# Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity

Laixi Shi [1] Gen Li [2] Yuting Wei [2] Yuxin Chen [2] Yuejie Chi [1]

## Abstract

Offline or batch reinforcement learning seeks to learn a near-optimal policy using history data without active exploration of the environment. To counter the insufficient coverage and sample scarcity of many offline datasets, the principle of pessimism has been recently introduced to mitigate high bias of the estimated values. While pessimistic variants of model-based algorithms (e.g., value iteration with lower confidence bounds) have been theoretically investigated, their model-free counterparts — which do not require explicit model estimation — have not been adequately studied, especially in terms of sample efficiency. To address this inadequacy, we study a pessimistic variant of Q-learning in the context of finite-horizon Markov decision processes, and characterize its sample complexity under the single-policy concentrability assumption which does not require the full coverage of the state-action space. In addition, a variance-reduced pessimistic Q-learning algorithm is proposed to achieve near-optimal sample complexity. Altogether, this work highlights the efficiency of model-free algorithms in offline RL when used in conjunction with pessimism and variance reduction.

## 1. Introduction

Reinforcement Learning (RL) has achieved remarkable success in recent years, including matching or surpassing human performance in robotics control and strategy games (Silver et al., 2017; Mnih et al., 2015). Nevertheless, these success stories often come with nearly prohibitive cost, where an astronomical number of samples are required to train the learning algorithm to a satisfactory level. Scaling up and replicating the RL success in many real-world problems face considerable challenges, due to limited access to large-scale simulation data. In applications such as online advertising and clinical trials, real-time data collection could be expensive, time-consuming, or constrained in sample sizes as a result of experimental limitations.

On the other hand, it is worth noting that tons of samples might have already been accumulated and stored — albeit not necessarily with the desired quality — during previous data acquisition attempts. It is therefore natural to wonder whether such history data can be leveraged to improve performance in future deployments. In reality, the history data was often obtained by executing some (possibly unknown) behavior policy, which is typically not the desired policy. This gives rise to the problem of offline RL or batch RL (Lange et al., 2012; Levine et al., 2020),[1] namely, how to make the best use of history data to learn an improved or even optimal policy, without further exploring the environment. In stark contrast to online RL that relies on active interaction with the environment, the performance of offline RL depends critically not only on the quantity, but also the quality of history data (e.g., coverage over the space-action space), given that the agent is no longer collecting new samples for the purpose of exploring the unknown environment.

Recently, the principle of pessimism (or conservatism) — namely, being conservative in Q-function estimation when there are not enough samples — has been put forward as an effective way to solve offline RL (Buckman et al., 2020; Kumar et al., 2020). This principle has been implemented in, for instance, a model-based offline value iteration algorithm, which modifies classical value iteration (Azar et al., 2017) by subtracting a penalty term in the estimated Q-values and has been shown to achieve appealing sample efficiency (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b). It is noteworthy that the model-based approach is built upon the construction of an empirical transition kernel, and therefore, requires specific representation of the environment (see, e.g. Agarwal et al., 2020; Li et al., 2020). It remains

[1] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA [2] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Correspondence to: Laixi Shi <laixishi@cmu.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] Throughout this paper, we will be using the term offline RL (resp. dataset) or batch RL (resp. dataset) interchangeably.

| Algorithm | Type | Sample complexity |
|---|---|---|
| VI-LCB<br>(Xie et al., 2021b) | model-based | $\frac{H^6 S C^?}{"^2}$ |
| PEVI-Adv<br>(Xie et al., 2021b) | model-based | $\frac{H^4 S C^?}{"^2}$ |
| Q-LCB<br>(this work) | model-free | $\frac{H^6 S C^?}{"^2}$ |
| Q-LCB-Adv<br>(this work) | model-free | $\frac{H^4 S C^?}{"^2}$ |
| lower bound<br>(Xie et al., 2021b) | n/a | $\frac{H^4 S C^?}{"^2}$ |

Table 1. Comparisons between our results and prior art for finding an "-optimal policy in finite-horizon non-stationary MDPs. The sample complexities included in the table are valid for sufficiently small ", with all logarithmic factors omitted.

unknown whether the pessimism principle can be incorporated into model-free algorithms — another class of popular algorithms that performs learning without model estimation — in a provably effective fashion for offline RL.

### 1.1. Main contributions

In this paper, we consider finite-horizon non-stationary Markov decision processes (MDPs) with $S$ states, $A$ actions, and horizon length $H$. The focal point is to pin down the sample efficiency for pessimistic variants of model-free algorithms, under the mild single-policy concentrability assumption (cf. Assumption 2.1) of the batch dataset introduced in Rashidinejad et al. (2021); Xie et al. (2021b) (in short, this assumption captures how close the batch dataset is to an expert dataset, and will be formally introduced in Section 2.2). Given $K$ episodes of history data each of length $H$ (which amounts to a total number of $T = K H$ samples), our main contributions are summarized as follows.

We first study a natural pessimistic variant of the Q-learning algorithm, which simply modifies the classical Q-learning update rule by subtracting a penalty term (via certain lower confidence bounds). We prove that pessimistic Q-learning finds an "-optimal policy as soon as the sample size $T$ exceeds the order of (up to log factor)

$$\frac{H^6 S C^?}{"^2};$$

where $C^?$ denotes the single-policy concentrability coefficient of the batch dataset. In comparison to the minimax lower bound $\frac{H^4 S C^?}{"^2}$ developed in Xie et al. (2021b), the sample complexity of pessimistic Q-learning is at most a factor of $H^2$ from optimal (modulo some log factor).

To further improve the sample efficiency of pessimistic model-free algorithms, we introduce a variance-reduced variant of pessimistic Q-learning. This algorithm is guaranteed to find an "-optimal policy as long as the sample size $T$ is above the order of

$$\frac{H^4 S C^?}{"^2} + \frac{H^5 S C^?}{"}$$

up to some log factor. In particular, this sample complexity is minimax-optimal (namely, as low as $\frac{H^4 S C^?}{"^2}$ up to log factor) for small enough " (namely, " (0; 1=H]). The "-range that enjoys near-optimality is much larger compared to " $0; 1=H^{2:5}$ established in Xie et al. (2021b) for model-based algorithms.

Both of the proposed algorithms achieve low computation cost (i.e., $O(T)$) and low memory complexities (i.e., $O(\min fT; SAHg)$). Additionally, more complete comparisons with prior sample complexities of pessimistic model-based algorithms (Xie et al., 2021b) are provided in Table 1. In comparison with model-based algorithms, model-free algorithms require drastically different technical tools to handle the complicated statistical dependency between the estimated Q-values at different time steps.

### 1.2. Related works

In this section, we discuss several lines of works which are related to ours, with an emphasis on value-based algorithms for tabular settings with finite state and action spaces.

Offline RL. One of the key challenges in offline RL lies in the insufficient coverage of the batch dataset, due to lack of interaction with the environment (Levine et al., 2020; Liu et al., 2020). To address this challenge, most of the recent works can be divided into two lines: 1) regularizing the policy to avoid visiting under-covered state and action pairs (Fujimoto et al., 2019; Dadashi et al., 2021); 2) penalizing the estimated values of the under-covered state-action pairs (Buckman et al., 2020; Kumar et al., 2020). Our work follows the latter line (also known as the principle of pessimism), which has garnered significant attention recently. In fact, pessimism has been incorporated into recent development of various offline RL approaches, such as policy-based approaches (Rezaeifar et al., 2021; Xie et al., 2021a; Zanette et al., 2021), model-based approaches (Rashidinejad et al., 2021; Uehara & Sun, 2021; Jin et al., 2021; Yu et al., 2020; Kidambi et al., 2020; Xie et al., 2021b; Yin & Wang, 2021; Uehara et al., 2021; Yan et al., 2022b; Yu et al., 2021b; Yin et al., 2022), and model-free approaches (Kumar et al., 2020; Yu et al., 2021a; Yan et al., 2022a).

Finite-sample guarantees for pessimistic approaches. While model-free approaches with pessimism (Kumar et al., 2020; Yu et al., 2021a) have achieved considerable empirical successes in offline RL, prior theoretical guarantees of pessimistic schemes have been confined almost exclusively to model-based approaches. Under the same single-policy concentrability assumption used in prior analyses of model-based approaches (Rashidinejad et al., 2021; Xie et al., 2021b; Yin et al., 2021b), the current paper provides the first finite-sample guarantees for model-free approaches with pessimism in the tabular case without explicit model construction. In addition, Yin & Wang (2021) directly employed the occupancy distributions of the behavior policy and the optimal policy in bounding the performance of a model-based approach, rather than the worst-case upper bound of their ratios as done under the single-policy concentrability assumption.

**Non-asymptotic guarantees for variants of Q-learning.** Q-learning, which is among the most famous model-free RL algorithms (Watkins, 1989; Jaakkola et al., 1994; Watkins & Dayan, 1992), has been adapted in a multitude of ways to deal with different RL settings. Theoretical analyses for Q-learning and its variants have been established in, for example, the online setting via regret analysis (Jin et al., 2018; Bai et al., 2019; Zhang et al., 2020b; Li et al., 2021b; Dong et al., 2019; Zhang et al., 2020a;c; Jafarnia-Jahromi et al., 2020; Yang et al., 2021), and the simulator setting via probably approximately correct (PAC) bounds (Chen et al., 2020; Wainwright, 2019; Li et al., 2021a). The variant that is most closely related to ours is asynchronous Q-learning, which aims to find the optimal Q-function from Markovian trajectories following some behavior policy (Even-Dar & Mansour, 2003; Beck & Srikant, 2012; Qu & Wierman, 2020; Li et al., 2021c; Yin et al., 2021a;b). Different from ours, these works typically require full coverage of the state-action space by the behavior policy, a much stronger assumption than the single-policy concentrability assumed in our offline RL setting.

**Variance reduction in RL.** Variance reduction, originally proposed to accelerate stochastic optimization (e.g., the SVRG algorithm proposed by Johnson & Zhang (2013)), has been successfully leveraged to improve the sample efficiency of various RL algorithms, including but not limited to policy evaluation (Du et al., 2017; Wai et al., 2019; Xu et al., 2019; Khamaru et al., 2020), planning (Sidford et al., 2018a;b), Q-learning and its variants (Wainwright, 2019; Zhang et al., 2020b; Li et al., 2021b;c; Yan et al., 2022a), and offline RL (Xie et al., 2021b; Yin et al., 2021b).

### 1.3. Notation and paper organization

Let us introduce a set of notation that will be used throughout. We denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$, and introduce the notation $[N] := \{1, \cdots, N\}$ for any integer $N > 0$. For any vector $x \in \mathbb{R}^{SA}$ (resp. $x \in \mathbb{R}^{S}$) that constitutes certain values for each of the state-action pairs (resp. state), we shall often use $x(s,a)$ (resp. $x(s)$) to denote the entry associated with the $(s,a)$ pair (resp. state $s$). Similarly, we shall denote by $\boldsymbol{x} := \{x_h\}_{h\in[H]}$ the set composed of certain vectors for each of the time step $h \in [H]$. We let $e_i$ represent the $i$-th standard basis vector, with the only non-zero element being in the $i$-th entry.

Let $\mathcal{X} := (S, A, H, T)$. The notation $f(\mathcal{X}) \lesssim g(\mathcal{X})$ (resp. $f(\mathcal{X}) \gtrsim g(\mathcal{X})$) means that there exists a universal constant $C_0 > 0$ such that $|f(\mathcal{X})| \le C_0|g(\mathcal{X})|$ (resp. $|f(\mathcal{X})| \ge C_0|g(\mathcal{X})|$). In addition, we often overload scalar functions and expressions to take vector-valued arguments, with the interpretation that they are applied in an entrywise manner. For example, for a vector $x = [x_i]_{1\le i\le n}$, we have $x^2 = [x_i^2]_{1\le i\le n}$. For any two vectors $x = [x_i]_{1\le i\le n}$ and $y = [y_i]_{1\le i\le n}$, the notation $x \le y$ (resp. $x \ge y$) means $x_i \le y_i$ (resp. $x_i \ge y_i$) for all $1 \le i \le n$.

**Paper organization.** The rest of this paper is organized as follows. Section 2 introduces the backgrounds on finite-horizon MDPs and formulates the offline RL problem. Section 3 starts by introducing a natural pessimistic variant of Q-learning along with its sample complexity bound, and further enhances the sample efficiency via variance reduction in Section 4. Section A presents the proof outline and key lemmas. Finally, we conclude in Section 5 with a discussion and defer the proof details to the supplementary material.

## 2. Background and problem formulation

### 2.1. Tabular finite-horizon MDPs

**Basics.** This work focuses on an episodic finite-horizon MDP as represented by

$$\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H\big),$$

where $H$ is the horizon length, $\mathcal{S}$ is a finite state space of cardinality $S$, $\mathcal{A}$ is a finite action space of cardinality $A$, and $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ (resp. $r_h : \mathcal{S}\times\mathcal{A} \to [0,1]$) represents the probability transition kernel (resp. reward function) at the $h$-th time step ($1 \le h \le H$). Throughout this paper, we shall adopt the following convenient notation

$$P_{h,s,a} := P_h(\cdot \,|\, s,a) \in [0,1]^{1\times S}, \tag{1}$$

which stands for the transition probability vector given the current state-action pair $(s,a)$ at time step $h$. The parameters $S$, $A$ and $H$ can all be quite large, allowing one to capture the challenges arising in MDPs with large state/action space and long horizon.

A policy (or action selection rule) of an agent is represented by $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the associated selection probability over the action space at time step h (or more precisely, we let $_h(a\,j\,s)$ represent the probability of selecting action a in state s at step h). When is a deterministic policy, we abuse the notation and let $_h(s)$ denote the action selected by policy in state s at step h. In each episode, the agent generates an initial state $s_1$ 2 S drawn from an initial state distribution 2 (S), and rolls out a trajectory over the MDP by executing a policy as follows:

$$fs_h; a_h; r_h g_{h=1}^H = fs_1; a_1; r_1; :::; s_H; a_H; r_H g; \qquad (2)$$

where at time step h, $a_h$ $_h(j\,s_h)$ indicates the action selected in state $s_h$, $r_h = r_h(s_h; a_h)$ denotes the deterministic immediate reward, and $s_{h+1}$ denotes the next state drawn from the transition probability vector $P_{h;s\ ;a}$ := $P_h(j\,s_h; a_h)$. In addition, let $d_h(s)$ and $d_h(s; a)$ denote respectively the occupancy distribution induced by at time step h 2 [H], namely,

$$d_h(s) := P(s_h = s\,j\,s_1 \ ;);$$
$$d_h(s; a) := P(s_h = s\,j\,s_1 \ ;)\,_h(a\,j\,s); \qquad (3)$$

here and throughout, we denote [H] := f1; ; Hg. Given that the initial state $s_1$ is drawn from , the above definition gives

$$d_1(s) = (s) \qquad \text{for any policy :} \qquad (4)$$

**Value function, Q-function, and optimal policy.** The value function $V_h(s)$ of policy in state s at step h is defined as the expected cumulative rewards when this policy is executed starting from state s at step h, i.e.,

$$V_h(s) := E\left[\sum_{t=h}^{H} r_t\ \ s_t; a_t\ \ s_h = s\right]; \qquad (5)$$

where the expectation is taken over the randomness of the trajectory (2) induced by the policy as well as the MDP transitions. Similarly, the Q-function $Q_h(;)$ of a policy at step h is defined as

$$Q_h(s; a) := r_h(s; a)$$
$$+\ E\left[\sum_{t=h+1}^{H} r_t(s_t; a_t)\ s_h = s; a_h = a\right]; \qquad (6)$$

where the expectation is again over the randomness induced by and the MDP except that the state-action pair at step h is now conditioned to be (s; a). By convention, we shall also set

$$V_{H+1}(s) = Q_{H+1}(s; a) = 0 \quad \text{for any and } (s; a)\ 2\ S\ A: \qquad (7)$$

A policy $^? = f^?_h g_{h=1}^H$ is said to be an optimal policy if it maximizes the value function (resp. Q-function) simultaneously for all states (resp. state-action pairs) among all policies, whose existence is always guaranteed (Puterman, 2014). The resulting optimal value function $V^? = fV_h^? g_{h=1}^H$ and optimal Q-functions $Q^? = fQ_h^? g_{h=1}^H$ are denoted respectively by

$$V_h^?(s) := V_h^{\ ?}(s) = \max V_h(s);$$
$$Q_h^?(s; a) := Q_h^{\ ?}(s; a) = \max Q_h(s; a)$$

for any (s; a; h) 2 S A [H]. Throughout this paper, we assume that $^?$ is a deterministic optimal policy, which always exists (Puterman, 2014).

Additionally, when the initial state is drawn from a given distribution , the expected value of a given policy and that of the optimal policy at the initial step are defined respectively by

$$V_1() = : \underset{s_1}{E} V_1(s_1);$$
$$V_1^?() = : \underset{s_1}{E} V_1^?(s_1) : \qquad (8)$$

**Bellman equations.** The Bellman equations play a fundamental role in dynamic programming (Bertsekas, 2017). Specifically, the value function and the Q-function of any policy satisfy the following Bellman consistency equation:

$$Q_h(s; a) = r_h(s; a) + \underset{s^0 P_{h;s;a}}{E} V_{h+1}(s^0) \qquad (9)$$

for all (s; a; h) 2 S A [H]. Moreover, the optimal value function and the optimal Q-function satisfy the Bellman optimality equation:

$$Q_h^?(s; a) = r_h(s; a) + \underset{s^0 P_{h;s;a}}{E} V_{h+1}^?(s^0) \qquad (10)$$

for all (s; a; h) 2 S A [H].

## 2.2. Offline RL under single-policy concentrability

Offline RL assumes the availability of a history dataset D containing K episodes each of length H. These episodes are independently generated based on a certain policy $= f_h g_{h=1}^H$ — called the behavior policy, resulting in a dataset

$$D = :\left\{ s_1^k; a_1^k; r_1^k; :::; s_H^k; a_H^k; r_H^k \right\}_{k=0}^{K-1}:$$

Here, the initial states $fs_1^k g_{k=1}^K$ are independently drawn from 2 (S) such that $s_1^k \overset{i:i:d}{\ }$ , while the remaining states and actions are generated by the MDP induced by the behavior policy . The total number of samples is thus given by

$$T = KH:$$

With the notation (8) in place, the goal of offline RL amounts to finding an "-optimal policy $b = fb_h g_{h=1}^H$ satisfying

$$V_1^?() \quad V_1^b() \quad "$$

with as few samples as possible, and ideally, in a computationally fast and memory-efficient manner.

Obviously, efficient offline RL cannot be accomplished without imposing proper assumptions on the behavior policy, which also provide means to gauge the difficulty of the offline RL task through the quality of the history dataset. Following the recent works Rashidinejad et al. (2021); Xie et al. (2021b), we assume that the behavior policy satisfies the following property called single-policy concentrability.

Assumption 2.1 (single-policy concentrability). The single-policy concentrability coefficient $C^?$ 2 $[1; 1)$ of a behavior policy is defined to be the smallest quantity that satisfies

$$\max_{(h;s;a)2[H]SA} \frac{d_h^?(s;a)}{d_h(s;a)} \quad C^?; \tag{11}$$

where we adopt the convention 0=0 = 0.

Intuitively, the single-policy concentrability coefficient measures the discrepancy between the optimal policy $^?$ and the behavior policy in terms of the resulting density ratio of the respective occupancy distributions. It is noteworthy that a finite $C^?$ does not necessarily require to cover the entire state-action space; instead, it can be attainable when its coverage subsumes that of the optimal policy $^?$. This is in stark contrast to, and in fact much weaker than, other assumptions that require either full coverage of the behavior policy (i.e., $\min_{(h;s;a)2[H]SA} d_h(s;a) > 0$ (Li et al., 2021c; Yin et al., 2021a;b)), or uniform concentrability over all possible policies (Chen & Jiang, 2019). Additionally, the single-policy concentrability coefficient is minimized (i.e., $C^? = 1$) when the behavior policy coincides with the optimal policy $^?$, a scenario closely related to imitation learning or behavior cloning (Rajaraman et al., 2020).

## 3. Pessimistic Q-learning: algorithms and theory

In the current paper, we present two model-free algorithms — namely, LCB-Q and LCB-Q-Advantage — for offline RL, along with their respective theoretical guarantees. The first algorithm can be viewed as a pessimistic variant of the classical Q-learning algorithm, while the second one further leverages the idea of variance reduction to boost the sample efficiency. In this section, we begin by introducing LCB-Q.

### 3.1. LCB-Q: a natural pessimistic variant of Q-learning

Before proceeding, we find it convenient to first review the classical Q-learning algorithm (Watkins, 1989; Watkins & Dayan, 1992), which can be regarded as a stochastic approximation scheme to solve the Bellman optimality equation (10). Upon receiving a sample transition $(s_h; a_h; r_h; s_{h+1})$ at time step h, Q-learning updates the corresponding entry in the Q-estimate as follows

$$Q_h(s_h; a_h) \quad (1 \quad )Q_h(s_h; a_h) + \quad \text{n} \; r_h(s_h; a_h) + V_{h+1}(s_{h+1}) \; \text{o} ; \tag{12}$$

where $Q_h$ (resp. $V_h$) indicates the running estimate of $Q_h^?$ (resp. $V_h^?$), and $0 < \quad < 1$ is the learning rate. In comparison to model-based algorithms that require estimating the probability transition kernel based on all the samples, Q-learning, as a popular kind of model-free algorithms, is simpler and enjoys more flexibility without explicitly constructing the model of the environment. The wide applicability of Q-learning motivates one to adapt it to accommodate offline RL.

Inspired by recent advances in incorporating the pessimism principle for offline RL (Rashidinejad et al., 2021; Jin et al., 2021), we study a pessimistic variant of Q-learning called LCB-Q, which modifies the Q-learning update rule as follows

$$Q_h(s_h; a_h) \quad (1 \quad _n)Q_h(s_h; a_h) + \, _n \; \text{n} \; r_h(s_h; a_h) + V_{h+1}(s_{h+1}) \quad b_n \; \text{o} ; \tag{13}$$

where $_n$ is the learning rate depending on the number of times n that the state-action pair $(s_h; a_h)$ has been visited at step h, and the penalty term $b_n > 0$ (cf. line 9 of Algorithm 1) reflects the uncertainty of the corresponding Q-estimate and implements pessimism in the face of uncertainty. The entire algorithm, which is a single-pass algorithm that only requires reading the offline dataset once, is summarized in Algorithm 1.

### 3.2. Theoretical guarantees for LCB-Q

The proposed LCB-Q algorithm manages to achieve an appealing sample complexity as formalized by the following theorem.

Theorem 3.1. Consider any 2 (0; 1). Suppose that the behavior policy satisfies Assumption 2.1 with single-policy concentrability coefficient $C^?$ 1. Let $c_b$ 0 be some sufficiently large constant, and take = log $\frac{SAT}{}$. Assume that $T > SC^?$, then the policy b returned by Algorithm 1 satisfies

$$V_1^?() \quad V_1^b() \quad c_a \sqrt{\frac{H^6SC^{?3}}{T}} \tag{14}$$

with probability at least 1 , where $c_a > 0$ is some universal constant.

As asserted by Theorem 3.1, the LCB-Q algorithm is guaranteed to find an "-optimal policy with high probability, as long as the total sample size T = K H exceeds

$$\text{Q}e \; \frac{H^6SC^?}{"^2} ; \tag{15}$$

Algorithm 1 LCB-Q for offline RL

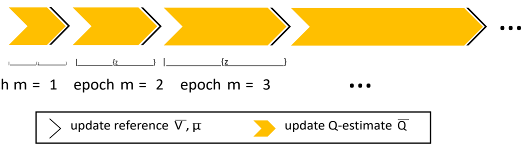
```
1: Parameters: some constant $c_b$ > 0, target success probability 1   2 (0; 1), and = log $\frac{SAT}{}$ .
2: Initialize: $Q_h(s; a)$   0; $N_h(s; a)$   0 for all (s; a; h) 2 S A [H]; $V_h(s)$   0 for all (s; h) 2 S [H + 1]; b s.t. $b_h(s)$ = 1 for all (h; s) 2 [H] S.
3: for Episode k = 1 to K do
4:     // sampling from batch dataset
       Sample a trajectory f$s_h$; $a_h$; $r_h$g$_{h=1}^{H}$ from D.
5:     for Step h = 1 to H do
6:         // update the counter
7:         $N_h(s_h; a_h)$   $N_h(s_h; a_h)$ + 1; n   $N_h(s_h; a_h)$.
8:         $_n$   $\frac{H+1}{H+n}$. // update the learning rate
9:         $b_n$   $c_b$ $\sqrt{\frac{H^{3}2}{n}}$. // update the bonus term
10:        // update the Q-estimates with LCB
11:        $Q_h(s_h; a_h)$   $Q_h(s_h; a_h)$ + $_n$ n $r_h(s_h; a_h)$ + $V_{h+1}(s_{h+1})$   $Q_h(s_h; a_h)$   $b_n$ o :
12:        // update the value estimates
13:        $V_h(s_h)$   max n $V_h(s_h)$; $\max_a Q_h(s_h; a)$ o .
14:        If $V_h(s_h)$ = $\max_a Q_h(s_h; a)$: update $b_h(s)$   arg $\max_a Q_h(s; a)$.
15:    end for
16: end for
17: Output: the policy b.
```

where Θ() hides logarithmic dependencies. When the behavior policy is close to the optimal policy, the single-policy concentrability coefficient $C^{?}$ is closer to 1; if this is the case, then our bound indicates that the sample complexity does not depend on the size A of the action space, which can be a huge saving when the action space is enormous.

Comparison with model-based pessimistic approaches. A model-based approach — called Value Iteration with Lower Confidence Bounds (VI-LCB) — has been recently proposed for offline RL (Rashidinejad et al., 2021; Xie et al., 2021b). In the finite-horizon case, VI-LCB incorporates an additional LCB penalty into the classical value iteration algorithm, and updates all the entries in the Q-estimate simultaneously as follows

$$Q_h(s; a) \quad r_h(s; a) + \widehat{P}_{h;s;a} V_{h+1} \quad b_h(s; a); \tag{16}$$

with the aim of tuning down the confidence on those state-action pairs that have only been visited infrequently. Here, $\widehat{P}_{h;s;a}$ represents the empirical estimation of the transition kernel $P_{h;s;a}$, and $b_h(s; a) > 0$ is chosen to capture the uncertainty level of $(\widehat{P}_{h;s;a} \quad P_{h;s;a})V_{h+1}$. Working backward, the algorithm estimates the Q-value $Q_h$ recursively over the time steps h = H; H 1; ;1. In comparison with VI-LCB, our sample complexity bound for LCB-Q matches the bound developed for VI-LCB by Xie et al. (2021b), while enjoying enhanced flexibility without the need of specifying the transition kernel of the environment (as model estimation might potentially incur a higher memory burden).



epoch m = 1
epoch m = 2
epoch m = 3
...
update reference $\overline{V}$, μ
update Q-estimate $\overline{Q}$

Figure 1. An illustration of the epoch-based LCB-Q-Advantage algorithm.

## 4. LCB-Q-Advantage for near-optimal offline RL: algorithm and theory

The careful reader might notice that the sample complexity (15) derived for LCB-Q remains a factor of $H^2$ away from the minimax lower bound (see Table 1). To further close the gap and improve the sample complexity, we propose a new variant called LCB-Q-Advantage, which leverages the idea of variance reduction to accelerate convergence (Johnson & Zhang, 2013; Sidford et al., 2018b; Wainwright, 2019; Zhang et al., 2020b; Xie et al., 2021b; Li et al., 2021c;b).

Inspired by the reference-advantage decomposition adopted in (Zhang et al., 2020b; Li et al., 2021b) for online Q-learning, LCB-Q-Advantage maintains a collection of reference values f$\overline{V}_h$g$_{h=1}^{H}$, which serve as running proxy for the optimal values f$V_h^?$g$_{h=1}^{H}$ and allow for reduced variability in each iteration. To be more specific, the LCB-Q-Advantage algorithm (cf. Algorithm 2 as well as the subroutines in Algorithm 3 that closely resemble Li et al. (2021b)) proceeds in an epoch-based style (the m-th epoch consists of $L_m = 2^m$ episodes of samples), where the reference values are updated at the end of each epoch to be used in the next epoch, and the Q-estimates are iteratively updated during the remaining time of each epoch. By maintaining two auxiliary sequences of pessimistic Q-estimates — that is, $Q^{LCB}$ constructed by the pessimistic Q-learning update, and $\overline{Q}$ constructed by the pessimistic Q-learning update based on the reference-advantage decomposition — the Q-estimate is updated by taking the maximum over the three candidates (cf. line 19 of Algorithm 2)

$$Q_h(s; a) \quad \max f Q_h^{LCB}(s; a); \overline{Q}_h(s; a); Q_h(s; a)g \tag{17}$$

when the state-action pair (s; a) is visited at the step h. We now take a moment to discuss the key ingredients of the proposed algorithm in further detail.

Updating the references $\overline{V}_h$ and $\overline{}_h$. At the end of each epoch, the reference values f$\overline{V}_h$g$_{h=1}^{H}$, as well as the associated running average f$\overline{}_h$g$_{h=1}^{H}$, are determined using what happens during the current epoch. More specifically, the

**Algorithm 2** Offline LCB-Q-Advantage RL

```
1:  Parameters: number of epochs M, universal constant $c_b > 0$, target success probability $1 \quad 2\ (0;1)$, and $= \log \frac{SAT}{}$;
2:  Initialize: $Q_h(s;a); Q^{LCB}_h(s;a); \overline{Q}_h(s;a); _h(s;a), {}^{next}_h(s;a), N_h(s;a) \quad 0$ for all $(s;a;h)\ 2\ S \quad A \quad [H]$;
    $V_h(s); \overline{V}_h(s); \overline{V}^{next}_h(s) \quad 0$ for all $(s;h)\ 2\ S \quad [H+1]$;
    ${}^{ref}_h(s;a); {}^{ref}_h(s;a); {}^{adv}_h(s;a); {}^{adv}_h(s;a); _h(s;a), B_h(s;a)$ 0 for all $(s;a;h)\ 2\ S \quad A \quad [H]$;
3:  for Epoch m = 1 to M do
4:      $L_m = 2^m$; // specify the number of episodes in the current epoch
5:      $\hat{N}_h(s;a) = 0$ for all $(h;s;a)\ 2\ [H]\ S \quad A$: // reset the epoch-wise counter
6:      /* Inner-loop: update value-estimates $V_h(s;a)$ and Q-estimates $Q_h(s;a)$
7:      for In-epoch Episode t = 1 to $L_m$ do
8:          Sample a trajectory $fs_h; a_h; r_h g^H_{h=1}$. // sampling
9:          for Step h = 1 to H do
10:             // update the overall counter
11:             $N_h(s_h;a_h) \quad N_h(s_h;a_h) + 1; n \quad N_h(s_h;a_h)$.
12:             $_n \quad \frac{H+1}{H+n}$; // update the learning rate
13:             // update the Q-estimate with LCB
14:             $Q^{LCB}_h(s_h;a_h) \quad$ update-lcb-q();
15:             // update the Q-estimate with LCB and reference-advantage
16:             $\overline{Q}_h(s_h;a_h) \quad$ update-lcb-q-ra();
17:             // update the estimates $Q_h$ and $V_h$
18:             $Q_h(s_h;a_h)$
19:                 $\max fQ^{LCB}_h(s_h;a_h); \overline{Q}_h(s_h;a_h); Q_h(s_h;a_h)g$:
20:             $V_h(s_h) \quad \max_a Q_h(s_h;a)$.
21:             // update the epoch-wise counter and ${}^{next}_h$ for the next epoch
22:             $\hat{N}_h(s_h;a_h) \quad \hat{N}_h(s_h;a_h) + 1$;
23:             ${}^{next}_h(s_h;a_h) \quad 1 \quad \frac{1}{\hat{N}_h(s_h;a_h)} \quad {}^{next}_h(s_h;a_h) + \frac{1}{\hat{N}_h(s_h;a_h)} V^{next}_{h+1}(s_{h+1})$.
24:         end for
25:     end for
26:     for $(s;a;h)\ 2\ S \quad A \quad [H+1]$ do
27:         // set $\overline{V}_h$ and $\overline{\ }_h$ for the next epoch
28:         $\overline{V}_h(s) \quad \overline{V}^{next}_h(s); \ \overline{\ }_h(s;a) \quad {}^{next}_h(s;a)$.
29:         // restart ${}^{next}_h$ and set $\overline{V}^{next}_h$ for the next epoch
30:         $\overline{V}^{next}_h(s) \quad V_h(s); \ {}^{next}_h(s;a) \quad 0$.
31:     end for
32: end for
33: Output: the policy b s.t. $b_h(s) = \arg\max_a Q_h(s;a)$ for any $(s;h)\ 2\ S \quad [H]$.
```

following update rules for $\overline{V}_h$ and $\overline{\ }_h$ are carried out at the end of the m-th epoch:

$$\overline{V}_h(s) \quad \overline{V}^{next}_h(s); \tag{18a}$$

**Algorithm 3** Auxiliary functions

```
1:  Function update-lcb-q():
2:      $Q^{LCB}_h(s_h;a_h) \quad (1 \quad _n)Q^{LCB}_h(s_h;a_h) + {}_n \ r(s_h;a_h) + V_{h+1}(s_{h+1}) \quad c_b \sqrt{\frac{H^3 \ ^2}{n}}$.
3:  Function update-lcb-q-ra():
        /* update the moment statistics of the interested terms
4:      $[{}^{ref}_h; {}^{ref}_h; {}^{adv}_h; {}^{adv}_h](s_h;a_h) \quad$ update-moments();
        /* update the bonus difference and accumulative bonus
5:      $[_h; \overline{B}_h](s_h;a_h) \quad$ update-bonus();
6:      $\overline{b}_h(s_h;a_h) \quad \overline{B}_h(s_h;a_h) + (1 \quad _n) \frac{_h}{(h\ a\ )_n} + c_b \frac{H^{7=4}}{n^{3=4}} + c_b \frac{H^2}{n}$;
        // update the Q-estimate based on reference-advantage
7:      $\overline{Q}_h(s_h;a_h) \quad (1 \quad _n)\overline{Q}_h(s_h;a_h) + {}_n \ r_h(s_h;a_h) + V_{h+1}(s_{h+1}) \quad \overline{V}_{h+1}(s_{h+1}) + \overline{\ }_h(s_h;a_h) \quad \overline{b}_h$;
8:  Function update-moments():
9:      ${}^{ref}_h(s_h;a_h) \quad (1 \quad \frac{1}{n})^{ref}_h(s_h;a_h) + \frac{1}{n}\overline{V}^{next}_{h+1}(s_{h+1})$;
        // mean of the reference
10:     ${}^{ref}_h(s_h;a_h) \quad (1 \quad \frac{1}{n})^{ref}_h(s_h;a_h) + \frac{1}{n}\overline{V}^{next}_{h+1}(s_{h+1})^{\ 2}$; // 2nd moment of the reference
11:     ${}^{adv}_h(s_h;a_h) \quad (1 \quad _n)^{adv}_h(s_h;a_h) + {}_n \ V_{h+1}(s_{h+1}) \quad \overline{V}_{h+1}(s_{h+1})$; // mean of the advantage
12:     ${}^{adv}_h(s_h;a_h) \quad (1 \quad _n)^{adv}_h(s_h;a_h) + {}_n \ V_{h+1}(s_{h+1}) \quad \overline{V}_{h+1}(s_{h+1})^{\ 2}$. // 2nd moment of the advantage
13: Function update-bonus():
14:     $B^{next}_h(s_h;a_h) \quad c_b \frac{p}{n} \sqrt{{}^{ref}_h(s_h;a_h) \quad {}^{ref}_h(s_h;a_h)^{\ 2}} + \ p\overline{H} \sqrt{{}^{adv}_h(s_h;a_h) \quad {}^{adv}_h(s_h;a_h)^{\ 2}}$;
15:     $_h(s_h;a_h) \quad B^{next}_h(s_h;a_h) \quad \overline{B}_h(s_h;a_h)$;
16:     $\overline{B}_h(s_h;a_h) \quad B^{next}_h(s_h;a_h)$:
```

$$\overline{\ }_h(s;a) \quad \frac{\sum_{t=1}^{L_m} 1(s^t_h = s; a^t_h = a)V_{h+1}(s^t_{h+1})}{\max \ n \sum_{t=1}^{L_m} 1(s^t_h = s; a^t_h = a) \ ;1 \ o} \tag{18b}$$

for all $(h;s;a)\ 2\ [H]\ S \quad A$. Here, $V_h(s)$ is assigned by $\overline{V}^{next}_h(s)$, which is maintained as the value estimate $V_h(s)$ at the end of the $(m \quad 1)$-th epoch, and the update of $\overline{\ }_h(s;a)$ is implemented in a recursive manner in the current m-th epoch. See also line 28 and line 30 of Algorithm 2.

**Learning Q-estimate $\overline{Q}_h$ based on the reference-advantage decomposition.** Armed with the references $\overline{V}_h$ and $\overline{\ }_h$ updated at the end of the previous $(m \quad 1)$-th epoch, LCB-Q-Advantage iteratively updates the Q-estimate $\overline{Q}_h$ in all episodes during the m-th epoch. At each time step h in any episode, whenever $(s;a)$ is visited, LCB-Q-Advantage updates the reference Q-value as follows:

$$\overline{Q}_h(s;a) \quad (1 \quad _n)\overline{Q}_h(s;a) + {}_n \ r_h(s;a)$$

$$+ \underbrace{\widehat{P}_{h;s;a}\left(V_{h+1} \quad \overline{V}_{h+1}\right)}_{\text{estimate of } P_{h;s;a}(V_{h+1} \quad \overline{V}_{h+1})} + \underbrace{\overline{}_{h}}_{\text{estimate of } P_{h;s;a}\overline{V}_{h+1}} \quad \overline{b}_h(s;a) \Big\} : \qquad (19)$$

Intuitively, we decompose the target $P_{h;s;a}V_{h+1}$ into a reference part $P_{h;s;a}\overline{V}_{h+1}$ and an advantage part $P_{h;s;a}(V_{h+1}$ $\overline{}_{h+1})$, and cope with the two parts separately. In the sequel, let us take a moment to discuss three essential ingredients of the update rule (19), which shed light on the design rationale of our algorithm.

Akin to LCB-Q, the term $\widehat{P}_{h;s;a}\left(V_{h+1} \quad \overline{V}_{h+1}\right)$ serves as an unbiased stochastic estimate of $P_{h;s;a}\left(V_{h+1} \quad \overline{V}_{h+1}\right)$ if a sample transition $(s; a; s_{h+1})$ at time step h is observed. If $V_{h+1}$ stays close to the reference $\overline{V}_{h+1}$ as the algorithm proceeds, the variance of this stochastic term can be lower than that of the stochastic term $\widehat{P}_{h;s;a}V_{h+1}$ in (13).

The auxiliary estimate $\overline{}_{h}$ introduced in (18b) serves as a running estimate of the reference part $P_{h;s;a}\overline{V}_{h+1}$. Based on the update rule (18b), we design $\overline{}_{h}(s;a)$ to estimate the running mean of the reference part $P_{h;s;a}\overline{V}_{h+1}$ using a number of previous samples. As a result, we expect the variability of this term to be well-controlled, particularly as the number of samples in each epoch grows exponentially (recall that $L_m = 2^m$).

In each episode, the term $\overline{b}_h(s;a)$ serves as the additional confidence bound on the error between the estimates of the reference/advantage and the ground truth. More specifically, $_h^{\text{ref}}(s;a)$ and $_h^{\text{ref}}(s;a)$ are respectively the running mean and 2nd moment of the reference part $P_{h;s;a}\overline{V}_{h+1}$ (cf. lines 9-10 of Algorithm 3); $_h^{\text{adv}}(s;a)$ and $_h^{\text{adv}}(s;a)$ represent respectively the running mean and 2nd moment of the advantage part $P_{h;s;a}(V_{h+1} \quad \overline{V}_{h+1})$ (cf. lines 11-12 of Algorithm 3); $\overline{B}_h(s;a)$ aggregates the empirical standard deviations of the reference and the advantage parts. The LCB penalty term $\overline{b}_h(s;a)$ is updated using $\overline{B}_h(s;a)$ and $\overline{}_h(s;a)$ (cf. lines 5-6 of Algorithm 3), taking into account the confidence bounds for both the reference and the advantage.

In a nutshell, the auxiliary sequences of the reference values are designed to help reduce the variance of the stochastic Q-learning updates, which taken together with the principle of pessimism play a crucial role in the improvement of sample complexity for offline RL.

### 4.1. Theoretical guarantees for LCB-Q-Advantage

Encouragingly, the proposed LCB-Q-Advantage algorithm provably achieves near-optimal sample complexity for sufficiently small ", as demonstrated by the following theorem.

**Theorem 4.1.** Consider any $2 (0;1)$, and recall that $= \log \frac{SAT}{}$ and $T = KH$. Suppose that $c_b > 0$ is chosen to be a sufficiently large constant, and that the behavior policy satisfies Assumption 2.1. Then there exists some universal constant $c_g > 0$ such that with probability at least 1 , V the policy b output by Algorithm 2 satisfies

$$V_1^?() \quad V_1^{b}() \quad c_g \sqrt{\frac{H^4SC^?{}^5}{T}} + \frac{H^5SC^?{}^4}{T} : \qquad (20)$$

As a consequence, Theorem 4.1 reveals that the LCB-Q-Advantage algorithm is guaranteed to find an "-optimal policy (i.e., $V_1^?() \quad V_1^b() \quad "$ ) as long as the total sample size T exceeds

$$\widetilde{O}\left(\frac{H^4SC^?}{"^2} + \frac{H^5SC^?}{"}\right) : \qquad (21)$$

For sufficiently small accuracy level " (i.e., " 1=H), this results in a sample complexity of

$$\widetilde{O}\left(\frac{H^4SC^?}{"^2}\right) ; \qquad (22)$$

thereby matching the minimax lower bound developed in Xie et al. (2021b) up to logarithmic factor. Compared with the minimax lower bound $\frac{H^4SA}{"^2}$ in the online RL setting (Domingues et al., 2021), this suggests that offline RL can be fairly sample-efficient when the behavior policy closely mimics the optimal policy in terms of the resulting state-action occupancy distribution (a scenario where $C^?$ is potentially much smaller than the size of the action space).

**Comparison with offline model-based approaches.** In the same offline finite-horizon setting, the state-of-art model-based approach called PEVI-Adv has been proposed by Xie et al. (2021b), which also leverage the idea of reference-advantage decomposition. In comparison with PEVI-Adv, LCB-Q-Advantage not only enjoys the flexibility of model-free approaches, but also achieves optimal sample complexity for a broader range of target accuracy level ". More precisely, the "-range for which the algorithm achieves sample optimality can be compared as follows:

$$\underbrace{" \quad 0; H^{\ 1}}_{\text{(Our LCB-Q-Advantage)}} \quad \text{vs.} \quad \underbrace{" \quad 0; H^{\ 2:5}}_{\text{(PEVI-Adv)}}; \qquad (23)$$

offering an improvement by a factor of $H^{1:5}$.

## 5. Discussions

Focusing on model-free paradigms, this paper has developed near-optimal sample complexities for some variants of pessimistic Q-learning algorithms — armed with lower confidence bounds and variance reduction — for offline RL. These sample complexity results, taken together with the analysis framework developed herein, open up a few exciting directions for future research. For example, the pessimistic Q-learning algorithms can be deployed in conjunction with their optimistic counterparts (e.g., Jin et al. (2018); Li et al. (2021b); Zhang et al. (2020b)), when ad-ditional online data can be acquired to fine-tune the policy (Xie et al., 2021b). In addition, the "-range for LCB-Q-Advantage to attain sample optimality remains somewhat limited (i.e., " 2 (0; 1=H])). Our concurrent work Li et al. (2022) suggests that a new variant of pessimistic model-based algorithm is sample-optimal for a broader range of ", which in turn motivates further investigation into whether model-free algorithms can accommodate a broader "-range too without compromising sample efficiency. Moving be-yond the tabular setting, it would be of great importance to extend the algorithmic and theoretical framework to accom-modate low-complexity function approximation (Nguyen-Tang et al., 2021).

## Acknowledgements

L. Shi and Y. Chi are supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778, CCF-2007911, and the CAREER award ECCS-1818571. Y. Wei is supported in part by the NSF grants CCF-2106778, DMS-2147546/2015447, and the CAREER award DMS-2143215. Y. Chen is supported in part by the Alfred P. Sloan Re-search Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, IIS-2218713, DMS-2014279 and IIS-2218773. L. Shi is also gratefully supported by the Leo Finzi Memorial Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship, and Liang Ji-Dian Graduate Fellowship at Carnegie Mellon University. Part of this work was done while G. Li, Y. Wei and Y. Chen were visiting the Simons Institute for the Theory of Computing.

## References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. Conference on Learning Theory, pp. 67–83, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, pp. 263–272. JMLR. org, 2017.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient Q-learning with low switching cost. arXiv preprint arXiv:1905.12849, 2019.

Beck, C. L. and Srikant, R. Error bounds for constant step-size Q-learning. Systems & control letters, 61(12): 1203–1208, 2012.

Bertsekas, D. P. Dynamic programming and optimal control (4th edition). Athena Scientific, 2017.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. In International Conference on Learning Representations, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In International Conference on Machine Learning, pp. 1042–1051. PMLR, 2019.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-sample analysis of stochastic approximation using smooth convex envelopes. arXiv preprint arXiv:2002.00874, 2020.

Dadashi, R., Rezaeifar, S., Vieillard, N., Hussenot, L., Pietquin, O., and Geist, M. Offline reinforcement learning with pseudometric learning. arXiv preprint arXiv:2103.01948, 2021.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In Algorithmic Learning Theory, pp. 578–598. PMLR, 2021.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. arXiv preprint arXiv:1901.09311, 2019.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1049–1058. JMLR. org, 2017.

Even-Dar, E. and Mansour, Y. Learning rates for Q-learning. Journal of machine learning Research, 5(Dec):1–25, 2003.

Freedman, D. A. On tail probabilities for martingales. the Annals of Probability, pp. 100–118, 1975.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In International Conference on Machine Learning, pp. 2052–2062. PMLR, 2019.

Jaakkola, T., Jordan, M. I., and Singh, S. P. Convergence of stochastic iterative dynamic programming algorithms. In Advances in neural information processing systems, pp. 703–710, 1994.

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. arXiv preprint arXiv:2006.04354, 2020.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In Advances in Neural Information Processing Systems, pp. 4863–4873, 2018.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In International Conference on Machine Learning, pp. 5084–5096, 2021.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in neural information processing systems, pp. 315–323, 2013.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. Is temporal difference learning optimal? an instance-dependent analysis. arXiv preprint arXiv:2003.07337, 2020.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. arXiv preprint arXiv:2005.05951, 2020.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In Reinforcement learning, pp. 45–73. Springer, 2012.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643, 2020.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In Advances in Neural Information Processing Systems, volume 33, 2020.

Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y., and Chi, Y. Is Q-learning minimax optimal? a tight sample complexity analysis. arXiv preprint arXiv:2102.06548, 2021a.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. Advances in Neural Information Processing Systems, 34, 2021b.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. IEEE Transactions on Information Theory, 68(1):448–473, 2021c.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. arXiv preprint arXiv:2204.05275, 2022.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. arXiv preprint arXiv:2007.08202, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., and Ostrovski, G. Human-level control through deep reinforcement learning. Nature, 518(7540): 529–533, 2015.

Nguyen-Tang, T., Gupta, S., and Venkatesh, S. Sample complexity of offline reinforcement learning with deep ReLU networks. arXiv preprint arXiv:2103.06671, 2021.

Puterman, M. L. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014.

Qu, G. and Wierman, A. Finite-time analysis of asynchronous stochastic approximation and Q-learning. Conference on Learning Theory, pp. 3185–3205, 2020.

Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. Toward the fundamental limits of imitation learning. Advances in Neural Information Processing Systems, 33, 2020.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imita-tion learning: A tale of pessimism. Neural Information Processing Systems (NeurIPS), 2021.

Rezaeifar, S., Dadashi, R., Vieillard, N., Hussenot, L., Bachem, O., Pietquin, O., and Geist, M. Offline reinforcement learning as anti-exploration. arXiv preprint arXiv:2106.06431, 2021.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In Advances in Neural Information Processing Systems, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In Proceedings of the Twenty-Ninth

Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 770–787. SIAM, 2018b.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. Nature, 550(7676):354–359, 2017.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. arXiv preprint arXiv:2107.06226, 2021.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in Low-rank MDPs. arXiv preprint arXiv:2110.04652, 2021.

Vershynin, R. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. Variance reduced policy evaluation with smooth function approximation. Advances in Neural Information Processing Systems, 32:5784–5795, 2019.

Wainwright, M. J. Variance-reduced Q-learning is minimax optimal. arXiv preprint arXiv:1906.04697, 2019.

Watkins, C. J. and Dayan, P. Q-learning. Machine learning, 8(3-4):279–292, 1992.

Watkins, C. J. C. H. Learning from delayed rewards. PhD thesis, King's College, University of Cambridge, 1989.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. arXiv preprint arXiv:2106.06926, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. arXiv preprint arXiv:2106.04895, 2021b.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. Reanalysis of variance reduced temporal difference learning. In International Conference on Learning Representations, 2019.

Yan, Y., Li, G., Chen, Y., and Fan, J. The efficacy of pessimism in asynchronous Q-learning. arXiv preprint arXiv:2203.07368, 2022a.

Yan, Y., Li, G., Chen, Y., and Fan, J. Model-based reinforcement learning is minimax optimal for offline zero-sum Markov games. arXiv preprint arXiv:2206.04044, 2022b.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In International Conference on Artificial Intelligence and Statistics, pp. 1576–1584. PMLR, 2021.

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. Advances in neural information processing systems, 34, 2021.

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

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. Advances in neural information processing systems, 34, 2021b.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. arXiv preprint arXiv:2203.05804, 2022.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. MOPO: Model-based offline policy optimization. arXiv preprint arXiv:2005.13239, 2020.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. Conservative data sharing for multi-task offline reinforcement learning. arXiv preprint arXiv:2109.08128, 2021a.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. Combo: Conservative offline model-based policy optimization. arXiv preprint arXiv:2102.08363, 2021b.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. arXiv preprint arXiv:2108.08812, 2021.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. arXiv preprint arXiv:2009.13503, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. Advances in Neural Information Processing Systems, 33, 2020b.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. arXiv preprint arXiv:2006.03864, 2020c.

**Algorithm 4** LCB-Q for offline RL (a rewrite of Algorithm 1 to specify the dependency on k)

```
Parameters: some constant $c_b > 0$, target success probability $1 \quad 2 (0; 1)$, and $= \log \frac{SAT}{\ }$ .
Initialize: $Q_h^1(s;a) \quad 0$; $N_h^1(s;a) \quad 0$ for all $(s;a;h)$ 2 S A [H]; $V_h^1(s) \quad 0$ for all $(s;h)$ 2 S [H + 1]; $^1$ s.t.
    $^1_h(s) = 1$ for all $(s;h)$ 2 S [H].
for Episode k = 1 to K do
    Sample the k-th trajectory $f s_h^k; a_h^k; r_h^k g_{h=1}^H$ from D. // sampling from batch dataset
    for Step h = 1 to H do
        for (s; a) 2 S A do
            // carry over the estimates and policy
            $N_h^{k+1}(s;a) \quad N_h^k(s;a)$; $Q_h^{k+1}(s;a) \quad Q_h^k(s;a)$; $V_h^{k+1}(s) \quad V_h^k(s)$; $^{k+1}_h(s) \quad ^k_h(s)$.
        end for
        $N_h^{k+1}(s_h^k;a_h^k) \quad N_h^k(s_h^k;a_h^k) + 1$. // update the counter
        $n \quad N_h^{k+1}(s_h^k;a_h^k)$; $_n \quad \frac{H+1}{H+n}$. // update the learning rate
        $b_n \quad c_b \sqrt{\frac{H^{3\,2}}{n}}$. // update the bonus term
        // update the Q-estimates with LCB
        $Q_h^{k+1}(s_h^k;a_h^k) \quad Q_h^k(s_h^k;a_h^k) + {}_n \left\{ r_h(s_h^k,a_h^k) + V_{h+1}^k(s_{h+1}^k) \quad Q_h(s_h^k,a_h^k) \quad b_n \right\}$:
        // update the value estimates
        $V_h^{k+1}(s_h^k) \quad \max \left\{ V_h^k(s_h^k); \max_a Q_h^{k+1}(s_h^k;a) \right\}$.
        //update the policy
        If $V_h^{k+1}(s_h^k) = \max_a Q_h^{k+1}(s_h^k;a)$: update $^{k+1}_h(s^k_h) = \arg\max_a Q^{k+1}_h(s^k_h;a)$.
    end for
end for
```

## A. Analysis

In this section, we outline the main steps needed to establish the main results in Theorem 3.1 and Theorem 4.1. Before proceeding, let us first recall the following rescaled learning rates

$$_n = \frac{H+1}{H+n} \tag{24}$$

for the $n$-th visit of a given state-action pair at a given time step $h$, which are adopted in both LCB-Q and LCB-Q-Advantage. For notational convenience, we further introduce two sequences of related quantities defined for any integers $N \quad 0$ and $n \quad 1$:

$$_0^N = \begin{cases} \prod_{i=1}^N (1 \quad _i) = 0; & \text{if } N > 0; \\ 1; & \text{if } N = 0; \end{cases} \quad \text{and} \quad _n^N = \begin{cases} _n \prod_{i=n+1}^N (1 \quad _i); & \text{if } N > n; \\ _n; & \text{if } N = n; \\ 0; & \text{if } N < n: \end{cases} \tag{25}$$

The following identity can be easily verified:

$$\sum_{n=0}^{N} {}_n^N = 1: \tag{26}$$

### A.1. Analysis of LCB-Q

To begin with, we intend to derive a recursive formula concerning the update rule of $Q_h^k$— the estimate of the Q-function at step $h$ at the beginning of the $k$-th episode. Note that we have omitted the dependency of all quantities on the episode index $k$ in Algorithm 1. For notational convenience and clearness, we rewrite Algorithm 1 as Algorithm 4 by specifying the dependency on the episode index $k$ and shall often use the following set of short-hand notation when it is clear from context.

$N_h^k(s;a)$, or the shorthand $N_h^k$; the number of episodes that has visited $(s;a)$ at step $h$ before the beginning of the $k$-th episode.

- $k_h^n(s,a)$, or the shorthand $k^n$: the index of the episode in which the state-action pair $(s,a)$ is visited at step $h$ for the $n$-th times. We also adopt the convention that $k^0 = 0$.

- $P_h^k \in \{0,1\}^{1\times S}$: a row vector corresponding to the empirical transition at step $h$ of the $k$-th episode, namely,

$$P_h^k(s) = \mathbb{1}\big\{s = s_{h+1}^k\big\} \qquad \text{for all } s \in \mathcal{S}. \tag{27}$$

- $\pi^k = \{\pi_h^k\}_{h=1}^H$ with $\pi_h^k(s) =: \arg\max_a Q_h^k(s,a), \forall (h,s) \in [H]\times\mathcal{S}$: the deterministic greedy policy at the beginning of the $k$-th episode.

- $\widehat{\pi}$: the final output $\widehat{\pi}$ of Algorithms 1 corresponds to $\pi^{K+1}$ defined above; for notational simplicity, we shall treat $\widehat{\pi}$ as $\pi^K$ in our analysis, which does not affect our result at all.

Consider any state-action pair $(s,a)$. According to the update rule in line 14 of Algorithm 4, we can express (with the assistance of the above notation)

$$Q_h^k(s,a) = Q_h^{k^{N_h^k}+1}(s,a) = \big(1-\eta_{N_h^k}\big)Q_h^{k^{N_h^k}}(s,a) + \eta_{N_h^k}\Big\{r_h(s,a) + V_{h+1}^{k^{N_h^k}}\big(s_{h+1}^{k^{N_h^k}}\big) - b_{N_h^k}\Big\}, \tag{28}$$

where the first identity holds since $k^{N_h^k}$ denotes the latest episode prior to $k$ that visits $(s,a)$ at step $h$, and the learning rate is defined in (24). Note that it always holds that $k > k^{N_h^k}$. Applying the above relation (28) recursively and using the notation (25) lead to

$$Q_h^k(s,a) = \eta_0^{N_h^k} Q_h^1(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big\{r_h(s,a) + V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) - b_n\Big\}. \tag{29}$$

As another important fact, the value estimate $V_h^k$ is monotonically non-decreasing in $k$, i.e.,

$$V_h^{k+1}(s) \ge V_h^k(s) \qquad \text{for all } (s,k,h) \in \mathcal{S}\times[K]\times[H], \tag{30}$$

which is an immediate consequence of the update rule in line 16 of Algorithm 4. Crucially, we observe that the iterate $V_h^k$ forms a "pessimistic view" of $V_h^{\pi^k}$ — and in turn $V_h^\star$ — resulting from suitable design of the penalty term. This observation is formally stated in the following lemma, with the proof postponed to Section C.1.

**Lemma A.1.** Consider any $\delta \in (0,1)$, and suppose that $c_b > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$,

$$\Bigg|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big)V_{h+1}^{k^n(s,a)}\Bigg| \le \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \tag{31}$$

holds simultaneously for all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$, and

$$V_h^k(s) \le V_h^{\pi^k}(s) \le V_h^\star(s) \tag{32}$$

holds simultaneously for all $(k,h,s) \in [K]\times[H]\times\mathcal{S}$.

In a nutshell, the result (32) in Lemma A.1 reveals that $V_h^k$ is a pointwise lower bound on $V_h^{\pi^k}$ and $V_h^\star$, thereby forming a pessimistic estimate of the optimal value function. In addition, the property (31) in Lemma A.1 essentially tells us that the weighted sum of the penalty terms dominates the weighted sum of the uncertainty terms, which plays a crucial role in ensuring the aforementioned pessimism property. As we shall see momentarily, Lemma A.1 forms the basis of the subsequent proof.

We are now ready to embark on the analysis for LCB-Q, which is divided into multiple steps as follows.

**Step 1: decomposing estimation errors.** With the aid of Lemma A.1, we can develop an upper bound on the performance difference of interest in (20) as follows

$$\begin{aligned}
V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &= \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^\star(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{\pi^K}(s_1)\big] \\
&\overset{(\mathrm{i})}{\leq} \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^\star(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^K(s_1)\big] \\
&\overset{(\mathrm{ii})}{\leq} \frac{1}{K}\sum_{k=1}^{K}\Big(\mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^\star(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^k(s_1)\big]\Big) \\
&= \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\Big(V_1^\star(s) - V_1^k(s)\Big), \qquad (33)
\end{aligned}$$

where (i) results from Lemma A.1 (i.e., $V_1^{\pi^K}(s) \geq V_1^K(s)$ for all $s\in\mathcal{S}$), (ii) follows from the monotonicity property in (30), and the last equality holds since $d_1^{\pi^\star}(s) = \rho(s)$ (cf. (4)).

We then attempt to bound the quantity on the right-hand side of (33). Given that $\pi^\star$ is assumed to be a deterministic policy, we have $d_h^{\pi^\star}(s) = d_h^{\pi^\star}(s,\pi_h^\star(s))$. Taking this together with the relations $V_h^k(s) \geq \max_a Q_h^k(s,a) \geq Q_h^k(s,\pi_h^\star(s))$ (see line 16 of Algorithm 4) and $V_h^\star(s) = Q_h^\star(s,\pi_h^\star(s))$, we obtain

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\Big(V_h^\star(s) - V_h^k(s)\Big) &= \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi_h^\star(s)\big)\Big(V_h^\star(s) - V_h^k(s)\Big) \\
&\leq \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi_h^\star(s)\big)\Big(Q_h^\star\big(s,\pi_h^\star(s)\big) - Q_h^k\big(s,\pi_h^\star(s)\big)\Big) \\
&= \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\Big(Q_h^\star(s,a) - Q_h^k(s,a)\Big) \qquad (34)
\end{aligned}$$

for any $h\in[H]$, where the last identity holds since $\pi^\star$ is deterministic and hence

$$d_h^{\pi^\star}(s,a) = 0 \qquad \text{for any } a \neq \pi_h^\star(s). \qquad (35)$$

In view of (34), we need to properly control $Q_h^\star(s,a) - Q_h^k(s,a)$. By virtue of (26), we can rewrite $Q_h^\star(s,a)$ as follows

$$\begin{aligned}
Q_h^\star(s,a) &= \sum_{n=0}^{N_h^k}\eta_n^{N_h^k} Q_h^\star(s,a) = \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} Q_h^\star(s,a) \\
&= \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(r_h(s,a) + P_{h,s,a}V_{h+1}^\star\Big), \qquad (36)
\end{aligned}$$

where the second line follows from Bellman's optimality equation (10). Combining (29) and (36) leads to

$$\begin{aligned}
&Q_h^\star(s,a) - Q_h^k(s,a) \\
&= \eta_0^{N_h^k}\Big(Q_h^\star(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s,a}V_{h+1}^\star - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n\Big) \\
&= \eta_0^{N_h^k}\Big(Q_h^\star(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} P_{h,s,a}\Big(V_{h+1}^\star - V_{h+1}^{k^n}\Big) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s,a} - P_h^{k^n}\Big)V_{h+1}^{k^n} \qquad (37) \\
&\leq \eta_0^{N_h^k} H + 2\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} P_{h,s,a}\Big(V_{h+1}^\star - V_{h+1}^{k^n}\Big), \qquad (38)
\end{aligned}$$

where we have made use of the definition in (27) by recognizing $P_h^{k^n} V_{h+1}^{k^n} = V_{h+1}^{k^n}(s_{h+1}^{k^n})$ in (37), and the last inequality follows from the fact $Q_h^{\star}(s,a) - Q_h^{1}(s,a) = Q_h^{\star}(s,a) - 0 \le H$ and the bound (31) in Lemma A.1. Substituting the above bound into (34), we arrive at

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\big(V_h^{\star}(s) - V_h^{k}(s)\big) \le \underbrace{\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\,\eta_0^{N_h^k(s,a)} H + 2\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)} b_n}_{=:I_h} + \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a) P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^{\star} - V_{h+1}^{k^n(s,a)}\big). \qquad (39)$$

**Step 2: establishing a crucial recursion.** As it turns out, the last term on the right-hand side of (39) can be used to derive a recursive relation that connects step $h$ with step $h+1$, as summarized in the next lemma.

**Lemma A.2.** *With probability at least $1-\delta$, the following recursion holds:*

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a) P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^{\star} - V_{h+1}^{k^n(s,a)}\big)$$
$$\le \Big(1+\frac{1}{H}\Big)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\big(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\big) + 24\sqrt{H^2 C^{\star} K \log\frac{2H}{\delta}} + 12HC^{\star}\log\frac{2H}{\delta}. \qquad (40)$$

Lemma A.2 taken together with (39) implies that

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\big(V_h^{\star}(s) - V_h^{k}(s)\big) \le \Big(1+\frac{1}{H}\Big)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\big(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\big)$$
$$+ I_h + 24\sqrt{H^2 C^{\star} K \log\frac{2H}{\delta}} + 12HC^{\star}\log\frac{2H}{\delta}. \qquad (41)$$

Invoking (41) recursively over the time steps $h = H, H-1, \cdots, 1$ with the terminal condition $V_{H+1}^{k} = V_{H+1}^{\star} = 0$, we reach

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\star}(s)\big(V_1^{\star}(s) - V_1^{k}(s)\big) \le \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\Big[V_h^{\star}(s) - V_h^{k}(s)\Big]$$
$$\le \sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\left(I_h + 24\sqrt{H^2 C^{\star} K \log\frac{2H}{\delta}} + 12HC^{\star}\log\frac{2H}{\delta}\right), \qquad (42)$$

which captures the estimation error resulting from the use of pessimism principle.

**Step 3: controlling the right-hand side of (42).** The right-hand side of (42) can be bounded through the following lemma, which will be proved in Appendix C.3.

**Lemma A.3.** *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, we have*

$$\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\left(I_h + 24\sqrt{H^2 C^{\star} K \log\frac{2H}{\delta}} + 12HC^{\star}\log\frac{2H}{\delta}\right) \lesssim H^2 S C^{\star}\iota + \sqrt{H^5 S C^{\star} K \iota^3}, \qquad (43)$$

*where we recall that $\iota := \log\frac{SAT}{\delta}$.*

Combining Lemma A.3 with (42) and (33) yields

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\star}(s)\big(V_1^{\star}(s) - V_1^{k}(s)\big)$$

$$\begin{aligned}
\frac{1}{K}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in S} d_h^{\star}(s)\big(V_h^{\star}(s)-V_h^{k}(s)\big)
&\le \frac{c_a}{2}\sqrt{\frac{H^5SC^{\star}\iota^3}{K}}+\frac{c_a}{2}\frac{H^2SC^{\star}}{K}=\frac{c_a}{2}\sqrt{\frac{H^6SC^{\star}\iota^3}{T}}+\frac{c_a}{2}\frac{H^3SC^{\star}}{T}\\
&\le c_a\sqrt{\frac{H^6SC^{\star}\iota^3}{T}}
\end{aligned}\tag{44}$$

for some sufficiently large constant $c_a > 0$, where the last inequality is valid as long as $T \geq SC^{\star}$. This concludes the proof of Theorem 3.1.

### A.2. Analysis of LCB-Q-Advantage

We now turn to the analysis of LCB-Q-Advantage. Thus far, we have omitted the dependency of all quantities on the epoch number $m$ and the in-epoch episode number $t$ in Algorithms 2 and 3. While it allows for a more concise description of our algorithm, it might hamper the clarity of our proofs. In the following, we introduce the notation $k$ to denote the current episode as follows:

$$k := \sum_{i=1}^{m-1} L_i + t, \tag{45}$$

which corresponds to the $t$-th in-epoch episode in the $m$-th epoch; here, $L_m = 2^m$ stands for the total number of in-epoch episodes in the $m$-th epoch. With this notation in place, we can rewrite Algorithm 2 as Algorithm 5 in order to make clear the dependency on the episode index $k$, epoch number $m$, and in-epoch episode index $t$.

Before embarking on our main proof, we make two crucial observations which play important roles in our subsequent analysis. First, similar to the property (30) for LCB-Q, the update rule (cf. lines 19-20 of Algorithm 5) ensures the monotonic non-decreasing property of $V_h(s)$ such that for all $k \in [K]$,

$$V_h^{k+1}(s) \geq V_h^{k}(s); \qquad \text{for all } (k, s, h) \in [K]\times S\times[H]. \tag{46}$$

Secondly, $V_h^k$ forms a "pessimistic view" of $V_h^\star$, which is formalized in the lemma below; the proof is deferred to Appendix D.1.

Lemma A.4. Let $\delta \in (0, 1)$. Suppose that $c_b > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$, the value estimates produced by Algorithm 2 satisfy

$$V_h^k(s) \le V_h^{\pi^k}(s) \le V^\star(s) \tag{47}$$

for all $(k, h, s) \in [K]\times[H + 1]\times S$.

With these two observations in place, we can proceed to present the analysis for LCB-Q-Advantage. To begin with, the performance difference of interest can be controlled similar to (33) as follows:

$$\begin{aligned}
V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) &= \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{\star}(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{\pi^K}(s_1)\big]\\
&\overset{(i)}{\le} \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{\star}(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{K}(s_1)\big]\\
&\overset{(ii)}{\le} \frac{1}{K}\sum_{k=1}^{K}\Big(\mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{\star}(s_1)\big] - \mathop{\mathbb{E}}_{s_1\sim\rho}\big[V_1^{k}(s_1)\big]\Big)\\
&= \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in S} d_1^{\star}(s)\big(V_1^{\star}(s) - V_1^{k}(s)\big);
\end{aligned}\tag{48}$$

where (i) follows from Lemma A.4 (i.e., $V_1^{\pi^K}(s) \geq V_1^K(s)$ for all $s \in S$), (ii) holds due to the monotonicity in (46) and the last equality holds since $d_1^{\star}(s) = \rho(s)$ (cf. (4)). It then boils down to controlling the right-hand side of (48). Towards this

**Algorithm 5** LCB-Q-Advantage (a rewrite of Algorithm 2 that specifies dependency on $k$ or $(m,t)$.)

```
Parameters: number of epochs $M$, universal constant $c_b > 0$, target success probability $1-\delta \in (0,1)$, and $\iota = \log\frac{SAT}{\delta}$;
Initialize: $Q_h^1(s,a); Q_h^{\mathrm{LCB},1}(s,a); \overline{Q}_h^1(s,a); \overline{\mu}_h^1(s,a); \mu_h^{\mathrm{next},1}(s,a); N_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$; $V_h^1(s); \overline{V}_h^1(s); \overline{V}_h^{\mathrm{next},1}(s) \leftarrow 0$ for all $(s,h) \in \mathcal{S}\times[H+1]$; $\mu_h^{\mathrm{ref},1}(s,a); \sigma_h^{\mathrm{ref},1}(s,a); \mu_h^{\mathrm{adv},1}(s,a), \sigma_h^{\mathrm{adv},1}(s,a), \delta_h^1(s,a); B_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
for Epoch $m = 1$ to $M$ do
    $L_m = 2^m$; // specify the number of episodes in the current epoch
    $\widehat{N}_h^{(m,1)}(s,a) = 0$ for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$; // reset the epoch-wise counter
    /* Inner-loop: update value-estimates $V_h(s,a)$ and Q-estimates $Q_h(s,a)$
    for In-epoch Episode $t = 1$ to $L_m$ do
        Set $k \leftarrow \sum_{i=1}^{m-1} L_i + t$; // set the episode index
        Sample the $k$-th trajectory $\{s_h^k, a_h^k\}_{h=1}^H$. // sampling from batch dataset
        Compute $\pi^k$ s.t. $\pi_h^k(s) = \arg\max_a Q_h^k(s,a)$ for all $(s,h) \in \mathcal{S}\times[H]$. // update the policy
        for Step $h = 1$ to $H$ do
            for $(s,a) \in \mathcal{S}\times\mathcal{A}$ do
                // carry over the estimates
                $N_h^{k+1}(s,a) \leftarrow N_h^k(s,a); \widehat{N}_h^{k+1}(s,a) \leftarrow \widehat{N}_h^k(s,a); V_h^{k+1}(s) \leftarrow V_h^k(s)$;
                $Q_h^{\mathrm{LCB},k+1}(s,a) \leftarrow Q_h^{\mathrm{LCB},k}(s,a)$ $\overline{Q}_h^{k+1}(s,a) \leftarrow \overline{Q}_h^k(s,a)$; $Q_h^{k+1}(s,a) \leftarrow Q_h^k(s,a)$;
                $\overline{V}_h^{k+1}(s) \leftarrow \overline{V}_h^k(s)$ $\overline{V}_h^{\mathrm{next},k+1}(s) \leftarrow \overline{V}_h^{\mathrm{next},k}(s)$; $\mu_h^{\mathrm{next},k+1}(s,a) \leftarrow \mu_h^{\mathrm{next},k}(s,a)$.
            end for
            $N_h^{k+1}(s_h^k,a_h^k) \leftarrow N_h^k(s_h^k,a_h^k) + 1$; $n \leftarrow N_h^{k+1}(s_h^k,a_h^k)$. // update the overall counter
            $\eta_n \leftarrow \frac{H+1}{H+n}$; // update the learning rate
            // update the Q-estimate with LCB
            $Q_h^{\mathrm{LCB},k+1}(s_h^k,a_h^k) \leftarrow$ update-lcb-q();
            // update the Q-estimate with LCB and reference-advantage
            $\overline{Q}_h^{k+1}(s_h^k,a_h^k) \leftarrow$ update-lcb-q-ra();
            // update the Q-estimate $Q_h$ and value estimate $V_h$
            $Q_h^{k+1}(s_h^k,a_h^k) \leftarrow \max\{Q_h^{\mathrm{LCB},k+1}(s_h^k,a_h^k), \overline{Q}_h^{k+1}(s_h^k,a_h^k), Q_h^k(s_h^k,a_h^k)\}$;
            $V_h^{k+1}(s_h^k) \leftarrow \max_a Q_h^{k+1}(s_h^k,a)$.
            // update epoch-wise counter and $\mu_h^{\mathrm{next}}(s,a)$ for the next epoch
            $\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k) \leftarrow \widehat{N}_h^{(m,t)}(s_h^k,a_h^k) + 1$;
            $\mu_h^{\mathrm{next},k+1}(s_h^k,a_h^k) \leftarrow \Big(1 - \frac{1}{\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k)}\Big)\mu_h^{\mathrm{next},k}(s_h,a_h) + \frac{1}{\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k)}\overline{V}_{h+1}^{\mathrm{next},k}(s_{h+1})$.
        end for
    end for
    /* Update the reference $(\overline{V}_h, \overline{V}_h^{\mathrm{next}})$ and $(\overline{\mu}_h, \mu_h^{\mathrm{next}})$
    for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H+1]$ do
        $\overline{V}_h^{k+1}(s) \leftarrow \overline{V}_h^{\mathrm{next},k+1}(s)$; $\overline{\mu}_h^{k+1}(s,a) \leftarrow \mu_h^{\mathrm{next},k+1}(s,a)$. // set $\overline{V}_h$ and $\overline{\mu}_h$ for the next epoch
        $\overline{V}_h^{\mathrm{next},k+1}(s) \leftarrow V_h^{k+1}(s)$; $\mu_h^{\mathrm{next},k+1}(s,a) \leftarrow 0$. // set $\mu^{\mathrm{next}}$ and $\overline{V}^{\mathrm{next}}$ for the next epoch
    end for
end for
Output: the policy $\widehat{\pi} = \pi^K$ with $K = \sum_{m=1}^M L_m$.
```

end, it turns out that one can control a more general counterpart, i.e.,

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\big(V_h^\star(s) - V_h^k(s)\big) \tag{49}$$

for any $h \in [H]$. This is accomplished via the following lemma, whose proof is postponed to Appendix D.2.

**Lemma A.5.** Let $\delta \in (0,1)$, and recall that $\iota := \log\frac{SAT}{\delta}$. Suppose that $c_a, c_b > 0$ are some sufficiently large constants.

Then with probability at least $1-\delta$, one has

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\big(V_h^\star(s) - V_h^k(s)\big) \le J_h^1 + J_h^2 + J_h^3, \tag{50}$$

where

$$J_h^1 := \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + \frac{4c_bH^{7/4}\iota}{\big(N_h^k(s,a)\vee 1\big)^{3/4}} + \frac{4c_bH^2\iota}{N_h^k(s,a)\vee 1}\right];$$

$$J_h^2 := 2\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\overline{B}_h^k(s,a);$$

$$J_h^3 := \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^\star(s) - V_{h+1}^k(s)\big) + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_aH^3C^\star\sqrt{S}\iota^2. \tag{51}$$

As a direct consequence of Lemma A.5, one arrives at a recursive relationship between time steps $h$ and $h+1$ as follows:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\big(V_h^\star(s) - V_h^k(s)\big) \le{}& \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^\star(s) - V_{h+1}^k(s)\big) + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_aH^3C^\star\sqrt{S}\iota^2 + J_h^1 + J_h^2.\end{aligned} \tag{52}$$

Recursing over time steps $h = H, H-1, \cdots, 1$ with the terminal condition $V_{H+1}^k = V_{H+1}^\star = 0$, we can upper bound the performance difference at $h = 1$ as follows

$$\begin{aligned}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\big(V_1^\star(s) - V_1^k(s)\big) &\le \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\Big[V_h^\star(s) - V_h^k(s)\Big] \\ &\le \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_aH^3C^\star\sqrt{S}\iota^2 + J_h^1 + J_h^2\right).\end{aligned} \tag{53}$$

To finish up, it suffices to upper bound each term in (53) separately. We summarize their respective upper bounds as follows; the proof is provided in Appendix D.3.

**Lemma A.6.** Fix $\delta\in(0,1)$, and recall that $\iota := \log\frac{SAT}{\delta}$. With probability at least $1-\delta$, we have

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^1 \lesssim H^{2.75}(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2 + H^3SC^\star\iota^3; \tag{54a}$$

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^2 \lesssim \sqrt{H^4SC^\star\iota^3\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\Big[V_h^\star(s) - V_h^k(s)\Big]} + \sqrt{H^3SC^\star K\iota^5} + H^4SC^\star\iota^4; \tag{54b}$$

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_aH^3C^\star\sqrt{S}\iota^2\right) \lesssim \sqrt{H^3C^\star K\log\frac{2H}{\delta}} + H^4C^\star\sqrt{S}\iota^2. \tag{54c}$$

Substituting the above upper bounds into (48) and (53) and recalling that $T = HK$, we arrive at

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{1}{K}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\big(V_h^\star(s) - V_h^k(s)\big)$$

$$
\begin{aligned}
&\lesssim \frac{1}{K}\left(\sqrt{H^4SC^\star\iota^3 \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\big(V_h^\star(s)-V_h^k(s)\big)}+\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4+H^{2.75}(SC^\star)^{\frac34}K^{\frac14}\iota^2\right)\\
&\overset{\text{(i)}}{\lesssim} \frac{1}{K}\left(\sqrt{H^4SC^\star\iota^3 \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\big(V_h^\star(s)-V_h^k(s)\big)}+\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4\right)\\
&\overset{\text{(ii)}}{\lesssim} \frac{1}{K}\left(\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4\right)\\
&\asymp \sqrt{\frac{H^4SC^\star\iota^5}{T}}+\frac{H^5SC^\star\iota^4}{T};
\end{aligned}
$$

where (i) has made use of the AM-GM inequality:

$$
2H^{2.75}(SC^\star)^{\frac34}K^{\frac14}\iota^{2} \le \left(H^{0.75}(SC^\star)^{\frac14}K^{\frac14}\iota^{\frac14}\right)^2+\left(H^2(SC^\star)^{\frac12}\iota^{\frac74}\right)^2=\sqrt{H^3SC^\star K\iota}+H^4SC^\star\iota^{\frac72};
$$

and (ii) holds by letting $x := \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\big(V_h^\star(s)-V_h^k(s)\big)$ and solving the inequality $x \lesssim \sqrt{H^4SC^\star\iota^3 x}+\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4$. This concludes the proof.

## B. Technical lemmas

### B.1. Preliminary facts

Our results rely heavily on proper choices of the learning rates. In what follows, we make note of several useful properties concerning the learning rates, which have been established in (Jin et al., 2018; Li et al., 2021b).

Lemma B.1 (Lemma 1 in (Li et al., 2021b)). For any integer $N > 0$, the following properties hold:

$$
\frac{1}{N^a} \le \sum_{n=1}^{N}\frac{\eta_n^N}{n^a} \le \frac{2}{N^a} \qquad \text{for all } \frac12 \le a \le 1; \tag{55a}
$$

$$
\max_{1\le n\le N}\eta_n^N \le \frac{2H}{N}, \qquad \sum_{n=1}^{N}\left(\eta_n^N\right)^2 \le \frac{2H}{N}, \qquad \sum_{N=n}^{\infty}\eta_n^N \le 1+\frac{1}{H}. \tag{55b}
$$

In addition, we gather a few elementary properties about the Binomial distribution, which will be useful throughout the proof. The lemma below is adapted from Xie et al. (2021b, Lemma A.1).

Lemma B.2. Suppose $N \sim \mathsf{Binomial}(n,p)$, where $n \ge 1$ and $p \in [0,1]$. For any $\delta \in (0,1)$, we have

$$
\frac{p}{N \vee 1} \le \frac{8\log\frac{1}{\delta}}{n}; \tag{56}
$$

and

$$
N \ge \frac{np}{8\log\frac{1}{\delta}} \qquad \text{if } np \ge 8\log\frac{1}{\delta}; \tag{57a}
$$

$$
N \le \begin{cases} e^2np & \text{if } np \ge \log\frac{1}{\delta};\\ 2e^2\log\frac{1}{\delta} & \text{if } np \le 2\log\frac{1}{\delta}: \end{cases} \tag{57b}
$$

with probability at least $1-4\delta$.

Proof. To begin with, we directly invoke Xie et al. (2021b, Lemma A.1) which yields the results in (56) and (57a). Regarding (57b), invoking the Chernoff bound (Vershynin, 2018, Theorem 2.3.1) with $\mathbb{E}[N] = np$, when $np \ge \log\frac{1}{\delta}$, it satisfies

$$
\mathbb{P}(N \ge e^2np) \le e^{-np}\left(\frac{enp}{e^2np}\right)^{e^2np} \le e^{-np}:
$$

Similarly, when $np \le 2\log\frac{1}{\delta}$, we have

$$\mathbb{P}\left\{N \ge 2e^2\log\frac{1}{\delta}\right\} \overset{(i)}{\le} e^{-np}\left(\frac{enp}{2e^2\log\frac{1}{\delta}}\right)^{2e^2\log(\frac{1}{\delta})} \overset{(ii)}{\le} e^{-np}\left(\frac{enp}{e^2np}\right)^{2e^2\log(\frac{1}{\delta})} \le e^{-2e^2\log(\frac{1}{\delta})} \le \delta;$$

where (i) results from Vershynin (2018, Theorem 2.3.1), and (ii) follows from the basic fact $e^2\log\frac{1}{\delta} \ge 2\log\frac{1}{\delta} \ge np$. Taking the union bound thus completes the proof.

□

### B.2. Freedman's inequality and its consequences

Both the samples collected within each episode and the algorithms analyzed herein exhibit certain Markovian structure. As a result, concentration inequalities tailored to martingales become particularly effective for our analysis. In this subsection, we collect a few useful concentration results that will be applied multiple times in the current paper. These results might be of independent interest.

To begin with, the following theorem provides a user-friendly version of Freedman's inequality (Freedman, 1975); see Li et al. (2021a, Section C) for more details.

**Theorem B.3** (Freedman's inequality). Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and let $\mathbb{E}_k$ stand for the expectation conditioned on $\mathcal{F}_k$. Suppose that $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}_{k-1}[X_k] = 0 \qquad \text{for all } k \ge 1$$

for some quantity $R < \infty$. We also define

$$W_n := \sum_{k=1}^{n} \mathbb{E}_{k-1}\left[X_k^2\right].$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically for some given quantity $\sigma^2 < \infty$. Then for any positive integer $m \ge 1$, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^m}\right\}\log\frac{2m}{\delta}} + \frac{4}{3}R\log\frac{2m}{\delta}. \tag{58}$$

We shall also record some immediate consequence of Freedman's inequality tailored to our problem. Recall that $N_h^i(s,a)$ denotes the number of times that $(s,a)$ has been visited at step $h$ before the beginning of the $i$-th episode, and $k^n(s,a)$ stands for the index of the episode in which $(s,a)$ is visited for the $n$-th time. The following concentration bound has been established in Li et al. (2021b, Lemma 7).

**Lemma B.4.** Let $\left\{W_h^i \in \mathbb{R}^S \mid 1 \le i \le K, 1 \le h \le H+1\right\}$ and $\left\{u_h^i(s,a,N) \in \mathbb{R} \mid 1 \le i \le K, 1 \le h \le H+1\right\}$ be a collections of vectors and scalars, respectively, and suppose that they obey the following properties:

- $W_h^i$ is fully determined by the samples collected up to the end of the $(h-1)$-th step of the $i$-th episode;
- $\|W_h^i\|_\infty \le C_w$;
- $u_h^i(s,a,N)$ is fully determined by the samples collected up to the end of the $(h-1)$-th step of the $i$-th episode, and a given positive integer $N \in [K]$;
- $0 \le u_h^i(s,a,N) \le C_u$;
- $0 \le \sum_{n=1}^{N_h^K(s,a)} u_h^{k_h^n(s,a)}(s,a,N) \le 2$.

In addition, consider the following sequence

$$X_i(s,a,h,N) := u_h^i(s,a,N)\left(P_h^i - P_{h,s,a}\right)W_{h+1}^i \mathbb{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}, \qquad 1 \le i \le K, \tag{59}$$

with $P_h^i$ defined in (27). Consider any $\delta \in (0,1)$. Then with probability at least $1-\delta$,

$$\sum_{i=1}^{k} X_i(s,a,h,N) \lesssim \sqrt{C_u \log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k_h^n(s,a)}\right)} + \left(C_u C_w + \sqrt{\frac{C_u}{N}}C_w\right)\log^2\frac{SAT}{\delta} \tag{60}$$

holds simultaneously for all $(k,h,s,a,N) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$.

Next, we make note of an immediate consequence of Lemma B.4 as follows.

Lemma B.5. Let $\{W_h^i \in \mathbb{R}^S \mid 1 \le i \le K, 1 \le h \le H+1\}$ be a collection of vectors satisfying the following properties:

- $W_h^i$ is fully determined by the samples collected up to the end of the $(h-1)$-th step of the $i$-th episode;
- $\|W_h^i\|_\infty \le C_w$.

For any positive $N \ge H$, we consider the following sequence

$$X_i(s,a,h,N) := \eta^N_{N_h^i(s,a)}\left(P_h^i - P_{h,s,a}\right)W_{h+1}^i \mathbb{1}\left\{(s_h^i,a_h^i)=(s,a)\right\}, \qquad 1 \le i \le K, \tag{61}$$

with $P_h^i$ defined in (27). Consider any $\delta \in (0,1)$. With probability at least $1-\delta$,

$$\sum_{i=1}^{k} X_i(s,a,h,N) \lesssim \sqrt{\frac{H}{N}C_w^2\log^2\frac{SAT}{\delta}} \tag{62}$$

holds simultaneously for all $(k,h,s,a,N) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$.

Proof. Taking $u_h^i(s,a,N) = \eta^N_{N_h^i(s,a)}$, one can see from (55b) in Lemma B.1 that

$$u_h^i(s,a,N) \le \frac{2H}{N} =: C_u.$$

Recognizing the trivial bound $\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k_h^n(s,a)}\right) \le C_w^2$, we can invoke Lemma B.4 to obtain that, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{i=1}^{k} X_i(s,a,h,N) &\lesssim \sqrt{C_u \log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)} \eta_n^N C_w^2} + \left(C_u C_w + \sqrt{\frac{C_u}{N}}C_w\right)\log^2\frac{SAT}{\delta}\\ &\lesssim \sqrt{\frac{H}{N}\log^2\frac{SAT}{\delta}}C_w + \frac{HC_w}{N}\log^2\frac{SAT}{\delta} \lesssim \sqrt{\frac{HC_w^2}{N}\log^2\frac{SAT}{\delta}}\end{aligned}$$

holds simultaneously for all $(k,h,s,a,N) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$, where the last line applies (55b) in Lemma B.1 once again. □

Finally, we introduce another lemma by invoking Freedman's inequality in Theorem B.3.

Lemma B.6. Let $\{W_h^k(s,a) \in \mathbb{R}^S \mid (s,a) \in \mathcal{S}\times\mathcal{A}, 1 \le k \le K, 1 \le h \le H+1\}$ be a collection of vectors satisfying the following properties:

- $W_h^k(s,a)$ is fully determined by the given state-action pair $(s,a)$ and the samples collected up to the end of the $(k-1)$-th episode;
- $\|W_h^k(s,a)\|_\infty \le C_w$.

For any positive $C_d \geq 0$, we consider the following sequences

$$X_{h,k} := C_d \left[ \frac{d_h^{\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a) P_{h,s,a} W_{h+1}^k(s,a) \right], \qquad 1 \leq k \leq K; \tag{63}$$

$$\overline{X}_{h,k} := C_d \left[ \frac{d_h^{\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a) P_{h,s,a} W_{h+1}^k(s,a) \right], \qquad 1 \leq k \leq K. \tag{64}$$

Consider any $\delta \in (0,1)$. Then with probability at least $1-\delta$,

$$\sum_{k=1}^K X_{h,k} \leq \sqrt{8C_d^2 C^\star \sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\left(P_{h,s,a}W_{h+1}^k(s,a)\right)^2 \log\frac{2H}{\delta}} + 2C_d C^\star C_w \log\frac{2H}{\delta} \tag{65}$$

$$\sum_{k=1}^K \overline{X}_{h,k} \leq \sqrt{8C_d^2 C^\star \sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\left(W_{h+1}^k(s,a)\right)^2 \log\frac{2H}{\delta}} + 2C_d C^\star C_w \log\frac{2H}{\delta} \tag{66}$$

hold simultaneously for all $h \in [H]$.

*Proof.* We intend to apply Freedman's inequality (cf. Theorem B.3) to control $\sum_{k=1}^K X_{h,k}$. Considering any given time step $h$, it is easily verified that

$$\mathbb{E}_{k-1}[X_{h,k}] = 0; \qquad \mathbb{E}_{k-1}[\overline{X}_{h,k}] = 0;$$

where $\mathbb{E}_{k-1}$ denotes the expectation conditioned on everything happening up to the end of the $(k-1)$-th episode. To continue, we observe that

$$|X_{h,k}| \leq C_d \left(\frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} + 1\right)\left\|W_{h+1}^k(s,a)\right\|_\infty \leq 2C_d C^\star C_w; \tag{67}$$

$$|\overline{X}_{h,k}| \leq C_d \left(\frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} + 1\right)\left\|W_{h+1}^k(s,a)\right\|_\infty \leq 2C_d C^\star C_w; \tag{68}$$

where we use the assumptions $\frac{d_h^\star(s,a)}{d_h^\mu(s,a)} \leq C^\star$ for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ (cf. Assumption 2.1) and $\left\|W_{h+1}^k(s_h^k,a_h^k)\right\|_\infty \leq C_w$. Recall that $\Delta(\mathcal{S}\times\mathcal{A})$ is the probability simplex over the set $\mathcal{S}\times\mathcal{A}$ of all state-action pairs, and we denote by $d_h^\mu \in \Delta(\mathcal{S}\times\mathcal{A})$ the state-action visitation distribution induced by the behavior policy $\mu$ at time step $h \in [H]$. With this in hand, we obtain

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}_{k-1}\left[|X_{h,k}|^2\right] &\leq \sum_{k=1}^K C_d^2\, \mathbb{E}_{k-1}\left[\left(\frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}W_{h+1}^k(s,a)\right)^2\right] \\
&\leq \sum_{k=1}^K C_d^2\, \mathbb{E}_{(s_h^k,a_h^k)\sim d_h^\mu}\left[\left(\frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k,a_h^k)\right)^2\right] \\
&= \sum_{k=1}^K C_d^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^\star(s,a)}{d_h^\mu(s,a)}\, d_h^\star(s,a)\left(P_{h,s,a}W_{h+1}^k(s,a)\right)^2 \\
&\overset{\text{(i)}}{\leq} \sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\left(P_{h,s,a}W_{h+1}^k(s,a)\right)^2 &(69)\\
&\leq \sum_{k=1}^K C_d^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} C^\star d_h^\star(s,a)\left\|W_{h+1}^k(s_h^k,a_h^k)\right\|_\infty^2 \leq C_d^2 C^\star C_w^2 K, &(70)
\end{aligned}$$

where (i) follows from $\frac{d_h^\star(s,a)}{d_h(s,a)} \le C^\star$ (see Assumption 2.1) and the assumption $\|W_{h+1}(s_h^k,a_h^k)\|_\infty \le C_w$. Similarly, we can derive

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}_{k-1}\left[|X_{h,k}|^2\right] &\le \sum_{k=1}^K C_d^2\, \mathbb{E}_{k-1}\left[\left\|\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)} P_h^k W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a} W_{h+1}^k(s,a)\right\|^2\right] \\
&\le \sum_{k=1}^K C_d^2\, \mathbb{E}_{(s_h,a_h^k)\sim d_h^k}\, \mathbb{E}_{P_h^k\sim P_{h,s_h^k,a_h}}\left[\left\|\frac{d_h^\star(s_h^k,a_h^k)}{d_h^k(s_h,a_h)} P_h^k W_{h+1}^k(s_h^k,a_h)\right\|^2\right] \\
&= \sum_{k=1}^K C_d^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^\star(s,a)}{d_h(s,a)} d_h^\star(s,a)\, \mathbb{E}_{P_h^k\sim P_{h,s,a}}\left\|P_h W_{h+1}^k(s,a)\right\|^2 \\
&\overset{(i)}{\le} \sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\, \mathbb{E}_{P_h^k\sim P_{h,s,a}}\left\|P_h W_{h+1}^k(s,a)\right\|^2 \qquad (71) \\
&= \sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a} \left\|W_{h+1}^k(s,a)\right\|^2 \qquad (72) \\
&\le \sum_{k=1}^K C_d^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} C^\star d_h^\star(s,a) \left\|W_{h+1}^k(s,a)\right\|_\infty^2 \le C_w^2 C^\star C_d^2 K, \qquad (73)
\end{aligned}$$

where (i) follows from $\frac{d_h^\star(s,a)}{d_h(s,a)} \le C^\star$ (see Assumption 2.1) and the assumption $\left\|W_{h+1}(s_h^k,a_h^k)\right\|_\infty \le C_w$.

Plugging in the results in (67) and (69) (resps. (68) and (72)) to control $\sum_{k=1}^K |X_{h,k}|$ (resps. $\sum_{k=1}^K X_{h,k}$), we invoke Theorem B.3 with $m = \lceil \log_2 K \rceil$ and take the union bound over $h \in [H]$ to show that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^K X_{h,k} &\le \sqrt{8 \max\left\{\sum_{k=1}^K C_d C^{\star 2} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a}\left\|W_{h+1}^k(s,a)\right\|^2, \frac{C_d^2 C^\star C_w^2 K}{2^m}\right\} \log\frac{2H}{\delta}} \\
&\quad + \frac{8}{3} C_d C^\star C_w \log\frac{2H}{\delta} \\
&\le \sqrt{8 \sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a}\left\|W_{h+1}^k(s,a)\right\|^2 \log\frac{2H}{\delta}} + 6 C_d C^\star C_w \log\frac{2H}{\delta}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{k=1}^K X_{h,k} &\le \sqrt{8 \max\left\{\sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a}\left\|W_{h+1}^k(s,a)\right\|^2, \frac{C_d^2 C^\star C_w^2 K}{2^m}\right\} \log\frac{2H}{\delta}} \\
&\quad + \frac{8}{3} C_d C^\star C_w \log\frac{2H}{\delta} \\
&\le \sqrt{8 \sum_{k=1}^K C_d^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a}\left\|W_{h+1}^k(s,a)\right\|^2 \log\frac{2H}{\delta}} + 6 C_d C^\star C_w \log\frac{2H}{\delta}
\end{aligned}$$

holds simultaneously for all $h \in [H]$. □

## C. Proof of main lemmas for LCB-Q (Theorem 3.1)

### C.1. Proof of Lemma A.1

#### C.1.1. Proof of inequality (31)

To begin with, we shall control $\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(P_{h,s,a} - P_h^{k^n(s,a)}\big) V_{h+1}^{k^n(s,a)}$ by invoking Lemma B.5. Let

$$W_{h+1}^i := V_{h+1}^i,$$

which satisfies

$$\|W_{h+1}^i\|_\infty \le H =: C_W.$$

Applying Lemma B.5 with $N = N_h^k(s,a)$ reveals that, with probability at least $1-\delta$,

$$\left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(P_{h,s,a} - P_h^{k^n(s,a)}\big) V_{h+1}^{k^n(s,a)}\right| = \left|\sum_{i=1}^{k} X_i\big(s,a,h,N_h^k(s,a)\big)\right| \le c_b \sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}} \tag{74a}$$

holds simultaneously for all $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$, provided that the constant $c_b > 0$ is large enough and that $N_h^k(s,a) > 0$. If $N_h^k(s,a) = 0$, then we have the trivial bound

$$\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(P_{h,s,a} - P_h^{k^n(s,a)}\big) V_{h+1}^{k^n(s,a)} = 0. \tag{74b}$$

Additionally, from the definition $b_n = c_b\sqrt{\frac{H^3\iota^2}{n}}$, we observe that

$$\begin{cases} \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \in \left[c_b\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}, 2c_b\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}\right], & \text{if } N_h^k(s,a) > 0 \\ \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n = 0, & \text{if } N_h^k(s,a) = 0 \end{cases} \tag{75}$$

holds simultaneously for all $s,a,h,k \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, which follows directly from the property (55a) in Lemma B.1. Combining the above bounds (74) and (75), we arrive at the advertised result

$$\left|\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(P_{h,s,a} - P_h^{k^n(s,a)}\big) V_{h+1}^{k^n(s,a)}\right| \le \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n.$$

#### C.1.2. Proof of inequality (32)

Note that the second inequality of (32) holds straightforwardly as

$$V_h^{\pi}(s) \le V^{\star}(s)$$

holds for any policy $\pi$. As a consequence, it suffices to establish the first inequality of (32), namely,

$$V_h^k(s) \le V_h^{\pi^k}(s) \qquad \text{for all } (s,h,k) \in \mathcal{S}\times[H]\times[K]. \tag{76}$$

Before proceeding, let us introduce the following auxiliary index

$$k_o(h,k,s) := \max\left\{ l : l < k \text{ and } V_h^l(s) = \max_a Q_h^l(s,a) \right\} \tag{77}$$

for any $(h,k,s) \in [H]\times[K]\times\mathcal{S}$, which denotes the index of the latest episode — before the end of the $(k-1)$-th episode — in which $V_h(s)$ has been updated. In what follows, we shall often abbreviate $k_o(h,k,s)$ as $k_o(h)$ whenever it is clear from the context.

Towards establishing the relation (76), we proceed by means of an inductive argument. In what follows, we shall first justify the desired inequality for the base case when $h+1 = H+1$ for all episodes $k$ 2 $[K]$, and then use induction to complete the argument for other cases. More specifically, consider any step $h$ 2 $[H]$ in any episode $k$ 2 $[K]$, and suppose that the first inequality of (32) is satisfied for all previous episodes as well as all steps $h^0$ $h+1$ in the current episode, namely,

$$V_{h^0}^{k^0}(s) \;\; V_{h^0}^{\;\;k^0}(s) \qquad \text{for all } (k^0; h^0; s) \; 2 \; [k \quad 1] \; [H+1] \; S; \tag{78a}$$

$$V_{h^0}^{k}(s) \;\; V_{h^0}^{\;\;k}(s) \qquad \text{for all } h^0 \;\; h+1 \text{ and } s \; 2 \; S: \tag{78b}$$

We intend to justify that the following is valid

$$V_h^k(s) \;\; V_h^{\;\;k}(s) \qquad \text{for all } s \; 2 \; S; \tag{79}$$

assuming that the induction hypothesis (78) holds.

Step 1: base case. Let us begin with the base case when $h+1 = H+1$ for all episodes $k$ 2 $[K]$. Recognizing the fact that $V_{H+1} = V_{H+1}^k = 0$ for any and any $k$ 2 $[K]$, we directly arrive at

$$V_{H+1}^k(s) \;\; V_{H+1}^{\;\;k}(s) \qquad \text{for all } (k; s) \; 2 \; [K] \; S: \tag{80}$$

Step 2: induction. To justify (79) under the induction hypothesis (78), we decompose the difference term to obtain

$$\begin{aligned} V_h^{\;\;k}(s) \quad V_h^k(s) &= V_h^{\;\;k}(s) \quad \max \max_a Q_h^k(s; a); V_h^{k\;\;1}(s) \\ &= Q_h^{\;\;k}\; s; {}_h^k(s) \quad \max \max_a Q_h^k(s; a); V_h^{k_o(h)}(s) \; ; \end{aligned} \tag{81}$$

where the last line holds since $V_h(s)$ has not been updated during episodes $k_o(h); k_o(h)+1; \;\; ; k \quad 1$ (in view of the definition of $k_o(h)$ in (77)). We shall prove that the right-hand side of (81) is non-negative by discussing the following two cases separately.

Consider the case where $V_h^k(s) = \max_a Q_h^k(s; a)$. Before continuing, it is easily observed from the update rule in line 16 and line 16 of Algorithm 1 that: $V_h(s)$ and ${}_h(s)$ are updated hand-in-hand for every $h$. Thus, it implies that

$$_h^k(s) = \arg\max_a Q_h^k(s; a); \qquad \text{when } V_h^k(s) = \max_a Q_h^k(s; a) \tag{82}$$

holds for all $(k; h)$ 2 $[K]$ $[H]$. As a result, we express the term of interest as follows:

$$V_h^{\;\;k}(s) \quad V_h^k(s) = Q_h^{\;\;k}\; s; {}_h^k(s) \quad \max_a Q_h^k(s; a) = Q_h^{\;\;k}\; s; {}_h^k(s) \quad Q_h^k\; s; {}_h^k(s) : \tag{83}$$

To continue, we turn to controlling a more general term $Q_h^{\;\;k}(s; a) \quad Q_h^k(s; a)$ for all $(s; a)$ 2 $S$ $A$. Invoking the fact ${}_0^{N_h^k} + \sum_{n=1}^{N_h^k} {}_n^{N_h^k} = 1$ (see (25) and (26)) leads to

$$Q_h^k(s; a) = {}_0^{N_h^k} Q_h^{k}(s; a) + \sum_{n=1}^{N_h^k} {}_n^{N_h^k} Q_h^{k_n}(s; a):$$

This relation combined with (29) allows us to express the difference between $Q_h^{\;\;k}$ and $Q_h^k$ as follows

$$\begin{aligned} Q_h^{\;\;k}(s; a) \quad Q_h^k(s; a) &= {}_0^{N_h^k} Q_h^{\;\;k}(s; a) \quad Q_h^1(s; a) + \sum_{n=1}^{N_h^k} {}_n^{N_h^k} Q_h^{\;\;k}(s; a) \quad r_h(s; a) \quad V_{h+1}(s_{h+1}^{k_n}) + b_n^{k_n} \; {}_{n=1}^{i} \\ &\overset{(i)}{=} {}_0^{N_h^k} \; Q_h^{\;\;k}(s; a) \quad Q_h^1(s; a) \; + \sum_{n=1}^{N_h^k} {}_n^{N_h^k} \; {}^{h} P_{h;s;a} V_{h+1}^{\;\;k} \quad V_{h+1}^{k_n}(s_{h+1}^{k_n}) + b_n \; {}^{i} \end{aligned}$$

$$\overset{(ii)}{} \sum_{n=1}^{N_h^k} {}_{n}^{N_h^k} \Big[ P_{h;s;a} V_{h+1}^{\;\;k} \quad V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n \Big]$$

$$\overset{(iii)}{=} \sum_{n=1}^{N_h^k} {}_{n}^{N_h^k} P_{h;s;a} \quad V_{h+1}^{\;\;k} \quad V_{h+1}^{k^n} \; + \sum_{n=1}^{N_h^k} {}_{n}^{N_h^k} \Big[ P_{h;s;a} \quad P_h^{k^n} \quad V_{h+1}^{k^n} + b_n \Big]$$

$$\overset{(iv)}{} \sum_{n=1}^{N_h^k} {}_{n}^{N_h^k} \Big[ P_{h;s;a} \quad P_h^{k^n} \quad V_{h+1}^{k^n} + b_n \Big] : \tag{84}$$

Here, (i) invokes the Bellman equation $Q_h^{\;k}(s;a) = r_h(s;a) + P_{h;s;a} V_{h+1}^{\;\;k}$; (ii) holds since $Q_h^{\;k}(s;a) \quad 0 = Q_h^1(s;a)$; (iii) relies on the notaion (27); and (iv) comes from the fact

$$V_{h+1}^{\;\;k} \quad V_{h+1}^{k} \quad V_{h+1;}^{k^n}$$

owing to the induction hypothesis in (78) as well as the monotonicity of $V_{h+1}$ in (30). Consequently, it follows from (84) that

$$Q_h^{\;k}(s;a) \quad Q_h^k(s;a) \quad \sum_{n=1}^{N_h^k(s;a)} {}_{n}^{N_h^k(s;a)} \quad P_{h;s;a} \quad P_h^{k^n(s;a)} \quad V_{h+1}^{k^n(s;a)} + \sum_{n=1}^{N_h^k(s;a)} {}_{n}^{N_h^k(s;a)} b_n$$
$$\sum_{n=1}^{N_h^k(s;a)} {}_{n}^{N_h^k(s;a)} b_n \quad \sum_{n=1}^{N_h^k(s;a)} {}_{n}^{N_h^k(s;a)} \quad P_{h;s;a} \quad P_h^{k^n(s;a)} \quad V_{h+1}^{k^n(s;a)} \qquad 0 \tag{85}$$

for all state-action pair $(s;a)$, where the last inequality holds due to the bound (31) in Lemma A.1. Plugging the above result into (83) directly establishes that

$$V_h^{\;k}(s) \quad V_h^{\;k}(s) = Q_h^{\;k}\; s; {}^k(s) \quad Q_h^k \; s; {}^k(s) \quad 0: \tag{86}$$

When $V_h^{\;k}(s) = V_h^{k_o(h)}(s)$, it indicates that

$$V_h^{k_o(h)}(s) = \max_a Q_h^{k_o(h)}(s;a); \qquad {}_h^{k_o(h)}(s) = \arg\max_a Q_h^{k_o(h)}(s;a); \tag{87}$$

which follows from the definition of $k_o(h)$ in (77) and the corresponding fact in (82). We also make note of the fact that

$$_h^k(s) = {}_h^{k_o(h)}(s); \tag{88}$$

which holds since $V_h(s)$ (and hence $_h(s)$) has not been updated during episodes $k_o(h); k_o(h)+1; \; ; k \quad 1$ (in view of the definition (77)). Combining the above two results, we can show that

$$V_h^{\;k}(s) \quad V_h^{\;k}(s) = Q_h^{\;k}\; s; {}^k(s) \quad V_h^{k_o(h)}(s) = Q_h^{\;k}\; s; {}^k(s) \quad \max_a Q_h^{k_o(h)}(s;a) = Q_h^{\;k}$$
$$s; {}_h^{k_o(h)}(s) \quad {}_h Q_h^{k_o(h)} \; s; {}_h^{k_o(h)}(s)$$
$$0; \tag{89}$$

where the final line can be verified using exactly the same argument as in the previous case to show (84) and then (86). Here, we omit the proof of this step for brevity.

To conclude, substituting the relations (86) and (89) in the above two cases back into (81), we arrive at

$$V_h^{\;k}(s) \quad V_h^k(s) \quad 0$$

as desired in (79). This immediately completes the induction argument.

### C.2. Proof of Lemma A.2

Observing that Lemma A.2 would follow immediately if we could establish the following relation:

$$A_h := \sum_{k=1}^{K} \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) P_{h,s,a} \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big)}_{=:A_{h,k}} \le \sum_{k=1}^{K} \underbrace{\Big(1+\frac{1}{H}\Big) \sum_{s\in\mathcal{S}} d_{h+1}^\star(s)\big(V_{h+1}^\star(s) - V_{h+1}^k(s)\big)}_{=:B_{h,k}} + 16\sqrt{H^2 C^\star K \log} + 8HC^\star \log \, ; \tag{90}$$

the remainder of the proof is thus dedicated to proving (90).

To continue, let us first consider two auxiliary sequences $\{Y_{h,k}\}_{k=1}^K$ and $\{Z_{h,k}\}_{k=1}^K$ which are the empirical estimation of $A_{h,k}$ and $B_{h,k}$ respectively. For any time step $h$ in episode $k$, $Y_{h,k}$ and $Z_{h,k}$ are defined as follows

$$Y_{h,k} := \frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \big(V_{h+1}^\star - V_{h+1}^{k^n(s_h^k,a_h^k)}\big);$$

$$Z_{h,k} := \Big(1+\frac{1}{H}\Big) \frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \big(V_{h+1}^\star - V_{h+1}^k\big).$$

To begin with, let us establish the relationship between $\{Y_{h,k}\}_{k=1}^K$ and $\{Z_{h,k}\}_{k=1}^K$:

$$\begin{aligned}\sum_{k=1}^K Y_{h,k} &= \sum_{k=1}^K \frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \big(V_{h+1}^\star - V_{h+1}^{k^n(s_h^k,a_h^k)}\big) \\ &\overset{(i)}{=} \sum_{l=1}^K \frac{d_h^\star(s_h^l,a_h^l)}{d_h(s_h^l,a_h^l)} P_{h,s_h^l,a_h^l} \left\{ \sum_{N=N_h^l(s_h^l,a_h^l)}^{N_h^K(s_h^l,a_h^l)} \eta^N_{N_h^l(s_h^l,a_h^l)} \right\} \big(V_{h+1}^\star - V_{h+1}^l\big) \qquad (91) \\ &\le \Big(1+\frac{1}{H}\Big) \sum_{k=1}^K \frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \big(V_{h+1}^\star - V_{h+1}^k\big) = \sum_{k=1}^K Z_{h,k}. \qquad (92)\end{aligned}$$

Here, (i) holds by replacing $k^n(s_h^k,a_h^k)$ with $l$ and gathering all terms that involve $V_{h+1}^\star - V_{h+1}^{k^n(s_h^k,a_h^k)}$; in the last line, we have invoked the property $\sum_{N=n}^{N_h^K(s,a)} \eta_n^N \le \sum_{N=n}^{\infty} \eta_n^N = 1 + 1/H$ (see (55b)) together with the fact $V_{h+1}^\star - V_{h+1}^l \ge 0$ (see Lemma A.1), and have further replaced $l$ with $k$.

With this relation in hand, to verify (90), we further decompose $A_h$ into several terms

$$\begin{aligned} A_h = \sum_{k=1}^K A_{h,k} &= \sum_{k=1}^K Y_{h,k} + \sum_{k=1}^K (A_{h,k} - Y_{h,k}) \overset{(i)}{\le} \sum_{k=1}^K Z_{h,k} + \sum_{k=1}^K (A_{h,k} - Y_{h,k}) \\ &= \sum_{k=1}^K B_{h,k} + \sum_{k=1}^K (Z_{h,k} - B_{h,k}) + \sum_{k=1}^K (A_{h,k} - Y_{h,k}) \end{aligned} \tag{93}$$

where (i) follows from (92).

As a result, it remains to control $\sum_{k=1}^K (Z_{h,k} - B_{h,k})$ and $\sum_{k=1}^K (A_{h,k} - Y_{h,k})$ separately in the following.

**Step 1: controlling $\sum_{k=1}^K (A_{h,k} - Y_{h,k})$.** We shall first control this term by means of Lemma B.6. Specifically, consider

$$W_{h+1}^k(s,a) := \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big); \qquad C_d := 1 \tag{94}$$

which satisfies

$$W^k_{h+1}(s;a)_1 \quad \sum_{n=1}^{N^k_h(s;a)} \quad N^k_n{}^{(s;a)}_h \quad V_{h+1}{}^?{}_1 + V^{k^n(s;a)}_{h+1} \quad {}_1 \qquad 2H = : C_w : \tag{95}$$

Here we use the fact that $N^k_0{}_h + \sum_{n=1}^{N^k_h} N^k_n{}_h = 1$ (see (25) and (26)). Then, applying Lemma B.6 with (94), we have with probability at least 1 , the following inequality holds true

$$\sum_{k=1}^{K}(A_{h;k} \quad Y_{h;k}) = \sum_{k=1}^{K} X_{h;k} \quad \sqrt{8C^2C^?_d \sum_{k=1}^{K} \sum_{(s;a)2SA} d_h^?(s;a)\,P_{h;s;a}W^k_{h+1}(s;a)^2 \log\frac{H}{} + 2C_dC^?C_w\log\frac{H}{}}$$

$$\overset{(i)}{} \sqrt{8C^? \sum_{k=1}^{K} W^k_{h+1}(s;a)_1^2 \log\frac{H}{}} + 4HC^?\log\frac{H}{}$$

$$8\sqrt{H^2C^?K\log\frac{H}{}} + 4HC^?\log\frac{H}{}; \tag{96}$$

where (i) holds by $P_{h;s;a}{}_1 = 1$.

Step 2: controlling $\sum_{k=1}^{K}(Z_{h;k} \quad B_{h;k})$. Similarly, we shall control $\sum_{k=1}^{K}(Z_{h;k} \quad B_{h;k})$ by invoking Lemma B.6. Recalling that

$$Z_{h;k} \quad B_{h;k} = \quad 1+\frac{1}{H} \quad \frac{d_h^?(s^k_h;a^k_h)}{d_h(s^k_h;a^k_h)}P_{h;s^k_h;a^k_h}\quad V^?_{h+1} \quad V^k_{h+1} \qquad 1+\frac{1}{H}\sum_{s2S} d^?_{h+1}(s)\quad V^?_{h+1}(s) \quad V^k_{h+1}(s) \quad ; \tag{97}$$

consider

$$W^k_{h+1}(s;a) := V^?_{h+1} \quad V^k_{h+1}; \qquad C_d := \quad 1+\frac{1}{H} \quad 2 \tag{98}$$

which satisfies

$$W^k_{h+1}(s;a)_1 \quad V_{h+1}{}^?_1 + V_{h+1}{}^k_1 \quad 2H = C_w: \quad : \tag{99}$$

Again, in view of Lemma B.6, we have with probability at least 1 ,

$$\sum_{k=1}^{K}(B_{h;k} \quad Z_{h;k}) = \sum_{k=1}^{K} X_{h;k} \quad \sqrt{8C^2_dC^? \sum_{k=1}^{K} \sum_{(s;a)2SA} d_h^?(s;a)\;P_{h;s;a}W^k_{h+1}(s;a)^2 \log\frac{H}{} + 2C_dC^?C_w\log\frac{H}{}}$$

$$\overset{(i)}{} \sqrt{32C^? \sum_{k=1}^{K} W^k_{h+1}(s;a)_1^2 \log\frac{H}{}} + 8HC^?\log\frac{H}{}$$

$$16\sqrt{H^2C^?K\log\frac{H}{}} + 8HC^?\log\frac{H}{}; \tag{100}$$

where (i) holds by $P_{h;s;a}{}_1 = 1$.

Step 3: putting together. Substitution results in (96) and (100) back into (93) completes the proof of Lemma A.2 by

$$A_h \quad \sum^{K} B_{h;k} + \sum_{k=1}^{K}(Z_{h;k} \quad B_{h;k}) + \sum_{k=1}^{K}(A_{h;k} \quad Y_{h;k}) \quad \sum_{k=1}^{K} B_{h;k} + 24\sqrt{H^2C^?K\log\frac{H}{}} + 12HC^?\log\frac{H}{} : {}_{k=1}$$

### C.3. Proof of Lemma A.3

Recall that the term of interest in (42) is given by

$$\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\left(24\sqrt{H^{2}C^{\star}K\log\frac{2H}{\delta}}+12HC^{\star}\log\frac{2H}{\delta}\right)+\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}I_h. \tag{101}$$

First, it is easily seen that

$$\Big(1+\frac{1}{H}\Big)^{h-1}\le\Big(1+\frac{1}{H}\Big)^{H}\le e\qquad\text{for every } h=1,\cdots,H; \tag{102}$$

which taken collectively with the expression of the first term in (101) yields

$$\begin{aligned}\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\left(24\sqrt{H^{2}C^{\star}K\log\frac{2H}{\delta}}+12HC^{\star}\log\frac{2H}{\delta}\right)&\le 24e\sum_{h=1}^{H}\left(\sqrt{H^{2}C^{\star}K\log\frac{2H}{\delta}}+HC^{\star}\log\frac{2H}{\delta}\right)\\&\lesssim\sqrt{H^{4}C^{\star}K\log\frac{H}{\delta}}+H^{2}C^{\star}\log\frac{H}{\delta}.\end{aligned} \tag{103}$$

As a result, it remains to control the second term in (101). Plugging the expression of $I_h$ (cf. (39)) and invoking the fact (102) give

$$\begin{aligned}\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}I_h&=\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\eta_0^{N_h^k(s,a)}H\\&\quad+2\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n\\&\le\underbrace{e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\eta_0^{N_h^k(s,a)}H}_{=:\mathcal{A}}+2e\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n}_{=:\mathcal{B}}.\end{aligned} \tag{104}$$

**Step 1: controlling the quantities $\mathcal{A}$ and $\mathcal{B}$ in (104).** We first develop an upper bound on the quantity $\mathcal{A}$ in (104). Recognizing the fact that $\eta_0^N=0$ for any $N>0$ (see (25)), we have

$$\begin{aligned}\mathcal{A}&=e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\eta_0^{N_h^k(s,a)}H\\&\le eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{k=1}^{K}\mathbb{1}\big\{N_h^k(s,a)<1\big\}\\&\le eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\frac{8\iota}{d_h^{\mu}(s,a)}+eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{k=\lceil\frac{8\iota}{d_h^{\mu}(s,a)}\rceil}^{K}\mathbb{1}\big\{N_h^k(s,a)<1\big\}\\&=eH\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}d_h^{\star}\big(s,\pi^{\star}(s)\big)\frac{8\iota}{d_h^{\mu}\big(s,\pi^{\star}(s)\big)}+eH\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}d_h^{\star}\big(s,\pi^{\star}(s)\big)\sum_{k=\lceil\frac{8\iota}{d_h^{\mu}(s,\pi^{\star}(s))}\rceil}^{K}\mathbb{1}\big\{N_h^k\big(s,\pi^{\star}(s)\big)<1\big\},\end{aligned}$$

where the last equality holds since $\pi^{\star}$ is a deterministic policy (so that $d_h^{\star}(s,a)=0$ only when $a\neq\pi^{\star}(s)$). Recalling $\frac{d_h^{\star}(s,a)}{d_h^{\mu}(s,a)}\le C^{\star}$ under Assumption 2.1, we can further bound $\mathcal{A}$ by

$$\mathcal{A}\le 8eH^{2}SC^{\star}\iota+eH\sum_{s\in\mathcal{S}}\sum_{h=1}^{H}d_h^{\star}\big(s,\pi^{\star}(s)\big)\sum_{k=\lceil\frac{8\iota}{d_h^{\mu}(s,\pi^{\star}(s))}\rceil}^{K}\mathbb{1}\big\{N_h^k\big(s,\pi^{\star}(s)\big)<1\big\}$$

$$= 8eH^2SC^\star; \tag{105}$$

where the last inequality follows since when $k \geq \frac{8\iota}{d_h^\mu(s,a)}$, one has — with probability at least $1-\delta$ — that

$$N_h^k(s,a) \geq \frac{k d_h^\mu(s,a)}{8} \geq 1;$$

holds simultaneously for all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$ (as implied by (57a)).

Turning to the quantity $\mathcal{B}$ in (104), one can deduce that

$$\begin{aligned} \mathcal{B} &= 2e \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \\ &\lesssim \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) \sqrt{\frac{H^3\iota^2}{N_h^k(s,a) \vee 1}} = \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^\star\big(s,\pi^\star(s)\big) \sqrt{\frac{H^3\iota^2}{N_h^k\big(s,\pi^\star(s)\big) \vee 1}}; \end{aligned} \tag{106}$$

where the inequality follows from inequality (75), and the last equality is valid since $\pi^\star$ is a deterministic policy. To further control the right hand side above, Lemma B.2 provides an upper bound for $\sqrt{1/\big(N_h^k\big(s,\pi^\star(s)\big) \vee 1\big)}$ which in turn leads to

$$\begin{aligned} \mathcal{B} &\lesssim \sqrt{H^3\iota^3} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^\star\big(s,\pi^\star(s)\big) \sqrt{\frac{1}{k d_h^\mu\big(s,\pi^\star(s)\big)}} \\ &\lesssim \sqrt{H^3C^\star\iota^3} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} \sqrt{d_h^\star\big(s,\pi^\star(s)\big)} \sqrt{\frac{1}{k}} \\ &\lesssim \sqrt{H^5C^\star K\iota^3} \max_h \sum_{s\in\mathcal{S}} \sqrt{d_h^\star\big(s,\pi^\star(s)\big)} \\ &\lesssim \sqrt{H^5C^\star K\iota^3} \left( \sqrt{S} \sqrt{\sum_{s\in\mathcal{S}} d_h^\star\big(s,\pi^\star(s)\big)} \right) \lesssim \sqrt{H^5SC^\star K\iota^3}; \end{aligned} \tag{107}$$

where the second inequality follows from the fact $\frac{d_h^\star(s,a)}{d_h^\mu(s,a)} \leq C^\star$ under Assumption 2.1, and the last line invokes the Cauchy-Schwarz inequality.

Taking the upper bounds on both $\mathcal{A}$ and $\mathcal{B}$ collectively establishes

$$\sum_{h=1}^{H} \left(1+\frac{1}{H}\right)^{h-1} I_h \leq \mathcal{A} + \mathcal{B} \lesssim H^2SC^\star\iota + \sqrt{H^5SC^\star K\iota^3}. \tag{108}$$

**Step 2: putting everything together.** Combining (103) and (108) allows us to establish that

$$\sum_{h=1}^{H} \left(1+\frac{1}{H}\right)^{h-1} \left( I_h + 16\sqrt{H^2C^\star K \log\frac{2H}{\delta}} + 8HC^\star \log\frac{2H}{\delta} \right) \lesssim H^2SC^\star\iota + \sqrt{H^5SC^\star K\iota^3};$$

as advertised.

## D. Proof of lemmas for LCB-Q-Advantage (Theorem 4.1)

**Additional notation for LCB-Q-Advantage.** Let us also introduce, and remind the reader of, several notation of interest in Algorithm 5 as follows.

- $N_h^k(s,a)$ (resp. $N_h^{(m,t)}(s,a)$) denotes the value of $N_h(s,a)$ — the number of episodes that has visited $(s,a)$ at step $h$ at the beginning of the $k$-th episode (resp. the beginning of $t$-th episode of the $m$-th epoch); for the sake of conciseness, we shall often abbreviate $N_h^k = N_h^k(s,a)$ (resp. $N_h^{(m,t)} = N_h^{(m,t)}(s,a)$) when it is clear from context.

$L_m = 2^m$: the total number of in-epoch episodes in the $m$-th epoch.

$k^n_h(s;a)$: the index of the episode in which $(s;a)$ is visited for the $n$-th time at time step $h$; $(m^n_h(s;a); t^n_h(s;a))$ denote respectively the index of the epoch and that of the in-epoch episode in which $(s;a)$ is visited for the $n$-th time at step $h$; for the sake of conciseness, we shall often use the shorthand $k^n = k^n_h(s;a)$, $(m^n;k^n) = (m^n_h(s;a); k^n_h(s;a))$ whenever it is clear from context.

$Q^k_h(s;a)$, $Q^{\mathrm{LCB};k}_h(s;a)$, $\overline{Q}^k_h(s;a)$ and $V^k_h(s)$ are used to denote $Q_h(s;a)$, $Q^{\mathrm{LCB}}_h(s;a)$, $\overline{Q}_h(s;a)$, and $V_h(s)$ at the beginning of the $k$-th episode, respectively.

$\overline{V}^k_h(s)$; $\overline{V}^{\mathrm{next};k}_h(s)$; ${}^k_h(s;a)$; ${}^{\mathrm{next};k}_h(s;a)$ denote the values of $\overline{V}_h(s)$; $\overline{V}^{\mathrm{next}}_h(s)$; $\overline{\phantom{x}}_h(s;a)$ and ${}^{\mathrm{next}}_h(s;a)$ at the begin-ning of the $k$-th episode, respectively.

$N^{(m;t)}_h(s;a)$ represents $N_h(s;a)$ at the beginning of the $t$-th in-epoch episode in the $m$-th epoch.

$N^{\mathrm{epo};m}_h(s;a)$ denotes $N^{(m;L_m+1)}_h(s;a)$, representing the number of visits to $(s;a)$ in the entire duration of the $m$-th epoch.

$[{}^{\mathrm{ref};k}_h; {}^{\mathrm{ref};k}_h; {}^{\mathrm{adv};k}_h; {}^{\mathrm{adv};k}_h; {}^k_h; B^k_h; b^k_h]$: the values of $[{}^{\mathrm{ref}}_h; {}^{\mathrm{ref}}_h; {}^{\mathrm{adv}}_h; {}^{\mathrm{adv}}_h; {}_h; B_h; b_h]$ at the beginning of the $k$-th episode, respectively.

In addition, for a fixed vector $V \; 2 \; \mathbb{R}^{jSj}$, let us define a variance parameter with respect to $P_{h;s;a}$ as follows

$$\mathrm{Var}_{h;s;a}(V) := \mathop{\mathbb{E}}_{s^0 P_{h;s;a}} \Big[ V(s^0) \quad P_{h;s;a}V \Big]^2 = P_{h;s;a}(V^2) \quad (P_{h;s;a}V)2: \tag{109}$$

This notation will be useful in the subsequent proof. We remind the reader that there exists a one-to-one mapping between the index of the episode $k$ and the index pair $(m;t)$ (i.e., the epoch $m$ and in-epoch episode $t$), as specified in (45). In the following, for any episode $k$, we recall the expressions of $\overline{V}_{h+1}$ and $\overline{\phantom{x}}_h$ (which is the running mean of $V_{h+1}$).

Recalling the update rule of $\overline{V}_h$ and $\overline{V}^{\mathrm{next}}_h$ in line 34 and line 35 of Algorithm 5, we observe that both the reference values for the current epoch $\overline{V}_h$ and for the next epoch $\overline{V}^{\mathrm{next}}_h$ remain unchanged within each epoch. Additionally, for any epoch $m$, $\overline{V}_h$ takes the value of $\overline{V}^{\mathrm{next}}_h$ in the previous $(m \quad 1)$-th epoch; namely, for any episode $k$ happening in the $m$-th epoch, we have

$$\overline{V}^k_h = \overline{V}^{\mathrm{next};k^0}_h \tag{110}$$

for all episode $k^0$ within the $(m \quad 1)$-th epoch.

${}^k_h$ serves as the estimate of $P_{h;s;a}\overline{V}^k_{h+1}$ constructed by the samples in the previous $(m \quad 1)$-th epoch (collected by updating $\overline{\phantom{x}}^{\mathrm{next}}_h$). Recall the update rule of $\overline{\phantom{x}}_h$ in line 34 and line 29 of Algorithm 5: for any $(s;a;h) \; 2 \; S \quad A \quad [H]$, we can write ${}_h$ as $\overline{\phantom{x}}^k$

$$\begin{aligned} {}^k_h(s;a) &= \overline{\phantom{x}}^{(m;1)}_h(s;a) = \overline{\phantom{x}}^{\mathrm{next};(m;1)}_h(s;a) = \overline{\phantom{x}}^{\mathrm{next};(m \quad 1;L_{m \quad 1})}_h(s;a) \; P^{N^{(m;1)}_h} \\ &= \frac{ {}_{i=N^{(m \quad 1;1)}_h+1}^{h} \overline{V}^{\mathrm{next};k^i}_{h+1} \; (s^{k^i}_{h+1}) }{ N^{\mathrm{epo};m \quad 1}_h(s;a) \_ 1 } = \frac{ P^{N^{(m;1)}_h}_{i=N^{(m \quad 1;1)}_h+1} \overline{V}^k_{h+1}(s^{k^i}_{h+1}) }{ N^{\mathrm{epo};m \quad 1}_h(s;a) \_ 1 }; \end{aligned} \tag{111}$$

where the last equality follows from (110) using the fact that the indices of episodes in which $(s;a)$ is visited within the $(m \quad 1)$-th epoch are $f i \; : i = N^{(m \quad 1;1)}_h + 1; N^{(m \quad 1;1)}_h + 2; \; ; N^{(m;1)}_h g$.

Finally, according to the update rules of ${}^{\mathrm{adv};k^{n+1}}_h(s_h^k;a_h^k)$ and ${}^{\mathrm{adv};k^{n+1}}_h(s_h^k;a_h^k)$ in lines 11-12 of Algorithm 3, we have

$${}^{\mathrm{adv};k^{n+1}}_h(s_h^k;a_h^k) = {}^{\mathrm{adv};k^n+1}_h(s^k_h;a^k_h) = (1 \quad {}_n)\,{}^{\mathrm{adv};k^n}_h(s_h;a^k_h) + {}_n \; V^k_{h+1}(s^{k^n}_{h+1}) \quad \overline{V}^k_{h+1}(s^{k^n}_{h+1}); \quad {}_{k^n}$$

$$ {}^{\mathrm{adv};k^{n+1}}_{h}(s_h; a_h) = {}^{\mathrm{adv};k^n+1}_{h}(s_h; a_h) = (1 \quad {}_n)^{\mathrm{adv};k^n}_{h}(s_h; a_h) + {}_n \; {}_{k}V_{h+1}(s_{h+1}) \quad {}_{k}\overline{V}_{h+1}(s_{h+1})^2 : {}_{k^n} $$

Applying this relation recursively and invoking the definitions of ${}_n^{N_h^k}$ in (25) give

$$ {}^{\mathrm{adv};k^{N_h^k}+1}_{h}(s; a) = \sum_{n=1}^{N_h^k} {}^{N_h^k}_{n} P_h^n \; {}_{k}V_{h+1}{}_{k^n} \quad \overline{V}_{h+1;k^n} ; \quad {}^{\mathrm{adv};k^{N_h^k}+1}_{h}(s; a) = \sum_{n=1}^{N_h^k} {}^{N_h^k}_{n} P_h^n \; {}_{k}V_{h+1}{}_{k^n} \quad \overline{V}_{h+1}{}_{k^n}^{2} : \qquad (112) $$

Similarly, according to the update rules of ${}^{\mathrm{ref};k^{n+1}}_{h}(s; a)$ and ${}^{\mathrm{ref};k^{n+1}}_{h}(s; a)$ in lines 9-10 of Algorithm 3, we obtain

$$ {}^{\mathrm{ref};k^{n+1}}_{h}(s; a) = {}^{\mathrm{ref};k^n+1}_{h}(s; a) = 1 \quad \frac{1}{n} {}^{\mathrm{ref};k^n}_{h}(s; a) + \frac{1}{n} V_{h+1}^{\mathrm{next};}{}_{k^n}(s_{h+1}); \quad {}^{\mathrm{ref};k^{n+1}}_{h}(s; a) $$

$$ = {}^{\mathrm{ref};k^n+1}_{h}(s; a) = 1 \quad \frac{1}{n} {}^{\mathrm{ref};k^n}_{h}(s; a) + \frac{1}{n} V_{h+1}^{\mathrm{next};}{}_{k^n}(s_{h+1})^2 : $$

Simple recursion leads to

$$ {}^{\mathrm{ref};k^{N_h^k}+1}_{h}(s; a) = \frac{1}{N_h^k} \sum_{n=1}^{N_h^k} P_h^n \; \overline{V}^{\mathrm{next};k^n}_{h+1} ; \quad {}^{\mathrm{ref};k^{N_h^k}+1}_{h}(s; a) = \frac{1}{N_h^k} \sum_{n=1}^{N_h^k} P_h^n \; \overline{V}^{\mathrm{next};k^n}_{h+1}{}^{2} : \qquad (113) $$

### D.1. Proof of Lemma A.4

Akin to the proof of Lemma A.1, the second inequality of (47) holds trivially since

$$ V_h(s) \quad V_h^{?}(s) $$

holds for any policy . Thus, it suffices to focus on justifying the first inequality of (47), namely,

$$ V_h^k(s) \quad V_h^{\;k}(s) \qquad 8(k; h; s) \; 2 \; [K] \quad [H] \quad S; \qquad (114) $$

which we shall prove by induction.

Step 1: introducing the induction hypothesis. For notational simplicity, let us define

$$ k_o(h; k; s) \coloneqq \max \Big\{ l : l < k \text{ and } V_h^l(s) = \max_a \max \Big\{ Q_h^{\mathrm{LCB};l}(s; a); \overline{Q}_h^l(s; a) \Big\}\Big\} \qquad (115) $$

for any $(h; k; s)$ 2 $[H]$ $[K]$ $S$. Here, $k_o(h; k; s)$ denotes the index of the latest episode — right at the end of the (k 1)-th episode — in which $V_h(s)$ has been updated, which shall be abbreviated as $k_o(h)$ whenever it is clear from context.

In what follows, we shall first justify the advertised inequality for the base case where h = H + 1 for all episodes k 2 [K], followed by an induction argument. Regarding the induction part, let us consider any k 2 [K] and any h 2 [H], and suppose that

$$ V_{h^0}^{\;0}(s) \quad V_{h^0}^{\;k^0}(s) \qquad \text{for all } (k^0; h^0; s) \; 2 \; [k \quad 1] \; [H + 1] \; S; \qquad (116a) $$

$$ V_{h^0}^{k}(s) \quad V_{h^0}^{\;k}(s) \qquad \text{for all } h^0 \quad h + 1 \text{ and } s \; 2 \; S: \qquad (116b) $$

We intend to justify

$$ V_h^k(s) \quad V_h^{\;k}(s) \qquad 8s \; 2 \; S; \qquad (117) $$

assuming that the induction hypotheses (116) hold.

**Step 2: controlling the confident bound** $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} b_h^{k^n+1}$. Before proceeding, we first introduce an auxiliary result on bounding $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}$, which plays a crucial role. For any $(s,a)$, it is easily seen that

$$N_h^k(s,a)=0 \quad\implies\quad \sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\overline{b}_h^{k^n(s,a)+1}=0. \tag{118}$$

When $N_h^k(s,a)>0$, expanding the definitions of $\overline{b}_h^{k^n+1}$ (cf. line 6 of Algorithm 3) and $\overline{B}_h^{k+1}$ (cf. line 15 of Algorithm 3) leads to

$$\begin{aligned}
&\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}\\
&=\sum_{n=1}^{N_h^k}\eta_n\prod_{i=n+1}^{N_h^k}(1-\eta_i)\left\{\Big(1-\frac{1}{\eta_n}\Big)\overline{B}_h^{k^n}(s,a)+\frac{1}{\eta_n}\overline{B}_h^{k^n+1}(s,a)\right\}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^{7/4}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}H^2\\
&=\sum_{n=1}^{N_h^k}\left(\prod_{i=n+1}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n+1}(s,a)-\prod_{i=n}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n}(s,a)\right)+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^{7/4}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}H^2\\
&\overset{\text{(i)}}{=}\sum_{n=1}^{N_h^k}\prod_{i=n+1}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n+1}(s,a)-\sum_{n=2}^{N_h^k}\prod_{i=n}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n}(s,a)+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^{7/4}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}H^2\\
&\overset{\text{(ii)}}{=}\sum_{n=1}^{N_h^k}\prod_{i=n+1}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n+1}(s,a)-\sum_{n=1}^{N_h^k-1}\prod_{i=n+1}^{N_h^k}(1-\eta_i)\overline{B}_h^{k^n+1}(s,a)+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^{7/4}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}H^2\\
&=\overline{B}_h^{k^{N_h^k}+1}(s,a)+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^{7/4}+c_b\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}H^2;
\end{aligned}\tag{119}$$

where we abuse the notation to let $\prod_{i=j+1}^{j}(1-\eta_i)=1$. Here, (i) holds since $\overline{B}_h^{k^1}(s,a)=0$, (ii) follows from the fact that $\overline{B}_h^{k^n+1}(s,a)=\overline{B}_h^{k^{n+1}}(s,a)$, since $(s,a)$ has not been visited at step $h$ during the episodes between the indices $k^n+1$ and $k^{n+1}-1$. Combining the above result in (119) with the properties $\frac{1}{(N_h^k)^{3/4}}\le\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}\le\frac{2}{(N_h^k)^{3/4}}$ and $\frac{1}{N_h^k}\le\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}\le\frac{2}{N_h^k}$ (see Lemma B.1), we arrive at

$$\overline{B}_h^{k^{N_h^k}+1}(s,a)+c_b\frac{H^{7/4}}{(N_h^k)^{3/4}}+c_b\frac{H^2}{N_h^k}\le\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}\le\overline{B}_h^{k^{N_h^k}+1}(s,a)+2c_b\frac{H^{7/4}}{(N_h^k)^{3/4}}+2c_b\frac{H^2}{N_h^k} \tag{120}$$

as long as $N_h^k(s,a)>0$.

**Step 3: base case.** Let us look at the base case with $h=H+1$ for any $k\in[K]$. Recalling the facts that $V_{H+1}^{\pi}=V_{H+1}^k=0$ for any $\pi$ and any $k\in[K]$, we reach

$$V_{H+1}^k(s)\le V_{H+1}^{\pi^k}(s)\qquad\text{for all }(k,s)\in[K]\times\mathcal{S}. \tag{121}$$

**Step 4: induction arguments.** We now turn to the induction arguments. Suppose that (116) holds for a pair $(k,h)\in[K]\times[H]$. Everything comes down to justifying (117) for time step $h$ in the episode $k$.

First, we recall the update rule of $V_h(s)$ in lines 25-26 of Algorithm 5:

$$V_h^k(s)=\max_a Q_h^k(s,a)=Q_h^k\big(s,\pi_h^k(s)\big)=\max\Big\{Q_h^{\mathrm{LCB},k}\big(s,\pi_h^k(s)\big),\overline{Q}_h^k\big(s,\pi_h^k(s)\big),Q_h^{k-1}\big(s,\pi_h^k(s)\big)\Big\}.$$

Then we shall verify (117) in three different cases.

When $V_h^k(s) = Q_h^{\mathrm{LCB};k}\big(s; _h^k(s)\big)$, the term of interest can be controlled by

$$V_h^{k}(s) \quad V_h^k(s) \overset{(i)}{=} Q_h^{k}\big(s; _h^k(s)\big) \quad Q_h^{\mathrm{LCB};k}\big(s; _h^k(s)\big) \quad 0;$$

where (i) holds since $^k$ is set to be the greedy policy such that $V_h^k(s) = Q_h^k(s; _h^k(s))$, and the last inequality follows directly from the analysis for LCB-Q (see (85)).

When $V_h^k(s) = \overline{Q}_h^k\big(s; _h^k(s)\big)$, we obtain

$$V_h^{k}(s) \quad V_h^k(s) = Q_h^{k}\big(s; _h^k(s)\big) \quad \overline{Q}_h^k\big(s; _h^k(s)\big); \tag{122}$$

To prove the term on the right-hand side of (122) is non-negative, we proceed by developing a more general lower bound on $Q_h^k(s;a) \quad \overline{Q}_h^k(s;a)$ for every $(s;a) \; 2 \; S \quad A$. Towards this, recalling the definition of $N_h^k$ and $k_h^n$, we can express

$$\overline{Q}_h^k(s;a) = \overline{Q}_h^{k^{N_h^k}+1}(s;a):$$

Thus, according to the update rule (cf. line 7 in Algorithm 3), we arrive at

$$\begin{aligned}\overline{Q}_h^k(s;a) &= \overline{Q}_h^{k^{N_h^k}+1}(s;a)\\ &= (1 \quad _{N_h^k})\overline{Q}_h^{k^{N_h^k}}(s;a) + _{N_h^k}\Big(r_h(s;a) + V_{h+1}^{k^{N_h^k}}\big(s_{h+1}^{k^{N_h^k}}\big) \quad \overline{V}_{h+1}^{k^{N_h^k}}\big(s_{h+1}^{k^{N_h^k}}\big) + _h^{k^{N_h^k}}(s;a) \quad b_h^{k^{N_h^k}+1}\Big):\end{aligned}$$

Applying this relation recursively and invoking the definitions of $_0^{N_h^k}$ and $_n^{N_h^k}$ in (25) give

$$\overline{Q}_h^k(s;a) = _0^{N_h^k}\overline{Q}_h^1(s;a) + \sum_{n=1}^{N_h^k} _n^{N_h^k}\Big(r_h(s;a) + V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) \quad \overline{V}_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) + _h^{k^n}(s;a) \quad b_h^{k^n+1}\Big): \tag{123}$$

Additionally, for any policy $^k$, the basic relation $_0^{N_h^k} + \sum_{n=1}^{N_h^k} _n^{N_h^k} = 1$ (see (26) and (25)) gives

$$Q_h^{k}(s;a) = _0^{N_h^k}Q_h^{k}(s;a) + \sum_{n=1}^{N_h^k} _n^{N_h^k}Q_h^{k}(s;a): \tag{124}$$

Combing (123) and (124) leads to

$$\begin{aligned}Q_h^{k}(s;a) \quad \overline{Q}_h^k(s;a) &= _0^{N_h^k}\Big(Q_h^{k}(s;a) \quad \overline{Q}_h^1(s;a)\Big)\\ &+ \sum_{n=1}^{N_h^k} _n^{N_h^k}\Big(Q_h^{k}(s;a) \quad r_h(s;a) \quad V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) + \overline{V}_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) \quad _h^{k^n}(s;a) + b_h^{k^n+1}\Big):\end{aligned} \tag{125}$$

Plugging in the construction of $\overline{}_h$ in (111) and invoking the Bellman equation

$$Q_h^{k}(s;a) = r_h(s;a) + P_{h;s;a}V_{h+1}^{k}; \tag{126}$$

we arrive at

$$\begin{aligned}&Q_h^{k}(s;a) \quad r_h(s;a) \quad V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) + \overline{V}_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) \quad \overline{}_h^{k^n}(s;a) + \overline{b}_h^{k^n+1}\\ &= P_{h;s;a}V_{h+1}^{k} + \overline{V}_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) \quad V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) \quad \frac{\sum_{i=N_h^{(m^n\;1;1)}+1}^{N_h^{(m^n;1)}} \overline{V}_{h+1}^{k^i}\big(s_{h+1}^{k^i}\big)}{N_h^{\mathrm{ep}} \; \mathrm{om}^n \; 1(s;a) \_ 1} + \overline{b}_h^{k^n+1}\end{aligned}$$

$$= P_{h;s;a} V^{k}_{h+1} \quad V^{k^n}_{h+1}(s^{k^n}_{h+1}) + \quad P^{k^n}_h \quad P_{h;s;a} \quad \overline{V}^{k^n}_{h+1} + \left( P_{h;s;a} \quad \frac{\sum_{i=N_h^{(m^n;1)}}^{N_h^{(m^n \; 1;1)}+1} P_h^{k^i}}{N_h^{\mathrm{epo};m^n \; 1}(s;a) \vee 1} \right) \overline{V}^{k^n}_{h+1} + b^{k^n+1}_h$$

$$= P_{h;s;a} V_{h+1}^{k} \quad V_{h+1}^{k^n} + b_h^{k^n+1} + {}_h^{k^n};$$

where

$$ {}_h^{k^n} := \quad P_h^{k^n} \quad P_{h;s;a} \quad \overline{V}^{k^n}_{h+1} \quad V^{k^n}_{h+1} \quad + \left( P_{h;s;a} \quad \frac{\sum_{i=N_h^{(m^n;1)}}^{N_h^{(m^n \; 1;1)}+1} P_h^{k^i}}{N_h^{\mathrm{ep}\;m^n \; 1}(s;a) \vee 1} \right) \overline{V}^{k^n}_{h+1}: \tag{127}$$

Inserting the above result into (125) leads to the following decomposition

$$Q_h^{k}(s;a) \quad Q_h^{k}(s;a) = {}^{N_h^k}_0 \; Q_h^{k}(s;a) \quad Q_h^{1}(s;a) + \sum_{n=1}^{N_h^k} {}^{N_h^k}_n \; P_{h;s;a} V_{h+1} \quad V^{k}_{h+1} + b_h^{k^n+1} + {}_h^{k^n} \tag{128}$$

$$\sum_{n=1}^{N_h^k} {}^{N_h^k}_n \left( b_h^{k^n+1} + {}_h^{k^n} \right); \tag{129}$$

which holds by virtue of the following facts:

(i) The initialization $\overline{Q}^1_h(s;a) = 0$ and the non-negativity of $Q_h(s;a)$ for any policy and $(s;a) \in S \times A$ lead to $Q_h^{k}(s;a) \quad \overline{Q}^1_h(s;a) = Q_h^{k}(s;a) \quad 0.$

(ii) For any episode $k^n$ appearing before $k$, making use of the induction hypothesis $V^{k}_{h+1}(s) \quad V^{k}_{h+1}(s)$ in (116b) and the monotonicity of $V_h(s)$ in (46), we obtain

$$V^{k}_{h+1}(s) \quad V^{k^n}_{h+1}(s) \quad V^{k}_{h+1}(s) \quad V^{k^n}_{h+1}(s) \quad 0: \tag{130}$$

The following lemma ensures that the right-hand side of (129) is non-negative. We postpone the proof of Lemma D.1 to Appendix D.4 to streamline our discussion.

Lemma D.1. For any $\in (0;1)$, there exists some sufficiently large constant $c_b > 0$, such that with probability at least $1 \quad$ ,

$$\sum_{n=1}^{N_h^k} {}^{N_h^k}_n \; {}_h^{k^n} \quad \sum_{n=1}^{N_h^k} {}^{N_h^k}_n \; b_h^{k^n+1}; \qquad \forall k \in [K]: \tag{131}$$

Taking this lemma together with the inequalities (122) and (129) yields

$$V_h^{k}(s) \quad V_h^{k}(s) = Q_h^{k}(s;a) \quad Q_h^{k}(s;a) \quad \sum_{n=1}^{N_h^k} {}^{N_h^k}_n \; b_h^{k^n+1} \quad \sum_{n=1}^{N_h^k} {}^{N_h^k}_n \; {}_h^{k^n} \quad 0:$$

Next, consider the case where $V_h^k(s) = Q_h^{k \; 1} \; s; {}_h^k(s)$. In view of the definition of $k_o(h)$ in (115), one has

$$V_h^k(s) = Q_h^{k \; 1} \; s; {}_h^k(s) = Q_h^{k_o(h)} \; s; {}_h^k(s) = \max \left\{ Q_h^{\mathrm{LCB};k_o(h)} \; s; {}_h^k(s) \; ; \; Q_h^{k_o(h)} \; s; {}_h^k(s) \right\};$$

since $Q_h \; s; {}^k(s)$ has not been updated during the episode $k_o(h)$ and remains unchanged in the episodes $k_o(h) + 1; k_o(h) + 2; \cdots; k \quad 1$. With this equality in hand, the term of interest in (117) can be controlled by

$$V_h^{k}(s) \quad V_h^{k}(s) = Q_h^{k}(s; {}^k_h(s)) \quad \max \left\{ Q_h^{\mathrm{LCB};k_o(h)} \; s; {}^k_h(s) ; \overline{Q}_h^{k_o(h)} \; s; {}^k_h(s) \right\} \quad 0;$$

where the last inequality follows from the facts

$$Q_h^{k}(s; {}^k_h(s)) \quad Q_h^{\mathrm{LCB};k_o(h)}(s; {}^k_h(s)) \overset{(i)}{\quad} 0;$$

$$Q_h^{k}\big(s,\pi_h^{k}(s)\big) - \overline{Q}_h^{k_o(h)}\big(s,\pi_h^{k}(s)\big) \overset{(ii)}{\geq} 0.$$

Here, (i) follows from the same analysis framework for showing (84) and (86); (ii) holds due to the following fact

$$Q_h^{k}(s,a) - \overline{Q}_h^{k_o(h)}(s,a) \geq \sum_{n=1}^{N_h^{k_o(h)}} \eta_n^{N_h^{k_o(h)}}\big(b_h^{k^n+1} + \overline{\xi}_h^{k^n}\big) \geq 0,$$

which is obtained directly by adapting (129) and then invoking (131) for $k = k_o(h)$; since the analysis follows verbatim, we omit their proofs here.

Combining the above three cases verifies the induction hypothesis in (117), provided that (116) is satisfied.

**Step 5: putting everything together.** Combining the base case in Step 3 and induction arguments in Step 4, we can readily verify the induction hypothesis in Step 1, which in turn establishes Lemma A.4.

### D.2. Proof of Lemma A.5

For every $h \in [H]$, we can decompose

$$\begin{aligned}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\big(V_h^{\star}(s) - V_h^{k}(s)\big) &\overset{(i)}{\leq} \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\Big(Q_h^{\star}\big(s,\pi_h^{\star}(s)\big) - \overline{Q}_h^{k}\big(s,\pi_h^{\star}(s)\big)\Big)\\ &= \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\big(Q_h^{\star}(s,a) - \overline{Q}_h^{k}(s,a)\big),\end{aligned} \tag{132}$$

where (i) follows from the fact $V_h^k(s) = \max_a Q_h^k(s,a) \geq \max_a \overline{Q}_h^k(s,a) \geq \overline{Q}_h^k\big(s,\pi_h^\star(s)\big)$ (see lines 25-26 in Algorithm 5). Here, the last equality is due to (35).

**Step 1: bounding $Q_h^\star(s,a) - \overline{Q}_h^k(s,a)$.** The basic relation $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} = 1$ (see (26) and (25)) gives

$$Q_h^\star(s,a) = \eta_0^{N_h^k}Q_h^\star(s,a) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}Q_h^\star(s,a), \tag{133}$$

which combined with (123) leads to

$$\begin{aligned}Q_h^\star(s,a) - \overline{Q}_h^k(s,a) &= \eta_0^{N_h^k}\big(Q_h^\star(s,a) - \overline{Q}_h^1(s,a)\big)\\ &\quad + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(Q_h^\star(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{\mu}_h^{k^n}(s,a) + b_h^{k^n+1}\Big).\end{aligned} \tag{134}$$

Invoking the Bellman optimality equation

$$Q_h^\star(s,a) = r_h(s,a) + P_{h,s,a}V_{h+1}^\star, \tag{135}$$

we can decompose $Q_h^\star(s,a) - \overline{Q}_h^k(s,a)$ similar to (128) by inserting (127) as follows:

$$\begin{aligned}Q_h^\star(s,a) - \overline{Q}_h^k(s,a) &= \eta_0^{N_h^k}\big(Q_h^\star(s,a) - \overline{Q}_h^1(s,a)\big) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s,a}\big(V_{h+1}^\star - \overline{V}_{h+1}^{k^n}\big) + b_h^{k^n+1} + \overline{\xi}_h^{k^n}\Big)\\ &\overset{(i)}{\leq} \eta_0^{N_h^k}H + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(b_h^{k^n+1} + \overline{\xi}_h^{k^n}\big) + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\big(V_{h+1}^\star - \overline{V}_{h+1}^{k^n}\big)\end{aligned}$$

$$\overset{(ii)}{\le} \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\big(V_{h+1}^\star - V_{h+1}^{k^n}\big) + 2\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1}$$

$$\overset{(iii)}{\le} \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\big(V_{h+1}^\star - V_{h+1}^{k^n}\big) + 2\Big(B_h^k(s,a) + 2c_b\frac{H^{7/4}}{(N_h^k-1)^{3/4}} + 2c_b\frac{H^2}{N_h^k-1}\Big); \tag{136}$$

where (i) follows from the initialization $\overline{Q}_h^1(s,a) = 0$ and the trivial upper bound $Q_h^\pi(s,a) \le H$ for any policy $\pi$, (ii) holds owing to the fact (see (131))

$$\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1} + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} b_h^{k^n+1} \le 2\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}; \tag{137}$$

and (iii) comes from (120) with the fact $\overline{B}_h^{k^{N_h^k}+1}(s,a) = \overline{B}_h^k(s,a)$.

**Step 2: decomposing the error in (132).** Plugging (136) into (132) and rearranging terms yield

$$\sum_{k=1}^K\sum_{s\in S} d_h^{\pi^\star}(s)\big(V_h^\star(s) - V_h^k(s)\big) \tag{138}$$

$$\le \sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + 2B_h^k(s,a) + \frac{4c_bH^{7/4}}{\big(N_h^k(s,a)-1\big)^{3/4}} + \frac{4c_bH^2}{N_h^k(s,a)-1}\right]$$

$$+ \sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big)$$

$$\le \underbrace{\sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + \frac{4c_bH^{7/4}}{\big(N_h^k(s,a)-1\big)^{3/4}} + \frac{4c_bH^2}{N_h^k(s,a)-1}\right]}_{=:J_1^h} + 2\underbrace{\sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)B_h^k(s,a)}_{=:J_2^h}$$

$$+ \sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big). \tag{139}$$

**Step 3: controlling the last term in (139).** If we could verify the following result

$$\sum_{k=1}^K\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big)$$

$$\le \underbrace{\Big(1+\frac{1}{H}\Big)\sum_{s\in S} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^\star(s) - V_{h+1}^k(s)\big)}_{=:J_3^h} + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_aH^3C^\star\sqrt{S^2}; \tag{140}$$

then combining this result with inequality (139) would immediately establish Lemma A.5. As a result, it suffices to verify the inequality (140), which shall be accomplished as follows.

**Proof of inequality (140).** We first make the observation that the left-hand side of inequality (140) is the same as what Lemma A.2 shows. Therefore, we shall establish this inequality following the same framework as in Appendix C.2. To begin with, let us recall several definitions in Appendix C.2:

$$A_h := \sum_{k=1}^K \underbrace{\sum_{(s,a)\in S\times A} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\big)}_{=:A_{h,k}};$$

$$
\begin{aligned}
B_{h,k} &:= \Big(1+\frac{1}{H}\Big)\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\big);\\
Y_{h,k} &= \frac{d_h^{\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_{h,s_h^k,a_h^k}\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)}\Big(V_{h+1}^{\star}-V_{h+1}^{k^n(s_h^k,a_h^k)}\Big);\\
Z_{h,k} &= \Big(1+\frac{1}{H}\Big)\frac{d_h^{\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\star}-V_{h+1}^{k}\big);
\end{aligned}
\tag{141}
$$

and we also remind the reader of the relation in (93) as follows

$$
A_h \le \sum_{k=1}^{K} B_{h,k} + \sum_{k=1}^{K}(Z_{h,k}-B_{h,k}) + \sum_{k=1}^{K}(A_{h,k}-Y_{h,k}). \tag{142}
$$

Equipped with these relations, we aim to control $\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})$ and $\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})$ respectively as follows.

We first bound $\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})$, which is similar to (96) (as controlled by Lemma B.6). Repeating the argument and tightening the bound from the second line of (96), we have for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$
\begin{aligned}
\sum_{k=1}^{K}(A_{h,k}-Y_{h,k}) &\le \sqrt{8C_dC^{\star}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}W_{h+1}^{k}(s,a)^2\log\frac{2H}{\delta}} + 2C_dC^{\star}C_w\log\frac{2H}{\delta}\\
&\le \sqrt{8C^{\star}\log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\Bigg[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\Big(V_{h+1}^{\star}-V_{h+1}^{k^n(s,a)}\Big)\Bigg]^2} + 4HC^{\star}\log\frac{2H}{\delta}\\
&\overset{(i)}{\le} \sqrt{8C^{\star}\log\frac{2H}{\delta}\big(36HK+3c_a^2H^6SC^{\star}\big)} + 4HC^{\star}\log\frac{2H}{\delta}\\
&\le 32\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 12c_aH^3C^{\star}\sqrt{S}^{2}.
\end{aligned}
\tag{143}
$$

Here, (i) holds by virtue of the following fact

$$
\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\Bigg[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\Big(V_{h+1}^{\star}-V_{h+1}^{k^n(s,a)}\Big)\Bigg]^2 \le 36HK+3c_a^2H^6SC^{\star}; \tag{144}
$$

whose proof is postponed to Appendix D.2.1.

Next, we turn to $\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})$, which can be bounded similar to (100) (as controlled via Lemma B.6). Repeating the argument and tightening the bound from the second line of (100) yield

$$
\begin{aligned}
\sum_{k=1}^{K}(B_{h,k}-Z_{h,k}) &\le \sqrt{8C_dC^{\star 2}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}W_{h+1}^{k}(s,a)^2\log\frac{2H}{\delta}} + 2C_dC^{\star}C_w\log\frac{2H}{\delta}\\
&\le 8\sqrt{C^{\star}\log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\Big(P_{h,s,a}\big(V_{h+1}^{\star}-V_{h+1}^{k}\big)\Big)^2} + 8HC^{\star}\log\frac{2H}{\delta}.
\end{aligned}
\tag{145}
$$

To further control (145), we have

$$
\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\Big(P_{h,s,a}\big(V_{h+1}^{\star}-V_{h+1}^{k}\big)\Big)^2 \overset{(i)}{\le} \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}\big(V_{h+1}^{\star}-V_{h+1}^{k}\big)^2
$$

$$\overset{\text{(ii)}}{\leq} H \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}\left(V_{h+1}^{\star} - V_{h+1}^{k}\right)$$

$$\overset{\text{(iii)}}{\leq} 2HK + c_a^2 H^6 SC^{\star}. \tag{146}$$

Here, (i) holds due to the non-negativity of the variance

$$\mathrm{Var}_{h,s,a}(V_{h+1}^{\star} - V_{h+1}^{k}) = P_{h,s,a}(V_{h+1}^{\star} - V_{h+1}^{k})^2 - \left(P_{h,s,a}(V_{h+1}^{\star} - V_{h+1}^{k})\right)^2 \geq 0; \tag{147}$$

(ii) follows from the basic property $V_{h+1}^{\star} - V_{h+1}^{k} \leq H$; to see why (iii) holds, we refer the reader to (154), which will be proven in Appendix D.2.1 as well. Inserting (146) back into (145) yields

$$\sum_{k=1}^{K}(B_{h,k} - Z_{h,k}) \leq 8\sqrt{C^{\star}\log\frac{2H}{\delta}\left(2KH + c_a^2H^6SC^{\star}\right)} + 8HC^{\star}\log\frac{2H}{\delta}$$

$$\leq 16\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 16c_aH^3C^{\star}\sqrt{S}. \tag{148}$$

Substituting the inequalities (143) and (148) into (142), and using the definitions in (141), we arrive at

$$A_h = \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^{\star} - V_{h+1}^{k^n(s,a)}\right)$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\left(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\right) + \sum_{k=1}^{K}(Z_{h,k} - B_{h,k}) + \sum_{k=1}^{K}(A_{h,k} - Y_{h,k})$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\left(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\right) + 32\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 12c_aH^3C^{\star}\sqrt{S^2}$$

$$+ 16\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 16c_aH^3C^{\star}\sqrt{S}$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\left(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\right) + 48\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 28c_aH^3C^{\star}\sqrt{S^2}; \tag{149}$$

which directly verifies (140) and completes the proof.

### D.2.1. Proof of inequality (144)

**Step 1: rewriting the term of interest.** We first invoke Jensen's inequality to obtain

$$\left[\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^{\star} - V_{h+1}^{k^n}\right)\right]^2 \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left[P_{h,s,a}\left(V_{h+1}^{k^n} - V_{h+1}^{\star}\right)\right]^2 \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^{k^n} - V_{h+1}^{\star}\right)^2,$$

where the first inequality follows from $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} = 1$ (see (26) and (25)), and the last inequality holds by the non-negativity of the variance $\mathrm{Var}_{h,s,a}[V_{h+1}^{\star} - V_{h+1}^{k^n}]$. This allows one to derive

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\left(V_{h+1}^{\star} - V_{h+1}^{k^n}\right)\right]^2$$

$$\leq \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n} - V_{h+1}^{\star}\right)^2$$

$$\overset{\text{(i)}}{\leq} \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\left(V_{h+1}^{\star}(s) - V_{h+1}^{k}(s)\right)^2 + 32\sqrt{H^4C^{\star}K\log\frac{2H}{\delta}} + 32H^2C^{\star}\log\frac{2H}{\delta}; \tag{150}$$

where (i) can be verified in a way similar to the proof of Lemma A.2 in Appendix C.2. We omit the details for conciseness.

**Step 2: controlling the first term in (150).** Let us introduce the following short-hand notation

$$k_{\mathsf{stop}} := c_a^2 H^5 S C^{\star};$$

and decompose the term in (150) as follows

$$\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s) \sum_{k=1}^{K} \big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big)^2 \overset{(i)}{\le} H \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big)$$

$$= H\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big) + H\sum_{k=k_{\mathsf{stop}}+1}^{K}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big). \qquad (151)$$

Here, (i) holds since $0 \le V^{\star}_{h+1}(s) - V^{k}_{h+1}(s) \le H$. The first term in (151) satisfies

$$H\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big) \le H c_a\left(\sqrt{H^5SC^{\star}k_{\mathsf{stop}}} + c_a H^2SC^{\star}\right) \le c_a^2 H^6 S C^{\star}; \qquad (152)$$

where the first inequality holds by applying the results of LCB-Q in (44) with $K = k_{\mathsf{stop}}$. The second term in (151) can be controlled as follows:

$$H\sum_{k=k_{\mathsf{stop}}+1}^{K}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big) \le HK\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k_{\mathsf{stop}}+1}_{h+1}(s)\big)$$

$$\le HK\frac{1}{k_{\mathsf{stop}}}\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big)$$

$$\le HK\, c_a\left(\sqrt{\frac{H^5SC^{\star}}{k_{\mathsf{stop}}}} + \frac{c_a H^2SC^{\star}}{k_{\mathsf{stop}}}\right) \le 2HK; \qquad (153)$$

where the first and the second inequalities hold by the monotonicity property $V^{k+1}_{h+1} \ge V^{k}_{h+1}$ introduced in (46), and the final inequality follows from applying (44).

Inserting the results in (152) and (153) into (151) yields

$$\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s) \sum_{k=1}^{K} \big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big)^2 \le H \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1}(s) - V^{k}_{h+1}(s)\big) \le 2HK + c_a H^6 S C^{\star}. \qquad (154)$$

**Step 3: combining the above results.** Inserting the above result (154) back into (150), we reach:

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\star}_h(s,a)\left[\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} P_{h,s,a}\big(V^{\star}_{h+1} - V^{k^n}_{h+1}\big)\right]^2$$

$$\le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d^{\star}_{h+1}(s)\big(V^{\star}_{h+1} - V^{k}_{h+1}\big)^2 + 32\sqrt{H^4C^{\star}K\log\frac{2H}{\delta}} + 32H^2C^{\star}\log\frac{2H}{\delta}$$

$$\overset{(i)}{\le} 4HK + 2c_a^2H^6SC^{\star} + 32\sqrt{H^4C^{\star}K\log\frac{2H}{\delta}} + 32H^2C^{\star}\log\frac{2H}{\delta}$$

$$\overset{(ii)}{\le} 36HK + 3c_a^2H^6SC^{\star}; \qquad (155)$$

where (i) holds due to (154) and $1 + \frac{1}{H} \le 2$, and (ii) results from the Cauchy-Schwarz inequality.

### D.3. Proof of Lemma A.6

We shall verify the three inequalities in (54) separately.

### D.3.1. Proof of inequality (54a)

We start by rewriting the term of interest using the expression of $J_h^1$ in (51) as

$$\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}J_h^1$$

$$=\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\left[\eta_0^{N_h^k(s,a)}H+\frac{4c_bH^{7/4}}{\big(N_h^k(s,a)\vee 1\big)^{3/4}}+\frac{4c_bH^2}{N_h^k(s,a)\vee 1}\right]$$

$$=\underbrace{\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\eta_0^{N_h^k(s,a)}H}_{=:J_1^1}+\underbrace{\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\frac{4c_bH^{7/4}}{\big(N_h^k(s,a)\vee 1\big)^{3/4}}}_{=:J_1^2}$$

$$+\underbrace{\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\frac{4c_bH^2}{N_h^k(s,a)\vee 1}}_{=:J_1^3}. \tag{156}$$

Invoking (105) and (102) yields

$$J_1^1\lesssim H^2SC^\star. \tag{157}$$

In terms of $J_1^2$, one has

$$J_1^2=\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\frac{4c_bH^{7/4}}{\big(N_h^k(s,a)\vee 1\big)^{3/4}}$$

$$\overset{(i)}{\lesssim}H^{7/4}\sum_{h=1}^H\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\frac{1}{\big(kd_h^{\mathsf{b}}(s,a)\big)^{3/4}}$$

$$\overset{(ii)}{\lesssim}H^{7/4}(C^\star)^{3/4}\sum_{h=1}^H\sum_{k=1}^K\frac{1}{k^{3/4}}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\big(d_h^\star(s,a)\big)^{\frac14}$$

$$=H^{7/4}(C^\star)^{3/4}\sum_{h=1}^H\sum_{k=1}^K\frac{1}{k^{3/4}}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}\big(a=\pi_h^\star(s)\big)\big(d_h^\star(s,a)\big)^{\frac14};$$

where (i) holds due to (102) and $\frac{1}{N_h^k(s,a)\vee 1}\le\frac{8}{kd_h^{\mathsf{b}}(s,a)}$ from Lemma B.2, and (ii) follows from the definition of $C^\star$ in Assumption 2.1. A direct application of Hölder's inequality leads to

$$J_1^2\lesssim H^{7/4}(C^\star)^{\frac34}\sum_{h=1}^H\sum_{k=1}^K\frac{1}{k^{\frac34}}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}\big(a=\pi_h^\star(s)\big)\right)^{3/4}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\right)^{1/4}$$

$$\overset{(iii)}{\lesssim}H^{7/4}(SC^\star)^{\frac34}\sum_{h=1}^H\sum_{k=1}^K\frac{1}{k^{\frac34}}\lesssim H^{2.75}(SC^\star)^{\frac34}K^{\frac14}; \tag{158}$$

where (iii) follows since $\pi^\star$ is assumed to be a deterministic policy.

Similarly, we can derive an upper bound on $J_1^3$ as follows:

$$J_1^3=\sum_{h=1}^H\Big(1+\frac{1}{H}\Big)^{h-1}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\frac{4c_bH^2}{N_h^k(s,a)\vee 1}$$

$$\overset{(i)}{\lesssim}H^2\sum_{h=1}^H\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^\star(s,a)}{kd_h^{\mathsf{b}}(s,a)}\lesssim H^3SC^\star; \tag{159}$$

where (i) follows from the result in (102) and the fact $\frac{N_h^k(s,a)\vee 1}{k}\gtrsim d_h^{\star}(s,a)$ (cf. Lemma B.2), and the last relation results from the definition of $C^\star$ (cf. Assumption 2.1) and the assumption that $\pi^\star$ is a deterministic policy.

Putting the preceding results (157), (158) and (159) together, we conclude that

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^1 \lesssim H^{2.75}(SC^\star)^{\frac{1}{4}}K^{\frac{1}{4}}\iota^2+H^3SC^\star\iota^3. \tag{160}$$

### D.3.2. Proof of inequality (54b)

Making use of the definition of $\overline{B}_h^k(s,a)$ (cf. (14)) in the expression of $J_h^2$ (cf. (51)), we obtain

$$\begin{aligned}
\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^2 &= 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\overline{B}_h^k(s,a)\\
&= 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}c_b\sqrt{H\iota}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s,a)-\big(\mu_h^{\mathrm{adv},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}}\\
&\quad+2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}c_b\sqrt{\iota}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s,a)-\big(\mu_h^{\mathrm{ref},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}}\\
&\lesssim \underbrace{\sqrt{H\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s,a)-\big(\mu_h^{\mathrm{adv},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}}}_{=:J_2^1}\\
&\quad+\underbrace{\sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s,a)-\big(\mu_h^{\mathrm{ref},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}}}_{=:J_2^2};
\end{aligned}\tag{161}$$

where the last inequality follows from (102). In the following, we shall look at the two terms in (161) separately.

**Step 1: controlling $J_2^1$.** Recalling the expressions of $\sigma_h^{\mathrm{adv},k}(s,a)=\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s,a)$ in (112), we observe that the main part of $J_2^1$ in (161) satisfies

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s,a)-\big(\mu_h^{\mathrm{adv},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}} \lesssim \sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{d_h^\star(s,a)}\sqrt{\frac{d_h^\star(s,a)\sigma_h^{\mathrm{adv},k}(s,a)}{kd_h(s,a)}}\\
&\lesssim\sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2}{kd_h(s,a)}}\,d_h^\star(s,a)\\
&\overset{(i)}{\lesssim}\sqrt{C^\star\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{1}{k}\mathbb{1}\big(a=\pi_h^\star(s)\big)d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2}\\
&\overset{(ii)}{\le}\sqrt{C^\star\iota}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}\big(a=\pi_h^\star(s)\big)\frac{1}{k}}\\
&\lesssim\sqrt{HSC^\star\iota^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2};
\end{aligned}\tag{162}$$

where the first inequality is due to the fact $\frac{N_h^k(s,a)}{k-1} \geq \frac{d_h^{\mathsf{b}}(s,a)}{8}$ from Lemma B.2, (i) follows from the definition of $C^\star$ in Assumption 2.1 and (35), and (ii) follows from the Cauchy-Schwarz inequality. To continue, we claim the following bound holds, which will be proven in Appendix D.3.4:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\Big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\Big)^2\\
&\qquad\lesssim H^2\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\Big(V_h^\star(s)-V_h^k(s)\Big)+K+H^5\sqrt{S}C^{\star 2}.
\end{aligned}\tag{163}$$

Combining the above inequality with (162), we arrive at

$$\begin{aligned}
J_2^1 &\lesssim \sqrt{H^2SC^{\star 3}}\sqrt{H^2\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\Big(V_h^\star(s)-V_h^k(s)\Big)+K+H^5\sqrt{S}C^{\star 2}}\\
&\lesssim \sqrt{H^4SC^{\star 3}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^\star(s)\Big(V_h^\star(s)-V_h^k(s)\Big)}+\sqrt{H^2SC^{\star 3}K}+H^{3.5}SC^{\star 2.5}.
\end{aligned}\tag{164}$$

**Step 2: controlling $J_2^2$.** Recalling the expressions of $\mu_h^{\mathrm{ref},k+1}(s,a) = \mu_h^{\mathrm{ref},k^{N_h^k}+1}(s,a)$ and $\sigma_h^{\mathrm{ref},k+1}(s,a) = \sigma_h^{\mathrm{ref},k^{N_h^k}+1}(s,a)$ in (113) to $J_2^2$ in (161), we can deduce that

$$\begin{aligned}
J_2^2 &= \sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s,a)-\big(\mu_h^{\mathrm{ref},k}(s,a)\big)^2}{N_h^k(s,a)\vee 1}}\\
&= \sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\Big(\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})\Big)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}}_{=:F_{h,k}}
\end{aligned}\tag{165}$$

We further decompose and bound $F_{h,k}$ as follows:

$$\begin{aligned}
F_{h,k} &\overset{(i)}{\leq} \sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^\star(s_{h+1}^{k^n})\Big)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}\\
&= \sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^\star(s_{h+1}^{k^n})\Big)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2+\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}\\
&\overset{(ii)}{\leq} \underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^\star(s_{h+1}^{k^n})\Big)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}}_{G_{h,k}}+\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}2H\Big(V_{h+1}^\star(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})\Big)}{N_h^k(s,a)\vee 1}}}_{=:L_{h,k}},
\end{aligned}\tag{166}$$

where (i) follows from the fact that for some $k'\in[K]$, $\overline{V}_{h+1}^{\mathrm{next},k^n} = V_{h+1}^{k'} \leq V_{h+1}^\star$ (see the update rule of $\overline{V}^{\mathrm{next}}$ in line 35 and the fact in (47)), and (ii) holds due to the fact that

$$\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2 \leq 2H\,\frac{\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^\star(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathrm{next},k^n}(s_{h+1}^{k^n})\Big)}{N_h^k(s,a)\vee 1}.$$

Inserting (166) back into (165), we arrive at

$$
\begin{aligned}
J_2^2 &\lesssim \sqrt{\ }\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\,(G_{h,k}+L_{h,k}) \\
&\overset{(i)}{\lesssim} \sqrt{\ }\sqrt{H^3SC^{\star}K^4 + H^4SC^{\star 3}} + \sqrt{H^3SC^{\star}K^2 + H^{2.5}SC^{\star 3}} \lesssim \sqrt{H^3SC^{\star}K^5 + H^4SC^{\star 4}};
\end{aligned}
\tag{167}
$$

where (i) follows from the following facts

$$
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\,L_{h,k} \lesssim \sqrt{H^3SC^{\star}K^4} + H^4SC^{\star 3}; \tag{168}
$$

$$
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\,G_{h,k} \lesssim \sqrt{H^3SC^{\star}K^2} + H^{2.5}SC^{\star 3}. \tag{169}
$$

We postpone the proofs of (168) and (169) to Appendix D.3.5 and Appendix D.3.6, respectively.

**Putting the bounds together.** Substitute (164) and (167) back into (161) to yield

$$
\begin{aligned}
\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1} J_h^2 &\lesssim \sqrt{H^4SC^{\star 3}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\big(V_h^{\star}(s)-V_h^{k}(s)\big)} + \sqrt{H^2SC^{\star}K^3} + H^{3.5}SC^{\star 2.5} \\
&\quad + \sqrt{H^3SC^{\star}K^5} + H^4SC^{\star 4} \\
&\lesssim \sqrt{H^4SC^{\star 3}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\star}(s)\big(V_h^{\star}(s)-V_h^{k}(s)\big)} + \sqrt{H^3SC^{\star}K^5} + H^4SC^{\star 4}.
\end{aligned}
$$

### D.3.3. Proof of inequality (54c)

Invoking inequality (102) directly leads to

$$
\sum_{h=1}^{H}\Big(1+\frac{1}{H}\Big)^{h-1}\left(48\sqrt{HC^{\star}K\log\frac{2H}{\delta}} + 28c_aH^3C^{\star}\sqrt{S}^{2}\right) \lesssim \sqrt{H^3C^{\star}K\log\frac{2H}{\delta}} + H^4C^{\star}\sqrt{S}^{2}
$$

as claimed.

### D.3.4. Proof of inequality (163)

We shall control the term in (163) in a way similar to the proof of Lemma A.2 in Appendix C.2.

**Step 1: decomposing the terms of interest.** Akin to Appendix C.2, let us introduce the terms of interest and definitions as follows:

$$
\begin{aligned}
A_h &:= \sum_{k=1}^{K}\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\Big(P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2}_{=:A_{h,k}}; \\
B_{h,k} &:= \Big(1+\frac{1}{H}\Big)\sum_{s\in\mathcal{S}} d_{h+1}^{\star}(s)\Big(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\Big)^2; \\
Y_{h,k} &= \frac{d_h^{\star}(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}\sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)}\Big(P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2;
\end{aligned}
$$

$$Z_{h,k} = \left(1+\frac{1}{H}\right)\frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k\left(V_{h+1}^k-\overline{V}_{h+1}^k\right)^2. \tag{170}$$

With these definitions in place, we directly adapt the argument in (93) to arrive at

$$A_h \le \sum_{k=1}^K B_{h,k} + \sum_{k=1}^K (Z_{h,k} - B_{h,k}) + \sum_{k=1}^K (A_{h,k} - Y_{h,k}). \tag{171}$$

As a consequence, it remains to control $\sum_{k=1}^K (Z_{h,k} - B_{h,k})$ and $\sum_{k=1}^K (A_{h,k} - Y_{h,k})$ separately.

**Step 2: controlling $\sum_{k=1}^K (A_{h,k} - Y_{h,k})$.** To control $\sum_{k=1}^K (A_{h,k} - Y_{h,k})$, we resort to Lemma B.6 by setting

$$W_{h+1}^k(s,a) := \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\right)^2; \qquad C_d := 1; \tag{172}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_1 \le 4H^2 =: C_w.$$

Applying Lemma B.6 with (172) yields that: with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^K (A_{h,k} - Y_{h,k}) &= \sum_{k=1}^K X_{h,k}\\ &\le \sqrt{8C_dC^\star\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)P_{h,s,a}W_{h+1}^k(s,a)^2\log\frac{2H}{\delta}} + 2C_dC^\star C_w\log\frac{2H}{\delta}\\ &\lesssim \sqrt{C^\star\log\frac{2H}{\delta}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right]^2} + C^\star H^2\log\frac{2H}{\delta}. \end{aligned}\tag{173}$$

To further control the first term in (173), it follows from Jensen's inequality that

$$P_{h,s,a}\left[\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right]^2 \le P_{h,s,a}\left[\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^4\right]; \tag{174}$$

which yields

$$\begin{aligned}&\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right]^2\\ &\le \sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^4\\ &\le \left(1+\frac{1}{H}\right)\sum_{k=1}^K\sum_{s\in\mathcal{S}} d_{h+1}^\star(s)\left(V_{h+1}^k(s)-\overline{V}_{h+1}^k(s)\right)^4 + 32\sqrt{H^8C^\star K\log\frac{2H}{\delta}} + 32H^4C^\star\log\frac{2H}{\delta}. \end{aligned}\tag{175}$$

This can be verified similar to the proof for Lemma A.2 in Appendix C.2. We omit the details for conciseness. To continue, it follows that

$$\sum_{k=1}^K\sum_{s\in\mathcal{S}} d_{h+1}^\star(s)\left(V_{h+1}^k(s)-\overline{V}_{h+1}^k(s)\right)^4$$

$$
\begin{aligned}
&\overset{(i)}{\le} \sum_{m=1}^{M}\sum_{t=1}^{L_m}\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\Big(V^\star_{h+1}(s) - \overline{V}^{(m,t)}_{h+1}(s)\Big)^4 \\
&\overset{(ii)}{=} \sum_{m=1}^{M}\sum_{t=1}^{L_m}\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\Big(V^\star_{h+1}(s) - V^{((m-1)\vee 1,1)}_{h+1}(s)\Big)^4 \\
&\overset{(iii)}{=} \sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{m=1}^{M} 2^m\Big(V^\star_{h+1}(s) - V^{((m-1)\vee 1,1)}_{h+1}(s)\Big)^4 \\
&= 4\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{m-2=-1}^{M-2} 2^{m-2}\Big(V^\star_{h+1}(s) - V^{((m-1)\vee 1,1)}_{h+1}(s)\Big)^4 \\
&= 4\sum_{-1}^{0} 2^{m-2}\Big(V^\star_{h+1}(s) - V^{(1,1)}_{h+1}(s)\Big)^4 + 4\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{m-2=1}^{M-2} 2^{m-2}\Big(V^\star_{h+1}(s) - V^{(m-1,1)}_{h+1}(s)\Big)^4 . 
\end{aligned}
$$

Here, (i) holds by using the pessimistic property $V^\star \ge V^k \ge \overline{V}^k$ for all $k \in [K]$ (see (47)) and by regrouping the summands; (ii) follows from the fact (see updating rules in line 34 and line 35) that for any $(m, s, h) \in [M] \times \mathcal{S} \times [H+1]$,

$$
\overline{V}^{(m,t)}_h(s) = V_h^{((m-1)\vee 1,1)}(s), \qquad t = 1, 2, \cdots, L_m; \tag{176}
$$

and (iii) results from the choice of the parameter $L_m = 2^m$. In addition, we can further control

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\Big(V^\star_{h+1}(s) - \overline{V}^k_{h+1}(s)\Big)^4 &\overset{(iv)}{\le} 8H^4 + 4\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{m=1}^{M-2}\sum_{t=1}^{L_m}\Big(V^\star_{h+1}(s) - V^{(m+1,1)}_{h+1}(s)\Big)^4 \\
&\overset{(v)}{\le} 8H^4 + 4\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{m=1}^{M-2}\sum_{t=1}^{L_m}\Big(V^\star_{h+1}(s) - V^{(m,t)}_{h+1}(s)\Big)^4 \\
&\le 8H^4 + 4\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{k=1}^{K}\Big(V^\star_{h+1}(s) - V^k_{h+1}(s)\Big)^4 \qquad (177)\\
&\le 8H^4 + 4H^3\sum_{s\in\mathcal{S}} d^\star_{h+1}(s)\sum_{k=1}^{K}\Big(V^\star_{h+1}(s) - V^k_{h+1}(s)\Big) \\
&\overset{(vi)}{\lesssim} H^3K + H^8SC^\star . \qquad (178)
\end{aligned}
$$

Here, (iv) follows from the fact $0 \le V^\star_{h+1}(s) - V^{(1,1)}_{h+1}(s) \le H - 0 = H$; (v) holds since $V_{h+1}^{(m+1,1)} = V^{(m,L_m)}_{h+1} \ge V^{(m,t)}_{h+1}$ for all $t \in [L_m]$ (using the monotonic increasing property of $V_{h+1}$ introduced in (46)); and (vi) follows from (154). Putting (178) and (175) together with (173), we arrive at

$$
\begin{aligned}
\sum_{k=1}^{K}(A_{h,k} - Y_{h,k}) &\lesssim \sqrt{C^\star \log\frac{2H}{\delta}\Big(H^3K + H^8SC^\star + \sqrt{H^8C^\star K\log\frac{2H}{\delta}} + H^4C^\star\log\frac{2H}{\delta}\Big)} + C^\star H^2\log\frac{2H}{\delta} \\
&\lesssim \sqrt{H^3C^\star K} + H^4\sqrt{SC^{\star 2}} . 
\end{aligned} \tag{179}
$$

Step 3: controlling $\sum_{k=1}^{K}(Z_{h,k} - B_{h,k})$. Similarly, we also invoke Lemma B.6 to control $\sum_{k=1}^{K}(Z_{h,k} - B_{h,k})$. Let's set

$$
W^k_{h+1}(s,a) := \Big(V^k_{h+1} - \overline{V}^k_{h+1}\Big)^2; \qquad C_d := \Big(1 + \frac{1}{H}\Big) \le 2; \tag{180}
$$

which satisfies

$$
\big\|W^k_{h+1}(s,a)\big\|_1 \le 4H^2 =: C_w .
$$

Applying Lemma B.6 with (180) yields that: with probability at least $1-\delta$,

$$
\begin{aligned}
\sum_{k=1}^{K}\big(B_{h,k}-Z_{h,k}\big)=\sum_{k=1}^{K}X_{h,k}
&\le\sqrt{8\sum_{k=1}^{K}C^{2}C_{d}^{\star}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)P_{h,s,a}\big(W_{h+1}^{k}(s,a)\big)^{2}\log\frac{2H}{\delta}}+2C_{d}C^{\star}C_{w}\log\frac{2H}{\delta}\\
&\lesssim\sqrt{C^{\star}\log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)P_{h,s,a}\Big[V_{h+1}^{k}-\overline{V}_{h+1}^{k}\Big]^{4}}+C^{\star}H^{2}\log\frac{2H}{\delta}\\
&\overset{(i)}{\lesssim}\sqrt{C^{\star}\log\frac{2H}{\delta}\big(H^{3}K+H^{8}SC^{\star}\big)}+C^{\star}H^{2}\log\frac{2H}{\delta}\lesssim\sqrt{H^{3}C^{\star}K}+H^{4}\sqrt{SC^{\star2}};
\end{aligned}
\tag{181}
$$

where (i) follows from (177) and (178).

**Step 4: combining the results.** Inserting (181) and (179) back into (171), we can conclude that

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\Big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\Big)^{2}=\sum_{h=1}^{H}A_h\\
&\quad\le\sum_{h=1}^{H}\sum_{k=1}^{K}B_{h,k}+\sum_{h=1}^{H}\sum_{k=1}^{K}\big(Z_{h,k}-B_{h,k}\big)+\sum_{h=1}^{H}\sum_{k=1}^{K}\big(A_{h,k}-Y_{h,k}\big)\\
&\quad\le\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(1+\frac{1}{H}\Big)\sum_{s\in\mathcal{S}}d_{h+1}^{\star}(s)\Big(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\Big)^{2}+\sum_{h=1}^{H}\sum_{k=1}^{K}\big(Z_{h,k}-B_{h,k}\big)+\sum_{h=1}^{H}\sum_{k=1}^{K}\big(A_{h,k}-Y_{h,k}\big)\\
&\quad\lesssim H\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(1+\frac{1}{H}\Big)\sum_{s\in\mathcal{S}}d_{h+1}^{\star}(s)\Big(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\Big)+\sqrt{H^{5}C^{\star}K}+H^{5}\sqrt{SC^{\star2}}\\
&\quad\overset{(i)}{\lesssim}H\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}d_{h+1}^{\star}(s)\Big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\Big)+K+H^{5}\sqrt{SC^{\star2}}\\
&\quad\lesssim H^{2}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}d_h^{\star}(s)\Big(V_h^{\star}(s)-V_h^{k}(s)\Big)+K+H^{5}\sqrt{SC^{\star2}};
\end{aligned}
\tag{182}
$$

where (i) follows from the same routine to obtain (177) and the Cauchy-Schwarz inequality.

### D.3.5. Proof of inequality (168)

**Step 1: decomposing the error in (168).** The term in (168) obeys

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee1}}L_{h,k}\\
&=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee1}}\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}2H\Big(V_{h+1}^{\star}\big(s_{h+1}^{k^n}\big)-\overline{V}_{h+1}^{\mathrm{next},k^n}\big(s_{h+1}^{k^n}\big)\Big)}{N_h^k(s,a)\vee1}}\\
&\overset{(i)}{\lesssim}\sqrt{H}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{d_h^{\star}(s,a)}{kd_h(s,a)}}\sqrt{\frac{d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^{\star}\big(s_{h+1}^{k^n}\big)-\overline{V}_{h+1}^{\mathrm{next},k^n}\big(s_{h+1}^{k^n}\big)\Big)}{kd_h(s,a)}}\\
&\overset{(ii)}{\lesssim}\sqrt{\frac{\phantom{x}}{HC^{\star}}}^{-2}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{\mathbb{1}\big(a=\pi^{\star}(s)\big)}{k}}\sqrt{\frac{d_h^{\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^{\star}\big(s_{h+1}^{k^n}\big)-\overline{V}_{h+1}^{\mathrm{next},k^n}\big(s_{h+1}^{k^n}\big)\Big)}{kd_h(s,a)}}
\end{aligned}
$$

$$\begin{aligned}
&\overset{(iii)}{\lesssim} \sqrt{HC^\star\iota^2}\sqrt{\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^K \frac{d_h^\star(s,a)\sum_{n=1}^{N_h^k(s,a)}\big(V_{h+1}^\star(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next};k^n}(s_{h+1}^{k^n})\big)}{k\,d_h(s,a)}}\\
&\lesssim \sqrt{H^2SC^\star\iota^2}\sqrt{\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^\star(s,a)}{d_h(s,a)}\sum_{k=1}^K\frac{1}{k}\sum_{n=1}^{N_h^k(s,a)}\Big(V_{h+1}^\star(s_{h+1}^{k^n(s,a)})-\overline{V}_{h+1}^{\mathsf{next};k^n}(s_{h+1}^{k^n(s,a)})\Big)}\\
&\overset{(iv)}{=} \sqrt{H^2SC^\star\iota^3}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}\sum_{k'=k}^K\frac{1}{k'}P_h^k\big(V_{h+1}^\star-\overline{V}_{h+1}^{\mathsf{next};k}\big)}\\
&\lesssim \sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}P_h^k\big(V_{h+1}^\star-\overline{V}_{h+1}^{\mathsf{next};k}\big)}.
\end{aligned}\tag{183}$$

Here, (i) follows from the fact $\frac{1}{N_h^k(s,a)\vee 1}\lesssim\frac{8\iota}{k d_h(s,a)}$ (cf. Lemma B.2); (ii) follows from the definition of $C^\star$ in Assumption 2.1; (iii) invokes the Cauchy-Schwarz inequality; (iv) can be obtained by regrouping the terms (the terms involving $(V_{h+1}^\star-\overline{V}_{h+1}^{\mathsf{next};k})$ associated with index $k$ will only been added during episodes $k'=k,k+1,\cdots,K$).

With this upper bound in hand, we further decompose

$$\begin{aligned}
&\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^K\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}L_{h,k} \lesssim \sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}P_h^k\big(V_{h+1}^\star-\overline{V}_{h+1}^{\mathsf{next};k}\big)}\\
&\overset{(i)}{\le}\sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}P_h^k\big(V_{h+1}^\star-\overline{V}_{h+1}^k\big)}\\
&\overset{(ii)}{\le}\sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\big(V_{h+1}^\star-\overline{V}_{h+1}^k\big)}\\
&\quad+\sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^H\sum_{k=1}^K\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}-\frac{d_h^\star(s_h^k,a_h^k)}{d_h(s_h^k,a_h^k)}P_h^k\right)\big(V_{h+1}^\star-\overline{V}_{h+1}^k\big)}.
\end{aligned}\tag{184}$$

Here (i) holds due to the following observation: denoting by $m$ the index of the epoch in which episode $k$ occurs, we have

$$\overline{V}_{h+1}^{\mathsf{next};k}=V_{h+1}^{(m,1)}\ge V_{h+1}^{((m-1)\vee 1),1)}=\overline{V}_{h+1}^k;\tag{185}$$

which invokes the monotonicity of $V_{h+1}^k$ in (46). In addition, (ii) arises from the Cauchy-Schwarz inequality.

**Step 2: controlling the first term in (184).** The first term in (184) satisfies

$$\begin{aligned}
\sum_{h=1}^H\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\big(V_{h+1}^\star-\overline{V}_{h+1}^k\big) &= \sum_{h=1}^H\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\big\langle P_h(\cdot\,|\,s,a),V_{h+1}^\star-\overline{V}_{h+1}^k\big\rangle\\
&\overset{(i)}{=}\sum_{h=1}^H\sum_{k=1}^K\sum_{s'\in\mathcal{S}} d_{h+1}^\star(s')\big(V_{h+1}^\star(s')-\overline{V}_{h+1}^k(s')\big)\\
&\overset{(ii)}{\lesssim} H^2+\sum_{h=1}^H\sum_{k=1}^K\sum_{s\in\mathcal{S}} d_{h+1}(s)\big(V_{h+1}(s)-V_{h+1}^\star(s)\big)\\
&\overset{(iii)}{\lesssim} HK+H^6SC^\star;
\end{aligned}\tag{186}$$

where (i) holds due to the fact $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_h(\cdot\,|\,s,a)=d_{h+1}^\star(\cdot)$, (ii) comes from the same argument employed to establish (177), and (iii) follows from (154).

**Step 3: controlling the second term in (184).** We shall invoke Lemma B.6 for this purpose. To proceed, let

$$W^k_{h+1}(s,a) := V^\star_{h+1} - \overline{V}^k_{h+1}, \qquad C_d =: 1, \tag{187}$$

which satisfies

$$\big\|W^k_{h+1}(s,a)\big\|_\infty \le H =: C_w.$$

Applying Lemma B.6 with (187) yields, for all $h \in [H]$, with probability at least $1-\delta$

$$\begin{aligned}
&\sum_{k=1}^K \left( \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a} - \frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k \right)\left(V^\star_{h+1} - \overline{V}^k_{h+1}\right) = \sum_{k=1}^K X_{h,k} \\
&\le \sqrt{8C_dC^\star \sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\big(W^k_{h+1}(s,a)\big)^2 \log\frac{2H}{\delta}} + 2C_dC^\star C_w \log\frac{2H}{\delta} \\
&\lesssim \sqrt{C^\star \log\frac{2H}{\delta}\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}} d_h^\star(s,a)P_{h,s,a}\big(V^\star_{h+1} - \overline{V}^k_{h+1}\big)^2} + HC^\star\log\frac{2H}{\delta} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{C^\star \log\frac{2H}{\delta}\left(H^2 + \sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a}\big(V^\star_{h+1} - V^k_{h+1}\big)^2\right)} + HC^\star\log\frac{2H}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{C^\star\log\frac{2H}{\delta}(HK + H^6SC^\star)} + HC^\star\log\frac{2H}{\delta} \\
&\lesssim \sqrt{HC^\star K} + H^3\sqrt{S}C^\star.
\end{aligned} \tag{188}$$

Here (i) follows from the same routine to arrive at (177), and (ii) comes from (154). As a result, the second term in (184) satisfies, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{h=1}^H\sum_{k=1}^K \left( \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a} - \frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k \right)\left(V^\star_{h+1} - \overline{V}^k_{h+1}\right) \\
&= \sum_{h=1}^H\sum_{k=1}^K \left( \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)P_{h,s,a} - \frac{d_h^\star(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k \right)\left(V^\star_{h+1} - \overline{V}^k_{h+1}\right) \lesssim \sqrt{H^3C^\star K} + H^4\sqrt{S}C^\star.
\end{aligned} \tag{189}$$

**Step 4: combining the results.** Finally, inserting (186) and (189) into (184), we arrive at

$$\begin{aligned}
&\sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^K \sqrt{\frac{1}{N_h^k(s,a)}}L_{h,k} \\
&\lesssim \sqrt{H^2SC^{\star 4}}\sqrt{HK + H^6SC^\star} + \sqrt{H^2SC^{\star 4}}\sqrt{\sqrt{H^3C^\star K + H^4}\sqrt{S}C^\star} \\
&\lesssim \sqrt{H^3SC^\star K^4 + H^4SC^{\star 3}} + \sqrt{\sqrt{H^2SC^{\star 4}}}\sqrt{HK + H^4\sqrt{S}C^\star} \lesssim \sqrt{H^3SC^\star K^4 + H^4SC^{\star 3}};
\end{aligned} \tag{190}$$

where the last two inequalities follow from the Cauchy-Schwarz inequality.

### D.3.6. Proof of inequality (169)

Recall the expression of $G_{h,k}$ in (166) as

$$G_{h,k}^2 = \frac{\sum_{n=1}^{N_h^k(s,a)} \big(V^\star_{h+1}(s_{h+1}^{k^n})\big)^2}{N_h^k(s,a)\vee 1} - \left(\frac{\sum_{n=1}^{N_h^k(s,a)} V^\star_{h+1}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2$$

$$= \frac{\sum_{n=1}^{N_h^k(s,a)} \big(P_h^{k^n} V_{h+1}^\star\big)^2}{N_h^k(s,a) \vee 1} - \left(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a) \vee 1}\right)^2. \tag{191}$$

To continue, we make the following observation

$$\begin{aligned} G_{h,k} &\le \Big(\big|G_{h,k}^2 - \mathsf{Var}_{h,s,a}(V_{h+1}^\star)\big| + \mathsf{Var}_{h,s,a}(V_{h+1}^\star)\Big)^{1/2} \\ &\le \big|G_{h,k}^2 - \mathsf{Var}_{h,s,a}(V_{h+1}^\star)\big|^{1/2} + \sqrt{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)} \end{aligned} \tag{192}$$

due to the elementary inequality $\sqrt{a^2+b^2} \le a + b$ for any $a, b \ge 0$. Here, we remind the reader that $\mathsf{Var}_{h,s,a}(V_{h+1}^\star) = P_{h,s,a}(V_{h+1}^\star)^2 - (P_{h,s,a}V_{h+1}^\star)^2$ (cf. (109)). This allows us to rewrite

$$\begin{aligned} &\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) \sum_{k=1}^{K} \sqrt{\frac{1}{N_h^k(s,a)\vee 1} G_{h,k}} \\ &\le \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) \sum_{k=1}^{K} \sqrt{\frac{\big|G_{h,k}^2 - \mathsf{Var}_{h,s,a}(V_{h+1}^\star)\big|}{N_h^k(s,a)\vee 1}} + \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a) \sum_{k=1}^{K} \sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{N_h^k(s,a)\vee 1}}, \end{aligned} \tag{193}$$

leaving us with two terms to cope with.

**Step 1: controlling the first term of (193).** By definition, we have

$$\begin{aligned} G_{h,k}^2 - \mathsf{Var}_{h,s,a}(V_{h+1}^\star) &= \frac{\sum_{n=1}^{N_h^k(s,a)} \big(P_h^{k^n} V_{h+1}^\star\big)^2}{N_h^k(s,a)\vee 1} - \left(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\right)^2 - P_{h,s,a}(V_{h+1}^\star)^2 + \big(P_{h,s,a}V_{h+1}^\star\big)^2 \\ &\le \left|\frac{\sum_{n=1}^{N_h^k(s,a)} \big(P_h^{k^n} V_{h+1}^\star\big)^2}{N_h^k(s,a)\vee 1} - P_{h,s,a}(V_{h+1}^\star)^2\right| + \left|\left(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\right)^2 - \big(P_{h,s,a}V_{h+1}^\star\big)^2\right| \\ &\le \left|\frac{\sum_{n=1}^{N_h^k(s,a)} \big(P_h^{k^n} V_{h+1}^\star\big)^2}{N_h^k(s,a)\vee 1} - P_{h,s,a}(V_{h+1}^\star)^2\right| + 2H\left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right|, \end{aligned} \tag{194}$$

where the last inequality holds due to

$$\begin{aligned} \left|\left(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\right)^2 - \big(P_{h,s,a}V_{h+1}^\star\big)^2\right| &= \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right| \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} + P_{h,s,a}V_{h+1}^\star\right| \\ &\le 2H\left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right|. \end{aligned}$$

We now control the two terms in (194) separately by invoking Lemma B.4. For the first term in (194), let us set

$$W_{h+1} := \big(V_{h+1}^\star\big)^2 \quad \text{and} \quad u_h(s,a;N) := \frac{1}{N \vee 1} =: C_u, \tag{195}$$

which indicates that

$$\|W_{h+1}^i\|_\infty \le H^2 =: C_w. \tag{196}$$

Applying Lemma B.4 with (195) and $N = N_h^k = N_h^k(s,a)$, with probability at least $1 - \frac{\delta}{2}$, we arrive at

$$\frac{1}{N_h^k(s,a)\vee 1} \sum_{n=1}^{N_h^k} \big(P_h^{k^n} - P_{h,s,a}\big)\big(V_{h+1}^\star\big)^2 = \sum_{i=1}^{k} X_i\big(s,a,h,N_h^k\big)$$

$$
\begin{aligned}
&\lesssim \sqrt{C_u \log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k} u_h^n(s,a;N_h^k)\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k^n}\right)} + \left(C_u C_w + \sqrt{\frac{C_u}{N_h^k-1}}C_w\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^2}{N_h^k-1}}\sqrt{\sum_{n=1}^{N_h^k}\frac{1}{N_h^k-1}\left\|W_{h+1}^{k^n}\right\|_\infty^2} + \frac{H^2\iota^2}{N_h^k-1} \lesssim H\iota^2\sqrt{\frac{1}{N_h^k-1}}. \qquad (197)
\end{aligned}
$$

Similarly, for the second term in (194), with $W_{h+1}^i = V_{h+1}^\star$, we have with probability at least $1-\frac{\delta}{2}$,

$$
\left|\frac{1}{N_h^k(s,a)-1}\sum_{n=1}^{N_h^k}\left(P_h^{k^n}-P_{h,s,a}\right)V_{h+1}^\star\right| \lesssim H\iota^2\sqrt{\frac{1}{N_h^k(s,a)-1}}. \qquad (198)
$$

Inserting (197) and (198) back into (194) yields

$$
\left|G_{h,k}^2 - \mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)\right| \lesssim H\iota^2\sqrt{\frac{1}{N_h^k(s,a)-1}}. \qquad (199)
$$

Consequently, the first term in (193) can be controlled as

$$
\begin{aligned}
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\left|G_{h,k}^2-\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)\right|}{N_h^k(s,a)-1}} &\lesssim H\iota\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\frac{1}{\left(N_h^k(s,a)\right)^{\frac{3}{4}}-1} \\
&\lesssim H^2(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2, \qquad (200)
\end{aligned}
$$

where the last inequality holds due to (158).

**Step 2: controlling the second term of (193).** The second term can be decomposed as

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)}{N_h^k(s,a)-1}} \\
&\overset{\text{(i)}}{\lesssim} \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}}\sum_{=1}^{A}\sum_{k=1}^{K}\sqrt{\frac{C^\star d_h^\star(s,a)\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)}{k}\iota}\,\mathbb{1}\left(a=\pi_h^\star(s)\right) \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{C^\star\iota}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}} d_h^\star(s,a)\sum_{k=1}^{K}\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}}\sum_{k=1}^{K}\frac{1}{k}\mathbb{1}\left(a=\pi_h^\star(s)\right)} \\
&\lesssim \sqrt{HSC^\star K\iota^2}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right)}; \qquad (201)
\end{aligned}
$$

where (i) follows from the facts $\frac{N_h^k(s,a)-1}{k} \gtrsim d_h^{\mathsf{b}}(s,a)\ \forall$ by Lemma B.2 and the definition of $C^\star$ in Assumption 2.1, (ii) holds by the Cauchy-Schwarz inequality, and the final inequality comes from the fact that $\pi^\star$ is deterministic.

We are then left with bounding $\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)\mathrm{Var}_{h,s,a}(V_{h+1}^\star)$. Note that

$$
\begin{aligned}
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\mathcal{A}} d_h^\star(s,a)\mathrm{Var}_{h,s,a}\left(V_{h+1}^\star\right) &= \mathbb{E}_{s_1\sim\rho,\, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\sum_{h=1}^{H}\mathrm{Var}_{h,s_h,\pi_h^\star(s_h)}\left(V_{h+1}^\star\right)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{s_1\sim\rho,\, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\sum_{h=1}^{H}\left(r_h\left(s_h,\pi_h^\star(s_h)\right)+V_{h+1}^\star(s_{h+1})-V_h^\star(s_h)\right)^2\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{s_1\sim\rho,\, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\left(\sum_{h=1}^{H} r_h\left(s_h,\pi_h^\star(s_h)\right)+V_{h+1}^\star(s_{h+1})-V_h^\star(s_h)\right)^2\right]
\end{aligned}
$$

$$\overset{(iii)}{=} \mathbb{E}_{s_1, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\left(\sum_{h=1}^{H} r_h(s_h,\pi_h^\star(s_h))\right) - V_1^\star(s_1)\right]^2 \overset{(iv)}{\le} H^2; \tag{202}$$

where (i) follows from Bellman's optimality equation, (ii) follows from the Markov property, (iii) holds due to the fact that $V_{H+1}^\star(s) = 0$ for all $s \in \mathcal{S}$, and (iv) arises from the fact $r_h(s,a) \le 1$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Substituting (202) back into (201), we get

$$\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{N_h^k(s,a)\vee 1}} \lesssim \sqrt{H^3SC^\star K^2}. \tag{203}$$

**Step 4: combing the results.** Combining (200) and (203) with (193) yields

$$\begin{aligned}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\star(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}G_{h,k} &\lesssim H^2(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2 + \sqrt{H^3SC^\star K^2}\\ &\lesssim \sqrt{H^3SC^\star K^2} + H^{2.5}SC^\star\iota^3.\end{aligned} \tag{204}$$

### D.4. Proof of Lemma D.1

In view of (127), we can decompose the term of interest into

$$\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\xi_h^{k^n} \le |U_1| + |U_2|,$$

where

$$U_1 := \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(P_h^{k^n} - P_{h,s,a}\right)\left(\overline{V}_{h+1}^{k^n} - V_{h+1}^{k^n}\right), \tag{205a}$$

$$U_2 := \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{N_h^{\mathsf{epo},m^n-1}(s,a)\vee 1}\right)\overline{V}_{h+1}^{k^n}. \tag{205b}$$

Next, we turn to controlling these two terms separately with the assistance of Lemma B.4.

**Step 1: controlling $U_1$.** In the following, we invoke Lemma B.4 to control $U_1$ in (205a). Let us set

$$W_{h+1}^i := \overline{V}_{h+1}^i - V_{h+1}^i, \qquad \text{and} \qquad u_h^i(s,a,N) := \eta^N_{N_h^i(s,a)} \ge 0;$$

which indicates that

$$\|W_{h+1}^i\|_\infty \le \|\overline{V}_{h+1}^i\|_\infty + \|V_{h+1}^i\|_\infty \le 2H =: C_w;$$

and

$$\max_{N,h,s,a\in\{0\}\cup[K]\times[H]\times\mathcal{S}\times\mathcal{A}} \eta^N_{N_h^i(s,a)} \overset{(206)}{\le} \frac{2H}{N\vee 1} =: C_u.$$

Here, the last inequality follows since (according to Lemma B.1 and the definition in (25))

$$\begin{cases}\eta^N_{N_h^i(s,a)} \le \dfrac{2H}{N\vee 1}, & \text{if } 0 \le N_h^i(s,a) \le N;\\ \eta^N_{N_h^i(s,a)} = 0, & \text{if } N_h^i(s,a) > N.\end{cases}$$

To continue, it can be seen from (26) that

$$0 \le \sum_{n=1}^{N} u_h^{k^n(s,a)}(s,a;N) = \sum_{n=1}^{N} \eta_n^N \le 1 \tag{207}$$

holds for all $(N,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$. Therefore, choosing $N = N_h^k(s,a) = N_h^k$ for any $(s,a)$ and applying Lemma B.4 with the above quantities, we arrive at

$$\begin{aligned}
|U_1| &= \left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(P_h^{k^n} - P_{h,s,a}\right)\left(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\right)\right| = \left|\sum_{i=1}^{N_h^k} X_i\left(s,a,h,N_h^k\right)\right| \\
&\lesssim \sqrt{C_u \log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s,a;N_h^k)\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k^n}\right)} + \left(C_u C_w + \sqrt{\frac{C_u}{N_h^k-1}}\,C_w\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{H\iota^2}{N_h^k-1}}\sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\mathrm{Var}_{h,s,a}\left(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\right)} + \frac{H^2\iota^2}{N_h^k-1} \qquad (208) \\
&\lesssim \sqrt{\frac{H\iota^2}{N_h^k-1}}\sqrt{\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s,a) - \left(\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s,a)\right)^2} + \frac{N H^{7/4}\iota^2}{(N_h^k-1)^{3/4}} + \frac{H^2\iota^2}{N_h^k-1}, \qquad (209)
\end{aligned}$$

with probability at least $1-\delta$. Here, the proof of the inequality (209) is postponed to Appendix D.4.1 in order to streamline the presentation of the analysis.

**Step 2: bounding $U_2$.** Making use of the result in (111), we arrive at

$$\frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\overline{V}_{h+1}^{k^n} = \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}\overline{V}_{h+1}^{\mathrm{next},k^i}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}.$$

To continue, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we rewrite and rearrange $U_2$ (cf. (205b)) as follows:

$$\begin{aligned}
U_2 &= \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\right)\overline{V}_{h+1}^{k^n} \\
&= \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(P_{h,s,a}\overline{V}_{h+1}^{k^n} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}\overline{V}_{h+1}^{\mathrm{next},k^i}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\right) \\
&\overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(\frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_{h,s,a}\overline{V}_{h+1}^{k^n}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}\overline{V}_{h+1}^{\mathrm{next},k^i}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\right) \\
&= \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}}\left(P_{h,s,a} - P_h^{k^i}\right)\overline{V}_{h+1}^{\mathrm{next},k^i} \\
&\overset{\text{(ii)}}{=} \sum_{i=1}^{N_h^k}\left(\sum_{n=N_h^{(m^i+1,1)}+1}^{N_h^{(m^i+2,1)} \wedge N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathrm{epo},m^n-1}(s,a) \vee 1}\right)\left(P_{h,s,a} - P_h^{k^i}\right)\overline{V}_{h+1}^{\mathrm{next},k^i} \\
&= \sum_{i=1}^{N_h^k}\left(\sum_{n=N_h^{(m_i+1,1)}+1}^{N_h^{(m^i+2,1)} \wedge N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathrm{epo},m^i} \vee 1}\right)\left(P_{h,s,a} - P_h^{k^i}\right)\overline{V}_{h+1}^{\mathrm{next},k^i};
\end{aligned}$$

where (i) follows from the fact that $N_h^{(m^n,1)} - N_h^{(m^n-1,1)} = N_h^{\mathsf{epo},m^n-1}(s,a)$, and (ii) is obtained by rearranging terms with respect to $i$ (the terms with respect to $\overline{V}_{h+1}^{\mathsf{next},k^i}$ will only be added during the epoch $m^i+1$), and the last equality holds since $m^n - 1 = m^i$ for all $n = N_h^{(m^i+1,1)}+1, N_h^{(m^i+1,1)}+2, \cdots, N_h^{(m^i+2,1)}$.

With the above relation in mind, we are ready to invoke Lemma B.4 to control $U_2$. To continue, for any episode $j \leq k$, let us denote by $m(j)$ the index of the epoch in which episode $j$ happens (with slight abuse of notation). Let us set

$$W_{h+1}^j := \overline{V}_{h+1}^{\mathsf{next},j}; \qquad \text{and} \qquad u_h^j(s,a,N) := \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\eta_n^N}{N_h^{\mathsf{epo},m(j)}(s,a) \vee 1}.$$

As a result, we see that

$$\|W_{h+1}^j\|_\infty \leq \|\overline{V}_{h+1}^{\mathsf{next},j}\|_\infty \leq H =: C_{\mathrm{w}}.$$

and the following fact (which will be established in Appendix D.4.2)

$$0 \leq u_h^j(s,a,N) = \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\eta_n^N}{N_h^{\mathsf{epo},m(j)}(s,a) \vee 1} \leq \frac{64e^2}{N \vee 1} =: C_{\mathrm{u}} \tag{210}$$

holds for all $(j,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$ with probability at least $1-\delta$.

Given that $N = N_h^k(s,a) = N_h^k$, applying Lemma B.4 with the above quantities, we can show that for any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
|U_2| &= \left| \sum_{i=1}^{N_h^k} \left( \sum_{n=N_h^{(m^i+1,1)}+1}^{N_h^{(m^i+2,1)}\wedge N_h^k} \frac{\eta_n^{N_h^k}}{N_h^{\mathsf{epo},m^i} \vee 1} \right) \left(P_{h,s,a} - P_h^{k^i}\right) \overline{V}_{h+1}^{\mathsf{next},k^i} \right| = \left| \sum_{j=1}^{k} X_j\left(s,a,h,N_h^k\right) \right| \\
&\lesssim \sqrt{C_{\mathrm{u}} \log^2\frac{SAT}{\delta} \sum_{i=1}^{N_h^k(s,a)} u_h^{k^i(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\left(W_{h+1}^{k^i(s,a)}\right)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N\vee 1}}\,C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^3}{N_h^k \vee 1}}\sqrt{\frac{1}{N_h^k \vee 1}\sum_{i=1}^{N_h^k}\mathsf{Var}_{h,s,a}\left(\overline{V}_{h+1}^{\mathsf{next},k^i}\right)} + \frac{H\iota^3}{N_h^k \vee 1} \\
&\lesssim \sqrt{\frac{\iota^3}{N_h^k \vee 1}}\sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k}+1}(s,a) - \left(\mu_h^{\mathsf{ref},k^{N_h^k}+1}(s,a)\right)^2} + \frac{H\iota^3}{(N_h^k \vee 1)^{3/4}}.
\end{aligned} \tag{211}$$

To streamline the presentation of the analysis, we shall postpone the proof of (211) to Appendix D.4.3.

**Step 3: summing up.** Combining the bounds in (209) and (211) yields that: for any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left|\xi_h^{k^n}\right| |U_1| + |U_2| \\
&\qquad \lesssim \sqrt{\frac{H\iota^2}{N_h^k \vee 1}}\sqrt{\sigma_h^{\mathsf{adv},k^{N_h^k}+1}(s,a) - \left(\mu_h^{\mathsf{adv},k^{N_h^k}+1}(s,a)\right)^2} \\
&\qquad\quad + \sqrt{\frac{\iota^3}{N_h^k \vee 1}}\sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k}+1}(s,a) - \left(\mu_h^{\mathsf{ref},k^{N_h^k}+1}(s,a)\right)^2} + c_{\mathrm{b}}\frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + c_{\mathrm{b}}\frac{H^2\iota^2}{N_h^k \vee 1} \\
&\qquad \leq \overline{B}_h^{k^{N_h^k}+1}(s,a) + c_{\mathrm{b}}\frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + c_{\mathrm{b}}\frac{H^2\iota^2}{N_h^k \vee 1}
\end{aligned} \tag{212}$$

holds for some sufficiently large constant $c_b > 0$, where the last line follows from the definition of $B_h^{k^{N_h^k}+1}(s,a)$ in line 14 of Algorithm 3. As a consequence of the inequality (212), for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, one has

$$\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{B}_h^{k^n} \le \overline{B}_h^{k^{N_h^k}+1}(s,a)+c_b\frac{H^{7/4}\iota^2}{(N_h^k-1)^{3/4}}+c_b\frac{H\iota^2}{N_h^k-1} \lesssim \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} b_h^{k^n+1};$$

where the last inequality holds due to (120). We have thus concluded the proof of Lemma D.1.

### D.4.1. Proof of inequality (209)

To establish the inequality (209), it is sufficient to consider the difference

$$W_1 := \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathrm{Var}_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)-\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s,a)+\big(\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s,a)\big)^2.$$

Before continuing, it is easily verified that if $N_h^k = N_h^k(s,a) = 0$, the basic fact $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} = 0$ leads to $W_1 = 0$, and therefore, (209) holds directly. The remainder of the proof is thus dedicated to controlling $W_1$ when $N_h^k = N_h^k(s,a) \ge 1$. Recalling the definition in (109)

$$\mathrm{Var}_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big) = P_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2-\Big(P_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2; \tag{213}$$

we can take this result together with (112) to yield

$$\begin{aligned}W_1 = {} & \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2-\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2\\ & +\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2-\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2\\ = {} & -\underbrace{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s,a}\big)\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)^2}_{=:W_1^1}\\ & +\underbrace{\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2-\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\big(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\big)\Big)^2}_{=:W_1^2}.\end{aligned} \tag{214}$$

It then boils down to control the above two terms in (214) separately when $N_h^k = N_h^k(s,a) \ge 1$.

**Step 1: controlling $W_1^1$.** To control $W_1^1$, we shall invoke Lemma B.4 by setting

$$W_{h+1}^i := \big(V_{h+1}^i-\overline{V}_{h+1}^i\big)^2; \qquad \text{and} \qquad u_h^i(s,a;N) := \eta_{N_h^i(s,a)}^N \ge 0;$$

which obey

$$\|W_{h+1}^i\|_\infty \le \|\overline{V}_{h+1}^i\|_\infty^2+\|V_{h+1}^i\|_\infty^2 \le 2H^2 =: C_\mathrm{w}.$$

Invoking the facts in (206) and (207), we arrive at

$$\frac{2H}{N-1} =: C_\mathrm{u}$$

and

$$0 \le \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a;N) \le 1; \qquad \forall (N,s,a) \in [K]\times\mathcal{S}\times\mathcal{A}.$$

Therefore, choosing $N = N_h^k(s,a) = N_h^k$ for any $(s,a)$ and applying Lemma B.4 with the above quantities, we arrive at, with probability at least $1-\delta$,

$$\begin{aligned}
|W_1^1| &= \left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}(P_h^{k^n} - P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2\right| = \left|\sum_{i=1}^{k} X_i(s,a,h,N_h^k)\right| \\
&\lesssim \sqrt{C_u \log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s,a;N_h^k)\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k^n}\right)} + \left(C_u C_w + \sqrt{\frac{C_u}{N-1}}C_w\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{H}{N_h^k-1}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\|W_{h+1}^{k^n}\|_\infty^2} + \frac{H^{3}}{N_h^k-1} \lesssim \sqrt{\frac{H^5}{N_h^k-1}} + \frac{H^{3}}{N_h^k-1}.
\end{aligned} \tag{215}$$

**Step 2: controlling $W_1^2$.** Observe that Jensen's inequality gives

$$\left(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)\right)^2 \le \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2; \tag{216}$$

due to the fact $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} = 1$ (see (26) and (25)). Plugging the above relation into (214) gives

$$\begin{aligned}
W_1^2 &\le \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2 - \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2 \\
&= \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}(P_h^{k^n}-P_{h,s,a})\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2 + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}(-P_h^{k^n}+P_{h,s,a})\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right).
\end{aligned} \tag{217}$$

Note that the first term in (217) is exactly $|U_1|$ defined in (205a), which can be controlled by invoking (208) to achieve that, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}(P_h^{k^n}-P_{h,s,a})\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right| \\
&\lesssim \sqrt{\frac{H^2}{N_h^k-1}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathrm{Var}_{h,s,a}\left(\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right)} + \frac{H^2}{N_h^k-1} \lesssim \sqrt{\frac{H^3}{N_h^k-1}} + \frac{H^2}{N_h^k-1};
\end{aligned} \tag{218}$$

where the final inequality holds since $\mathrm{Var}_{h,s,a}\left(\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right) \lesssim H^2$ and the fact in (26). In addition, the second term in (217) can be controlled straightforwardly by

$$\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}(-P_h^{k^n}+P_{h,s,a})\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)\right| \le \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(\|P_h^{k^n}\|_1+\|P_{h,s,a}\|_1\right)\left\|V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right\|_\infty \le 2H;$$

where we have used the fact in (26), $\|V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\|_\infty \le H$ and $\|P_h^{k^n}\|_1 = \|P_{h,s,a}\|_1 = 1$.

Taking the above two facts collectively with (217) yields

$$W_1^2 \lesssim \sqrt{\frac{H^5}{N_h^k-1}} + \frac{H^3}{N_h^k-1}. \tag{219}$$

**Step 3: summing up.** Plugging the results in (215) and (219) back into (214), we have

$$W_1 \le W_{1,1} + W_{1,2} \lesssim \sqrt{\frac{H^5\iota^2}{N_h^k \vee 1}} + \frac{H^3\iota^2}{N_h^k \vee 1};$$

which leads to the desired result (209) directly.

### D.4.2. Proof of inequality (210)

To begin with, let us recall two pieces of notation that shall be used throughout this proof:

1. $m(j)$: the index of the epoch in which the $j$-th episode occurs.
2. $\widehat{N}_h^{\mathsf{epo},m}(s,a)$: the value of $\widehat{N}_h^{(m,L_m+1)}(s,a)$, representing the number of visits to $(s,a)$ in the entire $m$-th epoch with length $L_m = 2^m$.

Applying (56) and taking the union bound over $(m(j),h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ yield

$$\widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \ge \frac{2^{m(j)}d_h(s,a)}{8\log\frac{SAT}{\delta}} \tag{220}$$

with probability at least $1-\delta/2$.

For any epoch $m$, if we denote by $k_{\mathsf{last}}(m)$ the index of the last episode in the $m$-th epoch, we can immediately see that

$$k_{\mathsf{last}}(m) = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} 2^i = 2^{m+1} - 2 \le 2^{m+1}. \tag{221}$$

Applying (56) again and taking the union bound over $(m(j),h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, one can guarantee that for every $n \in [N_h^{(m(j)+1,1)}, N_h^{(m(j)+2,1)}]$, with probability at least $1-\delta/2$,

$$N_h^{(m(j)+1,1)} \le n \le N_h^{(m(j)+2,1)} = N_h^{k_{\mathsf{last}}(m(j)+1)} \le N_h^{2^{m(j)+2}} \le \begin{cases} e^2 2^{m(j)+2} d_h(s,a) & \text{if } 2^{m(j)+2}d_h(s,a) \ge \log\frac{SAT}{\delta} \\ 2e^2\log\frac{SAT}{\delta} & \text{if } 2^{m(j)+2}d_h(s,a) \le 2\log\frac{SAT}{\delta} \end{cases}. \tag{222}$$

Combine the above results to yield

$$\begin{cases} \widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \overset{(i)}{\ge} \frac{2^{m(j)}d_h(s,a)}{8\log\frac{SAT}{\delta}} \overset{(ii)}{\ge} \frac{1}{32e^2\log\frac{SAT}{\delta}}n; & \text{if } 2^{m(j)+2}\, d_h(s,a) \ge \log\frac{SAT}{\delta}; \text{ if} \\ \widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \ge 1 \overset{(iii)}{\ge} \frac{1}{2e^2\log\frac{SAT}{\delta}}n & 2^{m(j)+2}\, d_h(s,a) \le 2\log\frac{SAT}{\delta}; \end{cases} \tag{223}$$

where (i) follows from (220), (ii) and (iii) hold due to (222). As a result, we arrive at

$$\sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\frac{N}{n}}{\widehat{N}_h^{\mathsf{epo},m(j)}(s,a)\vee 1} \le \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{32e^2\log\frac{SAT}{\delta}\,\frac{N}{n}}{n} \le \sum_{n=N_h^{(m(j)+1,1)}+1}^{N} \frac{32e^2\log\frac{SAT}{\delta}\,\frac{N}{n}}{n} \le \frac{64e^2\log\frac{SAT}{\delta}}{N\vee 1};$$

where the last inequality holds since $\sum_{i=1}^{N}\frac{\frac{N}{i}}{i} \le \frac{2}{N\vee 1}$ (see Lemma B.1).

### D.4.3. Proof of inequality (211)

In this subsection, we intend to control the following term

$$W_2 := \frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}\mathrm{Var}_{h,s,a}\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right) - \left(\sigma^{\mathrm{ref},k^{N_h^k}+1}_h(s,a) - \left(\mu^{\mathrm{ref},k^{N_h^k}+1}_h(s,a)\right)^2\right)$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. First, it is easily seen that if $N_h^k = 0$, then we have $W_2 = 0$ and thus (211) is satisfied. Therefore, the remainder of the proof is devoted to verifying (211) when $N_h^k = N_h^k(s,a) \geq 1$. Combining the expression (113) with the following definition

$$\mathrm{Var}_{h,s,a}\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right) = P_{h,s,a}\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 - \left(P_{h,s,a}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2;$$

we arrive at

$$\begin{aligned}
W_2 &\le \frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 - \left(P_{h,s,a}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2\right) \\
&\quad - \frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}P_h^{k^n}\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 + \left(\frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}P_h^{k^n}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 \\
&= \underbrace{\frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}-P_h^{k^n}\right)\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2}_{=:W_2^1} + \underbrace{\left(\frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}P_h^{k^n}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 - \frac{1}{N_h^k-1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2}_{=:W_2^2}.
\end{aligned} \tag{224}$$

In the sequel, we intend to control the terms in (224) separately.

**Step 1: controlling $W_2^1$.** The first term $W_2^1$ can be controlled by invoking Lemma B.4 and set

$$W_{h+1} := \left(\overline{V}^{\mathrm{next},i}_{h+1}\right)^2; \qquad \text{and} \qquad u_h(s,a;N) := \frac{1}{N} =: C_u.$$

To proceeding, with the fact

$$\left\|W_{h+1}\right\|_\infty \le \left\|\overline{V}^{\mathrm{next},i}_{h+1}\right\|_\infty^2 \le H^2 =: C_w,$$

and $N = N_h^k(s,a) = N_h^k$, applying Lemma B.4 with the above quantities, we have for all state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
W_2^1 &= \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}-P_h^{k^n}\right)\left(\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 = \sum_{i=1}^{k}X_i\left(s,a,h,N_h^k\right) \\
&\lesssim \sqrt{C_u\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)}u_h^{k_h^n(s,a)}(s,a;N)\,\mathrm{Var}_{h,s,a}\left(W_{h+1}^{k_h^n(s,a)}\right)} + \left(C_uC_w + \sqrt{\frac{C_u}{N}}C_w\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^2\sqrt{\left\|W_{h+1}\right\|_\infty^2}\,H^2\iota^2}{N_h^k}} + \sqrt{\frac{H^4\iota^2}{N_h^k}} + \frac{H^2\iota^2}{N_h^k}.
\end{aligned} \tag{225}$$

**Step 2: controlling $W_2^2$.** Towards controlling $W_2^2$ in (224), we observe that by Jensen's inequality,

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_{h,s,a}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2 \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\overline{V}^{\mathrm{next},k^n}_{h+1}\right)^2.$$

Equipped with this relation, $W_2^2$ satisfies

$$W_2^2 \le \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\overline{V}^{\mathsf{next};k^n}_{h+1}\right)^2 - \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_{h;s;a}\overline{V}^{\mathsf{next};k^n}_{h+1}\right)^2$$
$$= \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h;s;a}\right)\overline{V}^{\mathsf{next};k^n}_{h+1}\right)\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h;s;a}\right)\overline{V}^{\mathsf{next};k^n}_{h+1}\right). \qquad (226)$$

As for the first term in (226), let us set

$$W_{h+1}^i := \overline{V}^{\mathsf{next};i}_{h+1}; \qquad \text{and} \qquad u_h^i(s;a;N) := \frac{1}{N} =: C_u;$$

which satisfy

$$\|W_{h+1}\|_\infty \le \|\overline{V}^{\mathsf{next};i}_{h+1}\|_\infty \le H =: C_w.$$

For any $(s;a)$, Lemma B.4 together with the above quantities and $N = N_h^k = N_h^k(s;a)$ gives

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h;s;a}\right)\overline{V}^{\mathsf{next};k^n}_{h+1}\right|$$
$$\lesssim \sqrt{C_u\log^2\frac{SAT}{\delta}\sum_{n=1}^{N_h^k(s;a)} u_h^{k^n_h(s;a)}(s;a;N)\,\mathrm{Var}_{h;s;a}\left(W_{h+1}^{k^n_h(s;a)}\right)} + \left(C_uC_w + \sqrt{\frac{C_u}{N}}C_w\right)\log^2\frac{SAT}{\delta}$$
$$\lesssim \sqrt{\frac{\left\|W_{h+1}^{k^n_h(s;a)}\right\|_\infty^2 H^2}{N_h^k}} + \frac{H^2}{N_h^k} \lesssim \sqrt{\frac{H^2 N_h^k}{N_h^k}} + \frac{H^2}{N_h^k}$$

with probability at least $1-\delta$. In addition, the second term can be bounded straightforwardly by

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h;s;a}\right)\overline{V}^{\mathsf{next};k_n}_{h+1}\right| \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(\|P_h^{k^n}\|_1 + \|P_{h;s;a}\|_1\right)\|\overline{V}^{\mathsf{next};k_n}_{h+1}\|_\infty \le 2H;$$

where the last inequality is valid since $\|\overline{V}^{\mathsf{next};k^n}_{h+1}\|_\infty \le H$ and $\|P_h^{k^n}\|_1 = \|P_{h;s;a}\|_1 = 1$. Substitution of the above two observations back into (226) yields

$$W_2^2 \lesssim \sqrt{\frac{H^4}{N_h^k \vee 1}}^{\,2} + \frac{H^2}{N_h^k \vee 1}^{\,2}. \qquad (227)$$

**Step 3: combining the above results.** Plugging the results in (225) and (227) into (224), we reach

$$W_2 \le W_2^1 + W_2^2 \lesssim \sqrt{\frac{H^4}{N_h^k \vee 1}}^{\,2} + \frac{H^2}{N_h^k \vee 1}^{\,2};$$

thus establishing the desired inequality (211).