A Refined Analysis of UCBVI

Simone Drago SIMONE.DRAGO@POLIMI.IT
Marco Mussi MARCO.MUSSI@POLIMI.IT
Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano
Piazza Leonardo da Vinci 32, Milan, 20133, Italy

Abstract

In this work, we provide a refined analysis of the UCBVI algorithm (Azar et al., 2017), improving both the bonus terms and the regret analysis. Additionally, we compare our version of UCBVI with both its original version and the state-of-the-art MVP algorithm. Our empirical validation demonstrates that improving the multiplicative constants in the bounds has significant positive effects on the empirical performance of the algorithms.

1 Introduction

We focus on the problem of *finite-horizon tabular RL*, where the statistical complexity is characterized by a well-established lower bound of the order $\Omega(\sqrt{HSAT})$, where S is the number of states, A is the number of actions, H is the horizon of the episode, and $T = HK$ where K is the number of episodes (Domingues et al., 2021). State-of-the-art learning algorithms match the lower bound up to logarithmic factors. UCBVI (Azar et al., 2017) matches this lower bound by combining the classical value iteration approach with the *optimism in the face of uncertainty* mechanism, achieving $\widetilde{\mathcal{O}}(\sqrt{HSAT})$ under the condition $T \geqslant \Omega(H^3S^3A)$. Recently, MVP (Zhang et al., 2024) overcome this limitation by employing the so-called *doubling trick*, ensuring order-optimal regret (up to logarithmic factors) for every time horizon. However, this improvement comes at the expense of significantly larger constant factors, which may imply poor empirical performances (see Section 5). Consequently, UCBVI remains a competitive and practical solution for finite-horizon tabular RL.

In this work, we improve both the bound and the analysis of the UCBVI algorithm, with particular attention to its advanced form with Bernstein-Freedman optimistic bonus. Our goal is to design an exploration bonus that is as tight as possible and to conduct a regret upper bound analysis that minimizes also constants.

2 Setting

In this section, we introduce the notation and the setting we consider in the rest of the work.

Notation. Given a measurable set $\mathcal{X}$, we denote with $\Delta(\mathcal{X})$ the set of probability measures over $\mathcal{X}$, and with $|\mathcal{X}|$ it cardinality. For $n \in \mathbb{N}$, we denote the set $\{1, \dots, n\}$ as $[\![n]\!]$. We denote the L1 norm of a vector as $\|\cdot\|_1$. We denote the indicator function of event x as $\mathbb{I}\{x\}$.

Markov Decision Processes. An undiscounted, episodic Markov Decision Process (MDP, Puterman, 1990) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, R, H)$. In this tuple, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the state transition probability, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward function, and $H \in \mathbb{N}_{\geqslant 1}$ is the length of each episode.[1]

We assume the state space and the action space are finite sets, and we denote their cardinalities as $|\mathcal{S}| = S < +\infty$ and $|\mathcal{A}| = A < +\infty$. We assume the state transition probability and the reward do not depend on the stage. Moreover, we assume the reward to be deterministic, known, and bounded in $[0, 1]$.

Interaction with the Environment. The agent interacts with the environment in a sequence of K episodes. Denote as $x_{k,h}$ the state occupied by the agent at stage $h \in [\![H]\!]$ of episode $k \in [\![K]\!]$, and as $a_{k,h}^{\pi_k}$ the action played by the agent at stage h of episode k according to the policy π_k. We assume policies to be deterministic and stage-dependent, i.e., $\pi : \mathcal{S} \times [\![H]\!] \to \mathcal{A}$.

The interaction of the k-th episode is defined as follows. Starting from state $x_{k,1} \in \mathcal{S}$, the agent selects which action to play as $a_{k,h}^{\pi_k} = \pi_k(x_{k,h}, h)$ for every $h \in [\![H]\!]$, and observes a sequence of next-states and rewards, until the end of the episode.

The function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ denotes the value function at stage $h \in [\![H]\!]$, such that $V_h^\pi(x)$ represents the expected sum of the $H - h$ returns received under policy π starting from state $x \in \mathcal{S}$. Under the assumptions stated above, there exists a deterministic policy π^* which attains the best possible value function $V_h^*(x) := \sup_\pi V_h^\pi(x)$ for every state $x \in \mathcal{S}$. We measure the performance of a learning algorithm $\mathfrak{A}$ after K episodes by means of the *cumulative regret*:

$$\mathrm{Reg}(\mathfrak{A}, K) := \sum_{i=1}^{K} V_1^*(x_{i,1}) - V_1^{\pi_i}(x_{i,1}).$$

We denote as $T = KH$ the total number of interactions.

3 UCBVI

In this section, we consider the UCBVI algorithm, introduced in (Azar et al., 2017). First, we provide a more compact (but equivalent) pseudocode of the algorithm in Algorithm 1.

We start by initializing the counters needed in order to run the algorithm. Then, we start the continuous interaction procedure for every episode $k \in [\![K]\!]$. For every episode, before starting the interaction, the algorithm computes the optimistic estimate of the value function. Such estimate starts by computing all the transition probabilities $\widehat{P}_k(y|x, a)$ for every state, action and next state as:

$$\widehat{P}_k(y|x, a) = \frac{N_k(x, a, y)}{\max\{1, N_k(x, a)\}}, \tag{1}$$

where $N_k(x, a)$ is the number of times we play action $a \in \mathcal{A}$ in state $x \in \mathcal{S}$, and $N_k(x, a, y)$ is the number of times we do so and observe the next state $y \in \mathcal{S}$. Then, we compute the optimistic value iteration for finite horizon MDPs, starting from stage H backward, where the optimistic $Q_{k,h}(x, a)$ is defined as:

$$Q_{k,h}(x, a) = \min\{Q_{k-1,h}(x, a), R(x, a) + \sum_{y \in \mathcal{S}} \widehat{P}_k(y|x, a) V_{k,h+1}(y) + b_{k,h}(x, a)\}.$$

[1] Let $x, y \in \mathcal{S}$ and $a \in \mathcal{A}$, we denote as $P(y|x, a)$ the probability of observing y as the next state after playing action a in state x, and $R(x, a)$ the reward obtained after playing action a in state x.

Algorithm 1: UCBVI.

```
Initialize: $N_0(x,a,y) = 0,\ N_0(x,a) = 0,\ N'_{0,h}(x) = 0,\ \forall (x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$
            $Q_{0,h}(x,a) = H - h + 1,\ \forall (x,a,h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$
for $k \in \llbracket K \rrbracket$ do
    // Update the optimistic estimates for episode $k$
    Estimate $\widehat{P}_k(y|x,a) = N_k(x,a,y)/\max\{1, N_k(x,a)\}$
    Initialize $V_{k,H+1}(x) = 0,\ \forall x \in \mathcal{S}$
    for $h = \{H, H-1, \ldots, 1\}$ do
        for $x \in \mathcal{S}$ do
            for $a \in \mathcal{A}$ do
                $Q_{k,h}(x,a) = \min\{Q_{k-1,h}(x,a), R(x,a) + \sum_{y \in \mathcal{S}} \widehat{P}_k(y|x,a) V_{k,h+1}(y) + b_{k,h}(x,a)\}$
            end
            $V_{k,h}(x) = \max_{a \in \mathcal{A}} Q_{k,h}(x,a)$
        end
    end
    // Interact with the environment for episode $k$
    Agent observes state $x_{k,1}$
    for $h \in \llbracket H \rrbracket$ do
        Agent plays action $a_{k,h} \in \arg\max_{a \in \mathcal{A}} Q_{k,h}(x_{k,h}, a)$
        Environment returns reward $r_{k,h}$ and next state $x_{k,h+1}$
        Update for every $(x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:
                $N_k(x,a,y) = N_{k-1}(x,a,y) + \mathbb{1}\{x = x_{k,h}, a = a_{k,h}, y = x_{k,h+1}\}$
                $N_k(x,a) = N_{k-1}(x,a) + \mathbb{1}\{x = x_{k,h}, a = a_{k,h}\}$
                $N'_{k,h}(x) = N'_{k-1,h}(x) + \mathbb{1}\{x = x_{k,h}\}$
    end
end
```

This procedure mimics value iteration with an additive exploration term $b_{k,h}(x,a)$ which will be further characterized later in Section 4. Then, term $V_{k,h}(x)$ is computed as usual for value iteration:

$$V_{k,h}(x) = \max_{a \in \mathcal{A}} Q_{k,h}(x,a).$$

After that, we can start interacting with the environment and play greedily. More in detail, at every stage $h \in \llbracket H \rrbracket$ of episode $k \in \llbracket K \rrbracket$, we observe state $x_{k,h}$ and play the most promising action according to the optimistic estimate:

$$a_{k,h} \in \arg\max_{a \in \mathcal{A}} Q_{k,h}(x_{k,h}, a).$$

After having played $a_{k,h}$, we observe the reward $r_{k,h}$ and the next state $x_{k,h+1}$. Finally, we use the collected information to properly update counters.

4 A Refined UCBVI Analysis

In this section, we analyze the UCBVI algorithm. We start in Section 4.1 by providing the regret upper bound for UCBVI with *Chernoff-Hoeffding* bonus (Theorem 4.1). Then, we consider the more elaborated *Bernstein-Freedman* bonus, and we characterize the regret of `UCBVI` we get when we consider such a bonus (Theorem 4.2). Our contribution consists of refining the analysis to obtain tighter bonuses and, as a direct consequence, tighter regret bounds.

4.1 Chernoff-Hoeffding

Let us start with the *Chernoff-Hoeffding* bound, taking the opportunity to correct some typos of the original analysis of (Azar et al., 2017).

Theorem 4.1 (Regret for `UCBVI` with Chernoff-Hoeffding bonus) *Let $\delta \in (0,1)$. Considering:*[2]

$$b_{k,h}(x,a) = \frac{2HL}{\sqrt{\max\{N_k(x,a),1\}}},$$

then, w.p. at least $1-\delta$, the regret of `UCBVI-CH` is bounded by:

$$\mathrm{Reg}(\texttt{UCBVI-CH}, K) \leqslant 10eHL\sqrt{SAT} + \frac{8}{3}eH^2S^2AL^2,$$

where $L = \ln(5HSAT/\delta)$. For $T \geqslant \Omega(H^2S^3A)$, this bound translates to $\widetilde{\mathcal{O}}(H\sqrt{SAT})$.

Theorem 4.1 should be compared to Theorem 1 of (Azar et al., 2017). Since the analysis is a refinement of the original analysis in terms of constants, the order of the regret does not change between the two theorems. However, our analysis provides a smaller value for the constants.[3] Moreover, observe how the minimum value of T for which the regret bound holds according to our analysis is H times higher than the one reported in the original theorem. This is due to the fact that the minimum T in the statement of Theorem 1 of (Azar et al., 2017) is incorrect, although the derivation in the appendix provides the same minimum value of T we obtain.

4.2 Bernstein-Freedman

We now move to the *Bernstein-Freedman* bonus.

Theorem 4.2 (Regret for `UCBVI` with Bernstein-Freedman bonus) *Let $\delta \in (0,1)$. Considering:*[2]

$$b_{k,h}(x,a) = \underbrace{\sqrt{\frac{4L\,\mathbb{V}\mathrm{ar}_{y\sim\widehat{P}_k(\cdot|x,a)}(V_{k,h+1}(y))}{\max\{N_k(x,a),1\}}}}_{\texttt{(A)}} + \underbrace{\frac{7HL}{3\max\{N_k(x,a)-1,1\}}}_{\texttt{(B)}}$$
$$+ \underbrace{\sqrt{\frac{4\sum_{y\in\mathcal{S}}\left(\widehat{P}(y|x,a)\cdot\min\left\{\frac{84^2H^3S^2AL^2}{\max\{1,N'_{k,h+1}(y)\}}, H^2\right\}\right)}{\max\{N_k(x,a),1\}}}}_{\texttt{(C)}},$$

[2] We assume that, by definition, $b_{k,H}(s,a) = 0$, as at the last stage, there is no need for exploration, and the rewards are deterministic.

[3] To the best of the authors' knowledge, the original analysis of (Azar et al., 2017) is missing a multiplicative e factor in the regret bound.

then, w.p. at least $1-\delta$, the regret of `UCBVI-BF` *is bounded by:*

$$\mathrm{Reg}(\texttt{UCBVI-BF}, K) \leqslant 24eL\sqrt{HSAT} + 616eH^2S^2AL^2 + 4e\sqrt{H^2TL},$$

where $L = \ln(5HSAT/\delta)$. For $T \geqslant \Omega(H^3S^3A)$ and $SA \geqslant H$, this bound translates to $\widetilde{\mathcal{O}}(\sqrt{HSAT})$.

Theorem 4.2 should be compared to Theorem 2 of (Azar et al., 2017). Again, as the analysis is a refinement in terms of constants, the order of the regret does not change. Moreover, also the minimum value of T under which the regret bound matches the lower bound in unchanged between the two analyses. It is important to notice, however, that the constant terms of our analysis are strictly smaller than the ones of (Azar et al., 2017).[3] In particular, comparing the bonus term $b_{k,h}$ presented in Theorem 4.2 w.r.t. the one of (Azar et al., 2017, Algorithm 4) we observe the following differences:

- in term `(A)`, we have a $\sqrt{4}$ multiplicative factor instead of $\sqrt{8}$;
- in term `(B)`, we have a 7 multiplicative factor instead of 14;
- in term `(C)`, we have a $\sqrt{4}$ multiplicative factor instead of $\sqrt{8}$ and, inside the minimum, a term multiplicative factor 84^2 instead of 100^2.

The reader shall refer to Appendices D and E for the proofs of Theorems 4.1 and 4.2, respectively. The derivations provided in the appendices closely follow the proofs of (Azar et al., 2017), focusing on lowering the constant terms. A full description of the notation employed throughout the paper is reported in Appendix A. Both proofs are conducted under the condition that concentration inequalities hold for the next state estimator and its variance. Those conditions fall under event $\mathcal{E}$, which is presented in Appendix B.4 of (Azar et al., 2017) and restated in Appendix B. Finally, additional lemmas necessary to show the regret decomposition and to bound the summation of the term it comprises are demonstrated in Appendix C.

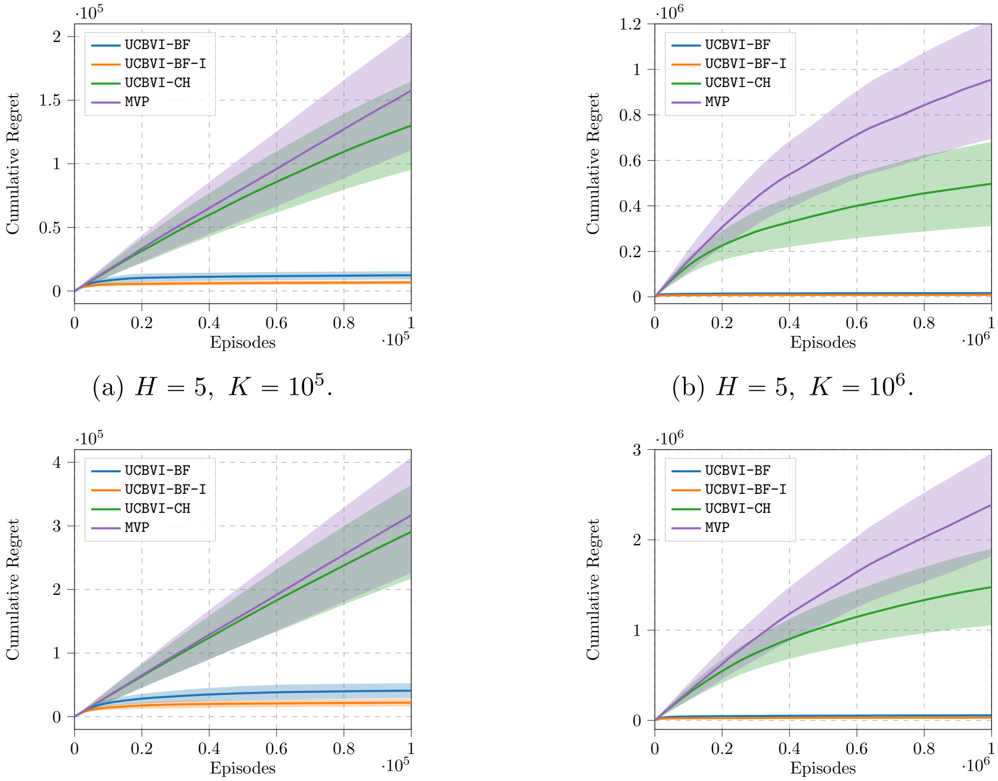
5 Numerical Validation

In this section, we numerically compare the performances of `UCBVI`, both with the Chernoff-Hoeffding and Bernstein-Freedman bonuses of (Azar et al., 2017) and with the improved Bernstein-Freedman bonus of this paper, against the `MVP` (Zhang et al., 2024) algorithm.

In order to fairly compare to the `MVP` algorithm, all the $N_h(x,a)$ terms are considered as $N(x,a)$, removing the discriminant of the stage from the algorithm, and the c_2 constant (which refers to the uncertainty in the estimation of the rewards) is set to 0, to remove the exploration factor needed due to the stochasticity of the reward in the original paper.

5.1 Illustrative Environments

As a first experimental evaluation, we consider a set of illustrative environments. We consider an MDP with parameters $S = 3$, $A = 3$, $H \in \{5, 10\}$, and we consider a number of episodes $K \in \{10^5, 10^6\}$. We evaluate each experiment by averaging over 10 runs. In each run, the rewards and transition probabilities of the MDP are randomly generated. Then, the clairvoyant optimum is calculated for the purpose of regret computation, and the algorithms are evaluated.



(a) $H = 5,\ K = 10^5$.

(b) $H = 5,\ K = 10^6$.

(c) $H = 10,\ K = 10^5$.

(d) $H = 10,\ K = 10^6$.

Figure 1: Performances in terms of cumulative regret in toy environments with $S = 3$ states and $A = 3$ actions (10 runs, mean ± 95% C.I.).

Results. Figure 1 represents the cumulative regret of the evaluated algorithms in the first experimental evaluation for different values of H and K. From the figures, we can observe that `UCBVI` with the Chernoff-Hoeffding bonus and `MVP` begin to show a sub-linear regret for $K = 10^6$, whereas both versions of `UCBVI` with the Bernstein-Freedman bonus greatly outperform the other algorithms in all the evaluated scenarios. In particular, the use of a tighter Bernstein-Freedman bonus (`UCBVI-BF-I`) translates into a cumulative regret that is, although of the same order, lower than with the usage of a larger bonus (`UCBVI-BF`), highlighting the importance of lower order terms and constants in empirical performance.

5.2 RiverSwim

We now consider the RiverSwim environment (Strehl and Littman, 2008). This environment emulates a swimmer that has to swim against the current, where the agent has 2 options: *(i)* to try to swim to the other side or *(ii)* to turn back. In this scenario, the rewards and the transition probabilities are designed such that the optimal policy corresponds to trying to swim and reach the other side of the "river". This is considered a challenging benchmark for exploration. We consider the scenario with $S = 5$ and $H = 10$. The reward model and the transition probability are designed such that the suboptimality gap between the optimal

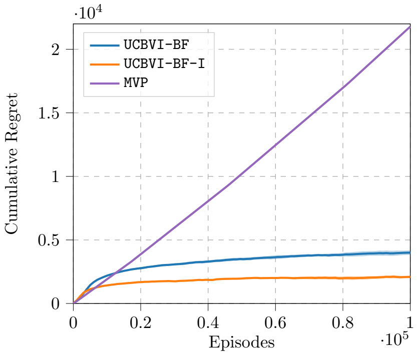


Figure 2: Performances in terms of cumulative regret in the RiverSwim environment with $S = 5$ states and horizon $H = 10$ (4 runs, mean ± 95% C.I.).

	Bonus ratio	Regret upper bound ratio	Empirical regret ratio
CH	$7/2$	2	-
BF	$\sqrt{2}$	$5/4$	1.87 ± 0.03

Table 1: Improvement ratios in the bonuses, regret upper bounds, and empirical regret between our analysis and the original of (Azar et al., 2017).

action and the other one in the initial state is very low ($\sim$0.1, with a scale of the problem in the order of $H = 10$).

Results. Figure 2 compares the results when using `MVP` and `UCBVI` in its original version (`UCBVI-BF`) and the one we propose with tighter bounds (`UCBVI-BF-I`). `MVP` confirms its poor empirical performance, failing to deliver a sublinear trend for the considered horizon. Instead, `UCBVI`, in both versions, shows a clear sublinear trend, with the improved version (`UCBVI-BF-I`) with a cumulative regret approximately half of the original one (`UCBVI-BF`).

6 Conclusions

A summary of the improvements, expressed in terms of improvement ratios in the bonuses, regret upper bounds, and empirical regret, is reported in Table 1. First, we compare our versions of the `UCBVI` algorithms with the original ones from (Azar et al., 2017). The algorithmic structure remains the same, though we re-derived the bonus terms to make them as tight as possible, resulting in an improvement of $7/2$ and $\sqrt{2}$ for the Chernoff-Hoeffding and Bernstein-Freedman bonuses, in the dominant terms, respectively. This reduction in over-exploration has significant empirical effects, as shown in Section 5, where, as reported in Table 1, we achieve an improvement in the empirical regret of 1.87 times. Additionally, this impacts the regret analysis, where we were able to reduce the regret bound by a factor of 2 and $5/4$ for the Chernoff-Hoeffding and Bernstein-Freedman bonuses, respectively, in terms of dominant terms. However, lower order terms also have an impact on the performance, and through a refined analysis we were able to reduce them by a factor of $\sim$90 and $\sim$4 for the Chernoff-Hoeffding and Bernstein-Freedman bonuses, respectively.

References

Mohammad G. Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability Second Edition*. Chapman and Hall/CRC, 2019.

Nicoló Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Omar D. Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Annual Conference on Learning Theory (COLT)*, 2009.

Martin L. Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

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

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125, 2003.

Zihan Zhang, Yuxin Chen, Jason D. Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. In *Annual Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pages 5213–5219. PMLR, 2024.

Appendix A. Notation

In this section, we collect the notation used throughout the main paper and the appendices.

Symbol	**Meaning**
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
P	Transition distribution
R	Reward function
H	Length of the episode
K	Total number of episodes
T	Total number of steps
T_k	Total number of steps up to episode k
S	Cardinality of the state space
A	Cardinality of the action space
$x_{k,h}$	State occupied at stage h of episode k
$a^{\pi}_{k,h}$	Action played at stage h of episode k under policy π
$R^{\pi}(x)$	Reward obtained by playing according to policy π in state x
$N_k(x,a)$	Number of visits to state-action pair (x,a) up to episode k
$N_k(x,a,y)$	Number of transitions to state y from x after playing a, up to episode k
$N'_{k,h}(x)$	Number of visits to state x at stage h up to episode k
$\widehat{P}_k$	Estimated transition distribution
$b_{k,h}$	Exploration bonus
$b'_{k,h}(x)$	$\min\{\frac{84^2H^3S^2AL^2}{N'_{k,h}(x)}, H^2\}$
π_k	Policy played during episode k
π^*	Optimal policy
$Q_{k,h}$	Optimistic state-action value function
V^*_h	Value function of the optimal policy at stage h
V^{π}_h	Value function under policy π at stage h
$V_{k,h}$	Optimistic estimator of the optimal value function at stage h of episode k
$\Delta_{k,h}(x)$	Regret in state x, at stage h of episode k, following policy π_k
$\widetilde{\Delta}_{k,h}(x)$	Pseudo-regret in state x, at stage h of episode k, following policy π_k
$\mathrm{Reg}(\texttt{UCBVI-CH}, k)$	Regret of `UCBVI` using Chernoff-Hoeffding bonus after k episodes
$\widetilde{\mathrm{Reg}}(\texttt{UCBVI-CH}, k)$	Pseudo-regret of `UCBVI` using Chernoff-Hoeffding bonus after k episodes
$\mathrm{Reg}(\texttt{UCBVI-BF}, k)$	Regret of `UCBVI` using Bernstein-Freedman bonus after k episodes
$\widetilde{\mathrm{Reg}}(\texttt{UCBVI-BF}, k)$	Pseudo-regret of `UCBVI` using Bernstein-Freedman bonus after k episodes
$\mathcal{E}$	Concentration inequalities event
$\Omega, \Omega_{k,h}$	Optimism events
ε, $\overline{\varepsilon}$	Martingale Difference Sequences
$[k]_{\mathrm{typ}}$, $[k]_{\mathrm{typ},x}$	Sets of typical episodes
$\mathcal{H}_{k,h}$	History of the interactions up to, and including, stage h of episode k
L	Logarithmic term $\ln(5HSAT/\delta)$
$\mathbb{V}^{\pi_k}_h(x,a)$	Next-state variance of V^{π_k}
$\mathbb{V}^*_h$	Next-state variance of V^*
$\widehat{\mathbb{V}}_{k,h}$	Empirical next-state variance of $V_{k,h}$
$\widehat{\mathbb{V}}^*_{k,h}$	Empirical next-state variance of V^*
$\xi_{k,j}(x,a)$	State-action wise model error $\xi_{k,j}(x,a) := \sum_{y\in\mathcal{S}}[\widehat{P}_k(y\vert x,a) - P(y\vert x,a)]V^*_{h+1}(y)$

Table 2: Table of notation.

Let us now state the definitions of the next-state variances employed in the analysis. We define the empirical next state variance of V as:

$$\widehat{\mathbb{V}}_{k,h+1}(x,a) \coloneqq \underset{y \sim \widehat{P}_k(\cdot|x,a)}{\mathbb{V}\mathrm{ar}} [V_{k,h+1}(y)].$$

We define the next state variance of V^* as:

$$\mathbb{V}^*_{h+1}(x,a) \coloneqq \underset{y \sim P(\cdot|x,a)}{\mathbb{V}\mathrm{ar}} [V^*_{h+1}(y)].$$

We define the next state variance of V^π as:

$$\mathbb{V}^\pi_{h+1}(x,a) \coloneqq \underset{y \sim P(\cdot|x,a)}{\mathbb{V}\mathrm{ar}} [V^\pi_{h+1}(y)].$$

Finally, we define the empirical next state variance of V^* as:

$$\widehat{\mathbb{V}}^*_{k,h+1}(x,a) \coloneqq \underset{y \sim \widehat{P}_k(\cdot|x,a)}{\mathbb{V}\mathrm{ar}} [V^*_{h+1}(y)].$$

Let us now state the definition of the Martingale Difference Sequences employed in the analysis:

$$\begin{aligned}
\varepsilon_{k,h} &\coloneqq P(\cdot|x_{k,h}, a^{\pi_k}_{k,h})^\top \widetilde{\Delta}_{k,h+1}(\cdot) - \widetilde{\Delta}_{k,h+1}(x_{k,h+1}), \\
\overline{\varepsilon}_{k,h} &\coloneqq \sum_{y \in \mathcal{S}} P(y|x_{k,h}, a^{\pi_k}_{k,h}) \sqrt{\frac{\mathbb{I}(y \in [y]_{k,h})}{N_k(x_{k,h}, a^{\pi_k}_{k,h}) P(y|x_{k,h}, a^{\pi_k}_{k,h})}} \widetilde{\Delta}_{k,h+1}(y) \\
&\quad - \sqrt{\frac{\mathbb{I}(x_{k,h+1} \in [y]_{k,h})}{N_k(x_{k,h}, a^{\pi_k}_{k,h}) P(x_{k,h+1}|x_{k,h}, a^{\pi_k}_{k,h})}} \widetilde{\Delta}_{k,h+1}(x_{k,h+1}).
\end{aligned}$$

Appendix B. High Probability Events

In this section, we restate the high probability event $\mathcal{E}$ under which the concentration inequalities hold, presented in Appendix B.4 of (Azar et al., 2017).
Event $\mathcal{E}$ is defined as:

$$\begin{aligned}
\mathcal{E} \coloneqq \mathcal{E}_{\widehat{P}} \bigcap \bigcap_{\substack{k \in [\![K]\!] \\ h \in [\![H]\!] \\ x \in \mathcal{S}}} \Bigg[& \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}_{\widetilde{\Delta},k,h}, H, L \right) \bigcap \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}'_{\widetilde{\Delta},k,h}, \frac{1}{\sqrt{L}}, L \right) \bigcap \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, L \right) \\
& \bigcap \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}'_{\widetilde{\Delta},k,h,x}, \frac{1}{\sqrt{L}}, L \right) \bigcap \mathcal{E}_{\mathrm{fr}} \left(\mathcal{G}_{\mathbb{V},k,h}, H^4 T, H^3, L \right) \\
& \bigcap \mathcal{E}_{\mathrm{fr}} \left(\mathcal{G}_{\mathbb{V},k,h,x}, H^5 N'_{k,h}(x), H^3, L \right) \bigcap \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}_{b',k,h}, H^2, L \right) \\
& \bigcap \mathcal{E}_{\mathrm{az}} \left(\mathcal{F}_{b',k,h,x}, H^2, L \right) \Bigg]
\end{aligned}$$

We refer the reader to Lemma 1 of (Azar et al., 2017) for the proof that event $\mathcal{E}$ holds with high probability. Let, for ease of reading $\overline{x} = x_{i,j}$, $\overline{x}' = x_{i,j+1}$, and $\overline{a} = a_{i,j}^{\pi_i}$ We now restate the definition of the events that compose $\mathcal{E}$:

$$\mathcal{E}_{\widehat{P}} := \left\{ \widehat{P}_k(y|x,a) \in \mathcal{P}(k,h,N_k(x,a),x,a,y), \forall k \in [\![K]\!], h \in [\![H]\!], (x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \right\},$$

where $\mathcal{P}(k,h,n,x,a,y)$ is defined as the subset of the set of all probability distributions $\mathcal{P}$ over $\mathcal{S}$ such that:

$$\mathcal{P}(k,h,n,x,a,y) := \Bigg\{ \widetilde{P}(\cdot|x,a) \in \mathcal{P} : \|\widetilde{P}(\cdot|x,a) - P(\cdot|x,a)\|_1 \leqslant 2\sqrt{\frac{SL}{n}}, \tag{2}$$

$$\begin{aligned} &\left| \sum_{y\in\mathcal{S}} (\widetilde{P}(y|x,a) - P(y|x,a)) V_h^*(y) \right| \\ &\leqslant \min\left(\sqrt{\frac{2\widehat{\mathbb{V}}_{k,h+1}^*(x,a)L}{n}} + \frac{7HL}{3(n-1)}, \sqrt{\frac{2\mathbb{V}_{h+1}^*(x,a)L}{n}} + \frac{2HL}{3n} \right), \end{aligned} \tag{3}$$

$$\left| \widetilde{P}(y|x,a) - P(y|x,a) \right| \leqslant \sqrt{\frac{2P(y|x,a)(1-P(y|x,a))L}{n}} + \frac{2L}{3n} \Bigg\}, \tag{4}$$

where Equation (2) follows by applying the result of Theorem 2.1 of Weissman et al. (2003), Equation (3) follows by applying both Bernstein's inequality (see, e.g., Cesa-Bianchi and Lugosi, 2006) and the empirical Bernstein inequality (Maurer and Pontil, 2009), and Equation (4) follows by applying Lemma C.1.

The remaining events concern the summation of Martingale difference sequences:

$$\begin{aligned} \mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{\widetilde{\Delta},k,h}, H, L\right) := \Bigg\{ &\sum_{i=1}^{k} \sum_{j=h}^{H-1} \left[\sum_{y\in\mathcal{S}} P(y|\overline{x},\overline{a}) \widetilde{\Delta}_{i,j+1}(y) - \widetilde{\Delta}_{i,j+1}(\overline{x}') \right] \\ &\leqslant 2\sqrt{k(H-h)H^2L} \Bigg\}, \end{aligned}$$

$$\begin{aligned} \mathcal{E}_{\mathrm{az}}\left(\mathcal{F}'_{\widetilde{\Delta},k,h}, \frac{1}{\sqrt{L}}, L\right) := \Bigg\{ &\sum_{i=1}^{k} \sum_{j=h}^{H} \Bigg[\sum_{y\in\mathcal{S}} P(y|\overline{x},\overline{a}) \sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(\overline{x},\overline{a}) P(y|\overline{x},\overline{a})}} \widetilde{\Delta}_{i,j+1}(y) \\ &- \sqrt{\frac{\mathbb{I}(\overline{x}' \in [y]_{i,j})}{N_i(\overline{x},\overline{a}) P(\overline{x}'|\overline{x},\overline{a})}} \widetilde{\Delta}_{i,j+1}(\overline{x}') \Bigg] \\ &\leqslant 2\sqrt{k(H-h)\frac{1}{\sqrt{L}^2}L} \Bigg\}, \end{aligned}$$

$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, L\right) := \Bigg\{ \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \left[\sum_{y\in\mathcal{S}} P(y|\overline{x},\overline{a}) \widetilde{\Delta}_{i,j+1}(y) - \widetilde{\Delta}_{i,j+1}(\overline{x}') \right]$$

$$\leqslant 2\sqrt{N'_{k,h}(x)(H-h)H^2L}\Bigg\},$$

$$\begin{aligned}\mathcal{E}_{\text{az}}\left(\mathcal{F}'_{\widetilde{\Delta},k,h,x},\frac{1}{\sqrt{L}},L\right) := \Bigg\{ \sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H}\Bigg[\Bigg[\sum_{y\in\mathcal{S}}P(y|\overline{x},\overline{a})\sqrt{\frac{\mathbb{I}(y\in[y]_{i,j})}{N_i(\overline{x},\overline{a})P(y|\overline{x},\overline{a})}}\widetilde{\Delta}_{i,j+1}(y)\Bigg] \\ -\sqrt{\frac{\mathbb{I}(\overline{x}'\in[y]_{i,j})}{N_i(\overline{x},\overline{a})P(y|\overline{x},\overline{a})}}\widetilde{\Delta}_{i,j+1}(\overline{x}')\Bigg] \\ \leqslant 2\sqrt{N'_{k,h}(x)(H-h)\frac{1}{\sqrt{L}^2}L}\Bigg\},\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\text{fr}}\left(\mathcal{G}_{\mathbb{V},k,h},H^4T,H^3,L\right) := \Bigg\{\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(\overline{x},\overline{a})|\mathcal{H}_{i,h}\right]-\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(\overline{x},\overline{a})\Bigg\} \\ \leqslant 2\sqrt{H^4T_kL}+\frac{4H^3L}{3},\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\text{fr}}\left(\mathcal{G}_{\mathbb{V},k,h,x},H^5N'_{k,h}(x),H^3,L\right) := \Bigg\{\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(\overline{x},\overline{a})|\mathcal{H}_{i,h}\right] \\ -\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(\overline{x},\overline{a})\Bigg\} \\ \leqslant 2\sqrt{H^5N'_{k,h}(x)L}+\frac{4H^3L}{3},\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\text{az}}\left(\mathcal{F}_{b',k,h},H^2,L\right) := \left\{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left[\sum_{y\in\mathcal{S}}P(y|\overline{x},\overline{a})b'_{i,j+1}(y)-b'_{i,j+1}(\overline{x}')\right]\right\} \\ \leqslant 2\sqrt{(H-h)H^3T_kL}\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\text{az}}\left(\mathcal{F}_{b',k,h,x},H^2,L\right) := \left\{\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\left[\sum_{y\in\mathcal{S}}P(y|\overline{x},\overline{a})b'_{i,j+1}(y)-b'_{i,j+1}(\overline{x}')\right]\right\} \\ \leqslant 2\sqrt{N'_{k,h}(x)(H-h)H^4L}\end{aligned}$$

Appendix C. Technical Lemmas

Lemma C.1 (Bernstein inequality for Bernoulli random variables) *Let p be the parameter of a Bernoulli random variable, and let $\widehat{p}$ be its estimator. Let $\delta > 0$. Then, w.p. at least $1 - \delta$, it holds that:*

$$|\widehat{p} - p| \leqslant \sqrt{\frac{2p(1-p)L}{n}} + \frac{2L}{3n},$$

where n represents the number of observations, and $L = \ln(2/\delta)$.

Proof Let $\{Y_i\}_{i=1\dots,n}$ be the set of i.i.d. realizations of a Bernoulli with parameter p. Define the auxiliary random variable:

$$X_i = \frac{Y_i}{n}.$$

Observe that $X_1, \dots, X_n$ are independent random variables, and that $0 \leqslant X_i \leqslant 1/n$. Let S_n be their sum, and E_n be the expected value of S_n:

$$S_n = \sum_{i=1}^{n} X_i = \widehat{p},$$
$$E_n = \mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = p.$$

Let V_n be the variance of S_n:

$$V_n = \mathbb{V}\text{ar}[S_n] = \sum_{i=1}^{n} \mathbb{V}\text{ar}[X_i] = \sum_{i=1}^{n} \left(\frac{1(1-p)}{n^2} \right) = \frac{p(1-p)}{n}.$$

By applying Bernstein's inequality, we obtain that:

$$\Pr(|S_n - E_n| > \epsilon) < 2 \exp\left(-\frac{\epsilon^2/2}{V_n + C(\epsilon/3)} \right), \tag{5}$$

where C is the range of values of the addends in S_n (i.e., $C = 1/n$). By setting this probability to be equal to δ, we can derive that:

$$\frac{\epsilon^2}{2} = V_n \ln\left(\frac{2}{\delta}\right) + \frac{\epsilon}{3n} \ln\left(\frac{2}{\delta}\right).$$

Let $L = \ln(2/\delta)$, by solving the second order polynomial we get that:

$$\epsilon = \frac{L}{3n} \pm \sqrt{\frac{L^2}{9n^2} + 2V_n L}.$$

We can discard the equation with the minus, as it would result in $\epsilon < 0$, thus resulting in the inequality in Equation (5) holding w.p. $1 - \delta$. As such, we derive that:

$$\begin{aligned}
\epsilon &= \frac{L}{3n} + \sqrt{\frac{L^2}{9n^2} + 2V_n L} \\
&\leqslant \frac{L}{3n} + \sqrt{\frac{L^2}{9n^2}} + \sqrt{\frac{2p(1-p)L}{n}} \\
&= \sqrt{\frac{2p(1-p)L}{n}} + \frac{2L}{3n},
\end{aligned}$$

thus completing the proof. ■

Lemma C.2 (Regret decomposition upper bound) *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Assume events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then the regret from stage h onward of all episodes up to k can be upper bounded as follows:*

$$\begin{aligned}
\sum_{i=1}^{k} \Delta_{i,h}(x_{i,h}) \leqslant \sum_{i=1}^{k} \widetilde{\Delta}_{i,h}(x_{i,h}) \leqslant e \sum_{i=1}^{k} \sum_{j=h}^{H-1} \Big[& \varepsilon_{i,j} + 2\sqrt{L}\overline{\varepsilon}_{i,j} + b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) \\
& + \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} \Big].
\end{aligned}$$

Proof We begin the proof by considering a single value of $k \in [\![K]\!]$. Under $\Omega_{k,h}$, we observe that:

$$\begin{aligned}
\Delta_{k,h}(x_{k,h}) &= V_h^*(x_{k,h}) - V_h^{\pi_k}(x_{k,h}) \\
&\leqslant V_{k,h}(x_{k,h}) - V_h^{\pi_k}(x_{k,h}) \\
&= \widetilde{\Delta}_{k,h}(x_{k,h}).
\end{aligned}$$

As such, we bound the pseudo-regret $\widetilde{\Delta}_{k,h}(x_{k,h})$:

$$\begin{aligned}
\widetilde{\Delta}_{k,h}(x_{k,h}) &= V_{k,h}(x_{k,h}) - V_h^{\pi_k}(x_{k,h}) \\
&= b_{k,h}(x_{k,h}, a_{k,h}^{\pi_k}) + \sum_{y \in \mathcal{S}} \widehat{P}_k(y|x_{k,h}, a_{k,h}^{\pi_k}) V_{k,h+1}(y) - \sum_{y \in \mathcal{S}} P(y|x_{k,h}, a_{k,h}^{\pi_k}) V_{h+1}^{\pi_k}(y) \\
&= b_{k,h}(x_{k,h}, a_{k,h}^{\pi_k}) + \sum_{y \in \mathcal{S}} \Big[\widehat{P}_k(y|x_{k,h}, a_{k,h}^{\pi_k}) - P(y|x_{k,h}, a_{k,h}^{\pi_k}) \Big] V_{k,h+1}(y) \\
&\quad + \sum_{y \in \mathcal{S}} P(y|x_{k,h}, a_{k,h}^{\pi_k}) \left[V_{k,h+1}(y) - V_{h+1}^{\pi_k}(y) \right] \\
&= b_{k,h}(x_{k,h}, a_{k,h}^{\pi_k}) + \sum_{y \in \mathcal{S}} \Big[\widehat{P}_k(y|x_{k,h}, a_{k,h}^{\pi_k}) - P(y|x_{k,h}, a_{k,h}^{\pi_k}) \Big] V_{h+1}^*(y) \\
&\quad + \sum_{y \in \mathcal{S}} \Big[\widehat{P}_k(y|x_{k,h}, a_{k,h}^{\pi_k}) - P(y|x_{k,h}, a_{k,h}^{\pi_k}) \Big] \left[V_{k,h+1}(y) - V_{h+1}^*(y) \right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{y\in\mathcal{S}} P(y|x_{k,h}, a^{\pi_k}_{k,h})\widetilde{\Delta}_{k,h+1}(y) \\
&= \widetilde{\Delta}_{k,h+1}(x_{k,h+1}) + b_{k,h}(x_{k,h}, a^{\pi_k}_{k,h}) + \varepsilon_{k,h} \qquad (6)\\
&\quad + \sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x_{k,h}, a^{\pi_k}_{k,h}) - P(y|x_{k,h}, a^{\pi_k}_{k,h})\right] V^*_{h+1}(y) \\
&\quad + \sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x_{k,h}, a^{\pi_k}_{k,h}) - P(y|x_{k,h}, a^{\pi_k}_{k,h})\right] \left[V_{k,h+1}(y) - V^*_{h+1}(y)\right] \\
&\leqslant \widetilde{\Delta}_{k,h+1}(x_{k,h+1}) + b_{k,h}(x_{k,h}, a^{\pi_k}_{k,h}) + \xi_{k,h}(x_{k,h}, a^{\pi_k}_{k,h}) \\
&\quad + \varepsilon_{k,h} + \underbrace{\sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x_{k,h}, a^{\pi_k}_{k,h}) - P(y|x_{k,h}, a^{\pi_k}_{k,h})\right] \left[V_{k,h+1}(y) - V^*_{h+1}(y)\right]}_{(a)}, \qquad (7)
\end{aligned}
$$

where, in Equation (6) we apply the definition $\varepsilon_{k,h} := P(\cdot|x_{k,h}, a^{\pi_k}_{k,h})^\top \widetilde{\Delta}_{k,h+1}(\cdot) - \widetilde{\Delta}_{k,h+1}(x_{k,h+1})$, and in Equation (7) we apply the definition of $\xi_{k,h}(x_{k,h}, a^{\pi_k}_{k,h})$:

$$
\xi_{k,h}(x_{k,h}, a^{\pi_k}_{k,h}) = \sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x_{k,h}, a^{\pi_k}_{k,h}) - P(y|x_{k,h}, a^{\pi_k}_{k,h})\right] V^*_{h+1}(y).
$$

Let $\mathcal{H}_{k,h}$ be the history of the interactions up to, and including, stage h of episode k. Observing that $|\varepsilon_{k,h}| \leqslant H \leqslant +\infty$ and $\mathbb{E}[\varepsilon_{k,h}|\mathcal{H}_{k,h}] = 0$, we can derive that $\varepsilon_{k,h}$ is a Martingale difference sequence.
We now focus on bounding term (a):

$$
\begin{aligned}
(a) &= \sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x_{k,h}, a^{\pi_k}_{k,h}) - P(y|x_{k,h}, a^{\pi_k}_{k,h})\right] \left[V_{k,h+1}(y) - V^*_{h+1}(y)\right] \\
&\leqslant \sum_{y\in\mathcal{S}} \left[\sqrt{\frac{2P(y|x_{k,h}, a^{\pi_k}_{k,h})(1-P(y|x_{k,h}, a^{\pi_k}_{k,h}))L}{N_k(x_{k,h}, a^{\pi_k}_{k,h})}} + \frac{2L}{3N_k(x_{k,h}, a^{\pi_k}_{k,h})}\right] \widetilde{\Delta}_{k,h+1}(y) \qquad (8)\\
&\leqslant \sum_{y\in\mathcal{S}} \sqrt{\frac{2P(y|x_{k,h}, a^{\pi_k}_{k,h})L}{N_k(x_{k,h}, a^{\pi_k}_{k,h})}} \widetilde{\Delta}_{k,h+1}(y) + \frac{2L}{3N_k(x_{k,h}, a^{\pi_k}_{k,h})} \sum_{y\in\mathcal{S}} \widetilde{\Delta}_{k,h+1}(y) \qquad (9)\\
&\leqslant \sqrt{2L}\underbrace{\sum_{y\in\mathcal{S}} \sqrt{\frac{P(y|x_{k,h}, a^{\pi_k}_{k,h})}{N_k(x_{k,h}, a^{\pi_k}_{k,h})}} \widetilde{\Delta}_{k,h+1}(y)}_{(b)} + \frac{2SHL}{3N_k(x_{k,h}, a^{\pi_k}_{k,h})}, \qquad (10)
\end{aligned}
$$

where Equation (8) is obtained by applying Lemma C.1 to bound $\widehat{P}_k - P$ and by ovserving that $V^*_{h+1}(y) \geqslant V^{\pi_k}_{h+1}(y)$ by definition, Equation (9) is obtained by splitting the terms and observing that $1 - P(y|x,a) \leqslant 1$ for every $x, y \in \mathcal{S}$ and $a \in \mathcal{A}$, and finally Equation (10) is obtained by upper bounding $\widetilde{\Delta}_{k,h+1}(y)$ with H. To bound term (b), we first need to define the following set of states:

$$
[y]_{k,h} := \{y \in \mathcal{S} : N_k(x_{k,h}, a^{\pi_k}_{k,h})P(y|x_{k,h}, a^{\pi_k}_{k,h}) \geqslant 2H^2L\}.
$$

As such, we can rewrite:

$$(b) = \underbrace{\sum_{y \in [y]_{k,h}} \sqrt{\frac{P(y|x_{k,h}, a_{k,h}^{\pi_k})}{N_k(x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y)}_{(c)} + \underbrace{\sum_{y \notin [y]_{k,h}} \sqrt{\frac{P(y|x_{k,h}, a_{k,h}^{\pi_k})}{N_k(x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y)}_{(d)}. \tag{11}$$

We now bound term (c) as:

$$\begin{aligned}(c) &= \sum_{y \in [y]_{k,h}} \sqrt{\frac{P(y|x_{k,h}, a_{k,h}^{\pi_k})}{N_k(x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y) \\ &= \sum_{y \in [y]_{k,h}} P(y|x_{k,h}, a_{k,h}^{\pi_k}) \sqrt{\frac{1}{N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(y|x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y) \\ &= \overline{\varepsilon}_{k,h} + \sqrt{\frac{\mathbb{I}(x_{k,h+1} \in [y]_{k,h})}{N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(x_{k,h+1}|x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(x_{k,h+1}) \qquad (12) \\ &\leqslant \overline{\varepsilon}_{k,h} + \sqrt{\frac{1}{2H^2 L}} \widetilde{\Delta}_{k,h+1}(x_{k,h+1}), \qquad (13)\end{aligned}$$

where Equation (12) is obtained by applying the definition of $\overline{\varepsilon}_{k,h}$:

$$\begin{aligned}\overline{\varepsilon}_{k,h} \coloneqq \sum_{y \in \mathcal{S}} P(y|x_{k,h}, a_{k,h}^{\pi_k}) \sqrt{\frac{\mathbb{I}(y \in [y]_{k,h})}{N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(y|x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y) \\ - \sqrt{\frac{\mathbb{I}(x_{k,h+1} \in [y]_{k,h})}{N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(x_{k,h+1}|x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(x_{k,h+1}),\end{aligned}$$

and Equation (13) is obtained by bounding the indicator function with 1, and by applying the definition of $[y]_{k,h}$. With the same reasoning of $\varepsilon_{k,h}$, we can prove that $\overline{\varepsilon}_{k,h}$ is also a Martingale difference sequence.

We can now bound term (d) as follows:

$$\begin{aligned}(d) &= \sum_{y \notin [y]_{k,h}} \sqrt{\frac{P(y|x_{k,h}, a_{k,h}^{\pi_k})}{N_k(x_{k,h}, a_{k,h}^{\pi_k})}} \widetilde{\Delta}_{k,h+1}(y) \\ &= \sum_{y \notin [y]_{k,h}} \sqrt{\frac{N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(y|x_{k,h}, a_{k,h}^{\pi_k})}{(N_k(x_{k,h}, a_{k,h}^{\pi_k}))^2}} \widetilde{\Delta}_{k,h+1}(y) \\ &\leqslant \frac{H^2 S \sqrt{2L}}{N_k(x_{k,h}, a_{k,h}^{\pi_k})}, \qquad (14)\end{aligned}$$

where Equation (14) is obtained by bounding $\widetilde{\Delta}_{k,h+1}(y)$ with H, and by applying the definition of $[y]_{k,h}$. We can now plug the bounds of (c) and (d) into Equation (11) to obtain that:

$$(b) \leqslant \bar{\varepsilon}_{k,h} + \sqrt{\frac{1}{2H^2L}}\widetilde{\Delta}_{k,h+1}(x_{k,h+1}) + \frac{H^2S\sqrt{2L}}{N_k(x_{k,h}, a^{\pi_k}_{k,h})}.$$

By plugging the bound of (b) into Equation (10), we obtain that:

$$(a) \leqslant \sqrt{2L}\bar{\varepsilon}_{k,h} + \frac{1}{H}\widetilde{\Delta}_{k,h+1}(x_{k,h+1}) + \frac{8H^2SL}{3N_k(x_{k,h}, a^{\pi_k}_{k,h})}.$$

Finally, substituting the bound on (a) into Equation (7), we obtain that:

$$\begin{aligned}\widetilde{\Delta}_{k,h}(x_{k,h}) \leqslant \left(1 + \frac{1}{H}\right)\widetilde{\Delta}_{k,h+1}(x_{k,h+1}) + b_{k,h}(x_{k,h}, a^{\pi_k}_{k,h}) + \xi_{k,h}(x_{k,h}, a^{\pi_k}_{k,h})\\ + \varepsilon_{k,h} + \sqrt{2L}\bar{\varepsilon}_{k,h} + \frac{8H^2SL}{3N_k(x_{k,h}, a^{\pi_k}_{k,h})}.\end{aligned}$$

We now apply an inductive argument on $\widetilde{\Delta}_{k,h}(x_{k,h})$ to isolate the term.
Observing that $\widetilde{\Delta}_{k,H+1}(x_{k,H+1}) = 0$ by definition, we can rewrite:

$$\begin{aligned}\widetilde{\Delta}_{k,h}(x_{k,h}) \leqslant \sum_{j=h}^{H-1}\gamma_{j-h}\bigg[& b_{k,j}(x_{k,j}, a^{\pi_k}_{k,j}) + \xi_{k,j}(x_{k,j}, a^{\pi_k}_{k,j})\\ & + \varepsilon_{k,j} + \sqrt{2L}\bar{\varepsilon}_{k,j} + \frac{8H^2SL}{3N_k(x_{k,j}, a^{\pi_k}_{k,j})}\bigg],\end{aligned}$$

where $\gamma_{j-h} = \left(1 + \frac{1}{H}\right)^{j-h}$. Notice that the summation is limited to $H - 1$. This will be recurrent throughout the paper and is due to the fact that, the reward being deterministic, there is no uncertainty at $h = H$. As such, we can assume that the policies π_k for $k \in [\![K]\!]$ always play greedily at the last stage of each episode.
Observing that $1 + \frac{1}{H} > 1$, we trivially derive that $\gamma_{j-h} \leqslant \gamma_H$ for $j \in [\![h, H]\!]$. Recalling that $\lim_{x\to+\infty}\left(1 + \frac{1}{x}\right)^x = e$, we can bound $\gamma_H \leqslant e$, and rewrite:

$$\begin{aligned}\widetilde{\Delta}_{k,h}(x_{k,h}) \leqslant e\sum_{j=h}^{H-1}\bigg[& b_{k,j}(x_{k,j}, a^{\pi_k}_{k,j}) + \xi_{k,j}(x_{k,j}, a^{\pi_k}_{k,j})\\ & + \varepsilon_{k,j} + \sqrt{2L}\bar{\varepsilon}_{k,j} + \frac{8H^2SL}{3N_k(x_{k,j}, a^{\pi_k}_{k,j})}\bigg],\end{aligned} \tag{15}$$

To conclude the proof, we need to show that this holds for any value of $k \in [\![K]\!]$. Recalling the definition of $\Omega_{k,h}$:

$$\Omega_{k,h} := \left\{V_{i,j}(x) \geqslant V^*_j(x), \forall (i,j) \in [k,h]_{\text{hist}}, x \in \mathcal{S}\right\},$$

where $[k,h]_{\text{hist}} := \{(i,j) : i \in [\![K]\!], j \in [\![H]\!], (i < k) \vee (i = k, j \geqslant h)\}$, we observe that, if $\Omega_{k,h}$ holds, then also the events $\Omega_{i,j}$ hold for $(i,j) \in [k,h]_{hist}$. As such, we can sum up the previous bound of Equation (15) over all the episodes $i \in [\![k]\!]$, thus concluding the proof. ■

Lemma C.3 *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Let events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then the following bounds hold:*

$$\sum_{i=1}^{k}\sum_{j=h}^{H} \varepsilon_{i,j} \leqslant 2\sqrt{H^2 T_k L},$$
$$\sum_{i=1}^{k}\sum_{j=h}^{H} \overline{\varepsilon}_{i,j} \leqslant 2\sqrt{T_k},$$

where $T_k = kH$.

Proof Let us first recall the definitions of $\varepsilon_{i,j}$ and $\overline{\varepsilon}_{i,j}$:

$$\begin{aligned}\varepsilon_{i,j} &\coloneqq P(\cdot|x_{i,j}, a_{i,j}^{\pi_i})^\top \widetilde{\Delta}_{i,j+1}(\cdot) - \widetilde{\Delta}_{i,j+1}(x_{i,j+1}),\\ \overline{\varepsilon}_{i,j} &\coloneqq \sum_{y\in\mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i})\sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i})P(y|x_{i,j}, a_{i,j}^{\pi_i})}}\widetilde{\Delta}_{i,j+1}(y)\\ &\quad - \sqrt{\frac{\mathbb{I}(x_{i,j+1} \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i})P(x_{i,j+1}|x_{i,j}, a_{i,j}^{\pi_i})}}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}),\end{aligned}$$

where:

$$[y]_{k,h} \coloneqq \{y \in \mathcal{S} : N_k(x_{k,h}, a_{k,h}^{\pi_k})P(y|x_{k,h}, a_{k,h}^{\pi_k}) \geqslant 2H^2L\}.$$

Under event $\mathcal{E}$ the following events hold:

$$\mathcal{E}_{\mathrm{az}}(\mathcal{F}_{\widetilde{\Delta},k,h}, H, L), \quad \text{and} \quad \mathcal{E}_{\mathrm{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h}, 1/\sqrt{L}, L).$$

Event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}_{\widetilde{\Delta},k,h}, H, L)$ is defined as the event such that:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left[\sum_{y\in\mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i})\widetilde{\Delta}_{i,j+1}(y) - \widetilde{\Delta}_{i,j+1}(x_{i,j+1})\right] &\leqslant 2\sqrt{k(H-1-h)H^2L}\\ &\leqslant 2\sqrt{H^2T_kL}.\end{aligned}$$

Under this event, we can apply the definition of $\varepsilon_{i,j}$ and derive that:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1} \varepsilon_{i,j} \leqslant 2\sqrt{H^2T_kL}.$$

Event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h}, 1/\sqrt{L}, L)$, on the other hand, is defined as the event such that:

$$\sum_{i=1}^{k}\sum_{j=h}^{H}\left[\sum_{y\in\mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i})\sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i})P(y|x_{i,j}, a_{i,j}^{\pi_i})}}\widetilde{\Delta}_{i,j+1}(y)\right]$$

$$
\begin{aligned}
&- \sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i}) P(y|x_{i,j}, a_{i,j}^{\pi_i})}} \widetilde{\Delta}_{i,j+1}(x_{i,j+1}) \\
&\leqslant 2\sqrt{k(H-h)\frac{1}{\sqrt{L}^2} L} \\
&\leqslant 2\sqrt{T_k}.
\end{aligned}
$$

Under this event, we can apply the definition of $\overline{\varepsilon}_{i,j}$ and derive that:

$$
\sum_{i=1}^{k} \sum_{j=h}^{H-1} \overline{\varepsilon}_{i,j} \leqslant 2\sqrt{T_k},
$$

thus concluding the proof. ■

Lemma C.4 *Let $k \in [\![K]\!]$, $h \in [\![H]\!]$, and $x \in \mathcal{S}$. Let events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then the following bounds hold:*

$$
\begin{aligned}
\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H} \varepsilon_{i,j} &\leqslant 2\sqrt{H^3 N'_{k,h}(x) L}, \\
\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H} \overline{\varepsilon}_{i,j} &\leqslant 2\sqrt{H N'_{k,h}(x)}.
\end{aligned}
$$

Proof In a similar way to the proof of Lemma C.3, we recall the definitions of $\varepsilon_{i,j}$ and $\overline{\varepsilon}_{i,j}$:

$$
\begin{aligned}
\varepsilon_{i,j} &:= P(\cdot|x_{i,j}, a_{i,j}^{\pi_i})^\top \widetilde{\Delta}_{i,j+1}(\cdot) - \widetilde{\Delta}_{i,j+1}(x_{i,j+1}), \\
\overline{\varepsilon}_{i,j} &:= \sum_{y \in \mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i}) \sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i}) P(y|x_{i,j}, a_{i,j}^{\pi_i})}} \widetilde{\Delta}_{i,j+1}(y) \\
&\quad - \sqrt{\frac{\mathbb{I}(x_{i,j+1} \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i}) P(x_{i,j+1}|x_{i,j}, a_{i,j}^{\pi_i})}} \widetilde{\Delta}_{i,j+1}(x_{i,j+1}),
\end{aligned}
$$

where:

$$
[y]_{k,h} := \{y \in \mathcal{S} : N_k(x_{k,h}, a_{k,h}^{\pi_k}) P(y|x_{k,h}, a_{k,h}^{\pi_k}) \geqslant 2H^2 L\}.
$$

Under event $\mathcal{E}$ the following events hold:

$$
\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, L), \quad \text{and} \quad \mathcal{E}_{\text{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h,x}, 1/\sqrt{L}, L).
$$

Event $\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, L)$ is defined as the event such that:

$$
\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \left[\sum_{y \in \mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i}) \widetilde{\Delta}_{i,j+1}(y) - \widetilde{\Delta}_{i,j+1}(x_{i,j+1}) \right] \leqslant 2\sqrt{H^3 N'_{k,h}(x) L}.
$$

Under this event, we can apply the definition of $\varepsilon_{i,j}$ and derive that:

$$\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \varepsilon_{i,j} \leqslant 2\sqrt{H^3 N'_{k,h}(x) L}.$$

Event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h,x}, 1/\sqrt{L}, L)$, on the other hand, is defined as the event such that:

$$\begin{aligned}\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H} \Bigg[\sum_{y \in \mathcal{S}} P(y|x_{i,j}, a_{i,j}^{\pi_i}) \sqrt{\frac{\mathbb{I}(y \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i}) P(y|x_{i,j}, a_{i,j}^{\pi_i})}} \widetilde{\Delta}_{i,j+1}(y) \Bigg] \\ - \sqrt{\frac{\mathbb{I}(x_{i,j+1} \in [y]_{i,j})}{N_i(x_{i,j}, a_{i,j}^{\pi_i}) P(x_{i,j+1}|x_{i,j}, a_{i,j}^{\pi_i})}} \widetilde{\Delta}_{i,j+1}(x_{i,j+1}) \\ \leqslant 2\sqrt{N'_{k,h}(x)(H-h) \frac{1}{\sqrt{L}^2} L} \\ \leqslant 2\sqrt{H N'_{k,h}(x)}.\end{aligned}$$

Under this event, we can apply the definition of $\overline{\varepsilon}_{i,j}$ and derive that:

$$\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \overline{\varepsilon}_{i,j} \leqslant 2\sqrt{H N'_{k,h}(x)},$$

thus concluding the proof.

■

Lemma C.5 *Let $k \in \llbracket K \rrbracket$ and $h \in \llbracket H \rrbracket$. Let π_k be the policy followed during episode k. Under the events $\mathcal{E}$ and $\Omega_{k,h}$, the following holds for every $x \in \mathcal{S}$:*

$$\sum_{i=1}^{k} \sum_{j=h}^{H-1} \mathbb{V}_{j+1}^{\pi_i}(x_{i,j}, a_{i,j}^{\pi_i}) \leqslant H T_k + 2\sqrt{H^4 T_k L} + \frac{4}{3} H^3 L,$$

$$\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \mathbb{V}_{j+1}^{\pi_i}(x_{i,j}, a_{i,j}^{\pi_i}) \leqslant H^2 N'_{k,h}(x) + 2\sqrt{H^5 N'_{k,h}(x) L} + \frac{4}{3} H^3 L.$$

Proof We begin the proof by restating the definition of $\mathbb{V}_{j+1}^{\pi_k}(x_{i,j}, a_{i,j}^{\pi_i})$:

$$\mathbb{V}_{j+1}^{\pi_k}(x_{i,j}, a_{i,j}^{\pi_i}) := \underset{y \sim P(\cdot|x_{i,j}, a_{i,j}^{\pi_i})}{\mathbb{V}\mathrm{ar}} [V_{j+1}^{\pi_k}(y)]$$

Under event $\mathcal{E}$, the following events hold:

$$\mathcal{E}_{\mathrm{fr}}(\mathcal{G}_{\mathbb{V},k,h}, H^4 T_k, H^3, L) \quad \text{and} \quad \mathcal{E}_{\mathrm{fr}}(\mathcal{G}_{\mathbb{V},k,h,x}, H^5 N'_{k,h}, H^3, L).$$

Event $\mathcal{E}_{\mathrm{fr}}(\mathcal{G}_{\mathbb{V},k,h}, H^4 T_k, H^3, L)$ is defined as the event such that:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})-\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})|\mathcal{H}_{k,h}\right]\leqslant 2\sqrt{H^4T_kL}+\frac{4H^3L}{3},$$

which implies that:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})\leqslant\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})|\mathcal{H}_{k,h}\right]+2\sqrt{H^4T_kL}+\frac{4H^3L}{3}. \tag{16}$$

On the other hand, event $\mathcal{E}_{\text{fr}}(\mathcal{G}_{\mathbb{V},k,h,x},H^5N'_{k,h},H^3,L)$ is defined as the event such that:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})-\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})|\mathcal{H}_{k,h}\right]&\\ \leqslant 2\sqrt{H^5N'_{k,h}(x)L}+\frac{4H^3L}{3}&,\end{aligned}$$

which implies that:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})\leqslant\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}_{j+1}^{\pi_k}(x_{i,j},a_{i,j}^{\pi_i})|\mathcal{H}_{k,h}\right]&\\ +2\sqrt{H^5N'_{k,h}(x)L}+\frac{4H^3L}{3}&.\end{aligned} \tag{17}$$

Observe that by applying the *law of total variance* (LTV, see, e.g., Theorem 9.5.5 of Blitzstein and Hwang, 2019), we can write:

$$\begin{aligned}\underset{x_{i,h+1},\dots,x_{i,H-1}}{\mathbb{V}\mathrm{ar}}\left[\sum_{j=h}^{H-1}R^{\pi}(x_{i,j})\right]&=\underbrace{\underset{x_{i,h+1}}{\mathbb{V}\mathrm{ar}}\left[\underset{x_{i,h+2},\dots,x_{i,H-1}}{\mathbb{E}}\left[\sum_{j=h}^{H-1}R^{\pi}(x_{i,j})\middle|x_{i,h+1}\right]\right]}_{(a)}\\ &\quad+\underset{x_{i,h+1}}{\mathbb{E}}\left[\underbrace{\underset{x_{i,h+2},\dots,x_{i,H-1}}{\mathbb{V}\mathrm{ar}}\left[\sum_{j=h}^{H-1}R^{\pi}(x_{i,j})\middle|x_{i,h+1}\right]}_{(b)}\right].\end{aligned} \tag{18}$$

Term (a) can be rewritten as:

$$(a)=\underset{x_{i,h+1}}{\mathbb{V}\mathrm{ar}}\left[R^{\pi}(x_{i,h})+\underset{x_{i,h+2},\dots,x_{i,H-1}}{\mathbb{E}}\left[\sum_{j=h+1}^{H-1}R^{\pi}(x_{i,j})\middle|x_{i,h+1}\right]\right]$$

$$= \underset{x_{i,h+1}}{\mathbb{V}\text{ar}}\left[V^{\pi_k}_{h+1}(x_{i,h+1})\right] \tag{19}$$

$$= \mathbb{V}^{\pi_k}_{h+1}(x_{i,h}, a^{\pi_i}_{i,h}), \tag{20}$$

where Equation (19) is obtained by observing that $R^{\pi}(x_{i,h})$ has zero variance w.r.t. $x_{i,h+1}$, and by applying the definition of value function.
We can then recursively apply the LTV to term (b) and, considering the expectation over the trajectory generated following policy π from stage h onward, we can write:

$$\underset{x_{i,h+1},\ldots,x_{i,H-1}}{\mathbb{V}\text{ar}}\left[\sum_{j=h}^{H-1} R^{\pi}(x_{i,j})\right] = \mathbb{E}\left[\sum_{j=h}^{H-1} \mathbb{V}^{\pi_k}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})\right]. \tag{21}$$

By applying the result of Equation (21) to Equations (16) and (17), we get:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})|\mathcal{H}_{k,h}\right] &= \sum_{i=1}^{k}\mathbb{V}\text{ar}\left[\sum_{j=h+1}^{H-1} R^{\pi}(x_{i,j})\right]\\ &\leqslant k(H-h)^2\\ &\leqslant HT_k, \end{aligned}\tag{22}$$

and:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathbb{E}\left[\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})|\mathcal{H}_{k,h}\right] &= \sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\,\mathbb{V}\text{ar}\left[\sum_{j=h+1}^{H-1} R^{\pi}(x_{i,j})\right]\\ &\leqslant N'_{k,h}(x)(H-h)^2\\ &\leqslant H^2 N'_{k,h}(x). \end{aligned}\tag{23}$$

Finally, we can plug Equations (22) and (23) into Equations (16) and (17), respectively, obtaining:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) \leqslant HT_k + 2\sqrt{H^4 T_k L} + \frac{4}{3}H^3 L,$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{V}^{\pi_k}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) \leqslant H^2 N'_{k,h}(x) + 2\sqrt{H^5 N'_{k,h}(x) L} + \frac{4}{3}H^3 L,$$

thus concluding the proof. ■

Lemma C.6 *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Let π_k be the policy played during episode k. Under the events $\mathcal{E}$ and $\Omega_{k,h}$, the following holds for every $x \in \mathcal{S}$:*

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left(\mathbb{V}^{*}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})\right) \leqslant 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j}) + 4H^2\sqrt{T_k L},$$

$$\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \Big(\mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) \Big) \leqslant 2H \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,h}(x_{i,h}) + 4H^2 \sqrt{HN'_{k,h}(x)L}.$$

Proof We demonstrate the result by providing an upper bound to $\mathbb{V}^*_{j+1} - \mathbb{V}^{\pi_i}_{j+1}$ first, and then bounding its summation over episodes and stages. We can demonstrate that:

$$\begin{aligned} \mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) &= \underset{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})}{\mathbb{V}\mathrm{ar}} [V^*_{j+1}(y)] - \underset{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})}{\mathbb{V}\mathrm{ar}} [V^{\pi_i}_{j+1}(y)] \\ &\leqslant \mathbb{E}_{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} [(V^*_{j+1}(y))^2 - (V^{\pi_i}_{j+1}(y))^2] \qquad (24) \\ &\leqslant 2H \mathbb{E}_{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} [V^*_{j+1}(y) - V^{\pi_i}_{j+1}(y)], \qquad (25) \end{aligned}$$

where Equation (24) is obtained by applying the definition of variance and observing that $V^*_{j+1}(x) \geqslant V^{\pi_i}_{j+1}(x)$ by definition, and Equation (25) is obtained by expanding the square and by observing that $V^{\pi}_{j+1}(x) \leqslant V^*_{j+1}(x) \leqslant H$.
Using the argument of Equation (25), we obtain the following inequalities:

$$\begin{aligned} &\sum_{i=1}^{k} \sum_{j=h}^{H-1} \left(\mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) \right) \\ &\qquad \leqslant 2H \underbrace{\sum_{i=1}^{k} \sum_{j=h}^{H-1} \mathbb{E}_{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} [\Delta_{i,j+1}(y)]}_{(a)}, \qquad (26) \end{aligned}$$

$$\begin{aligned} &\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \left(\mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) \right) \\ &\qquad \leqslant 2H \underbrace{\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \mathbb{E}_{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} [\Delta_{i,j+1}(y)]}_{(b)}. \qquad (27) \end{aligned}$$

We now bound term (a) as follows:

$$\begin{aligned} (a) &\leqslant \sum_{i=1}^{k} \sum_{j=h}^{H-1} \mathbb{E}_{y \sim P(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} [\widetilde{\Delta}_{i,j+1}(y)] \qquad (28) \\ &\leqslant 2\sqrt{H^2 T_k L} + \sum_{i=1}^{k} \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1}(x_{i,j+1}) \qquad (29) \end{aligned}$$

where Equation (28) is obtained because, under $\Omega_{k,h}$, it holds that $V^*_{j+1}(y) \leqslant V_{i,j+1}(y)$. Equation (29) is obtained by considering that, under event $\mathcal{E}$, the event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}_{\widetilde{\Delta},k,h}, H, L)$ holds, as shown in Lemma C.3.

Following a similar procedure, we bound term (b) by considering event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, L)$, thus obtaining:

$$(b) \leqslant 2H\sqrt{HN'_{k,h}(x)L} + \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1}(x_{i,j+1}). \tag{30}$$

We can then plug Equations (29) and (30) into Equations (26) and (27), respectively, to write:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left(\mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})\right) \leqslant 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j}) + 4H^2\sqrt{T_k L},$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h} = x)\sum_{j=h}^{H-1}\left(\mathbb{V}^*_{j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})\right) \leqslant 2H\sum_{i=1}^{k}\mathbb{I}(x_{i,h} = x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j}) + 4H^2\sqrt{HN'_{k,h}(x)L},$$

thus concluding the proof. ∎

Lemma C.7 *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Let π_k denote the policy followed during episode k. Under events $\mathcal{E}$ and $\Omega_{k,h}$, the following inequalities hold for every $x \in \mathcal{S}$:*

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})$$
$$\leqslant 7H^2S\sqrt{AT_kL} + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}),$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,j} = x)\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})$$
$$\leqslant 7H^2S\sqrt{HAN'_{k,h}(x)L} + 2H\sum_{i=1}^{k}\mathbb{I}(x_{i,h} = x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}).$$

Proof Similarly to the proof of Lemma C.6, we demonstrate the result by providing an upper bound to $\widehat{\mathbb{V}}_{i,j+1} - \mathbb{V}^{\pi_i}_{j+1}$ first, and then bounding its summation over episodes and stages.

$$\begin{aligned}
&\widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})\\
&\qquad = \underset{y\sim\widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}{\mathbb{V}\mathrm{ar}}[V_{i,j+1}(y)] - \underset{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}{\mathbb{V}\mathrm{ar}}[V^{\pi_i}_{j+1}(y)]\\
&\qquad = \mathbb{E}_{y\sim\widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim\widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[V_{i,j+1}(y)]^2
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V^{\pi_i}_{j+1}(y))^2] + \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[V^{\pi_i}_{j+1}(y)]^2 \\
\leqslant\ & \mathbb{E}_{y\sim \widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V^{\pi_i}_{j+1}(y))^2] \\
& + \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[V^{*}_{j+1}(y)]^2 - \mathbb{E}_{y\sim \widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[V^{*}_{j+1}(y)]^2 \qquad (31) \\
\leqslant\ & \mathbb{E}_{y\sim \widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] \\
& + \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V^{\pi_i}_{j+1}(y))^2] \qquad (32) \\
& + 2H \sum_{y\in\mathcal{S}} (P(y|x_{i,j},a^{\pi_i}_{i,j}) - \widehat{P}_i(y|x_{i,j},a^{\pi_i}_{i,j}))V^{*}_{j+1}(y) \\
\leqslant\ & \mathbb{E}_{y\sim \widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] \\
& + \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V^{\pi_i}_{j+1}(y))^2] \qquad (33) \\
& + 4H\sqrt{\frac{H^2L}{N_i(x_{i,j},a^{\pi_i}_{i,j})}}
\end{aligned}
$$

where Equation (31) follows from the fact that, under $\Omega_{k,h}$, $V_{i,j}(y) \geqslant V^*_j(y) \geqslant V^{\pi_i}_j(y)$. Equation (32) is obtained by adding and subtracting $\mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2]$, and by observing that $V^*_j(y) \leqslant H$. Equation (33) is obtained by bounding the model error via Hoeffding's inequality.

Putting this result into the double summation, we get:

$$
\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=h}^{H-1} \widehat{\mathbb{V}}_{i,j+1}(x_{i,j},a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j},a^{\pi_i}_{i,j}) \\
&\quad \leqslant \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left[\mathbb{E}_{y\sim \widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2]\right]}_{(a)} \\
&\quad + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V_{i,j+1}(y))^2] - \mathbb{E}_{y\sim P(\cdot|x_{i,j},a^{\pi_i}_{i,j})}[(V^{\pi_i}_{j+1}(y))^2]}_{(b)} \qquad (34) \\
&\quad + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1} 4H\sqrt{\frac{H^2L}{N_i(x_{i,j},a^{\pi_i}_{i,j})}}}_{(c)}.
\end{aligned}
$$

We begin by bounding term (a):

$$
\begin{aligned}
(a) &\leqslant \sum_{i=1}^{k}\sum_{j=h}^{H-1} H^2\|\widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j}) - P(\cdot|x_{i,j},a^{\pi_i}_{i,j})\|_1 \\
&\leqslant \sum_{i=1}^{k}\sum_{j=h}^{H-1} 2H^2\sqrt{\frac{SL}{N_i(x_{i,j},a^{\pi_i}_{i,j})}} \qquad (35)
\end{aligned}
$$

$$
\begin{aligned}
&= 2H^2\sqrt{SL}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_i(x,a)} n^{-1/2}\\
&\leqslant 2H^2\sqrt{SL}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{\frac{kH}{SA}} n^{-1/2}\\
&\leqslant H^2 S\sqrt{AT_kL},
\end{aligned}
$$

where Equation (35) follows by applying the result of Theorem 2.1 of Weissman et al. (2003), which holds under event $\mathcal{E}$.

We now bound term (b):

$$
\begin{aligned}
(b) &= \sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim P(\cdot|x_{i,j},a_{i,j}^{\pi_i})}[(V_{i,j+1}(y)+V_{j+1}^{\pi_k}(y))(V_{i,j+1}(y)-V_{j+1}^{\pi_k}(y))]\\
&\leqslant 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim P(\cdot|x_{i,j},a_{i,j}^{\pi_i})}[\widetilde{\Delta}_{i,j+1}(y)]\\
&\leqslant 2H(\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})+2H\sqrt{T_kL}),
\end{aligned}
\tag{36}
$$

where Equation (36) is obtained by observing that, under event $\mathcal{E}$, event $\mathcal{E}_{\mathrm{az}}(\mathcal{F}_{\widetilde{\Delta},k,h},H,L)$ holds. We now bound term (c):

$$
\begin{aligned}
(c) &\leqslant 4H^2\sqrt{L}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{\frac{kH}{SA}} n^{-1/2}\\
&\leqslant 2H^2\sqrt{SAT_kL}.
\end{aligned}
$$

Finally, by plugging the bounds of terms (a), (b), and (c) into Equation (34), we get:

$$
\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1}(x_{i,j},a_{i,j}^{\pi_i})-\mathbb{V}_{j+1}^{\pi_i}(x_{i,j},a_{i,j}^{\pi_i})\\
&\qquad\leqslant H^2S\sqrt{AT_kL}+2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})\\
&\qquad\quad+4H^2\sqrt{T_kL}+2H^2\sqrt{SAT_kL}\\
&\qquad\leqslant 7H^2S\sqrt{AT_kL}+2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}).
\end{aligned}
$$

Using the same procedure, we can bound the following summation as:

$$
\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1}(x_{i,j},a_{i,j}^{\pi_i})-\mathbb{V}_{j+1}^{\pi_i}(x_{i,j},a_{i,j}^{\pi_i})
$$

$$\leqslant 7H^2S\sqrt{HAN'_{k,h}(x)L} + 2H\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}),$$

thus concluding the proof. ■

Lemma C.8 (Summation over typical episodes of state-action wise model errors) *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Let π_k be the policy followed during episode k. Under events $\mathcal{E}$ and $\Omega_{k,h}$, the following inequalities hold for every $x \in \mathcal{S}$:*

$$\sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ}})\sum_{j=h}^{H-1}\left[\widehat{P}_i(\cdot|x_{i,j},a_{i,j}^{\pi_i}) - P(\cdot|x_{i,j},a_{i,j}^{\pi_i})\right]^\top V^*_{j+1}(\cdot)$$
$$\leqslant \sqrt{6HSAT_kL^2} + \frac{2}{3}HSAL^2 \tag{37}$$
$$+ 2\sqrt{HSAL^2\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})},$$

$$\sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ},x}, x_{i,h}=x)\sum_{j=h}^{H-1}\left[\widehat{P}_i(\cdot|x_{i,j},a_{i,j}^{\pi_i}) - P(\cdot|x_{i,j},a_{i,j}^{\pi_i})\right]^\top V^*_{j+1}(\cdot)$$
$$\leqslant \sqrt{6H^2SAN'_{k,h}(x)L^2} + \frac{2}{3}HSAL^2 \tag{38}$$
$$+ 2\sqrt{HSAL^2\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})},$$

where:

$$[k]_{\text{typ}} := \{i \in [\![k]\!] : (x_{i,h},a_{i,h}^{\pi_i}) \in [(x,a)]_k, i \geqslant 250HS^2AL, \forall h \in [\![H]\!]\},$$
$$[k]_{\text{typ},x} := \{i \in [\![k]\!] : (x_{i,h},a_{i,h}^{\pi_i}) \in [(x,a)]_k, N'_{k,h}(x) \geqslant 250HS^2AL, \forall h \in [\![H]\!]\},$$
$$[(x,a)]_k := \{(x,a) \in \mathcal{S}\times\mathcal{A} : N_k(x,a) \geqslant H, N'_{k,h}(x) \geqslant H, \forall h \in [\![H]\!]\}.$$

Proof We begin by demonstrating the bound of Equation (37):

$$\sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ}})\sum_{j=h}^{H-1}\left[\widehat{P}_i(\cdot|x_{i,j},a_{i,j}^{\pi_i}) - P(\cdot|x_{i,j},a_{i,j}^{\pi_i})\right]^\top V^*_{j+1}(\cdot)$$
$$\leqslant \sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ}})\sum_{j=h}^{H-1}\left[\sqrt{\frac{2\mathbb{V}^*_{j+1}(x_{i,j},a_{i,j}^{\pi_i})L}{N_i(x_{i,j},a_{i,j}^{\pi_i})}} + \frac{2HL}{3N_i(x_{i,j},a_{i,j}^{\pi_i})}\right] \tag{39}$$

$$\begin{aligned}&\leqslant \sqrt{2L}\sqrt{\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^*_{j+1}(x_{i,j},a^{\pi_i}_{i,j})}_{(a)}}\sqrt{\underbrace{\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\frac{1}{N_i(x_{i,j},a^{\pi_i}_{i,j})}}_{(b)}}\\ &\quad+\underbrace{\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\frac{2HL}{3N_i(x_{i,j},a^{\pi_i}_{i,j})}}_{(c)},\end{aligned} \tag{40}$$

where Equation (39) is obtained by applying Bernstein's inequality (see, e.g., Cesa-Bianchi and Lugosi, 2006), and Equation (40) is obtained by applying Cauchy-Schwarz's inequality. We now bound terms (a), (b), and (c).
By adding and subtracting $\mathbb{V}^{\pi_i}_{j+1}(x_{i,j},a^{\pi_i}_{i,j})$ to term (a), we can rewrite it as:

$$(a)=\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{\pi_i}_{j+1}(x_{i,j},a^{\pi_i}_{i,j})}_{(d)}+\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left(\mathbb{V}^*_{j+1}(x_{i,j},a^{\pi_i}_{i,j})-\mathbb{V}^{\pi_i}_{j+1}(x_{i,j},a^{\pi_i}_{i,j})\right)}_{(e)}.$$

As events $\mathcal{E}$ and $\Omega_{k,h}$ hold, we can apply Lemmas C.5 and C.6 to bound terms (d) and (e), respectively, thus obtaining:

$$\begin{aligned}(a)&\leqslant HT_k+2H^2\sqrt{T_kL}+\frac{4}{3}H^3L+2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})+4H^2\sqrt{T_kL}\\ &\leqslant 3T_kH+2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j}),\end{aligned} \tag{41}$$

where Equation (41) holds under the condition of $[k]_{\text{typ}}$.
We now bound terms (b) and (c) as follows:

$$\begin{aligned}(b)&\leqslant\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{kH}n^{-1}\\ &\leqslant SAL,\end{aligned} \tag{42}$$

$$\begin{aligned}(c)&\leqslant\frac{2}{3}HL\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{kH}n^{-1}\\ &\leqslant\frac{2}{3}HSAL^2.\end{aligned} \tag{43}$$

Finally, by plugging the results of Equations (41), (42), and (43) into Equation (40), we get:

$$\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\left[\widehat{P}_i(\cdot|x_{i,j},a^{\pi_i}_{i,j})-P(\cdot|x_{i,j},a^{\pi_i}_{i,j})\right]^\top V^*_{j+1}(\cdot)$$

$$\leqslant \sqrt{2L}\sqrt{3T_kH + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})}\sqrt{SAL} + \frac{2}{3}HSAL^2$$

$$\leqslant \sqrt{6HSAT_kL^2} + 2\sqrt{HSAL^2(\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j}))} + \frac{2}{3}HSAL^2, \tag{44}$$

where Equation (44) is obtained by computing the product of the square roots and by the sub-additivity of the square root. Following the same procedure, we can obtain the upper bound of Equation (38) by substituting terms T_k with $HN'_{k,h}(x)$, and terms $\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})$ with $\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j}(x_{i,j})$. ■

Lemma C.9 (Summation over typical episodes of bonus terms) *Let $k \in [\![K]\!]$ and $h \in [\![H]\!]$. Let π_k be the policy followed during episode k. Let the UCB bonus be defined as:*

$$b_{k,h}(x,a) = \sqrt{\frac{4L\,\mathbb{V}\mathrm{ar}_{y\sim\widehat{P}_k(\cdot|x,a)}[V_{k,h+1}(y)]}{N_k(x,a)}} + \frac{7HL}{3(N_k(x,a)-1)}$$
$$+\sqrt{\frac{4\mathbb{E}_{y\sim\widehat{P}_k(\cdot|x,a)}[\min\{\frac{84^2H^3S^2AL^2}{N'_{k,h+1}(y)}, H^2\}]}{N_k(x,a)}}.$$

Under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following inequalities hold for every $x \in \mathcal{S}$:

$$\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\mathrm{typ}})\sum_{j=h}^{H-1}b_{i,j}(x_{i,j},a_{i,j}^{\pi_i})$$
$$\leqslant \sqrt{28HSAT_kL^2} + \frac{7}{3}HSAL^2 + 2\sqrt{84^2H^3S^4A^2L^4} \tag{45}$$
$$+\sqrt{8HSAL^2\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})},$$

$$\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\mathrm{typ},x}, x_{i,h}=x)\sum_{j=h}^{H-1}b_{i,j}(x_{i,j},a_{i,j}^{\pi_i})$$
$$\leqslant \sqrt{28H^2SAN'_{k,h}(x)L^2} + \frac{7}{3}HSAL^2 + 2\sqrt{84^2H^3S^4A^2L^4} \tag{46}$$
$$+\sqrt{8HSAL^2\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})},$$

where:

$$[k]_{\mathrm{typ}} := \{i\in[\![k]\!] : (x_{i,h},a_{i,h}^{\pi_i})\in[(x,a)]_k, i \geqslant 250HS^2AL, \forall h\in[\![H]\!]\},$$

$$[k]_{\text{typ},x} := \{i \in [\![k]\!] : (x_{i,h}, a^{\pi_i}_{i,h}) \in [(x,a)]_k, N'_{k,h}(x) \geqslant 250HS^2AL, \forall h \in [\![H]\!]\},$$
$$[(x,a)]_k := \{(x,a) \in \mathcal{S} \times \mathcal{A} : N_k(x,a) \geqslant H, N'_{k,h}(x) \geqslant H, \forall h \in [\![H]\!]\}.$$

Proof We begin by demonstrating the bound of Equation (45). We can rewrite the summation as:

$$\begin{aligned}\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} b_{i,j}(x_{i,j}, a^{\pi_i}_{i,j}) \leqslant & \underbrace{\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} \sqrt{\frac{4L\widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j})}{N_i(x_{i,j}, a^{\pi_i}_{i,j})}}}_{(a)} \\ & + \underbrace{\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} \frac{7HL}{3(N_i(x_{i,j}, a^{\pi_i}_{i,j}) - 1)}}_{(b)} \\ & + \underbrace{\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} \sqrt{\frac{4\mathbb{E}_{y \sim \widehat{P}_i(\cdot|x_{i,j}, a^{\pi_i}_{i,j})} b'_{i,j+1}(y)}{N_i(x_{i,j}, a^{\pi_i}_{i,j})}}}_{(c)},\end{aligned} \tag{47}$$

where $b'_{i,j+1}(y) := \min\{\frac{84^2H^2S^2AL^2}{N'_{i,j+1}(y)}, H^2\}$. First of all, we observe that we can bound term (b) by using a pigeonhole argument as:

$$(b) \leqslant \frac{7}{3}HSAL^2. \tag{48}$$

We now bound term (a). By applying Cauchy-Schwarz's inequality, we obtain:

$$(a) \leqslant \sqrt{4L} \sqrt{\underbrace{\sum_{i=1}^{k} \sum_{j=h}^{H-1} \widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j})}_{(d)}} \sqrt{\underbrace{\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} \frac{1}{N_i(x_{i,j}, a^{\pi_i}_{i,j})}}_{(e)}}. \tag{49}$$

By applying the same argument as that of Equation (42) of Lemma C.8, we bound term (e) with SAL.
We can rewrite term (d) as follows:

$$(d) = \underbrace{\sum_{i=1}^{k} \sum_{j=h}^{H+1} \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})}_{(f)} + \underbrace{\sum_{i=1}^{k} \sum_{j=h}^{H-1} [\widehat{\mathbb{V}}_{i,j+1}(x_{i,j}, a^{\pi_i}_{i,j}) - \mathbb{V}^{\pi_i}_{j+1}(x_{i,j}, a^{\pi_i}_{i,j})]}_{(g)} \tag{50}$$

Under events $\mathcal{E}$ and $\Omega_{k,h}$, we can apply Lemmas C.5 and C.7 to upper bound terms (f) and (g) respectively, obtaining the following:

$$(f) \leqslant HT_k + 2\sqrt{H^4T_kL} + \frac{4H^3L}{3},$$

$$(g) \leqslant 7H^2S\sqrt{AT_kL} + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}).$$

Plugging the bounds of (f) and (g) into Equation (50), we get:

$$\begin{aligned}(d) &\leqslant HT_k + 2\sqrt{H^4T_kL} + \frac{4H^3L}{3} + 7H^2S\sqrt{AT_kL} + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}) \\ &\leqslant 4HT_k + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1}), \qquad (51)\end{aligned}$$

where Equation (51) holds under the condition of $[k]_{\mathrm{typ}}$. Combining the bounds of terms (d) and (e), we can rewrite Equation (49) as:

$$\begin{aligned}(a) &\leqslant \sqrt{4L}\sqrt{4HT_k + 2H\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})}\sqrt{SAL} \\ &\leqslant \sqrt{16HSAT_kL^2} + \sqrt{8HSAL^2\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\Delta}_{i,j+1}(x_{i,j+1})}, \qquad (52)\end{aligned}$$

where Equation (52) is obtained by expanding the products and applying the subadditivity of the square root.

To bound term (c), we apply Cauchy-Schwarz's inequality, obtaining:

$$(c) \leqslant 2\sqrt{\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim\widehat{P}_i(\cdot|x_{i,j},a_{i,j}^{\pi_i})}b'_{i,j+1}(y)}_{(h)}}\sqrt{\underbrace{\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\mathrm{typ}})\sum_{j=h}^{H-1}\frac{1}{N_i(x_{i,j},a_{i,j}^{\pi_i})}}_{(i)}}. \qquad (53)$$

Similar to term (e), we can bound term (i) with SAL. We now bound term (h). We can rewrite the term as:

$$\begin{aligned}(h) &= \sum_{i=1}^{k}\sum_{j=h}^{H-1}\sum_{y\in\mathcal{S}}\widehat{P}_i(y|x_{i,j},a_{i,j}^{\pi_i})b'_{i,j+1}(y) \\ &= \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\sum_{y\in\mathcal{S}}(\widehat{P}_i(y|x_{i,j},a_{i,j}^{\pi_i}) - P(y|x_{i,j},a_{i,j}^{\pi_i}))b'_{i,j+1}(y)}_{(j)} + \sum_{i=1}^{k}\sum_{j=h}^{H-1}\sum_{y\in\mathcal{S}}P(y|x_{i,j},a_{i,j}^{\pi_i})b'_{i,j+1}(y) \\ &= (j) + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim P(\cdot|x_{i,j},a_{i,j}^{\pi_i})}b'_{i,j+1}(y) - b'_{i,j+1}(x_{i,j+1})}_{(k)} + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}b'_{i,j+1}(x_{i,j+1})}_{(l)}. \qquad (54)\end{aligned}$$

We bound term (j) as follows:

$$\begin{aligned}
(j) &\leqslant H^2 \sum_{i=1}^{k} \sum_{j=h}^{H-1} \|\widehat{P}_i(\cdot|x_{i,j}, a_{i,j}^{\pi_i}) - P(\cdot|x_{i,j}, a_{i,j}^{\pi_i})\|_1 && (55)\\
&\leqslant 2H^2\sqrt{SL} \sum_{i=1}^{k} \sum_{j=h}^{H-1} (N_i(x_{i,j}, a_{i,j}^{\pi_i}))^{-1/2} && (56)\\
&\leqslant 2H^2\sqrt{SL} \sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{n=1}^{\frac{kH}{SA}} n^{-1/2}\\
&\leqslant H^2 S\sqrt{AT_k L}, && (57)
\end{aligned}$$

where Equation (55) is obtained by bounding $b'_{i,j+1}(y)$ with H^2, Equation (56) follows by applying the result of Theorem 2.1 of Weissman et al. (2003), which holds under event $\mathcal{E}$, and Equation (57) follows from a derivation similar to that of term (a) of Lemma C.7.
To bound term (k), we first observe that it is a Martingale difference sequence, and as such we can bound it via the event $\mathcal{E}_{\text{az}}(\mathcal{F}_{b',k,h}, H^2, L)$, which holds under $\mathcal{E}$, obtaining:

$$(k) \leqslant 2H^2\sqrt{T_k L}.$$

By applying the definition of b', we can bound term (l) as:

$$\begin{aligned}
(l) &\leqslant 84^2 H^3 S^2 A L^2 \sum_{i=1}^{k} \sum_{j=h}^{H-1} \frac{1}{N'_{i,j+1}(x_{i,j+1})}\\
&\leqslant 84^2 H^3 S^2 A L^2 \sum_{x\in\mathcal{S}} \sum_{n=1}^{T} n^{-1}\\
&\leqslant 84^2 H^3 S^3 A L^3.
\end{aligned}$$

Plugging the bounds of terms (j), (k), and (l) into Equation (54), we get:

$$(h) \leqslant H^2 S\sqrt{AT_k L} + 2H^2\sqrt{T_k L} + 84^2 H^3 S^3 A L^3.$$

By applying the bounds of terms (h) and (i) to Equation (53), we get:

$$\begin{aligned}
(c) &\leqslant 2\sqrt{H^2 S\sqrt{AT_k L} + 2H^2\sqrt{T_k L} + 84^2 H^3 S^3 A L^3}\sqrt{SAL}\\
&\leqslant 2\sqrt{3HSAT_k L} + 2\sqrt{84^2 H^3 S^4 A^2 L^4}, && (58)
\end{aligned}$$

where Equation (58) is obtained by expanding the products, applying the subadditivity of the square root, and applying the definition of $[k]_{\text{typ}}$.
Finally, we can combine the bounds of terms (a), (b), and (c) into Equation (47), obtaining the following bound:

$$
\begin{aligned}
&\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) \\
&\qquad\leqslant \sqrt{16HSAT_kL^2} + \sqrt{8HSAL^2 \sum_{i=1}^{k} \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1}(x_{i,j+1})} \\
&\qquad\quad + \frac{7}{3}HSAL^2 + 2\sqrt{3HSAT_kL} + 2\sqrt{84^2H^3S^4A^2L^4} \\
&\qquad\leqslant \sqrt{28HSAT_kL^2} + \frac{7}{3}HSAL^2 \\
&\qquad\quad + \sqrt{8HSAL^2 \sum_{i=1}^{k} \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1}(x_{i,j+1})} + 2\sqrt{84^2H^3S^4A^2L^4},
\end{aligned}
$$

thus demonstrating the result of Equation (45). By following the same procedure, substituting T_k with $HN'_{k,h}(x)$ and $\sum_{i=1}^{k}\sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j}(x_{i,j})$ with $\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j}(x_{i,j})$, we can obtain an upper bound to Equation (46) as:

$$
\begin{aligned}
&\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ},x}, x_{i,h} = x) \sum_{j=h}^{H-1} b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) \\
&\qquad\leqslant \sqrt{28H^2SAN'_{k,h}(x)L^2} + \frac{7}{3}HSAL^2 \\
&\qquad\quad + \sqrt{8HSAL^2 \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1}(x_{i,j+1})} + 2\sqrt{84^2H^3S^4A^2L^4},
\end{aligned}
$$

thus concluding the proof.

■

Appendix D. Proof of Theorem 4.1

Theorem 4.1 (Regret for UCBVI with Chernoff-Hoeffding bonus) *Let $\delta \in (0,1)$. Considering:*[2]

$$b_{k,h}(x,a) = \frac{2HL}{\sqrt{\max\{N_k(x,a), 1\}}},$$

then, w.p. at least $1-\delta$, the regret of UCBVI-CH is bounded by:

$$\text{Reg}(\texttt{UCBVI-CH}, K) \leqslant 10eHL\sqrt{SAT} + \frac{8}{3}eH^2S^2AL^2,$$

where $L = \ln(5HSAT/\delta)$. For $T \geqslant \Omega(H^2S^3A)$, this bound translates to $\widetilde{\mathcal{O}}(H\sqrt{SAT})$.

We begin the proof by demonstrating optimism under the `UCBVI-CH` algorithm (i.e., every optimistic value function is an upper bound of the true optimal value function), which requires us to show that, with high probability, the event $\Omega := \{V_{k,h}(x) \geqslant V^*_h(x), \forall k \in [\![K]\!], h \in [\![H]\!], x \in \mathcal{S}\}$ holds.

Lemma D.1 (Optimism under Chernoff-Hoeffding bonus) *Let the optimistic bonus be defined as:*

$$b_{k,h}(x,a) = \frac{2HL}{\sqrt{N_k(x,a)}}.$$

Then, under event $\mathcal{E}$, the following event holds:

$$\Omega := \{V_{k,h}(x) \geqslant V^*_h(x), \forall k \in [\![K]\!], h \in [\![H]\!], x \in \mathcal{S}\}.$$

Proof We demonstrate the result by induction. Let $V_{k,h}$ be the optimistic value function at stage h computed using the history up to the end of episode $k-1$, and let V^*_h be the true optimal value function at stage h.
By definition, $V_{k,H+1}(x) = V^*_{H+1}(x) = 0$ for every $x \in \mathcal{S}$, and thus the inequality $V_{k,H+1} \geqslant V^*_{H+1}$ trivially holds. To prove the inductive step, we need to demonstrate that, if $V_{k,h+1} \geqslant V^*_{h+1}$ holds, then it also holds that $V_{k,h} \geqslant V^*_h$. We can derive this result as follows:

$$\begin{aligned}
V_{k,h}(x) - V^*_h &= \max_{a\in\mathcal{A}} Q_{k,h}(x,a) - V^*_h(x) \\
&\geqslant Q_{k,h}(x, a^{\pi^*}_{k,h}) - V^*_h(x) \\
&= \sum_{y\in\mathcal{S}} \widehat{P}_k(y|x, a^{\pi^*}_{k,h}) V_{k,h+1}(y) + b_{k,h}(x, a^{\pi^*}_{k,h}) - \sum_{y\in\mathcal{S}} P(y|x, a^{\pi^*}_{k,h}) V^*_{h+1}(y) \\
&\geqslant \sum_{y\in\mathcal{S}} \left[\widehat{P}_k(y|x, a^{\pi^*}_{k,h}) - P(y|x, a^{\pi^*}_{k,h})\right] V^*_{h+1}(y) + b_{k,h}(x, a^{\pi^*}_{k,h}) &\quad (59) \\
&\geqslant b_{k,h}(x, a^{\pi^*}_{k,h}) - 2\sqrt{\frac{H^2 L}{N_k(x_{k,h}, a^{\pi^*}_{k,h})}} &\quad (60) \\
&\geqslant 2\frac{HL}{\sqrt{N_k(x_{k,h}, a^{\pi^*}_{k,h})}} - 2\sqrt{\frac{H^2 L}{N_k(x_{k,h}, a^{\pi^*}_{k,h})}} \\
&\geqslant 0,
\end{aligned}$$

where Equation (59) follows by the inductive hypothesis, and Equation (60) is obtained because, under $\mathcal{E}$, we can bound $|\widehat{P}_k(y|x, a^{\pi^*}_{k,h}) - P(y|x, a^{\pi^*}_{k,h}) V^*_{h+1}(y)|$ by applying Azuma-Hoeffding's inequality, allowing us to simplify terms and show optimism. ■

Our objective is to bound the regret after K episodes (i.e., $\mathrm{Reg}(\texttt{UCBVI-CH}, K)$). We can observe that, under event Ω, it holds that:

$$\mathrm{Reg}(\texttt{UCBVI-CH}, K) = \sum_{k\in[\![K]\!]} V^*_1(x_{k,1}) - V^{\pi_k}_1(x_{k,1})$$

$$
\begin{aligned}
&\leqslant \sum_{k \in [\![K]\!]} V_{k,1}(x_{k,1}) - V_1^{\pi_k}(x_{k,1}) \\
&= \sum_{k \in [\![K]\!]} \widetilde{\Delta}_{k,1}(x_{k,1}) \\
&= \widetilde{\operatorname{Reg}}(\texttt{UCBVI-CH}, K).
\end{aligned}
$$

As such, we can now focus on finding an upper bound to $\widetilde{\operatorname{Reg}}(\texttt{UCBVI-CH}, K)$. By applying Lemma C.2, we can write:

$$
\begin{aligned}
\widetilde{\operatorname{Reg}}(\texttt{UCBVI-CH}, K) &= \sum_{k \in [\![K]\!]} \widetilde{\Delta}_{k,1}(x_{k,1}) \\
&\leqslant e \sum_{i=1}^{K} \sum_{j=1}^{H-1} \Bigg[\varepsilon_{i,j} + 2\sqrt{L}\overline{\varepsilon}_{i,j} + b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) \\
&\quad + \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} \Bigg].
\end{aligned} \tag{61}
$$

To find an upper bound to the regret, we can thus bound the summation of each of the terms individually.
By applying Lemma C.3, we obtain the following bounds:

$$
\begin{aligned}
\sum_{i=1}^{K} \sum_{j=1}^{H-1} \varepsilon_{i,j} &\leqslant 2\sqrt{H^2TL}, \\
\sum_{i=1}^{K} \sum_{j=1}^{H-1} 2\sqrt{L}\overline{\varepsilon}_{i,j} &\leqslant 4\sqrt{TL}.
\end{aligned}
$$

Then, we can derive the following bound:

$$
\begin{aligned}
\sum_{i=1}^{K} \sum_{j=1}^{H} \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} &= \frac{8}{3} H^2SL \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{n=1}^{N_K(x,a)} n^{-1} && (62) \\
&\leqslant \frac{8}{3} H^2SL \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{n=1}^{\frac{KH}{SA}} n^{-1} && (63) \\
&\leqslant \frac{8}{3} H^2S^2AL^2
\end{aligned}
$$

where Equation (62) is obtained by rearranging the terms to isolate the summation of n^{-1} for n from 1 to $N_K(x,a)$ (i.e., the total number of times each state-action pair has been observed up to the end of episode K), and Equation (63) derives from the observation that the summation can be upper bounded by considering a uniform state-action visit distribution. This derivation produces the same result as applying the well-known pigeonhole principle.

By applying a similar reasoning, we bound the remaining summations over the bonus terms:

$$\begin{aligned}\sum_{i=1}^{K}\sum_{j=1}^{H} b_{i,j}(x_{i,j}) &= \sum_{i=1}^{K}\sum_{j=1}^{H} 2H\sqrt{\frac{L}{N_i(x_{i,j}, a_{i,j}^{\pi_i})}}\\ &= 2H\sqrt{L}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_K(x,a)} n^{-1/2}\\ &\leqslant 2H\sqrt{L}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{\frac{KH}{SA}} n^{-1/2}\\ &\leqslant 2\sqrt{H^2SATL},\end{aligned}$$

and over the model error terms:

$$\begin{aligned}\sum_{i=1}^{K}\sum_{j=1}^{H} \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) &\leqslant \sum_{i=1}^{K}\sum_{j=1}^{H} 2H\sqrt{\frac{L}{N_i(x_{i,j}, a_{i,j}^{\pi_i})}} \qquad (64)\\ &= 2H\sqrt{L}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_K(x,a)} n^{-1/2}\\ &\leqslant 2H\sqrt{L}\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{\frac{KH}{SA}} n^{-1/2}\\ &\leqslant 2\sqrt{H^2SATL},\end{aligned}$$

where Equation (64) is obtained by bounding $\xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i})$ using the Chernoff-Hoeffding inequality. Finally, we can put all the bounds together and rewrite Equation (61) as:

$$\begin{aligned}\widetilde{\operatorname{Reg}}(\texttt{UCBVI-CH}, K) &\leqslant e\left[2\sqrt{H^2TL} + 4\sqrt{TL} + 2\sqrt{H^2SATL} + 2\sqrt{H^2SATL} + \frac{8}{3}H^2S^2AL^2\right]\\ &\leqslant e\left[10\sqrt{H^2SATL} + \frac{8}{3}H^2S^2AL^2\right],\end{aligned}$$

thus completing the proof.

Appendix E. Proof of Theorem 4.2

Theorem 4.2 (Regret for `UCBVI` with Bernstein-Freedman bonus) *Let* $\delta \in (0,1)$*. Considering:*[2]

$$b_{k,h}(x,a) = \underbrace{\sqrt{\frac{4L\,\mathbb{V}\mathrm{ar}_{y\sim\widehat{P}_k(\cdot|x,a)}(V_{k,h+1}(y))}{\max\{N_k(x,a),1\}}}}_{\texttt{(A)}} + \underbrace{\frac{7HL}{3\max\{N_k(x,a)-1,1\}}}_{\texttt{(B)}}$$

$$+\underbrace{\sqrt{\frac{4\sum_{y\in\mathcal{S}}\left(\widehat{P}(y|x,a)\cdot\min\left\{\frac{84^2H^3S^2AL^2}{\max\{1,N'_{k,h+1}(y)\}},H^2\right\}\right)}{\max\{N_k(x,a),1\}}}}_{\texttt{(C)}},$$

then, w.p. at least $1-\delta$, the regret of `UCBVI-BF` *is bounded by:*

$$\mathrm{Reg}(\texttt{UCBVI-BF},K)\leqslant 24eL\sqrt{HSAT}+616eH^2S^2AL^2+4e\sqrt{H^2TL},$$

where $L=\ln(5HSAT/\delta)$. For $T\geqslant\Omega(H^3S^3A)$ and $SA\geqslant H$, this bound translates to $\widetilde{\mathcal{O}}(\sqrt{HSAT})$.

Similarly to the proof of Theorem 4.1 in Appendix D, in order to demonstrate the upper bound of `UCBVI-BF`, we first need to demonstrate optimism. However, in order to remove the additional $\sqrt{H}$ term, we are required to both demonstrate optimism as well as to bound by how much the optimistic value function estimator exceeds the true optimal value function. We start by observing that:

$$\begin{aligned}\mathrm{Reg}(\texttt{UCBVI-BF},K)&=\sum_{k\in[\![K]\!]}V_1^*(x_{k,1})-V_1^{\pi_k}(x_{k,1})\\&\leqslant\sum_{k\in[\![K]\!]}V_{k,1}(x_{k,1})-V_1^{\pi_k}(x_{k,1})\\&=\sum_{k\in[\![K]\!]}\widetilde{\Delta}_{k,1}(x_{k,1})\\&=\widetilde{\mathrm{Reg}}(\texttt{UCBVI-BF},K).\end{aligned}$$

According to Lemma C.2, under the events $\mathcal{E}$ and $\Omega_{k,h}$, we can decompose the pseudo-regret as:

$$\sum_{i=1}^{k}\widetilde{\Delta}_{i,h}(x_{i,h})\leqslant e\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left[\varepsilon_{i,j}+2\sqrt{L}\overline{\varepsilon}_{i,j}+b_{i,j}(x_{i,j},a_{i,j}^{\pi_i})+\xi_{i,j}(x_{i,j},a_{i,j}^{\pi_i})+\frac{8H^2SL}{3N_i(x_{i,j},a_{i,j}^{\pi_i})}\right].\tag{65}$$

We also define, by trivially modifying the derivation of Lemma C.2, the pseudo-regret considering only the episodes in which, at stage $h\in[\![H]\!]$ a specific state $x\in\mathcal{S}$ was occupied:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\Delta_{i,h}(x_{i,h})&\leqslant\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\widetilde{\Delta}_{i,h}(x_{i,h})\\&\leqslant e\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\Bigg[\varepsilon_{i,j}+2\sqrt{L}\overline{\varepsilon}_{i,j}+b_{i,j}(x_{i,j},a_{i,j}^{\pi_i})\\&\qquad+\xi_{i,j}(x_{i,j},a_{i,j}^{\pi_i})+\frac{8H^2SL}{3N_i(x_{i,j},a_{i,j}^{\pi_i})}\Bigg].\end{aligned}\tag{66}$$

By applying Lemmas C.3 and C.4, we can upper bound Equations (65) and (66) as:

$$
\begin{aligned}
\sum_{i=1}^{k} \widetilde{\Delta}_{i,h}(x_{i,h}) \leqslant e \sum_{i=1}^{k} \sum_{j=h}^{H-1} &\left[b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} \right] \\
&+ 2e\sqrt{H^2T_kL} + 4e\sqrt{T_kL} \\
&= U_{k,h},
\end{aligned}
$$

and:

$$
\begin{aligned}
\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\widetilde{\Delta}_{i,h}(x_{i,h}) \leqslant e \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} &\left[b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} \right] \\
&+ 2e\sqrt{H^3N'_{k,h}(x)L} + 4e\sqrt{HN'_{k,h}(x)L} \\
&= U_{k,h,x},
\end{aligned}
$$

where we denote the upper bounds of $\sum_{i=1}^{k} \widetilde{\Delta}_{i,h}(x_{i,h})$ and $\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\widetilde{\Delta}_{i,h}(x_{i,h})$ as $U_{k,h}$ and $U_{k,h,x}$, respectively, for ease of notation.
We now demonstrate optimism, which requires us to show that, with high probability, the event $\Omega_{k,h}$ holds.

Lemma E.1 (Optimism under Bernstein-Freedman bonus) *Let the optimistic bonus be defined as:*

$$
\begin{aligned}
b_{k,h}(x,a) = &\sqrt{\frac{4L\,\mathbb{V}\mathrm{ar}_{y\sim\widehat{P}_k(\cdot|x,a)}[V_{k,h+1}(y)]}{N_k(x,a)}} + \frac{7HL}{3(N_k(x,a)-1)} \\
&+ \sqrt{\frac{4\min\{\mathbb{E}_{y\sim\widehat{P}_k(\cdot|x,a)}[\frac{84^2H^3S^2AL^2}{N'_{k,h+1}(y)}], H^2\}}{N_k(x,a)}}.
\end{aligned}
$$

Then, under event $\mathcal{E}$, the following set of events hold:

$$
\Omega_{k,h} := \left\{ V_{i,j}(x) \geqslant V_j^*(x), \forall (i,j) \in [k,h]_{\mathrm{hist}}, x \in \mathcal{S} \right\},
$$

for $k \in [\![K]\!]$ and $h \in [\![H]\!]$, where:

$$
[k,h]_{\mathrm{hist}} := \{(i,j) \in [\![K]\!] \times [\![H]\!] : i < k \vee (i = h, j \geqslant h)\}.
$$

Proof We demonstrate the result by induction. We begin by observing that $V_{k,H+1}(x) = V_{H+1}^*(x) = 0$ for every $k \in [\![K]\!]$ and $x \in \mathcal{S}$. To prove the induction, we need to prove that, if $\Omega_{k,h}$ holds, then also $\Omega_{k,h-1}$ holds. We prove this for a generic $k \in [\![K]\!]$, and we can then apply this procedure for increasing values of k, starting from $k = 1$.
If $\Omega_{k,h}$ holds, then $V_{k,h}(x) \geqslant V_h^*(x)$ for every $x \in \mathcal{S}$. We now bound the estimation error due to the optimistic approach:

$$
\begin{aligned}
V_{k,h}(x) - V_h^*(x) &= \frac{1}{N'_{k,h}(x)} \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)(V_{k,h}(x) - V_h^*(x)) \\
&\leqslant \frac{1}{N'_{k,h}(x)} \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)(V_{i,h}(x) - V_h^{\pi_i}(x)) \qquad (67) \\
&= \frac{1}{N'_{k,h}(x)} \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\widetilde{\Delta}_{i,h}(x_{i,h}), \qquad (68)
\end{aligned}
$$

where Equation (67) follows from the fact that $V_{k,h}$ is monotonically decreasing in k by definition, and by observing that $V_h^* \geqslant V_h^{\pi_i}$.
Recalling the upper bound of $\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\widetilde{\Delta}_{i,h}(x_{i,h})$:

$$
\begin{aligned}
\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\widetilde{\Delta}_{i,h}(x_{i,h}) \leqslant U_{k,h,x} \\
&= e \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) \sum_{j=h}^{H-1} \Bigg[b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) + \xi_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) \\
&\quad + \frac{8H^2SL}{3N_i(x_{i,j}, a_{i,j}^{\pi_i})} \Bigg] + 2e\sqrt{H^3 N'_{k,h}(x)L} + 4e\sqrt{HN'_{k,h}(x)L},
\end{aligned}
$$

we now bound the summations over the terms in the summation over episodes and stages. By applying Lemma C.9, we can bound the summation over typical episodes of the bonus terms as:

$$
\begin{aligned}
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\mathrm{typ},x}, x_{i,h} = x) \sum_{j=h}^{H-1} b_{i,j}(x_{i,j}, a_{i,j}^{\pi_i}) &\leqslant \sqrt{28H^2SAN'_{k,h}(x)L^2} + \frac{7}{3}HSAL^2 \\
&\quad + 2\sqrt{84^2H^3S^4A^2L^4} + \sqrt{8H^2SAL^2U_{k,h,x}},
\end{aligned}
$$

by observing that $\sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\Delta_{i,h}(x_{i,h}) \leqslant U_{k,h,x}$ and that the series of $U_{k,h,x}$ terms is decreasing in h, as each term $U_{k,h,x}$ is a summation of elements each of which includes the next term, and as such we can upper bound $\sum_{j=h}^{H-1} U_{k,j,x}$ with $HU_{k,h,x}$.
In a similar way, we can apply the result of Lemma C.8 to bound the summation over typical episodes of the state-action wise model error terms as:

$$
\begin{aligned}
\sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\mathrm{typ},x}, x_{i,h} = x) \sum_{j=h}^{H-1} \left[\widehat{P}_i(\cdot|x_{i,j}, a_{i,j}^{\pi_i}) - P(\cdot|x_{i,j}, a_{i,j}^{\pi_i})\right]^\top V_{j+1}^*(\cdot) \\
\leqslant \sqrt{6H^2SAN'_{k,h}(x)L^2} + \frac{2}{3}HSAL^2 + 2\sqrt{H^2SAL^2U_{k,h,x}}.
\end{aligned}
$$

With the same procedure as in the proof of Lemma D.1, we obtain the following upper bound:

$$\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{8H^2SL}{3N_i(x_{i,j},a_{i,j}^{\pi_i})}\leqslant\frac{8}{3}H^2S^2AL^2. \tag{69}$$

By combining these result, and accounting for the regret on non-typical episodes, we can write:

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\Delta_{i,h}(x_{i,h})\leqslant U_{k,h,x}&\\ &\leqslant e\Bigg[\sqrt{28H^2SAN'_{k,h}(x)L^2}+\frac{7}{3}HSAL^2+168\sqrt{H^3S^4A^2L^4}\\ &\quad+\sqrt{8H^2SAL^2U_{k,h,x}}+\sqrt{6H^2SAN'_{k,h}(x)L^2}+\frac{2}{3}HSAL^2\\ &\quad+2\sqrt{H^2SAL^2U_{k,h,x}}+\frac{8}{3}H^2S^2AL^2+2\sqrt{H^3N'_{k,h}(x)L}\\ &\quad+4\sqrt{HN'_{k,h}(x)L}+100H^2S^2AL^2\Bigg]\\ &\leqslant e\Bigg[12\sqrt{H^2SAN'_{k,h}(x)L^2}+5\sqrt{H^2SAL^2U_{k,h,x}}\\ &\quad+\frac{821}{3}H^2S^2AL^2+2\sqrt{H^3N'_{k,h}(x)L}\Bigg].\end{aligned}$$

Letting:

$$\begin{aligned}\alpha&=e\left[12\sqrt{H^2SAN'_{k,h}(x)L^2}+\frac{821}{3}H^2S^2AL^2+2\sqrt{H^3N'_{k,h}(x)L}\right],\\ \beta&=5e\sqrt{H^2S^2AL^2},\end{aligned}$$

we can solve for $U_{k,h,x}$ and obtain the following upper bound:

$$U_{k,h,x}\leqslant\beta^2+2\alpha,$$

which we can write as:

$$\begin{aligned}U_{k,h,x}&\leqslant 25e^2H^2S^2AL^2+24e\sqrt{H^2SAN'_{k,h}(x)L^2}+\frac{1642}{3}eH^2S^2AL^2+4e\sqrt{H^3N'_{k,h}(x)L}\\ &\leqslant 24e\sqrt{H^2SAN'_{k,h}(x)L^2}+\frac{1846}{3}eH^2S^2AL^2+4e\sqrt{H^3N'_{k,h}(x)L}\\ &\leqslant 28e\sqrt{H^2SAN'_{k,h}(x)L^2}+\frac{1846}{3}eH^2S^2AL^2 &&(70)\\ &\leqslant 28\cdot\frac{12}{11}e\sqrt{H^3S^2AN'_{k,h}(x)L^2} &&(71)\end{aligned}$$

$$\leqslant 84\sqrt{H^3S^2AN'_{k,h}(x)L^2},$$

where Equation (70) holds if $SA \geqslant H$, and Equation (71) holds under the condition of $[k]_{\text{typ},x}$.
Plugging this result into Equation (68), and observing that the error cannot be greater than H, we get the following upper bound to the estimation error due to the optimistic approach:

$$V_{k,h}(x) - V_h^*(x) \leqslant \min\left\{84\sqrt{\frac{H^3S^2AL^2}{N'_{k,h}(x)}}, H\right\}. \tag{72}$$

Using this result, we now prove that $V_{k,h-1}(x) \geqslant V_{h-1}^*(x)$. Let us recall the definition of $V_{k,h-1}(x)$:

$$V_{k,h-1}(x) = \min\left\{V_{k-1,h-1}(x), H, \mathcal{T}_{h-1}^{\pi_k}V_{k,h}\right\},$$

where $\mathcal{T}_{h-1}^{\pi_k}V_{k,h} := R^{\pi_k}(x_{k,h-1}) + b_{k,h-1}(x_{k,h-1}, a_{k,h-1}^{\pi_k}) + \mathbb{E}_{y\sim\widehat{P}_k(\cdot|x_{k,h-1},a_{k,h-1}^{\pi_k})}V_{k,h}(y)$. Observe that, if $V_{k,h-1}(x) = H$, then the optimism holds trivially. Also, if $V_{k,h-1}(x) = V_{k-1,h-1}(x)$, the optimism holds trivially under $\Omega_{k,h}$. As such, we only need to demonstrate the case in which $V_{k,h-1}(x) = \mathcal{T}_{h-1}^{\pi_k}V_{k,h}$. As such, we derive the following:

$$\begin{aligned}
&V_{k,h-1}(x) - V_{h-1}^*(x) \\
&\qquad = \max_{a\in\mathcal{A}}\left\{R(x,a) + b_{k,h-1}(x,a) + \sum_{y\in\mathcal{S}}\widehat{P}_k(y|x,a)V_{k,h}(y)\right\} \\
&\qquad\quad - R(x, a_{k,h-1}^{\pi^*}) - \sum_{y\in\mathcal{S}}P(y|x, a_{k,h-1}^{\pi^*})V_h^*(y) \\
&\qquad \geqslant b_{k,h-1}(x, a_{k,h-1}^{\pi^*}) + \sum_{y\in\mathcal{S}}\widehat{P}_k(y|x, a_{k,h-1}^{\pi^*})V_{k,h}(y) \\
&\qquad\quad - \sum_{y\in\mathcal{S}}P(y|x, a_{k,h-1}^{\pi^*})V_h^*(y) \\
&\qquad = b_{k,h-1}(x, a_{k,h-1}^{\pi^*}) + \sum_{y\in\mathcal{S}}\widehat{P}_k(y|x, a_{k,h-1}^{\pi^*})\left[V_{k,h}(y) - V_h^*(y)\right] \\
&\qquad\quad + \sum_{y\in\mathcal{S}}\left[\widehat{P}_k(y|x, a_{k,h-1}^{\pi^*}) - P(y|x, a_{k,h-1}^{\pi^*})\right]V_h^*(y) \\
&\qquad \geqslant b_{k,h-1}(x, a_{k,h-1}^{\pi^*}) + \sum_{y\in\mathcal{S}}\left[\widehat{P}_k(y|x, a_{k,h-1}^{\pi^*}) - P(y|x, a_{k,h-1}^{\pi^*})\right]V_h^*(y) \qquad (73)
\end{aligned}$$

where Equation (73) follows from the induction assumption.
Under event $\mathcal{E}$, we can apply the empirical Bernstein inequality (Maurer and Pontil, 2009):

$$\left|\sum_{y\in\mathcal{S}}\left[\widehat{P}_k(y|x,a) - P(y|x,a)\right]V_h^*(y)\right| \leqslant \sqrt{\frac{2\widehat{\mathbb{V}}_{k,h}^*(x,a)L}{N_k(x,a)}} + \frac{7HL}{3(N_k(x,a)-1)},$$

where $\widehat{\mathbb{V}}_{k,h}^*(x,a) := \mathbb{V}\mathrm{ar}_{y\sim\widehat{P}_k(\cdot|x,a)}[V_h^*(y)]$. As such, we obtain:

$$\begin{aligned} V_{k,h-1}(x) - V^*_{h-1}(x) &\geqslant b_{k,h-1}(x, a^{\pi^*}_{k,h-1}) - \sqrt{\frac{2\widehat{\mathbb{V}}^*_{k,h}(x, a^{\pi^*}_{k,h-1})L}{N_k(x, a^{\pi^*}_{k,h-1})}} - \frac{7HL}{3(N_k(x, a^{\pi^*}_{k,h-1}) - 1)} \\ &= \sqrt{\frac{4\widehat{\mathbb{V}}_{k,h}(x, a^{\pi^*}_{k,h-1})L}{N_k(x, a^{\pi^*}_{k,h-1})}} + \sqrt{\frac{4L\mathbb{E}_{y\sim\widehat{P}_k(\cdot|x,a^{\pi^*}_{k,h-1})} b'_{k,h}(y)}{N_k(x, a^{\pi^*}_{k,h-1})}} \\ &\quad - \sqrt{\frac{2\widehat{\mathbb{V}}^*_{k,h}(x, a^{\pi^*}_{k,h-1})L}{N_k(x, x, a^{\pi^*}_{k,h-1})}}. \end{aligned} \tag{74}$$

We now bound $\widehat{\mathbb{V}}^*_{k,h}$ in terms of $\widehat{\mathbb{V}}_{k,h}$. Observing that:

$$\begin{aligned} \mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X - \mathbb{E}[X]]^2 \\ &= \mathbb{E}[X \pm Y - \mathbb{E}[X] \pm \mathbb{E}[Y]]^2 \\ &= \mathbb{E}[(X - Y) - \mathbb{E}[X - Y] + Y - \mathbb{E}[Y]]^2 \\ &\leqslant 2\mathbb{E}[(X - Y) - \mathbb{E}[X - Y]]^2 + 2\mathbb{E}[Y - \mathbb{E}[Y]]^2 \\ &= \mathbb{V}\mathrm{ar}[X - Y] + 2\,\mathbb{V}\mathrm{ar}[Y], \end{aligned}$$

we can then rewrite:

$$\begin{aligned} \widehat{\mathbb{V}}^*_{k,h}(x, a^{\pi^*}_{k,h-1}) &\leqslant 2\widehat{\mathbb{V}}_{k,h}(x, a^{\pi^*}_{k,h-1}) + 2 \underset{y\sim\widehat{P}_k(\cdot|x,a^{\pi^*}_{k,h-1})}{\mathbb{V}\mathrm{ar}}[V^*_h(y) - V_{k,h}(y)] \\ &\leqslant 2\widehat{\mathbb{V}}_{k,h}(x, a^{\pi^*}_{k,h-1}) + 2\sum_{y\in\mathcal{S}} \widehat{P}_k(\cdot|x, a^{\pi^*}_{k,h-1})(V_{k,h}(y) - V^*_y)^2. \end{aligned}$$

By plugging this result into Equation (74), we get:

$$V_{k,h-1}(x) - V^*_{h-1}(x) \geqslant \sqrt{\frac{4L\mathbb{E}_{y\sim\widehat{P}_k(\cdot|x,a^{\pi^*}_{k,h-1})} b'_{k,h}(y)}{N_k(x, a^{\pi^*}_{k,h-1})}} - \sqrt{\frac{4L\mathbb{E}_{y\sim\widehat{P}_k(\cdot|x,a^{\pi^*}_{k,h-1})}(V_{k,h}(y) - V^*_h(y))^2}{N_k(x, a^{\pi^*}_{k,h-1})}}.$$

By applying the result of Equation (72) and the definition of $b'_{k,h}(y)$, we finally obtain that $V_{k,h-1}(x) - V^*_{h-1}(x) \geqslant 0$, thus demonstrating optimism. ■

Having demonstrated optimism, we now prove the upper bound of the regret $\mathrm{Reg}(\texttt{UCBVI-BF}, K)$:

$$\widetilde{\mathrm{Reg}}(\texttt{UCBVI-BF}, K) \leqslant U_{K,1}$$

$$
\begin{aligned}
&= e\bigg[\sqrt{28HSATL^2} + \frac{7}{3}HSAL^2 + 2\sqrt{84^2H^3S^4A^2L^4} \\
&\quad + \sqrt{8HSAL^2U_{K1}} + \sqrt{6HSATL^2} + \frac{2}{3}HSAL^2 \\
&\quad + 2\sqrt{HSAL^2U_{K,1}} + \frac{8}{3}H^2S^2AL^2 + 2\sqrt{H^2TL} \\
&\quad + 4\sqrt{TL} + 100H^2S^2AL\bigg] \qquad (75) \\
&\leqslant e\bigg[12\sqrt{HSATL^2} + 5\sqrt{H^2SAL^2U_{K,1}} \\
&\quad + \frac{821}{3}H^2S^2AL^2 + 2\sqrt{H^2TL}\bigg] \qquad (76)
\end{aligned}
$$

where Equation (75) is obtained by applying the results of Lemmas C.9 and C.8, by applying the result of Equation (69), and by accounting for the regret of non-typical episodes.
As done in Lemma E.1, by letting:

$$
\begin{aligned}
\alpha &= e\left[12\sqrt{HSATL^2} + \frac{821}{3}H^2S^2AL^2 + 2\sqrt{H^2TL}\right], \\
\beta &= 5e\sqrt{H^2S^2AL^2},
\end{aligned}
$$

we can solve for $U_{K,1}$ and obtain:

$$
\begin{aligned}
\widetilde{\mathrm{Reg}}(\texttt{UCBVI-BF}, K) &\leqslant 24e\sqrt{HSATL^2} + \frac{1846}{3}eH^2S^2AL^2 + 4e\sqrt{H^2TL} \\
&\leqslant 24e\sqrt{HSATL^2} + 616eH^2S^2AL^2 + 4e\sqrt{H^2TL}
\end{aligned}
$$

thus completing the proof.