounds in Equations (48) to (50) in Lemma 39 such that $\sum_{(t,h)\in\mathcal{I}_1\cup\mathcal{I}_2\cup\mathcal{I}_3}\min\left(1, \bar{\sigma}_t^h\cdot\left(\bar{\sigma}_t^h\right)^{-1}D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)\le O\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot Hd_\alpha\right)$.

For terms restricting on $\mathcal{I}_4$ and $\mathcal{I}_5$ we do a tighter analysis different from Lemma 39. For summations terms in $\mathcal{I}_5$, we have

$$\begin{aligned}
&\sum_{(t,h)\in\mathcal{I}_5} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \le \sum_{(t,h)\in\mathcal{I}_5} \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \\
&\quad = \sum_{(t,h)\in\mathcal{I}_5} \sqrt{2}\iota(\delta_{t,h}) \cdot \sqrt{f_{t,2}^h(z_t^h) - f_{t,2}^h(z_t^h)} \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \\
&\quad \le O\left(\sqrt{\log \frac{\mathcal{N}\mathcal{N}_b TH}{\delta}} \sqrt{\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)} \cdot \sqrt{H \cdot d_\alpha}\right).
\end{aligned} \tag{53}$$

Here for the last inequality we use Cauchy-Schwarz inequality together with the definition of $\iota(\delta)$ as in (15).

Restricting on $\mathcal{I}_4$, by Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{(t,h)\in\mathcal{I}_4} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) |\mathcal{E}_{\le T}\right] \\
&\le \mathbb{E}\left[\sum_{(t,h)\in\mathcal{I}_4} \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)|\mathcal{E}_{\le T}\right] \\
&\overset{(o)}{\le} \sqrt{\mathbb{E}\left[\sum_{t,h\in\mathcal{I}_4} \left(\sigma_t^h\right)^2 |\mathcal{E}_{\le T}\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{t,h\in\mathcal{I}_4} \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)|\mathcal{E}_{\le T}\right]} \\
&\overset{(i)}{\le} \sqrt{\mathbb{E}\left[\sum_{t,h\in\mathcal{I}_4} \left(\sigma_t^h\right)^2 |\mathcal{E}_{\le T}\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{t,h\in\mathcal{I}_4} \min\left(1, \left(\bar{\sigma}_t^h\right)^{-2} D^2_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) |\mathcal{E}_{\le T}\right]} \\
&\overset{(ii)}{\le} O\left(\sqrt{\mathbb{E}\left[\sum_{t,h} \mathbb{V}_{r^h, x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right] + \sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) |\mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(\sqrt{\mathbb{E}\left[\sum_{t,h} \min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h)\right) \sqrt{\log \frac{\mathcal{N}\mathcal{N}_b TH}{\delta}} + T\epsilon + TH\epsilon |\mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\overset{(iii)}{\le} O\left(\sqrt{2\mathbb{E}\sum_{t,h} \mathbb{V}_{r^h, x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1})|z_t^h\right] + \mathbb{E}\left[\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) |\mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(\sqrt{\mathbb{E}\left[\sum_{t,h} \min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \mathbf{1}_{[t-1]}^h)\right) \sqrt{\log \frac{\mathcal{N}\mathcal{N}_b TH}{\delta}} + T\epsilon + TH\epsilon |\mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right).
\end{aligned}$$

Here we use (o) the condition that $\left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \le 1$ by definition of $\mathcal{I}_4$, (i) definition of Eluder dimension, (ii) Lemma 34 and (iii) the fact that event $\mathcal{E}_{\le T}$ happens with at least $1-5\delta$ probability so that

$$\mathbb{E}\left[\sum_{t,h} \mathbb{V}_{r^h, x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right] | \mathcal{E}_{\le T}\right] \le \frac{1}{1-5\delta} \mathbb{E} \sum_{t,h} \mathbb{V}_{r^h, x^{h+1}}\left[r^h + f_{t,1}^{h+1}(x^{h+1}) | z_t^h\right].$$

Further, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{(t,h)\in\mathcal{I}_4} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) | \mathcal{E}_{\le T}\right] \\
&\overset{(i)}{\le} O\left(\sqrt{T + H^2 \cdot \mathbb{E}|\mathcal{T}_{\text{oo}}| + TH^2(\epsilon+\delta) + H \cdot \mathbb{E}\left[\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(\sqrt{\mathbb{E}\left[\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) | \mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(\sqrt{\left(H \cdot d_\alpha + H\sqrt{T \cdot d_\alpha}\right)\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon}} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\overset{(ii)}{\le} O\left(\sqrt{T + H^2 \mathbb{E}[|\mathcal{T}_{\text{oo}}| \mid \mathcal{E}_{\le T}] + TH^2(\epsilon+\delta)} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(\sqrt{H \cdot \mathbb{E}\left[\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) | \mathcal{E}_{\le T}\right]} \cdot \sqrt{H \cdot d_\alpha}\right) \\
&\quad + O\left(H\sqrt{d_\alpha \cdot \left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right)} \cdot \sqrt{H \cdot d_\alpha}\right).
\end{aligned}$$

Here for (i) we have plugged in bounds in Corollary 40 and Equation (47). For (ii) in the first line we note $\mathcal{E}_{\le T}$ happens with probability $1-5\delta$ so that $\mathbb{E}[\sum_{t,h}(f_{t,2}^h - f_{t,-2}^h)] \le O(\delta TH + \mathbb{E}[\sum_{t,h}(f_{t,2}^h - f_{t,-2}^h)|\mathcal{E}_{\le T}])$ and $\mathbb{E}|\mathcal{T}_{\text{oo}}| \le O(\delta T + \mathbb{E}[|\mathcal{T}_{\text{oo}}||\mathcal{E}_{\le T}])$ since with probability 1 we have $\sum_{t,h}(f_{t,2}^h - f_{t,-2}^h) \le O(TH)$ and $|\mathcal{T}_{\text{oo}}| \le T$. In the third line of (ii) we also use AM-GM inequality such that

$$H\sqrt{T \cdot d_\alpha\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right)} \le T + H^2 d_\alpha \cdot \left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right). \tag{54}$$

Summing all terms together and taking conditional expectation we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)|\mathcal{E}_{\le T}\right] \\
&= O\left(\sqrt{T + H^2\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}| \mid \mathcal{E}_{\le T}] + TH^2(\epsilon+\delta)}\cdot\sqrt{Hd_\alpha}\right) \\
&\quad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\cdot\sqrt{Hd_\alpha}\cdot\sqrt{H\cdot\mathbb{E}\left[\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)|\mathcal{E}_{\le T}\right]}\right) \\
&\quad + O\left(\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right)\cdot H^{1.5}\cdot d_\alpha\right).
\end{aligned}$$

Now plugging in the bounds proven in Lemma 43 we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1, D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)|\mathcal{E}_{\le T}\right] \\
&= O\left(\sqrt{T}\cdot\sqrt{Hd_\alpha}\right) \\
&\quad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\sqrt{H^3\cdot\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}| \mid \mathcal{E}_{\le T}] + TH^3(\epsilon+\delta) + H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\right) \quad (55)\\
&\quad + O\left(\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right)\cdot H^{1.5}\cdot d_\alpha\right) \\
&\quad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\cdot\sqrt{Hd_\alpha}\cdot\sqrt{H^2\cdot\mathbb{E}\left[I|\mathcal{E}_{\le T}\right] + H^2\cdot\mathbb{E}\left[II|\mathcal{E}_{\le T}\right]}\right).
\end{aligned}$$

We can further bound the last term in the RHS of above inequality as

$$\begin{aligned}
&H^2\cdot\mathbb{E}\left[I|\mathcal{E}_{\le T}\right] + H^2\cdot\mathbb{E}\left[II|\mathcal{E}_{\le T}\right] \\
&\le O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\cdot H^3\sqrt{Td_\alpha}\right) \\
&\quad + O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\log\frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta}\cdot H^2 d_\alpha + \sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T^2H\epsilon}\cdot TH^2\epsilon_{\mathrm{b}}\right) \\
&\le O\left(\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}} + \left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}\right)^2 H^6 d_\alpha\right) \\
&\quad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T^2H\epsilon}\cdot TH^2\epsilon_{\mathrm{b}}\right).
\end{aligned} \quad (56)$$

Here for the first inequality we plug in crude bounds in Lemma 38 and Lemma 39 given the definition of bonus terms, and absorb low-order terms. For the second inequality we use the AM-GM inequality such that

$$\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot H^3\sqrt{Td_\alpha}$$
$$\leq\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}+\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\right)^2H^6d_\alpha,$$

and absorb other low-order terms.

Plugging (56) back to the original bound in (55), rearranging terms and absorbing low-order terms we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1,D_{\mathcal{F}^h}(z_t^h;z_{[t-1]}^h,\bar{\sigma}_{[t-1]}^h)\right)|\mathcal{E}_{\leq T}\right]\\
&\leq O\left(\sqrt{T}\cdot\sqrt{Hd_\alpha}\right)\\
&\quad+O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^3\cdot\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}|\mid\mathcal{E}_{\leq T}]+TH^3(\epsilon+\delta)+H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\right)\\
&\quad+O\left(\sqrt{\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\log^3\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot H^{7/2}d_\alpha+TH\epsilon_{\mathrm{b}}\right).
\end{aligned}$$

Here the low-order $TH\epsilon_{\mathrm{b}}$ term comes from applying AM-GM inequality on the $\mathrm{poly}(\epsilon_{\mathrm{b}})$ term and absorbing other low-order terms by $O\left(\sqrt{(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon)\log^3\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot H^{7/2}d_\alpha\right)$.

Plugging the bound back to (52) and rearranging terms, we have the claimed bounds. ■

Theorem 15 (Bound on expected regret) *Suppose function class $\{\mathcal{F}^h\}_{h\in[H]}$ satisfies Assumption 1 with $\epsilon\in[0,1]$ and Definition 2 with $\lambda=1$, and given consistent bonus oracle $\mathcal{B}$ (output function in class $\mathcal{W}$) satisfying Definition 5, VOQL with $\alpha=\sqrt{1/TH}$, $\delta<1/(T+10)$, $\epsilon\leq 1$ and*

$$u_t=C\cdot\left(\frac{\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}\cdot H^{5/2}\sqrt{d_\alpha}+\sqrt{t}H\epsilon_{\mathrm{b}}\right)}{\sqrt{t}}+H^2\epsilon+H\delta\right)$$

for sufficiently large constant $C<\infty$, achieves a total regret $\mathbb{E}R_T=O\Big(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\sqrt{THd_\alpha}+\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\cdot\left(\log^2\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^5d_\alpha+T^2\epsilon_{\mathrm{b}}^2\right)\Big)$.

Proof Following Equation (51), we have

$$\mathbb{E}R_T=O\left(1+\delta HT+\epsilon HT\right)+2\mathbb{E}\left[I|\mathcal{E}_{\leq T}\right]+2\mathbb{E}\left[III|\mathcal{E}_{\leq T}\right].$$

Now plugging in the guarantees in Lemma 44 for bounding $\mathbb{E}[I|\mathcal{E}_{\le T}]$, and Corollary 42 for bounding $\mathbb{E}\left[III|\mathcal{E}_{\le T}\right]$, we have

$$\mathbb{E}R_T = O\left(1 + TH(\epsilon+\delta) + \sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\sqrt{T}\cdot\sqrt{Hd_\alpha}\right)$$
$$+ O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{TH^3(\epsilon+\delta) + H^3\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}] + H^2\sum_{t\in\mathcal{T}_{\mathrm{o}}} u_t}\cdot\sqrt{Hd_\alpha}\right)$$
$$+ O\left(\left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\log^{1.5}\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^{7/2}d_\alpha + \sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta} + T^2H\epsilon}\cdot TH\epsilon_{\mathrm{b}}\right).$$

Now further plugging in the choice of u_t and $\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}]$ due to Lemma 41, we have

$$\begin{aligned}
&\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^2\mathbb{E}[|\mathcal{T}_{\mathrm{oo}}||\mathcal{E}_{\le T}] + H^2\sum_{t\in\mathcal{T}_{\mathrm{o}}} u_t} \\
&\quad = O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}} + \sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot TH^3\epsilon_{\mathrm{b}} + TH^4\epsilon + TH^3\delta}\right) \\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot\sqrt{\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}H^{4.5}\sqrt{d_\alpha}\sqrt{T}}\right) \\
&\quad \le O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}} + \sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot TH^3\epsilon_{\mathrm{b}} + TH^4\epsilon + TH^3\delta}\right) \\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot\sqrt{\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}} + \left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\cdot\log^3\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}H^9d_\alpha}\right) \\
&\quad = O(\sqrt{T}) \\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot\sqrt{\left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\cdot\left(\log^3\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}H^9d_\alpha + TH^3(\epsilon_{\mathrm{b}}+\delta)\right)}\right)
\end{aligned}$$

which by multiplying both sides with $\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\sqrt{Hd_\alpha}$, plugging back, using AM-GM inequality to simplify the $\mathrm{poly}(\epsilon_{\mathrm{b}},\delta)$ terms gives

$$\begin{aligned}
\mathbb{E}R_T = O&\left(1 + TH(\epsilon+\delta) + \sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\sqrt{T}\cdot\sqrt{Hd_\alpha}\right) \\
&+ O\left(\left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\cdot\left(\log^2\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}H^5d_\alpha + T^2\epsilon_{\mathrm{b}}^2 + T\delta\right)\right).
\end{aligned}$$

Adjusting the constant $\delta \leftarrow \frac{\delta}{5}$, using the range of δ and omitting low-order terms of $\mathrm{poly}(\epsilon)$ conclude the final bound for regret. ∎

Appendix G. Regret with High Probability.

In this section, we provide the full analysis of the high-probability guarantee stated in Theorem 6 of Algorithm 1. We first provide a complete statement of the guarantee as follows.

Theorem 45 (Regret bound with high probability) *Suppose function class $\{\mathcal{F}^h\}_{h\in[H]}$ satisfy Assumption 1 with $\epsilon \in [0,1]$ and Definition 2 with $\lambda = 1$, given consistent bonus oracle $\mathcal{B}$ satisfying Definition 5, Algorithm 1 with $\alpha = \sqrt{1/TH}$, $\delta \le 1/(H^2+11)$, $\epsilon \le 1$ and*

$$u_t = C \cdot \left(\frac{\sqrt{\log \frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\log \frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H^{5/2}\sqrt{d_\alpha} + \sqrt{t}H^2\epsilon_{\mathrm{b}}\right)}{\sqrt{t}} + H^2\epsilon \right)$$

for sufficiently large constant $C < \infty$, with high probability $1-\delta$ event $\mathcal{E} = \mathcal{E}_{\le T} \cap \mathcal{E}_{\xi_1} \cap \mathcal{E}_{\xi_{-2}} \cap \mathcal{E}_{\xi_2} \cap \mathcal{E}_{\xi_{\mathrm{dif}}} \cap \mathcal{E}_{\mathbb{V}}$ happens. Further, when conditioning on $\mathcal{E}$ the algorithm achieves a total regret R_T of

$$O\left(\sqrt{\log \frac{\mathcal{N}TH}{\delta} + T^2H\epsilon} \cdot \sqrt{T} \cdot \sqrt{Hd_\alpha} + \left(\log \frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right) \cdot \left(\log^2 \frac{\mathcal{N}\mathcal{N}_b TH}{\delta} H^5 d_\alpha + T^2\epsilon_{\mathrm{b}}^2\right)\right).$$

The notation of this section is the same as in Appendix F; see Table 2 and Table 3 for the formal definitions. It also builds on Appendices F.1 to F.4. The section can be viewed as an alternative of the analysis in Appendix F.5 for the high-probability setting. It is organized as follows: In Appendix G.1 we prove some concentration properties of the martingale sequences used in bounding the final regret. In Appendix G.2 we bound the size of $|\mathcal{T}_{\mathrm{oo}}|$, showing that with high probability the agent doesn't use $f_{t,2}$ too often in the exploration. In Appendix G.3, we bound the summation of variances incurred in the total exploration in high probability, using the concentration property together with expectation bound shown in Corollary 40. In Appendix G.4, we give the high-probability bound on the summation of differences $\sum_{t,h}(f_{t,2}^h - f_{t,-2}^h)$ which are used in the definition of $\bar{\sigma}_t^h$. Finally, we combine all parts together and bound the summation of bonus terms in Appendix G.5 to prove the final high-probability regret bound.

G.1. Concentration of Random Variables

To turn the in-expectation bound into a high probability argument, we first provide a few concentration results on the random variables ξs, building on their MDS property as stated in Lemma 37.

Lemma 46 *Recall the simplified notation of $\mathbb{V}[\cdot|z_t^h] = \mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$. For given $\delta \in (0,1)$, we have:*

- *For $\{\xi_t^h\}_{t,h}$, we let $\mathcal{E}_\xi$ be the event such that*

$$\left|\sum_{t\in[T],h\in[H]} \xi_t^h\right| \le \sqrt{\sum_{t\in[T],h\in[H]} \mathbb{V}\left[r^h + V_t^{h+1}|z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right] \log\frac{2}{\delta}} + 2\log\frac{2}{\delta}$$
$$\le \sqrt{4TH\log\frac{2}{\delta}} + 2\log\frac{2}{\delta}, \quad (57)$$

  we thus have event $\mathcal{E}_\xi$ happens with probability at least $1-\delta$.

- *For* $\{\xi_{t,1}^h\}_{t,h}$*, we let* $\mathcal{E}_{\xi_1}$ *be the event such that*

$$\left|\sum_{t\in[T],h'\geq h} \xi_{t,1}^{h'}\right| \leq \sqrt{\sum_{t\in[T],h'\geq h} \mathbb{V}[r^{h'} + f_{t,1}^{h'+1}(x^{h'+1})|z_t^{h'}] \log\frac{2H}{\delta}} + 2\log\frac{2H}{\delta}$$
$$\leq \sqrt{4TH\log\frac{2H}{\delta}} + 2\log\frac{2H}{\delta} \text{ for all } h\in[H], \tag{58}$$

we thus have event $\mathcal{E}_{\xi_1}$ *happens with probability at least* $1-\delta$*.*

- *For* $\{\xi_{t,2}^h\}_{t,h}$*, we let* $\mathcal{E}_{\xi_2}$ *be the event such that*

$$\left|\sum_{t\in[T],h\in[H]} \xi_{t,2}^{h}\right| \leq \sqrt{\sum_{t\in[T],h'\geq h} \mathbb{V}[r^{h'} + f_{t,2}^{h'+1}(x^{h'+1})|z_t^{h'}] \log\frac{2}{\delta}} + 2\log\frac{2}{\delta}$$
$$\leq \sqrt{4TH\log\frac{2}{\delta}} + 2\log\frac{2}{\delta}, \tag{59}$$

we thus have event $\mathcal{E}_{\xi_2}$ *happens with probability at least* $1-\delta$*.*

- *For* $\{\xi_{t,-2}\}_{t,h}$*, we let* $\mathcal{E}_{\xi_{-2}}$ *be the event such that*

$$\left|\sum_{t\in[T],h'\geq h} \xi_{t,-2}^{h'}\right| \leq \sqrt{\sum_{t\in[T],h'\geq h} \mathbb{V}[r^{h'} + f_{t,-2}^{h'+1}(x^{h'+1})|z_t^{h'}] \log\frac{2H}{\delta}} + 2\log\frac{2H}{\delta}$$
$$\leq \sqrt{4TH\log\frac{2H}{\delta}} + 2\log\frac{2H}{\delta} \text{ for all } h\in[H], \tag{60}$$

we thus have event $\mathcal{E}_{\xi_{-2}}$ *happens with probability at least* $1-\delta$*.*

The proof of this lemma is an immdiate application of Lemma 16.

G.2. Size of $|\mathcal{T}_{\rm oo}|$

We consider the following lemma due to martingale concentration, which will be useful to give a with high probability argument for bounding the size of $\mathcal{T}_{\rm oo}$.

Lemma 47 (Concentration with indicators) *Let* $\mathsf{D}_t^h = \left(\xi_t^h - \xi_{t,2}^h\right)\mathbf{1}_{\{h\geq h_t\}}$ *for any* $t\in[T]$*,* $h\in[H]$*. We have* D_t^h *is a martingale difference sequence and with probability* $1-\delta$*,*

$$\sum_{t\in[T],h\in[H]} \left(\xi_t^h - \xi_{t,2}^h\right)\mathbf{1}_{\{h\geq h_t\}} \leq O\left(\sqrt{|\mathcal{T}_{\rm oo}|H\log\frac{TH}{\delta}} + \log\frac{TH}{\delta}\right),$$
$$\text{and also} \sum_{t\in[T],h\in[H]} \left(\xi_{t,1}^h - \xi_{t,2}^h\right)\mathbf{1}_{\{h\geq h_t\}} \leq O\left(\sqrt{|\mathcal{T}_{\rm oo}|H\log\frac{TH}{\delta}} + \log\frac{TH}{\delta}\right).$$

We call this event $\mathcal{E}_{\xi_{\rm dif}}$*.*

Proof We first prove the first inequality. Recall $\xi_t^h = r_t^h + V_t^{h+1} - \mathbb{E}\left[r^h + V_t^{h+1} | z_t^{[h]}, f_{t,1}^{[H]}, f_{t,2}^{[H]}\right]$ and $\xi_{t,2}^h = r_t^h + f_{t,2}^{h+1}(x_t^{h+1}) - \mathbb{E}_{r^h, x^{h+1}}\left[r^h + f_{t,2}^{h+1}(x^{h+1}) | z_t^h\right]$. We also recall the filtration defined earlier as $\mathcal{H}_t^h = \sigma(x_1^1, r_1^1, x_1^2, \cdots, r_1^H, x_1^{H+1}; x_2^1, r_2^1, x_2^2, \cdots, r_2^H, x_2^{H+1}; \cdots, x_t^1, r_t^1, \cdots, r_t^h, x_t^{h+1})$ for any $t \in [T], h = 0, 1, \ldots, H$.

Thus following Lemma 37 we have

$$\mathbb{E}[\mathsf{D}_t^h | \mathcal{H}_t^{h-1}] = \mathbf{1}_{\{h \geq h_t\}} \cdot \mathbb{E}[\xi_t^h - \xi_{t,2}^h | \mathcal{H}_t^{h-1}] = 0,$$

which by definition shows that D_t^h as defined is a martingale difference sequence.

Further, applying Lemma 17 to $\{\mathsf{D}_t^h\}_{t \in [T]}^{h \in [H]}$, we have with probability at least $1 - \delta/2$, it holds that

$$\sum_{t \in [T], h \in [H]} \left(\xi_t^h - \xi_{t,2}^h\right) \mathbf{1}_{\{h \geq h_t\}} \leq O\left(\sqrt{|\mathcal{T}_{\mathrm{oo}}| H \log \frac{TH}{\delta}} + \log \frac{TH}{\delta}\right),$$

where we use the variance of ξ_t^h and $\xi_{t,2}^h$ are constants for each t, h.

Similarly we can show with probability at least $1 - \delta/2$ the second inequality holds true too. ■

Lemma 48 (Bounding size of $\mathcal{T}_{\mathrm{oo}}$) *Suppose $\alpha \leq 1$, we set*

$$u_t \geq C \cdot \left(\frac{\sqrt{\log \frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\log \frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H^{5/2}\sqrt{d_\alpha} + \sqrt{t} H \epsilon_{\mathrm{b}}\right)}{\sqrt{t}} + H^2 \epsilon\right),$$

for some large enough constant $C < \infty$ and $\epsilon \leq 1$, then we have the following facts about $\mathcal{T}_{\mathrm{oo}}$ holds true:

$$|\mathcal{T}_{\mathrm{oo}}| \leq O\left(\frac{T}{\log \frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H^3}\right) \text{ when } \mathcal{E}_{\leq T} \cap \mathcal{E}_{\xi_{\mathrm{dif}}} \text{ happens.}$$

Proof Similar to the proof of Lemma 41, when the events $\mathcal{E}_{\leq T}$ and also $\mathcal{E}_{\xi_{\mathrm{dif}}}$ by Lemma 47 happen, we have

$$\sum_{t \in \mathcal{T}_{\mathrm{oo}}} \left(f_{t,2}^{h_t}(x_t^{h_t}) - f_{t,1}^{h_t}(x_t^{h_t})\right)$$
$$\geq \frac{C}{4}\left(\sqrt{\log \frac{\mathcal{N}TH}{\alpha\delta} + \frac{T}{\alpha^2}\epsilon} \cdot \left(\log \frac{\mathcal{N}\mathcal{N}_b TH}{\alpha\delta} \cdot H^{5/2}\sqrt{d_\alpha}\frac{|\mathcal{T}_{\mathrm{oo}}|}{\sqrt{T}} + |\mathcal{T}_{\mathrm{oo}}| H \epsilon_{\mathrm{b}}\right) + |\mathcal{T}_{\mathrm{oo}}| H^2 \epsilon + |\mathcal{T}_{\mathrm{oo}}| H \delta\right)$$

and also

$$\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\left(f_{t,2}^{h_t}(x_t^{h_t})-f_{t,1}^{h_t}(x_t^{h_t})\right)\le\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\left(f_{t,2}^{h_t}(x_t^{h_t})-V_t^{h_t}\right)$$

$$\le O\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}}H\sqrt{|\mathcal{T}_{\mathrm{oo}}|\cdot d_\alpha}+\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}H\cdot d_\alpha+|\mathcal{T}_{\mathrm{oo}}|H\epsilon_{\mathrm{b}}\right)\right)$$

$$+\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\sum_{h_t\le h\le H}\left(\xi_t^h-\xi_{t,2}^h\right)+O(|\mathcal{T}_{\mathrm{oo}}|H^2\epsilon)$$

$$\le O\left(\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}}H\sqrt{|\mathcal{T}_{\mathrm{oo}}|\cdot d_\alpha}+\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}H\cdot d_\alpha+|\mathcal{T}_{\mathrm{oo}}|H\epsilon_{\mathrm{b}}\right)\right)$$

$$+O\left(\sqrt{|\mathcal{T}_{\mathrm{oo}}|H\log\frac{TH}{\delta}}+\log\frac{TH}{\delta}+|\mathcal{T}_{\mathrm{oo}}|H^2\epsilon\right).$$

Here in last inequality we use the bound in Lemma 47 conditioning on $\mathcal{E}_{\xi_{\mathrm{dif}}}$. For sufficiently large constant $C<\infty$, the above two inequalities hold true only when $|\mathcal{T}_{\mathrm{oo}}|\le O\left(T/(H^3\cdot\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta})\right)$. ■

Building on this bound of $|\mathcal{T}_{\mathrm{oo}}|$, we show the next corollary on a tighter bound for the summation terms in III.

Corollary 49 (Fine-grained bound on III) *Given $b_{t,2}(\cdot)\le C\cdot\big(D_{\mathcal{F}^h}(\cdot,z_{[t-1]}^h,\mathbf{1}_{[t-1]}^h)\sqrt{(\beta_{t,2}^h)^2+\lambda}+\epsilon_{\mathrm{b}}\cdot\beta_{t,2}^h\big)$ and using the particular choice of u_t as in Lemma 48, when $\lambda=\Theta(1)$, $\alpha\le 1$, we have when $\mathcal{E}_{\le T}$ and $\mathcal{E}_{\xi_{\mathrm{dif}}}$ happen,*

$$III:=\sum_{t\in\mathcal{T}_{\mathrm{oo}}}\sum_{h\in[H]}\min\left(1+L,b_{t,2}^h(z_t^h)\right)$$

$$=O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T\epsilon}\cdot\sqrt{T\cdot d_\alpha}+\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}+T\epsilon}\cdot(H\cdot d_\alpha+T\epsilon_{\mathrm{b}})\right).$$

Proof This is an immediate corollary by combining Lemma 38 and Lemma 48. ■

G.3. Sum of Variances

Further, we will provide a lemma showing that the concentration of the summation variance terms in the exploration happens with high probability, which we denote as event $\mathcal{E}_{\mathbb{V}}$. The result builds on law of total variance due to Proposition 20 and the in-expectation bounds due to Corollary 40. Recall the definition of $\mathcal{H}_{t-1}^H=\sigma(x_1^1,r_1^1,x_1^2,\cdots,r_1^H,x_1^{H+1};\cdots,x_{t-1}^1,r_{t-1}^1,\cdots,r_{t-1}^H,x_{t-1}^{H+1})$ and exploration rule (22).

Corollary 50 (Corollary from adapted version using LTV, high probability) *Recall the simplified expression of $\mathbb{V}[\cdot|z_t^h]=\mathbb{V}_{r^h,x^{h+1}}[\cdot|z_t^h]$. When $L=O(1)$ we have with probability at least*

$1-\delta$,

$$\sum_{t\in[T]}\sum_{h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]$$
$$\le O\left(H^4\log^2\frac{TH}{\delta}+T+TH^2\delta+H^2|\mathcal{T}_{\rm oo}|+H\cdot\sum_{t\in[T]}\sum_{h\in[H]}\left(f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right)\right).$$

We denote such event as $\mathcal{E}_{\mathbb{V}}$ hereinafter.

Proof For the high probability bounds, we consider applying Freedman's inequality in Corollary 18 to

$$\mathsf{D}_t=\left(\sum_{h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|x_t^h,a_t^h\right]-\mathbb{E}\left[\sum_{h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]\mid\mathcal{H}_{t-1}^H\right]\right),$$
$$\text{with }\ M=O(H),\ \ V^2=O(TH^2),$$
$$\text{and }\sum_{t\in[T]}\mathbb{E}\left[\mathsf{D}_t^2\mid\mathcal{H}_{t-1}^H\right]=\ H\cdot\sum_{t\in[T]}\mathbb{E}\left[\sum_{h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]\mid\mathcal{H}_{t-1}^H\right].$$

Thus Corollary 18 gives that with probability at least $1-\delta$, it holds that

$$\sum_{t\in[T],h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]$$
$$\overset{(i)}{\le}O\left(H\cdot\log(TH/\delta)+\sum_{t\in[T]}\mathbb{E}\left[\sum_{h\in[H]}\mathbb{V}\left[r^h+f_{t,1}^{h+1}(x^{h+1})|z_t^h\right]\mid\mathcal{H}_{t-1}^H\right]\right)$$
$$\overset{(ii)}{\le}O\left(H\log\frac{TH}{\delta}+T+TH^2\delta+H^2\cdot\sum_{t\in[T]}\mathbb{E}[\mathbf{1}_{\{t\in\mathcal{T}_{\rm oo}\}}\mid\mathcal{H}_{t-1}^H]\right)$$
$$+O\left(H\cdot\sum_{t\in[T]}\mathbb{E}\Big[\sum_{h\in[H]}\left(f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right)\mid\mathcal{H}_{t-1}^H\Big]\right)$$
$$\overset{(iii)}{\le}O\left((H+H^2\sqrt{T})\log\frac{TH}{\delta}+T+TH^2\delta+H^2\cdot|\mathcal{T}_{\rm oo}|+H\sum_{t\in[T]}\sum_{h\in[H]}\left(f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right)\right),$$

where for (i) we also use the AM-GM inequality, for (ii) we use Corollary 40, and for (iii) we use the Azuma-Hoeffding concentration of martingale. The claim follows by again applying the AM-GM inequality. ■

G.4. Difference of Overly Optimistic Sequence

Lemma 51 *When good events $\mathcal{E}_{\le T}$, $\mathcal{E}_{\xi_1}$, $\mathcal{E}_{\xi_{-2}}$ and $\mathcal{E}_{\xi_2}$ happen, and when $\lambda = \Theta(1)$, $\alpha \le 1$, $\epsilon \le 1$ we have*

$$\sum_{t\in[T]}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right]$$
$$\le O\left(H\cdot[I] + H\cdot[II] + H\sqrt{HT\log(H/\delta)} + H\log(H/\delta) + |\mathcal{T}_{\rm oo}|H^2 + TH^2\epsilon\right) + H\cdot\sum_{t\in\mathcal{T}_{\rm o}} u_t.$$

Proof Similar to the proof of Lemma 43, for $t\in\mathcal{T}_{\rm oo}$ it holds that

$$\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right] = O(|\mathcal{T}_{\rm oo}|H).$$

Otherwise, for iterations $t\in\mathcal{T}_{\rm o}$, we also have

$$\sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right]$$
$$\le O\left(\sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\min\left(1, \sum_{h\le h'\le H} b_{t,1}^{h'}(z_t^{h'})\right) + \sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\min\left(1, \sum_{h\le h'\le H} b_{t,2}^{h'}(z_t^{h'})\right)\right)$$
$$+ \sum_{t\in\mathcal{T}_{\rm o}} H\cdot u_t + O\left(\sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'} - \sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'} + TH^2\epsilon\right),$$

given event $\mathcal{E}_{\le T}$ happens.

Summing over all iterations $t\in\mathcal{T}_{\rm o}$, this implies

$$\sum_{t\in\mathcal{T}_{\rm o}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right]$$
$$\overset{(i)}{\le} O\left(H\cdot I + H\cdot II + TH^2\epsilon\right) + \sum_{t\in\mathcal{T}_{\rm o}} H\cdot u_t$$
$$+ O\left(\left|\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}\right| + \left|\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}\right| + \left|\sum_{t\in[T]}\sum_{h\in[H]}\xi_{t,2}^{h}\right|\right)$$
$$+ O\left(\left|\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}\right| + \left|\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}\right| + \left|\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h\in[H]}\xi_{t,2}^{h}\right|\right)$$
$$\le O\left(H\cdot I + H\cdot II + TH^2\epsilon + |\mathcal{T}_{\rm oo}|H^2\right) + \sum_{t\in\mathcal{T}_{\rm o}} H\cdot u_t$$
$$+ O\left(\left|\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,-2}^{h'}\right| + \left|\sum_{t\in[T]}\sum_{h\in[H]}\sum_{h\le h'\le H}\xi_{t,1}^{h'}\right| + \left|\sum_{t\in[T]}\sum_{h\in[H]}\xi_{t,2}^{h}\right|\right).$$

Now since also events $\mathcal{E}_{\xi_1}$, $\mathcal{E}_{\xi_2}$ and $\mathcal{E}_{\xi_{-2}}$ happen, plugging in bounds in (58), (59) and (60) we have

$$\sum_{t\in\mathcal{T}_\mathrm{o}}\sum_{h\in[H]}\left[f_{t,2}^h(z_t^h)-f_{t,-2}^h(z_t^h)\right]$$
$$\leq O\left(H\cdot[I]+H\cdot[II]+H\sqrt{HT\log(H/\delta)}+H\log(H/\delta)+|\mathcal{T}_\mathrm{oo}|H^2+TH^2\epsilon\right)+H\cdot\sum_{t\in\mathcal{T}_\mathrm{o}}u_t.$$

Summing the two cases gives the claimed bound. ■

G.5. Bounding the Regret in High Probability

Now we will continue working on bounding $I=\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1+L,b_{t,1}^h(z_t^h)\right)$.

Lemma 52 (Fine-grained bound on I) *Recall the definition of $b_{t,1}$ and $b_{t,2}$ as in Lemma 39 and Corollary 49. When $\lambda=1$, $\alpha=1/\sqrt{TH}$, $\epsilon\leq 1$, when event $\mathcal{E}=\mathcal{E}_{\leq T}\cap\mathcal{E}_{\xi_1}\cap\mathcal{E}_{\xi_{-2}}\cap\mathcal{E}_{\xi_2}\cap\mathcal{E}_{\xi_\mathrm{dif}}\cap\mathcal{E}_{\mathbb{V}}$ happens, we have the following inequality holds true:*

$$I:=\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1+L,b_{t,1}^h(z_t^h)\right)$$
$$=O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\sqrt{T+TH^2\delta}\cdot\sqrt{H\cdot d_\alpha}\right)$$
$$+O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^3|\mathcal{T}_\mathrm{oo}|+TH^3\epsilon+H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\right)$$
$$+O\left(\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\log^{1.5}\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^{7/2}d_\alpha+\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot TH\epsilon_\mathrm{b}\right).$$

Proof Again we note that by assumption and definition,

$$\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1+L,b_{t,1}^h(z_t^h)\right)$$
$$=O\left(\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1,D_{\mathcal{F}^h}(z_t^h;z_{[t-1]}^h,\bar{\sigma}_{[t-1]}^h)\cdot\sqrt{\left(\beta_{t,1}^h\right)^2+\lambda}\right)+TH\epsilon_\mathrm{b}\cdot\max_{t,h}\beta_{t,1}^h\right)$$
$$=O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\left(\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1,D_{\mathcal{F}^h}(z_t^h;z_{[t-1]}^h,\bar{\sigma}_{[t-1]}^h)\right)+TH\epsilon_\mathrm{b}\right)\right).$$

Treating $L=O(1)$ as defined (see Assumption 1), we now bound the summation terms

$$\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1,D_{\mathcal{F}^h}(z_t^h;z_{[t-1]}^h,\bar{\sigma}_{[t-1]}^h)\right)=\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1,\bar{\sigma}_t^h\cdot\left(\bar{\sigma}_t^h\right)^{-1}D_{\mathcal{F}^h}(z_t^h;z_{[t-1]}^h,\bar{\sigma}_{[t-1]}^h)\right)$$

by dividing into cases, same as we do in Lemma 39 and Lemma 44.

$$\mathcal{I}_1 = \{(t,h) \in \mathcal{T} \times [H] \mid \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h) \geq 1\},$$
$$\mathcal{I}_2 = \{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}_t^h = \alpha\},$$
$$\mathcal{I}_3 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}_t^h = 2\big(\sqrt{\upsilon(\delta_{t,h})} + \iota(\delta_{t,h})\big) \cdot \sqrt{D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)}\right\},$$
$$\mathcal{I}_4 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}_t^h = \sigma_t^h\right\},$$
$$\mathcal{I}_5 = \left\{(t,h) \in \mathcal{T} \times [H] \mid (t,h) \notin \mathcal{I}_1,\ \bar{\sigma}_t^h = \sqrt{2}\iota(\delta_{t,h})\sqrt{f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)}\right\}.$$

Now same as in Lemma 39 and Lemma 44 we could bound the first three terms as

$$\sum_{(t,h)\in\mathcal{I}_1\cup\mathcal{I}_2\cup\mathcal{I}_3} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right) \leq O\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} \cdot H d_\alpha\right).$$

For terms restricting on $\mathcal{I}_4$ and $\mathcal{I}_5$ we do a similar tighter analysis in correspondence to Lemma 44. For summation terms in $\mathcal{I}_5$, recall Equation (53) already shows that

$$\sum_{(t,h)\in\mathcal{I}_5} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)$$
$$\leq O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\sqrt{\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)} \cdot \sqrt{H \cdot d_\alpha}\right).$$

Restricting on $\mathcal{I}_4$, when event $\mathcal{E} = \mathcal{E}_{\leq T} \cap \mathcal{E}_{\xi_1} \cap \mathcal{E}_{\xi_{-2}} \cap \mathcal{E}_{\xi_2} \cap \mathcal{E}_{\xi_{\mathrm{dif}}} \cap \mathcal{E}_{\mathbb{V}}$ happens, by using Cauchy-Schwarz inequality, Lemma 34, Corollary 50, and AM-GM inequality, similar to the in-expectation proof we have

$$\sum_{(t,h)\in\mathcal{I}_4} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)$$
$$\leq O\left(\sqrt{T + H^2|\mathcal{T}_{\mathrm{oo}}| + TH^2(\epsilon+\delta)} \cdot \sqrt{H \cdot d_\alpha}\right)$$
$$+ O\left(\sqrt{H \cdot \sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right) + H^2 d_\alpha \cdot \left(\log^2\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right)} \cdot \sqrt{H \cdot d_\alpha}\right).$$

Summing all cases together we have

$$\sum_{t\in[T]}\sum_{h\in[H]} \min\left(1, \bar{\sigma}_t^h \cdot \left(\bar{\sigma}_t^h\right)^{-1} D_{\mathcal{F}^h}(z_t^h; z_{[t-1]}^h, \bar{\sigma}_{[t-1]}^h)\right)$$
$$\leq O\left(\sqrt{T + H^2|\mathcal{T}_{\mathrm{oo}}| + TH^2(\epsilon+\delta)} \cdot \sqrt{H \cdot d_\alpha}\right)$$
$$+ O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta}}\sqrt{H\sum_{t,h}\left(f_{t,2}^h(z_t^h) - f_{t,-2}^h(z_t^h)\right)} \cdot \sqrt{H \cdot d_\alpha}\right)$$
$$+ O\left(\left(\log\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} + T\epsilon\right) \cdot H^{1.5} \cdot d_\alpha\right).$$

Now plugging in the bound on $\sum_{t,h}\left(f^h_{t,2}(z^h_t) - f^h_{t,-2}(z^h_t)\right)$ in Lemma 51, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{h\in[H]} \min\left(1, \bar{\sigma}^h_t \cdot \left(\bar{\sigma}^h_t\right)^{-1} D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})\right)\\
&\quad \le O\left(\sqrt{T+TH^2\delta}\cdot\sqrt{H\cdot d_\alpha}\right)\\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^3|\mathcal{T}_{\text{oo}}| + TH^3\epsilon + H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\right)\\
&\qquad + O\left(\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta} + T\epsilon\right)\cdot H^{1.5}\cdot d_\alpha\right)\\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^2\cdot[I] + H^2\cdot[II] + H^2\sqrt{HT\log\frac{H}{\delta}}}\cdot\sqrt{H\cdot d_\alpha}\right).
\end{aligned}
$$

To bound the last term in the RHS of inequality above, we plug in crude bounds in Lemma 38 and Lemma 39, note the crude bound of I dominates that of II and absorbing the low-order terms we have

$$
\begin{aligned}
&H^2\cdot[I] + H^2\cdot[II] + H^2\sqrt{HT\log\frac{H}{\delta}}\\
&\quad \le O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\cdot H^3\sqrt{Td_\alpha}\right)\\
&\qquad\quad + O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}\cdot H^2d_\alpha + TH^2\epsilon_{\text{b}}\right)\right)\\
&\quad \le O\left(\frac{T}{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}} + \left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\right)^2 H^6d_\alpha\right)\\
&\qquad\quad + O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon}\cdot TH^2\epsilon_{\text{b}}\right).
\end{aligned}
$$

Plugging this back, rearranging terms, absorbing lower terms and again use AM-GM inequality we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{h\in[H]} \min\left(1, D_{\mathcal{F}^h}(z^h_t; z^h_{[t-1]}, \bar{\sigma}^h_{[t-1]})\right)\\
&\quad \le O\left(\sqrt{T+TH^2\delta}\cdot\sqrt{H\cdot d_\alpha}\right)\\
&\qquad + O\left(\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^3|\mathcal{T}_{\text{oo}}| + TH^3\epsilon + H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\right)\\
&\qquad + O\left(\sqrt{\left(\log\frac{\mathcal{N}TH}{\delta} + T^2H\epsilon\right)\log^{1.5}\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^{7/2}d_\alpha + TH\epsilon_{\text{b}}}\right).
\end{aligned}
$$

■

With these bounds we are ready to prove formally the regret bounds for Algorithm 1.

Theorem 45 (Regret bound with high probability) *Suppose function class* $\{\mathcal{F}^h\}_{h\in[H]}$ *satisfy Assumption 1 with* $\epsilon \in [0,1]$ *and Definition 2 with* $\lambda = 1$*, given consistent bonus oracle* $\mathcal{B}$ *satisfying Definition 5, Algorithm 1 with* $\alpha = \sqrt{1/TH}$*,* $\delta \le 1/(H^2+11)$*,* $\epsilon \le 1$ *and*

$$u_t = C\cdot\left(\frac{\sqrt{\log\frac{\mathcal{N}TH}{\alpha\delta}+\frac{T}{\alpha^2}\epsilon}\cdot\left(\log\frac{\mathcal{N}\mathcal{N}_bTH}{\alpha\delta}\cdot H^{5/2}\sqrt{d_\alpha}+\sqrt{t}H^2\epsilon_{\rm b}\right)}{\sqrt{t}}+H^2\epsilon\right)$$

for sufficiently large constant $C<\infty$*, with high probability* $1-\delta$ *event* $\mathcal{E} = \mathcal{E}_{\le T}\cap\mathcal{E}_{\xi_1}\cap\mathcal{E}_{\xi_{-2}}\cap\mathcal{E}_{\xi_2}\cap\mathcal{E}_{\xi_{\rm dif}}\cap\mathcal{E}_{\mathbb{V}}$ *happens. Further, when conditioning on* $\mathcal{E}$ *the algorithm achieves a total regret* R_T *of*

$$O\left(\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\sqrt{T}\cdot\sqrt{Hd_\alpha}+\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\cdot\left(\log^2\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}H^5d_\alpha+T^2\epsilon_{\rm b}^2\right)\right).$$

Proof When event $\mathcal{E} = \mathcal{E}_{\le T}\cap\mathcal{E}_{\xi}\cap\mathcal{E}_{\xi_1}\cap\mathcal{E}_{\xi_{-2}}\cap\mathcal{E}_{\xi_2}\cap\mathcal{E}_{\xi_{\rm dif}}\cap\mathcal{E}_{\mathbb{V}}$ happen (with probability $1-11\delta$), we recall the upper bound on regret as:

$$\begin{aligned}R_T &\le O(1+HT\epsilon)+\underbrace{2\sum_{t\in[T]}\sum_{h\in[H]}\min\left(1+L, b_{t,1}^h(z_t^h)\right)}_{I}+2\sum_{t\in\mathcal{T}_{\rm oo}}\min\left(1+L,\sum_{h_t\le h\le H}b_{t,2}^h(z_t^h)\right)\\ &\quad+\left[\sum_{t\in[T],h\in[H]}\xi_t^h-\sum_{t\in[T]}\sum_{h\in[H]}\xi_{t,1}+\sum_{t\in\mathcal{T}_{\rm oo}}\left(\sum_{h_t\le h\le H}\xi_{t,1}^h-\sum_{h_t\le h\le H}\xi_{t,2}^h\right)\right]\\ &\le O\left(1+HT\epsilon+I+III\right)+\left[\sum_{t\in[T],h\in[H]}\xi_t^h-\sum_{t\in[T]}\sum_{h\in[H]}\xi_{t,1}\right]+\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}\left(\xi_{t,1}^h-\xi_{t,2}^h\right).\end{aligned}$$

Now plugging in guarantees of Lemma 52 for bounding I, Corollary 49 for bounding III, Equation (57) for bounding $\sum_{t,h}\xi_t^h$, Equation (58) for bounding $\sum_{t,h}\xi_{t,1}^h$, Lemma 47 for bounding $\sum_{t\in\mathcal{T}_{\rm oo}}\sum_{h_t\le h\le H}\left(\xi_{t,1}^h-\xi_{t,2}^h\right)$, we have

$$\begin{aligned}R_T = O\Bigg(&TH\epsilon+\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\cdot\left(\sqrt{T+TH^2\delta}\cdot\sqrt{Hd_\alpha}\right)\\ &+\sqrt{\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon}\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}}\sqrt{H^3|\mathcal{T}_{\rm oo}|+TH^3\epsilon+H^2\sum_{t\in[T]}u_t}\cdot\sqrt{H\cdot d_\alpha}\\ &+\left(\log\frac{\mathcal{N}TH}{\delta}+T^2H\epsilon\right)\cdot\log^{1.5}\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}\cdot H^{7/2}\cdot d_\alpha+\sqrt{\log\frac{\mathcal{N}\mathcal{N}_bTH}{\delta}+T^2H\epsilon}\cdot TH\epsilon_{\rm b}\Bigg).\end{aligned}$$

Now further plugging in the assumption of δ, choice of u_t and the bound of $\mathcal{T}_{\text{oo}}$ under such choice as in Lemma 48, we have

$$R_T = O\left(1 + \sqrt{\log\frac{\mathcal{N}TH}{\delta} + T^2 H\epsilon} \cdot \sqrt{T} \cdot \sqrt{Hd_\alpha}\right) + O\left(\left(\log\frac{\mathcal{N}TH}{\delta} + T^2 H\epsilon\right) \cdot \left(\log^2\frac{\mathcal{N}\mathcal{N}_b TH}{\delta} H^5 d_\alpha + T^2 \epsilon_{\text{b}}^2\right)\right).$$

Adjusting the constant $\delta \leftarrow \frac{\delta}{11}$ and absorbing the low-order terms of $\text{poly}(\epsilon)$ conclude the final bound for regret. ■