

MULTI-LEVEL REGRESSION FOR NONLINEAR CONTEXTUAL BANDITS AND RL: SECOND-ORDER AND HORIZON-FREE REGRET BOUNDS

006 **Anonymous authors**

007 Paper under double-blind review

ABSTRACT

013 Recent works have established second-order regret bounds for nonlinear context-
 014 tual bandits. However, these results exhibit a suboptimal dependence on the com-
 015 plexity of the function class. To close this gap, we propose a novel algorithm fea-
 016 turing a multi-level regression structure. This method partitions data by their un-
 017 certainty and variance, then performs separate regressions on each level, enabling
 018 adaptive, instance-dependent learning. Our method achieves a tight second-order
 019 regret bound of $\tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}}\right)$, which matches the
 020 theoretical lower bound. Here, $d_{\mathcal{F}}$ and $\log N_{\mathcal{F}}$ represent the Eluder dimension and
 021 log-covering number of the reward function class \mathcal{F} , σ_t^2 is the unknown variance
 022 of the reward at round t , and R is the range of rewards. The proposed algorithm
 023 is computationally efficient assuming access to a regression oracle. We further
 024 extend our framework to model-based reinforcement learning, achieving a regret
 025 bound that is both second-order and horizon-free. The underlying multi-level re-
 026 gression technique is of independent interest and applicable to a broad range of
 027 online decision-making problems.

1 INTRODUCTION

031 In the realm of online decision-making problems, contextual bandits serve as a foundational model,
 032 where an agent interacts with the environment to learn and act optimally in the face of uncertainty.
 033 This paradigm is central to numerous real-world applications, including personalized recom-
 034 mendation systems (Li et al., 2010; Covington et al., 2016), dynamic pricing (Kleinberg & Leighton,
 035 2003; Ferreira et al., 2018), and online advertising (Agarwal et al., 2014; Chapelle et al., 2014).
 036 A central goal in this field is to design algorithms with strong performance guarantees measured
 037 by regret—the difference in rewards between the algorithm’s choices and those of an optimal pol-
 038 icy. Although worst-case regret bounds have been well studied (Auer et al., 2002; Abbasi-Yadkori
 039 et al., 2011), the field has increasingly focused on developing more nuanced, instance-dependent
 040 guarantees (Zhou & Gu, 2022; Li & Sun, 2024; Huang et al., 2023). Second-order regret bounds,
 041 which incorporate the *unknown* variance of the rewards, are particularly valuable as they adapt to
 042 the problem’s intrinsic statistical difficulty rather than relying on pessimistic worst-case guarantees.

043 Despite significant progress in linear contextual bandits (Zhao et al., 2023), a fundamental chal-
 044 lenge has persisted in the setting of general function approximation, which is critical for capturing
 045 the complex relationships present in real-world scenarios. Current algorithms often suffer from a
 046 suboptimal dependence on the complexity of the reward function class, such as the Eluder dimen-
 047 sion $d_{\mathcal{F}}$. For instance, the best-known algorithms from Pacchiano (2025); Jia et al. (2024) achieve
 048 regret bounds of the form $\tilde{O}\left(d_{\mathcal{F}} \sqrt{\log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}}\right)$, which falls short of the theo-
 049 retical lower bound by Jia et al. (2024) that suggests a $\sqrt{d_{\mathcal{F}}}$ dependency. While Wang et al. (2024b)
 050 achieve a $O(\sqrt{d_{\mathcal{P}}})$ regret bound, their algorithm requires a stronger realizability assumption: access
 051 to the full reward distribution. This discrepancy raises a crucial open question:

052 *Can we design an algorithm for nonlinear contextual bandits that achieves a minimax-optimal,
 053 second-order regret with the standard realizability assumption?*

054
055 Table 1: Regret bounds of algorithms for contextual bandits with *unknown* reward variances. Here,
056 d denotes the dimension for linear function approximation, \mathcal{P} represents the reward distribution
057 class, \mathcal{F} is the the reward function class, $d_{\mathcal{F}}, d_{\mathcal{P}}$ are the Eluder dimension, $N_{\mathcal{F}}, N_{\mathcal{P}}$ are the covering
058 number, T is the number of rounds, σ_t is the variance of the reward at round t , and R is the range of
059 rewards. \tilde{O} omits logarithmic terms.

| Algorithm | Function Type | Regret Bound | Computational Efficiency |
|--|---------------|--|--------------------------|
| SAVE (Zhao et al., 2023) | Linear | $\tilde{O}\left(d\sqrt{\sum_{t \in [T]} \sigma_t^2} + Rd\right)$ | Yes |
| DistUCB (Wang et al., 2024b) | Nonlinear | $\tilde{O}\left(\sqrt{d_{\mathcal{P}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + Rd_{\mathcal{P}} \log N_{\mathcal{P}}\right)$ | Yes |
| Unknown-Variance SOOLS (Pacchiano, 2025) | Nonlinear | $\tilde{O}\left(d_{\mathcal{F}} \sqrt{\log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + Rd_{\mathcal{F}} \log N_{\mathcal{F}}\right)$ | Yes |
| VarUCB (Jia et al., 2024) | Nonlinear | $\tilde{O}\left(d_{\mathcal{F}} \sqrt{\log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + d_{\mathcal{F}} (\log N_{\mathcal{F}})^{3/4}\right)$ | No |
| VACB (Ye et al., 2025) | Nonlinear | $\tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + Rd_{\mathcal{F}} \log N_{\mathcal{F}}\right)$ | Yes |
| UCB-MLR (Theorem 4.2) | Nonlinear | $\tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + Rd_{\mathcal{F}} \log N_{\mathcal{F}}\right)$ | Yes |

072
073 We give an affirmative answer to this question by delving into the problem of nonlinear contextual
074 bandits. Specifically, we consider the setting with *heteroscedastic* noise—where the variance of
075 rewards changes over time—and, critically, we assume this variance is *unknown* to the agent, a
076 common scenario in real-world applications. Our contributions are summarized as follows:

- 077 • We propose a novel Multi-Level Regression (MLR) structure, which significantly advances
078 prior multi-layer algorithms inspired from Zhao et al. (2023). A key innovation lies in
079 our data partitioning method, ADALEVEL, which leverages both uncertainty and variance
080 rather than just uncertainty. By running separate regressions on each level, our algorithm
081 learns in an adaptive and instance-dependent way, leading to a more accurate function
082 estimate. The principles of this multi-level regression technique are broadly applicable and
083 may be of independent interest for other online decision-making problems.
- 084 • Leveraging our new technique, we propose UCB-MLR, a novel algorithm for nonlinear
085 contextual bandits. Through the use of a tighter Bernstein-style bound for nonlinear re-
086 gression and a detailed analysis of estimation error at different levels, we theoretically
087 establish a regret bound of $\tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + Rd_{\mathcal{F}} \log N_{\mathcal{F}}\right)$. This result is sig-
088 nificant because it is the first to match the second-order lower bound from Jia et al. (2024),
089 effectively resolving a suboptimal dependency on $d_{\mathcal{F}}$. Our algorithm also achieves compu-
090 tational efficiency with access to a regression oracle.
- 091 • We further demonstrate the effectiveness and generality of our algorithmic framework by
092 applying it to model-based Reinforcement Learning (RL), where an agent learns to act
093 optimally by building a model of the environment. Our proposed algorithm, ML-VTR,
094 is the first to achieve a regret bound of $\tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + d_{\mathcal{F}} \log N_{\mathcal{F}}\right)$ for Markov
095 Decision Processes (MDPs) with general function approximation. This result is notable
096 because it is simultaneously second-order, horizon-free, and computationally efficient. As
097 a special case, it reduces to $\tilde{O}(d\sqrt{\text{Var}_K^*} + d^2)$ for linear mixture MDPs. This bound
098 matches the state-of-the-art from Zhao et al. (2023), suggesting that our novel algorithm
099 and fine-grained analysis are effective for a wide range of general RL problems.

100 For a comprehensive comparison with state-of-the-art results, we summarize the regrets in Table 1
101 for contextual bandits and Table 2 for RL.

102 **Notations** Let $[n] := \{1, 2, \dots, n\}$, $\overline{[n]} := \{0, 1, \dots, n\}$, and $X_{\mathcal{I}} := \{X_i\}_{i \in \mathcal{I}}$. Denote
103 $\min_{x \in \mathcal{X}} \{c, f(x)\} := \min\{c, \min_{x \in \mathcal{X}} f(x)\}$ for short. Denote the ϵ -covering number of \mathcal{F} w.r.t.
104 ℓ_{∞} -norm as $N_{\mathcal{F}}(\epsilon)$. $\tilde{O}(\cdot)$ omits logarithmic terms in $O(\cdot)$.

108
109 Table 2: Regret bounds of algorithms for model-based RL that achieve *instance-dependent* and
110 *horizon-free*. Here, d denotes the dimension for linear function approximation, \mathcal{P} represents the
111 transition model class, \mathcal{F} is the the function class induced by \mathcal{P} , $d_{\mathcal{F}}, d_{\mathcal{P}}$ are the Eluder dimension,
112 $N_{\mathcal{F}}, N_{\mathcal{P}}$ are the covering number, quantity Var_K^* defined in (5.3) is the total variance of the optimal
113 value functions, and \mathcal{Q}^* is a higher-order moments quantity defined in Huang et al. (2024). \tilde{O} omits
114 logarithmic terms.

| Algorithm | Function Type | Regret Bound | Computational Efficiency |
|--------------------------------|---------------|--|--------------------------|
| UCRL-AVE (Zhao et al., 2023) | Linear | $\tilde{O}(d\sqrt{\text{Var}_K^*} + d^2)$ | Yes |
| UCRL-WVTR (Huang et al., 2024) | Nonlinear | $\tilde{O}(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \mathcal{Q}^*} + d_{\mathcal{F}} \log N_{\mathcal{F}})$ | Yes |
| O-MBRL (Wang et al., 2025) | Nonlinear | $\tilde{O}(\sqrt{d_{\mathcal{P}} \log N_{\mathcal{P}} \text{Var}_K^*} + d_{\mathcal{P}} \log N_{\mathcal{P}})$ | No |
| ML-VTR (Theorem 5.1) | Nonlinear | $\tilde{O}(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + d_{\mathcal{F}} \log N_{\mathcal{F}})$ | Yes |

2 RELATED WORK

126 **Second-Order Regret in Nonlinear Contextual Bandits** Designing algorithms with second-
127 order regret has become a central theme in contextual bandits literature. While the linear setting
128 is well-understood (Zhao et al., 2023), the nonlinear setting with *unknown* variances presents sub-
129 stantially greater challenges, revealing a distinct gap to statistical optimality.

130 Several attempts, such as Unknown-Variance SOOLS (Pacchiano, 2025) and VarUCB (Jia et al.,
131 2024), have been made to generalize the multi-layer technique developed in (Zhao et al., 2023) to
132 nonlinear settings. Furthermore, VACB (Ye et al., 2025) utilizes Catoni estimator to handle the
133 heavy-tailedness of noise, removing the R dependence on the lower order. However, due to the
134 intrinsic difficulty caused by nonlinear structure, they only obtain a regret that is suboptimal on the
135 function complexity, thereby leaving a gap to optimality. A different line of work (Foster et al., 2018;
136 Wang et al., 2024b;a), exemplified by DistUCB (Wang et al., 2024b), pursue variance-adaptivity us-
137 ing MLE for the full reward distribution. However, this distributional approach requires the stronger
138 and often impractical modeling assumption that the entire reward distribution—not just the expected
139 reward—is realizable by the model class. Our multi-level regression framework, by contrast, oper-
140 ates under the standard, less restrictive realizability assumption.

141 **Instance-dependent and Horizon-free Regret in Model-based RL** The principles of instance-
142 dependent learning are also paramount in the more complex domain of RL, where the additional
143 challenges of long-planning horizons must be addressed. A key goal in modern RL theory is to de-
144 velop algorithms that are not only second-order but also horizon-free, meaning their regret bounds
145 scale at most polylogarithmically with the planning horizon H (Jiang & Agarwal, 2018). For MDPs
146 with linear function approximation, also known as linear mixture MDPs, Zhao et al. (2023) pro-
147 vide an efficient, second-order and horizon-free algorithm. However, extending these successes to
148 general function approximation presents significant challenges.

149 To name a few, Huang et al. (2024) made the first attempt to propose an algorithm, UCRL-
150 WVTR, using weighted value-targeted regression for estimating the model and achieves an instance-
151 dependent and horizon-free regret. Despite worst-case optimal when specialized to linear mixture
152 MDPs, their regret bound has a suboptimal dependence on the higher-order moments of the optimal
153 value functions. Conversely, O-MBRL (Wang et al., 2025) extends DistUCB to RL and achieves a
154 tight, second-order and horizon-free statistical guarantee. However, it is generally computationally
155 intractable and requires the stronger assumption of access to the full distribution.

3 PRELIMINARIES

160 **Nonlinear Contextual Bandits** We consider a T -round contextual bandit problem. At each round
161 $t \in [T]$, the environment provides a candidate decision set $\mathcal{X}_t \subseteq \mathcal{X}$. This framework includes the
162 classic contextual bandit setting given context z_t and action set \mathcal{A} , by setting $\mathcal{X}_t = \{z_t\} \times \mathcal{A}$. The

162 agent selects an action $x_t \in \mathcal{X}_t$ and receives a reward $y_t = f_*(x_t) + \varepsilon_t$. We assume $y_t \in [0, R]$,
 163

$$164 \mathbb{E}[\varepsilon_t|x_t] = 0, \quad \text{Var}[\varepsilon_t|x_t] = \text{Var}[y_t|x_t] = \sigma_t^2 \leq \sigma^2.$$

165 To enable the utilization of a priori unknown variance information, we make Assumption 3.2, which
 166 is also adopted by Ye et al. (2025).

167 **Assumption 3.1** (Realizability). We are given access to a function class \mathcal{F} such that $f_* \in \mathcal{F}$.

168 **Assumption 3.2.** We are given access to a function class \mathcal{G} and constant $c_v > 0$ such that $g_* \in \mathcal{G}$,
 169 and for all rounds $t \in [T]$,

$$171 \mathbb{E}[y_t^2|x_t] = g_*(x_t), \quad \text{Var}[y_t^2|x_t] \leq c_v^2 R^2 \cdot \text{Var}[y_t|x_t] = c_v^2 R^2 \sigma_t^2.$$

173 We use the standard Eluder dimension and covering number to measure the complexity of \mathcal{F} . Recall
 174 the definition of Eluder dimension (Russo & Van Roy, 2013):

175 **Definition 3.3** (Eluder Dimension). Let \mathcal{F} be a function class defined on \mathcal{X} and $\epsilon > 0$. The Eluder
 176 dimension $\dim_{\mathcal{F}}(\epsilon)$ of \mathcal{F} is the length of the longest sequence $x_{[n]} \subseteq \mathcal{X}$ such that for some $\epsilon' \geq \epsilon$,
 177 for all $t \leq n$, x_t is ϵ' -independent of $x_{[t-1]}$ given \mathcal{F} . That is, there exists $f, f' \in \mathcal{F}$ such that

$$179 \sum_{s \in [t-1]} [f(x_s) - f'(x_s)]^2 \leq \epsilon'^2 \text{ while } |f(x_t) - f'(x_t)| > \epsilon'.$$

182 We also use the notation $d_{\mathcal{F}} := \dim_{\mathcal{F}}(\epsilon)$ and $N_{\mathcal{F}} := N_{\mathcal{F}}(\epsilon)$ for short when ϵ is clear from the
 183 context. Let $\lambda > 0$. We quantify uncertainty of x given dataset $x_{[t-1]}$ and weights $w_{[t-1]}$ w.r.t. \mathcal{F} as:

$$185 D_{\mathcal{F}}(x; x_{[t-1]}, w_{[t-1]}) := \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(x) - f_2(x))^2}{\sum_{s \in [t-1]} w_s^2 (f_1(x_s) - f_2(x_s))^2 + \lambda}. \quad (3.1)$$

187 **MDPs with General Function Approximation** We consider episodic MDPs defined by a tuple
 188 $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \{r_h\}_{h \in [H]})$. Here, \mathcal{S} and \mathcal{A} are the state space and action spaces, H is the planning
 189 horizon, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the transition dynamics, $r_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the h -th step reward
 190 function known to the agents¹. We assume a bounded reward setting where $\sum_{h=1}^H r_h(s_h, a_h) \leq 1$
 191 for any trajectory. We use a deterministic policy throughout this paper, which is a collection of H
 192 mappings from the state space to the action space, denoted as $\pi = \{\pi_h : \mathcal{S} \rightarrow \mathcal{A}\}_{h \in [H]}$. For any
 193 state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the action value function $Q_h^\pi(s, a)$ and the (state) value function
 194 $V_h^\pi(s)$ are defined as:

$$196 Q_h^\pi(s, a) := \mathbb{E} \left[\sum_{h'=h}^H r(s_{h'}, a_{h'}) \middle| s_h = s, a_h = a \right], \quad V_h^\pi(s) := Q_h^\pi(s, \pi_h(s)),$$

199 where the expectation is taken w.r.t. the transition kernel \mathbb{P} and the agent's policy π . We denote the
 200 optimal value functions as $V_h^*(s) := \sup_{\pi} V_h^\pi(s)$ and $Q_h^*(s, a) := \sup_{\pi} Q_h^\pi(s, a)$. For simplicity,
 201 we introduce the following shorthands. Let \mathcal{V} be the set of all value functions $V : \mathcal{S} \rightarrow [0, 1]$. For
 202 any $V \in \mathcal{V}$, we denote the conditional expectation and variance of V as

$$203 [\mathbb{P}V](s, a) := \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, a)} [V(s')], \quad [\mathbb{V}V](s, a) := [\mathbb{P}V^2](s, a) - [\mathbb{P}V]^2(s, a)$$

205 Our objective is to design efficient algorithms that minimize the K -episode regret, defined as

$$207 \text{Regret}(K) := \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)).$$

209 To solve problems of large state spaces, we consider MDPs with general function approximation.
 210 We adopt the following assumptions to accurately estimate the variance of value functions, which is
 211 reasonable since a small variance of a next-state value function often indicates more deterministic
 212 transitions, thus suggesting a small variance for the squared next-state value function.

213 **Assumption 3.4** (Realizability). Let \mathcal{P} be a general function class consisting of transition kernels
 214 that map state-action pairs to measures over \mathcal{S} . We assume the MDP's transition model $\mathbb{P} \in \mathcal{P}$.

215 ¹We consider deterministic rewards since our result can be easily generalized to the unknown-reward cases.

216

Algorithm 1 UCB-MLR

217

Require: $\alpha, \tilde{\alpha}, \gamma, \tilde{\gamma}, L = \lceil \log_2 \frac{R}{\alpha} \rceil, \tilde{L} = \lceil \log_2 \frac{R^2}{\alpha} \rceil, \{\beta_{t,l}\}_{t \geq 1, l \in [L]}, \{\tilde{\beta}_{t,\ell}\}_{t \geq 1, \ell \in [\tilde{L}]}$
1: $\Psi_{1,l} \leftarrow \emptyset$ for $l \in [\tilde{L}], \tilde{\Psi}_{1,\ell} \leftarrow \emptyset$ for $\ell \in [\tilde{L}]$
2: $\hat{f}_{1,l} \leftarrow 0$ for $l \in [L], \hat{g}_{1,\ell} \leftarrow 0$ for $\ell \in [\tilde{L}]$
3: **for** $t = 1, \dots, T$ **do**
4: Observe \mathcal{X}_t
5: Choose $x_t \leftarrow \operatorname{argmax}_{x \in \mathcal{X}_t} \min_{l \in [L]} (\hat{f}_{t,l}(x) + \min\{R, \beta_{t,l} D_{t,l}(x)\})$, receive y_t
6: Update $\bar{\sigma}_t$ according to (4.3).
7: Set $l_t, w_t \leftarrow \text{ADALEVEL}(\{D_{t,l}(x_t)\}_{l \in [L]}, \bar{\sigma}_t, \alpha, \gamma)$
8: Set $\ell_t, \tilde{w}_t \leftarrow \text{ADALEVEL}(\{\tilde{D}_{t,\ell}(x_t)\}_{\ell \in [\tilde{L}]}, c_v \bar{\sigma}_t, \tilde{\alpha}, \tilde{\gamma})$
9: Update $\Psi_{t+1,l} \leftarrow \Psi_{t,l} \cup \{t\}, \Psi_{t+1,l} \leftarrow \Psi_{t,l}$ for $l \in [\tilde{L}], l \neq l_t$
10: Update $\tilde{\Psi}_{t+1,\ell} \leftarrow \tilde{\Psi}_{t,\ell} \cup \{t\}, \tilde{\Psi}_{t+1,\ell} \leftarrow \tilde{\Psi}_{t,\ell}$ for $\ell \in [\tilde{L}], \ell \neq \ell_t$
11: Update $\hat{f}_{t+1,l}$ for $l \in [L], \hat{g}_{t+1,\ell}$ for $\ell \in [\tilde{L}]$ according to (4.1), (4.2)
12: **end for**

232

Algorithm 2 ADALEVEL

233

Require: $\{D_{t,l}\}_{l \in [L]}, \bar{\sigma}_t, \alpha, \gamma$
Ensure: Level l_t , weight w_t
1: Set $l_t \leftarrow \max\{l \in [L] : \gamma D_{t,l} > 2^l \alpha\}$
2: **if** $l_t = -\infty$ **then**
3: Update $l_t \leftarrow 0$
4: **else**
5: **if** $\bar{\sigma}_t \leq 2^{l_t} \alpha$ **then**
6: Set $w_t \leftarrow \frac{2^{l_t} \alpha}{\gamma D_{t,l_t}}$
7: **else**
8: Update $l_t \leftarrow \min\{l \in [L], l > l_t : \bar{\sigma}_t \leq 2^l \alpha\}$
9: Set $w_t \leftarrow 1$
10: **end if**
11: **end if**

247

248

Assumption 3.5. There exists a constant $c_v > 0$ such that for all steps $(k, h) \in [K] \times [H]$ and all $V_{h+1} \in \mathcal{V}$, the following holds:

249

250

$$[\mathbb{V}V_{h+1}^2](s_h^k, a_h^k) \leq c_v^2 [\mathbb{V}V_{h+1}](s_h^k, a_h^k).$$

251

252

We use the covering number and Eluder dimension to measure the complexity of the function class \mathcal{F} , which is induced from the model class \mathcal{P} . \mathcal{F} is generally smaller than \mathcal{P} , since we only require the expectation instead of the distribution information.

253

254

$$\mathcal{F} := \{f : \mathcal{S} \times \mathcal{A} \times \mathcal{V} \rightarrow \mathbb{R} \mid \exists \mathbb{P} \in \mathcal{P}, f(s_h^k, a_h^k, V_{h+1}) = [\mathbb{P}V_{h+1}](s_h^k, a_h^k)\},$$

255

256

4 MULTI-LEVEL REGRESSION FOR CONTEXTUAL BANDITS

257

258

In this section, we propose a new algorithm for nonlinear contextual bandits, UCB-MLR, which is formally presented in Algorithm 1. We introduce the notation $D_{t,l}(x) := D_{\mathcal{F}}(x; x_{\Psi_{t,l}}, w_{\Psi_{t,l}})$ and $\tilde{D}_{t,\ell}(x) := D_{\mathcal{G}}(x; x_{\tilde{\Psi}_{t,\ell}}, w_{\tilde{\Psi}_{t,\ell}})$ for conciseness. We first outline the high-level idea, then analyze the computational complexity and regret bound.

259

260

4.1 ALGORITHM DESCRIPTION

261

262

263

264

265

266

267

268

269

UCB-MLR improves upon the multi-layer structure proposed by Zhao et al. (2023). Their approach partitions data into $L + 1$ layers based on uncertainty, performs regressions within each layer $l \in [L]$, and combine L results to form a more accurate estimate of the reward function. In contrast, our

leveling algorithm, ADALEVEL, partitions date using both uncertainty and variance. We highlight the primary enhancements of UCB-MLR in as follows:

Adaptive Leveling In Line 7 of Algorithm 1, ADALEVEL adaptively chooses the level l_t for each data point x_t at round t , as detailed in Algorithm 2. This selection, based on its uncertainty within each level $\{D_{t,l}(x_t)\}_{l \in [L]}$ and the estimated variance $\bar{\sigma}_t^2$, leverages the concentration inequality in Lemma 4.3 to reduce the estimation error of reward function f_* .

We use $\Psi_{t+1,l}$ to denote the index set of all date partitioned into level $l \in \overline{[L]}$ up to time t . The detailed properties of ADALEVEL are listed in Property 1. In general, for all $t \in [T]$ such that $t \in \Psi_{T+1,l}$ with $l \in [L]$, we set weight

$$w_t = \min \left\{ 1, \frac{2^l \alpha}{\gamma D_{t,l}(x_t)} \right\}.$$

This is done to avoid a sharp change in uncertainty between adjacent levels. Consequently, we have:

$$w_t D_{t,l} \leq 2^l \alpha / \gamma, \quad w_t \bar{\sigma}_t \leq 2^l \alpha,$$

where α and γ are prespecified parameters. This ensures that the data at level l have roughly the same uncertainty and variance, both on the order of $2^l \alpha$. We use ADALEVEL similarly to construct $\{\tilde{\Psi}_{T,\ell}\}_{\ell \in \overline{[L]}}$ for estimating the squared-reward function g_* .

Multi-Level Regression and Upper Confidence Bound (UCB) At round t , after updating $\{\Psi_{t+1,l}\}_{l \in \overline{[L]}}$ and $\{\Psi_{t+1,\ell}\}_{\ell \in \overline{[L]}}$, we utilize weighted least squares regression to estimate f_* for level $l \in [L]$ and g_* for level $\ell \in \overline{[L]}$:

$$\hat{f}_{t+1,l} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s \in \Psi_{t+1,l}} w_s^2 (f(x_s) - y_s)^2, \quad (4.1)$$

$$\hat{g}_{t+1,\ell} = \operatorname{argmin}_{g \in \mathcal{F}} \sum_{s \in \tilde{\Psi}_{t+1,\ell}} \tilde{w}_s (g(x_s) - y_s^2)^2. \quad (4.2)$$

As shown in Line 5, for any $x \in \mathcal{X}$, we can construct L high-probability UCBs for $f_*(x)$ and take their minimum to choose the action optimistically:

$$f_*(x) \leq \min_{l \in [L]} (\hat{f}_{t+1,l}(x) + \min\{R, \beta_{t+1,l} D_{t+1,l}(x)\}),$$

Similarly, we can set $\bar{\sigma}_t^2$ as the upper bound of σ_t^2 :

$$\bar{\sigma}_t^2 = \min_{l \in [L], \ell \in \overline{[L]}} \{\sigma^2, \hat{g}_{t,\ell}(x_t) - \hat{f}_{t,l}(x_t) + R \min\{R, 2\beta_{t,l} D_{t,l}(x_t)\} + \min\{R^2, \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x_t)\}\}. \quad (4.3)$$

According to Lemma 4.3, $\beta_{t,l} = \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}})$ and $\tilde{\beta}_{t,\ell} = \tilde{O}(2^\ell \tilde{\alpha} \sqrt{\log N_{\mathcal{G}}})$.

Computational Complexity We analyze the computational complexity of UCB-MLR, relying on a regression oracle defined in Assumption 4.1 for solving the weighted nonlinear least squares regression. By adopting the techniques from Li et al. (2023); Huang et al. (2024), we can leverage this oracle to compute the uncertainty $\mathcal{D}_{\mathcal{F}}$ defined in (3.1) through a binary search procedure, requiring only $\tilde{O}(1)$ calls to the oracle.

Assumption 4.1 (Regression Oracle). We assume access to a weighted least squares *regression oracle*, which takes a function class \mathcal{F} and t weighted examples $\{(X_s, w_s, Y_s)\}_{s \in [t]} \subseteq \mathcal{X} \times \mathbb{R}^+ \times \mathbb{R}$ as input. It then outputs the solution to the weighted least squares problem, \hat{f} , within \mathcal{R} time, where

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s=1}^t w_s (f(X_s) - Y_s)^2.$$

We now consider the computation cost for a single round t . First, computing the UCB of $f_*(x_t)$ in Line 5 requires $\tilde{O}(L\mathcal{R})$ time, as it involves calculating $\mathcal{D}_{t,l}$ for $l \in [L]$. To select the best action over the set \mathcal{X}_t , the algorithm must compute the UCB for at most $|\mathcal{X}|$ actions. Next, the estimated variance $\bar{\sigma}_t^2$ in (4.3) can be computed within $\tilde{O}((L + \tilde{L})\mathcal{R})$ time. And ADALEVEL takes $\tilde{O}(L + \tilde{L})$ time. Finally, it takes $(L + \tilde{L})\mathcal{R}$ time to calculate the regression estimates $\hat{f}_{t+1,l}$ in (4.1) for $l \in [L]$ and $\hat{g}(t+1, \ell)$ in (4.2). Therefore, the total computational cost of UCB-MLR is $\tilde{O}(T|\mathcal{X}|\mathcal{R})$.

324 4.2 REGRET BOUND
325326 **Theorem 4.2.** For contextual bandit with general function approximation as defined in Section 3, if
327 the parameters in Algorithm 1 are set according to Section B, then with probability at least $1 - (L +$
328 $\tilde{L})\delta$, UCB-MLR achieves

329
$$\text{Regret}(T) = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + \max\{1, C\} R d_{\mathcal{F}} \log N_{\mathcal{F}}\right),$$

330

331 where $C = \max\{1, c_v\} \sqrt{\frac{d_{\mathcal{G}} \log N_{\mathcal{G}}}{d_{\mathcal{F}} \log N_{\mathcal{F}}}}$.
332

333 *Proof.* The proof uses a tighter Bernstein-style bound for the estimated function and a detailed
334 analysis of the summation of bonuses within each level. See Section 4.3 for a proof sketch and
335 Section B for a detailed proof. \square 336 Our result matches the second-order lower bound established by Jia et al. (2024), therefore success-
337 fully eliminating the gap related to $d_{\mathcal{F}}$. We leave the removal of C in the lower order term as an
338 open problem for future work.339 As a special case, for a d -dimensional linear contextual bandit, where $d_{\mathcal{F}}, \log N_{\mathcal{F}} = O(d)$ (Jia et al.,
340 2024), our algorithm achieves a regret of $\tilde{O}\left(d \sqrt{\sum_{t \in [T]} \sigma_t^2} + R d^2\right)$. This matches the state-of-the-
341 art result of Zhao et al. (2023) for the main term.342 4.3 PROOF SKETCH
343344 **Concentration of the Estimated Function** Our primary effort is to establish a tight UCB for the
345 true reward function f_* . This relies on the concentration inequality presented in Lemma 4.3.346 **Lemma 4.3.** Let $\{X_t\}_{t \geq 1} \subseteq \mathcal{X}$ and $\{Y_t\}_{t \geq 1} \subseteq [0, R]$ be sequences of random elements, and let
347 $\{w_t\}_{t \geq 1}$ be a sequence of weights. Let $f_* \in \mathcal{F}$ with function class $\mathcal{F} : \mathcal{X} \rightarrow [0, R]$. Suppose for all
348 $s \in [t]$, $\mathbb{E}[Y_s | X_s] = f_*(X_s)$, $|w_s| \leq W$, and $w_s^2 \text{Var}[Y_s | X_s] \leq \sigma^2$. Let the estimated function be

349
$$\hat{f}_{t+1} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s=1}^t w_s^2 (f(X_s) - Y_s)^2. \quad (4.4)$$

350

351 Then for any $\delta, \epsilon > 0$, with probability at least $1 - \delta$, we have for all $t \geq 1$,

352
$$\sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq \beta_{t+1}^2 \text{ with}$$

353

354
$$\beta_{t+1} = 3\sqrt{\iota_t \sigma} + 2\iota_t R \min\left\{1, \max_{s \in [t]} w_s^2 D_{\mathcal{F}}(X_s; X_{[s-1]}, w_{[s-1]})\right\} + \sqrt{\lambda} + \sqrt{6W^2 R t \epsilon},$$

355

356 where $\iota_t = 16 \log \frac{2N_{\mathcal{F}}(\epsilon)t^2(\log(\sigma^2 W^2 R^2 t) + 2)(\log(W^2 R^2) + 2)}{\delta} = \tilde{O}(\log N_{\mathcal{F}})$.
357

358 *Proof.* See Section A for a detailed proof. \square 359 **Remark 4.4.** Lemma 4.3 improves upon the Bernstein-style bound for nonlinear regression
360 from Huang et al. (2024) by tightening the term concerning uncertainty. Here, we denote
361 $D_t := \max_{s \in [t]} D_{\mathcal{F}}(X_s; X_{[s-1]}, w_{[s-1]})$ for short. This implies the confidence radius $\beta_{t+1} =$
362 $\tilde{O}(\sigma \sqrt{\log N_{\mathcal{F}}} + R D_t \log N_{\mathcal{F}})$. Compared to the bound $\tilde{O}(D_t \sqrt{\sum_{s \in [t]} \sigma_s^2 \log N_{\mathcal{F}}} + R D_t \log N_{\mathcal{F}})$
363 used in previous multi-layer algorithms (Pacchiano, 2025; Jia et al., 2024), our result improves the
364 first term by a factor of $\sqrt{d_{\mathcal{F}}/t}$ when the reward variances are roughly equal, since D_t is of order
365 $\sqrt{d_{\mathcal{F}}/t}$ under certain conditions according to Lemma E.4. This is a key step in removing the $\sqrt{d_{\mathcal{F}}}$
366 gap in regret bound.367 Recall that ADALEVEL ensures $w_t D_{t,l} \leq 2^l \alpha / \gamma$ and $w_t \bar{\sigma}_t \leq 2^l \alpha$ if $t \in \Psi_{T+1,l}$. This implies the
368 confidence radius $\beta_{t,l} = \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}})$.
369

378

Algorithm 3 ML-VTR

379 **Require:** $\alpha, \gamma, L = \lceil \log_2 \frac{1}{\alpha} \rceil$, confidence radius $\{\beta_{k,l}\}_{k \geq 1, l \in [L]}$

380 1: $\hat{f}_{1,l} \leftarrow 0$ for $l \in [L]$

381 2: $\Psi_{1,l}, \tilde{\Psi}_{1,l} \leftarrow \emptyset$ for $l \in [L]$, $\tilde{\Psi}_{1,\ell}, \tilde{\tilde{\Psi}}_{1,1,\ell} \leftarrow \emptyset$ for $\ell \in [L]$

382 3: **for** $k = 1, \dots, K$ **do**

383 4: $V_{k,H+1} \leftarrow 0$

384 5: **for** $h = H, \dots, 1$ **do**

385 6: $Q_{k,h}(\cdot, \cdot) \leftarrow \min_{l \in [L]} \{1, r_h(\cdot, \cdot) + \hat{f}_{k,l}(\cdot, \cdot, V_{k,h+1}) + \min\{1, \beta_{k,l} D_{k,l}(\cdot, \cdot, V_{k,h+1})\}\}$

386 7: $V_{k,h} \leftarrow \max_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$

387 8: $\pi_h^k \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$

388 9: **end for**

389 10: Receive s_1^k

390 11: **for** $h = 1, \dots, H$ **do**

391 12: Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive s_{h+1}^k

392 13: Update $z_{k,h} \leftarrow (s_h^k, a_h^k, V_{k,h+1})$, $\tilde{z}_{k,h} \leftarrow (s_h^k, a_h^k, V_{k,h+1}^2)$, $y_{k,h} \leftarrow V_{k,h+1}(s_{h+1}^k)$

393 14: Update $\bar{\sigma}_{k,h}^2$ according to (5.2)

394 15: Update $l_{k,h}, w_{k,h} \leftarrow \text{ADALEVEL}(\{D_{k,h,l}(z_{k,h})\}_{l \in [L]}, \bar{\sigma}_{k,h}, \alpha, \gamma)$

395 16: Update $\ell_{k,h}, \tilde{w}_{k,h} \leftarrow \text{ADALEVEL}(\{D_{k,h,l}(\tilde{z}_{k,h})\}_{l \in [L]}, c_v \bar{\sigma}_{k,h}, \alpha, \gamma)$

396 17: Update $\Psi_{k,h+1,l_{k,h}} \leftarrow \Psi_{k,h,l_{k,h}} \cup \{(k, h)\}$, $\Psi_{k,h+1,l} \leftarrow \Psi_{k,h,l}$ for $l \in [L], l \neq l_{k,h}$

397 18: Update $\tilde{\Psi}_{k,h+1,\ell_{k,h}} \leftarrow \tilde{\Psi}_{k,h,\ell_{k,h}} \cup \{(k, h)\}$, $\tilde{\Psi}_{k,h+1,\ell} \leftarrow \tilde{\Psi}_{k,h,\ell}$ for $\ell \in [L], \ell \neq \ell_{k,h}$

398 19: **end for**

400 20: Update $\Psi_{k+1,l}, \Psi_{k+1,1,l} \leftarrow \Psi_{k,H+1,l}$ for $l \in [L]$, $\tilde{\Psi}_{k+1,\ell}, \tilde{\Psi}_{k+1,1,\ell} \leftarrow \tilde{\Psi}_{k,H+1,\ell}$ for $\ell \in [L]$

401 21: Update $\hat{f}_{k+1,l}$ according to (5.1) for $l \in [L]$

402 22: **end for**

404
405 **Summation of Bonuses in Each Level** The regret can be related to the summation of bonuses
406 across each level, as follows:

407
$$\text{Regret}(T) \leq 2 \sum_{l \in [L]} \sum_{t \in \Psi_{T+1,l}} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(x_t)\}.$$

408

409 Thanks to ADALEVEL, the properties in Property 1 hold. Specifically, for any $l \in [L-1]$, if
410 $t \in \Psi_{T+1,l}$, the maximum over uncertainty $D_{t,l}(x_t)$ and estimated variance $\bar{\sigma}_t$ is of the order 2^l .
411 For high-uncertainty data, $\beta_{t,l+1} D_{t,l+1}(x_t) \approx 2^{2l}$, while Lemma E.4 implies $|\Psi_{T+1,l}| \approx 2^{-2l} d_{\mathcal{F}}$,
412 which leads to a lower order term in the final regret. For high-variance data, $\beta_{t,l} D_{t,l}(x_t) \approx$
413 $\bar{\sigma}_t D_{t,l}(x_t)$, and Lemma E.4 implies $D_{t,l}(x_t) \approx \sqrt{d_{\mathcal{F}}/|\Psi_{T+1,l}|}$, resulting a second-order term in the
414 final regret. We provide a more fine-grained analysis in Lemma E.3 to prove that for any $l \in [L-1]$,
415

416
$$\sum_{t \in \Psi_{T+1,l}} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(x_t)\} = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in \Psi_{T+1,l}} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}}\right).$$

417

418 The complete proof requires an in-depth analysis of the summation over different levels, and a
419 careful treatment of estimated variance to eliminate lower-order terms.

421 **5 MULTI-LEVEL REGRESSION FOR REINFORCEMENT LEARNING**

424 In this section, we extend our multi-level regression framework to MDPs with general function
425 approximation. This yields a new algorithm, ML-VTR, as detailed in Algorithm 3. We denote
426 $D_{k,l}(\cdot) := D_{\mathcal{F}}(\cdot; z_{\Psi_{k,l}} \cup \tilde{z}_{\tilde{\Psi}_{k,l}}, w_{\Psi_{k,l}} \cup \tilde{w}_{\tilde{\Psi}_{k,l}})$, $D_{k,h,l}(\cdot) := D_{\mathcal{F}}(\cdot; z_{\Psi_{k,h,l}} \cup \tilde{z}_{\tilde{\Psi}_{k,h,l}}, w_{\Psi_{k,h,l}} \cup \tilde{w}_{\tilde{\Psi}_{k,h,l}})$.
427 We first outline the high-level idea, then analyze the computational complexity and regret bound.

428
429 **5.1 ALGORITHM DESCRIPTION**

430 ML-VTR features a novel combination of the Multi-Level regression framework in Section 4
431 and Value-Targeted Regression (VTR) developed in Ayoub et al. (2020). Specifically, similar to

UCB-MLR, in Line 15, we leverage ADALEVEL to partition data into sets $\{\Psi_{K+1,l}\}_{l \in [L]}$ based on their uncertainty $\{D_{k,h,l}(z_{k,h})\}_{l \in [L]}$ and estimated variance $\bar{\sigma}_{k,h}$ for data points $z_{k,h}$. A similar process is applied to create the sets $\{\tilde{\Psi}_{K+1,l}\}_{l \in [L]}$ similarly for data points $\tilde{z}_{k,h}$. Here, $z_{i,h}$ and $\tilde{z}_{i,h}$ is defined in Line 13. Then we adopt Multi-Level VTR to estimate the model. Since all data share the same transition model f_* , we can estimate it in a combined manner to reduce error:

$$\hat{f}_{k+1,l} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{(i,h) \in \Psi_{k+1,l}} w_{i,h}^2 (f(z_{i,h}) - y_{i,h})^2 + \sum_{(i,h) \in \tilde{\Psi}_{k+1,l}} \tilde{w}_{i,h}^2 (f(\tilde{z}_{i,h}) - y_{i,h}^2)^2. \quad (5.1)$$

Once the estimate $\{\hat{f}_{k,l}\}_{l \in [L]}$ are obtained, we construct the action value functions $\{Q_{k,h}\}_{h \in [H]}$ as in Line 6. And the upper bound of $\operatorname{Var}[y_{k,h}|z_{k,h}] = [\mathbb{V}V_{k,h+1}](s_h^k, a_h^k)$ is then set as

$$\sigma_{k,h}^2 = \min_{l \in [L], \ell \in [L]} \{1, \hat{f}_{k,\ell}(\tilde{z}_{k,h}) - \hat{f}_{k,l}^2(z_{k,h}) + \min\{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\} + \min\{1, \beta_{k,\ell} D_{k,\ell}(\tilde{z}_{k,h})\}\}. \quad (5.2)$$

Computational Complexity We analyze the computational complexity of ML-VTR under the assumption that for any $(s, a, V) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V}$, the function $f_{\mathbb{P}}(s, a, V) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) V(s')$ can be evaluated in \mathcal{O} time. Recall \mathcal{R} represents the computational cost of the regression oracle. We consider the computation cost for a single step (k, h) . First, computing the action value function $Q_{k,h}$ in Line 6 for a given state-action pair (s, a) requires $\tilde{\mathcal{O}}(L(\mathcal{O} + \mathcal{R}))$ time, since it involves evaluating the estimated function $\hat{f}_{k,l}$ and computing the uncertainty $D_{k,l}$ for $l \in [L]$. To take an action based on π_h^k , the algorithm needs to compute $Q_{k,h}$ for $|\mathcal{A}|$ actions. Next, the estimated variance $\sigma_{k,h}^2$ in (5.2) can be computed within $\tilde{\mathcal{O}}(L(\mathcal{O} + \mathcal{R}))$ time. And ADALEVEL takes $\tilde{\mathcal{O}}(L)$ time. Finally, it takes $L\mathcal{R}$ time to calculate $\hat{f}_{k+1,l}$ in (5.1) for $l \in [L]$. Therefore, the total computational cost of ML-VTR is $\tilde{\mathcal{O}}(KH|\mathcal{A}|(\mathcal{O} + \mathcal{R}))$.

5.2 REGRET BOUND

Theorem 5.1. For MDP with general function approximation defined in Section 3, if the parameters in Algorithm 3 are set according to Section C, then with probability at least $1 - (L+2)\delta$, ML-VTR achieves

$$\operatorname{Regret}(K) = \tilde{\mathcal{O}}(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \operatorname{Var}_K^*} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}}),$$

where Var_K^* is the total variance of the optimal value functions $\{V_h^*\}_{h \in [H]}$:

$$\operatorname{Var}_K^* = \sum_{k=1}^K \sum_{h=1}^H [\mathbb{V}V_{h+1}^*](s_h^k, a_h^k). \quad (5.3)$$

Proof. The proof combines the technique used in proving contextual bandits with a fine-grained analysis of the higher-order moments of value functions, which eliminates polynomial dependence on the horizon H . See Section C for a detailed proof. \square

Our second-order result from Theorem 5.1 is also horizon-free, as its dependence on the horizon H is up to logarithmic factors. As a special case, for a d -dimensional linear mixture MDP, we have $d_{\mathcal{F}} \log N_{\mathcal{F}} = \mathcal{O}(d)$ (Huang et al., 2024). Our bound therefore simplifies to $\tilde{\mathcal{O}}(d\sqrt{\operatorname{Var}_K^*} + d^2)$, which matches the state-of-the-art result by Zhao et al. (2023). This demonstrates that our novel algorithm design and fine-grained analysis effectively and sharply handle general RL problems.

6 CONCLUSION

This paper presents a novel multi-level regression framework, ADALEVEL, that resolves a key challenge in online learning by partitioning data based on both uncertainty and variance. Our UCB-MLR algorithm for nonlinear contextual bandits, is the first to achieve an optimal second-order regret bound with computational efficiency. We extend this framework to reinforcement learning with general function approximation, where our ML-VTR algorithm provides the first horizon-free, second-order, and efficient regret bound. This multi-level regression technique is of independent interest and applicable to a broad range of online decision-making problems.

486 ETHICS STATEMENT
487488 The authors have read and adhere to the ICLR Code of Ethics.
489490 REPRODUCIBILITY STATEMENT
491492 For reproducibility, we have included all necessary details in the main body and appendix. The full
493 proofs for our theoretical results are in appendix.
494495 REFERENCES
496497 Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic
498 bandits. *Advances in neural information processing systems*, 24, 2011.
499500 Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming
501 the monster: A fast and simple algorithm for contextual bandits. In *International conference on
502 machine learning*, pp. 1638–1646. PMLR, 2014.503 Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with
504 nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*,
505 pp. 987–1063. PMLR, 2023.506 Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit
507 problem. *Machine learning*, 47(2):235–256, 2002.509 Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement
510 learning with value-targeted regression. In *International Conference on Machine Learning*, pp.
511 463–474. PMLR, 2020.513 Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for
514 display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–34,
515 2014.516 Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations.
517 In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.519 Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management
520 using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.521 Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical con-
522 textual bandits with regression oracles. In *International Conference on Machine Learning*, pp.
523 1539–1548. PMLR, 2018.525 Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforce-
526 ment learning with function approximation: Minimax optimal and instance-dependent regret
527 bounds. *Advances in Neural Information Processing Systems*, 36:56576–56588, 2023.528 Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Horizon-free and instance-dependent regret
529 bounds for reinforcement learning with general function approximation. In *International Confer-
530 ence on Artificial Intelligence and Statistics*, pp. 3673–3681. PMLR, 2024.532 Zeyu Jia, Jian Qian, Alexander Rakhlin, and Chen-Yu Wei. How does variance shape the regret
533 in contextual bandits? *Advances in Neural Information Processing Systems*, 37:83730–83785,
534 2024.535 Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds
536 on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398. PMLR, 2018.538 Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret
539 for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer
Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.

540 Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to
 541 personalized news article recommendation. In *Proceedings of the 19th international conference*
 542 *on World wide web*, pp. 661–670, 2010.

543

544 Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under
 545 heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

546 Yunfan Li, Yiran Wang, Yu Cheng, and Lin Yang. Low-switching policy gradient with exploration
 547 via online sensitivity sampling. In *International Conference on Machine Learning*, pp. 19995–
 548 20034. PMLR, 2023.

549

550 Aldo Pacchiano. Second order bounds for contextual bandits with function approximation. In *The*
 551 *Thirteenth International Conference on Learning Representations*, 2025.

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

554

555 Kaiwen Wang, Nathan Kallus, and Wen Sun. The central role of the loss function in reinforcement
 556 learning. *arXiv preprint arXiv:2409.12799*, 2024a.

557 Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being
 558 distributional: Second-order bounds for reinforcement learning. In *International Conference on*
 559 *Machine Learning*, pp. 51192–51213. PMLR, 2024b.

560

561 Zhiyong Wang, Dongruo Zhou, John C.S. Lui, and Wen Sun. Model-based RL as a minimalist
 562 approach to horizon-free and second-order bounds. In *The Thirteenth International Conference*
 563 *on Learning Representations*, 2025.

564 Chenlu Ye, Yujia Jin, Alekh Agarwal, and Tong Zhang. Catoni contextual bandits are robust to
 565 heavy-tailed rewards. In *Forty-second International Conference on Machine Learning*, 2025.

566

567 Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret
 568 bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In
 569 *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4977–5020. PMLR, 2023.

570 Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning
 571 for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349,
 572 2022.

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594 THE USE OF LARGE LANGUAGE MODELS (LLMs)
595596 LLM was used as a general-purpose writing assistant for tasks like grammar and spelling correction,
597 and to refine sentence structure. All authors take full responsibility for the final content.
598599 A PROOF OF LEMMA 4.3
600601 Proof of Lemma 4.3. We define filtration $\{\mathcal{G}_t\}_{t \geq 1}$ such that $X_t \in \mathcal{G}_{t-1}$, $Y_t \in \mathcal{G}_t$. Recall the definition
602 of \hat{f}_{t+1} in (4.4), which implies
603

604
$$\sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq 2 \sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s)) (Y_s - f_*(X_s)).$$

605

606 For any fixed $f \in \mathcal{F}$, denote $E_s(f) = w_s^2 (f(X_s) - f_*(X_s)) (Y_s - f_*(X_s))$, which is a martingale
607 difference sequence adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$. Note $|w_s| \leq W$ and $f(X_s), f_*(X_s), Y_s$ are
608 bounded in $[0, R]$, thereby the expectation and summation of variances are upper bounded by
609

610
$$|E_s(f)| \leq W^2 R^2, \quad \sum_{s=1}^t \mathbb{E}[E_s^2(f) | \mathcal{G}_{s-1}] \leq \sigma^2 W^2 R^2 t.$$

611

612 We denote $D_s = D_{\mathcal{F}}(X_s; X_{[s]}, w_{[s]})$ for short. Furthermore, we have
613

614
$$\max_{s \in [t]} |E_s(f)| \stackrel{(a)}{\leq} R \max_{s \in [t]} w_s^2 D_s \sqrt{\sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \lambda},$$

615
616
$$\sum_{s=1}^t \mathbb{E}[E_s^2(f) | \mathcal{G}_{s-1}] \leq \sigma^2 \sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2,$$

617

618 where (a) holds due to the definition of $D_{\mathcal{F}}$ in (3.1). Let $\epsilon > 0$ and \mathcal{V} be a ϵ -covering net of \mathcal{F} .
619 Applying Lemma E.1 with $m = v = \sigma^2$, $\iota_t = 16 \log \frac{2N_{\mathcal{F}}(\epsilon)t^2(\log(\sigma^2 W^2 R^2 t) + 2)(\log(W^2 R^2) + 2)}{\delta}$, and a
620 union bound over $f \in \mathcal{V}$, for any $t \geq 1$, with probability at least $1 - \delta/(2t^2)$, we have for all $f \in \mathcal{V}$,
621

622
$$\begin{aligned} 623 & 2 \sum_{s=1}^t E_s(f) \\ 624 & \leq \sqrt{\iota_t} \cdot \sqrt{\sigma^2 \sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \sigma^4} \\ 625 & \quad + \iota_t \left(L \max_{s \in [t]} w_s^2 D_s \cdot \sqrt{\sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \lambda + \sigma^2} \right) \\ 626 & \stackrel{(a)}{\leq} \left(\sqrt{\iota_t} \sigma + \iota_t L \max_{s \in [t]} w_s^2 D_s \right) \sqrt{\sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \sqrt{\lambda} \iota_t L \max_{s \in [t]} w_s^2 D_s + 2\iota_t \sigma^2} \\ 627 & \stackrel{(b)}{\leq} \frac{1}{2} \sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \frac{1}{2} \left(\sqrt{\iota_t} \sigma + \iota_t L \max_{s \in [t]} w_s^2 D_s \right)^2 + \frac{1}{2} \left(\iota_t L \max_{s \in [t]} w_s^2 D_s \right)^2 \\ 628 & \quad + \frac{1}{2} \lambda + 2\iota_t \sigma^2 \\ 629 & \stackrel{(c)}{\leq} \frac{1}{2} \sum_{s=1}^t w_s^2 (f(X_s) - f_*(X_s))^2 + \frac{1}{2} \left(\sqrt{\iota_t} \sigma + 2\iota_t L \max_{s \in [t]} w_s^2 D_s \right)^2 + \frac{1}{2} \lambda + 2\iota_t \sigma^2, \end{aligned}$$

630 where (a) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$ and $\iota_t \geq 1$, (b) holds due to $\sqrt{ab} \leq$
631 $a/2 + b/2$ for any $a, b \geq 0$, and (c) holds due to $a^2 + b^2 \leq (a+b)^2$ for any $a, b \geq 0$. Let $g \in \mathcal{V}$

648 such that $\|g - \hat{f}_{t+1}\|_\infty \leq \epsilon$, then
649

$$\begin{aligned}
& \sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))^2 \\
& \leq 2 \sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))(Y_s - f_*(X_s)) \\
& \leq 2 \sum_{s=1}^t w_s^2 (g(X_s) - f_*(X_s))(Y_s - f_*(X_s)) + 2W^2 R t \epsilon \\
& \leq \frac{1}{2} \sum_{s=1}^t w_s^2 (g(X_s) - f_*(X_s))^2 + \frac{1}{2} \left(\sqrt{\iota_t} \sigma + 2\iota_t R \max_{s \in [t]} w_s^2 D_s \right)^2 + \frac{1}{2} \lambda + 2\iota_t \sigma^2 + 2W^2 R t \epsilon \\
& \leq \frac{1}{2} \sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))^2 + \frac{1}{2} \left(\sqrt{\iota_t} \sigma + 2\iota_t R \max_{s \in [t]} w_s^2 D_s \right)^2 + \frac{1}{2} \lambda + 2\iota_t \sigma^2 + 3W^2 R t \epsilon.
\end{aligned}$$

664 That is for any fixed $t > 0$, we have
665

$$\begin{aligned}
& \sum_{s=1}^t w_s^2 (\hat{f}_{t+1}(X_s) - f_*(X_s))^2 \leq \left(\sqrt{\iota_t} \sigma + 2\iota_t R \max_{s \in [t]} w_s^2 D_s \right)^2 + \lambda + 4\iota_t \sigma^2 + 6W^2 R t \epsilon \\
& \leq \left(3\sqrt{\iota_t} \sigma + 2\iota_t R \max_{s \in [t]} w_s^2 D_s + \sqrt{\lambda} + \sqrt{6W^2 R t \epsilon} \right)^2,
\end{aligned}$$

671 where the second inequality is due to $2\sqrt{ab} \leq a + b$ and $a + b \leq (\sqrt{a} + \sqrt{b})^2$ for any $a, b \geq 0$. Note
672

$$\begin{aligned}
w_t^2 D_t &= w_t^2 D_{\mathcal{F}}(X_t; X_{[t]}, w_{[t]}) \\
&= \sup_{f_1, f_2 \in \mathcal{F}} \frac{w_t^2 (f_1(X_t) - f_2(X_t))^2}{\sum_{s \in [t]} w_s^2 (f_1(X_s) - f_2(X_s))^2 + \lambda} \\
&\leq \min\{1, w_t^2 D_{\mathcal{F}}(X_t; X_{[t-1]}, w_{[t-1]})\}.
\end{aligned}$$

673 Finally, the result holds through a union bound over all $t \geq 1$ and $\sum_{t=1}^{\infty} \frac{1}{2t^2} \leq 1$. \square
674

681 B PROOF OF THEOREM 4.2

683 **Parameters in Algorithm 1** For any $t \in [T]$, $l \in [L]$, $\ell \in [\tilde{L}]$, let $\mathcal{B}_{t,l}$, $\tilde{\mathcal{B}}_{t,\ell}$ denote the confidence
684 region as follows:
685

$$\begin{aligned}
\mathcal{B}_{t,l} &:= \left\{ f \in \mathcal{F} : \sum_{s \in \Psi_{t,l}} w_s^2 (\hat{f}_{t,l}(x_s) - f(x_s))^2 \leq \beta_{t,l}^2 \right\}, \\
\tilde{\mathcal{B}}_{t,\ell} &:= \left\{ g \in \mathcal{G} : \sum_{s \in \tilde{\Psi}_{t,\ell}} \tilde{w}_s^2 (\hat{g}_{t,\ell}(x_s) - g(x_s))^2 \leq \tilde{\beta}_{t,\ell}^2 \right\}.
\end{aligned}$$

691 Here

$$\beta_{t,l} = 2^l \alpha \left(3\sqrt{\iota_t} + 2 \frac{\iota_t R}{\gamma} \right) + \sqrt{\lambda} + \sqrt{6Rt\epsilon}, \quad (\text{B.1})$$

$$\tilde{\beta}_{t,\ell} = 2^\ell \tilde{\alpha} \left(3\sqrt{\iota_t} + 2 \frac{\tilde{\iota}_t R^2}{\tilde{\gamma}} \right) + \sqrt{\tilde{\lambda}} + \sqrt{6R^2 t \epsilon}, \quad (\text{B.2})$$

697 where

$$\iota_t = 16 \log \frac{2N_{\mathcal{F}}(\epsilon) t^2 (\log(\sigma^2 R^2 t) + 2)(\log(R^2) + 2)}{\delta} = \tilde{O}(\log N_{\mathcal{F}})$$

$$\tilde{\iota}_t = 16 \log \frac{2N_{\mathcal{G}}(\tilde{\epsilon}) t^2 (\log(c_v^2 \sigma^2 R^4 t) + 2)(\log(R^4) + 2)}{\delta} = \tilde{O}(\log N_{\mathcal{G}}).$$

Furthermore, setting

$$\gamma = R\sqrt{\log N_{\mathcal{F}}}, \quad \tilde{\gamma} = R^2\sqrt{\log N_{\mathcal{G}}}, \quad (\text{B.3})$$

$$\lambda = \alpha^2 \log N_{\mathcal{F}}, \quad \tilde{\lambda} = \tilde{\alpha}^2 \log N_{\mathcal{G}}, \quad (\text{B.4})$$

$$\epsilon = \frac{\alpha^2 \log N_{\mathcal{F}}}{RT}, \quad \tilde{\epsilon} = \frac{\tilde{\alpha}^2 \log N_{\mathcal{G}}}{R^2 T}, \quad (\text{B.5})$$

we have

$$\beta_{t,l} = \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}}), \quad \tilde{\beta}_{t,\ell} = \tilde{O}(2^\ell \tilde{\alpha} \sqrt{\log N_{\mathcal{G}}}).$$

Property 1 (Properties of ADALEVEL). For any $t \in [T]$, suppose $l_t = l$, then

1. If $l = 0$:

$$D_{t,1}(X_t) \leq 2\alpha/\gamma.$$

2. If $l \in [L-1]$:

$$\begin{cases} w_t = \frac{2^l \alpha}{\gamma D_{t,l}(X_t)}, \\ D_{t,l+1}(X_t) \leq 2^{l+1} \alpha / \gamma, \\ \bar{\sigma}_t \leq 2^l \alpha; \end{cases} \quad \text{or} \quad \begin{cases} w_t = 1, \\ D_{t,l}(X_t) \leq 2^l \alpha / \gamma, \\ 2^{l-1} \alpha < \bar{\sigma}_t \leq 2^l \alpha. \end{cases}$$

3. If $l = L$:

$$\begin{cases} w_t = \frac{2^L \alpha}{\gamma D_{t,L}(X_t)}, \\ \bar{\sigma}_t \leq 2^L \alpha; \end{cases} \quad \text{or} \quad \begin{cases} w_t = 1, \\ D_{t,L}(X_t) \leq 2^L \alpha / \gamma, \\ 2^{L-1} \alpha < \bar{\sigma}_t \leq 2^L \alpha. \end{cases}$$

Proof of Theorem 4.2. For $t \in [T]$, we define events $\mathcal{E}_t, \mathcal{E}$ as

$$\mathcal{E}_t = \{\forall l \in [L], f_* \in \mathcal{B}_{t,l} \text{ and } \forall \ell \in [\tilde{L}], g_* \in \tilde{\mathcal{B}}_{t,\ell}\}, \quad \mathcal{E} = \bigcap_{k \in [K]} \mathcal{E}_t.$$

The following lemmas hold.

Lemma B.1. On event \mathcal{E}_t , we have for all $l \in [L], \ell \in [\tilde{L}]$,

$$\begin{aligned} |\hat{f}_{t,l}(x) - f_*(x)| &\leq \beta_{t,l} D_{t,l}(x), \\ |\hat{g}_{t,\ell}(x) - g_*(x)| &\leq \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x). \end{aligned}$$

Furthermore,

$$f_*(x) \leq \min_{l \in [L]} (\hat{f}_{t,l}(x) + \min\{R, \beta_{t,l} D_{t,l}(x)\}),$$

and

$$\sigma_t^2 \leq \bar{\sigma}_t^2.$$

Proof. On event \mathcal{E}_t , for any $l \in [L]$, we have

$$\begin{aligned} |\hat{f}_{t,l}(x) - f_*(x)| &\leq D_{t,l}(x) \sqrt{\sum_{s \in \Psi_{t,l}} w_s^2 (\hat{f}_{t,l}(x_s) - f_*(x))^2 + \lambda} \\ &\leq D_{t,l}(x) \sqrt{\beta_{t,l}^2 + \lambda} \\ &\approx \beta_{t,l} D_{t,l}(x), \end{aligned}$$

since $\sqrt{\lambda} = O(\beta_{t,l})$ according to (B.4). Similarly, for any $\ell \in [\tilde{L}]$,

$$|\hat{g}_{t,\ell}(x) - g_*(x)| \leq \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x).$$

Furthermore, since this holds for all $l \in [L]$, we can choose the upper confidence bound of $f_*(x)$ as

$$\min_{l \in [L]} (\hat{f}_{t,l}(x) + \min\{R, \beta_{t,l} D_{t,l}(x)\}) \geq f_*(x).$$

Recall the definition of $\bar{\sigma}_t$ in (4.3), we have for all $l \in [L], \ell \in [\tilde{L}]$,

$$\begin{aligned} & |(\hat{g}_{t,\ell}(x_t) - \hat{f}_{t,l}^2(x_t)) - (g_*(x_t) - f_*^2(x_t))| \\ & \leq |\hat{g}_{t,\ell}(x_t) - g_*(x_t)| + |\hat{f}_{t,l}(x_t) + f_*(x_t)| \cdot |\hat{f}_{t,l}(x_t) - f_*(x_t)| \\ & \leq \min\{R^2, \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x_t)\} + \min\{R^2, 2R\beta_{t,l} D_{t,l}(x_t)\}. \end{aligned}$$

Therefore, σ_t^2 is bounded by $\bar{\sigma}_t^2$:

$$\begin{aligned} \sigma_t^2 &= \text{Var}[y_t^2|x_t] = \mathbb{E}[y_t^2|x_t] - \mathbb{E}^2[y_t|x_t] = g_*(x_t) - f_*^2(x_t) \\ &\leq \min_{l \in [L], \ell \in [\tilde{L}]} \{\sigma^2, \hat{g}_{t,\ell}(x_t) - \hat{f}_{t,l}^2(x_t) + R \min\{R, 2\beta_{t,l} D_{t,l}(x_t)\} + \min\{R^2, \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x_t)\}\} \\ &= \bar{\sigma}_t^2. \end{aligned}$$

□

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

Proof. By a union bound, with probability at least $1 - (L + \tilde{L})\delta$, the result follows from Lemma 4.3 using $\{X_t, Y_t, w_t\}_t = \{x_t, y_t, w_t\}_{t \in \Psi_{T+1,l}}, \mathcal{F}$ for $l \in [L]$, and using $\{X_t, Y_t, w_t\}_t = \{x_t, y_t^2, \tilde{w}_t\}_{t \in \tilde{\Psi}_{T+1,\ell}}, \mathcal{F} = \mathcal{G}$ for $\ell \in [\tilde{L}]$. We will check the conditions of Lemma 4.3 for all $t \in [T]$ by induction.

First, for $t = 1$, the result holds trivially.

Next, for $t > 1$, suppose event $\bigcap_{s \in [t]} \mathcal{E}_s$ holds, by Lemma B.1, we have for all $s \in [t]$,

$$\sigma_s^2 \leq \bar{\sigma}_s^2.$$

Thus from Property 1, for all $l \in [L], \ell \in [\tilde{L}]$,

$$\begin{aligned} \text{Var}[y_s|x_s] &= \sigma_s^2 \leq \bar{\sigma}_s^2 \leq 2^l \alpha, \quad w_s D_{s,l}(x_s) \leq \frac{2^l \alpha}{\gamma}, \quad \forall s \in \Psi_{t+1,l}, \\ \text{Var}[y_s^2|x_s] &\leq c_v^2 \sigma_s^2 \leq c_v^2 \bar{\sigma}_s^2 \leq 2^\ell \tilde{\alpha}, \quad \tilde{w}_s \tilde{D}_{s,\ell}(x_s) \leq \frac{2^\ell \tilde{\alpha}}{\tilde{\gamma}}, \quad \forall s \in \tilde{\Psi}_{t+1,\ell}. \end{aligned}$$

Applying Lemma 4.3 with $\sigma = 2^l \alpha, \max_{s \in [t]} w_s^2 D_{\mathcal{F}}(X_s; X_{[s-1]}, w_{[s-1]}) = \frac{2^l \alpha}{\gamma}$, we have

$$\sum_{s \in \Psi_{t+1,l}} w_s^2 (\hat{f}_{t+1,l}(x_s) - f_*(x_s))^2 \leq \beta_{t+1,l}^2,$$

that is $f_* \in \mathcal{B}_{t+1,l}$ for all $l \in [L]$. Applying Lemma 4.3 again with $\sigma = 2^\ell \tilde{\alpha}, \max_{s \in [t]} w_s^2 D_{\mathcal{F}}(X_s; X_{[s-1]}, w_{[s-1]}) = \frac{2^\ell \tilde{\alpha}}{\tilde{\gamma}}$, we have

$$\sum_{s \in \tilde{\Psi}_{t+1,\ell}} \tilde{w}_s^2 (\hat{g}_{t+1,\ell}(x_s) - g_*(x_s))^2 \leq \tilde{\beta}_{t+1,\ell}^2,$$

that is $g_* \in \tilde{\mathcal{B}}_{t+1,\ell}$ for all $\ell \in [\tilde{L}]$, so event \mathcal{E}_{t+1} holds.

Then the proof is completed by induction over $t \in [T]$. □

We define

$$U := \sum_{t \in [T]} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(x_t)\}, \quad (\text{B.6})$$

$$\tilde{U} := \sum_{t \in [T]} \min_{\ell \in [\tilde{L}]} \{R^2, \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x_t)\}. \quad (\text{B.7})$$

810 On event \mathcal{E} , which holds with probability at least $1 - (L + \tilde{L})\delta$ by Lemma B.2, by the optimism of
 811 x_t implied by Lemma B.1, regret is bounded as
 812

$$\begin{aligned}
 813 \text{Regret}(T) &= \sum_{t \in [T]} (f_*(x_t^*) - f_*(x_t)) \\
 814 &\leq \sum_{t \in [T]} \left[\min_{l \in [L]} (\hat{f}_{t,l}(x_t) + \min\{R, \beta_{t,l} D_{t,l}(x_t)\}) - f_*(x_t) \right] \\
 815 &\leq 2 \sum_{t \in [T]} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(x_t)\} \\
 816 &\stackrel{(B.6)}{=} 2U. \\
 817 \\
 818 \\
 819 \\
 820 \\
 821 \\
 822
 \end{aligned} \tag{B.8}$$

823 Setting

$$\alpha = R\sqrt{\frac{d_{\mathcal{F}} \log N_{\mathcal{F}}}{T}}, \quad \tilde{\alpha} = R^2\sqrt{\frac{d_{\mathcal{G}} \log N_{\mathcal{G}}}{T}}, \tag{B.9}$$

825 applying Lemma E.3 to U and \tilde{U} , we have
 826

$$U = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}}\right), \tag{B.10}$$

$$\tilde{U} = \tilde{O}\left(c_v R \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R^2 d_{\mathcal{G}} \log N_{\mathcal{G}}\right). \tag{B.11}$$

830 Furthermore,
 831

$$\begin{aligned}
 832 \sum_{t \in [T]} \bar{\sigma}_t^2 &\leq \sum_{t \in [T]} (\sigma_t^2 + 2R \min_{l \in [L]} \{R, 2\beta_{t,l} D_{t,l}(x_t)\} + 2 \min_{\ell \in [\tilde{L}]} \{R^2, \tilde{\beta}_{t,\ell} \tilde{D}_{t,\ell}(x_t)\}) \\
 833 &\leq \sum_{t \in [T]} \sigma_t^2 + 4RU + 2\tilde{U}. \\
 834 \\
 835
 \end{aligned} \tag{B.12}$$

836 So we have
 837

$$\begin{aligned}
 838 \tilde{U} &\stackrel{(B.11)}{\lesssim} c_v R \sqrt{d_{\mathcal{G}} \log N_{\mathcal{G}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R^2 d_{\mathcal{G}} \log N_{\mathcal{G}} \\
 839 &\stackrel{(B.12)}{\lesssim} c_v R \sqrt{d_{\mathcal{G}} \log N_{\mathcal{G}} \left(\sum_{t \in [T]} \sigma_t^2 + RU + \tilde{U} \right)} + R^2 d_{\mathcal{G}} \log N_{\mathcal{G}} \\
 840 &\lesssim c_v R \sqrt{d_{\mathcal{G}} \log N_{\mathcal{G}} \left(\sum_{t \in [T]} \sigma_t^2 + RU \right)} + \max\{1, c_v^2\} R^2 d_{\mathcal{G}} \log N_{\mathcal{G}}, \\
 841 \\
 842
 \end{aligned} \tag{B.13}$$

843 where the last inequality holds since $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ for any $x \geq 0$.
 844

845 And

$$\begin{aligned}
 846 U &\stackrel{(B.10)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 847 &\stackrel{(B.12)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \left(\sum_{t \in [T]} \sigma_t^2 + RU + \tilde{U} \right)} + R d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 848 &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \left(\sum_{t \in [T]} \sigma_t^2 + RU \right)} + R d_{\mathcal{F}} \log N_{\mathcal{F}} + \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \tilde{U}}, \\
 849 \\
 850
 \end{aligned}$$

864 where the last inequality holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Here
 865

$$\begin{aligned}
 866 \quad & \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \tilde{U}} \\
 867 \quad &= \sqrt{\max\{1, c_v\} R \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} d_{\mathcal{G}} \log N_{\mathcal{G}}} \cdot \frac{\sqrt{\frac{d_{\mathcal{F}} \log N_{\mathcal{F}}}{d_{\mathcal{G}} \log N_{\mathcal{G}}}}}{\max\{1, c_v\} R} \tilde{U} \\
 868 \quad &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \left(\sum_{t \in [T]} \sigma_t^2 + RU \right)} + \max\{1, c_v\} R \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} d_{\mathcal{G}} \log N_{\mathcal{G}}} \\
 869 \quad &\approx \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \left(\sum_{t \in [T]} \sigma_t^2 + RU \right)} + CR d_{\mathcal{F}} \log N_{\mathcal{F}},
 \end{aligned}$$

870 where the inequality holds due to Cauchy-Schwartz inequality and (B.13), and the last equality holds
 871 since $C := \max\{1, c_v\} \sqrt{\frac{d_{\mathcal{G}} \log N_{\mathcal{G}}}{d_{\mathcal{F}} \log N_{\mathcal{F}}}}$. Plugin back, we have
 872

$$\begin{aligned}
 873 \quad U &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \left(\sum_{t \in [T]} \sigma_t^2 + RU \right)} + \max\{1, C\} R d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 874 \quad &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + \max\{1, C\} R d_{\mathcal{F}} \log N_{\mathcal{F}}. \tag{B.14}
 \end{aligned}$$

875 Finally, combining (B.8) and (B.14), we have
 876

$$\text{Regret}(T) = \tilde{O} \left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \sigma_t^2} + \max\{1, C\} R d_{\mathcal{F}} \log N_{\mathcal{F}} \right).$$

□

C PROOFS FOR REINFORCEMENT LEARNING

892 **Parameters in Algorithm 3** For $k \in [K], l \in [L]$, let $\mathcal{B}_{k,l}$ denote the confidence region as follows:

$$\mathcal{B}_{k,l} := \left\{ f \in \mathcal{F} : \sum_{(i,h) \in \Psi_{k,l}} w_{i,h}^2 (\hat{f}_{k,l}(z_{i,h}) - f(z_{i,h}))^2 + \sum_{(i,h) \in \tilde{\Psi}_{k,l}} \tilde{w}_{i,h}^2 (\hat{f}_{k,l}(\tilde{z}_{i,h}) - f(\tilde{z}_{i,h}))^2 \leq \beta_{k,l}^2 \right\}.$$

893 Here

$$\beta_{k,l} = 2^l \alpha \left(3\sqrt{\iota_k} + 2 \frac{\iota_k}{\gamma} \right) + \sqrt{\lambda} + \sqrt{12kH\epsilon} \tag{C.1}$$

894 where

$$\iota_k = 16 \log \frac{16N_{\mathcal{F}}(\epsilon)k^2H^2(\log(2kH) + 2)}{\delta} = \tilde{O}(\log N_{\mathcal{F}}).$$

895 Furthermore, setting

$$\gamma = \sqrt{\log N_{\mathcal{F}}}, \tag{C.2}$$

$$\lambda = \alpha^2 \log N_{\mathcal{F}}, \tag{C.3}$$

$$\epsilon = \frac{\alpha^2 \log N_{\mathcal{F}}}{KH}, \tag{C.4}$$

896 we have

$$\beta_{k,l} = \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}}).$$

C.1 OPTIMISM

897 For $k \in [K]$, we define events \mathcal{E}_k , and \mathcal{E} as

$$\mathcal{E}_k = \{\forall l \in [L], f_* \in \mathcal{B}_{k,l}\}, \quad \mathcal{E} = \bigcap_{k \in [K]} \mathcal{E}_k.$$

900 The following lemmas hold.

918 **Lemma C.1.** On event \mathcal{E}_k , we have for all $l \in [L]$,

$$|\widehat{f}_{k,l}(z) - f_*(z)| \leq \beta_{k,l} D_{k,l}(z).$$

919 Furthermore, for all $h \in [H]$,

$$r_h(s_h^k, a_h^k) + [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \leq V_{k,h}(s_h^k),$$

$$V_{k,h}(s_h^k) - r_h(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \leq 2 \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(z_{k,h})\},$$

920 and

$$[\mathbb{V}V_{k,h+1}](s_h^k, a_h^k) \leq \bar{\sigma}_{k,h}^2,$$

$$\bar{\sigma}_{k,h}^2 - [\mathbb{V}V_{k,h+1}](s_h^k, a_h^k) \leq 2 \min_{l \in [L]} \{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\} + 2 \min_{\ell \in [L]} \{1, \beta_{k,\ell} D_{k,\ell}(\tilde{z}_{k,h})\}.$$

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

931 **Lemma C.2.** Event \mathcal{E} holds with probability at least $1 - L\delta$.

932 *Proof.* See Appendix D.2 for a detailed proof. \square

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

934 *Proof.* See Appendix D.3 for a detailed proof. \square

935 C.2 HIGHER-ORDER QUANTITIES IN MDPs

936 Inspired by Zhao et al. (2023); Huang et al. (2024), we define the following higher-order quantities
937 of MDPs. Let $M = \lceil \log_2(3KH) \rceil$.

938 We define \mathcal{K} to be a set of episodes when the sum of uncertainty within each level grows smoothly:

$$\mathcal{K} := \{k \in [K] : \forall l \in [L], \sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) \leq 1\}. \quad (\text{C.5})$$

939 Let $\tilde{\mathcal{K}} := [K] \setminus \mathcal{K}$. We can prove the number of episodes when uncertainty grows sharply is at most
940 $|\tilde{\mathcal{K}}| = \tilde{O}(Ld_{\mathcal{F}})$.

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

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

$$\tilde{V}_{k,h}(s) = V_h^*(s) - V_h^{\pi^k}(s), \quad \forall s \in \mathcal{S}, (k, h) \in [K] \times [H]. \quad (\text{C.7})$$

943 In addition, we use \check{S}_m , \tilde{S}_m to represent the total variance of 2^m -th order value functions $\check{V}_{k,h+1}^{2^m}$,
944 $\tilde{V}_{k,h+1}^{2^m}$:

$$\check{S}_m = \sum_{k \in \mathcal{K}} \sum_h [\mathbb{V}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k), \quad (\text{C.8})$$

$$\tilde{S}_m = \sum_{k \in \mathcal{K}} \sum_h [\mathbb{V}\tilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k). \quad (\text{C.9})$$

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

$$\check{A}_m = \left| \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^m}(s_{h+1}^k) \right] \right|, \quad (\text{C.10})$$

$$\tilde{A}_m = \left| \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\tilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \tilde{V}_{k,h+1}^{2^m}(s_{h+1}^k) \right] \right|. \quad (\text{C.11})$$

972 Finally, we use the R, \tilde{R} to denote the summation of bonuses:
 973

$$974 R = \sum_{k \in \mathcal{K}} \sum_h \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(z_{k,h})\}, \quad (C.12)$$

$$975 \tilde{R} = \sum_{k \in \mathcal{K}} \sum_h \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(\tilde{z}_{k,h})\} \quad (C.13)$$

979 Now, we introduce the following lemmas to build the connection between these quantities.
 980

981 **Lemma C.4.** We have

$$982 |\tilde{\mathcal{K}}| \leq 2Ld_{\mathcal{F}}. \quad (C.14)$$

984 *Proof.* See Appendix D.4 for a detailed proof. \square
 985

986 **Lemma C.5.** On event \mathcal{E} , we have for all $m \in \overline{[M]}$,
 987

$$988 \check{S}_m \leq \check{A}_{m+1} + 2^{m+1} \cdot (2R), \quad (C.15)$$

$$989 \tilde{S}_m \leq \tilde{A}_{m+1} + 2^{m+1} \cdot (2R + \check{A}_0). \quad (C.16)$$

992 *Proof.* See Appendix D.5 for a detailed proof. \square
 993

994 **Lemma C.6.** With probability at least $1 - 2\delta$, we have for all $m \in \overline{[M]}$,
 995

$$996 \check{A}_m \leq \sqrt{\zeta \check{S}_m} + \zeta, \quad (C.17)$$

$$997 \tilde{A}_m \leq \sqrt{\zeta \tilde{S}_m} + \zeta, \quad (C.18)$$

1000 where $\zeta = 8 \log(2(M+1)(\log(KH) + 2)/\delta)$. We denote the corresponding event by \mathcal{A} .
 1001

1002 *Proof.* See Appendix D.6 for a detailed proof. \square
 1003

1004 **Lemma C.7.** On event $\mathcal{E} \cap \mathcal{A}$, we have

$$1006 \check{A}_0 \leq 2\sqrt{2\zeta R} + 7\zeta, \quad (C.19)$$

$$1007 \check{A}_1 \leq 4\sqrt{\zeta R} + 7\zeta. \quad (C.20)$$

1010 *Proof.* See Appendix D.7 for a detailed proof. \square
 1011

1012 **Lemma C.8.** On event $\mathcal{E} \cap \mathcal{A}$, we have

$$1013 \tilde{A}_0 \leq 4\sqrt{2\zeta R} + 15\zeta. \quad (C.21)$$

1016 *Proof.* See Appendix D.8 for a detailed proof. \square
 1017

1018 **Lemma C.9.** Setting

$$1019 \alpha = \frac{d_{\mathcal{F}} \log N_{\mathcal{F}}}{KH}, \quad (C.22)$$

1021 on event $\mathcal{E} \cap \mathcal{A}$, we have

$$1023 R = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}}\right). \quad (C.23)$$

1025 *Proof.* See Appendix D.9 for a detailed proof. \square

1026 **C.3 REGRET ANALYSIS**
1027

1028 *Proof of Theorem 5.1.* On event $\mathcal{E} \cap \mathcal{A}$, which holds with probability at least $1 - (L + 2)\delta$ by
1029 Lemma C.2, C.6 and a union bound, we have all lemmas in this section hold. By the optimism
1030 implied by Lemma C.3, we have

1031
$$\text{Regret}(K) = \sum_{k=1}^K (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)) \leq \sum_{k=1}^K (V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)). \quad (\text{C.24})$$

1032
1033
1034

1035 We further use Lemma C.10 to bound the regret with the quantities defined in Section C.2.

1036 **Lemma C.10.** On event \mathcal{E} , we have

1037
$$\sum_{k=1}^K (V_{k,1}(s_1^k) - V^{\pi^k}(s_1^k)) \leq 2R + \check{A}_0 + \tilde{A}_0 + |\tilde{\mathcal{K}}|. \quad (\text{C.25})$$

1038
1039
1040

1041 *Proof.* First, we decompose $V_{k,1}(s_1^k)$ and $V_1^{\pi^k}(s_1^k)$ as follows

1042
$$\begin{aligned} V_{k,1}(s_1^k) &= \sum_{h=1}^H [V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k)] \\ &= \sum_{h=1}^H r(s_h^k, a_h^k) + \sum_{h=1}^H [V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)] \\ &\quad + \sum_{h=1}^H [[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k)], \end{aligned}$$

1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065

1042
$$\begin{aligned} V_1^{\pi^k}(s_1^k) &= \sum_{h=1}^H [V_h^{\pi^k}(s_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \\ &= \sum_{h=1}^H r(s_h^k, a_h^k) + \sum_{h=1}^H [V_h^{\pi^k}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k)] \\ &\quad + \sum_{h=1}^H [[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \\ &= \sum_{h=1}^H r(s_h^k, a_h^k) + \sum_{h=1}^H [[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)], \end{aligned}$$

1066 Thus it follows that

1067
$$\begin{aligned} V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k) &= \sum_{h=1}^H [V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)] \\ &\quad + \sum_{h=1}^H [[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k)] - \sum_{h=1}^H [[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \\ &= \sum_{h=1}^H [V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)] \\ &\quad + \sum_{h=1}^H [[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k)] + \sum_{h=1}^H [[\mathbb{P}\tilde{V}_{k,h+1}](s_h^k, a_h^k) - \tilde{V}_{k,h+1}(s_{h+1}^k)]. \end{aligned}$$

1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

1080 Then we have

$$\begin{aligned}
 & \sum_{k=1}^K \left[V_{k,1}(s_1^k) - V^{\pi^k}(s_1^k) \right] \\
 & \stackrel{(a)}{\leq} |\tilde{\mathcal{K}}| + 2 \sum_{k \in \mathcal{K}} \sum_h \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(z_{k,h})\} + \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k) \right] \\
 & \quad + \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\tilde{V}_{k,h+1}](s_h^k, a_h^k) - \tilde{V}_{k,h+1}(s_{h+1}^k) \right] \\
 & = 2R + \check{A}_0 + \tilde{A}_0 + |\tilde{\mathcal{K}}|,
 \end{aligned}$$

1092 where (a) is due to Lemma C.1. □

1093

1094 Then we have

$$\begin{aligned}
 & 2R + \check{A}_0 + \tilde{A}_0 + |\tilde{\mathcal{K}}| \\
 & \stackrel{(a)}{\lesssim} R + \sqrt{R} + d_{\mathcal{F}} \\
 & \stackrel{(b)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}},
 \end{aligned} \tag{C.26}$$

1101 where (a) holds due to (C.19), (C.21), (C.14), and (b) holds due to (C.23).

1102

1103 Finally, Combining (C.24), (C.25) and (C.26), the high-probability regret bound is given by

$$\text{Regret}(K) = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}}\right).$$

1106

1107

1108

D MISSING PROOFS IN SECTION C

1110

D.1 PROOF OF LEMMA C.1

1112

1113 *Proof of Lemma C.1.* On event \mathcal{E}_k , for any $l \in [L]$, $f_* \in \mathcal{B}_{k,l}$, it follows that

$$\begin{aligned}
 & |\hat{f}_{k,l}(z) - f_*(z)| \\
 & \stackrel{(a)}{\leq} D_{k,l}(z) \sqrt{\sum_{(i,h) \in \Psi_{k,l}} w_{i,h}^2 (\hat{f}_{k,l}(z_{i,h}) - f_*(z_{i,h}))^2 + \sum_{(i,h) \in \tilde{\Psi}_{k,l}} \tilde{w}_{i,h}^2 (\hat{f}_{k,l}(\tilde{z}_{i,h}) - f_*(\tilde{z}_{i,h}))^2 + \lambda} \\
 & \stackrel{(b)}{\leq} D_{k,l}(z) \sqrt{\beta_{k,l}^2 + \lambda} \\
 & \stackrel{(c)}{\approx} \beta_{k,l} D_{k,l}(z),
 \end{aligned}$$

1123

1124 where (a) holds due to the definition of $D_{\mathcal{F}}$ in (3.1), (b) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any
1125 $a, b \geq 0$, (c) holds due to $\sqrt{\lambda} = O(\beta_{k,l})$ by (C.3).

1126

Furthermore, since this holds for all $l \in [L]$, we have

$$\begin{aligned}
 & r_h(s_h^k, a_h^k) + [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \\
 & = r_h(s_h^k, a_h^k) + f_*(z_{k,h}) \\
 & \leq \min_{l \in [L]} \{1, r_h(s_h^k, a_h^k) + \hat{f}_{k,l}(z_{k,h}) + \beta_{k,l} D_{k,l}(z_{k,h})\} \\
 & = V_{k,h}(s_h^k).
 \end{aligned}$$

1134 Thus,

$$\begin{aligned}
 & V_{k,h}(s_h^k) - r_h(s_h^k, a_h^k) + [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \\
 &= \min_{l \in [L]} \{1, r_h(s_h^k, a_h^k) + \hat{f}_{k,l}(z_{k,h}) + \beta_{k,l} D_{k,l}(z_{k,h})\} - r_h(s_h^k, a_h^k) - f_*(z_{k,h}) \\
 &\leq 2 \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(z_{k,h})\}.
 \end{aligned}$$

1140

1141 Recall the definition of $\bar{\sigma}_{k,h}$ in (5.2), we have for all $l \in [L], \ell \in [L]$,

$$\begin{aligned}
 & |[\hat{f}_{k,\ell}(\tilde{z}_{k,h}) - \hat{f}_{k,l}^2(z_{k,h})] - [f_*(\tilde{z}_{k,h}) - f_*^2(z_{k,h})]| \\
 &\leq |\hat{f}_{k,\ell}(\tilde{z}_{k,h}) - f_*(\tilde{z}_{k,h})| + |\hat{f}_{k,l}^2(z_{k,h}) - f_*^2(z_{k,h})| \\
 &= |\hat{f}_{k,\ell}(\tilde{z}_{k,h}) - f_*(\tilde{z}_{k,h})| + |\hat{f}_{k,l}(z_{k,h}) + f_*(z_{k,h})| \cdot |\hat{f}_{k,l}(z_{k,h}) - f_*(z_{k,h})| \\
 &\leq \min\{1, \beta_{k,l} D_{k,\ell}(\tilde{z}_{k,h})\} + \min\{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\},
 \end{aligned}$$

1147

1148 where the last inequality holds due to $\hat{f}_{k,l}, f_* \in [0, 1]$. Therefore, $[\mathbb{V}V_{k,h+1}](s_h^k, a_h^k)$ is bounded by
1149 $\bar{\sigma}_{k,h}^2$:

$$\begin{aligned}
 & [\mathbb{V}V_{k,h+1}](s_h^k, a_h^k) \\
 &= [\mathbb{P}V_{k,h+1}^2](s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}]^2(s_h^k, a_h^k) = f_*(\tilde{z}_{k,h}) - f_*^2(z_{k,h}) \\
 &\leq \min_{l \in [L], \ell \in [L]} \left\{ 1, \hat{f}_{k,\ell}(\tilde{z}_{k,h}) - \hat{f}_{k,l}^2(z_{k,h}) + \min\{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\} + \min\{1, \beta_{k,\ell} D_{k,\ell}(\tilde{z}_{k,h})\} \right\} \\
 &= \bar{\sigma}_{k,h}^2.
 \end{aligned}$$

1156

1157 Thus

$$\begin{aligned}
 & \bar{\sigma}_{k,h}^2 - [\mathbb{V}V_{k,h+1}](s_h^k, a_h^k) \\
 &= \min_{l \in [L], \ell \in [L]} \left\{ 1, \hat{f}_{k,\ell}(\tilde{z}_{k,h}) - \hat{f}_{k,l}^2(z_{k,h}) + \min\{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\} + \min\{1, \beta_{k,\ell} D_{k,\ell}(\tilde{z}_{k,h})\} \right\} \\
 &\quad - f_*(\tilde{z}_{k,h}) + f_*^2(z_{k,h}) \\
 &\leq 2 \min_{l \in [L]} \{1, 2\beta_{k,l} D_{k,l}(z_{k,h})\} + 2 \min_{\ell \in [L]} \{1, \beta_{k,\ell} D_{k,\ell}(\tilde{z}_{k,h})\}.
 \end{aligned}$$

1164

1165

D.2 PROOF OF LEMMA C.2

1167

1168 *Proof of Lemma C.2.* By a union bound, with probability at least $1 - L\delta$, the result follows from
1169 Lemma 4.3 using $\{X_{2t-1}, Y_{2t-1}, w_{2t-1}\}_t \cup \{X_{2t}, Y_{2t}, w_{2t}\}_t = \{z_{k,h}, y_{k,h}, w_{k,h}\}_{(k,h) \in \Psi_{K+1,l}} \cup$
1170 $\{\tilde{z}_{k,h}, y_{k,h}^2, \tilde{w}_{k,h}\}_{(k,h) \in \tilde{\Psi}_{K+1,l}}$, \mathcal{F} for $l \in [L]$. We will check the conditions of Lemma 4.3 for all
1171 $k \in [K]$ by induction.

1172

First, for $k = 1$, the result holds trivially.

1173

Next, for $k > 1$, suppose event $\bigcap_{i \in [k]} \mathcal{E}_i$ holds, by Lemma C.1, we have for all $(i, h) \in [k] \times [H]$,

$$[\mathbb{V}V_{i,h+1}](s_h^i, a_h^i) \leq \bar{\sigma}_{i,h}^2.$$

1176

Thus from Property 1, for all $l \in [L]$,

1177

$$\text{Var}[y_{i,h}|z_{i,h}] = [\mathbb{V}V_{i,h+1}](s_h^i, a_h^i) \leq \bar{\sigma}_{i,h}^2 \leq 2^l \alpha, \quad w_{i,h} D_{k,h,l}(z_{i,h}) \leq \frac{2^l \alpha}{\gamma} \quad \forall (i, h) \in \Psi_{k+1,l},$$

1179

1180

$$\text{Var}[y_{i,h}^2|\tilde{z}_{i,h}] = [\mathbb{V}V_{i,h+1}^2](s_h^i, a_h^i) \leq c_v^2 \bar{\sigma}_{i,h}^2 \leq 2^l \alpha, \quad \tilde{w}_{i,h} D_{k,h,l}(\tilde{z}_{i,h}) \leq \frac{2^l \alpha}{\gamma} \quad \forall (i, h) \in \tilde{\Psi}_{k+1,l}.$$

1181

1182

Applying Lemma 4.3 with $\sigma = 2^l \alpha, \max_{s \in [t]} w_s^2 D_{\mathcal{F}}(X_s; X_{[s-1]}, w_{[s-1]}) = \frac{2^l \alpha}{\gamma}$, we have

1183

1184

1185

$$\sum_{(i,h) \in \Psi_{k+1,l}} w_{i,h}^2 (\hat{f}_{k+1,l}(z_{i,h}) - f_*(z_{i,h}))^2 + \sum_{(i,h) \in \tilde{\Psi}_{k+1,l}} \tilde{w}_{i,h}^2 (\hat{f}_{k+1,l}(\tilde{z}_{i,h}) - f_*(\tilde{z}_{i,h}))^2 \leq \beta_{k+1,l}^2,$$

1186

1187

that is $f_* \in \mathcal{B}_{k+1,l}$ for all $l \in [L]$, so event \mathcal{E}_{k+1} holds.

Then the proof is completed by induction over $k \in [K]$. \square

1188 D.3 PROOF OF LEMMA C.3
1189

1190 *Proof of Lemma C.3.* We prove the optimism by induction. When $h = H + 1$, we have $V_{k,H+1}(\cdot) =$
1191 $V_{H+1}^*(\cdot) = 0$, and the result holds trivially. We assume the statement is true for all $h + 1$, and prove
1192 the case of h . For any (s, a) , if $Q_{k,h}(s, a) = 1$, then $Q_{k,h}(s, a) = 1 \geq Q_h^*(s, a)$. Otherwise, we
1193 have

$$\begin{aligned} & Q_{k,h}(s, a) - Q_h^*(s, a) \\ &= \min_{l \in [L]} \left\{ 1, r_h(s, a) + \hat{f}_{k,l}(s, a, V_{k,h+1}) + \min\{1, \beta_{k,l} D_{k,l}(s, a, V_{k,h+1})\} \right\} \\ &\quad - r_h(s, a) - f_*(s, a, V_{h+1}^*) \\ &\geq \min_{l \in [L]} \left(\hat{f}_{k,l}(s, a, V_{k,h+1}) + \min\{1, \beta_{k,l} D_{k,l}(s, a, V_{k,h+1})\} \right) - f_*(s, a, V_{k,h+1}) \\ &\geq 0, \end{aligned}$$

1202 where the first inequality holds due to $V_{k,h+1}(\cdot) \geq V_{h+1}^*(\cdot)$ and the second holds due to Lemma C.1.
1203 That is, we have $Q_{k,h}(\cdot, \cdot) \geq Q_h^*(\cdot, \cdot)$ and therefore $V_{k,h}(\cdot) \geq V_h^*(\cdot)$. Then the proof is completed
1204 by induction. \square

1205 D.4 PROOF OF LEMMA C.4

1206 *Proof of Lemma C.4.* Recall the definition of $\tilde{\mathcal{K}}$, we have

$$k \in \tilde{\mathcal{K}} \iff \exists l \in [L], \sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) > 1.$$

1207 Let $\tilde{\mathcal{K}}_l$ denote the indices k such that

$$\tilde{\mathcal{K}}_{l,1} := \left\{ k \in [K] : \sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) > 1 \right\}.$$

1208 Then we have $|\tilde{\mathcal{K}}| \leq |\bigcup_{l \in [L]} \tilde{\mathcal{K}}_l| \leq \sum_{l \in [L]} |\tilde{\mathcal{K}}_l|$. For any $l \in [L]$, we have

$$\begin{aligned} |\tilde{\mathcal{K}}_l| &\leq \sum_{k=1}^K \min \left\{ 1, \sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) \right\} \\ &\leq \sum_{(k,h) \in \Psi_{K+1,l}} \min \left\{ 1, w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) \right\} + \sum_{(k,h) \in \tilde{\Psi}_{K+1,l}} \min \left\{ 1, \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) \right\} \\ &\leq 2d_{\mathcal{F}}. \end{aligned}$$

1209 Taking the summation over $l \in [L]$ gives the upper bound of $|\tilde{\mathcal{K}}|$. \square

1210 D.5 PROOF OF LEMMA C.5

1211 *Proof of Lemma C.5.* We are to bound \check{S}_m and \tilde{S}_m with similar arguments.

1212 Recall the definition of \check{S}_m in (C.8), we have

$$\begin{aligned} \check{S}_m &= \sum_{k \in \mathcal{K}} \sum_h [\mathbb{V} \check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) \\ &= \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P} \check{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\ &= \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P} \check{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) \right] + \sum_{k \in \mathcal{K}} \sum_h \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\ &\quad + \sum_{k \in \mathcal{K}} \sum_h \left(\check{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) - \check{V}_{k,h}^{2^{m+1}}(s_h^k) \right) \\ &\leq \check{A}_{m+1} + \sum_{k \in \mathcal{K}} \sum_h \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P} \check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]. \end{aligned} \tag{D.1}$$

1242 For the second term in (D.1), we have
 1243
 1244
 1245
 1246

$$\begin{aligned}
 & \sum_{k \in \mathcal{K}} \sum_h \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\
 & \stackrel{(a)}{\leq} \sum_{k \in \mathcal{K}} \sum_h \left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}]^{2^{m+1}}(s_h^k, a_h^k) \right] \\
 & = \sum_{k \in \mathcal{K}} \sum_h \left[\check{V}_{k,h}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) \right] \prod_{i=0}^m \left[\check{V}_{k,h}^{2^i}(s_h^k) + [\mathbb{P}\check{V}_{k,h+1}]^{2^i}(s_h^k, a_h^k) \right] \\
 & \leq 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \max \{ \check{V}_{k,h}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k), 0 \} \\
 & \stackrel{(b)}{\leq} 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \max \{ V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k), 0 \} \\
 & \stackrel{(c)}{\leq} 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h 2 \min_{l \in [L]} \{ 1, \beta_{k,l} \mathcal{D}_{k,l}(z_{k,h}) \} \\
 & = 2^{m+1} \cdot (2R), \tag{D.2}
 \end{aligned}$$

1264
 1265
 1266
 1267

1268 where (a) holds due to $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, (b) holds due to the definition of $\check{V}_{k,h}$ and $V_h^*(s_h^k) \geq$
 1269 $r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)$, while (c) is due to Lemma C.1. Substituting (D.2) into (D.1), we have
 1270

1271
 1272
 1273

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

1274
 1275
 1276
 1277
 1278
 1279

1280 Next, we proceed to bound \tilde{S}_m . Recall the definition of \tilde{S}_m in (C.9), we have
 1281
 1282
 1283
 1284

$$\begin{aligned}
 \tilde{S}_m &= \sum_{k \in \mathcal{K}} \sum_h [\mathbb{V}\tilde{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) \\
 &= \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\tilde{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\
 &= \sum_{k \in \mathcal{K}} \sum_h \left[[\mathbb{P}\tilde{V}_{k,h+1}^{2^{m+1}}](s_h^k, a_h^k) - \tilde{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) \right] + \sum_{k \in \mathcal{K}} \sum_h \left[\tilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\
 &\quad + \sum_{k \in \mathcal{K}} \sum_h \left(\tilde{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k) - \tilde{V}_{k,h}^{2^{m+1}}(s_h^k) \right) \\
 &\leq \tilde{A}_{m+1} + \sum_{k \in \mathcal{K}} \sum_h \left[\tilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right]. \tag{D.3}
 \end{aligned}$$

1296 For the second term in (D.3), we have

$$\begin{aligned}
& \sum_{k \in \mathcal{K}} \sum_h \left[\tilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}^{2^m}]^2(s_h^k, a_h^k) \right] \\
& \stackrel{(a)}{\leq} \sum_{k \in \mathcal{K}} \sum_h \left[\tilde{V}_{k,h}^{2^{m+1}}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}]^{2^{m+1}}(s_h^k, a_h^k) \right] \\
& = \sum_{k \in \mathcal{K}} \sum_h \left[\tilde{V}_{k,h}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}](s_h^k, a_h^k) \right] \prod_{i=0}^m \left[\tilde{V}_{k,h}^{2^i}(s_h^k) + [\mathbb{P}\tilde{V}_{k,h+1}]^{2^i}(s_h^k, a_h^k) \right] \\
& \leq 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \max \left\{ \tilde{V}_{k,h}(s_h^k) - [\mathbb{P}\tilde{V}_{k,h+1}](s_h^k, a_h^k), 0 \right\} \\
& \stackrel{(b)}{=} 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \max \left\{ V_h^*(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k), 0 \right\} \\
& \stackrel{(c)}{\leq} 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \left[\max \left\{ V_h(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}](s_h^k, a_h^k), 0 \right\} \right. \\
& \quad \left. + |[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k)| \right] \\
& \stackrel{(d)}{\leq} 2^{m+1} \sum_{k \in \mathcal{K}} \sum_h \left[2 \min_{l \in [L]} \{1, \beta_{k,l} D_{k,l}(z_{k,h})\} + |[\mathbb{P}\check{V}_{k,h+1}](s_h^k, a_h^k) - \check{V}_{k,h+1}(s_{h+1}^k)| \right] \\
& \leq 2^{m+1} \cdot (2R + \check{A}_0), \tag{D.4}
\end{aligned}$$

1319 where (a) holds due to $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, (b) holds due to the definition of $\tilde{V}_{k,h}$ and $V_h^{\pi^k}(s_h^k) =$
1320 $r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k)$, (c) holds due to $V_h^*(s_h^k) \geq r(s_h^k, a_h^k) + [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)$ and the
1321 definition of $\check{V}_{k,h}$, while (d) is due to Lemma C.1. Substituting (D.4) into (D.3), we have

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

□

D.6 PROOF OF LEMMA C.6

1328 *Proof of Lemma C.6.* Let $X_{k,h} = [\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - \check{V}_{k,h+1}^{2^m}(s_{h+1}^k)$, then we have $\mathbb{E}[X_{k,h} | \mathcal{G}_{k,h}] =$
1329 0 , $|X_{k,h}| \leq 2$ and $\mathbb{E}[X_{k,h}^2 | \mathcal{G}_{k,h}] = [\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k, a_h^k)$. Therefore, for any $m \in \overline{[M]}$, applying
1330 variance-aware Freedman's inequality in Lemma E.2, with probability at least $1 - \frac{1}{M+1}\delta$, we have

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

1332 Thus, taking a union bound over $m \in \overline{[M]}$, with probability at least $1 - \delta$, (C.17) holds. The proofs
1333 for (C.18) follow the same arguments as (C.17). □

D.7 PROOF OF LEMMA C.7

1339 *Proof of Lemma C.7.* On event $\mathcal{E} \cap \mathcal{A}$, we have (C.15) and (C.17) hold by Lemma C.5 and C.6.
1340 Substituting the bound of \check{S}_m in (C.15) into (C.17), we have for all $m \in \overline{[M]}$,

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

1343 And we have for all $m \in \overline{[M]}$, $\check{A}_m \leq 2KH$. Then the result follows by Lemma E.6. □

D.8 PROOF OF LEMMA C.8

1347 *Proof of Lemma C.8.* On event $\mathcal{E} \cap \mathcal{A}$, we have (C.16) and (C.18) hold by Lemma C.5 and C.6.
1348 Substituting the bound of \tilde{S}_m in (C.16) into (C.18), we have for all $m \in \overline{[M]}$,

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

1350 And we have for all $m \in \overline{[M]}$, $\tilde{A}_m \leq 2KH$. Applying Lemma E.6, we have
 1351

$$\begin{aligned} \tilde{A}_0 &\leq 2\sqrt{\zeta(2R + \check{A}_0)} + 7\zeta \\ &\stackrel{(a)}{\leq} 2\sqrt{2\zeta R} + 2\sqrt{\zeta \check{A}_0} + 7\zeta \\ &\stackrel{(b)}{\leq} 2\sqrt{2\zeta R} + \zeta + \check{A}_0 + 7\zeta \\ &\stackrel{(c)}{\leq} 4\sqrt{2\zeta R} + 15\zeta, \end{aligned}$$

1360 where (a) holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, (b) holds due to $2\sqrt{ab} \leq a+b$ for $a, b \geq 0$
 1361 and (c) holds due to (C.19) in Lemma C.7. \square
 1362

1363 D.9 PROOF OF LEMMA C.9

1364
 1365 *Proof of Lemma C.9.* We have for all $k \in \mathcal{K}, l \in [L]$,

$$\sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) \leq 1.$$

1366 By Lemma E.5, it follows that for all $h \in [H]$,
 1367

$$\begin{aligned} &D_{k,l}(z_{k,h}) \\ &\leq \exp \left\{ \frac{1}{2} \left(\sum_{j \in \Psi_{k,h,l} \setminus \Psi_{k,l}} D_{k,j,l}(z_{k,j}) + \sum_{j \in \tilde{\Psi}_{k,h,l} \setminus \tilde{\Psi}_{k,l}} D_{k,j,l}(\tilde{z}_{k,j}) \right) \right\} D_{k,h,l}(z_{k,h}) \\ &\leq \exp \left\{ \frac{1}{2} \left(\sum_{h \in \Psi_{k+1,l} \setminus \Psi_{k,l}} w_{k,h}^2 D_{k,h,l}^2(z_{k,h}) + \sum_{h \in \tilde{\Psi}_{k+1,l} \setminus \tilde{\Psi}_{k,l}} \tilde{w}_{k,h}^2 D_{k,h,l}^2(\tilde{z}_{k,h}) \right) \right\} D_{k,h,l}(z_{k,h}) \\ &\leq 2D_{k,h,l}(z_{k,h}). \end{aligned}$$

1379 Similarly, for all $k \in \mathcal{K}, h \in [H], \ell \in [L]$,
 1380

$$D_{k,\ell}(\tilde{z}_{k,h}) \leq 2D_{k,h,\ell}(\tilde{z}_{k,h}).$$

1381 Then we have
 1382

$$\begin{aligned} R &\leq 2 \sum_{k \in \mathcal{K}} \sum_h \min_{l \in [L]} \{1, \beta_{k,l} D_{k,h,l}(z_{k,h})\}, \\ \tilde{R} &\leq 2 \sum_{k \in \mathcal{K}} \sum_h \min_{\ell \in [L]} \{1, \beta_{k,\ell} D_{k,h,\ell}(\tilde{z}_{k,h})\}. \end{aligned}$$

1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403
 1404
 1405
 1406
 1407
 1408
 1409
 1410
 1411
 1412
 1413
 1414
 1415
 1416
 1417
 1418
 1419
 1420
 1421
 1422
 1423
 1424
 1425
 1426
 1427
 1428
 1429
 1430
 1431
 1432
 1433
 1434
 1435
 1436
 1437
 1438
 1439
 1440
 1441
 1442
 1443
 1444
 1445
 1446
 1447
 1448
 1449
 1450
 1451
 1452
 1453
 1454
 1455
 1456
 1457
 1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511
 1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565
 1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581
 1582
 1583
 1584
 1585
 1586
 1587
 1588
 1589
 1590
 1591
 1592
 1593
 1594
 1595
 1596
 1597
 1598
 1599
 1600
 1601
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619
 1620
 1621
 1622
 1623
 1624
 1625
 1626
 1627
 1628
 1629
 1630
 1631
 1632
 1633
 1634
 1635
 1636
 1637
 1638
 1639
 1640
 1641
 1642
 1643
 1644
 1645
 1646
 1647
 1648
 1649
 1650
 1651
 1652
 1653
 1654
 1655
 1656
 1657
 1658
 1659
 1660
 1661
 1662
 1663
 1664
 1665
 1666
 1667
 1668
 1669
 1670
 1671
 1672
 1673
 1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727
 1728
 1729
 1730
 1731
 1732
 1733
 1734
 1735
 1736
 1737
 1738
 1739
 1740
 1741
 1742
 1743
 1744
 1745
 1746
 1747
 1748
 1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
 1795
 1796
 1797
 1798
 1799
 1800
 1801
 1802
 1803
 1804
 1805
 1806
 1807
 1808
 1809
 1810
 1811
 1812
 1813
 1814
 1815
 1816
 1817
 1818
 1819
 1820
 1821
 1822
 1823
 1824
 1825
 1826
 1827
 1828
 1829
 1830
 1831
 1832
 1833
 1834
 1835
 1836
 1837
 1838
 1839
 1840
 1841
 1842
 1843
 1844
 1845
 1846
 1847
 1848
 1849
 1850
 1851
 1852
 1853
 1854
 1855
 1856
 1857
 1858
 1859
 1860
 1861
 1862
 1863
 1864
 1865
 1866
 1867
 1868
 1869
 1870
 1871
 1872
 1873
 1874
 1875
 1876
 1877
 1878
 1879
 1880
 1881
 1882
 1883
 1884
 1885
 1886
 1887
 1888
 1889
 1890
 1891
 1892
 1893
 1894
 1895
 1896
 1897
 1898
 1899
 1900
 1901
 1902
 1903
 1904
 1905
 1906
 1907
 1908
 1909
 1910
 1911
 1912
 1913
 1914
 1915
 1916
 1917
 1918
 1919
 1920
 1921
 1922
 1923
 1924
 1925
 1926
 1927
 1928
 1929
 1930
 1931
 1932
 1933
 1934
 1935
 1936
 1937
 1938
 1939
 1940
 1941
 1942
 1943
 1944
 1945
 1946
 1947
 1948
 1949
 1950
 1951
 1952
 1953
 1954
 1955
 1956
 1957
 1958
 1959
 1960
 1961
 1962
 1963
 1964
 1965
 1966
 1967
 1968
 1969
 1970
 1971
 1972
 1973
 1974
 1975
 1976
 1977
 1978
 1979
 1980
 1981
 1982
 1983
 1984
 1985
 1986
 1987
 1988
 1989
 1990
 1991
 1992
 1993
 1994
 1995
 1996
 1997
 1998
 1999
 2000
 2001
 2002
 2003
 2004
 2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024
 2025
 2026
 2027
 2028
 2029
 2030
 2031
 2032
 2033
 2034
 2035
 2036
 2037
 2038
 2039
 2040
 2041
 2042
 2043
 2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051
 2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060
 2061
 2062
 2063
 2064
 2065
 2066
 2067
 2068
 2069
 2070
 2071
 2072
 2073
 2074
 2075
 2076
 2077
 2078
 2079
 2080
 2081
 2082
 2083
 2084
 2085
 2086
 2087
 2088
 2089
 2090
 2091
 2092
 2093
 2094
 2095
 2096
 2097
 2098
 2099
 2100
 2101
 2102
 2103
 2104
 2105
 2106
 2107
 2108
 2109
 2110
 2111
 2112
 2113
 2114
 2115
 2116
 2117
 2118
 2119
 2120
 2121
 2122
 2123
 2124
 2125
 2126
 2127
 2128
 2129
 2130
 2131
 2132
 2133
 2134
 2135
 2136
 2137
 2138
 2139
 2140
 2141
 2142
 2143
 2144
 2145
 2146
 2147
 2148
 2149
 2150
 2151
 2152
 2153
 2154
 2155
 2156
 2157
 2158
 2159
 2160
 2161
 2162
 2163
 2164
 2165
 2166
 2167
 2168
 2169
 2170
 2171
 2172
 2173
 2174
 2175
 2176
 2177
 2178
 2179
 2180
 2181
 2182
 2183
 2184
 2185
 2186
 2187
 2188
 2189
 2190
 2191
 2192
 2193
 2194
 2195
 2196
 2197
 2198
 2199
 2200
 2201
 2202
 2203
 2204
 2205
 2206
 2207
 2208
 2209
 2210
 2211
 2212
 2213
 2214
 2215
 2216
 2217
 2218
 2219
 2220
 2221
 2222
 2223
 2224
 2225
 2226
 2227
 2228
 2229
 2230
 2231
 2232
 2233
 2234
 2235
 2236
 2237
 2238
 2239
 2240
 2241
 2242
 2243
 2244
 2245
 2246
 2247
 2248
 2249
 2250
 2251
 2252
 2253
 2254
 2255
 2256
 2257
 2258
 2259
 2260
 2261
 2262
 2263
 2264
 2265
 2266
 2267
 2268
 2269
 2270
 2271
 2272
 2273
 2274
 2275
 2276
 2277
 2278
 2279
 2280
 2281
 2282
 2283
 2284
 2285
 2286
 2287
 2288
 2289
 2290
 2291
 2292
 2293
 2294
 2295
 2296
 2297
 2298
 2299
 2300
 2301
 2302
 2303
 2304
 2305
 2306
 2307
 2308
 2309
 2310
 2311
 2312
 2313
 2314
 2315
 2316
 2317
 2318
 2319
 2320
 2321
 2322
 2323
 2324
 2325
 2326
 2327
 2328
 2329
 2330
 2331
 2332
 2333
 2334
 2335
 2336
 2337
 2338
 2339
 2340
 2341
 2342
 2343
 2344
 2345
 2346
 2347
 2348
 2349
 2350
 2351
 2352
 2353
 2354
 2355
 2356
 2357
 2358
 2359
 2360
 2361
 2362
 2363
 2364
 2365
 2366
 2367
 2368
 2369
 2370
 2371
 2372
 2373
 2374
 2375
 2376
 2377
 2378
 2379
 2380
 2381
 2382
 2383
 2384
 2385
 2386
 2387
 2388
 2389
 2390
 2391
 2392
 2393
 2394
 2395
 2396
 2397
 2398
 2399
 2400
 2401
 2402
 2403
 2404
 2405
 2406
 2407
 2408
 2409
 2410
 2411
 2412
 2413
 2414
 2415
 2416
 2417
 2418
 2419
 2420
 2421
 2422
 2423
 2424
 2425
 2426
 2427
 2428
 2429
 2430
 2431
 2432
 2433
 2434
 2435
 2436
 2437
 2438
 2439
 2440
 2441
 2442
 2443
 2444
 2445
 2446
 2447
 2448
 2449
 2450
 2451
 2452
 2453
 2454
 2455
 2456
 2457
 2458
 2459
 2460
 2461
 2462
 2463
 2464
 2465
 2466
 2467
 2468
 2469
 2470
 2471
 2472
 2473
 2474
 2475
 2476
 2477
 2478
 2479
 2480
 2481
 2482
 2483
 2484
 2485
 2486
 2487
 2488
 2489
 2490
 2491
 2492
 2493
 2494
 2495
 2496
 2497
 2498
 2499
 2500
 2501
 2502
 2503
 2504
 2505
 2506
 2507
 2508
 2509
 2510
 2511
 2512
 2513
 2514
 2515
 2516
 2517
 2518
 2519
 2520
 2521
 2522
 2523
 2524
 2525
 2526
 2527
 2528
 2529
 2530
 2531
 2532
 2533
 2534
 2535
 2536
 2537
 2538
 2539
 2540
 2541
 2542
 2543
 2544
 2545
 2546
 2547
 2548
 2549
 2550
 2551
 2552
 2553
 2554
 2555
 2556
 2557
 2558
 2559
 2560
 2561
 2562
 2563
 2564
 2565
 2566
 2567
 2568
 2569
 2570
 2571
 2572
 2573
 2574
 2575
 2576
 2577
 2578
 2579
 2580
 2581
 2582
 2583
 2584
 2585
 2586
 2587
 2588
 2589
 2590
 2591
 2592
 2593
 2594
 2595
 2596
 2597
 2598
 2599
 2600
 2601
 2602
 2603
 2604
 2605
 2606
 2607<br

1404 So we have

$$\begin{aligned}
 \tilde{R} &\stackrel{(D.6)}{\lesssim} c_v \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{k \in \mathcal{K}} \sum_h \bar{\sigma}_{k,h}^2} + d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\stackrel{(D.7)}{\lesssim} c_v \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R + \tilde{R})} + d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\lesssim c_v \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R)} + \max\{1, c_v^2\} d_{\mathcal{F}} \log N_{\mathcal{F}},
 \end{aligned} \tag{D.8}$$

1411 where the last inequality holds since $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ for any $x \geq 0$.

1412 And thus

$$\begin{aligned}
 R &\stackrel{(D.5)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{k \in \mathcal{K}} \sum_h \bar{\sigma}_{k,h}^2} + d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\stackrel{(D.7)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R + \tilde{R})} + d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R)} + d_{\mathcal{F}} \log N_{\mathcal{F}} + \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \tilde{R}},
 \end{aligned}$$

1420 where the last inequality holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Here

$$\begin{aligned}
 &\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \tilde{R}} \\
 &= \sqrt{\max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}} \cdot \frac{1}{\max\{1, c_v\}} \tilde{R}} \\
 &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R)} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}},
 \end{aligned}$$

1428 where the inequality holds due to Cauchy-Schwartz inequality and (D.8). Plugin back, we have

$$\begin{aligned}
 R &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (S_0 + R)} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\stackrel{(a)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (\text{Var}_K^* + \check{S}_0 + R)} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\stackrel{(C.15)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (\text{Var}_K^* + \check{A}_1 + R)} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\stackrel{(C.20)}{\lesssim} \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} (\text{Var}_K^* + \sqrt{R} + R)} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}} \\
 &\lesssim \sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \text{Var}_K^*} + \max\{1, c_v\} d_{\mathcal{F}} \log N_{\mathcal{F}},
 \end{aligned}$$

1439 where the (a) holds due to $\text{Var}[X + Y] \leq 2 \text{Var}[X] + 2 \text{Var}[Y]$. \square

E AUXILIARY LEMMAS

1443 **Lemma E.1** (Variance-aware and range-aware Freedman's inequality, Corollary 2 in Agarwal et al. (2023)). Let $M \geq m > 0, V \geq v > 0$ be fixed constants, and $\{X_s\}_{s \in [t]}$ be a stochastic process adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$, such that X_s is \mathcal{G}_s -measurable. Suppose $\mathbb{E}[X_s | \mathcal{G}_{s-1}] = 0, |X_s| \leq M$ and $\sum_{s=1}^t \mathbb{E}[X_s^2 | \mathcal{G}_{s-1}] \leq V^2$ almost surely. Then for any $\delta > 0$, with probability at least $1 - (\log(V^2/v^2) + 2)(\log(M/m) + 2)\delta$, we have

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

1452 **Lemma E.2** (Variance-aware Freedman's inequality). Let $M > 0$ be fixed constants, and $\{X_s\}_{s \in [t]}$ be a stochastic process adapted to the filtration $\{\mathcal{G}_s\}_{s \in [t]}$, such that X_s is \mathcal{G}_s -measurable. Suppose $\mathbb{E}[X_s | \mathcal{G}_{s-1}] = 0$ and $|X_s| \leq M$ almost surely. Then for any $\delta > 0$, with probability at least $1 - 2(\log t + 2)\delta$, we have

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

1458 *Proof.* The result follows by applying Lemma E.1 with $V^2 = M^2t, m = v = M$. \square
 1459

1460 **Lemma E.3.** Let $R, \alpha, \gamma = R\sqrt{\log N_{\mathcal{F}}}, L = \lceil \log_2 \frac{R}{\alpha} \rceil, \beta_{t,l} = \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}})$. For $t \in [T]$, let
 1461 disjoint sets $\{\Psi_{t+1,l}\}_{l \in \overline{[L]}}$ be constructed according to ADALEVEL with $D_{t,l} = D_{t,l}(X_t)$. Here,
 1462 $D_{t,l}(X_t) := D_{\mathcal{F}}(X_t; X_{\Psi_{t,l}}, w_{\Psi_{t,l}})$, where $D_{\mathcal{F}}$ is defined in (3.1). Then we have

1463

$$1464 \sum_{t \in [T]} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}} + \frac{\alpha^2 T}{\gamma} \sqrt{\log N_{\mathcal{F}}}\right).$$

1465

1466 Furthermore, setting $\alpha = R\sqrt{\frac{d_{\mathcal{F}} \log N_{\mathcal{F}}}{T}}$ yields
 1467

1468

$$1469 \sum_{t \in [T]} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} = \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}}\right).$$

1470

1471 *Proof.* According to Lemma E.4, we have for all $l \in [L]$,

1472

1473

$$1474 \sum_{t \in \Psi_{T+1,l}} \min\{1, w_t^2 D_{t,l}^2(X_t)\} = O\left(\dim_{\mathcal{F}}\left(\sqrt{\lambda/T}\right) \log T \log(T/\lambda)\right) = \tilde{O}(d_{\mathcal{F}}). \quad (\text{E.1})$$

1475

1476 For all $t \in [T]$, Property 1 holds true. Next, we decompose $[T] = \bigcup_{l \in \overline{[L]}} \Psi_{T+1,l}$ and carefully
 1477 bound summations within each level.

1478

1479 **Level $l = 0$** For any $t \in \Psi_{T+1,0}, D_{t,1}(X_t) \leq \frac{2\alpha}{\gamma}, \beta_{t,1} = \tilde{O}(2\alpha \sqrt{\log N_{\mathcal{F}}})$, therefore
 1480

1481

$$1482 \sum_{t \in \Psi_{T+1,0}} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} \leq \sum_{t \in \Psi_{T+1,0}} \beta_{t,1} D_{t,1}(X_t)$$

1483

$$1484 \lesssim T \cdot 2\alpha \sqrt{\log N_{\mathcal{F}}} \cdot \frac{2\alpha}{\gamma}$$

1485

$$1486 = \tilde{O}\left(\frac{\alpha^2 T}{R}\right).$$

1487

1488 **Level $l = L$** We decompose $\Psi_{T+1,L} = \mathcal{J}_{L,1} \cup \mathcal{J}_{L,2}$ where
 1489

1490

$$1491 \mathcal{J}_{L,1} := \left\{t \in \Psi_{T+1,L} : w_t = \frac{2^L \alpha}{\gamma D_{t,L}(X_t)}\right\}, \quad \mathcal{J}_{L,2} := \{t \in \Psi_{T+1,L} : w_t = 1\}.$$

1492

1493 Thus

1494

$$1495 \sum_{t \in \Psi_{T+1,L}} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} \leq R |\mathcal{J}_{L,1}| + \sum_{t \in \mathcal{J}_{L,2}} \beta_{t,L} D_{t,L}(X_t).$$

1496

1497 **Summation over $\mathcal{J}_{L,1}$** For any $t \in \mathcal{J}_{L,1}, w_t D_{t,L}(X_t) = \frac{2^L \alpha}{\gamma} \geq \frac{R}{\gamma}$. Thus $1 \leq \frac{\gamma^2}{R^2} w_t^2 D_{t,L}^2(X_t)$.
 1498 Then we have

1499

$$1500 R |\mathcal{J}_{L,1}| \leq R \sum_{t \in \mathcal{J}_{L,1}} \frac{\gamma^2}{R^2} w_t^2 D_{t,L}^2(X_t)$$

1501

$$1502 = \frac{\gamma^2}{R} \sum_{t \in \mathcal{J}_{L,1}} \min\{1, w_t^2 D_{t,L}^2(X_t)\}$$

1503

$$1504 \stackrel{(\text{E.1})}{=} \tilde{O}\left(R d_{\mathcal{F}} \log N_{\mathcal{F}}\right).$$

1505

1506 **Summation over $\mathcal{J}_{L,2}$** For any $t \in \mathcal{J}_{L,2}$,

1507

1508

$$1509 \beta_{t,L} = \tilde{O}(2^L \alpha \sqrt{\log N_{\mathcal{F}}}) \lesssim \tilde{O}(2\bar{\sigma}_t \log N_{\mathcal{F}}),$$

1510

$$1511 D_{t,L} \leq \frac{2^L \alpha}{\gamma}.$$

1512

1512 Thus

$$\begin{aligned}
\sum_{t \in \mathcal{J}_{L,2}} \beta_{t,L} D_{t,L}(X_t) &\lesssim \sum_{t \in \mathcal{J}_{L,2}} \bar{\sigma}_t \sqrt{\log N_{\mathcal{F}}} D_{t,L}(X_t) \\
&\lesssim \sqrt{\log N_{\mathcal{F}}} \cdot \sqrt{\sum_{t \in \mathcal{J}_{L,2}} \bar{\sigma}_t^2} \cdot \sqrt{\sum_{t \in \mathcal{J}_{L,2}} \min\{1, w_t^2 D_{t,L}^2(X_t)\}} \\
&\stackrel{(E.1)}{=} \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in \mathcal{J}_{L,2}} \bar{\sigma}_t^2}\right).
\end{aligned}$$

1521 **Level** $l \in [L-1]$ We decompose $\Psi_{T+1,l} = \mathcal{J}_{l,1} \cup \mathcal{J}_{l,2}$ where

$$\mathcal{J}_{l,1} := \left\{t \in \Psi_{T+1,l} : w_t = \frac{2^l \alpha}{\gamma D_{t,l}(X_t)}\right\}, \quad \mathcal{J}_{l,2} := \{t \in \Psi_{T+1,l} : w_t = 1\}.$$

1525 Thus

$$\sum_{t \in \Psi_{T+1,l}} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} \leq \sum_{t \in \mathcal{J}_{l,1}} \beta_{t,l+1} D_{t,l+1}(X_t) + \sum_{t \in \mathcal{J}_{l,2}} \beta_{t,l} D_{t,l}(X_t).$$

1529 **Summation over** $\mathcal{J}_{l,1}$ For any $t \in \mathcal{J}_{l,1}$,

$$\begin{aligned}
\beta_{t,l+1} &= \tilde{O}(2^{l+1} \alpha \sqrt{\log N_{\mathcal{F}}}) = \tilde{O}(4\gamma w_t D_{t,l}(X_t) \sqrt{\log N_{\mathcal{F}}}) \\
D_{t,l+1}(X_t) &\leq \frac{2^{l+1} \alpha}{\gamma} = 2w_t D_{t,l}(X_t).
\end{aligned}$$

1534 Thus

$$\begin{aligned}
\sum_{t \in \mathcal{J}_{l,1}} \beta_{t,l+1} D_{t,l+1}(X_t) &\lesssim \sum_{t \in \mathcal{J}_{l,1}} \gamma w_t D_{t,l}(X_t) \sqrt{\log N_{\mathcal{F}}} \cdot w_t D_{t,l}(X_t) \\
&\approx \gamma \sqrt{\log N_{\mathcal{F}}} \sum_{t \in \mathcal{J}_{l,1}} \min\{1, w_t^2 D_{t,l}^2(X_t)\} \\
&\stackrel{(E.4)}{=} \tilde{O}(R d_{\mathcal{F}} \log N_{\mathcal{F}}).
\end{aligned}$$

1542 **Summation over** $\mathcal{J}_{l,2}$ For any $t \in \mathcal{J}_{l,2}$,

$$\begin{aligned}
\beta_{t,l} &= \tilde{O}(2^l \alpha \sqrt{\log N_{\mathcal{F}}}) \lesssim \tilde{O}(2\bar{\sigma}_t \log N_{\mathcal{F}}), \\
D_{t,l} &\leq \frac{2^l \alpha}{\gamma}.
\end{aligned}$$

1548 Thus

$$\begin{aligned}
\sum_{t \in \mathcal{J}_{l,2}} \beta_{t,l} D_{t,l}(X_t) &\lesssim \sum_{t \in \mathcal{J}_{l,2}} \bar{\sigma}_t \sqrt{\log N_{\mathcal{F}}} D_{t,l}(X_t) \\
&\lesssim \sqrt{\log N_{\mathcal{F}}} \cdot \sqrt{\sum_{t \in \mathcal{J}_{l,2}} \bar{\sigma}_t^2} \cdot \sqrt{\sum_{t \in \mathcal{J}_{l,2}} \min\{1, w_t^2 D_{t,l}^2(X_t)\}} \\
&= \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in \mathcal{J}_{l,2}} \bar{\sigma}_t^2}\right).
\end{aligned}$$

1557 Now we put pieces together:

$$\begin{aligned}
\sum_{t \in [T]} \min_{l \in [L]} \{R, \beta_{t,l} D_{t,l}(X_t)\} &= \sum_{l \in [L]} \sum_{t \in \Psi_{T+1,l}} \min_{l' \in [L]} \{R, \beta_{t,l'} D_{t,l'}(X_t)\} \\
&= \tilde{O}\left(\sqrt{L d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{l \in [L]} \sum_{t \in \mathcal{J}_{l,2}} \bar{\sigma}_t^2} + L R d_{\mathcal{F}} \log N_{\mathcal{F}} + \frac{\alpha^2 T}{\gamma} \sqrt{\log N_{\mathcal{F}}}\right) \\
&= \tilde{O}\left(\sqrt{d_{\mathcal{F}} \log N_{\mathcal{F}} \sum_{t \in [T]} \bar{\sigma}_t^2} + R d_{\mathcal{F}} \log N_{\mathcal{F}} + \frac{\alpha^2 T}{\gamma} \sqrt{\log N_{\mathcal{F}}}\right).
\end{aligned}$$

□

1566

Lemma E.4 (Lemma D.6 in Jia et al. (2024)). Let $D_{\mathcal{F}}$ be defined in (3.1), and $w_t \in [0, 1]$ for all $t \in [T]$. Then we have

1568

1569

1570

1571

1572

1573

$$\sum_{t \in [T]} \min\{1, w_t^2 D_{\mathcal{F}}(X_t; X_{[t-1]}, w_{[t-1]})\} = O\left(\dim_{\mathcal{F}}\left(\sqrt{\lambda/T}\right) \log T \log(T/\lambda)\right).$$

Lemma E.5 (Lemma H.4 in Huang et al. (2024)). Let $D_{\mathcal{F}}$ be defined in (3.1). Then for any $t > t_0 \geq 1$, we have

1574

1575

1576

$$D_{\mathcal{F}}^2(X_t; X_{[t_0]}, w_{[t_0]}) \leq \exp\left\{\sum_{s=t_0+1}^{t-1} w_s^2 D_{\mathcal{F}}^2(X_s; X_{[s-1]}, w_{[s-1]})\right\} D_{\mathcal{F}}^2(X_t; X_{[t-1]}, w_{[t-1]}).$$

1577

1578

1579

1580

Lemma E.6 (Lemma H.6 in Huang et al. (2024)). Let $\lambda_1, \lambda_2, \lambda_4 > 0$, $\lambda_3 \geq 1$ and $i' = \lceil \log_2 \lambda_1 \rceil$. Let $a_0, a_1, a_2, \dots, a_{i'}$ be non-negative reals such that $a_i \leq \lambda_1$ for any $0 \leq i \leq i'$, and $a_i \leq \lambda_2 \sqrt{a_{i+1} + 2^{i+1} \cdot \lambda_3} + \lambda_4$ for any $0 \leq i < i'$. Then we have

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619