Information and Inference: A Journal of the IMA (2023) **12**, 969–1043
https://doi.org/10.1093/imaiai/iaac034
Advance Access publication on 17 December 2022

Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning

GEN LI
Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

LAIXI SHI
Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

YUXIN CHEN
Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

AND

YUEJIE CHI
Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
†Corresponding author. Email: yuejiechi@cmu.edu —This paper was presented in part at the 2021 Conference on Neural Information Processing Systems (NeurIPS).

[Received on 10 October 2021; revised on 22 August 2022; accepted on 16 October 2022]

Abstract

Achieving sample efficiency in online episodic reinforcement learning (RL) requires optimally balancing exploration and exploitation. When it comes to a finite-horizon episodic Markov decision process with S states, A actions and horizon length H, substantial progress has been achieved toward characterizing the minimax-optimal regret, which scales on the order of $\sqrt{H^2SAT}$ (modulo log factors) with T the total number of samples. While several competing solution paradigms have been proposed to minimize regret, they are either memory-inefficient, or fall short of optimality unless the sample size exceeds an enormous threshold (e.g. $S^6A^4\,\mathrm{poly}(H)$ for existing model-free methods).

To overcome such a large sample size barrier to efficient RL, we design a novel model-free algorithm, with space complexity $O(SAH)$, that achieves near-optimal regret as soon as the sample size exceeds the order of $SA\,\mathrm{poly}(H)$. In terms of this sample size requirement (also referred to the initial burn-in cost), our method improves—by at least a factor of S^5A^3—upon any prior memory-efficient algorithm that is asymptotically regret-optimal. Leveraging the recently introduced variance reduction strategy (also called *reference-advantage decomposition*), the proposed algorithm employs an *early-settled* reference update rule, with the aid of two Q-learning sequences with upper and lower confidence bounds. The design

principle of our early-settled variance reduction method might be of independent interest to other RL settings that involve intricate exploration–exploitation trade-offs.

Keywords: model-free RL; memory efficiency; variance reduction; Q-learning; upper confidence bounds; lower confidence bounds;

2010 Math Subject Classification: 68T09; 68T37; 68W27; 90C40.

1. Introduction

Contemporary reinforcement learning (RL) has to deal with unknown environments with unprecedentedly large dimensionality. How to make the best use of samples in the face of high-dimensional state/action space lies at the core of modern RL practice. An ideal RL algorithm would learn to act favorably even when the number of available data samples scales sub-linearly in the ambient dimension of the model, i.e. the number of parameters needed to describe the transition dynamics of the environment. The challenge is further compounded when this task needs to be accomplished with limited memory.

Simultaneously achieving the desired sample and memory efficiency is particularly challenging when it comes to online episodic RL scenarios. In contrast to the simulator setting that permits sampling of any state–action pair, an agent in online episodic RL is only allowed to draw sample trajectories by executing a policy in the unknown Markov decision process (MDP), where the initial states are pre-assigned and might even be chosen by an adversary. Careful deliberation needs to be undertaken when deciding what policies to use to allow for effective interaction with the unknown environment, how to optimally balance exploitation and exploration and how to process and store the collected information intelligently without causing redundancy.

1.1 *Regret-optimal model-free RL? A sample size barrier*

In order to evaluate and compare the effectiveness of RL algorithms in high dimension, a recent body of works sought to develop a finite-sample theoretical framework to analyze the algorithms of interest, with the aim of delineating the dependency of algorithm performance on all salient problem parameters in a non-asymptotic fashion [14, 33]. Such finite-sample guarantees are brought to bear toward understanding and tackling the challenges in the sample-starved regime commonly encountered in practice. To facilitate discussion, let us take a moment to summarize the state-of-the-art theory for episodic finite-horizon MDPs with non-stationary transition kernels, focusing on minimizing cumulative regret—a metric that quantifies the performance difference between the learned policy and the true optimal policy—with the fewest number of samples. Here and throughout, we denote by S, A and H the size of the state space, the size of the action space and the horizon length of the MDP, respectively, and let T represent the sample size. In addition, the immediate reward gained at each time step is assumed to lie between 0 and 1.

Fundamental regret lower bound. Following the arguments in [3, 28], the recent works [15, 29] developed a fundamental lower bound[1] on the expected total regret for this setting. Specifically, this lower bound claims that no matter what algorithm to use, one can find an MDP such that the accumulated

[1] It is worth emphasizing that [15] adopts the notation T to represent the number of trajectories (with each trajectory containing H samples), while this paper employs K to denote the number of sample trajectories and $T = KH$ the total number of samples.

regret incurred by the algorithm necessarily exceeds the order of

$$\text{(lower bound)} \qquad \sqrt{H^2SAT}, \tag{1.1}$$

as long as $T \geq H^2SA$.[2] This sublinear regret lower bound in turn imposes a sampling limit if one wants to achieve ε average regret.

Model-based RL. Moving beyond the lower bound, let us examine the effectiveness of model-based RL—an approach that can be decoupled into a model estimation stage (i.e. estimating the transition kernel using available data) and a subsequent stage of planning using the learned model [2, 7, 20, 28, 48]. In order to ensure a sufficient degree of exploration, [7] came up with an algorithm called UCB-VI that blends model-based learning and the optimism principle, which achieves a regret bound[3] $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ that nearly attains the lower bound (1.1) as T tends to infinity. Caution needs to be exercised, however, that existing theory does not guarantee the near optimality of this algorithm unless the sample size T surpasses

$$T \geq S^3AH^6,$$

a threshold that is significantly larger than the dimension of the underlying model. This threshold can also be understood as the initial *burn-in cost* of the algorithm, namely, a sampling burden needed for the algorithm to exhibit the desired performance. In addition, model-based algorithms typically require storing the estimated probability transition kernel, resulting in a space complexity that could be as high as $O(S^2AH)$ [7].

Model-free RL. Another competing solution paradigm is the model-free approach, which circumvents the model estimation stage and attempts to learn the optimal values directly [8, 29, 56, 70]. In comparison with the model-based counterpart, the model-free approach holds the promise of low space complexity, as it eliminates the need of storing a full description of the model. In fact, in a number of previous works (e.g. [29, 56]), an algorithm is declared to be model-free only if its space complexity is $o(S^2AH)$ regardless of the sample size T.

- *Memory-efficient model-free methods.* [29] proposed the first memory-efficient model-free algorithm—which is an optimistic variant of classical Q-learning—that achieves a regret bound proportional to $\sqrt{T}$ with a space complexity $O(SAH)$. Compared with the lower bound (1.1), however, the regret bound in [29] is off by a factor of $\sqrt{H}$ and hence suboptimal for problems with long horizon. This drawback has recently been overcome in [75] by leveraging the idea of variance reduction (or the so-called 'reference-advantage decomposition') for large enough T. While the resulting regret matches the information-theoretic limit asymptotically, its optimality in the non-asymptotic regime is not guaranteed unless the sample size T exceeds (see (75, Lemma 7))

$$T \geq S^6A^4H^{28},$$

Consequently, the lower bound developed in [15] for non-stationary finite-horizon MDPs reads $\Omega(\sqrt{H^3SAK})$, or equivalently, $\Omega(\sqrt{H^2SAT})$ using the notation adopted herein.

[2] Given that a trivial upper bound on the regret is T, one needs to impose a lower bound $T \geq H^2SA$ in order for (1.1) to be meaningful.

[3] Here and throughout, we use the standard notation $f(n) = O(g(n))$ to indicate that $f(n)/g(n)$ is bounded above by a constant as n grows. The notation $\widetilde{O}(\cdot)$ resembles $O(\cdot)$ except that it hides any logarithmic scaling. The notation $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$.

TABLE 1 *Comparisons between prior art and our results for non-stationary episodic MDPs when $T \geq H^2SA$. The table includes the order of the regret bound, the range of sample sizes that achieve the optimal regret $\widetilde{O}(\sqrt{H^2SAT})$, and the memory complexity, with all logarithmic factors omitted for simplicity of presentation. The red text highlights the suboptimal part of the respective algorithms.*

Algorithm	Regret	Range of sample sizes T that attain optimal regret	Space complexity
UCB-VI [7]	$\sqrt{H^2SAT} + H^4S^2A$	$[S^3AH^6, \infty)$	S^2AH
UCB-Q-Hoeffding [29]	$\sqrt{H^4SAT}$	never	SAH
UCB-Q-Bernstein [29]	$\sqrt{H^3SAT} + \sqrt{H^9S^3A^3}$	never	SAH
UCB2-Q-Bernstein [8]	$\sqrt{H^3SAT} + \sqrt{H^9S^3A^3}$	never	SAH
UCB-Q-Advantage [75]	$\sqrt{H^2SAT} + H^8S^2A^{\frac{3}{2}}T^{\frac{1}{4}}$	$[S^6A^4H^{28}, \infty)$	SAH
UCB-M-Q [44]	$\sqrt{H^2SAT} + H^4SA$	$[SAH^6, \infty)$	S^2AH
Q-EarlySettled-Advantage **(this work)**	$\sqrt{H^2SAT} + H^6SA$	$[SAH^{10}, \infty)$	SAH
Lower bound [15]	$\sqrt{H^2SAT}$	n/a	n/a

a requirement that is even far more stringent than the burn-in cost imposed by [7].

- *A memory-inefficient 'model-free' variant.* The recent work [44] put forward a novel sample-efficient variant of Q-learning called UCB-M-Q, which relies on a carefully chosen momentum term for bias reduction. This algorithm is guaranteed to yield near-optimal regret $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ as soon as the sample size exceeds $T \geq SA\mathrm{poly}(H)$, which is a remarkable improvement vis-à-vis previous regret-optimal methods [7, 75]. Nevertheless, akin to the model-based approach, it comes at a price in terms of the space and computation complexities, as the space required to store all bias-value function is $O(S^2AH)$ and the computation required is $O(ST)$, both of which are larger by a factor of S than other model-free algorithms like [75]. In view of this memory inefficiency, UCB-M-Q falls short of fulfilling the definition of model-free algorithms in [29, 56]. See (44, Section 3.3) for more detailed discussions.

A more complete summary of prior results can be found in Table 1.

1.2 *A glimpse of our contributions*

In brief, while it is encouraging to see that both model-based and model-free approaches allow for near-minimal regret as T tends to infinity, they are either memory-inefficient, or require the sample size to exceed a threshold substantially larger than the model dimensionality. In fact, no prior algorithms have been shown to be *simultaneously regret-optimal and memory-efficient* unless

$$T \geq S^6A^4\,\mathrm{poly}(H),$$

which constitutes a stringent sample size barrier constraining their utility in the sample-starved and memory-limited regime. The presence of this sample complexity barrier motivates one to pose a natural question:

> *Is it possible to design an algorithm that accommodates a significantly broader sample size range without compromising regret optimality and memory efficiency?*

In this paper, we answer this question affirmatively, by designing a new model-free algorithm, dubbed as Q-EarlySettled-Advantage, that enjoys the following performance guarantee.

THEOREM 1.1. informalThe proposed Q-EarlySettled-Advantage algorithm, which has a space complexity $O(SAH)$, achieves near-optimal regret $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ as soon as the sample size exceeds $T \geq SA\,\mathrm{poly}(H)$.

As can be seen from Table 1, the space complexity of the proposed algorithm is $O(SAH)$, which is far more memory-efficient than both the model-based approach in [7] and the UCB-M-Q algorithm in [44] (both of these prior algorithms require S^2AH units of space). In addition, the sample size requirement $T \geq SA\,\mathrm{poly}(H)$ of our algorithm improves—by a factor of at least S^5A^3—upon that of any prior algorithm that is simultaneously regret-optimal and memory-efficient. In fact, this requirement is nearly sharp in terms of the dependency on both S and A, and was previously achieved only by the UCB-M-Q algorithm at a price of a much higher storage burden.

Let us also briefly highlight the key ideas of our algorithm. As an optimistic variant of variance-reduced Q-learning, Q-EarlySettled-Advantage leverages the recently introduced reference-advantage decompositions for variance reduction [75]. As a distinguishing feature from prior algorithms, we employ an *early-stopped* reference update rule, with the assistance of two Q-learning sequences that incorporate upper and lower confidence bounds (LCBs), respectively. The design of our early-stopped variance reduction scheme, as well as its analysis framework, might be of independent interest to other settings that involve managing intricate exploration–exploitation trade-offs.

1.3 *Related works*

We now take a moment to discuss a small sample of other related works. We limit our discussions primarily to RL algorithms in the tabular setting with finite state and action spaces, which are the closest to our work. The readers interested in those model-free variants with function approximation are referred to [19, 22, 45] and the references therein.

Probably approximately correct (PAC) bounds for synchronous and asynchronous Q-learning. Q-learning is arguably among the most famous model-free algorithms developed in the RL literature [26, 57, 60, 65], which enjoys a space complexity $O(SAH)$. Non-asymptotic sample analysis and PAC bounds have seen extensive developments in the last several years, including but not limited to the works of [10, 12, 21, 37, 62] for the synchronous setting (the case with access to a generative model or a simulator), and the works of [10, 13, 21, 42, 50] for the asynchronous setting (where one observes a single Markovian trajectory induced by a behavior policy). Finite-time guarantees of other variants of Q-learning have also been developed; partial examples include speedy Q-learning [5], double Q-learning [68], variance-reduced Q-learning [42, 63], momentum Q-learning [67], pessimistic Q-learning [53] and Q-learning for linearly parameterized MDPs [64]. This line of works did not account for exploration, and hence the success of Q-learning in these settings heavily relies on the access to a simulator or a behavior policy with sufficient coverage over the state-action space.

Regret analysis for model-free RL with exploration. When it comes to online episodic RL (so that a simulator is unavailable), regret analysis is the prevailing analysis paradigm employed to capture the

trade-off between exploration and exploitation. A common theme is to augment the original model-free update rule (e.g. the Q-learning update rule) by an exploration bonus, which typically takes the form of, say, certain upper confidence bounds (UCBs) motivated by the bandit literature [4, 35]. In addition to the ones in Table 1 for episodic finite-horizon settings, sample-efficient model-free algorithms have been investigated for infinite-horizon MDPs as well [16, 27, 43, 70, 74, 76].

Variance reduction in RL. The seminal idea of variance reduction was originally proposed to accelerate finite-sum stochastic optimization, e.g. [24, 32, 46]. Thereafter, the variance reduction strategy has been imported to RL, which assists in improving the sample efficiency of RL algorithms in multiple contexts, including but not limited to policy evaluation [17, 34, 61, 69], RL with a generative model [54, 55, 63], asynchronous Q-learning [42] and offline RL [53, 71].

Model-based approach. Model-based RL is known to be minimax-optimal in the presence of a simulator [1, 6, 41], beating the state-of-the-art model-free algorithms by achieving optimality for the entire sample size range [41]. When it comes to online episodic RL, [7] was the first work that managed to achieve near-optimal regret (at least for large T); in fact, this was also the first result (for any algorithm) matching existing lower bounds for large T. The sample efficiency of the model-based approach has subsequently been established for other settings, including but not limited to discounted infinite-horizon MDPs [25], MDPs with bounded total reward [72, 74], offline RL [40, 52] and Markov games [39, 73].

Regret lower bound. Inspired by the classical lower bound argument developed for multi-armed bandits [3], the work [28] established a regret lower bound for MDPs with finite diameters (so that for an arbitrary pair of states, the expected time to transition between them is assumed to be finite as long as a suitable policy is used), which has been reproduced in the note [47] with the purpose of facilitating comparison with [9]. The way to construct hard MDPs in [28] has since been adapted by [29] to exhibit a lower bound on episodic MDPs (with a sketched proof provided therein). It was recently revisited in [15], which presented a detailed and rigorous proof argument with a different construction.

2. Problem formulation

In this section, we formally describe the problem setting. Here and throughout, we denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$, and introduce the notation $[M] := \{1, \cdots, M\}$ for any integer $M > 0$.

Basics of finite-horizon MDPs. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$ represent a finite-horizon MDP, where $\mathcal{S} := \{1, \cdots S\}$ is the state space of size S, $\mathcal{A} := \{1, \cdots, A\}$ is the action space of size A, H denotes the horizon length 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, $1 \le h \le H$, respectively. More specifically, $P_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ stands for the transition probability vector from state s at time step h when action a is taken, while $r_h(s, a)$ indicates the immediate reward received at time step h for a state–action pair (s, a) (which is assumed to be deterministic and fall within the range $[0, 1]$). The MDP is said to be non-stationary when the P_h's are not identical across $1 \le h \le H$. A policy of an agent is represented by $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{S} \to \mathcal{A}$ the action selection rule at time step h, so that $\pi_h(s)$ specifies which action to execute in state s at time step h. Throughout this paper, we concentrate on deterministic policies.

Value functions, Q-functions and Bellman equations. The value function $V_h^\pi(s)$ of a (deterministic) policy π at step h is defined as the expected cumulative rewards received between time steps h and H when executing this policy from an initial state s at time step h, namely,

$$V_h^\pi(s) := \mathop{\mathbb{E}}_{s_{t+1} \sim P_t(\cdot \mid s_t, \pi_t(s_t)), t \ge h} \left[\sum_{t=h}^{H} r_t\big(s_t, \pi_t(s_t)\big) \,\Big|\, s_h = s \right], \tag{2.1}$$

where the expectation is taken over the randomness of the MDP trajectory $\{s_t \mid h \le t \le H\}$. The action-value function (a.k.a. the Q-function) $Q_h^\pi(s,a)$ of a policy π at step h can be defined analogously except that the action at step h is fixed to be a, that is,

$$Q_h^\pi(s,a) := r_h(s,a) + \mathop{\mathbb{E}}_{\substack{s_{h+1}\sim P_h(\cdot|s,a),\\ s_{t+1}\sim P_t(\cdot|s_t,\pi_t(s_t)),\, t>h}} \left[\sum_{t=h+1}^{H} r_t\big(s_t,\pi_t(s_t)\big) \,\Big|\, s_h = s, a_h = a \right]. \tag{2.2}$$

In addition, we define $V_{H+1}^\pi(s) = Q_{H+1}^\pi(s,a) = 0$ for any policy π and any state–action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. By virtue of basic properties in dynamic programming [11], the value function and the Q-function satisfy the following Bellman equation:

$$Q_h^\pi(s,a) = r_h(s,a) + \mathop{\mathbb{E}}_{s'\sim P_h(\cdot|s,a)} \big[V_{h+1}^\pi(s')\big]. \tag{2.3}$$

A policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ is said to be an optimal policy if it maximizes the value function simultaneously for all states among all policies. The resulting optimal value function $V^\star = \{V_h^\star\}_{h=1}^H$ and optimal Q-functions $Q^\star = \{Q_h^\star\}_{h=1}^H$ satisfy

$$V_h^\star(s) = V_h^{\pi^\star}(s) = \max_\pi V_h^\pi(s) \qquad \text{and} \qquad Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a) = \max_\pi Q_h^\pi(s,a) \tag{2.4}$$

for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. It is well known that the optimal policy always exists [49], and satisfies the Bellman optimality equation:

$$\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]: \qquad Q_h^\star(s,a) = r_h(s,a) + \mathop{\mathbb{E}}_{s'\sim P_h(\cdot|s,a)} \big[V_{h+1}^\star(s')\big]. \tag{2.5}$$

Online episodic RL. This paper investigates the online episodic RL setting, where the agent is allowed to execute the MDP sequentially in a total number of K episodes each of length H. This amounts to collecting

$$T = KH \text{ samples}$$

in total. More specifically, in each episode $k = 1,\ldots,K$, the agent is assigned an arbitrary initial state s_1^k (possibly by an adversary), and selects a policy $\pi^k = \{\pi_h^k\}_{h=1}^H$ learned based on the information collected up to the $(k-1)$-th episode. The k-th episode is then executed following the policy π^k and the dynamic of the MDP $\mathcal{M}$, leading to a length-H sample trajectory.

Goal: regret minimization. In order to evaluate the quality of the learned policies $\{\pi^k\}_{1\le k\le K}$, a frequently used performance metric is the cumulative regret defined as follows:

$$\mathrm{Regret}(T) := \sum_{k=1}^{K} \big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\big). \tag{2.6}$$

In words, the regret reflects the sub-optimality gaps between the values of the optimal policy and those of the learned policies aggregated over K episodes. A natural objective is thus to design a sample-optimal algorithm, namely, an algorithm whose resulting regret scales optimally in the sample size T.

Accomplishing this goal requires carefully managing the trade-off between exploration and exploitation, which is particularly challenging in the sample-limited regime.

Notation. Before presenting our main results, we take a moment to introduce some convenient notation to be used throughout the remainder of this paper. For any vector $x \in \mathbb{R}^{SA}$ that constitutes certain quantities for all state–action pairs, we shall often use $x(s,a)$ to denote the entry associated with the state–action pair (s,a), as long as it is clear from the context. We shall also let

$$P_{h,s,a} = P_h(\cdot \mid s,a) \in \mathbb{R}^{1\times S} \tag{2.7}$$

abbreviate the transition probability vector given the (s,a) pair at time step h. Additionally, we denote by e_i the i-th standard basis vector, with the only non-zero element being in the i-th entry and equal to 1.

3. Algorithm and theoretical guarantees

In this section, we present the proposed algorithm called Q-EarlySettled-Advantage, as well as the accompanying theory confirming its sample and memory efficiency.

3.1 *Review: Q-learning with UCB exploration and reference advantage*

This subsection briefly reviews the Q-learning algorithm with UCB exploration proposed in [29], as well as a variant that further exploits the idea of variance reduction [75]. These two model-free algorithms inspire the algorithm design in this paper.

Q-learning with UCB exploration (UCB-Q or UCB-Q-Hoeffding). Recall that the classical Q-learning algorithm has been proposed as a stochastic approximation scheme [51] to solve the Bellman optimality equation (2.5), which consists of the following update rule [65, 66]:

$$Q_h(s,a) \leftarrow (1-\eta)Q_h(s,a) + \eta\Big\{r_h(s,a) + \underbrace{\widehat{P}_{h,s,a}V_{h+1}}_{\text{stochastic estimate of } P_{h,s,a}V_{h+1}}\Big\}. \tag{3.1}$$

Here, Q_h (resp. V_h) indicates the running estimate of $Q_h^\star$ (resp. $V_h^\star$), η is the (possibly iteration-varying) learning rate or stepsize and $\widehat{P}_{h,s,a}V_{h+1}$ is a stochastic estimate of $P_{h,s,a}V_{h+1}$ (cf. (2.7)). For instance, if one has available a sample (s,a,s') transitioning from state s at step h to s' at step $h+1$ under action a, then a stochastic estimate of $P_{h,s,a}V_{h+1}$ can be taken as $V_{h+1}(s')$, which is unbiased in the sense that

$$\mathbb{E}\big[V_{h+1}(s')\big] = P_{h,s,a}V_{h+1}.$$

To further encourage exploration, the algorithm proposed in [29]—which shall be abbreviated as UCB-Q or UCB-Q-Hoeffding hereafter—augments the Q-learning update rule (3.1) in each episode via an additional exploration bonus:

$$Q_h^{\mathsf{UCB}}(s,a) \leftarrow (1-\eta)Q_h^{\mathsf{UCB}}(s,a) + \eta\big\{r_h(s,a) + \widehat{P}_{h,s,a}V_{h+1} + b_h\big\}. \tag{3.2}$$

The bonus term $b_h \geq 0$ is chosen to be a certain UCB for $(\widehat{P}_{h,s,a} - P_{h,s,a})V_{h+1}$, an exploration-efficient scheme that originated from the bandit literature [35, 36]. The algorithm then proceeds to the next episode by executing/sampling the MDP using a greedy policy w.r.t. the updated Q-estimate. These steps are repeated until the algorithm is terminated.

Q-learning with UCB exploration and reference advantage (UCB-Q-Advantage). The regret bounds derived for UCB-Q [29], however, fall short of being optimal, as they are at least a factor of $\sqrt{H}$ away from the fundamental lower bound. In order to further shave this $\sqrt{H}$ factor, one strategy is to leverage the idea of variance reduction to accelerate convergence [32, 42, 55, 63]. An instantiation of this idea for the regret setting is a variant of UCB-Q based on reference-advantage decomposition, which was put forward in [75] and shall be abbreviated as UCB-Q-Advantage throughout this paper.

To describe the key ideas of UCB-Q-Advantage, imagine that we are able to maintain a collection of reference values $V^{\mathrm{R}} = \{V_h^{\mathrm{R}}\}_{h=1}^H$, which form reasonable estimates of $V^\star = \{V_h^\star\}_{h=1}^H$ and become increasingly more accurate as the algorithm progresses.

At each time step h, the algorithm adopts the following update rule:

$$Q_h^{\mathrm{R}}(s,a) \leftarrow (1-\eta)Q_h^{\mathrm{R}}(s,a) + \eta\Big\{r_h(s,a) + \underbrace{\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathrm{R}}\big)}_{\text{stochastic estimate of } P_{h,s,a}\left(V_{h+1}-V_{h+1}^{\mathrm{R}}\right)} + \big[\widehat{P_h V_{h+1}^{\mathrm{R}}}\big](s,a) + b_h^{\mathrm{R}}\Big\}. \tag{3.3}$$

Two ingredients of this update rule are worth noting:

- Akin to the UCB-Q case, we can take $\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathrm{R}}\big)$ to be the stochastic estimate $V_{h+1}(s') - V_{h+1}^{\mathrm{R}}(s')$ if we observe a sample transition (s,a,s') at time step h. If V_{h+1} is fairly close to the reference V_{h+1}^{R}, then this stochastic term can be less volatile than the stochastic term $\widehat{P}_{h,s,a}V_{h+1}$ in (3.2).
- Additionally, the term $\widehat{P_h V_{h+1}^{\mathrm{R}}}$ indicates an estimate of the one-step look-ahead value $P_h V_{h+1}^{\mathrm{R}}$, which shall be computed using a batch of samples.
- The variability of $\widehat{P_h V_{h+1}^{\mathrm{R}}}$ can be well controlled through the use of batch data, at the price of an increased sample size.

Accordingly, the exploration bonus term b_h^{R} is taken to be a UCB for the above-mentioned two terms combined. Given that the uncertainty of (3.3) largely stems from these two terms (which can both be much smaller than the variability in (3.2)), the incorporation of the reference term helps accelerate convergence.

3.2 *The proposed algorithm:* Q-EarlySettled-Advantage

As alluded to previously, however, the sample size required for UCB-Q-Advantage to achieve optimal regret needs to exceed a large polynomial S^6A^4 in the size of the state/action space. To overcome this sample complexity barrier, we come up with an improved variant called Q-EarlySettled-Advantage.

Motivation: early settlement of a reference value. An important insight obtained from previous algorithm designs is that in order to achieve low regret, it is desirable to maintain an estimate of Q-function that (i) provides an optimistic view (namely, an over-estimate) of the truth $Q^\star$, and (ii) mitigates the bias $Q - Q^\star$ as much as possible. With two additional optimistic Q-estimates in hand—Q_h^{UCB} based on UCB-Q, and a reference Q_h^{R}—it is natural to combine them as follows to further reduce the bias without violating the optimism principle:

$$Q_h(s_h,a_h) \leftarrow \min\Big\{Q_h^{\mathrm{R}}(s_h,a_h),\, Q_h^{\mathsf{UCB}}(s_h,a_h),\, Q_h(s_h,a_h)\Big\}. \tag{3.4}$$

This is precisely what is conducted in UCB-Q-Advantage. However, while the estimate Q_h^{R} obtained with the aid of reference-advantage decomposition provides great promise, fully realizing its potential in the sample-limited regime relies on the ability to quickly *settle on* a desirable 'reference' during the initial stage of the algorithm. This leads us to a dilemma that requires careful thinking. On the one hand, the reference value V^{R} needs to be updated in a timely manner in order to better control the stochastic estimate of $P_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathrm{R}}\big)$. On the other hand, updating V^{R} too frequently incurs an overly large sample size burden, as new samples need to be accumulated whenever V^{R} is updated.

Built upon the above insights, it is advisable to prevent frequent updating of the reference value V^{R}. In fact, it would be desirable to stop updating the reference value once a point of sufficient quality—denoted by V^{R}, final—has been obtained. Locking on a reasonable reference value early on means that (a) fewer samples will be wasted on estimating a drifting target $P_h V_{h+1}^{\mathrm{R}}$, and (b) all ensuing samples can then be dedicated to estimating the key quantity of interest $P_h V^{\mathrm{R}}$, final_{h+1}.

REMARK 1. In [75], the algorithm UCB-Q-Advantage requires collecting $\widetilde{O}\big(SAH^6\big)$ samples *for each state* before settling on the reference value, which inevitably contributes to the large burn-in cost.

Algorithm 1 Q-EarlySettled-Advantage

Parameters: some universal constant $c_{\mathrm{b}} > 0$ and probability of failure $\delta \in (0,1)$;
Initialize $Q_h(s,a), Q_h^{\mathsf{UCB}}(s,a), Q_h^{\mathsf{R}}(s,a) \leftarrow H$; $V_h(s), V_h^{\mathsf{R}}(s) \leftarrow H$; $Q_h^{\mathsf{LCB}}(s,a) \leftarrow 0$; $V_h^{\mathsf{LCB}}(s) \leftarrow 0$; $N_h(s,a) \leftarrow 0$; $\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \delta_h^{\mathsf{R}}(s,a), B_h^{\mathsf{R}}(s,a) \leftarrow 0$; and $u_{\mathrm{ref}}(s) = \mathsf{True}$ for all $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$.
for Episode $k = 1$ **to** K **do**
  Set initial state $s_1 \leftarrow s_1^k$.
  for Step $h = 1$ **to** H **do**
    Take action $a_h = \pi_h^k(s_h) = \arg\max_a Q_h(s_h,a)$, and draw $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$. `// sampling`
    $N_h(s_h,a_h) \leftarrow N_h(s_h,a_h) + 1$; $n \leftarrow N_h(s_h,a_h)$. `// update the counter`
    $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`
    $Q_h^{\mathsf{UCB}}(s_h,a_h) \leftarrow$ `update-ucb-q()`. `// run UCB-Q; see Algorithm 2`
    $Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow$ `update-lcb-q()`. `// run LCB-Q; see Algorithm 2`
    $Q_h^{\mathsf{R}}(s_h,a_h) \leftarrow$ `update-ucb-q-advantage()`. `// estimate` Q_h^{R}`; see Algorithm 2`
      `/* update Q-estimates using all estimates in hand, and update value estimates */`
    $Q_h(s_h,a_h) \leftarrow \min\big\{Q_h^{\mathsf{R}}(s_h,a_h), Q_h^{\mathsf{UCB}}(s_h,a_h), Q_h(s_h,a_h)\big\}$.
    $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$.
    $V_h^{\mathsf{LCB}}(s_h) \leftarrow \max\big\{\max_a Q_h^{\mathsf{LCB}}(s_h,a), V_h^{\mathsf{LCB}}(s_h)\big\}$.
      `/* update reference values */`
    if $V_h(s_h) - V_h^{\mathsf{LCB}}(s_h) > 1$ **then**
      $V_h^{\mathsf{R}}(s_h) \leftarrow V_h(s_h)$, $u_{\mathrm{ref}}(s_h) = \mathsf{True}$,
    else if $u_{\mathrm{ref}}(s_h) = \mathsf{True}$ **then**
      $V_h^{\mathsf{R}}(s_h) \leftarrow V_h(s_h)$, $u_{\mathrm{ref}}(s_h) = \mathsf{False}$.
    end if
  end for
end for

The proposed Q-EarlySettled-Advantage algorithm. We now propose a new model-free algorithm that allows for early settlement of the reference value. A few key ingredients are as follows.

- *An auxiliary sequence based on LCB.* In addition to the two optimistic Q-estimates Q_h^{R} and Q_h^{UCB} described previously, we intend to maintain another *pessimistic* estimate $Q_h^{\mathsf{LCB}} \le Q_h^\star$ using the subroutine `update-lcb-q`, based on LCBs. We will also maintain the corresponding value function V_h^{LCB}, which lower bounds $V_h^\star$.
- *Termination rules for reference updates.* With $V_h^{\mathsf{LCB}} \le V_h^\star$ in place, the updates of the references (lines 15-18 of Algorithm 1) are designed to terminate when

$$V_h(s_h) \le V_h^{\mathsf{LCB}}(s_h) + 1 \le V_h^\star(s_h) + 1. \tag{3.5}$$

  Note that V_h^{R} keeps tracking the value of V_h before it stops being updated. In effect, when the additional condition in lines 15 is violated and thus (3.5) is satisfied, we claim that it is unnecessary to update the reference V_h^{R} afterwards, since it is of sufficient quality (being close enough to the optimal value $V_h^\star$) and further drifting the reference does not appear beneficial. As we will make it rigorous shortly, this reference update rule is sufficient to ensure that $|V_h - V_h^{\mathrm{R}}| \le 2$ throughout the execution of the algorithm, which in turn suggests that the standard deviation of $\widehat{P}_{h,s,a}(V_{h+1} - V_{h+1}^{\mathrm{R}})$ might be $O(H)$ times smaller than that of $\widehat{P}_{h,s,a}V_{h+1}$ (i.e. the stochastic term used in (3.1) of UCB-Q). This is a key observation that helps shave the addition factor H in the regret bound of UCB-Q.
- *Update rules for Q_h^{UCB} and Q_h^{R}.* The two optimistic Q-estimates Q_h^{UCB} and Q_h^{R} are updated using the subroutine `update-ucb-q` (following the standard Q-learning with Hoeffding bonus [29]) and `update-ucb-q-advantage`, respectively. Note that Q_h^{R} continues to be updated even after V_h^{R} is no longer updated.

Q-learning with reference-advantage decomposition. The rest of this subsection is devoted to explaining the subroutine `update-ucb-q-advantage`, which produces a Q-estimate Q^{R} based on the reference-advantage decomposition similar to [75]. To facilitate the implementation, let us introduce the parameters associated with a reference value V^{R}, which include six different components, i.e.

$$\left[\mu_h^{\mathrm{ref}}(s,a), \sigma_h^{\mathrm{ref}}(s,a), \mu_h^{\mathrm{adv}}(s,a), \sigma_h^{\mathrm{adv}}(s,a), \delta_h^{\mathrm{R}}(s,a), B_h^{\mathrm{R}}(s,a)\right], \tag{3.6}$$

for all $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$. Here, $\mu_h^{\mathrm{ref}}(s,a)$ and $\sigma_h^{\mathrm{ref}}(s,a)$ estimate the running mean and the second moment of the reference $\left[P_h V_{h+1}^{\mathrm{R}}\right](s,a)$; $\mu_h^{\mathrm{adv}}(s,a)$ and $\sigma_h^{\mathrm{adv}}(s,a)$ estimate the running (weighted) mean and the second moment of the advantage $\left[P_h(V_{h+1} - V_{h+1}^{\mathrm{R}})\right](s,a)$; $B_h^{\mathrm{R}}(s,a)$ aggregates the empirical standard deviations of the reference and the advantage combined; and last but not least, $\delta_h^{\mathrm{R}}(s,a)$ is the temporal difference between $B_h^{\mathrm{R}}(s,a)$ and its previous value.

As alluded to previously, the Q-function estimation follows the strategy (3.3) at a high level. Upon observing a sample transition (s_h, a_h, s_{h+1}), we compute the following estimates to update $Q^{\mathrm{R}}(s_h,a_h)$.

- The term $\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathrm{R}}\big)$ is set to be $V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{R}}(s_{h+1})$, which is an unbiased stochastic estimate of $P_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathrm{R}}\big)$.

- The term $\left[P_h V_{h+1}^{\mathrm{R}}\right](s,a)$ is estimated via $\mu^{\mathrm{ref}}, \mathrm{R}_h$ (cf. line 11). Given that this is estimated using all previous samples, we expect the variability of this term to be well-controlled as the sample size increases (especially after V^{R} is locked).
- The exploration bonus $b_h^{\mathrm{R}}(s,a)$ is updated using $B_h^{\mathrm{R}}(s_h,a_h)$ and $\delta_h^{\mathrm{R}}(s_h,a_h)$ (cf. lines 7-8 of Algorithm 2), which is a confidence bound accounting for both the reference and the advantage. Let us also explain line 8 of Algorithm 2 a bit. If we augment the notation by letting $b_h^{\mathrm{R},n+1}(s,a)$ and $B_h^{\mathrm{R},n+1}(s,a)$ denote, respectively, $b_h^{\mathrm{R}}(s,a)$ and $B_h^{\mathrm{R}}(s,a)$ after (s,a) is visited for the n-th time, then this line is designed to ensure that

$$\eta_n b_h^{\mathrm{R},n+1}(s,a) + (1-\eta_n) B_h^{\mathrm{R,n}}(s,a) \approx B_h^{\mathrm{R},n+1}(s,a).$$

With the above updates implemented properly, Q_h^{R} provides the advantage-based update of the Q-function at time step h, according to the update rule (3.3).

Algorithm 2 Auxiliary functions

1: **function** `update-ucb-q()`:
2: $Q_h^{\mathsf{UCB}}(s_h,a_h) \leftarrow (1-\eta_n) Q_h^{\mathsf{UCB}}(s_h,a_h) + \eta_n\Big(r_h(s_h,a_h) + V_{h+1}(s_{h+1}) + c_{\mathrm{b}}\sqrt{\frac{H^3 \log\frac{SAT}{\delta}}{n}}\Big)$.
3: **function** `update-lcb-q()`:
4: $Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow (1-\eta_n) Q_h^{\mathsf{LCB}}(s_h,a_h) + \eta_n\Big(r_h(s_h,a_h) + V_{h+1}^{\mathsf{LCB}}(s_{h+1}) - c_{\mathrm{b}}\sqrt{\frac{H^3 \log\frac{SAT}{\delta}}{n}}\Big)$.
5: **function** `update-ucb-q-advantage()`:
`/* update the moment statistics of` V_h^{R} `*/`
6: $[\mu_h^{\mathsf{ref}}, \sigma_h^{\mathsf{ref}}, \mu_h^{\mathsf{adv}}, \sigma_h^{\mathsf{adv}}](s_h,a_h) \leftarrow$ `update-moments()`;
`/* update the accumulative bonus and bonus difference */`
7: $[\delta_h^{\mathrm{R}}, B_h^{\mathrm{R}}](s_h,a_h) \leftarrow$ `update-bonus()`;
8: $b_h^{\mathrm{R}} \leftarrow B_h^{\mathrm{R}}(s_h,a_h) + (1-\eta_n)\frac{\delta_h^{\mathrm{R}}(s_h,a_h)}{\eta_n} + c_{\mathrm{b}}\frac{H^2\log\frac{SAT}{\delta}}{n^{3/4}}$;
`/* update the Q-estimate based on reference-advantage decomposition */`
9: $Q_h^{\mathrm{R}}(s_h,a_h) \leftarrow (1-\eta_n) Q_h^{\mathrm{R}}(s_h,a_h) + \eta_n\big(r_h(s_h,a_h) + V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{R}}(s_{h+1}) + \mu_h^{\mathsf{ref}}(s_h,a_h) + b_h^{\mathrm{R}}\big)$;
10: **function** `update-moments()`:
11: $\mu_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\mu_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n} V_{h+1}^{\mathrm{R}}(s_{h+1})$; `// mean of the reference`
12: $\sigma_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\sigma_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n}\big(V_{h+1}^{\mathrm{R}}(s_{h+1})\big)^2$; `//` 2^{nd} `moment of the reference`
13: $\mu_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\mu_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{R}}(s_{h+1})\big)$; `// weighted average of the advantage`
14: $\sigma_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\sigma_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{R}}(s_{h+1})\big)^2$. `//` 2^{nd} `moment of the advantage`
15: **function** `update-bonus()`:
16: $B_h^{\mathsf{next}}(s_h,a_h) \leftarrow$
17: $c_{\mathrm{b}}\sqrt{\frac{\log\frac{SAT}{\delta}}{n}}\Big(\sqrt{\sigma_h^{\mathsf{ref}}(s_h,a_h) - \big(\mu_h^{\mathsf{ref}}(s_h,a_h)\big)^2} + \sqrt{H}\sqrt{\sigma_h^{\mathsf{adv}}(s_h,a_h) - \big(\mu_h^{\mathsf{adv}}(s_h,a_h)\big)^2}\Big)$;
18: $\delta_h^{\mathrm{R}}(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h) - B_h^{\mathrm{R}}(s_h,a_h)$;
19: $B_h^{\mathrm{R}}(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h)$.

3.3 *Main results*

Encouragingly, the proposed Q-EarlySettled-Advantage algorithm manages to achieve near-optimal regret even in the sample-limited and memory-limited regime, as formalized by the following theorem.

THEOREM 3.1. Consider any $\delta \in (0,1)$, and suppose that $c_b > 0$ is chosen to be a sufficiently large universal constant. Then there exists some absolute constant $C_0 > 0$ such that Algorithm 1 achieves

$$\text{Regret}(T) \le C_0\left(\sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^6SA\log^3\frac{SAT}{\delta}\right) \tag{3.7}$$

with probability at least $1-\delta$.

Theorem 3.1 delivers a non-asymptotic characterization of the performance of our algorithm Q-EarlySettled-Advantage. Several appealing features of the algorithm are noteworthy.

- *Regret optimality.* Our regret bound (3.7) simplifies to

$$\text{Regret}(T) \le \widetilde{O}\big(\sqrt{H^2SAT}\big) \tag{3.8}$$

  as long as the sample size T exceeds

$$T \ge SA\,\text{poly}(H). \tag{3.9}$$

  This sublinear regret bound (3.8) is essentially optimal, as it coincides with the existing lower bound (1.1) modulo some logarithmic factor.

- *Sample complexity and substantially reduced burn-in cost.* As an interpretation of our theory (3.8), our algorithm attains ε average regret (i.e. $\frac{1}{K}\text{Regret}(T) \le \varepsilon$) with a sample complexity

$$\widetilde{O}\Big(\frac{SAH^4}{\varepsilon^2}\Big).$$

  Crucially, the burn-in cost (3.9) is significantly lower than that of the state-of-the-art memory-efficient model-free algorithm [75] (whose optimality is guaranteed only in the range $T \ge S^6A^4\,\text{poly}(H)$).

- *Memory efficiency.* Our algorithm, which is model-free in nature, achieves a low space complexity $O(SAH)$. This is basically un-improvable for the tabular case, since even storing the optimal Q-values alone takes $O(SAH)$ units of space. In comparison, while [44] also accommodates the sample size range (3.9), the algorithm proposed therein incurs a space complexity of $O(S^2AH)$ that is S times higher than ours.

- *Computational complexity.* An additional intriguing feature of our algorithm is its low computational complexity. The runtime of Q-EarlySettled-Advantage is no larger than $O(T)$, which is proportional to the time taken to read the samples. This matches the computational cost of the model-free algorithm UCB-Q proposed in [29], and is considerably lower than that of the UCB-M-Q algorithm in [44] (which has a computational cost of at least $O(ST)$).

4. Analysis

In this section, we outline the main steps needed to prove our main result in Theorem 3.1.

4.1 *Preliminaries: basic properties about learning rates*

Before continuing, let us first state some basic facts regarding the learning rates. Akin to [29], the proposed algorithm adopts the linearly rescaled learning rate

$$\eta_n = \frac{H+1}{H+n} \tag{4.1}$$

for the n-th visit of a state–action pair at any time step h. For notation convenience, we further introduce two sequences of related quantities defined for any integer $N \geq 0$ and $n \geq 1$:

$$\eta_n^N := \begin{cases} \eta_n \prod_{i=n+1}^N (1-\eta_i), & \text{if } N > n, \\ \eta_n, & \text{if } N = n, \\ 0, & \text{if } N < n \end{cases} \quad \text{and} \quad \eta_0^N := \begin{cases} \prod_{i=1}^N (1-\eta_i) = 0, & \text{if } N > 0, \\ 1, & \text{if } N = 0. \end{cases} \tag{4.2}$$

As can be easily verified, we have

$$\sum_{n=1}^N \eta_n^N = \begin{cases} 1, & \text{if } N > 0, \\ 0, & \text{if } N = 0. \end{cases} \tag{4.3}$$

The following properties play an important role in the analysis.

LEMMA 1. For any integer $N > 0$, the following properties hold:

$$\frac{1}{N^a} \leq \sum_{n=1}^N \frac{\eta_n^N}{n^a} \leq \frac{2}{N^a}, \qquad \text{for all} \quad \frac{1}{2} \leq a \leq 1, \tag{4.4a}$$

$$\max_{1\leq n\leq N} \eta_n^N \leq \frac{2H}{N}, \qquad \sum_{n=1}^N (\eta_n^N)^2 \leq \frac{2H}{N}, \qquad \sum_{N=n}^\infty \eta_n^N \leq 1 + \frac{1}{H}. \tag{4.4b}$$

Proof. See Appendix B. □

4.2 *Additional notation used in the proof*

In order to enable a more concise description of the algorithm, we have suppressed the dependency of many quantities on the episode number k in Algorithms 1 and 2. This, however, becomes notationally inconvenient when presenting the proof. As a consequence, we shall adopt, throughout the analysis, a more complete set of notation, detailed below.

- (s_h^k, a_h^k): the state–action pair encountered and chosen at time step h in the k-th episode.
- $k_h^n(s,a)$: the index of the episode in which (s,a) is visited for the n-th time at time step h; for the sake of conciseness, we shall sometimes use the shorthand $k^n = k_h^n(s,a)$ whenever it is clear from the context.
- $k_h^n(s)$: the index of the episode in which state s is visited for the n-th time at time step h; we might sometimes abuse the notation by abbreviating $k^n = k_h^n(s)$.

- $P_h^k \in \{0,1\}^{1\times|\mathscr{S}|}$: the empirical transition at time step h in the k-th episode, namely,

$$P_h^k(s) = \mathbb{1}\big(s = s_{h+1}^k\big). \tag{4.5}$$

In addition, for several parameters of interest in Algorithm 1, we introduce the following set of augmented notation.

- $N_h^k(s,a)$ denotes $N_h(s,a)$ *by the end* of the k-th episode; for the sake of conciseness, we shall often abbreviate $N^k = N_h^k(s,a)$ or $N^k = N_h^k(s_h^k, a_h^k)$ (depending on which result we are proving).
- $Q_h^k(s,a)$, $V_h^k(s)$ and $Q_h^{\mathsf{UCB},k}(s,a)$ denote, respectively, $Q_h(s,a)$, $V_h(s)$ and $Q_h^{\mathsf{UCB}}(s,a)$ *at the beginning* of the k-th episode.
- $Q_h^{\mathsf{LCB},k}(s,a)$ and $V_h^{\mathsf{LCB},k}(s)$ denote, respectively, $Q_h^{\mathsf{LCB}}(s,a)$ and $V_h^{\mathsf{LCB}}(s)$ *at the beginning* of the k-th episode.
- $Q_h^{\mathrm{R}}, k(s,a)$, $V_h^{\mathrm{R}}, k(s)$ and $u_{\mathrm{ref}}^k(s)$ denote, respectively, $Q_h^{\mathrm{R}}(s,a)$, $V_h^{\mathrm{R}}(s)$ and $u_{\mathrm{ref}}(s)$ *at the beginning* of the k-th episode.
- $\big[\mu^{\mathrm{ref}}, k_h, \sigma^{\mathrm{ref}}, k_h, \mu^{\mathrm{adv}}, k_h, \sigma^{\mathrm{adv}}, k_h, \delta^{\mathrm{R}}, k_h, B^{\mathrm{R}}, k_h\big]$ denotes $\big[\mu_h^{\mathrm{ref}}, \sigma_h^{\mathrm{ref}}, \mu_h^{\mathrm{adv}}, \sigma_h^{\mathrm{adv}}, \delta_h^{\mathrm{R}}, B_h^{\mathrm{R}}\big]$ *at the beginning* of the k-th episode.

Further, for any matrix $P = [P_{i,j}]_{1\le i\le m, 1\le j\le n}$, we define $\|P\|_1 := \max_{1\le i\le m}\sum_{j=1}^n |P_{i,j}|$. For any vector $V = [V_i]_{1\le i\le n}$, we define its ℓ_∞ norm as $\|V\|_\infty := \max_{1\le i\le n}|V_i|$. We often overload scalar functions and expressions to take vector-valued arguments, with the understanding that they are applied in an entrywise manner. For example, for a vector $x = [x_i]_{1\le i\le n}$, we denote $x^2 = [x_i^2]_{1\le i\le n}$. 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 $1 \le i \le n$. For any given vector $V \in \mathbb{R}^S$, we define the variance parameter w.r.t. $P_{h,s,a}$ (cf. (2.7)) as follows:

$$\mathrm{Var}_{h,s,a}(V) := \mathop{\mathbb{E}}_{s'\sim P_{h,s,a}}\Big[\big(V(s') - P_{h,s,a}V\big)^2\Big] = P_{h,s,a}\big(V^2\big) - \big(P_{h,s,a}V\big)^2. \tag{4.6}$$

Finally, let $\mathscr{X} := \big(S, A, H, T, \frac{1}{\delta}\big)$. The notation $f(\mathscr{X}) \lesssim g(\mathscr{X})$ (resp. $f(\mathscr{X}) \gtrsim g(\mathscr{X})$) means that there exists a universal constant $C_0 > 0$ such that $f(\mathscr{X}) \le C_0 g(\mathscr{X})$ (resp. $f(\mathscr{X}) \ge C_0 g(\mathscr{X})$); the notation $f(\mathscr{X}) \asymp g(\mathscr{X})$ means that $f(\mathscr{X}) \lesssim g(\mathscr{X})$ and $f(\mathscr{X}) \gtrsim g(\mathscr{X})$ hold simultaneously.

4.3 *Key properties of Q-estimates and auxiliary sequences*

In this subsection, we introduce several key properties of our Q-estimates and value estimates, which play a crucial role in the proof of Theorem 3.1. The proofs for this subsection are deferred to Appendix C.s

Properties of the Q-estimate Q_h^k: monotonicity and optimism. We first make an important observation regarding the monotonicity of the value estimates Q_h^k and V_h^k. To begin with, it is straightforward to see that the update rule in Algorithm 3 (cf. line 12) ensures the following monotonicity property:

$$Q_h^{k+1}(s,a) \le Q_h^k(s,a) \qquad \text{for all } (s,a,k,h) \in \mathscr{S}\times\mathscr{A}\times[K]\times[H], \tag{4.7a}$$

Algorithm 3 Q-EarlySettled-Advantage (rewrite of Algorithm 1)

Parameters: some universal constant $c_{\mathrm{b}} > 0$ and probability of failure $\delta \in (0,1)$;
Initialize $Q_h^1(s,a), Q_h^{\mathsf{UCB},1}(s,a), Q_h^{\mathsf{R},1}(s,a) \leftarrow H$; $Q_h^{\mathsf{LCB},1}(s,a) \leftarrow 0$; $N_h^0(s,a) \leftarrow 0$; $V_h^1(s), V_h^{\mathsf{R},1}(s) \leftarrow H$; $\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \delta_h^{\mathsf{R}}(s,a), B_h^{\mathsf{R}}(s,a) \leftarrow 0$; and $u_{\mathrm{ref}}^1(s) = \mathsf{True}$, for all $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$.
for Episode $k = 1$ **to** K **do**
  Set initial state $s_1 \leftarrow s_1^k$.
  for Step $h = 1$ **to** H **do**
    Take action $a_h^k = \pi_h^k(s_h) = \arg\max_a Q_h^k(s_h^k, a)$, and draw $s_{h+1}^k \sim P_h(\cdot \,|\, s_h^k, a_h^k)$. `// sampling`
    $N_h^k(s_h^k, a_h^k) \leftarrow N_h^{k-1}(s_h^k, a_h^k) + 1$; $n \leftarrow N_h^k(s_h^k, a_h^k)$. `// update the counter`
    $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`
    $Q_h^{\mathsf{UCB},k+1}(s_h^k, a_h^k) \leftarrow$ `update-ucb-q()`. `// run` UCB-Q;
    $Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) \leftarrow$ `update-lcb-q()`. `// run` LCB-Q;
    $Q_h^{\mathsf{R},k+1}(s_h^k, a_h^k) \leftarrow$ `update-ucb-q-advantage()`. `// estimate` Q_h^{R};
      `/* update Q-estimates using all estimates in hand, and update value estimates */`
    $Q_h^{k+1}(s_h^k, a_h^k) \leftarrow \min\{Q_h^{\mathsf{R},k+1}(s_h^k, a_h^k), Q_h^{\mathsf{UCB},k+1}(s_h^k, a_h^k), Q_h^k(s_h^k, a_h^k)\}$;
    $V_h^{k+1}(s_h^k) \leftarrow \max_a Q_h^{k+1}(s_h^k, a)$.
    $V_h^{\mathsf{LCB},k+1}(s_h^k) \leftarrow \max\left\{\max_a Q_h^{\mathsf{LCB},k+1}(s_h^k, a), V_h^{\mathsf{LCB},k}(s_h^k)\right\}$.
      `/* update reference values */`
    if $V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) > 1$ **then**
      $V_h^{\mathsf{R},k+1}(s_h^k) \leftarrow V_h^{k+1}(s_h^k)$, $u_{\mathrm{ref}}^{k+1}(s_h^k) = \mathsf{True}$;
    else if $u_{\mathrm{ref}}^k(s_h^k) = \mathsf{True}$ **then**
      $V_h^{\mathsf{R},k+1}(s_h) \leftarrow V_h^{k+1}(s_h)$, $u_{\mathrm{ref}}^{k+1}(s_h^k) = \mathsf{False}$.
    end if
  end for
end for

which combined with line 13 of Algorithm 3 leads to monotonicity of $V_h(s)$ as follows:

$$V_h^{k+1}(s) = Q_h^{k+1}\big(s, \pi_h^{k+1}(s)\big) \le Q_h^k\big(s, \pi_h^{k+1}(s)\big) \le V_h^k(s). \tag{4.7b}$$

Moreover, by virtue of the update rule in line 12 of Algorithm 3, we can immediately obtain (via induction) the following useful property:

$$Q_h^{\mathsf{R}}, k(s,a) \ge Q_h^k(s,a) \qquad \text{for all} (k,h,s,a) \in [K] \times [H] \times \mathscr{S} \times \mathscr{A}. \tag{4.8}$$

In addition, Q_h^k and V_h^k form an 'optimistic view' of $Q_h^\star$ and $V_h^\star$, respectively, as asserted by the following lemma.

LEMMA 2. Consider any $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$,

$$Q_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{and} \qquad V_h^k(s) \geq V_h^\star(s) \tag{4.9}$$

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

Lemma 2 implies that Q_h^k (resp. V_h^k) is a pointwise upper bound on $Q_h^\star$ (resp. $V_h^\star$). Taking this result together with the non-increasing property (4.7), we see that Q_h^k (resp. V_h^k) becomes an increasingly tighter estimate of $Q_h^\star$ (resp. $V_h^\star$) as the number of episodes k increases. This important fact forms the basis of the subsequent proof, allowing us to replace $V_h^\star$ with V_h^k when upper bounding the regret. Combining Lemma 2 with (4.8), we can straightforwardly see that with probability at least $1-\delta$:

$$Q_h^{\mathrm{R}},k(s,a) \geq Q_h^\star(s,a) \qquad \text{for all}(k,h,s,a) \in [K] \times [H] \times \mathscr{S} \times \mathscr{A}. \tag{4.10}$$

Properties of the Q-estimate $Q_h^{\mathsf{LCB},k}$: pessimism and proximity. In parallel, we formalize the fact that $Q_h^{\mathsf{LCB},k}$ and $V_h^{\mathsf{LCB},k}$ provide a 'pessimistic view' of $Q_h^\star$ and $V_h^\star$, respectively. Furthermore, it becomes increasingly more likely for $Q_h^{\mathsf{LCB},k}$ and Q_h^k to stay close to each other as k increases, which indicates that the confidence interval that contains the optimal value $Q_h^\star$ becomes shorter and shorter. These properties are summarized in the following lemma.

LEMMA 3. Consider any $\delta \in (0,1)$, and suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$,

$$Q_h^{\mathsf{LCB},k}(s,a) \leq Q_h^\star(s,a) \qquad \text{and} \qquad V_h^{\mathsf{LCB},k}(s) \leq V_h^\star(s) \tag{4.11}$$

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

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > \varepsilon\right) \lesssim \frac{H^6SA\log\frac{SAT}{\delta}}{\varepsilon^2} \tag{4.12}$$

holds for all $\varepsilon \in (0,H]$.

Interestingly, the upper bound (4.12) only scales logarithmically in the number K of episodes, thus implying the closeness of $Q_h^{\mathsf{LCB},k}$ and Q_h^k for a large fraction of episodes. Note that it is straightforward to ensure the monotonicity property of $V_h^{\mathsf{LCB},k}$ from the update rule in Algorithm 3 (cf. line 14):

$$V_h^{\mathsf{LCB},k+1}(s) \geq V_h^{\mathsf{LCB},k}(s) \qquad \text{for all}(s,k,h) \in \mathscr{S} \times [K] \times [H], \tag{4.13}$$

which in conjunction with (4.11), implies that $V_h^{\mathsf{LCB},k}(s)$ gets closer to $V_h^\star(s)$ as the number of episodes k increases. Together with the monotonicity of V_h^k (cf. (4.7b)), an important consequence is that the reference value V_h^{R} will stop being updated shortly after the following condition is met for the first time

(according to lines 15–18 of Algorithm 1)

$$V_h^k(s) \le V_h^{\mathsf{LCB},k}(s) + 1 \le V_h^\star(s) + 1 \qquad \text{for all} s \in \mathscr{S}. \tag{4.14}$$

Properties of the reference V_h^{R}, k. The above fact ensures that V_h^{R}, k will not be updated too many times. In fact, its value stays reasonably close to V_h^k even after being locked to a fixed value, which ensures its fidelity as a reference signal. Moreover, the aggregate difference between V^{R}, k_h and the final reference V^{R}, k_h over the entire trajectory can be bounded in a reasonably tight fashion (owing to (4.12)), as formalized in the next lemma. These properties play a key role in reducing the burn-in cost of the proposed algorithm.

LEMMA 4. Consider any $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability exceeding $1-\delta$, one has

$$\left|V_h^k(s) - V^{\mathrm{R}}, k_h(s)\right| \le 2 \tag{4.15}$$

for all $(k,h,s) \in [K] \times [H] \times \mathscr{S}$, and

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V^{\mathrm{R}}, k_h(s_h^k) - V^{\mathrm{R}}, k_h(s_h^k)\right) \\
&\quad \le H^2 S + \sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > 1\right) \qquad (4.16)\\
&\quad \lesssim H^6 SA \log\frac{SAT}{\delta}. \qquad (4.17)
\end{aligned}$$

In words, Lemma 4 guarantees that (i) our value function estimate and the reference value are always sufficiently close (cf. (4.15)), and (ii) the aggregate difference between V_h^{R}, k and the final reference value V^{R}, k_h is nearly independent of the sample size T (except for some logarithmic scaling).

4.4 *Main steps of the proof*

We are now ready to embark on the regret analysis for Q-EarlySettled-Advantage, which consists of multiple steps as follows.

Step 1: regret decomposition. Lemma 2 allows one to upper bound the regret as follows:

$$\mathrm{Regret}(T) := \sum_{k=1}^{K}\left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \le \sum_{k=1}^{K}\left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right). \tag{4.18}$$

To continue, it boils down to controlling $V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)$. Toward this end, we intend to examine $V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)$ across all time steps $1 \le h \le H$, which admits the following decomposition:

$$\begin{aligned}
V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) &\overset{\text{(i)}}{=} Q_h^k(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) \\
&= Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) + Q_h^\star(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) \\
&\overset{\text{(ii)}}{=} Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) + P_{h,s_h^k,a_h^k}\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) \\
&\overset{\text{(iii)}}{=} Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) + \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) + V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \\
&\le Q^{\mathrm{R}}, k_h(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) + \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) + V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k).
\end{aligned} \tag{4.19}$$

Here, (i) holds since π_h^k is a greedy policy w.r.t. Q_h^k and $\pi_h^k(s_h^k) = a_h^k$, (ii) comes from the Bellman equations

$$Q_h^{\pi^k}(s,a) - Q_h^\star(s,a) = \big(r_h(s,a) + P_{h,s,a}V_{h+1}^{\pi^k}\big) - \big(r_h(s,a) + P_{h,s,a}V_{h+1}^\star\big) = P_{h,s,a}\big(V_{h+1}^{\pi^k} - V_{h+1}^\star\big),$$

(iii) follows from $P_h^k(V_{h+1}^\star - V_{h+1}^{\pi^k}) = V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)$ (see the notation (4.5)), whereas the last inequality comes from (4.8). Summing (4.19) over $1 \le k \le K$ and using Lemma 2, we obtain

$$\begin{aligned}
\sum_{k=1}^K \big(V_h^\star(s_h^k) - V_h^{\pi^k}(s_h^k)\big) &\le \sum_{k=1}^K \big(V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\big) \\
&\le \sum_{k=1}^K \big(Q^{\mathrm{R}}, k_h(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\big) + \sum_{k=1}^K \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) \\
&\quad + \sum_{k=1}^K \big(V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big).
\end{aligned} \tag{4.20}$$

This allows us to establish a connection between $\sum_k \big(V_h^\star(s_h^k) - V_h^{\pi^k}(s_h^k)\big)$ for step h and $\sum_k \big(V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$ for step $h+1$.

Step 2: managing regret by recursion. The regret can be further manipulated by leveraging the update rule of Q^{R}, k_h as well as recursing over the time steps $h = 1, 2, \cdots, H$ with the terminal condition $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$. This leads to a key decomposition as summarized in the lemma below, whose proof is provided in Appendix D.

LEMMA 5. Fix $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$, one has

$$\sum_{k=1}^{K} \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \le \mathscr{R}_1 + \mathscr{R}_2 + \mathscr{R}_3, \tag{4.21}$$

where

$$\mathscr{R}_1 := \sum_{h=1}^{H} \left(1+\frac{1}{H}\right)^{h-1} \left(HSA + 8c_{\mathrm{b}}H^2(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta} + \sum_{k=1}^{K}\left(P_{h,s_h^k,a_h^k} - P_h^k\right)\left(V_{h+1}^{\star} - V_{h+1}^{\pi^k}\right)\right), \tag{4.22a}$$

$$\mathscr{R}_2 := \sum_{h=1}^{H} \left(1+\frac{1}{H}\right)^{h-1} \sum_{k=1}^{K} B^{\mathrm{R}}, k_h(s_h^k, a_h^k), \tag{4.22b}$$

$$\begin{aligned}\mathscr{R}_3 := \sum_{h=1}^{H}\sum_{k=1}^{K} \lambda_h^k \Bigg(& (P_h^k - P_{h,s_h^k,a_h^k})(V_{h+1}^{\star} - V^{\mathrm{R}}, k_{h+1}) \\ & + \frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^{\mathrm{R},k_h^i(s_h^k,a_h^k)}(s_{h+1}^{k_h^i(s_h^k,a_h^k)}) - P_{h,s_h^k,a_h^k} V^{\mathrm{R}}, k_{h+1}\right)}{N_h^k(s_h^k,a_h^k)}\Bigg),\end{aligned} \tag{4.22c}$$

with

$$\lambda_h^k := \left(1+\frac{1}{H}\right)^{h-1} \sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \eta_{N_h^k(s_h^k,a_h^k)}^{n}.$$

This lemma attempts to upper bound the target quantity $\sum_{k=1}^{K}\left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right)$ via three terms (see (4.21)). Informally, these terms reflect (i) the influence of the initialization as well as the finite-sample uncertainty of $P_h^k(V_{h+1}^{\star} - V_{h+1}^{\pi^k})$, (ii) the influence of the size of the bonus terms and (iii) the discrepancy term when the running value iterates are replaced by the reference values. As we shall see in the analysis, the key to obtaining these terms lies in properly expanding the component $\sum_{k=1}^{K}\left(Q^{\mathrm{R}}, k_h(s_h^k,a_h^k) - Q_h^{\star}(s_h^k,a_h^k)\right)$ in (4.20), as well as applying induction across all $h = 1, \cdots, H$.

Step 3: controlling the terms in (4.22) separately. As it turns out, each of the terms in (4.22) can be well controlled. We provide the bounds for these terms in the following lemma.

LEMMA 6. Consider any $\delta \in (0,1)$. With probability at least $1-\delta$, we have the following upper bounds:

$$\mathscr{R}_1 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^{4.5}SA\log^2\frac{SAT}{\delta}\right\},$$
$$\mathscr{R}_2 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^{4}SA\log^2\frac{SAT}{\delta}\right\},$$
$$\mathscr{R}_3 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^{6}SA\log^3\frac{SAT}{\delta}\right\}$$

for some universal constant $C_{\mathrm{r}} > 0$.

In order to derive the above bounds, the main strategy is to apply the Bernstein-type concentration inequalities carefully, and to upper bound the sum of variance in a careful manner. The proofs are deferred to Appendix E.

Step 4: putting all this together. We now have everything in place to establish our main result. Taking the preceding bounds in Lemma 6 together with (4.22), we see that with probability exceeding $1-\delta$, one has

$$\mathrm{Regret}(T) \le \mathscr{R}_1 + \mathscr{R}_2 + \mathscr{R}_3 \lesssim \sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^6SA\log^3\frac{SAT}{\delta}$$

as claimed.

5. Discussion

In this paper, we have proposed a novel model-free RL algorithm—tailored to online episodic settings—that attains near-optimal regret $\widetilde{O}(\sqrt{H^2SAT})$ and near-minimal memory complexity $O(SAH)$ at once. Remarkably, the near-optimality of the algorithm comes into effect as soon as the sample size rises above $O(SA\,\mathrm{poly}(H))$, which has significantly improved upon the sample size requirements (or burn-in cost) for any prior regret-optimal model-free algorithm (based on the definition of the model-free algorithm in [29]). Given that online data collection could be expensive, time-consuming or high-stakes in a variety of contemporary applications (e.g. clinical trials, autonomous driving, online advertisement), reducing burn-in sample sizes compromising sample optimality is crucial in enabling sample-efficient solutions in these sample-constrained applications.

The results in this paper naturally suggest a number of possible extensions and directions for future investigation. We close the paper by listing a few of them.

- While the proposed algorithm provably enables minimal burn-in cost in terms of the dependency on S and A, our current theory falls short of delivering optimal horizon dependency of the burn-in cost. More specifically, even though our burn-in cost improves upon the state-of-the-art theory for sample-optimal model-free algorithms by a factor of at least $S^5A^3H^{18}$ (see [75]), the way we cope with the dependency on H remains inadequate. This calls for more refined analysis tools to optimize the horizon dependency.
- This paper focuses primarily on MDPs with non-stationary probability transition kernels. Another important scenario is concerned with MDPs with stationary transition kernels (i.e. the case where

P_h is identical across different h). It is worth noting that the algorithm developed herein is incapable of attaining optimal regret for the stationary case (i.e. the resulting regret might be off by a factor of $\sqrt{H}$). While our analysis already contains multiple key ingredients that are useful for analyzing the stationary case, how to complete the picture is non-trivial, which we leave for future work.

- Admittedly, even though we are now able to settle the sample size dependency on the state-action space, the size of SA might remain prohibitively large in many modern RL applications. As a result, parsimonious function representation/approximation of the underlying MDP is needed in order to further reduce the sample complexity. Prominent examples of this kind include linearly parameterized or realizable MDPs [18, 31, 38]. We hope that the method and analysis framework developed herein might inspire further development of sample-efficient algorithms that can effectively accommodate low-dimensional function approximation.

5. Data availability

No new data were generated or analyzed in support of this research.

Acknowledgment

L. Shi and Y. Chi are supported in part by the grants Office of Naval Research N00014-19-1-2404, National Science Foundation CCF-2106778, CCF-2007911 and DMS-2134080. 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. Y. Chen is supported in part by the the Alfred P. Sloan Research Fellowship, the Air Force Office of Scientific Research grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, IIS-2218713 and IIS-2218773, and the Google Research Scholar Award. Part of this work was done while Y. Chen and G. Li were visiting the Simons Institute for the Theory of Computing.

References

1. Agarwal, A., Kakade, S. & Yang, L. F. (2020) Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory*, **125**, 67–83.
2. Agrawal, S. & Jia, R. (2017) Posterior sampling for reinforcement learning: worst-case regret bounds. arXiv preprint arXiv:1705.07041.
3. Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. E. (2002) The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, **32**, 48–77.
4. Auer, P. & Ortner, R. (2010) UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hungar.*, **61**, 55–65.
5. Azar, M. G., Kappen, H. J., Ghavamzadeh, M. & Munos, R. (2011) Speedy Q-learning. *Advances in neural information processing systems*, pp. 2411–2419.
6. Azar, M. G., Munos, R. & Kappen, H. J. (2013) Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, **91**, 325–349.
7. Azar, M. G., Osband, I. & Munos, R. (2017) Minimax regret bounds for reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning*, vol. **70**. JMLR. org, pp. 263–272.
8. Bai, Y., Xie, T., Jiang, N. & Wang, Y.-X. (2019) Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, pp. 8002–8011.
9. Bartlett, P. & Tewari, A. (2009) Regal: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*. AUAI Press, pp. 35–42.

10. Beck, C. L. & Srikant, R. (2012) Error bounds for constant step-size Q-learning. *Systems Control Lett.*, **61**, 1203–1208.
11. Bertsekas, D. P. (2017) *Dynamic programming and optimal control (4th edition)*. Athena Scientific.
12. Chen, Z., Maguluri, S. T., Shakkottai, S. & Shanmugam, K. (2020) Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*.
13. Chen, Z., Maguluri, S. T., Shakkottai, S. & Shanmugam, K. (2021) A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*.
14. Dann, C., Lattimore, T. & Brunskill, E. (2017) Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*.
15. Domingues, O. D., Ménard, P., Kaufmann, E. & Valko, M. (2021) Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. *Algorithmic Learning Theory*. PMLR, pp. 578–598.
16. Dong, K., Wang, Y., Chen, X. & Wang, L. (2019) Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.
17. Du, S. S., Chen, J., Li, L., Xiao, L. & Zhou, D. (s2017) Stochastic variance reduction methods for policy evaluation. *Proceedings of the 34th International Conference on Machine Learning*, vol. **70**. JMLR. org, pp. 1049–1058.
18. Du, S. S., Kakade, S. M., Wang, R. & Yang, L. F. (2020) Is a good representation sufficient for sample efficient reinforcement learning? *International Conference on Learning Representations*.
19. Du, S. S., Luo, Y., Wang, R. & Zhang, H. (2019) Provably efficient Q-learning with function approximation via distribution shift error checking oracle. *Advances in Neural Information Processing Systems*, pp. 8058–8068.
20. Efroni, Y., Merlis, N., Ghavamzadeh, M. & Mannor, S. (2019) *Tight regret bounds for model-based reinforcement learning with greedy policies*. arXiv preprint arXiv:1905.11527.
21. Even-Dar, E. & Mansour, Y. (2003) Learning rates for Q-learning. *Journal of machine learning Research*, **5**, 1–25.
22. Fan, J., Wang, Z., Xie, Y. & Yang, Z. (2020) A theoretical analysis of deep Q-learning. *Learning for Dynamics and Control*. PMLR, 486–489.
23. Freedman, D. A. (1975) On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118.
24. Gower, R. M., Schmidt, M., Bach, F. & Richtárik, P. (2020) Variance-reduced methods for machine learning. *Proceedings of the IEEE*, **108**, 1968–1983.
25. He, J., Zhou, D. & Gu, Q. (2020) *Nearly minimax optimal reinforcement learning for discounted MDPs*. arXiv preprint arXiv:2010.00587.
26. Jaakkola, T., Jordan, M. I. & Singh, S. P. (1994) Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, pp. 703–710.
27. Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R. & Luo, H. (2020) *A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret*. arXiv preprint arXiv:2006.04354.
28. Jaksch, T., Ortner, R. & Auer, P. (2010) Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, **11**, 1563–1600.
29. Jin, C., Allen-Zhu, Z., Bubeck, S. & Jordan, M. I. (2018a) Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, pp. 4863–4873.
30. Jin, C., Allen-Zhu, Z., Bubeck, S. & Jordan, M. I. (2018b) *Is Q-learning provably efficient?* arXiv preprint arXiv:1807.03765.
31. Jin, C., Yang, Z., Wang, Z. & Jordan, M. I. (2020) Provably efficient reinforcement learning with linear function approximation. *Conference on Learning Theory*. PMLR, pp. 2137–2143.
32. Johnson, R. & Zhang, T. (2013) Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, pp. 315–323.
33. Kakade, S. (2003) *On the sample complexity of reinforcement learning*PhD thesis,. University of London.
34. Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J. & Jordan, M. I. (2021) Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, SIAM, **3**(4), 1013–1040.

35. Lai, T. L. & Robbins, H. (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**, 4–22.
36. Lattimore, T. & Szepesvári, C. (2020) *Bandit algorithms*. Cambridge University Press.
37. Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y. & Chi, Y. (2021a) *Is Q-learning minimax optimal? a tight sample complexity analysis*. arXiv preprint arXiv:2102.06548.
38. Li, G., Chen, Y., Chi, Y., Gu, Y. & Wei, Y. (2021b) Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *Advances in Neural Information Processing Systems*, **34**, 16671–16685.
39. Li, G., Chi, Y., Wei, Y. & Chen, Y. (2022a) *Minimax-optimal multi-agent RL in markov games with a generative model*. arXiv preprint arXiv:2208.10458.
40. Li, G., Shi, L., Chen, Y., Chi, Y. & Wei, Y. (2022b) *Settling the sample complexity of model-based offline reinforcement learning*. arXiv preprint arXiv:2204.05275.
41. Li, G., Wei, Y., Chi, Y., Gu, Y. & Chen, Y. (2020a) Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, vol. **33**.
42. Li, G., Wei, Y., Chi, Y., Gu, Y. & Chen, Y. (2020b) Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in Neural Information Processing Systems (NeurIPS)*.
43. Liu, S. & Su, H. (2020) *γ-regret for non-episodic reinforcement learning*. arXiv:2002.05138.
44. Ménard, P., Domingues, O. D., Shang, X. & Valko, M. (2021) UCB Momentum Q-learning: Correcting the bias without forgetting. *International Conference on Machine Learning*, PMLR, 7609–7618.
45. Murphy, S. (2005) A generalization error for Q-learning. *Journal of Machine Learning Research*, **6**, 1073–1097.
46. Nguyen, L. M., Liu, J., Scheinberg, K. & Takáč, M. (2017) SARAH: A novel method for machine learning problems using stochastic recursive gradient. *International Conference on Machine Learning*. PMLR, pp. 2613–2621.
47. Osband, I. & Van Roy, B. (2016) *On lower bounds for regret in reinforcement learning*. arXiv preprint arXiv:1608.02732.
48. Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K. & Roberts, S. (2020) *On optimism in model-based reinforcement learning*. arXiv preprint arXiv:2006.11911.
49. Puterman, M. L. (2014) *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
50. Qu, G. & Wierman, A. (2020) Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conference on Learning Theory*, **125**, 3185–3205.
51. Robbins, H. & Monro, S. (1951) A stochastic approximation method. *The annals of mathematical statistics*, **22**(3), 400–407.
52. Shi, L. & Chi, Y. (2022) *Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity*. arXiv preprint arXiv:2208.05767.
53. Shi, L., Li, G., Wei, Y., Chen, Y. & Chi, Y. (2022) Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *Proceedings of the 39th International Conference on Machine Learning*, PMLR, **162**, 19967–20025.
54. Sidford, A., Wang, M., Wu, X., Yang, L. & Ye, Y. (2018a) Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, pp. 5186–5196.
55. Sidford, A., Wang, M., Wu, X. & Ye, Y. (2018b) Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 770–787.
56. Strehl, A. L., Li, L., Wiewiora, E., Langford, J. & Littman, M. L. (2006) PAC model-free reinforcement learning. *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888.
57. Szepesvári, C. (1997) The asymptotic convergence-rate of Q-learning. *NIPS*, vol. **10**. Citeseer, pp. 1064–1070.

58. Tao, T. (2012). *Topics in Random Matrix Theory*. Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island.
59. Tropp, J. (2011) Freedman's inequality for matrix martingales. *Electron. Comm. Probab.*, **16**, 262–270.
60. Tsitsiklis, J. N. (1994) Asynchronous stochastic approximation and Q-learning. *Machine learning*, **16**, 185–202.
61. Wai, H.-T., Hong, M., Yang, Z., Wang, Z. & Tang, K. (2019) Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, **32**, 5784–5795.
62. Wainwright, M. J. (2019a) *Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning*. arXiv preprint arXiv:1905.06265.
63. Wainwright, M. J. (2019b) *Variance-reduced Q-learning is minimax optimal*. arXiv preprint arXiv:1906.04697.
64. Wang, B., Yan, Y. & Fan, J. (2021) Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in Neural Information Processing Systems*, **34**, 23009–23022.
65. Watkins, C. J. & Dayan, P. (1992) Q-learning. *Machine learning*, vol. **8**, pp. 279–292.
66. Watkins, C. J. C. H. (1989) Learning from delayed rewards. *PhD thesis, King's College, University of Cambridge*.
67. Weng, B., Xiong, H., Zhao, L., Liang, Y. & Zhang, W. (2020) *Momentum Q-learning with finite-sample convergence guarantee*. arXiv preprint arXiv:2007.15418.
68. Xiong, H., Zhao, L., Liang, Y. & Zhang, W. (2020) Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, **33**.
69. Xu, T., Wang, Z., Zhou, Y. & Liang, Y. (2019) Reanalysis of variance reduced temporal difference learning. *International Conference on Learning Representations*.
70. Yang, K., Yang, L. & Du, S. (2021) Q-learning with logarithmic regret. *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1576–1584.
71. Yin, M., Bai, Y. & Wang, Y.-X. (2021) Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, **34**, 7677–7688.
72. Zanette, A. & Brunskill, E. (2019) Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *International Conference on Machine Learning*. PMLR, pp. 7304–7312.
73. Zhang, K., Kakade, S., Basar, T. & Yang, L. (2020a) Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, **33**.
74. Zhang, Z., Ji, X. & Du, S. S. (2021) Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *Conference on Learning Theory*, PMLR, 4528–4531.
75. Zhang, Z., Zhou, Y. & Ji, X. (2020) Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, **33**, 15198–15207.
76. Zhang, Z., Zhou, Y. & Ji, X. (2021) Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *International Conference on Machine Learning*, PMLR, 12653–12662.

A. Freedman's inequality

A.1 *A user-friendly version of Freedman's inequality*

Due to the Markovian structure of the problem, our analysis relies heavily on the celebrated Freedman's inequality [23, 59], which extends the Bernstein's inequality to accommodate martingales. For ease of reference, we state below a user-friendly version of Freedman's inequality as provided in (37, Section C).

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

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

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

$$\big|Y_n\big| \le \sqrt{8 \max\left\{W_n, \frac{\sigma^2}{2^m}\right\} \log \frac{2m}{\delta}} + \frac{4}{3} R \log \frac{2m}{\delta}. \tag{A.1}$$

A.2 *Application of Freedman's inequality*

We now develop several immediate consequences of Freedman's inequality, which lend themseleves well to our context. Before proceeding, we recall that $N_h^i(s,a)$ denotes the number of times that the state–action pair (s,a) has been visited at step h by the end of the i-th episode, and $k_h^n(s,a)$ stands for the episode index when (s,a) is visited at step h for the n-th time (see Section 4.2).

Our first result is concerned with a martingale concentration bound as follows:

LEMMA 7. 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 the 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;
- $\|W_h^i\|_\infty \le C_{\mathrm{w}}$;
- $u_h^i(s,a,N)$ 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 $N \in [K]$;
- $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$;
- $0 \le \sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N) \le 2$.

In addition, consider the following sequence:

$$X_i(s,a,h,N) := u_h^i(s,a,N)\big(P_h^i - P_{h,s,a}\big)W_{h+1}^i \mathbb{1}\big\{(s_h^i, a_h^i) = (s,a)\big\}, \qquad 1 \le i \le K, \tag{A.2}$$

with P_h^i defined in (4.5). Consider any $\delta \in (0,1)$. Then with probability at least $1-\delta$,

$$\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N) \mathrm{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}} C_{\mathrm{w}}\right) \log^2 \frac{SAT}{\delta} \quad \text{(A.3)}$$

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

Proof. For the sake of notational convenience, we shall abbreviate $X_i(s,a,h,N)$ as X_i throughout the proof of this lemma, as long as it is clear from the context. The plan is to apply Freedman's inequality (cf. Theorem A.1) to control the term $\sum_{i=1}^{k} X_i$ of interest.

Consider any given $(k,h,s,a,N) \in [K] \times [H] \times \mathscr{S} \times \mathscr{A} \times [K]$. It can be easily verified that

$$\mathbb{E}_{i-1}\left[X_i\right] = 0,$$

where $\mathbb{E}_{i-1}$ denotes the expectation conditioned on everything happening up to the end of the $(h-1)$-th step of the i-th episode. Additionally, we make note of the following crude bound:

$$\begin{aligned} \left|X_i\right| &\le u_h^i(s,a,N)\left|\left(P_h^i - P_{h,s,a}\right) W_{h+1}^i\right| \\ &\le u_h^i(s,a,N)\left(\left\|P_h^i\right\|_1 + \left\|P_{h,s,a}\right\|_1\right)\left\|W_{h+1}^i\right\|_\infty \le 2C_{\mathrm{w}} C_{\mathrm{u}}, \end{aligned} \quad \text{(A.4)}$$

which results from the assumptions $\|W_{h+1}^i\|_\infty \le C_{\mathrm{w}}$, $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$ as well as the basic facts $\left\|P_h^i\right\|_1 = \left\|P_{h,s,a}\right\|_1 = 1$. To continue, recalling the definition of the variance parameter in (4.6), we obtain

$$\begin{aligned} \sum_{i=1}^{k} \mathbb{E}_{i-1}\left[\left|X_i\right|^2\right] &= \sum_{i=1}^{k} \big(u_h^i(s,a,N)\big)^2 \mathbb{1}\big\{(s_h^i, a_h^i) = (s,a)\big\} \mathbb{E}_{i-1}\left[\left|(P_h^i - P_{h,s,a}) W_{h+1}^i\right|^2\right] \\ &= \sum_{n=1}^{N_h^k(s,a)} \big(u_h^{k_h^n(s,a)}(s,a,N)\big)^2 \mathrm{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big) \\ &\le C_{\mathrm{u}}\left(\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\right)\left\|W_{h+1}^{k_h^n(s,a)}\right\|_\infty^2 \\ &\le 2C_{\mathrm{u}} C_{\mathrm{w}}^2, \end{aligned} \quad \text{(A.5)}$$

where the inequalities hold true due to the assumptions $\|W_h^i\|_\infty \le C_{\mathrm{w}}$, $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$, and $0 \le \sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N) \le 1$.

With (A.4) and (A.5) in place, we can invoke Theorem A.1 (with $m = \lceil \log_2 N \rceil$) and take the union bound over all $(k, h, s, a, N) \in [K] \times [H] \times \mathscr{S} \times \mathscr{A} \times [K]$ to show that, with probability at least $1 - \delta$,

$$\begin{aligned}\left|\sum_{i=1}^{k} X_i\right| &\lesssim \sqrt{\max\left\{C_{\mathrm{u}} \sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathrm{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big), \frac{C_{\mathrm{u}}C_{\mathrm{w}}^2}{N}\right\} \log\frac{SAT^2 \log N}{\delta}} \\ &\quad + C_{\mathrm{u}}C_{\mathrm{w}} \log\frac{SAT^2 \log N_h^k}{\delta} \\ &\lesssim \sqrt{C_{\mathrm{u}} \log^2\frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathrm{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta}\end{aligned}$$

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

The next result is concerned with martingale concentration bounds for another type of sequences of interest.

LEMMA 8. Let $\big\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]\big\}$ be a collection of positive integers, and let $\{c_h : 0 \le c_h \le e, h \in [H]\}$ be a collection of fixed and bounded universal constants. Moreover, let $\big\{W_h^i \in \mathbb{R}^S \mid 1 \le i \le K, 1 \le h \le H+1\big\}$ and $\big\{u_h^i(s_h^i, a_h^i) \in \mathbb{R} \mid 1 \le i \le K, 1 \le h \le H+1\big\}$ represent, respectively, the collections of random vectors and scalars, which 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;
- $\|W_h^i\|_\infty \le C_{\mathrm{w}}$ and $W_h^i \ge 0$;
- $u_h^i(s_h^i, a_h^i)$ is fully determined by the integer $N(s_h^i, a_h^i, h)$ and all samples collected up to the end of the $(h-1)$-th step of the i-th episode;
- $0 \le u_h^i(s_h^i, a_h^i) \le C_{\mathrm{u}}$.

Consider any $\delta \in (0,1)$, and introduce the following sequences:

$$X_{i,h} := u_h^i(s_h^i, a_h^i)\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i, \qquad 1 \le i \le K, 1 \le h \le H+1, \tag{A.6}$$

$$Y_{i,h} := c_h\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i, \qquad 1 \le i \le K, 1 \le h \le H+1. \tag{A.7}$$

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

$$\left|\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}\right| \lesssim \sqrt{C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i}) W_{h+1}^i\right|^2\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{T^{HSA}}{\delta}$$

$$\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{T^{HSA}}{\delta}$$

$$\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right| \lesssim \sqrt{T C_{\mathrm{w}}^2 \log\frac{1}{\delta}} + C_{\mathrm{w}} \log\frac{1}{\delta}$$

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

Proof. This lemma can be proved by Freedman's inequality (cf. Theorem A.1).

- We start by controlling the first term of interest $\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}$. As can be easily seen, $a_h^i = \arg\max Q_h^i(s_h^i, a)$ is fully determined by what happens before step h of the i-th episode. Consider any given $\left\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathscr{S}\times\mathscr{A}\times[H]\right\}$. It is readily seen that

$$\mathbb{E}_{i,h-1}\left[X_i\right] = \mathbb{E}_{i,h-1}\left[u_h^i(s_h^i,a_h^i)\left(P_h^i - P_{h,s_h^i,a_h^i}\right) W_{h+1}^i\right] = 0,$$

  where $\mathbb{E}_{i,h-1}$ denotes the expectation conditioned on everything happening before step h of the i-th episode. In addition, we make note of the following crude bound:

$$\begin{aligned}\left|X_{i,h}\right| &\le u_h^i(s_h^i,a_h^i)\left|\left(P_h^i - P_{h,s_h^i,a_h^i}\right) W_{h+1}^i\right| \\ &\le u_h^i(s_h^i,a_h^i)\left(\left\|P_h^i\right\|_1 + \left\|P_{h,s_h^i,a_h^i}\right\|_1\right)\left\|W_{h+1}^i\right\|_\infty \le 2C_{\mathrm{w}}C_{\mathrm{u}}, \end{aligned} \tag{A.8}$$

  which arises from the assumptions $\|W_{h+1}^i\|_\infty \le C_{\mathrm{w}}$, $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$ together with the basic facts $\left\|P_h^i\right\|_1 = \left\|P_{h,s_h^i,a_h^i}\right\|_1 = 1$. Additionally, we can calculate that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|X_{i,h}\right|^2\right] &= \sum_{h=1}^{H}\sum_{i=1}^{K} \left(u_h^i(s_h^i,a_h^i)\right)^2 \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i}) W_{h+1}^i\right|^2\right] \\ &\overset{\text{(i)}}{\le} C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i}) W_{h+1}^i\right|^2\right] \end{aligned} \tag{A.9}$$

$$\leq C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|P_h^i W_{h+1}^i\right|^2\right]$$

$$\overset{\text{(ii)}}{=} C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i \left(W_{h+1}^i\right)^2\right]$$

$$\overset{\text{(iii)}}{\leq} C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \left\|W_{h+1}^i\right\|_\infty \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right]$$

$$\overset{\text{(iv)}}{\leq} C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \tag{A.10}$$

$$\leq C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \left\|W_{h+1}^i\right\|_\infty \overset{\text{(v)}}{\leq} HKC_{\mathrm{u}}^2 C_{\mathrm{w}}^2 = TC_{\mathrm{u}}^2 C_{\mathrm{w}}^2. \tag{A.11}$$

Here, (i) holds true due to the assumption $0 \leq u_h^i(s_h^i, a_h^i) \leq C_{\mathrm{u}}$, (ii) is valid since P_h^i only has one non-zero entry (cf. (4.5)), (iii) relies on the assumptions that W_h^i is non-negative, whereas (iv) and (v) follow since $\|W_h^i\|_\infty \leq C_{\mathrm{w}}$,

- With (A.8), (A.10) and (A.11) in mind, we can invoke Theorem A.1 (with $m = \lceil \log_2 T \rceil$) and take the union bound over all possible collections $\left\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]\right\}$—which has at most K^{HSA} possibilities—to show that, with probability at least $1-\delta$,

$$\begin{aligned}
\left|\sum_{h=1}^{H}\sum_{i=1}^{k} X_{i,h}\right| &\lesssim \sqrt{\max\left\{C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i}) W_{h+1}^i\right|^2\right], \frac{TC_{\mathrm{u}}^2 C_{\mathrm{w}}^2}{2^m}\right\} \log\frac{K^{HSA}\log T}{\delta}} \\
&\quad + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{K^{HSA}\log T}{\delta} \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i}) W_{h+1}^i\right|^2\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{T^{HSA}}{\delta} \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{T^{HSA}}{\delta}
\end{aligned}$$

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

- Then we turn to control the second term $\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right|$ of interest. Similar to $\left|\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}\right|$, we have

$$|Y_{i,h}| \leq 2eC_{\mathrm{w}},$$

$$\sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|Y_{i,h}\right|^2\right] \leq e^2 TC_{\mathrm{w}}^2.$$

Invoke Theorem A.1 (with $m = 1$) to arrive at

$$\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right| \lesssim \sqrt{TC_{\mathrm{w}}^2 \log\frac{1}{\delta}} + C_{\mathrm{w}} \log\frac{1}{\delta} \tag{A.12}$$

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

B. Proof of Lemma 1

First of all, the properties in (4.4b) follow directly from (29, Lemma 4.1). Therefore, it suffices to establish the property in (4.4a), which forms the remainder of this subsection.

When $N = 1$, the statement holds trivially since

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \eta_1^1 = 1 \in [1, 2].$$

Now suppose that $N \geq 2$. Using the basic relation $\eta_n^N = (1-\eta_N)\eta_n^{N-1}$ for all $n = 1, \cdots, N-1$, we observe the following identity:

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N)\sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a}. \tag{B.1}$$

We now prove the property in (4.4a) by induction. Suppose for the moment that the property holds for $N-1$, namely,

$$\frac{1}{(N-1)^a} \leq \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} \leq \frac{2}{(N-1)^a}. \tag{B.2}$$

Then it is readily seen from (B.1) that

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N)\sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} \geq \frac{\eta_N}{N^a} + \frac{1-\eta_N}{(N-1)^a} \geq \frac{\eta_N}{N^a} + \frac{1-\eta_N}{N^a} = \frac{1}{N^a}, \tag{B.3}$$

where the first inequality comes from (B.2). Similarly, one can upper bound

$$\begin{aligned}\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N)\sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} &\overset{\text{(i)}}{\leq} \frac{\eta_N}{N^a} + \frac{2(1-\eta_N)}{(N-1)^a} \overset{\text{(ii)}}{=} \frac{H+1}{N^a(H+N)} + \frac{2(N-1)^{1-a}}{H+N} \\ &\overset{\text{(iii)}}{\leq} \frac{H+1}{N^a(H+N)} + \frac{2N^{1-a}}{H+N} = \frac{1}{N^a}\left(\frac{H+1}{H+N} + \frac{2N}{H+N}\right) \overset{\text{(iv)}}{\leq} \frac{2}{N^a},\end{aligned}$$

where (i) arises from (B.2), (ii) follows from the choice $\eta_N = \frac{H+1}{H+N}$, (iii) holds since $a \leq 1$ and (iv) follows since $H \geq 1$. Consequently, we can immediately establish the advertised property (4.4a) by induction.

C. Proof of key lemmas in Section 4.3

C.1 *Proof of Lemma 2*

To begin with, suppose that we can prove

$$Q_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{for all}(k,h,s,a) \in [K]\times[H]\times\mathscr{S}\times\mathscr{A}. \tag{C.1}$$

Then this property would immediately lead to the claim w.r.t. V_h^k, namely,

$$V_h^k(s) \geq Q_h^k\big(s,\pi_h^\star(s)\big) \geq Q_h^\star\big(s,\pi_h^\star(s)\big) = V_h^\star(s) \qquad \text{for all}(k,h,s) \in [K]\times[H]\times\mathscr{S}. \tag{C.2}$$

As a result, it suffices to focus on justifying the claim (C.1), which we shall accomplish by induction.

- *Base case.* Given that the initialization obeys $Q_h^1(s,a) = H \geq Q_h^\star(s,a)$ for all $(h,s,a) \in [H]\times\mathscr{S}\times\mathscr{A}$, the claim (C.1) holds trivially when $k = 1$.
- *Induction.* Suppose that the claim (C.1) holds all the way up to the k-th episode, and we wish to establish it for the $(k+1)$-th episode as well. To complete the induction argument, it suffices to justify

$$\min\left\{Q_h^{\mathsf{UCB},k+1}(s,a), Q_h^{\mathrm{R},k+1}(s,a)\right\} \geq Q_h^\star(s,a)$$

according to line 12 of Algorithm 3. Recognizing that $Q_h^{\mathsf{UCB},k+1}$ is computed via the standard UCB-Q update rule (see line 2 of Algorithm 2), we can readily invoke the argument in (29, Lemma 4.3) to show that with probability at least $1-\delta$,

$$Q_h^{\mathsf{UCB},k+1}(s,a) \geq Q_h^\star(s,a)$$

holds simultaneously for all $(k,h,s,a) \in [K]\times[H]\times\mathscr{S}\times\mathscr{A}$. Therefore, it is sufficient to prove that

$$Q_h^{\mathrm{R},k+1}(s,a) \geq Q_h^\star(s,a). \tag{C.3}$$

The remainder of the proof is thus devoted to justifying (C.3), assuming that the claim (C.1) holds all the way up to k.

Since $Q_h^{\mathrm{R}}, k(s_h^k, a_h^k)$ is updated in the k-th episode while other entries of Q_h^{R}, k remain fixed, it suffices to verify

$$Q_h^{\mathrm{R},k+1}(s_h^k, a_h^k) \geq Q_h^\star(s_h^k, a_h^k).$$

We remind the readers of two important short-hand notation that shall be used when it is clear from the context:

- $N_h^k = N_h^k(s_h^k, a_h^k)$ denotes the number of times that the state–action pair (s_h^k, a_h^k) has been visited at step h by the end of the k-th episode;
- $k^n = k_h^n(s_h^k, a_h^k)$ denotes the index of the episode in which the state–action pair (s_h^k, a_h^k) is visited for the n-th time at step h.

Step 1: decomposing $Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)$.
To begin with, the above definition of N_h^k and k^n allows us to write

$$Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)=Q_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k), \tag{C.4}$$

since $k^{N_h^k}=k^{N_h^k(s_h^k,a_h^k)}=k$. According to the update rule (i.e. line 11 in Algorithm 3 and line 9 in Algorithm 2), we obtain

$$\begin{aligned}
Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)&=Q_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k)=(1-\eta_{N_h^k})Q_h^{\mathrm{R},k^{N_h^k}}(s_h^k,a_h^k)\\
&\quad+\eta_{N_h^k}\Big\{r_h(s_h^k,a_h^k)+V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})-V_{h+1}^{\mathrm{R},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})+\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k)+b_h^{\mathrm{R},k^{N_h^k}+1}\Big\}\\
&=(1-\eta_{N_h^k})Q_h^{\mathrm{R},k^{N_h^k-1}+1}(s_h^k,a_h^k)\\
&\quad+\eta_{N_h^k}\Big\{r_h(s_h^k,a_h^k)+V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})-V_{h+1}^{\mathrm{R},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})+\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k)+b_h^{\mathrm{R},k^{N_h^k}+1}\Big\},
\end{aligned}$$

where the last identity again follows from our argument for justifying (C.4). Applying this relation recursively and invoking the definitions of η_0^N and η_n^N in (4.2), we are left with

$$\begin{aligned}
Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)&=\eta_0^{N_h^k}Q_h^{\mathrm{R},1}(s_h^k,a_h^k)\\
&\quad+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big\{r_h(s_h^k,a_h^k)+V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+\mu_h^{\mathrm{ref},k^n+1}(s_h^k,a_h^k)+b_h^{\mathrm{R},k^n+1}\Big\}.
\end{aligned} \tag{C.5}$$

Additionally, the basic relation $\eta_0^{N_h^k}+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}=1$ (see (4.2) and (4.3)) tells us that

$$Q_h^\star(s_h^k,a_h^k)=\eta_0^{N_h^k}Q_h^\star(s_h^k,a_h^k)+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}Q_h^\star(s_h^k,a_h^k), \tag{C.6}$$

which combined with (C.5) leads to

$$\begin{aligned}
Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)&=\eta_0^{N_h^k}\big(Q_h^{\mathrm{R},1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\big)\\
&\quad+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big\{r_h(s_h^k,a_h^k)+V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+\mu_h^{\mathrm{ref},k^n+1}(s_h^k,a_h^k)+b_h^{\mathrm{R},k^n+1}-Q_h^\star(s_h^k,a_h^k)\Big\}.
\end{aligned} \tag{C.7}$$

To continue, invoking the Bellman optimality equation

$$Q_h^\star(s_h^k,a_h^k)=r_h(s_h^k,a_h^k)+P_{h,s_h^k,a_h^k}V_{h+1}^\star \tag{C.8}$$

and using the construction of μ_h^{ref} in line 11 of Algorithm 2 (which is the running mean of V_{h+1}^{R}), we reach

$$r_h(s_h^k,a_h^k)+V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+\mu_h^{\mathrm{ref},k^n+1}(s_h^k,a_h^k)+b_h^{\mathrm{R},k^n+1}-Q_h^\star(s_h^k,a_h^k)$$
$$=V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+\frac{\sum_{i=1}^n V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})}{n}-P_{h,s_h^k,a_h^k}V_{h+1}^\star+b_h^{\mathrm{R},k^n+1} \tag{C.9}$$

$$=P_{h,s_h^k,a_h^k}\left\{V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\right\}+\frac{\sum_{i=1}^n P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R},k^i}\right)}{n}-P_{h,s_h^k,a_h^k}V_{h+1}^\star+b_h^{\mathrm{R},k^n+1}+\xi_h^{k^n},$$
$$=P_{h,s_h^k,a_h^k}\left\{V_{h+1}^{k^n}-V_{h+1}^\star+\frac{\sum_{i=1}^n\left(V_{h+1}^{\mathrm{R},k^i}-V_{h+1}^{\mathrm{R},k^n}\right)}{n}\right\}+b_h^{\mathrm{R},k^n+1}+\xi_h^{k^n}. \tag{C.10}$$

Here, we have introduced the following quantity:

$$\xi_h^{k^n}:=\left(P_h^{k^n}-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\right)+\frac{1}{n}\sum_{i=1}^n\left(P_h^{k^i}-P_{h,s_h^k,a_h^k}\right)V_{h+1}^{\mathrm{R},k^i}, \tag{C.11}$$

with the notation P_h^k defined in (4.5). Putting (C.10) and (C.7) together leads to the following decomposition:

$$Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)=\eta_0^{N_h^k}\left(Q_h^{\mathrm{R},1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\right)$$
$$+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left\{P_{h,s_h^k,a_h^k}\left(V_{h+1}^{k^n}-V_{h+1}^\star+\frac{\sum_{i=1}^n\left(V_{h+1}^{\mathrm{R},k^i}-V_{h+1}^{\mathrm{R},k^n}\right)}{n}\right)+b_h^{\mathrm{R},k^n+1}+\xi_h^{k^n}\right\}. \tag{C.12}$$

Step 2: two key quantities for lower bounding $Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)$.

In order to develop a lower bound on $Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)$ based on the decomposition (C.12), we make note of several simple facts as follows:

(i) The initialization satisfies $Q_h^{\mathrm{R},1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\geq 0$.

(ii) For any $1\leq k^n\leq k$, one has

$$V_{h+1}^{k^n}\geq V_{h+1}^\star, \tag{C.13}$$

owing to the induction hypotheses (C.1) and (C.2) that hold up to k.

(iii) For all $0\leq i\leq n$ and any $s\in\mathcal{S}$, one has

$$V_{h+1}^{\mathrm{R},k^i}(s)-V_{h+1}^{\mathrm{R},k^n}(s)\geq 0, \tag{C.14}$$

which holds since the reference value $V_h^{\mathrm{R}}(s)$ is monotonically non-increasing in view of the monotonicity of $V_h(s)$ in (4.7b) and the update rule in line 16 of Algorithm 3.

The above three facts taken collectively with (C.12) allow one to drop several terms and yield

$$Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\geq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(b_h^{\mathrm{R},k^n+1}+\xi_h^{k^n}\big). \tag{C.15}$$

In the sequel, we aim to establish $Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)\geq Q_h^\star(s_h^k,a_h^k)$ based on this inequality (C.15).

As it turns out, if one could show that

$$\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\xi_h^{k^n}\right|\leq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathrm{R},k^n+1}, \tag{C.16}$$

then taking this together with (C.15) and the triangle inequality would immediately lead to the desired result

$$Q_h^{\mathrm{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\geq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathrm{R},k^n+1}-\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\xi_h^{k^n}\right|\geq 0. \tag{C.17}$$

As a result, the remaining steps come down to justifying the claim (C.16). In order to do so, we need to control the following two quantities (in view of (C.11)):

$$I_1:=\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big), \tag{C.18a}$$

$$I_2:=\sum_{n=1}^{N_h^k}\frac{1}{n}\eta_n^{N_h^k}\sum_{i=1}^{n}\big(P_h^{k^i}-P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathrm{R},k^i} \tag{C.18b}$$

separately, which constitutes the next two steps. As will be seen momentarily, these two terms can be controlled in a similar fashion using Freedman's inequality.

Step 3: controlling I_1. In the following text, we intend to invoke Lemma 7 to control the term I_1 defined in (C18a). To begin with, consider any $(N,h)\in[K]\times[H]$, and introduce

$$W_{h+1}^i:=V_{h+1}^i-V_{h+1}^{\mathrm{R},i}\qquad\text{and}\qquad u_h^i(s,a,N):=\eta_{N_h^i(s,a)}^N\geq 0. \tag{C.19}$$

Accordingly, we can derive and define

$$\|W_{h+1}^i\|_\infty\leq\|V_{h+1}^{\mathrm{R},i}\|_\infty+\|V_{h+1}^i\|_\infty\leq 2H=:C_{\mathrm{w}}, \tag{C.20}$$

and

$$\max_{N,h,s,a\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}}\eta_{N_h^i(s,a)}^N\leq\frac{2H}{N}=:C_{\mathrm{u}}, \tag{C.21}$$

where the last inequality follows since (according to Lemma 1 and the definition in (4.2))

$$\eta_{N_h^i(s,a)}^N\leq\frac{2H}{N},\qquad \text{if}1\leq N_h^i(s,a)\leq N;$$
$$\eta_{N_h^i(s,a)}^N=0,\qquad \text{if}N_h^i(s,a)>N.$$

Moreover, observed from (4.3), we have

$$0 \le \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a,N) = \sum_{n=1}^{N} \eta_n^N \le 1 \tag{C.22}$$

holds for all $(N,s,a) \in [K]\times\mathscr{S}\times\mathscr{A}$. Therefore, choosing $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ and applying Lemma 7 with the quantities (C.19) implies that, with probability at least $1-\delta$,

$$\begin{aligned}
|I_1| &= \left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big)\right| = \left|\sum_{i=1}^{k} X_i(s_h^k, a_h^k, h, N_h^k)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s_h^k, a_h^k, N_h^k)\mathrm{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\asymp \sqrt{\frac{H}{N_h^k}\log^2\frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big) + \frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k}} \qquad \text{(C.23)} \\
&\lesssim \sqrt{\frac{H}{N_h^k}\log^2\frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2} + \frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}}, \qquad \text{(C.24)}
\end{aligned}$$

where the proof of the last inequality (C.24) needs additional explanation and is postponed to Appendix C.1.1 to streamline the presentation.

Step 4: controlling I_2. Next, we turn attention to the quantity I_2 defined in (C18b). Rearranging terms in the definition (C18b), we are left with

$$I_2 = \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \frac{\sum_{i=1}^{n}\big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathrm{R},k^i}}{n} = \sum_{i=1}^{N_h^k}\left(\sum_{n=i}^{N_h^k}\frac{\eta_n^{N_h^k}}{n}\right)\big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathrm{R},k^i},$$

which can again be controlled by invoking Lemma 7. To do so, we abuse the notation by taking

$$W_{h+1}^i := V_{h+1}^{\mathrm{R},i} \qquad \text{and} \qquad u_h^i(s,a,N) := \sum_{n=N_h^i(s,a)}^{N} \frac{\eta_n^N}{n} \ge 0. \tag{C.25}$$

These quantities satisfy

$$\big\|W_{h+1}^i\big\|_\infty \le \big\|V_{h+1}^{\mathrm{R},i}\big\|_\infty \le H =: C_{\mathrm{w}} \tag{C.26}$$

and, according to Lemma 1,

$$\max_{N,h,s,a\in[K]\times[H]\times\mathscr{S}\times\mathscr{A}} \sum_{n=N_h^i(s,a)}^{N} \frac{\eta_n^N}{n} \le \sum_{n=1}^{N}\frac{\eta_n^N}{n} \le \frac{2}{N} =: C_{\mathrm{u}}. \tag{C.27}$$

Then it is readily seen from (C.27) that

$$0 \le \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a,N) \le \sum_{n=1}^{N} \frac{2}{N} \le 2 \tag{C.28}$$

holds for all $(N,s,a) \in [K] \times \mathscr{S} \times \mathscr{A}$.

With the above relations in mind, Taking $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ and applying Lemma 7 w.r.t. the quantities (C.25) reveals that

$$|I_2| = \left| \sum_{i=1}^{N_h^k} \sum_{n=i}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} \left(P_h^{k^i} - P_{h,s_h^k,a_h^k}\right) V_{h+1}^{\mathrm{R},k^i} \right| = \left| \sum_{i=1}^{k} X_i(s_h^k, a_h^k, h, N_h^k) \right| \tag{C.29}$$

$$\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s_h^k, a_h^k, N_h^k) \mathrm{Var}_{h,s_h^k,a_h^k}\left(W_{h+1}^{k^n}\right)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}} C_{\mathrm{w}} \right) \log^2 \frac{SAT}{\delta}$$

$$\lesssim \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\frac{1}{N_h^k} \sum_{n=1}^{N_h^k} \mathrm{Var}_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R},k^n}\right)} + \frac{H}{N_h^k} \log^2 \frac{SAT}{\delta} \tag{C.30}$$

$$\lesssim \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k, a_h^k) - \left(\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k, a_h^k)\right)^2} + \frac{H}{(N_h^k)^{3/4}} \log^2 \frac{SAT}{\delta} \tag{C.31}$$

with probability exceeding $1-\delta$, where the proof of the last inequality (C.31) is deferred to Appendix C.1.2 in order to streamline presentation.

Step 5: combining the above bounds. Summing up the results in (C.24) and (C.31), we arrive at an upper bound on $\left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n} \right|$ as follows:

$$\begin{aligned}
\left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n} \right| &\le |I_1| + |I_2| \\
&\lesssim \sqrt{\frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k, a_h^k) - \left(\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k, a_h^k)\right)^2} \\
&\quad + \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k, a_h^k) - \left(\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k, a_h^k)\right)^2} + \frac{H^2 \log^2 \frac{SAT}{\delta}}{(N_h^k)^{3/4}} \\
&\le B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k, a_h^k) + c_{\mathrm{b}} \frac{H^2 \log^2 \frac{SAT}{\delta}}{(N_h^k)^{3/4}}
\end{aligned} \tag{C.32}$$

for some sufficiently large constant $c_{\mathrm{b}} > 0$, where the last line follows from the definition of $B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k, a_h^k)$ in line 17 of Algorithm 2.

In order to establish the desired bound (C.16), we still need to control the sum $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathrm{R},k^n+1}$. Toward this end, the definition of b_h^{R,k^n+1} (resp. δ_h^{R}) in line 8 (resp. line 18) of Algorithm 2 yields

$$b_h^{\mathrm{R},k^n+1} = \Big(1-\frac{1}{\eta_n}\Big)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k) + \frac{1}{\eta_n}B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k) + \frac{c_{\mathrm{b}}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta}. \tag{C.33}$$

This taken collectively with the definition (4.2) of η_n^N allows us to expand

$$\begin{aligned}
&\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathrm{R},k^n+1} \\
&= \sum_{n=1}^{N_h^k} \eta_n \prod_{i=n+1}^{N_h^k}(1-\eta_i)\left(\Big(1-\frac{1}{\eta_n}\Big)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k) + \frac{1}{\eta_n}B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k)\right) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta} \\
&= \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k}(1-\eta_i)\left(-\big(1-\eta_n\big)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k) + B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k)\right) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta} \\
&= \sum_{n=1}^{N_h^k}\left(\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k) - \prod_{i=n}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k)\right) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta} \\
&\overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k}\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k) - \sum_{n=2}^{N_h^k}\prod_{i=n}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta} \\
&\overset{\text{(ii)}}{=} \sum_{n=1}^{N_h^k}\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k) - \sum_{n=1}^{N_h^k-1}\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta} \\
&= B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k) + c_{\mathrm{b}}\sum_{n=1}^{N_h^k}\frac{\eta_n^{N_h^k}}{n^{3/4}}H^2\log^2\frac{SAT}{\delta}.
\end{aligned} \tag{C.34}$$

Here, (i) is valid due to the fact that $B_h^{\mathrm{R},k^1}(s_h^k,a_h^k)=0$; (ii) follows from the fact that

$$\begin{aligned}
\sum_{n=2}^{N_h^k}\prod_{i=n}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n}(s_h^k,a_h^k) &= \sum_{n=1}^{N_h^k-1}\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^{n+1}}(s_h^k,a_h^k) \\
&= \sum_{n=1}^{N_h^k-1}\prod_{i=n+1}^{N_h^k}(1-\eta_i)B_h^{\mathrm{R},k^n+1}(s_h^k,a_h^k),
\end{aligned}$$

where the first relation can be seen by replacing n with $n+1$, and the last relation holds true since the state–action pair (s_h^k,a_h^k) has not been visited at step h between the (k^n+1)-th episode and the $(k^{n+1}-1)$-th episode. Combining the above identity (C.34) with the following property (see Lemma 1)

$$\frac{1}{(N_h^k)^{3/4}} \le \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} \le \frac{2}{(N_h^k)^{3/4}},$$

we can immediately demonstrate that

$$B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k) + c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}} \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathrm{R},k^n+1} \le B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k) + 2c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}}. \tag{C.35}$$

Taking (C.32) and (C.35) collectively demonstrates that

$$\left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\xi_h^{k^n}\right| \le B_h^{\mathrm{R},k^{N_h^k}+1}(s_h^k,a_h^k) + c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}} \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathrm{R},k^n+1} \tag{C.36}$$

as claimed in (C.16). We have thus concluded the proof of Lemma 2 based on the argument in Step 2.

C.1.1 *Proof of the inequality (C24).* In order to establish the inequality (C.24), it suffices to look at the following term:

$$I_3 := \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big) - \sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k) + \big(\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k)\big)^2, \tag{C.37}$$

which forms the main content of this subsection.

First of all, the update rules of $\mu_h^{\mathrm{adv},k^{n+1}}$ and $\sigma_h^{\mathrm{adv},k^{n+1}}$ in lines 13-14 of Algorithm 2 tell us that

$$\mu_h^{\mathrm{adv},k^{n+1}}(s_h^k,a_h^k) = \mu_h^{\mathrm{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\mu_h^{\mathrm{adv},k^n}(s_h^k,a_h^k) + \eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big),$$
$$\sigma_h^{\mathrm{adv},k^{n+1}}(s_h^k,a_h^k) = \sigma_h^{\mathrm{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\sigma_h^{\mathrm{adv},k^n}(s_h^k,a_h^k) + \eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2.$$

Applying this relation recursively and invoking the definitions of η_n^N (resp. P_h^k) in (4.2) (resp. (4.5)) give

$$\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k) \overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big) = \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big), \tag{C.38a}$$

$$\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k) \overset{\text{(ii)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2 = \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big)^2. \tag{C.38b}$$

Recognizing that $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.3)), we can immediately apply Jensen's inequality to the expressions (i) and (ii) to yield

$$\sigma_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k) \ge \Big(\mu_h^{\mathrm{adv},k^{N_h^k}+1}(s_h^k,a_h^k)\Big)^2. \tag{C.39}$$

Further, in view of the definition (4.6), we have

$$\mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)=P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2-\Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2,$$

which allows one to decompose and bound I_3 as follows:

$$\begin{aligned}
I_3 &= \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2-\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\\
&\qquad+\bigg(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\bigg)^2-\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2\\
&\le \underbrace{\bigg|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\bigg|}_{=:I_{3,1}}\\
&\qquad+\underbrace{\bigg(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\bigg)^2-\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2}_{=:I_{3,2}}. \qquad \text{(C.40)}
\end{aligned}$$

It then boils down to controlling the above two terms in (C.40) separately.

Step 1: bounding $I_{3,1}$. To upper bound the term $I_{3,1}$ in (C.40), we resort to Lemma 7 by setting

$$W_{h+1}^i:=\big(V_{h+1}^i-V_{h+1}^{\mathrm{R},i}\big)^2 \qquad \text{and} \qquad u_h^i(s,a,N):=\eta^N_{N_h^i(s,a)}. \qquad \text{(C.41)}$$

It is easily seen that

$$\|W_{h+1}^i\|_\infty\le\Big(\big\|V_{h+1}^{\mathrm{R},i}\big\|_\infty+\big\|V_{h+1}^i\big\|_\infty\Big)^2\le 4H^2=:C_{\mathrm{w}}, \qquad \text{(C.42)}$$

and it follows from (C.21) that

$$\max_{N,h,s,a\in[K]\times[H]\times\mathscr{S}\times\mathscr{A}}\eta^N_{N_h^i(s,a)}\le\frac{2H}{N}=:C_{\mathrm{u}}. \qquad \text{(C.43)}$$

Armed with the properties (C.42) and (C.43) and recalling (C.28), we can invoke Lemma 7 w.r.t. (C.41) and set $(N,s,a)=(N_h^k,s_h^k,a_h^k)$ to yield

$$\begin{aligned}
I_{3,1} &= \bigg|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\bigg|=\bigg|\sum_{i=1}^{k}X_i(s_h^k,a_h^k,h,N_h^k)\bigg|\\
&\lesssim\sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}u_h^{k^n}(s_h^k,a_h^k,N_h^k)\mathrm{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)}+\bigg(C_{\mathrm{u}}C_{\mathrm{w}}+\sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\bigg)\log^2\frac{SAT}{\delta}
\end{aligned}$$

$$\lesssim \sqrt{\frac{H}{N_h^k}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\Big)+\frac{H^3\log^2\frac{SAT}{\delta}}{N_h^k}}$$

$$\lesssim \sqrt{\frac{H^5}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta} \tag{C.44}$$

with probability at least $1-\delta$. Here, the last inequality results from the fact $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\le 1$ (see (4.3)) and the following trivial result:

$$\mathrm{Var}_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\Big)\le\big\|\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^4\big\|_\infty\le 16H^4. \tag{C.45}$$

Step 2: bounding $I_{3,2}$. Jensen's inequality tells us that

$$\begin{aligned}\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2&=\Big(\sum_{n=1}^{N_h^k}\big(\eta_n^{N_h^k}\big)^{1/2}\cdot\big(\eta_n^{N_h^k}\big)^{1/2}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2\\&\le\Big\{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big\}\Big\{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2\Big\}\\&\le\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2,\end{aligned}$$

where the last line arises from (4.3). Substitution into $I_{3,2}$ (cf. (C.40)) gives

$$\begin{aligned}I_{3,2}&\le\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2-\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big)^2\\&=\Big\{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big\}\Big\{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}+P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big\}.\end{aligned} \tag{C.46}$$

In what follows, we would like to use this relation to show that

$$I_{3,2}\le C_{32}\Big\{\sqrt{\frac{H^5}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}\Big\} \tag{C.47}$$

for some universal constant $C_{32}>0$.

If $I_{3,2}\le 0$, then (C.47) holds true trivially. Consequently, it is sufficient to study the case where $I_{3,2}>0$. To this end, we first note that the term in the first pair of curly brakets of (C.46) is exactly I_1

(see (C18a)), which can be bounded by recalling (C.23):

$$
\begin{aligned}
|I_1| &\lesssim \sqrt{\frac{H}{N_h^k}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)}+\frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k}\\
&\lesssim \sqrt{\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}}+\frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k}\\
&\lesssim \sqrt{\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta},
\end{aligned}
\tag{C.48}
$$

with probability at least $1-\delta$. Here, the second inequality arises from the following property:

$$
\mathrm{Var}_{h,s_h^k,a_h^k}\Big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\Big)\le\big\|\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)^2\big\|_\infty\le 4H^2, \tag{C.49}
$$

whereas the last inequality (C.48) holds as a result of the fact $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\le 1$ (see (4.3)).

Moreover, the term in the second pair of curly brakets of (C.46) can be bounded straightforwardly as follows:

$$
\begin{aligned}
&\Big|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}+P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\Big|\\
&\qquad\le\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(\big\|P_h^{k^n}\big\|_1+\big\|P_{h,s_h^k,a_h^k}\big\|_1\Big)\big\|V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big\|_\infty\le 2H,
\end{aligned}
\tag{C.50}
$$

where we have used the property (4.3), as well as the elementary facts $\big\|V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big\|_\infty\le H$ and $\big\|P_h^{k^n}\big\|_1=\big\|P_{h,s_h^k,a_h^k}\big\|_1=1$. Substituting the above two results (C.48) and (C.50) back into (C.46), we arrive at the bound (C.47) as long as $I_{3,2}>0$. Putting all cases together, we have established the claim (C.47).

Step 3: putting all this together. To finish up, plugging the bounds (C.44) and (C.47) into (C.40), we can conclude that

$$
I_3\le I_{3,1}+I_{3,2}\le C_3\left\{\sqrt{\frac{H^5}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}\right\}
$$

for some constant $C_3>0$. This together with the definition (C.37) of I_3 results in

$$
\begin{aligned}
&\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathrm{R},k^n}\big)\\
&\qquad\le\Big\{\sigma_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2\Big\}+C_3\left(\sqrt{\frac{H^5}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}\right),
\end{aligned}
$$

which combined with the elementary inequality $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for any $u, v \ge 0$ and (C.39) yields

$$\begin{aligned}&\left\{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathrm{R},k^n}\big)\right\}^{1/2} \\ &\quad \lesssim \left\{\sigma_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathrm{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2\right\}^{1/2} + \frac{H^{5/4}}{\big(N_h^k\big)^{1/4}} \log^{1/2}\frac{SAT}{\delta} + \frac{H^{3/2}}{\big(N_h^k\big)^{1/2}} \log\frac{SAT}{\delta}.\end{aligned}$$

Substitution into (C.23) establishes the desired result (C.24).

C.1.2 *Proof of the inequality (C31)* In order to prove the inequality (C.31), it suffices to look at the following term:

$$I_4 := \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} \mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathrm{R},k^n}) - \Big(\sigma_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2\Big). \tag{C.51}$$

In view of the update rules of $\mu_h^{\mathrm{ref},k^{n+1}}$ and $\sigma_h^{\mathrm{ref},k^{n+1}}$ in lines 11–12 of Algorithm 2, we have

$$\begin{aligned}\mu_h^{\mathrm{ref},k^{n+1}}(s_h^k,a_h^k) &= \mu_h^{\mathrm{ref},k^n+1}(s_h^k,a_h^k) = \left(1-\frac{1}{n}\right)\mu_h^{\mathrm{ref},k^n}(s_h^k,a_h^k) + \frac{1}{n}V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}), \\ \sigma_h^{\mathrm{ref},k^{n+1}}(s_h^k,a_h^k) &= \sigma_h^{\mathrm{ref},k^n+1}(s_h^k,a_h^k) = \left(1-\frac{1}{n}\right)\sigma_h^{\mathrm{ref},k^n}(s_h^k,a_h^k) + \frac{1}{n}\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2,\end{aligned}$$

Through simple recursion, these identities together with the definition (4.5) of P_h^k lead to

$$\mu_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(i)}}{=} \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n} V_{h+1}^{\mathrm{R},k^n}, \tag{C.52a}$$

$$\sigma_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(ii)}}{=} \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} \big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2 = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\big(V_{h+1}^{\mathrm{R},k^n}\big)^2, \tag{C.52b}$$

The expressions (i) and (ii) combined with Jensen's inequality give

$$\sigma_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \ge \Big(\mu_h^{\mathrm{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\Big)^2. \tag{C.53}$$

Taking these together with the definition

$$\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathrm{R},k^n}) = P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathrm{R},k^n}\big)^2 - \big(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathrm{R},k^n}\big)^2,$$

we obtain

$$I_4 = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s_h^k,a_h^k}(V_{h+1}^{\mathrm{R},k^n})^2 - \left(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathrm{R},k^n}\right)^2\right) - \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_h^{k^n}\left(V_{h+1}^{\mathrm{R},k^n}\right)^2 + \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_h^{k^n}V_{h+1}^{\mathrm{R},k^n}\right)^2$$
$$= \underbrace{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s_h^k,a_h^k} - P_h^{k^n}\right)\left(V_{h+1}^{\mathrm{R},k^n}\right)^2}_{=:I_{4,1}} + \underbrace{\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_h^{k^n}V_{h+1}^{\mathrm{R},k^n}\right)^2 - \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathrm{R},k^n}\right)^2}_{=:I_{4,2}}. \tag{C.54}$$

In what follows, we shall bound the terms $I_{4,1}$ and $I_{4,2}$ in (C.54) separately.

Step 1: bounding $I_{4,1}$. The first term $I_{4,1}$ in (C.54) can be bounded by means of Lemma 7 in an almost identical fashion as $I_{3,1}$ in (C.44). Specifically, let us set

$$W_{h+1}^i := (V_{h+1}^{\mathrm{R},i})^2 \qquad \text{and} \qquad u_h^i(s,a,N) := \frac{1}{N},$$

which clearly obey

$$|u_h^i(s,a,N)| = \frac{1}{N} =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H^2 =: C_{\mathrm{w}}.$$

It is easily verified that

$$\sum_{n=1}^{N} u_h^{k^n(s,a)}(s,a,N) = \sum_{n=1}^{N}\frac{1}{N} = 1$$

holds for all $(N,s,a) \in [K]\times\mathscr{S}\times\mathscr{A}$. Hence we can take $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ and apply Lemma 7 to yield

$$|I_{4,1}| = \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\mathrm{R},k^n}\right)^2\right| = \left|\sum_{i=1}^{k}X_i(s_h^k,a_h^k,h,N_h^k)\right|$$
$$\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}u_h^{k^n}(s_h^k,a_h^k,N_h^k)\mathrm{Var}_{h,s_h^k,a_h^k}\left(W_{h+1}^{k^n}\right)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta}$$
$$\lesssim \sqrt{\frac{H^4\log^2\frac{SAT}{\delta}}{N_h^k}} + \frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k} \tag{C.55}$$

with probability at least $1-\delta$, where the last inequality results from the fact that

$$\mathrm{Var}_{h,s_h^k,a_h^k}\left(W_{h+1}^{k^n}\right) \le \left\|W_{h+1}^{k^n}\right\|_\infty^2 \le C_{\mathrm{w}}^2 = H^4.$$

Step 2: bounding $I_{4,2}$. We now turn to the other term $I_{4,2}$ defined in (C.54). Toward this, we first make the observation that

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_{h,s_h^k,a_h^k} V_{h+1}^{\mathrm{R},k^n}\right)^2 \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s_h^k,a_h^k} V_{h+1}^{\mathrm{R},k^n}\right)^2, \tag{C.56}$$

which follows from Jensen's inequality. Equipped with this relation, we can upper bound $I_{4,2}$ as follows:

$$\begin{aligned} I_{4,2} &\le \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n} V_{h+1}^{\mathrm{R},k^n}\right)^2 - \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_{h,s_h^k,a_h^k} V_{h+1}^{\mathrm{R},k^n}\right)^2 \\ &= \left\{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right) V_{h+1}^{\mathrm{R},k^n}\right\}\left\{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h,s_h^k,a_h^k}\right) V_{h+1}^{\mathrm{R},k^n}\right\}. \end{aligned} \tag{C.57}$$

In the following text, we would like to apply this relation to prove

$$I_{4,2} \le C_{42}\left(\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right) \tag{C.58}$$

for some constant $C_{42} > 0$.

When $I_{4,2} \le 0$, the claim (C.58) holds trivially. As a result, we shall focus on the case where $I_{4,2} > 0$. Let us begin with the term in the first pair of curly brackets of (C.57). Toward this, let us abuse the notation and set

$$W_{h+1}^i := V_{h+1}^{\mathrm{R},i} \qquad \text{and} \qquad u_h^i(s,a,N) := \frac{1}{N},$$

which satisfy

$$|u_h^i(s,a,N)| = \frac{1}{N} =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Akin to our argument for bounding $I_{4,1}$, invoking Lemma 7 and setting $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ imply that

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}(P_h^{k^n} - P_{h,s_h^k,a_h^k}) V_{h+1}^{\mathrm{R},k^n}\right| \lesssim \sqrt{\frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k}} + \frac{H\log^2\frac{SAT}{\delta}}{N_h^k}$$

with probability at least $1-\delta$. In addition, the term in the second pair of curly brackets of (C.57) can be bounded straightforwardly by

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h,s_h^k,a_h^k}\right) V_{h+1}^{\mathrm{R},k^n}\right| \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(\left\|P_h^{k^n}\right\|_1 + \left\|P_{h,s_h^k,a_h^k}\right\|_1\right)\left\|V_{h+1}^{\mathrm{R},k^n}\right\|_\infty \le 2H,$$

where we have used $\left\|V_{h+1}^{\mathrm{R},k^n}\right\|_\infty \le H$ and $\left\|P_h^{k^n}\right\|_1 = \left\|P_{h,s_h^k,a_h^k}\right\|_1 = 1$. Substituting the preceding facts into (C.57) validates the bound (C.58) as long as $I_{4,2} > 0$. We have thus finished the proof of the claim (C.58).

Step 3: putting all pieces together. Combining the results (C.55) and (C.58) with (C.54) yields

$$I_4 \le |I_{4,1}| + I_{4,2} \le C_4\left\{\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right\}$$

for some constant $C_4 > 0$. This bound taken together with the definition (C.51) of I_4 gives

$$\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathrm{R},k^n}) \le \left\{\sigma_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k) - \left(\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k)\right)^2\right\} + C_4\left\{\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right\}.$$

Invoke the elementary inequality $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for any $u, v \ge 0$ and use the property (C.53) to obtain

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathrm{R},k^n})\right)^{1/2} \lesssim \left\{\sigma_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k) - \left(\mu_h^{\mathrm{ref},k^{N_h^k}+1}(s_h^k,a_h^k)\right)^2\right\}^{1/2} + \frac{H}{(N_h^k)^{1/4}}\log^{1/2}\frac{SAT}{\delta} + \frac{H}{(N_h^k)^{1/2}}\log\frac{SAT}{\delta}.$$

Substitution into (C.30) directly establishes the desired result (C.31).

C.2 *Proof of Lemma 3*

C.2.1 *Proof of the inequalities (4.11)*

Suppose that we can verify the following inequality:

$$Q_h^{\mathsf{LCB},k}(s,a) \le Q_h^\star(s,a) \qquad \text{for all}(s,a,k,h) \in \mathscr{S}\times\mathscr{A}\times[K]\times[H], \tag{C.59}$$

which in turn yields

$$\max_a Q_h^{\mathsf{LCB},k}(s,a) \le \max_a Q_h^\star(s,a) = V_h^\star(s) \qquad \text{for all}(k,h,s) \in [K]\times[H]\times\mathscr{S}. \tag{C.60}$$

In addition, the construction of $V_h^{\mathsf{LCB},k}$ (see line 14 of Algorithm 3) allows us to show that

$$V_h^{\mathsf{LCB},k+1}(s) \le \max\left\{\max_{j:j\le k+1}\max_a Q_h^{\mathsf{LCB},j}(s,a),\ \max_{j:j\le k} V_h^{\mathsf{LCB},j}(s)\right\}.$$

This taken together with the initialization $V_h^{\mathsf{LCB},1} = 0$ and a simple induction argument yields

$$V_h^{\mathsf{LCB},k}(s) \le V_h^\star(s) \qquad \text{for all}(k,h,s) \in [K]\times[H]\times\mathscr{S}. \tag{C.61}$$

As a consequence, everything comes down to proving the claim (C.59), which we shall accomplish by induction.

Base case. Given our initialization, we have

$$Q_h^{\mathsf{LCB},1}(s,a) - Q_h^\star(s,a) = 0 - Q_h^\star(s,a) \le 0,$$

and hence the claim (C.59) holds trivially when $k = 1$.

Induction step. Suppose now that the claim (C.59) holds all the way up to k for all (s,a,h), and we would like to validate it for the $(k+1)$-th episode as well. Toward this end, recall that the state–action pair (s_h^k, a_h^k) is visited in the k-th episode at time step h; this means that $Q_h^{\mathsf{LCB}}(s_h^k, a_h^k)$ is updated once we collect samples in the k-th episode, with all other entries Q_h^{LCB} frozen. It thus suffices to verify that

$$Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) \le Q_h^\star(s_h^k, a_h^k).$$

In what follows, we shall adopt the short-hand notation (see also Section 4.2)

$$N_h^k = N_h^k(s_h^k, a_h^k) \qquad \text{and} \qquad k^n = k_h^n(s_h^k, a_h^k)$$

which will be used throughout this subsection as long as it is clear from the context.

The update rule of $Q_h^{\mathsf{LCB},k}$ (cf. line 4 of Algorithm 2) and the Bellman optimality equation in (C.8) tell us the following identities:

$$\begin{aligned}
Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) &= Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s_h^k, a_h^k) \\
&= (1-\eta_{N_h^k}) Q_h^{\mathsf{LCB},k^{N_h^k}}(s_h^k, a_h^k) + \eta_{N_h^k}\Big(r_h(s_h^k, a_h^k) + V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - b_h^{k^{N_h^k}}\Big), \\
Q_h^\star(s_h^k, a_h^k) &= (1-\eta_{N_h^k}) Q_h^\star(s_h^k, a_h^k) + \eta_{N_h^k} Q_h^\star(s_h^k, a_h^k) \\
&= (1-\eta_{N_h^k}) Q_h^\star(s_h^k, a_h^k) + \eta_{N_h^k}\Big(r(s_h^k, a_h^k) + P_{h,s_h^k,a_h^k} V_{h+1}^\star\Big),
\end{aligned}$$

which taken collectively lead to the following identity

$$\begin{aligned}
&Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) = Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \\
&= (1-\eta_{N_h^k})\Big(Q_h^{\mathsf{LCB},k^{N_h^k}}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \eta_{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^{N_h^k}}\Big) \\
&= (1-\eta_{N_h^k})\Big(Q_h^{\mathsf{LCB},k^{N_h^k-1}+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \eta_{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^{N_h^k}}\Big).
\end{aligned}$$

Recall the definitions of η_0^N and η_n^N in (4.2). Applying the above relation recursively and using the decomposition of $Q_h^\star(s_h^k, a_h^k)$ in (C.6) result in

$$\begin{aligned}
&Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \\
&\quad = \eta_0^{N_h^k}\Big(Q_h^{\mathsf{LCB},1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^n}\Big) \\
&\quad \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n}) + \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big) V_{h+1}^\star - b_h^{k^n}\Big),
\end{aligned} \tag{C.62}$$

where the inequality follows from the initialization $Q_h^{\mathsf{LCB},1}(s_h^k, a_h^k) = 0 \le Q_h^\star(s_h^k, a_h^k)$ and the definition of P_h^k in (4.5). To continue, we invoke a result established in (29, proof of Lemma 4.3), which guarantees

that with probability at least $1-\delta$,

$$\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right) V_{h+1}^\star \lesssim \sqrt{\frac{H^3 \log(\frac{SAT}{\delta})}{N_h^k}} \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{k^n},$$

provided that c_{b} is some sufficiently large constant. Substituting the above relation into (C.62) implies that

$$Q_h^{\mathsf{LCB},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n})\right) \le 0, \tag{C.63}$$

where the last inequality follows from the induction hypothesis

$$V_{h+1}^{\mathsf{LCB},j}(s) \le V_{h+1}^\star(s) \qquad \text{for all } s \in \mathscr{S} \text{ and } j \le k.$$

The proof is thus completed by induction.

C.2.2 *Proof of the inequality (4.12)* The proof of (4.12) essentially follows the same arguments of (70, Lemma 4.2) (see also (30, Lemma C.7)), an algebraic result leveraging certain relations w.r.t. the Q-value estimates. Accounting for the difference between our algorithm and the one in [70], we paraphrase (70, Lemma 4.2) into the following form that is convenient for our purpose.

LEMMA 9. paraphrased from Lemma 4.2 in [70]Assume that there exists a constant $c_{\mathrm{b}} > 0$ such that for all $(s,a,k,h) \in \mathscr{S} \times \mathscr{A} \times [K] \times [H]$, it holds that

$$\begin{aligned} 0 &\le Q_h^{k+1}(s,a) - Q_h^{\mathsf{LCB},k+1}(s,a) \\ &\le \eta_0^{N_h^k(s,a)} H + \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n})\right) + 4c_{\mathrm{b}} \sqrt{\frac{H^3 \log \frac{SAT}{\delta}}{N_h^k(s,a)}}. \end{aligned} \tag{C.64}$$

Consider any $\varepsilon \in (0,H]$. Then for all $\beta = 1, \ldots, \left\lceil \log_2 \frac{H}{\varepsilon} \right\rceil$, one has

$$\left| \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) \in \left[2^{\beta-1}\varepsilon, 2^\beta \varepsilon\right)\right) \right| \lesssim \frac{H^6 SA \log \frac{SAT}{\delta}}{4^\beta \varepsilon^2}. \tag{C.65}$$

We first show how to justify (4.12) if the inequality (C.65) holds. As can be seen, the fact (C.65) immediately leads to

$$\sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > \varepsilon\right) \lesssim \sum_{\beta=1}^{\left\lceil \log_2 \frac{H}{\varepsilon} \right\rceil} \frac{H^6 SA \log \frac{SAT}{\delta}}{4^\beta \varepsilon^2} \le \frac{H^6 SA \log \frac{SAT}{\delta}}{2\varepsilon^2} \tag{C.66}$$

as desired.

We now return to justify the claim (C.65), toward which it suffices to demonstrate that (C.64) holds. Lemma 2 and Lemma 3 directly verify the left-hand side of (C.64) since

$$Q_h^k(s,a) \ge Q_h^\star(s,a) \ge Q_h^{\mathsf{LCB},k}(s,a) \qquad \text{for all} (s,a,k,h) \in \mathscr{S} \times \mathscr{A} \times [K] \times [H]. \tag{C.67}$$

The remainder of the proof is thus devoted to justifying the upper bound on $Q_h^{k+1}(s,a) - Q_h^{\mathsf{LCB},k+1}(s,a)$ in (C.64). In view of the update rule in line 12 of Algorithm 3, we have the following basic fact:

$$Q_h^{k+1}(s,a) \le Q_h^{\mathsf{UCB},k+1}(s,a).$$

This enables us to obtain

$$Q_h^{k+1}(s,a) - Q_h^{\mathsf{LCB},k+1}(s,a) \le Q_h^{\mathsf{UCB},k+1}(s,a) - Q_h^{\mathsf{LCB},k+1}(s,a) = Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a) - Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a), \tag{C.68}$$

where we abbreviate

$$N_h^k = N_h^k(s,a)$$

throughout this subsection as long as it is clear from the context. Using the update rules of $Q_h^{\mathsf{UCB},k}$ and $Q_h^{\mathsf{LCB},k}$ in line 2 and line 4 of Algorithm 2, we reach

$$\begin{aligned}
&Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a) - Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a) \\
&= (1-\eta_{N_h^k}) Q_h^{\mathsf{UCB},k^{N_h^k}}(s,a) + \eta_{N_h^k}\left(r_h(s,a) + V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right) \\
&\quad - (1-\eta_{N_h^k}) Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a) - \eta_{N_h^k}\left(r_h(s,a) + V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right) \\
&= (1-\eta_{N_h^k})\left(Q_h^{\mathsf{UCB},k^{N_h^k}}(s,a) - Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a)\right) \\
&\quad + \eta_{N_h^k}\left(V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + 2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right) \\
&= (1-\eta_{N_h^k})\left(Q_h^{\mathsf{UCB},k^{N_h^k-1}+1}(s,a) - Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a)\right) \\
&\quad + \eta_{N_h^k}\left(V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + 2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right).
\end{aligned}$$

Applying this relation recursively leads to the desired result

$$
\begin{aligned}
&Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a) - Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a) \\
&\quad = \eta_0^{N_h^k}\Big(Q_h^{\mathsf{UCB},1}(s,a) - Q_h^{\mathsf{LCB},1}(s,a)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) + 2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{n}}\right) \\
&\quad \le \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n})\right) + 4c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}.
\end{aligned}
$$

Here, the last line is valid due to the property $0 \le Q_h^{\mathsf{LCB},1}(s,a) \le Q_h^{\mathsf{UCB},1}(s,a) \le H$ and the following fact:

$$
\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}} \le 2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}},
$$

which is an immediate consequence of the elementary property $\sum_{n=1}^{N}\frac{\eta_n^N}{\sqrt{n}} \le \frac{2}{\sqrt{N}}$ (see Lemma 1). This combined with (C.68) establishes the condition (C.64), thus concluding the proof of the inequality (4.12).

C.3 *Proof of Lemma 4*

C.3.1 *Proof of the inequality (4.15).* Consider any state s that has been visited at least once during the K episodes. Throughout this proof, we shall adopt the shorthand notation

$$
k^i = k_h^i(s),
$$

which denotes the index of the episode in which state s is visited for the i-th time at step h. Given that $V_h(s)$ and $V_h^{\mathrm{R}}(s)$ are only updated during the episodes with indices coming from $\{i \mid 1 \le k^i \le K\}$, it suffices to show that for any s and the corresponding $1 \le k^i \le K$, the claim (4.15) holds in the sense that

$$
\left|V_h^{k^i+1}(s) - V_h^{\mathrm{R},k^i+1}(s)\right| \le 2. \tag{C.69}
$$

Toward this end, we look at three scenarios separately.

Case 1. Suppose that k^i obeys

$$
V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) > 1 \tag{C.70}
$$

or

$$
V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) \le 1 \qquad \text{and} \qquad u_{\mathrm{ref}}^{k^i}(s) = \mathsf{True} \tag{C.71}
$$

The above conditions correspond to the ones in line 15 and line 17 of Algorithm 3 (meaning that V_h^{R} is updated during the k^i-th episode), thus resulting in

$$
V_h^{k^i+1}(s) = V_h^{\mathrm{R},k^i+1}(s).
$$

This clearly satisfies (C.69).

Case 2. Suppose that k^{i_0} is the first time such that (C.70) and (C.71) are violated, namely,

$$i_0 := \min \left\{ j \mid V_h^{k^j+1}(s) - V_h^{\mathsf{LCB},k^j+1}(s) \le 1 \ \text{ and } u_{\mathrm{ref}}^{k^j}(s) = \mathsf{False} \right\}. \tag{C.72}$$

We make three observations.

- The definition (C.72) taken together with the update rules (lines 15–18 of Algorithm 3) reveals that V_h^{R} has been updated in the k^{i_0-1}-th episode, thus indicating that

$$V_h^{\mathrm{R},k^{i_0}}(s) = V_h^{\mathrm{R},k^{i_0-1}+1}(s) = V_h^{k^{i_0-1}+1}(s) = V_h^{k^{i_0}}(s). \tag{C.73}$$

- Additionally, note that under the definition (C.72), $V_h^{\mathrm{R}}(s)$ is not updated during the k^{i_0}-th episode, namely,

$$V_h^{\mathrm{R},k^{i_0}+1}(s) = V_h^{\mathrm{R},k^{i_0}}(s). \tag{C.74}$$

- The definition of k^{i_0} indicates that either (C.70) or (C.71) is satisfied in the previous episode $k^i = k^{i_0-1}$ in which s was visited. If (C.70) is satisfied, then lines 15–16 in Algorithm 3 tell us that

$$\mathsf{True} = u_{\mathrm{ref}}^{k^{i_0-1}+1}(s) = u_{\mathrm{ref}}^{k^{i_0}}(s), \tag{C.75}$$

which, however, contradicts the assumption $u_{\mathrm{ref}}^{k^{i_0}}(s) = \mathrm{False}$ in (C.72). Therefore, in the k^{i_0-1}-th episode, (C.71) is satisfied, thus leading to

$$V_h^{k^{i_0}}(s) - V_h^{\mathsf{LCB},k^{i_0}}(s) = V_h^{k^{i_0-1}+1}(s) - V_h^{\mathsf{LCB},k^{i_0-1}+1}(s) \le 1. \tag{C.76}$$

We see from (C.73), (C.74) and (C.76) that

$$V_h^{\mathrm{R},k^{i_0}+1}(s) - V_h^{k^{i_0}+1}(s) = V_h^{\mathrm{R},k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) = V_h^{k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) \tag{C.77}$$

$$\overset{\text{(i)}}{\le} V_h^{k^{i_0}}(s) - V_h^{\mathsf{LCB},k^{i_0}}(s) \overset{\text{(ii)}}{\le} 1, \tag{C.78}$$

where (i) holds since $V_h^{k^{i_0}+1}(s) \ge V_h^{\star}(s) \ge V_h^{\mathsf{LCB},k^{i_0}}(s)$, and (ii) follows from (C.76). In addition, we make note of the fact that

$$V_h^{\mathrm{R},k^{i_0}+1}(s) - V_h^{k^{i_0}+1}(s) = V_h^{k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) \ge 0, \tag{C.79}$$

which follows from (C.77) and the monotonicity of $V_h^k(s)$ in k. With the above results in place, we arrive at the advertised bound (C.69) when $i = i_0$.

Case 3. Consider any $i > i_0$. It is easily verified that

$$V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) \le 1 \qquad \text{and} \qquad u_{\mathrm{ref}}^{k^i}(s) = \mathsf{False}. \tag{C.80}$$

It then follows that

$$V_h^{\mathrm{R},k^i+1}(s) \overset{\text{(i)}}{\leq} V_h^{\mathrm{R},k^{i_0}+1}(s) \overset{\text{(ii)}}{\leq} V_h^{k^{i_0}+1}(s)+1 \overset{\text{(iii)}}{\leq} V_h^{\mathsf{LCB},k^{i_0}+1}(s)+2$$
$$\overset{\text{(iv)}}{\leq} V_h^{\star}(s)+2 \overset{\text{(v)}}{\leq} V_h^{k^i+1}(s)+2. \tag{C.81}$$

Here, (i) holds due to the monotonicity of V_h^{R} and V_h^k (see line 14 of Algorithm 3), (ii) is a consequence of (C.78), (iii) comes from the definition (C.72) of i_0, (iv) arises since V_h^{LCB} is a lower bound on $V_h^{\star}$ (see Lemma 3) and (v) is valid since $V_h^{k^i+1}(s) \geq V_h^{\star}(s)$ (see Lemma 2). In addition, in view of the monotonicity of V_h^k (see line 14 of Algorithm 3) and the update rule in line 16 of Algorithm 3, we know that

$$V_h^{\mathrm{R},k^i+1}(s) \geq V_h^{k^i+1}(s).$$

The preceding two bounds taken collectively demonstrate that

$$0 \leq V_h^{\mathrm{R},k^i+1}(s) - V_h^{k^i+1}(s) \leq 2,$$

thus justifying (C.69) for this case.

Therefore, we have established (C.69)—and hence (4.15)—for all cases.

C.3.2 *Proof of the inequality (4.16)* Suppose that

$$V^{\mathrm{R}},k_h(s_h^k) - V^{\mathrm{R}},k_h(s_h^k) \neq 0 \tag{C.82}$$

holds for some $k < K$. Then there are two possible scenarios to look at:

(a) *Case 1: the condition in line 15 and line 17 of Algorithm 3 are violated at step h of the k-th episode.* This means that we have

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1 \quad \text{and} \quad u_{\mathrm{ref}}^k(s_h^k) = \mathsf{False} \tag{C.83}$$

in this case. Then for any $k' > k$, one necessarily has

$$\begin{cases} V_h^{k'}(s_h^k) - V_h^{\mathsf{LCB},k'}(s_h^k) \leq V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1, \\ u_{\mathrm{ref}}^{k'}(s_h^k) = u_{\mathrm{ref}}^k(s_h^k) = \mathsf{False}, \end{cases} \tag{C.84}$$

where the first property uses the monotonicity of V_h^k and $V_h^{\mathsf{LCB},k}$ (see (4.7b) and line 14 of Algorithm 3). In turn, Condition (C.84) implies that V_h^{R} will no longer be updated after the k-th episode (see line 15 of Algorithm 3), thus indicating that

$$V^{\mathrm{R}},k_h(s_h^k) = V_h^{\mathrm{R},k+1}(s_h^k) = \cdots = V^{\mathrm{R}},k_h(s_h^k). \tag{C.85}$$

This, however, contradicts the assumption (C.82).

(b) *Case 2: the condition in either line 15 or line 17 of Algorithm 3 is satisfied at step h of the k-th episode.* If this occurs, then the update rule in line 15 of Algorithm 3 implies that

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) > 1, \tag{C.86}$$

or

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1 \qquad \text{and} \quad u_{\mathrm{ref}}^k(s_h^k) = \mathsf{True}. \tag{C.87}$$

To summarize, the above argument demonstrates that (C.82) can only occur if either (C.86) or (C.87) holds.

With the above observation in place, we can proceed with the following decomposition:

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\right)&=\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\neq 0\right)\\&\le\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathrm{LCB},k+1}(s_h^k)\le 1\ \text{and}\ u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)\\&\quad+\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^k(s_h^k)-V_h^{\mathrm{LCB},k}(s_h^k)\right)\mathbb{1}\left(V_h^k(s_h^k)-V_h^{\mathrm{LCB},k}(s_h^k)>1\right)}_{=:\omega}.\end{aligned}\tag{C.88}$$

Regarding the first term in (C.88), it is readily seen that for all $s\in\mathscr{S}$,

$$\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s)-V_h^{\mathrm{LCB},k+1}(s)\le 1\ \text{and}\ u_{\mathrm{ref}}^k(s)=\mathsf{True}\right)\le 1,\tag{C.89}$$

which arises since, for each $s\in\mathscr{S}$, the above condition is satisfied in at most one episode, owing to the monotonicity property of V_h, V_h^{LCB} and the update rule for u_{ref} in (17). As a result, one has

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathrm{LCB},k+1}(s_h^k)\le 1\ \text{and}\ u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)\\&\quad\le H\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathrm{LCB},k+1}(s_h^k)\le 1\ \text{and}\ u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)\\&\quad=H\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s)-V_h^{\mathrm{LCB},k+1}(s)\le 1\ \text{and}\ u_{\mathrm{ref}}^k(s)=\mathsf{True}\right)\\&\quad\le H\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}1=H^2S,\end{aligned}$$

where the first inequality holds since $\|V_h^{\mathrm{R}},k-V_h^{\mathrm{R}},k\|_\infty\le H$. Substitution into (C.88) yields

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathrm{R},k}(s_h^k)-V_h^{\mathrm{R},K}(s_h^k)\right)\le H^2S+\omega.\tag{C.90}$$

To complete the proof, it boils down to bounding the term ω defined in (C.88). To begin with, note that

$$V^{\mathrm{R}},k_h(s_h^k)\ge V_h^\star(s_h^k)\ge V_h^{\mathrm{LCB},k}(s_h^k),$$

where we use the optimism of $V_h^{\mathsf{R}}, k(s_h^k)$ stated in Lemma 2 (cf. (4.9)) and the pessimism of V_h^{LCB} in Lemma 3 (see (4.11)). As a result, we can obtain

$$\begin{aligned}\omega &= \sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^k(s_h^k) - V_h^{\mathsf{LCB},k}(s_h^k)\right)\mathbb{1}\left(V_h^k(s_h^k) - V_h^{\mathsf{LCB},k}(s_h^k) > 1\right)\\ &\le \sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > 1\right), \end{aligned} \tag{C.91}$$

where the second line arises from the properties $V_h^k(s_h^k) = Q_h^k(s_h^k,a_h^k)$ (given that $a_h^k = \arg\max_a Q_h^k(s_h^k,a)$) as well as the following fact (see line 14 of Algorithm 3)

$$V_h^{\mathsf{LCB},k}(s_h^k) \ge \max_a Q_h^{\mathsf{LCB},k}(s_h^k,a) \ge Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k).$$

Further, let us make note of the following elementary identity:

$$Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) = \int_0^\infty \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > t\right)\mathrm{d}t.$$

This allows us to obtain

$$\begin{aligned}\omega &\le \sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\int_0^\infty \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > t\right)\mathrm{d}t\right\}\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > 1\right)\\ &= \int_1^H \sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > t\right)\mathrm{d}t\\ &\lesssim \int_1^H \frac{H^6SA\log\frac{SAT}{\delta}}{t^2}\mathrm{d}t \lesssim H^6SA\log\frac{SAT}{\delta}, \end{aligned} \tag{C.92}$$

where the last line follows from the property (4.12) in Lemma 3. Combining the above bounds (C.91) and (C.92) with (C.90) establishes

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V^{\mathsf{R}},k_h(s_h^k) - V^{\mathsf{R}},k_h(s_h^k)\right)\\ &\le H^2S + \sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > 1\right)\\ &\le H^6SA\log\frac{SAT}{\delta}\end{aligned}$$

as claimed.

D. Proof of Lemma 5

For notational simplicity, we shall adopt the short-hand notation

$$k^n = k_h^n(s_h^k,a_h^k)$$

throughout this section. A starting point for proving this lemma is the upper bound already derived in (4.20), and we intend to further bound the first term on the right-hand side of (4.20). Recalling the expression of $Q_h^{\mathrm{R},k+1}(s_h^k, a_h^k)$ in (C.7) and (C.9), we can derive

$$
\begin{aligned}
Q_h^{\mathrm{R},k}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) &= Q_h^{\mathrm{R},k^{N_h^{k-1}(s_h^k,a_h^k)}+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \qquad \text{(D.1)}\\
&= \eta_0^{N_h^{k-1}(s_h^k,a_h^k)}\left(Q_h^{\mathrm{R},1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\right) + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)} b_h^{\mathrm{R},k^n+1}\\
&\quad + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right)\\
&\le \eta_0^{N_h^{k-1}(s_h^k,a_h^k)} H + B_h^{\mathrm{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\left(N_h^{k-1}(s_h^k,a_h^k)\right)^{3/4}}\log\frac{SAT}{\delta}\\
&\quad + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right),
\end{aligned}
$$

where the last line follows from (C.35) with $B_h^{\mathrm{R},k^{N_h^{k-1}}+1} = B^{\mathrm{R}}, k_h$ and the initialization $Q_h^{\mathrm{R},1}(s_h^k,a_h^k) = H$. Summing over all $1 \le k \le K$ gives

$$
\begin{aligned}
&\sum_{k=1}^K \left(Q^{\mathrm{R}}, k_h(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\right)\\
&\le \sum_{k=1}^K \left(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B^{\mathrm{R}}, k_h(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\left(N_h^{k-1}(s_h^k,a_h^k)\right)^{3/4}}\log\frac{SAT}{\delta}\right)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) + \frac{\sum_{i=1}^n V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})}{n} - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right)\\
&\le \sum_{k=1}^K \left(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B^{\mathrm{R}}, k_h(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\left(N_h^{k-1}(s_h^k,a_h^k)\right)^{3/4}}\log\frac{SAT}{\delta}\right)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n})\right)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^\star(s_{h+1}^{k^n}) - V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right).
\end{aligned} \qquad \text{(D.2)}
$$

Next, we control each term in (D.2) separately.

- Regarding the first term of (D.2), we make two observations. To begin with,

$$\sum_{k=1}^{K} \eta_0^{N_h^{k-1}(s_h^k,a_h^k)} \le \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \sum_{n=0}^{N_h^{K-1}(s,a)} \eta_0^n \le SA, \tag{D.3}$$

where the last inequality follows since $\eta_0^n = 0$ for all $n > 0$ (see (4.2)). Next, it is also observed that

$$\begin{aligned}\sum_{k=1}^{K} \frac{1}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}} &= \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \sum_{n=1}^{N_h^{K-1}(s,a)} \frac{1}{n^{3/4}} \\ &\le \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} 4\big(N_h^{K-1}(s,a)\big)^{1/4} \le 4(SA)^{3/4}K^{1/4}, \end{aligned}\tag{D.4}$$

where the last inequality comes from Holder's inequality

$$\sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \big(N_h^{K-1}(s,a)\big)^{1/4} \le \bigg[\sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} 1\bigg]^{3/4} \bigg[\sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} N_h^{K-1}(s,a)\bigg]^{1/4} \le (SA)^{3/4}K^{1/4}.$$

Combining the above bounds yields

$$\begin{aligned}&\sum_{k=1}^{K}\bigg(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B_h^{\mathrm{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}}\log\frac{SAT}{\delta}\bigg) \\ &\quad\le HSA + \sum_{k=1}^{K} B_h^{\mathrm{R},k}(s_h^k,a_h^k) + 8c_{\mathrm{b}}(SA)^{3/4}K^{1/4}H^2\log\frac{SAT}{\delta}. \end{aligned}\tag{D.5}$$

- We now turn to the second term of (D.2). A little algebra gives

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\big) \\ &= \sum_{l=1}^{K}\sum_{N=N_h^l(s_h^l,a_h^l)}^{N_h^{K-1}(s_h^l,a_h^l)} \eta_{N_h^l(s_h^l,a_h^l)}^{N}\big(V_{h+1}^{l}(s_{h+1}^{l}) - V_{h+1}^{\star}(s_{h+1}^{l})\big) \\ &\le \bigg(1+\frac{1}{H}\bigg)\sum_{l=1}^{K}\big(V_{h+1}^{l}(s_{h+1}^{l}) - V_{h+1}^{\star}(s_{h+1}^{l})\big) \\ &= \bigg(1+\frac{1}{H}\bigg)\bigg[\sum_{k=1}^{K}\big(V_{h+1}^{k}(s_{h+1}^{k}) - V_{h+1}^{\pi^k}(s_{h+1}^{k})\big) - \sum_{k=1}^{K}\big(V_{h+1}^{\star}(s_{h+1}^{k}) - V_{h+1}^{\pi^k}(s_{h+1}^{k})\big)\bigg]. \end{aligned}\tag{D.6}$$

Here, the second line replaces k^n (resp. n) with l (resp. $N_h^l(s_h^l,a_h^l)$), the third line is due to the property $\sum_{N=n}^{\infty}\eta_n^N \le 1 + 1/H$ (see Lemma 1), while the last relation replaces l with k again.

- When it comes to the last term of (D.2), we can derive

$$\sum_{k=1}^{K}\sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)}\eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(V_{h+1}^{\star}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+\frac{1}{n}\sum_{i=1}^{n}V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V_{h+1}^{\star}\right)$$

$$=\sum_{k=1}^{K}\sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)}\eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\left(\left(P_h^{k^n}-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\star}-V_{h+1}^{\mathrm{R},k^n}\right)+\frac{1}{n}\sum_{i=1}^{n}\left(V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V_{h+1}^{\mathrm{R},k^n}\right)\right)$$

$$=\sum_{k=1}^{K}\sum_{N=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\eta_{N_h^k(s_h^k,a_h^k)}^{N}\left(\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\star}-V^{\mathrm{R}},k_{h+1}\right)\right.$$

$$\left.+\frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V^{\mathrm{R}},k_{h+1}\right)}{N_h^k(s_h^k,a_h^k)}\right).$$

Here, the first equality holds since $V_{h+1}^{\star}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})=P_h^{k^n}\left(V_{h+1}^{\star}-V_{h+1}^{\mathrm{R},k^n}\right)$ (in view of the definition of P_h^k in (4.5)), the second equality can be seen via simple rearrangement of the terms, while in the last line we replace k^n (resp. n) with k (resp. $N_h^k(s_h^k,a_h^k)$).

Taking the above bounds together with (D.2) and (4.20), we can rearrange terms to reach

$$\begin{aligned}
&\sum_{k=1}^{K}\left(V_h^k(s_h^k)-V_h^{\pi^k}(s_h^k)\right)\\
&\le\left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\left(V_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)\right)+\sum_{k=1}^{K}B^{\mathrm{R}},k_h(s_h^k,a_h^k)\\
&+HSA+8c_{\mathrm{b}}H^2(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta}+\sum_{k=1}^{K}\left(P_{h,s_h^k,a_h^k}-P_h^k\right)\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right)\\
&+\sum_{k=1}^{K}\sum_{N=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\eta_{N_h^k(s_h^k,a_h^k)}^{N}\left[\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\star}-V^{\mathrm{R}},k_{h+1}\right)\right.\\
&\left.+\frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V^{\mathrm{R}},k_{h+1}\right)}{N_h^k(s_h^k,a_h^k)}\right],
\end{aligned}\tag{D.7}$$

where we have dropped the term $-\frac{1}{H}\sum_k\left(V_{h+1}^{\star}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)\right)$ owing to the fact that $V_{h+1}^{\star}\ge V_{h+1}^{\pi^k}$.

Thus far, we have established a crucial connection between $\sum_{k=1}^{K}\left(V_h^k(s_h^k)-V_h^{\pi^k}(s_h^k)\right)$ at step h and $\sum_{k=1}^{K}\left(V_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)\right)$ at step $h+1$. Clearly, the term $V_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)$ can be further bounded in the same manner. As a result, by recursively applying the above relation (D.7) over the time steps $h=1,2,\cdots,H$ and using the terminal condition $V_{H+1}^k=V_{H+1}^{\pi^k}=0$, we can immediately arrive at the advertised bound in Lemma 5.

E. Proof of Lemma 6

E.1 *Bounding the term* $\mathscr{R}_1$

First of all, let us look at the first two terms of $\mathscr{R}_1$ in (4.22a). Recognizing the following elementary inequality:

$$\left(1+\frac{1}{H}\right)^{h-1} \le \left(1+\frac{1}{H}\right)^{H} \le e \quad \text{for all} h = 1, 2, \cdots, H+1, \tag{E.1}$$

we are allowed to upper bound the first two terms in (4.22a) as follows:

$$\begin{aligned}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left\{HSA + 8c_{\mathrm{b}}H^2(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta}\right\} &\lesssim H^2SA + H^3(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta}\\ &\lesssim H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK} = H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^2SAT},\end{aligned} \tag{E.2}$$

where the last inequality can be shown using the AM–GM inequality as follows:

$$H^3(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta} = \left(H^{9/4}\sqrt{SA}\log\frac{SAT}{\delta}\right)\cdot(H^3SAK)^{1/4} \le H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK}.$$

We are now left with the last term of $\mathscr{R}_1$ in (4.22a). Toward this, we resort to Lemma 8 by setting

$$W_{h+1}^i := V_{h+1}^{\star} - V_{h+1}^{\pi^k} \qquad \text{and} \qquad c_h := \left(1+\frac{1}{H}\right)^{h-1}.$$

In view of (E.1) and the property $H \ge V^{\star}(s) \ge V^{\pi}(s) \ge 0$, we see that

$$0 \le c_h \le e, \qquad W_{h+1}^i \ge 0, \qquad \text{and} \qquad \|W_{h+1}^i\|_{\infty} \le H =: C_{\mathrm{w}}.$$

Therefore, applying Lemma 8 yields

$$\begin{aligned}\left|\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\left(P_{h,s_h^k,a_h^k} - P_h^k\right)\left(V_{h+1}^{\star} - V_{h+1}^{\pi^k}\right)\right| &= \left|\sum_{h=1}^{H}\sum_{k=1}^{K}Y_{k,h}\right|\\ &\lesssim \sqrt{TC_{\mathrm{w}}^2\log\frac{1}{\delta}} + C_{\mathrm{w}}\log\frac{1}{\delta} = \sqrt{H^2T\log\frac{1}{\delta}} + H\log\frac{1}{\delta}\end{aligned} \tag{E.3}$$

with probability exceeding $1-\delta$.

Combining (E.2) and (E.3) with the definition (4.22a) of $\mathscr{R}_1$ immediately leads to the claimed bound.

E.2 *Bounding the term* $\mathscr{R}_2$

In view of the definition of $B^{\mathrm{R}}, k_h(s_h^k, a_h^k)$ in line ?? of Algorithm 2, we can decompose $\mathscr{R}_2$ (cf. (4.22b)) as follows:

$$\begin{aligned}\mathscr{R}_2 &= \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{H\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}}\\ &\quad+\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}}\\ &\lesssim \sqrt{H\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}}\\ &\quad+\sqrt{\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}},\end{aligned} \tag{E.4}$$

where the last relation holds due to (E.1). In what follows, we intend to bound these two terms separately.

Step 1: upper bounding the first term in (E.4). Toward this, we make the observation that

$$\begin{aligned}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} &\le \sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s_h^k,a_h^k)}{N_h^k(s_h^k,a_h^k)}}\\ &= \sum_{k=1}^{K}\sqrt{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}},\end{aligned} \tag{E.5}$$

where the second line follows from the update rule of $\sigma_h^{\mathrm{adv}}, k$ in (C38). Combining the relation $|V_{h+1}^k(s_h^k) - V^{\mathrm{R}}, k_{h+1}(s_h^k)| \le 2$ (cf. (4.15)) and the property $\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)} \le 1$ (cf. (4.3)) with (E.5) yields

$$\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \le \sum_{k=1}^{K}\sqrt{\frac{4}{N_h^k(s_h^k,a_h^k)}} \le 2\sqrt{SAK}. \tag{E.6}$$

Here, the last inequality holds due to the following fact:

$$\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{1}{n}}\le 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)}$$
$$\le 2\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}1}\cdot\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_h^K(s,a)}=2\sqrt{SAK}, \tag{E.7}$$

where the last line arises from Cauchy–Schwarz and the basic fact that $\sum_{(s,a)}N_h^K(s,a)=K$.

Step 2: upper bounding the second term in (E.4). Recalling the update rules of $\mu_h^{\mathrm{ref},k}$ and $\sigma_h^{\mathrm{ref},k}$ in (C52), we have

$$\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathrm{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}}$$
$$=\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2}}_{=:J_h^k}. \tag{E.8}$$

Additionally, the quantity J_h^k defined in (E.8) obeys

$$(J_h^k)^2\le\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2-\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}$$
$$+\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\star}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2$$
$$\le\underbrace{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}2H\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big)}{N_h^k(s_h^k,a_h^k)}}_{=:J_1}$$
$$+\underbrace{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\star}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2}_{=:J_2}, \tag{E.9}$$

which arises from the fact that $H\ge V_{h+1}^{\mathrm{R},k^n}\ge V_{h+1}^{\star}\ge 0$ for all $k^n\le K$ and hence

$$\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})\big)^2-\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2=\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})+V_{h+1}^{\star}(s_{h+1}^{k^n})\big)\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big)$$
$$\le 2H\big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big).$$

With (E.9) in mind, we shall proceed to bound each term in (E.9) separately.

- The first term J_1 can be straightforwardly bounded as follows:

$$\begin{aligned} J_1 &= \frac{2H}{N_h^k(s_h^k,a_h^k)}\Bigg(\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\Big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\Big) \\ &\qquad \mathbb{1}\Big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n}) \le 3\Big) + \Phi_h^k(s_h^k,a_h^k)\Bigg) \\ &\le 6H + \frac{2H}{N_h^k(s_h^k,a_h^k)}\Phi_h^k(s_h^k,a_h^k), \end{aligned} \tag{E.10}$$

where $\Phi_h^k(s_h^k,a_h^k)$ is defined as

$$\Phi_h^k(s_h^k,a_h^k) := \sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\Big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\Big)\mathbb{1}\Big(V_{h+1}^{\mathrm{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n}) > 3\Big). \tag{E.11}$$

- When it comes to the second term J_2, we claim that

$$J_2 \lesssim \mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star}) + H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k(s_h^k,a_h^k)}}, \tag{E.12}$$

which will be justified in Appendix E.2.1.

Plugging (E.10) and (E.12) into (E.9) and (E.8) allows one to demonstrate that

$$\begin{aligned} &\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s_h^k,a_h^k) - \big(\mu_h^{\mathrm{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\ &\lesssim \sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}\sqrt{H + \frac{H\Phi_h^k(s_h^k,a_h^k)}{N_h^k(s_h^k,a_h^k)} + \mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star}) + H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k(s_h^k,a_h^k)}}} \\ &\lesssim \sum_{k=1}^{K}\left(\sqrt{\frac{H}{N_h^k(s_h^k,a_h^k)}} + \frac{\sqrt{H\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} + \sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} + \frac{H\log^{1/4}\frac{SAT}{\delta}}{\big(N_h^k(s_h^k,a_h^k)\big)^{3/4}}\right) \\ &\lesssim \sqrt{HSAK} + \sum_{k=1}^{K}\frac{\sqrt{H\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} + \sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} + H(SA)^{3/4}\left(K\log\frac{SAT}{\delta}\right)^{1/4}, \end{aligned} \tag{E.13}$$

where the last line follows from (E.7) and (D.4).

Step 3: putting together the preceding results.

Finally, the above results in (E.6) and (E.13) taken collectively with (E.4) lead to

$$\begin{aligned}
\mathscr{R}_2 &\lesssim \sqrt{H^3SAK\log\frac{SAT}{\delta}} + \sum_{h=1}^{H}\sqrt{\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathrm{ref},k}(s_h^k,a_h^k) - \big(\mu_h^{\mathrm{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\lesssim \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} + \sqrt{\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} \\
&\quad + \sqrt{H\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} + H^4SA\log^2\frac{SAT}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta} = \sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta}.
\end{aligned}$$

Here, (i) holds due to the following two claimed inequalities:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} \lesssim \sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^4SA\log\frac{SAT}{\delta}, \tag{E.14}$$

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} \lesssim H^{7/2}SA\log^{3/2}\frac{SAT}{\delta}, \tag{E.15}$$

whose proofs are postponed to Appendix E.2.2 and Appendix E.2.3, respectively. Additionally, the inequality (ii) above is valid since

$$\begin{aligned}
H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} &= \left(H^{5/4}(SA)^{1/2}\log\frac{SAT}{\delta}\right)\cdot\left(H^3SAK\log\frac{SAT}{\delta}\right)^{1/4} \\
&\lesssim H^{2.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK\log\frac{SAT}{\delta}} = H^{2.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^2SAT\log\frac{SAT}{\delta}}
\end{aligned}$$

due to the Cauchy–Schwarz inequality. This concludes the proof of the advertised upper bound on $\mathscr{R}_2$.

E.2.1 *Proof of the inequality (E12).* Akin to the proof of I_4^1 in (C.55), let

$$W_{h+1}^i := (V_{h+1}^\star)^2 \qquad \text{and} \qquad u_h^i(s,a,N) := \frac{1}{N}.$$

By observing and setting

$$C_{\mathrm{u}} := \frac{1}{N}, \qquad \|W_{h+1}^i\|_\infty \le H^2 =: C_{\mathrm{w}},$$

we can apply Lemma 7 to yield

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2\right| = \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^\star\right)^2\right| \lesssim H^2\sqrt{\frac{\log^2\frac{SAT}{\delta}}{N_h^k}}$$

with probability at least $1-\delta$. Similarly, by applying the trivial bound $\|V_{h+1}^\star\|_\infty \le H$ and Lemma 7, we can obtain

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right| = \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right)V_{h+1}^\star\right| \lesssim H\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}}$$

with probability at least $1-\delta$.

Recalling from (4.6) the definition

$$\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) = P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2 - \left(P_{h,s_h^k,a_h^k}V_{h+1}^\star\right)^2,$$

we can use the preceding two bounds and the triangle inequality to show that

$$\begin{aligned}
&\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})^2 - \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - \mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)\right| \\
&\quad\le \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})^2 - P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2\right| + \left|\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - (P_{h,s_h^k,a_h^k}V_{h+1}^\star)^2\right| \\
&\quad\lesssim H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}} + \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right|\cdot\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) + P_{h,s_h^k,a_h^k}V_{h+1}^\star\right| \\
&\quad\lesssim H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}}
\end{aligned}$$

with probability at least $1-\delta$, where the last line also uses the fact that $\|V_{h+1}^\star\|_\infty \le H$.

E.2.2 *Proof of the inequality (E14)* To begin with, we make the observation that

$$\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{\mathrm{Var}_{h,s,a}(V_{h+1}^\star)}{n}} \le 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)\mathrm{Var}_{h,s,a}(V_{h+1}^\star)},$$

which relies on the fact that $\sum_{n=1}^{N} 1/\sqrt{n} \le 2\sqrt{N}$. It then follows that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} &\le 2\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)\mathrm{Var}_{h,s,a}(V_{h+1}^{\star})} \\ &\le 2\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}1}\cdot\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_h^K(s,a)\mathrm{Var}_{h,s,a}(V_{h+1}^{\star})} \\ &= 2\sqrt{HSA}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}\,, \end{aligned} \tag{E.16}$$

where the second inequality invokes the Cauchy–Schwarz inequality.

The rest of the proof is then dedicated to bounding (E.16). Toward this end, we first decompose

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star}) &\le \sum_{h=1}^{H}\sum_{k=1}^{K}\mathrm{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{\pi^k}\big)+\sum_{h=1}^{H}\sum_{k=1}^{K}\Big|\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})-\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\Big| \\ &\overset{\text{(ii)}}{\lesssim} HT+H^3\log\frac{SAT}{\delta}+\sum_{h=1}^{H}\sum_{k=1}^{K}\Big|\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})-\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\Big|, \end{aligned} \tag{E.17}$$

where (ii) follows directly from (30, Lemma C.5). The second term on the right-hand side of (E.17) can be bounded as follows:

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{k=1}^{K}\Big|\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})-\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\Big| \\ &=\sum_{h=1}^{H}\sum_{k=1}^{K}\Big|P_{h,s_h^k,a_h^k}(V_{h+1}^{\star})^2-\big(P_{h,s_h^k,a_h^k}V_{h+1}^{\star}\big)^2-P_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})^2+\big(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\big)^2\Big| \\ &\le\sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\Big|P_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\big)\big(V_{h+1}^{\star}+V_{h+1}^{\pi^k}\big)\Big)\Big|+\Big|\big(P_{h,s_h^k,a_h^k}V_{h+1}^{\star}\big)^2-\big(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\big)^2\Big|\right\} \\ &\overset{\text{(i)}}{\le} 4H\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\big) \\ &=4H\sum_{h=1}^{H}\sum_{k=1}^{K}\Big\{V_{h+1}^{\star}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)+\big(P_{h,s_h^k,a_h^k}-P_h^k\big)\big(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\big)\Big\}\end{aligned}$$

$$\overset{\text{(ii)}}{\le} 4H\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\phi_{h+1}^{k}+\delta_{h+1}^{k}\right) \overset{\text{(iii)}}{\lesssim} H^2\sqrt{T\log\frac{SAT}{\delta}}+H^4\sqrt{SAT\log\frac{SAT}{\delta}}+H^4SA$$

$$\asymp H^4\sqrt{SAT\log\frac{SAT}{\delta}}+H^4SA, \tag{E.18}$$

where we define

$$\delta_{h+1}^{k}:=V_{h+1}^{\mathsf{UCB},k}(s_{h+1}^{k})-V_{h+1}^{\pi^k}(s_{h+1}^{k}),\qquad \phi_{h+1}^{k}:=\left(P_{h,s_h^k,a_h^k}-P_h^k\right)\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right). \tag{E.19}$$

We shall take a moment to explain how we derive (E.18). The inequality (i) holds by observing that $V_{h+1}^{\star}-V_{h+1}^{\pi^k}\ge 0$ and

$$\begin{aligned}\left|P_{h,s_h^k,a_h^k}\left(\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right)\left(V_{h+1}^{\star}+V_{h+1}^{\pi^k}\right)\right)\right| &\le P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right)\left(\left\|V_{h+1}^{\star}\right\|_\infty+\left\|V_{h+1}^{\pi^k}\right\|_\infty\right)\\ &\le 2HP_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right),\\ \left|\left(P_{h,s_h^k,a_h^k}V_{h+1}^{\star}\right)^2-\left(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\right)^2\right| &\le \left|P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right)\right|\cdot\left|P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}+V_{h+1}^{\pi^k}\right)\right|\\ &\le 2HP_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}-V_{h+1}^{\pi^k}\right);\end{aligned}$$

(ii) is valid since $V_{h+1}^{\mathsf{UCB}}\ge V_{h+1}^{\star}$; and (iii) results from the following two bounds:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\delta_{h+1}^{k}\lesssim H^3\sqrt{SAT\log\frac{SAT}{\delta}}+H^3SA, \tag{E.20a}$$

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\phi_{h+1}^{k}\lesssim H\sqrt{T\log\frac{SAT}{\delta}}, \tag{E.20b}$$

which come, respectively, from (30, Eqn. (C.13)) and the argument for (30, Eqn. (C.12)).[4]

As a consequence, substituting (E.17) and (E.18) into (E.16), we reach

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} &\lesssim \sqrt{HSA}\sqrt{HT+H^4\sqrt{SAT\log\frac{SAT}{\delta}}+H^4SA}\\ &\lesssim \sqrt{H^2SAT}+H^{5/2}(SA)^{3/4}\left(T\log\frac{SAT}{\delta}\right)^{1/4}+H^{2.5}SA\\ &= \sqrt{H^2SAT}+\left(H^2SAT\log\frac{SAT}{\delta}\right)^{1/4}\left(H^4SA\right)^{1/2}+H^{2.5}SA\\ &\lesssim \sqrt{H^2SAT\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta},\end{aligned}$$

[4] Note that the notation δ_h^k used in ([30], Section C.2) and the one in the proof of ([30], Theorem 1) are different; here, we need to adopt the notation used in the proof of ([30], Theorem 1).

where we have applied the basic inequality $2ab \le a^2 + b^2$ for any $a, b \ge 0$.

E.2.3 *Proof of the inequality (E15)* First, it is observed that

$$\sum_{k=1}^{K} \frac{\sqrt{\Phi_h^k(s_h^k, a_h^k)}}{N_h^k(s_h^k, a_h^k)} = \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \sum_{n=1}^{N_h^K(s,a)} \frac{\sqrt{\Phi_h^{k^n(s,a)}(s,a)}}{n}$$
$$\le \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \sqrt{\Phi_h^{N_h^K(s,a)}(s,a)} \log T \le \sqrt{SA \sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \Phi_h^{N_h^K(s,a)}(s,a)} \log T. \quad \text{(E.21)}$$

Here, the first inequality holds by the monotonicity property of $\Phi_h^k(s_h, a_h)$ with respect to k (see its definition in (E.11)) due to the same property of V_{h+1}^{R}, k, while the second inequality comes from Cauchy–Schwarz.

To continue, note that

$$\sum_{h=1}^{H} \sqrt{\sum_{(s,a)\in\mathscr{S}\times\mathscr{A}} \Phi_h^{N_h^K(s,a)}(s,a)}$$
$$= \sum_{h=1}^{H} \sqrt{\sum_{k=1}^{K} \left(V_{h+1}^{\mathrm{R},k}(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k)\right) \mathbb{1}\left(V_{h+1}^{\mathrm{R},k}(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k) > 3\right)}$$
$$\le \sum_{h=1}^{H} \sqrt{\sum_{k=1}^{K} \left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right) \mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 3\right)}$$
$$= \sum_{h=1}^{H} \sqrt{\sum_{k=1}^{K} \left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right) \mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}$$
$$\le \sum_{h=1}^{H} \sqrt{\sum_{k=1}^{K} 3\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right) \mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}$$
$$\le \sqrt{H} \sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K} 3\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right) \mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}, \quad \text{(E.22)}$$

where the first inequality follows from Lemma 4 (cf. (4.15)) and Lemma 3 (so that $V_{h+1}^{\mathrm{R}}, k(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k) \le V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)$), the penultimate inequality holds since $1 \le V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)$ when $\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right) \ne 0$, and the last inequality is a consequence of the Cauchy–Schwarz inequality.

Combining the above relation with (C.91) and applying the triangle inequality, we can demonstrate that

$$\sum_{h=1}^{H}\sqrt{\sum_{(s,a)\in\mathscr{S}\times\mathscr{A}}\Phi_h^{N_h^K(s,a)}(s,a)}$$
$$\lesssim\sqrt{H}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_{h+1}^k(s_{h+1}^k,a_{h+1}^k)-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k,a_{h+1}^k)\right)\mathbb{1}\left(Q_{h+1}^k(s_{h+1}^k,a_{h+1}^k)-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k,a_{h+1}^k)>1\right)}$$
$$\lesssim\sqrt{H^7SA\log\frac{SAT}{\delta}},$$

where the second inequality follows directly from (4.16), and the first inequality is valid since

$$V_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\le Q_{h+1}^k(s_{h+1}^k,a_{h+1}^k)-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k,a_{h+1}^k).$$

Substitution into (E.21) gives

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)}\lesssim\left(\sqrt{SA}\log T\right)\cdot\sqrt{H^7SA\log\frac{SAT}{\delta}}\asymp H^{7/2}SA\log^{3/2}\frac{SAT}{\delta},$$

thus concluding the proof.

E.3 *Bounding the term* $\mathscr{R}_3$

For notational convenience, we shall use the short-hand notation

$$k^i:=k_h^i(s_h^k,a_h^k)$$

whenever it is clear from the context. This allows us to decompose the expression of $\mathscr{R}_3$ in (4.22c) as follows:

$$\mathscr{R}_3:=\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^\star-V^{\mathrm{R}},k_{h+1}\right)}_{=:\mathscr{R}_3^1}$$
$$+\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\frac{\sum_{i\le N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V^{\mathrm{R}},k_{h+1}\right)}{N_h^k(s_h^k,a_h^k)}}_{=:\mathscr{R}_3^2}$$

with

$$\lambda_h^k:=\left(1+\frac{1}{H}\right)^{h-1}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\eta_{N_h^k(s_h^k,a_h^k)}^n\le\left(1+\frac{1}{H}\right)^h\le\left(1+\frac{1}{H}\right)^H\le e. \tag{E.23}$$

Here, the first inequality in (E.23) follows from the property $\sum_{N=n}^{\infty}\eta_n^N \le 1+1/H$ in Lemma 1, while the last inequality in (E.23) results from (E.1). In the sequel, we shall control each of these two terms separately.

Step 1: upper bounding $\mathscr{R}_3^1$. We plan to control this term by means of Lemma 8. For notational simplicity, let us define

$$N(s,a,h) := N_h^{K-1}(s,a)$$

and set

$$W_{h+1}^i := V^{\mathrm{R}}, k_{h+1} - V_{h+1}^\star \qquad \text{and} \qquad u_h^i(s_h^i,a_h^i) := \lambda_h^i = \left(1+\frac{1}{H}\right)^{h-1} \sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \eta_{N_h^i(s_h^i,a_h^i)}^n.$$

Given the fact that $V^{\mathrm{R}}, k_{h+1}(s), V_{h+1}^\star(s) \in [0,H]$ and the condition (E.23), it is readily seen that

$$\left|u_h^i(s_h^i,a_h^i)\right| \le e =: C_{\mathrm{u}} \qquad \text{and} \qquad \left\|W_{h+1}^i\right\|_\infty \le H =: C_{\mathrm{w}}.$$

Apply Lemma 8 to yield

$$\begin{aligned}
\left|\sum_{h=1}^H\sum_{k=1}^K \lambda_h^k\left(P_h^k - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\mathrm{R},k}\right)\right| &= \left|\sum_{h=1}^H\sum_{k=1}^K X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} HSA \sum_{h=1}^H\sum_{i=1}^K \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right]\log\frac{K}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}HSA\log\frac{K}{\delta} \\
&\lesssim \sqrt{H^2SA\sum_{h=1}^H\sum_{k=1}^K \mathbb{E}_{i,h-1}\left[P_h^k\left(V_{h+1}^{\mathrm{R},k} - V_{h+1}^\star\right)\right]\log\frac{T}{\delta}} + H^2SA\log\frac{T}{\delta} \\
&\asymp \sqrt{H^2SA\left\{\sum_{h=1}^H\sum_{k=1}^K P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R},k} - V_{h+1}^\star\right)\right\}\log\frac{T}{\delta}} + H^2SA\log\frac{T}{\delta} \qquad \text{(E.24)}
\end{aligned}$$

with probability at least $1-\delta/2$.

It then comes down to controlling the sum $\sum_{h=1}^H\sum_{k=1}^K P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R}}, k - V_{h+1}^\star\right)$. Toward this end, we first single out the following useful fact:

$$\begin{aligned}
\sum_{h=1}^H\sum_{k=1}^K P_h^k\left(V^{\mathrm{R}}, k_{h+1} - V_{h+1}^\star\right) &\overset{\text{(i)}}{\le} \sum_{h=1}^H\sum_{k=1}^K P_h^k\left(V_{h+1}^k + 2 - V_{h+1}^\star\right) \\
&\le 2HK + \sum_{h=1}^H\sum_{k=1}^K\left(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\star(s_{h+1}^k)\right) \overset{\text{(ii)}}{\lesssim} \sqrt{H^7SAK\log\frac{SAT}{\delta}} + H^3SA + HK \qquad \text{(E.25)}
\end{aligned}$$

with probability at least $1-\delta/4$, where (i) holds according to (4.15), and (ii) is valid since

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_{h+1}^{k}(s_{h+1}^{k})-V_{h+1}^{\star}(s_{h+1}^{k})\right)\le\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_{h+1}^{\mathsf{UCB},k}(s_{h+1}^{k})-V_{h+1}^{\pi^k}(s_{h+1}^{k})\right)$$
$$\lesssim\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA,$$

where the first inequality follows since $V_{h+1}^{\mathsf{UCB},k}\ge V_{h+1}^{k}$ and $V_{h+1}^{\star}\ge V_{h+1}^{\pi^k}$, and the second inequality comes from (E20a). Additionally, invoking Freedman's inequality (see Lemma 8) with $c_h=1$ and $\widetilde{W}_h^i=V_{h+1}^{\mathrm{R}},k-V_{h+1}^{\star}$ (so that $0\le\widetilde{W}_h^i(s)\le H$) directly leads to

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\right)\right|\lesssim\sqrt{TH^2\log\frac{1}{\delta}}+H\log\frac{1}{\delta}\asymp\sqrt{H^3K\log\frac{1}{\delta}}$$

with probability at least $1-\delta/4$, which taken collectively with (E.25) reveals that

$$\sum_{h=1}^{H}\sum_{k=1}^{K}P_{s_h^k,a_h^k,h}\left(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\right)\le\sum_{h=1}^{H}\sum_{k=1}^{K}P_h^k\left(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\right)+\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\left(P_h^k-P_{s_h^k,a_h^k,h}\right)\left(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\right)\right|$$
$$\lesssim\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK \tag{E.26}$$

with probability at least $1-\delta/2$. Substitution into (E.24) then gives

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\star}-V_{h+1}^{\mathrm{R},k}\right)\right|$$
$$\lesssim\sqrt{H^2SA\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\right)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\lesssim\sqrt{H^2SA\left(\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK\right)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\asymp\sqrt{H^2SA\left(H^6SA\log\frac{SAT}{\delta}+H^3SA+HK\right)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\lesssim\sqrt{H^3SAK\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta}$$
$$=\sqrt{H^2SAT\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta} \tag{E.27}$$

with probability exceeding $1-\delta$, where the third line holds since (due to Cauchy–Schwarz)

$$\sqrt{H^7SAK\log\frac{SAT}{\delta}}=\sqrt{H^6SA\log\frac{SAT}{\delta}}\sqrt{HK}\lesssim H^6SA\log\frac{SAT}{\delta}+HK.$$

Step 2: upper bounding $\mathscr{R}_3^2$. We start by making the following observation:

$$\begin{aligned}
\mathscr{R}_3^2 &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \frac{\lambda_h^k}{N_h^k(s_h^k,a_h^k)} \sum_{i\le N_h^k(s_h^k,a_h^k)} \big(V_{h+1}^{\mathrm{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k} V^{\mathrm{R}},k_{h+1}\big) \\
&= \sum_{h=1}^{H}\sum_{k=1}^{K} \sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\Big(V^{\mathrm{R}},k_{h+1}(s_{h+1}^k) - V^{\mathrm{R}},k_{h+1}(s_{h+1}^k) + \big(P_h^k - P_{h,s_h^k,a_h^k}\big)V^{\mathrm{R}},k_{h+1}\Big) \\
&\le (e\log T)\sum_{h=1}^{H}\sum_{k=1}^{K}\big(V^{\mathrm{R}},k_{h+1}(s_{h+1}^k) - V^{\mathrm{R}},k_{h+1}(s_{h+1}^k)\big) + \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)V_{h+1}^\star \\
&\quad + \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V^{\mathrm{R}},k_{h+1} - V_{h+1}^\star\big), \qquad \text{(E.28)}
\end{aligned}$$

where the first inequality comes from the monotonicity property $V^{\mathrm{R}},k_{h+1} \ge V_{h+1}^{\mathrm{R},k+1} \ge \cdots \ge V^{\mathrm{R}},k_{h+1}$, and the last line follows from the facts that $\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{1}{n} \le \log T$ and $\lambda_h^k \le e$ (cf. (E.23)). In what follows, we shall control the three terms in (E.28) separately.

- The first term in (E.28) can be controlled by Lemma 4 (cf. (4.16)) as follows:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\big(V^{\mathrm{R}},k_{h+1}(s_{h+1}^k) - V^{\mathrm{R}},k_{h+1}(s_{h+1}^k)\big) \lesssim H^6 SA\log\frac{SAT}{\delta} \qquad \text{(E.29)}$$

  with probability at least $1-\delta/3$.

- To control the second term in (E.28), we abuse the notation by setting

$$N(s,a,h) := N_h^{K-1}(s,a)$$

  and

$$W_{h+1}^i := V_{h+1}^\star, \qquad \text{and} \qquad u_h^i(s_h^i,a_h^i) := \sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \frac{\lambda_h^i}{n},$$

  which clearly satisfy

$$\big|u_h^i(s_h^i,a_h^i)\big| \le e\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \frac{1}{n} \le e\log T =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Here, we have used the properties $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^{K-1}(s_h^i,a_h^i)} \frac{1}{n} \le \log T$ and $\lambda_h^k \le e$ (cf. (E.23)). With these in place, applying Lemma 8 reveals that

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)V_{h+1}^\star\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{K} X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 HSA\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right]\log\frac{T}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}HSA\log\frac{T}{\delta} \\
&\overset{\text{(i)}}{\asymp} \sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\mathrm{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)\cdot HSA\log^3\frac{T}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{HSA\big(HT + H^4\sqrt{SAT}\big)\log^4\frac{SAT}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\lesssim \sqrt{HSA\big(HT + H^7SA\big)\log^4\frac{SAT}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta}
\end{aligned} \tag{E.30}$$

with probability at least $1-\delta/3$. Here, (i) comes from the definition in (4.6), (ii) holds due to (E.17) and (E.18) and (iii) is valid since

$$HT + H^4\sqrt{SAT} = HT + \sqrt{H^7SA}\cdot\sqrt{HT} \lesssim HT + H^7SA$$

due to the Cauchy–Schwarz inequality.

- Turning attention the third term of (E.28), we need to properly cope with the dependency between P_h^k and V_{h+1}^{R}, k. Toward this, we shall resort to the standard epsilon-net argument (see, e.g. [58]), which will be presented in Appendix E.3.1. The final bound reads like

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{R},K} - V_{h+1}^\star\big)\right| \lesssim H^4SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK\log^3\frac{SAT}{\delta}}. \tag{E.31}$$

- Combining (E.29), (E.30) and (E.31) with (E.28), we can use the union bound to demonstrate that

$$\mathscr{R}_3^2 \le C_{3,2}\left\{H^6SA\log^3\frac{SAT}{\delta} + \sqrt{H^2SAT\log^4\frac{SAT}{\delta}}\right\} \tag{E.32}$$

with probability at least $1-\delta$, where $C_{3,2} > 0$ is some constant.

Step 3: final bound of $\mathscr{R}_3$. Putting the above results (E.27) and (E.32) together, we immediately arrive at

$$\mathscr{R}_3 \le \left|\mathscr{R}_3^1\right| + \mathscr{R}_3^2 \le C_{\mathrm{r},3}\left\{H^6SA\log^3\frac{SAT}{\delta} + \sqrt{H^2SAT\log^4\frac{SAT}{\delta}}\right\} \tag{E.33}$$

with probability at least $1-2\delta$, where $C_{\mathrm{r},3} > 0$ is some constant. This immediately concludes the proof.

E.3.1 *Proof of (E31).* **Step 1: concentration bounds for a fixed group of vectors.** Consider a fixed group of vectors $\{V_{h+1}^{\mathrm{d}} \in \mathbb{R}^S \mid 1 \le h \le H\}$ obeying the following properties:

$$V_{h+1}^{\star} \le V_{h+1}^{\mathrm{d}} \le H \qquad \text{for} 1 \le h \le H. \tag{E.34}$$

We intend to control the following sum:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\left(P_h^k - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}\right).$$

To do so, we shall resort to Lemma 8. For the moment, let us take $N(s,a,h) := N_h^{K-1}(s,a)$ and

$$W_{h+1}^i := V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}, \qquad u_h^i(s_h^i,a_h^i) := \sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \frac{\lambda_h^i}{n}.$$

It is easily seen that

$$\left|u_h^i(s_h^i,a_h^i)\right| \le e\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \frac{1}{n} \le e\log T =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}},$$

which hold due to the facts $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^K(s_h^i,a_h^i)} \frac{1}{n} \le \log T$ and $\lambda_h^k \le e$ (cf. (E.23)) as well as the property that $V_{h+1}^{\mathrm{d}}(s), V_{h+1}^{\star}(s) \in [0,H]$. Thus, invoking Lemma 8 yields

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\left(P_h^k - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}\right)\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{K} X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}}\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right]\log\frac{K^{HSA}}{\delta_0}} + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{K^{HSA}}{\delta_0} \\
&\lesssim \sqrt{H\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{K^{HSA}}{\delta_0}} + H\left(\log T\right)\log\frac{K^{HSA}}{\delta_0}
\end{aligned} \tag{E.35}$$

with probability at least $1-\delta_0$, where the choice of δ_0 will be revealed momentarily.

Step 2: constructing and controlling an epsilon net. Our argument in Step 1 is only applicable to a fixed group of vectors. The next step is then to construct an epsilon net that allows one to cover the set

of interest. Specifically, let us construct an epsilon net $\mathscr{N}_{h+1,\alpha}$ (the value of α will be specified shortly) for each $h \in [H]$ such that:

a) for any $V_{h+1} \in [0,H]^S$, one can find a point $V_{h+1}^{\mathrm{net}} \in \mathscr{N}_{h+1,\alpha}$ obeying

$$0 \le V_{h+1}(s) - V_{h+1}^{\mathrm{net}}(s) \le \alpha \qquad \text{for all} s \in \mathscr{S};$$

b) its cardinality obeys

$$\left|\mathscr{N}_{h+1,\alpha}\right| \le \left(\frac{H}{\alpha}\right)^S. \tag{E.36}$$

Clearly, this also means that

$$\left|\mathscr{N}_{2,\alpha} \times \mathscr{N}_{3,\alpha} \times \cdots \times \mathscr{N}_{H+1,\alpha}\right| \le \left(\frac{H}{\alpha}\right)^{SH}.$$

Set $\delta_0 = \frac{1}{6}\delta/\left(\frac{H}{\alpha}\right)^{SH}$. Taking (E.35) together the union bound implies that: with probability at least $1 - \delta_0\left(\frac{H}{\alpha}\right)^{SH} = 1 - \delta/6$, one has

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\left(P_h^k - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\mathrm{net}} - V_{h+1}^{\star}\right)\right| \\
&\lesssim \sqrt{H\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{net}} - V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{K^{HSA}}{\delta_0}} + H(\log T)\log\frac{K^{HSA}}{\delta_0} \\
&\lesssim \sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{net}} - V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{SAT}{\delta\alpha}} + H^2SA\log^2\frac{SAT}{\delta\alpha}
\end{aligned} \tag{E.37}$$

simultaneously for all $\{V_{h+1}^{\mathrm{net}} \mid 1 \le h \le H\}$ obeying $V_{h+1}^{\mathrm{d}} \in \mathscr{N}_{h+1,\alpha}$ $(h \in [H])$.

Step 3: obtaining uniform bounds. We are now positioned to establish a uniform bound over the entire set of interest. Consider an arbitrary group of vectors $\{V_{h+1}^{\mathrm{u}} \in \mathbb{R}^S \mid 1 \le h \le H\}$ obeying (E.34). By construction, one can find a group of points $\left\{V_{h+1}^{\mathrm{net}} \in \mathscr{N}_{h+1,\alpha} \mid h \in [H]\right\}$ such that

$$0 \le V_{h+1}^{\mathrm{u}}(s) - V_{h+1}^{\mathrm{net}}(s) \le \alpha \qquad \text{for all}(h,s) \in \mathscr{S} \times [H]. \tag{E.38}$$

It is readily seen that

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big)\right|\\
&\le\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\Big(\big\|P_h^k\big\|_1+\big\|P_{h,s_h^k,a_h^k}\big\|_1\Big)\big\|V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big\|_\infty\right|\\
&\le 2eK\alpha\log T, \qquad \text{(E.39)}
\end{aligned}$$

where the last inequality follows from $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^{K-1}(s_h^i,a_h^i)}\frac{1}{n}\le\log T$ and $\lambda_h^k\le e$ (cf. (E.23)). Consequently, by taking $\alpha=1/(SAT)$, we can deduce that

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\star}\big)\right|\\
&\quad\le\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\big)\right|\\
&\qquad+\sum_{h=1}^{H}\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big)\right|\\
&\quad\lesssim\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\big)\right|+HK\alpha\log T\\
&\quad\lesssim\sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\big)\big(\log^2T\big)\log\frac{SAT}{\delta\alpha}}+H^2SA\log^2\frac{SAT}{\delta\alpha}+HK\alpha\log T\\
&\quad\asymp\sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\star}\big)\big(\log^2T\big)\log\frac{SAT}{\delta}}+H^2SA\log^2\frac{SAT}{\delta}, \qquad \text{(E.40)}
\end{aligned}$$

where the last line holds due to the condition (E.38) and our choice of α. To summarize, with probability exceeding $1-\delta/6$, the property (E.40) holds simultaneously for all $\{V_{h+1}^{\mathrm{u}}\in\mathbb{R}^S\mid 1\le h\le H\}$ obeying (E.34).

Step 4: controlling the original term of interest. With the above union bound in hand, we are ready to control the original term of interest

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V^{\mathrm{R}},k_{h+1}-V_{h+1}^{\star}\big). \tag{E.41}$$

To begin with, it can be easily verified using (4.10) that

$$V_{h+1}^{\star}\le V^{\mathrm{R}},k_{h+1}\le H \qquad \text{for all} 1\le h\le H. \tag{E.42}$$

Moreover, we make the observation that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathrm{R},K}-V_{h+1}^{\star}\big)&\overset{\text{(i)}}{\le}\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathrm{R},k}-V_{h+1}^{\star}\big)\\&\overset{\text{(ii)}}{\le}\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK\end{aligned} \tag{E.43}$$

with probability exceeding $1-\delta/6$, where (i) holds because V_{h+1}^{R} is monotonically non-increasing (in view of the monotonicity of $V_h(s)$ in (4.7b) and the update rule in line 16 of Algorithm 3), and (ii) follows from (E.26). Substitution into (E.40) yields

$$\begin{aligned}&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{R},K}-V_{h+1}^{\star}\big)\right|\\&\lesssim\sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K}P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{R},K}-V_{h+1}^{\star}\right)\left(\log^2T\right)\log\frac{SAT}{\delta}}+H^2SA\log^2\frac{SAT}{\delta}\\&\lesssim\sqrt{H^2SA\left\{\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK\right\}\left(\log^2T\right)\log\frac{SAT}{\delta}}+H^2SA\log^2\frac{SAT}{\delta}\\&\lesssim\sqrt{H^2SA\left\{H^6SA\log\frac{SAT}{\delta}+H^3SA+HK\right\}\log^3\frac{SAT}{\delta}}+H^2SA\log^2\frac{SAT}{\delta}\\&\lesssim H^4SA\log^2\frac{SAT}{\delta}+\sqrt{H^3SAK\log^3\frac{SAT}{\delta}},\end{aligned} \tag{E.44}$$

where the penultimate line holds since

$$\sqrt{H^7SAK\log\frac{SAT}{\delta}}=\sqrt{H^6SA\log\frac{SAT}{\delta}}\sqrt{HK}\lesssim H^6SA\log\frac{SAT}{\delta}+HK.$$