Appendices

We begin by introducing some notation in Sect. B and Sect. A. We then provide the full analysis of UCBVI in Sect. C.

A. Table of Notation

Symbol	Explanation
$\mathcal{S}$	The state space
$\mathcal{A}$	The action space
π_k	The policy at episode k
P	The transition distribution
R	The reward function
S	Size of state space
A	Size of action space
H	The horizon length
T and T_k	The total number of steps and number of steps up to episode k
K	The total number of episodes
$N_k(x,a)$	Number of visits to state-action pair (x,a) up to episode k
V_h^*	Optimal value function V^*
$\mathcal{T}$	Bellman operator
$V_{k,h}$	The estimate of value function at step h of episode k
$Q_{k,h}$	The estimate of action value function at step h of episode k
b	The exploration bonus
L	$\ln(5SAT/\delta)$
$N_k(x,a,y)$	Number of transitions from x to y upon taking action a up to episode k
$N'_{k,h}(x,a)$	Number of visits to state-action pair (x,a) at step h up to episode k
$N'_{k,h}(x)$	Number of visits to state x at step h up to episode k
$\widehat{P}_k(y\|x,a)$	The empirical transition distribution from x to y upon taking action a up to episode k
$\widehat{\mathbb{V}}_{k,h}(x,a)$	The empirical next-state variance of $V_{k,h}$ for every (x,a)
$\mathbb{V}_h^*(x,a)$	The next-state variance of V^* for every state-action pair (x,a)
$\widehat{\mathbb{V}}_{k,h}^*(x,a)$	The empirical next-state variance of V_h^* for every state-action pair (x,a) at episode k
$\mathbb{V}_h^\pi(x,a)$	The next-state variance of V_h^π for every state-action pair (x,a)
$b'_{i,j}(x)$	$\min\left(\frac{100^2 S^2 H^2 AL^2}{N'_{i,j}(x)}, H^2\right)$
$[(x,a)]_k$	Set of typical state-action pairs
$[k]_{\text{typ}}$	Set of typical episodes
$[y]_{k,x,a}$	Set of typical next states at every episode k for every (x,a)
$\mathrm{Regret}(k)$	The regret after k episodes
$\widetilde{\mathrm{Regret}}(k)$	The upper-bound regret after k episodes
$\mathrm{Regret}(k,x,h)$	The regret upon encountering state x at step h after k episodes
$\widetilde{\mathrm{Regret}}(k,x,h)$	The regret upon encountering state x at step h after k episodes
$\Delta_{k,h}$	One step regret at step h of episode k
$\widetilde{\Delta}_{k,h}$	One step upper-bound regret at step h of episode k
$\widetilde{\Delta}_{\text{typ},k,h}$	One step upper-bound regret at step h of episode k for typical episodes
$\mathcal{M}_t$	The martingale operator
$\varepsilon_{k,h}$ and $\bar{\varepsilon}_{k,h}$	Martingale difference terms
$c_1(v,n)$, $c_2(p,n)$ and $c_3(n)$	The confidence intervals for the value function and transition distribution
$C_{k,h}$	Sum of confidence intervals c_1 up to step h of episode k
$B_{k,h}$	Sum of exploration bonuses b up to step h of episode k
$\mathcal{E}$	The high probability event under which the concentration inequality holds
Ω	The high probability event under which the estimates $V_{k,h}$ are ucbs
$\mathcal{H}_t$	The history of all random events up to time step t

B. Notation

Let denote the total number of times that we visit state x while taking action a at step h of all episodes up to episode k by $N'_{k,h}(x,a)$. We also use the notation $N'_{k,h}(x) = \sum_{a\in\mathcal{A}} N'_{k,h}(x,a)$ for the total number of visits to state x at time step h up to episode k. Also define the empirical next-state variance $\widehat{\mathbb{V}}_{k,h}(x,a)$, the next-state variance of optimal value function $\mathbb{V}^*_h(x,a)$ the next-state empirical variance of optimal value function $\widehat{\mathbb{V}}^*_{k,h}(x,a)$ and the next-state variance of V^π_h as

$$\begin{aligned}
\widehat{\mathbb{V}}_{k,h}(x,a) &\stackrel{\text{def}}{=} \mathrm{Var}_{y\sim\widehat{P}_k(\cdot|x,a)}(V_{k,h+1}(y)),\\
\mathbb{V}^*_h(x,a) &\stackrel{\text{def}}{=} \mathrm{Var}_{y\sim P(\cdot|x,a)}(V^*_h(y)),\\
\widehat{\mathbb{V}}^*_{k,h}(x,a) &\stackrel{\text{def}}{=} \mathrm{Var}_{y\sim\widehat{P}_k(\cdot|x,a)}(V^*_h(y)),\\
\mathbb{V}^\pi_h(x,a) &\stackrel{\text{def}}{=} \mathrm{Var}_{y\sim P(\cdot|x,a)}(V^\pi_h(y)).
\end{aligned}$$

for every $(x,a)\in\mathcal{S}\times\mathcal{A}$ and $k\in[K]$ and $h\in[H]$. We further introduce some short-hand notation: we use the lower case to denote the functions evaluated at the current state-action pair, e.g., we write $n_{k,h}$ for $N_k(x_{k,h},\pi_k(x_{k,h},h))$ and $v_{k,h}$ for $V_{k,h}(x_{k,h})$. Let also denote $\mathbb{V}^*_{k,h} = \mathbb{V}^*_{k,h}(x_{k,h},\pi_k(x_{k,h},h))$ and $\mathbb{V}^{\pi_k}_{k,h} = \mathbb{V}^{\pi_k}_{k,h}(x_{k,h},\pi_k(x_{k,h},h))$ for every $k\in[K]$ and $h\in[H]$. Also define $b'_{i,j}(x) = \min\left(\frac{100^2S^2H^2AL^2}{N'_{i,j+1}(x)}, H^2\right)$ for every $x\in\mathcal{S}$.

B.1. "Typical" state-actions and steps

In our analysis we split the episodes into 2 sets: the set of "typical" episodes in which the number of visits to the encountered state-actions are large and the rest of the episodes. We then prove a tight regret bound for the typical episodes. As the total count of other episodes is bounded this technique provides us with the desired result. The set of typical state-actions pairs for every episode k is defined as follows

$$[(x,a)]_k \stackrel{\text{def}}{=} \{(x,a) : (x,a)\in\mathcal{S}\times\mathcal{A}, N_h(x,a)\geq H, N'_{k,h}(x)\geq H\}.$$

Based on the definition of $[(x,a)]_{\text{typ}}$ we define the set of typical episodes and the set of typical state-dependent episodes as follow

$$\begin{aligned}
[k]_{\text{typ}} &\stackrel{\text{def}}{=} \{i : i\in[k], \forall h\in[H], (x_{i,h},\pi_i(x_{i,h},h))\in[(x,a)]_k, i\geq 250HS^2AL\},\\
[k]_{\text{typ},x} &\stackrel{\text{def}}{=} \{i : i\in[k], \forall h\in[H], (x_{i,h},\pi_i(x_{i,h},h))\in[(x,a)]_k, N'_{k,h}(x)\geq 250HS^2AL\}.
\end{aligned}$$

Also for every $(x,a)\in\mathcal{S}\times\mathcal{A}$ the set of typical next states at every episode k is defined as follows

$$[y]_{k,x,a} \stackrel{\text{def}}{=} \{y : y\in\mathcal{S}, N_k(x,a)P(y|x,a)\geq 2H^2L\}.$$

Finally let denote $[y]_{k,h} = [y]_{k,x_{k,h},\pi_k(x_{x_{k,h}})}$ for every $k\in[K]$ and $h\in[H]$.

B.2. Surrogate regrets

Our ultimate goal is to prove bound on the regret $\mathrm{Regret}(k)$. However in our analysis we mostly focus on bounding the surrogate regrets. Let $\widetilde{\Delta}_{k,h}(x) \stackrel{\text{def}}{=} V_{k,h}(x) - V^{\pi_k}_h(x)$ for every $x\in\mathcal{S}$, $h\in[H]$ and $k\in[K]$. Then the upper-bound regret $\widetilde{\mathrm{Regret}}$ defined as follows

$$\widetilde{\mathrm{Regret}}(k) \stackrel{\text{def}}{=} \sum_{i=1}^{k}\widetilde{\delta}_{i,1}.$$

$\widetilde{\mathrm{Regret}}(k)$ is useful in our analysis since it provides an upperbound on the true regret $\mathrm{Regret}(k)$. So one can bound $\widetilde{\mathrm{Regret}}(k)$ as a surrogate for $\mathrm{Regret}(k)$.

We also define the corresponding per state-step regret and upper-bound regret for every state $x \in \mathcal{X}$ and step $h \in [H]$, respectively, as follows

$$\begin{aligned}\mathrm{Regret}(k,x,h) &\stackrel{\mathrm{def}}{=} \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\delta_{i,h},\\ \widetilde{\mathrm{Regret}}(k,x,h) &\stackrel{\mathrm{def}}{=} \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\widetilde{\delta}_{i,h}.\end{aligned}$$

B.3. Martingale difference sequences

In our analysis we rely heavily on the theory of martingale sequences to prove bound on the regret incurred due to encountering a random sequence of states. We now provide some definitions and notation in that regard.

We define the following martingale operator for every $k \in [K]$, $h \in [H]$ and $F : \mathcal{S} \to \Re$. Also let $t = (k-1)H + h$ denote the time stamp at step h of episode k then

$$\mathcal{M}_t F \stackrel{\mathrm{def}}{=} P_h^{\pi_k} F - F(x_{k,h+1}).$$

Let $\mathcal{H}_t$ be the history of all random events up to (and including) step h of episode k then we have that $\mathbb{E}(\mathcal{M}_t F | \mathcal{H}_t) = 0$. Thus $\mathcal{M}_t F$ is a martingale difference w.r.t. $\mathcal{H}_t$. Also let G be a real-value function depends on H_{t+s} for some integer $s > 0$. Then we generalize our definition of operator $\mathcal{M}_t$ as

$$\mathcal{M}_t G \stackrel{\mathrm{def}}{=} \mathbb{E}\left(G(\mathcal{H}_{t+s}) \middle| \mathcal{H}_t\right) - G(\mathcal{H}_{t+s}),$$

where $\mathbb{E}$ is over the randomization of the sequence of states generated by the sequence of policies $\pi_k, \pi_{k+1}, \ldots$. Here also $\mathcal{M}_t G$ is a martingale difference w.r.t. $\mathcal{H}_t$.

Let define $\Delta_{\mathrm{typ},k,h} : \mathcal{S} \to \Re$ as follows for every $k \in [K]$ and $h \in [H]$ and $y \in \mathcal{S}$

$$\widetilde{\Delta}_{\mathrm{typ},k,h+1}(y) \stackrel{\mathrm{def}}{=} \sqrt{\frac{\mathbb{I}_{k,h}(y)}{n_{k,h} p_{k,h}(y)}}\widetilde{\Delta}_{k,h+1}(y),$$

where the function $p_{k,h} : \mathcal{S} \to [0,1]$ is defined as $p_{k,h}(y) = P_h^{\pi_k}(y|x_{k,h})$ and $\mathbb{I}_{k,h}(y)$ writes for $\mathbb{I}_{k,h}(y) = \mathbb{I}(y \in [y]_{k,h})$ for every $y \in \mathcal{X}$. We also define the following martingale differences which we use frequently

$$\begin{aligned}\varepsilon_{k,h} &\stackrel{\mathrm{def}}{=} \mathcal{M}_t \widetilde{\Delta}_{k,h+1},\\ \bar{\varepsilon}_{k,h} &\stackrel{\mathrm{def}}{=} \mathcal{M}_t \widetilde{\Delta}_{\mathrm{typ},k,h+1}.\end{aligned}$$

B.4. High probability events

We now introduce the high probability events $\mathcal{E}$ and $\Omega_{k,h}$ under which the regret is small.

Let use the shorthand notation $L \stackrel{\mathrm{def}}{=} \ln\left(\frac{5SAT}{\delta}\right)$. Also for every $v > 0$, $p \in [0,1]$ and $n > 0$ let define the confidence intervals c_1, c_2 and c_3, respectively, as follow

$$
\begin{aligned}
c_1(v,n) &\stackrel{\text{def}}{=} 2\sqrt{\frac{vL}{n}} + \frac{14HL}{3n}, \\
c_2(p,n) &\stackrel{\text{def}}{=} 2\sqrt{\frac{p(1-p)L}{n}} + \frac{2L}{3n}, \\
c_3(n) &\stackrel{\text{def}}{=} 2\sqrt{\frac{SL}{n}}.
\end{aligned}
$$

Let $\mathcal{P}$ be the set of all probability distributions on $\mathcal{S}$. Define the following confidence set for every $k = 1, \ldots, K$, $n > 0$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$

$$
\begin{aligned}
\mathcal{P}(k,h,n,x,a,y) \stackrel{\text{def}}{=} \Big\{ & \widetilde{P}(\cdot|x,a) \in \mathcal{P} : |(\widetilde{P} - P)V_h^*(x,a)| \le \min\left(c_1\left(\mathbb{V}_h^*(x,a),n\right), c_1\left(\widehat{\mathbb{V}}_{k,h}^*(x,a),n\right)\right) \\
& |\widetilde{P}(y|x,a) - P(y|x,a)| \le c_2\left(P(y|x,a),n\right), \\
& \|\widetilde{P}(\cdot|x,a) - P(\cdot|x,a)\|_1 \le c_3(n) \Big\}.
\end{aligned}
$$

We now define the random event $\mathcal{E}_{\widehat{P}}$ as follows

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

Let t be a positive integer. Let $\mathcal{F} = \{f_s\}_{s \in [t]}$ be a set of real-value functions on $\mathcal{H}_{t+s}$, for some integer $s > 0$. We now define the following random events for every $\bar{w} > 0$ and $\bar{u} > 0$ and $\bar{c} > 0$:

$$
\begin{aligned}
\mathcal{E}_{\text{az}}(\mathcal{F},\bar{u},\bar{c}) &\stackrel{\text{def}}{=} \left\{ \sum_{s=1}^{t} \mathcal{M}_s f_s \le 2\sqrt{t\bar{u}^2\bar{c}} \right\}, \\
\mathcal{E}_{\text{fr}}(\mathcal{F},\bar{w},\bar{u},\bar{c}) &\stackrel{\text{def}}{=} \left\{ \sum_{s=1}^{t} \mathcal{M}_s f_s \le 4\sqrt{\bar{w}\bar{c}} + \frac{14\bar{u}\bar{c}}{3} \right\}.
\end{aligned}
$$

We also use the short-hand notation $\mathcal{E}_{\text{az}}(\mathcal{F},\bar{u})$ and $\mathcal{E}_{\text{fr}}(\mathcal{F},\bar{w},\bar{u})$ for $\mathcal{E}_{\text{az}}(\mathcal{F},\bar{u},L)$ and $\mathcal{E}_{\text{fr}}(\mathcal{F},\bar{w},\bar{u},L)$, respectively.

Now let define the following sets of random variables for every $k \in [K]$ and $h \in [H]$:

$$
\begin{aligned}
\mathcal{F}_{\widetilde{\Delta},k,h} &\stackrel{\text{def}}{=} \left\{ \widetilde{\Delta}_{i,j} : i \in [k], h < j \in [H-1] \right\}, \\
\mathcal{F}'_{\widetilde{\Delta},k,h} &\stackrel{\text{def}}{=} \left\{ \widetilde{\Delta}_{\text{typ},i,j} : i \in [k], h < j \in [H] \right\}, \\
\mathcal{F}_{\widetilde{\Delta},k,h,x} &\stackrel{\text{def}}{=} \left\{ \widetilde{\Delta}_{i,j}\mathbb{I}(x_{i,h} = x) : i \in [k], h < j \in [H] \right\}, \\
\mathcal{F}'_{\widetilde{\Delta},k,h,x} &\stackrel{\text{def}}{=} \left\{ \widetilde{\Delta}_{\text{typ},i,j}\mathbb{I}(x_{i,h} = x) : i \in [k], h < j \in [H] \right\}, \\
\mathcal{G}_{\mathbb{V},k,h} &\stackrel{\text{def}}{=} \left\{ \sum_{j=h+1}^{H} \mathbb{V}_j^{\pi_i} : i \in [k], h < j \in [H] \right\}, \\
\mathcal{G}_{\mathbb{V},k,h,x} &\stackrel{\text{def}}{=} \left\{ \sum_{j=h+1}^{H} \mathbb{V}_j^{\pi_i}\mathbb{I}(x_{i,h} = x) : i \in [k], h < j \in [H] \right\}, \\
\mathcal{F}_{b',k,h} &\stackrel{\text{def}}{=} \left\{ b'_{i,j} : i \in [k], h < j \in [H-1] \right\}, \\
\mathcal{F}_{b',k,h,x} &\stackrel{\text{def}}{=} \left\{ b'_{i,j}\mathbb{I}(x_{i,h} = x) : i \in [k], h < j \in [H] \right\}.
\end{aligned}
$$

We now define the high probability event $\mathcal{E}$ as follows

$$\mathcal{E} \stackrel{\text{def}}{=} \mathcal{E}_{\widehat{P}} \bigcap \bigcap_{\substack{k\in[K]\\ h\in[H]\\ x\in\mathcal{S}}} \Big[\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h}, H) \bigcap \mathcal{E}_{\text{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h}, 1/\sqrt{L}) \bigcap \mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H) \bigcap \mathcal{E}_{\text{az}}(\mathcal{F}'_{\widetilde{\Delta},k,x,h}, 1/\sqrt{L})$$
$$\bigcap \mathcal{E}_{\text{fr}}(\mathcal{G}_{\mathbb{V},k,h}, H^4T, H^3) \bigcap \mathcal{E}_{\text{az}}(\mathcal{G}_{\mathbb{V},k,h,x}, H^5 N'_{k,h}(x), H^3) \bigcap \mathcal{E}_{\text{az}}(\mathcal{F}_{b',k,h}, H^2) \bigcap \mathcal{E}_{\text{az}}(\mathcal{F}_{b',k,h,x}, H^2) \Big].$$

The following lemma shows that the event $\mathcal{E}$ holds with high probability:

Lemma 1. *Let $\delta > 0$ be a real scalar. Then the event $\mathcal{E}$ holds w.p. at least $1-\delta$.*

Proof. To prove this result we need to show that a set of concentration inequalities with regard to the empirical model $\widehat{P}_k$ holds simultaneously. For every $h \in [H]$ the Bernstein inequality combined with a union bound argument, to take into account that $N_k(x,a) \in [T]$ is a random number, leads to the following inequality w.p. $1-\delta$ (see, e.g., Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012, for the statement of the Bernstein inequality and the application of the union bound in similar cases, respectively.)

$$\left| \left[(P - \widehat{P}_k) V_h^* \right](x,a) \right| \quad \leq \quad \sqrt{\frac{2\mathbb{V}_h^*(x,a) \ln\left(\frac{2T}{\delta}\right)}{N_k(x,a)}} + \frac{2H \ln\left(\frac{2T}{\delta}\right)}{3N_k(x,a)}, \tag{9}$$

where we rely on the fact that V_h^* is uniformly bounded by H. Using the same argument but this time with the Empirical Bernstein inequality (see, e.g., Maurer & Pontil, 2009), for $N_k(x,a) > 1$, leads to

$$\left| \left[(P - \widehat{P}_k) V_h^* \right](x,a) \right| \quad \leq \quad \sqrt{\frac{2\widehat{\mathbb{V}}_{k,h}^*(x,a) \ln\left(\frac{2T}{\delta}\right)}{N_k(x,a)}} + \frac{7H \ln\left(\frac{2T}{\delta}\right)}{3N_k(x,a)}. \tag{10}$$

The Bernstein inequality combined with a union bound argument on $N_k(x,a)$ also implies the following bound w.p. $1-\delta$

$$|N_k(y,x,a) - N_k(x,a)P(y|x,a)| \leq \sqrt{2N_k(x,a)\text{Var}_{z\sim P(\cdot|x,a)}(\mathbf{1}(z=y)) \ln\left(\frac{2T}{\delta}\right)} + \frac{2\ln\left(\frac{2T}{\delta}\right)}{3},$$

which implies the following bound w.p. $1-\delta$:

$$\left| \widehat{P}_k(y|x,a) - P(y|x,a) \right| \quad \leq \quad \sqrt{\frac{P(y|x,a)(1-P(y|x,a)) \ln\left(\frac{2T}{\delta}\right)}{N_k(x,a)}} + \frac{2\ln\left(\frac{2T}{\delta}\right)}{3N_k(x,a)}. \tag{11}$$

A similar result holds on ℓ_1-normed estimation error of the transition distribution. The result of (Weissman et al., 2003) combined with a union bound on $N_k(x,a) \in [T]$ implies w.p. $1-\delta$

$$\left\| \widehat{P}_k(\cdot|x,a) - P(\cdot|x,a) \right\|_1 \quad \leq \quad \sqrt{\frac{2S \ln\left(\frac{2T}{\delta}\right)}{N_k(x,a)}}. \tag{12}$$

We now focus on bounding the sequence of martingales. Let $n > 0$ be an integer and $u, \delta > 0$ be some real scalars. Let the sequence of random variables $\{X_1, X_2, \ldots, X_n\}$ be a sequence of martingale differences w.r.t. to some filtration $\mathcal{F}_n$. Let

this sequence be uniformly bounded from above and below by u. Then the Azuma's inequality (see, e.g., Cesa-Bianchi & Lugosi, 2006) implies that w.p. $1-\delta$

$$\sum_{i=1}^{n} X_i \quad \leq \quad \sqrt{2nu \ln\left(\frac{1}{\delta}\right)}. \tag{13}$$

When the sum of the variances $\sum_{i=1}^{n} \mathrm{Var}(X_i|\mathcal{F}_i) \leq w$ for some $w > 0$ then the following sharper bound due to Freedman (1975) holds w.p. $1-\delta$

$$\sum_{i=1}^{n} X_i \quad \leq \quad \sqrt{2w \ln\left(\frac{1}{\delta}\right)} + \frac{2u \ln\left(\frac{1}{\delta}\right)}{3}. \tag{14}$$

Let $k \in [K]$, $h \in [H]$ and $x \in \mathcal{X}$. Then the inequality of Eq. 13 immediately implies that the following events holds w.p. $1-\delta$:

$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{\widetilde{\Delta},k,h}, H, \ln\left(1/\delta\right)\right), \tag{15}$$
$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}'_{\widetilde{\Delta},k,h}, 1/\sqrt{L}, \ln\left(1/\delta\right)\right), \tag{16}$$
$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{b',k,h}, H^2, \ln\left(1/\delta\right)\right). \tag{17}$$

Also Eq. 13 combined with a union bound argument over all $N'_{k,h}(x) \in [T]$ (see, e.g., Bubeck et al., 2011, for the full description of the application of union bound argument in the case of martigale process with random stopping time) implies that the following events hold w.p. $1-\delta$

$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H, \ln\left(T/\delta\right)\right), \tag{18}$$
$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}'_{\widetilde{\Delta},k,h,x}, 1/\sqrt{L}, \ln\left(T/\delta\right)\right), \tag{19}$$
$$\mathcal{E}_{\mathrm{az}}\left(\mathcal{F}_{b',k,h,x}, H^2, \ln\left(T/\delta\right)\right). \tag{20}$$

Similarly the inequality of Eq. 14 leads to the following events hold w.p. $1-\delta$

$$\mathcal{E}_{\mathrm{fr}}\left(\mathcal{G}_{\mathbb{V},k,h}, \bar{w}_{k,h}, , H^3, \ln\left(T/\delta\right)\right), \tag{21}$$
$$\mathcal{E}_{\mathrm{fr}}\left(\mathcal{G}_{\mathbb{V},k,h,x}, \bar{w}_{k,h,x}, , H^3, \ln\left(1/\delta\right)\right), \tag{22}$$

where $\bar{w}_{k,h}$ and $\bar{w}_{k,h,x}$ are upper bounds on $W_{k,h}$ and $W_{k,h,x}$, respectively, defined as

$$W_{k,h} \quad \stackrel{\mathrm{def}}{=} \quad \sum_{i=1}^{k} \mathrm{Var}\left(\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \middle| \mathcal{H}_{i,1}\right), \tag{23}$$

$$W_{k,h,x} \quad \stackrel{\mathrm{def}}{=} \quad \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)\mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \middle| \mathcal{H}_{i,1}\right). \tag{24}$$

So to establish a value for $\bar{w}_{k,h}$ and $\bar{w}_{k,h,x}$ we need to prove bound on $W_{k,h}$ and $W_{k,h,x}$. Here we only prove this bound for $W_{k,h}$ as the proof techniques to bound $W_{k,h,x}$ is identical to the way we bound $W_{k,h}$.

$$W_{k,h} \leq \sum_{i=1}^{k} \mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}_{i,j+1}^{\pi_k} \middle| \mathcal{H}_k \right)^2 \leq H^3 \sum_{i=1}^{k} \mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}_{i,j+1}^{\pi_k} \middle| \mathcal{H}_k \right). \tag{25}$$

Now let the sequence $\{x_1, x_2, \ldots, x_H\}$ be the sequence of states encountered by following some policy π throughout an episode k. Then the recursive application of LTV leads to (see e.g., Munos & Moore, 1999; Lattimore & Hutter, 2012, for the proof.)

$$\mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}^{\pi}(x_j, \pi(x_j, j)) \right) \quad = \quad \mathrm{Var}\left(\sum_{j=h}^{H-1} r^{\pi}(x_j) \right). \tag{26}$$

By combining Eq. 26 into Eq. 25 we deduce

$$W_{k,h} \quad \leq \quad H^3 \sum_{i=1}^{k} \mathrm{Var}\left(\sum_{j=h}^{H-1} r_{k,h} \middle| \mathcal{H}_k \right) \leq H^5 k = H^4 T_k. \tag{27}$$

Similarly the following bound holds on $W_{k,h,x}$

$$W_{k,h,x} \quad \leq \quad H^5 N_{k,h}(x). \tag{28}$$

Plugging the bounds of Eq. 27 and Eq. 28 in to the bounds of Eq. 21 and Eq. 22 and a union bound over all $N_{k,h}(x) \in [T]$ leads to the following events hold w.p. $1-\delta$:

$$\mathcal{E}_{\mathrm{fr}}\left(\mathcal{G}_{\mathbb{V},k,h}, H^4 T, H^3, \ln\left(1/\delta\right)\right), \tag{29}$$
$$\mathcal{E}_{\mathrm{fr}}\left(\mathcal{G}_{\mathbb{V},k,h,x}, H^5 N_{k,h}(x), H^3, \ln\left(T/\delta\right)\right). \tag{30}$$

Combining the results of Eq. 9, Eq. 10, Eq. 11, Eq. 12, Eq. 15, Eq. 16 Eq. 17, Eq. 18, Eq. 19, Eq. 20, Eq. 29 and Eq. 30 and taking a union bound over these random events as well as all possible $k \in [K]$, $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ proves the result.

□

B.4.1. UCB EVENTS

Let $k \in [K]$ and $h \in [H]$. Denote the set of steps for which the value functions are obtained before $V_{k,h}$ as

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

Let $\Omega_{k,h} = \{V_{i,j} \geq V_h^*, \forall (i,j) \in [k,h]_{\mathrm{hist}}\}$ be the event under which $V_{i,j}$ prior to $V_{k,h}$ computation are upper bounds on the optimal value functions. Using backward induction on h (and standard concentration inequalities) we will prove that $\Omega_{k,h}$ holds under the event $\mathcal{E}$ (see Lem. 19).

B.5. Other useful notation

Here we define some other notation that we use throughout the proof. We denote the total count of steps up to episode $k \in [K]$ by $T_k \stackrel{\text{def}}{=} H(k-1)$. We first define $c_{4,k,h}$, for every $h \in [H]$ and $k \in [K]$, as follow

$$c_{4,k,h} = \frac{4H^2SAL}{n_{k,h}}.$$

for every $k \in [K]$, $h \in [H]$ and $x \in [x]$ we also introduce the following notation which we use later when we sum up the regret:

$$\begin{aligned}
C_{k,h} &\stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} c_{1,i,j}, \\
B_{k,h} &\stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ}}) \sum_{j=h}^{H-1} b_{i,j}, \\
C_{k,h,x} &\stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ},x}, x_{k,h} = x) \sum_{j=h}^{H-1} c_{1,i,j}, \\
B_{k,h,x} &\stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathbb{I}(i \in [k]_{\text{typ},x}, x_{k,h} = x) \sum_{j=h}^{H-1} b_{i,j},
\end{aligned}$$

where $c_{1,k,h}$ is the shorthand-notation for $c_1(v^*_{k,h}, n_{k,h})$. We also define the upper bound $U_{k,h}$ and $U_{k,h,x}$ for every $k \in [K]$, $h \in [H]$ and $x \in \mathcal{S}$ as follows, respectively

$$\begin{aligned}
U_{k,h} &\stackrel{\text{def}}{=} e \sum_{i=1}^{k} \sum_{j=h}^{H-1} [b_{i,j} + c_{1,i,j} + c_{4,i,j}] + (H+1)\sqrt{T_k L}, \\
U_{k,h,x} &\stackrel{\text{def}}{=} e \sum_{i=1}^{k} \sum_{j=h}^{H-1} [b_{i,j} + c_{1,i,j} + c_{4,i,j}] + (H+1)^{3/2}\sqrt{N'_{k,h}(x) L},
\end{aligned}$$

C. Proof of the Regret Bounds

Before we start the main analysis we state the following useful lemma that will be used frequently in the analysis:

Lemma 2. *let $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ be two random variables. Then following bound holds for their variances*

$$\mathit{Var}(X) \leq 2\left[\mathit{Var}(Y) + \mathit{Var}(X - Y)\right].$$

Proof. The following sequence of inequalities hold

$$\text{Var}(X) = \mathbb{E}(X - Y - \mathbb{E}(X - Y) + Y - \mathbb{E}(Y))^2 \leq 2\mathbb{E}(X - Y - \mathbb{E}(X - Y))^2 + 2\mathbb{E}(Y - \mathbb{E}(Y))^2.$$

The result follows from the definition of variance. □

We proceed by proving the following key lemma which shows that proves bound on $\Delta_{k,h}$ under the assumption that $V_{k,h}$ is UCB w.r.t. V^*_h.

Lemma 3. *Let $k \in [K]$ and $h \in [H]$. Let the events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then the following bound holds on $\delta_{k,h}$ and $\widetilde{\delta}_{k,h}$:*

$$\delta_{k,h} \le \widetilde{\delta}_{k,h} \le e \sum_{i=h}^{H-1} \left[\varepsilon_{k,i} + 2\sqrt{L}\bar{\varepsilon}_{k,i} + c_{1,k,i} + b_{k,i} + c_{4,k,i} \right]. \tag{31}$$

Proof. For the ease of exposition we abuse the notation and drop the dependencies on k, e.g., we write x_1, π and V_1 for $x_{k,1}$, π_k and $V_{k,1}$, respectively. We proceed by bounding $\widetilde{\delta}_h$ under the event $\mathcal{E}$ at every step $0 < h < H$:

$$\begin{aligned}
\widetilde{\delta}_h &= \mathcal{T}_h V_{h+1}(x_h) - \mathcal{T}_h^\pi V_{h+1}^\pi(x_h) \\
&= [\widehat{P}_h^\pi V_{h+1}](x_h) + b_h - [P_h^\pi V_{h+1}^\pi](x_h) \\
&= b_h + [(\widehat{P}_h^\pi - P_h^\pi)V_{h+1}^*](x_h) + [(\widehat{P}_h^\pi - P_h^\pi)(V_{h+1} - V_{h+1}^*)](x_h) + [P_h^\pi(V_{h+1} - V_{h+1}^\pi)](x_h) \\
&\le \widetilde{\delta}_{h+1} + \varepsilon_h + b_h + c_{1,h} + \underbrace{[(\widehat{P}_h^\pi - P_h^\pi)(V_{h+1} - V_{h+1}^*)](x_h)}_{(a)},
\end{aligned} \tag{32}$$

where the last inequality follows from the fact that under the event $\mathcal{E}$ we have that $[(\widehat{P}_h^\pi - P_h^\pi)V_{h+1}^*](x_h) \le c_{1,h}$. We now bound (a):

$$\begin{aligned}
(a) &= \sum_{y \in \mathcal{S}} (\widehat{P}_h^\pi(y|x_h) - P_h^\pi(y|x_h))(V_{h+1}(y) - V_{h+1}^*(y)) \\
&\overset{(I)}{\le} \sum_{y \in \mathcal{S}} \left[2\sqrt{\frac{p_h(y)(1-p_h(y))L}{n_h}} + \frac{4L}{3n_h} \right] \Delta_{h+1}(y) \\
&\le \underbrace{2\sqrt{L} \sum_{y \in \mathcal{S}} \sqrt{\frac{p_h(y)}{n_h}} \widetilde{\Delta}_{h+1}(y)}_{(b)} + \frac{4SHL}{3n_h},
\end{aligned}$$

where (I) holds under the event $\mathcal{E}$. We proceed by bounding (b):

$$(b) = \underbrace{\sum_{y \in [y]_h} \sqrt{\frac{p_h(y)}{n_h}} \widetilde{\Delta}_{h+1}(y)}_{(c)} + \underbrace{\sum_{y \notin [y]_h} \sqrt{\frac{p_h(y)}{n_h}} \widetilde{\Delta}_{h+1}(y)}_{(d)}. \tag{33}$$

The term (c) can be bounded as follows

$$\begin{aligned}
(c) &= \sum_{y \in [y]_h} P_h^\pi(y|x_h) \sqrt{\frac{1}{n_h p_h(y)}} \widetilde{\Delta}_{h+1}(y) = \bar{\varepsilon}_h + \sqrt{\frac{1}{n_h p_h(x_{h+1})}} \mathbb{I}(x_{h+1} \in [y]_h) \widetilde{\delta}_{h+1} \\
&\le \bar{\varepsilon}_h + \sqrt{\frac{1}{4LH^2}} \widetilde{\delta}_{h+1},
\end{aligned} \tag{34}$$

where in the last line we rely on the definition of $[y]_h$. We now bound (d):

$$(d) = \sum_{y \notin [y]_h} \sqrt{\frac{p_h(y) n_h}{n_h^2}} \widetilde{\Delta}_{h+1}(y) \le \frac{SH\sqrt{4LH^2}}{n_h}. \tag{35}$$

By combining Eq. 34 and Eq. 35 into Eq. 33 we deduce

$$(b) \leq \frac{SH\sqrt{4LH^2}}{n_h} + \sqrt{\frac{1}{4LH^2}}\widetilde{\delta}_{h+1} + \bar{\varepsilon}_h. \tag{36}$$

By combining Eq. 36 and Eq. 33 into Eq. C we deduce

$$\widetilde{\delta}_h \leq \varepsilon_h + 2\sqrt{L}\bar{\varepsilon}_h + b_h + c_{1,h} + c_{4,h} + \left(1 + \frac{1}{H}\right)\widetilde{\delta}_{h+1}.$$

Let denote $\gamma_h = (1 + 1/H)^h$. The previous bound combined with an induction argument implies that

$$\widetilde{\delta}_h \quad \leq \sum_{i=h}^{H-1} \gamma_{i-h}\left[\varepsilon_i + 2\sqrt{L}\bar{\varepsilon}_i + c_{1,i} + c_{4,i} + b_i\right].$$

The inequality $\ln(1 + x) \leq x$ for every $x > -1$ leads to $\gamma_h \leq \gamma_H \leq e$ for every $h \in [H]$. This combined with the assumption that $v_h \geq v_h^*$ under the event Ω_h completes the proof. □

Lemma 4. *Let $k \in [k]$ and $h \in [H]$. Let the events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then*

$$\sum_{i=1}^{k-1}\delta_{i,h} \leq \sum_{i=1}^{k-1}\widetilde{\delta}_{i,h} \leq e\sum_{i=1}^{k-1}\sum_{j=h}^{H-1}\left[\varepsilon_{i,j} + 2\sqrt{L}\bar{\varepsilon}_{i,j} + b_{i,j} + c_{1,i,j} + c_{4,i,j}\right].$$

Proof. The proof follows by summing up the bounds of Lem. 3 and taking into acoount the fact if $\Omega_{k,h}$ holds then $\Omega_{i,j}$ for all $(i, j) \in [k, h]_{hist}$ hold. □

To simplify the bound of Lem. 4 we prove bound on sum of the martingales $\varepsilon_{k,h}$ and $\bar{\varepsilon}_{k,h}$

Lemma 5. *Let $k \in [k]$ and $h \in [H]$. Let the events $\mathcal{E}$ and $\Omega_{k,h}$ hold. Then the following bound holds*

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\varepsilon_{i,j} \quad \leq \quad H\sqrt{(H-h)kL} \leq H\sqrt{T_k L}, \tag{37}$$

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\bar{\varepsilon}_{i,j} \quad \leq \quad \sqrt{(H-h)k} \leq \sqrt{T_k}. \tag{38}$$

Also the following bounds holds for every $x \in \mathcal{X}$ and $h \in \mathcal{H}$:

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h} = x)\sum_{j=h}^{H-1}\varepsilon_{i,j} \quad \leq \quad H\sqrt{(H-h)N'_{k,h}(x)L}, \tag{39}$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h} = x)\sum_{j=h}^{H-1}\bar{\varepsilon}_{i,j} \quad \leq \quad \sqrt{(H-h)N'_{k,h}(x)}. \tag{40}$$

Proof. The fact that the event $\mathcal{E}$ holds implies that the events $\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h}, H)$, $\mathcal{E}_{\text{az}}(\mathcal{F}'_{\widetilde{\Delta},k,h}, \frac{1}{\sqrt{L}})$, $\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H)$ and $\mathcal{E}_{\text{az}}(\mathcal{F}'_{\widetilde{\Delta},x,k,h}, \frac{1}{\sqrt{L}})$ hold. Under these events the inequalities of the statement hold. This combined with the fact that $(H - h)k \leq T_k$ completes the proof.

□

We now bound the sum of δs in terms of the upper-bound U:

Lemma 6. *Let $k \in [K]$ and $h \in [H]$. Let the events $\mathcal{E}$ and $\Omega_{k,h}$ holds. Then the following bounds hold for every $h \in [H]$ $x \in \mathcal{S}$*

$$\begin{aligned}\sum_{i=1}^{k} \delta_{i,h} &\leq \sum_{i=1}^{k} \widetilde{\delta}_{i,h} \leq U_{k,h} \leq U_{k,1}, \\ \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\delta_{i,h} &\leq \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\widetilde{\delta}_{i,h} \leq U_{k,h,x}. \leq U_{k,1,x}.\end{aligned}$$

Proof. The proof follows by incorporating the result of Lem. 5 into Lem. 4 and taking into account that for every $h \in [H]$ the term $U_{k,h}$ ($U_{k,h,x}$) is a summation of non-negative terms which are also contained in $U_{k,1}$ ($U_{1,h,x}$). □

Lemma 7. *Let $k \in [K]$ and $h \in [H]$. Let the events $\mathcal{E}$ and $\Omega_{k,h}$ holds. Then the following bounds hold for every $x \in \mathcal{S}$*

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=h}^{H} \delta_{i,j} &\leq \sum_{i=1}^{k}\sum_{j=h}^{H} \widetilde{\delta}_{i,j} \leq HU_{k,1}, \\ \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H} \delta_{i,j} &\leq \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H} \widetilde{\delta}_{i,j} \leq HU_{k,1,x}.\end{aligned}$$

Proof. The proof follows by summing up the bounds of Lem. 6. □

We now focus on bounding the terms $C_{k,h}$ ($C_{k,h,x}$) and $B_{k,h}$ ($B_{k,h,x}$) in Lem. 11 and Lem. 12, respectively. Before we proceed with the proof of Lem. 11 and Lem. 12. we prove the following key result which bounds sum of the variances of $V^{\pi}_{k,h}$ using an LTV argument:

Lemma 8. *Let $k \in [K]$ and $h \in [H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$*

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} &\leq T_k H + 2\sqrt{H^4 T_k L} + \frac{4H^3 L}{3}, \\ \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} &\leq N'_{k,h}(x)H^2 + 2\sqrt{H^5 N'_{k,h}(x) L} + \frac{4H^3 L}{3}.\end{aligned}$$

Proof. Under $\mathcal{E}$ the events $\mathcal{E}_{\text{fr}}(\mathcal{G}_{\mathbb{V},k,h}, H^4 T_k, H^3)$ and $\mathcal{E}_{\text{fr}}(\mathcal{G}_{\mathbb{V},k,h,x}, H^5 N_{k,h}(x), H^3)$ hold which then imply:

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \leq \sum_{i=1}^{k} \mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \middle| \mathcal{H}_{k,h}\right) + 2\sqrt{H^4 T_k L} + \frac{4H^3 L}{3}, \tag{41}$$

$$\sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \leq \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x)\mathbb{E}\left(\sum_{j=h}^{H-1} \mathbb{V}^{\pi}_{i,j+1} \middle| \mathcal{H}_{k,h}\right) + 2\sqrt{H^5 N'_{k,h} L} + \frac{4H^3 L}{3}. \tag{42}$$

The LTV argument of Eq. 26 then leads to

$$\sum_{i=1}^{k}\mathbb{E}\left(\sum_{j=h}^{H-1}\mathbb{V}^{\pi}_{i,j+1}\middle|\mathcal{H}_{i,h}\right) = \sum_{i=1}^{k}\mathrm{Var}\left(\sum_{j=h+1}^{H}r^{\pi}_{i,j}\right)\leq KH^2 = TH, \tag{43}$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathbb{E}\left(\sum_{j=h}^{H-1}\mathbb{V}^{\pi}_{i,j+1}\middle|\mathcal{H}_{i,h}\right) = \sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\mathrm{Var}\left(\sum_{j=h+1}^{H}r^{\pi}_{i,j}\right)\leq N'_{k,h}(x)H^2. \tag{44}$$

Eq. 41 and Eq. 42 combined with Eq. 43 and Eq. 44, respectively, complete the proof.

□

Lemma 9. *Let $k\in[K]$ and $h\in[H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x\in\mathcal{S}$*

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1}\left(\mathbb{V}^{*}_{i,j+1}-\mathbb{V}^{\pi}_{i,j+1}\right) \leq 2H^2U_{k,h}+4H^2\sqrt{T_kL}, \tag{45}$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\left(\mathbb{V}^{*}_{i,j+1}-\mathbb{V}^{\pi}_{i,j+1}\right) \leq 2H^2U_{k,h,x}+4H^2\sqrt{HN'_{k,h}(x,a)L}. \tag{46}$$

Proof. We begin by the following sequence of inequalities:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{*}_{i,j+1}-\mathbb{V}^{\pi}_{i,j+1} &\overset{(I)}{\leq} \sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim p_{i,j}}\left[\left(V^{*}_{i,j+1}(y)\right)^2-\left(V^{\pi_i}_{i,j+1}(y)\right)^2\right]\\ &= \sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim p_{i,j}}\left[(V^{*}_{j+1}(y)-V^{\pi_i}_{j+1}(y)(V^{*}_{j+1}(y)+V^{\pi_i}_{j+1}(y))\right]\\ &\leq 2H\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{E}_{y\sim p_{i,j}}\left(V^{*}_{j+1}(y)-V^{\pi_i}_{j+1}(y)\right)}_{(a)},\end{aligned} \tag{47}$$

where (I) is obtained from the definition of the variance as well as the fact that $V^{*}_{i,j}\geq V^{\pi}_{k,h}$. The last line also follows from the fact that $V^{\pi_k}\leq V^{*}_{h}\leq H$.

Using an identical argument we can also prove the following bound for state-dependent difference:

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{V}^{*}_{i,j+1}-\mathbb{V}^{\pi}_{i,j+1} \leq 2H\underbrace{\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1}\mathbb{E}_{y\sim p_{i,j}}\left(V^{*}_{j+1}(y)-V^{\pi_i}_{j+1}(y)\right)}_{(b)}, \tag{48}$$

To bound (a) we use the fact that under the event $\mathcal{E}$ the event $\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h},H)$ also holds. This combined with the fact that under the event $\Omega_{k,h}$ the inequality $\delta_{k,h}\leq\widetilde{\delta}_{k,h}$ holds implies that

$$\begin{aligned}(a) &\leq \sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\delta}_{i,j+1}+2H\sqrt{T_kL}\\ &\leq HU_{1,h}+2H\sqrt{T_kL},\end{aligned} \tag{49}$$

where in the last line we rely on the result of Lem. 7. Similarly we can prove the following bound for (b) under the events $\Omega_{k,h}$ and $\mathcal{E}_{\text{az}}(\mathcal{F}_{\widetilde{\Delta},k,h,x}, H)$:

$$\begin{aligned}(b) &\leq \sum_{i=1}^{k} \mathbb{I}(x_{i,h}=x) \sum_{j=h}^{H-1} \widetilde{\Delta}_{i,j+1} + 2H^{1.5}\sqrt{N'_{k,h}(x)L} \\ &\leq HU_{k,h,x} + 2H^{1.5}\sqrt{N'_{k,h}(x)L}. \end{aligned} \tag{50}$$

The result then follows by incorporating the results of Eq. 49 and Eq. 50 into Eq. 47 and Eq. 48, respectively.

□

Lemma 10. *Let $k \in [K]$ and $h \in [H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$*

$$\sum_{i=1}^{k}\sum_{j=h}^{H-1} \widehat{\mathbb{V}}_{i,j+1} - \mathbb{V}^{\pi}_{i,j+1} \leq 2H^2U_{k,1} + 15H^2S\sqrt{AT_kL}, \tag{51}$$

$$\sum_{i=1}^{k}\mathbb{I}(x_{i,h}=x)\sum_{j=h}^{H-1} \widehat{\mathbb{V}}_{i,j+1} - \mathbb{V}^{\pi}_{i,j+1} \leq 2H^2U_{k,h,x} + 15H^2S\sqrt{HAN'_{k,h}(x)L}. \tag{52}$$

Proof. Here we only prove the bound on Eq. 51. The proof for the bound of Eq. 52 can be done in a very similar manner, as it is shown in the previous lemmas (the only difference is that $HN'_{k,h}(x)$ and $U_{k,h,x}$ replace T_k and $U_{k,1}$, respectively). The following sequence of inequalities hold:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=h}^{H-1} \widehat{\mathbb{V}}_{i,j+1} - \mathbb{V}^{\pi}_{i,j+1} &\overset{(I)}{\leq} \sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{E}_{y\sim\widehat{p}_{i,j}}\left(V_{i,j+1}(y)\right)^2 - \mathbb{E}_{y\sim p_{i,j}}\left(V^{\pi_i}_{j+1}(y)\right)^2 \\ &\quad + \sum_{i=1}^{k}\sum_{j=h}^{H-1} \left(\mathbb{E}_{y\sim p_{i,j}} V^{*}_{j+1}(y)\right)^2 - \left(\mathbb{E}_{y\sim\widehat{p}_{i,j}} V^{*}_{j+1}(y)\right)^2 \\ &\overset{(II)}{\leq} \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{E}_{y\sim\widehat{p}_{i,j}}\left(V_{i,j+1}(y)\right)^2 - \sum_{j=1}^{H-1}\mathbb{E}_{y\sim p_{i,j}}\left(V_{i,j+1}(y)\right)^2}_{(a)} \\ &\quad + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1} \mathbb{E}_{y\sim p_{i,j}}\left[\left(V_{i,j+1}(y)\right)^2 - \left(V^{\pi_i}_{j+1}(y)\right)^2\right]}_{(b)} \\ &\quad + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1} 4H^2\sqrt{\frac{L}{n_{k,h}}}}_{(c)}, \end{aligned} \tag{53}$$

where (I) holds due to the fact that under $\Omega_{k,h}$, $V_{i,j} \geq V^*_j \geq V^{\pi_i}_j$ and (II) holds under the event $\mathcal{E}$.

We now bound (a):

$$(a) \overset{(I)}{\leq} \sum_{i=1}^{k}\sum_{j=h}^{H-1} 2H^2\sqrt{\frac{SL}{n_{k,h}}} \overset{(II)}{\leq} 3H^2S\sqrt{AT_kL},$$

where (I) holds under the event $\mathcal{E}$ and (II) holds due to the pigeon-hole argument (see, e.g., Jaksch et al., 2010, for the proof).

Using an identical analysis to the one in Lem. 10 and taking into account that $V_{i,j} \geq V^*_j$ under the event $\Omega_{k,h}$ and $\mathcal{E}$ we can bound (b)

$$(b) \overset{(I)}{\leq} 2H\left(\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widetilde{\delta}_{i,j+1} + 2H\sqrt{T_kL})\right) \leq 2H^2\left(U_{k,1} + 2\sqrt{T_kL}\right),$$

where (I) holds since under the event $\mathcal{E}$ the event $\mathcal{E}_{az}(\mathcal{F}_{\widetilde{\Delta},k,h}, H)$ holds. Another application of pigeon-hole principle leads to a bound of $6H^2\sqrt{SAT_kL}$ on (c). We then combine this with the bounds on (a) and (b) to bound Eq. 53, which proves the result.

□

We now bound $C_{k,h}$ and $C_{k,h,x}$:

Lemma 11. *Let $k \in [K]$ and $h \in [H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$*

$$C_{k,h} \leq 4\sqrt{HSAT_k} + 4\sqrt{H^2U_{k,1}SAL^2}, \tag{54}$$

$$C_{k,h,x} \leq 4\sqrt{H^2SAN_{k,h}(x)} + 4\sqrt{H^2U_{k,h,x}SAL^2}. \tag{55}$$

Proof. Here we only prove the bound on Eq. 54. The proof for the bound of Eq. 55 can be done in a very similar manner, as it is shown in the previous lemmas (the only difference is that $HN'_{k,h}(x)$ and $U_{k,h,x}$ replace T_k and $U_{k,1}$, respectively). The Cauchy–Schwarz inequality leads to the following sequence of inequalities:

$$\begin{aligned} C_{k,h} &= \sum_{i=1}^{k}\mathbb{I}(j \in [k]_{\text{typ}})\left(\sum_{j=h}^{H-1} 2\sqrt{\frac{\mathbb{V}^*_{i,j+1}L}{n_{i,j}}} + \frac{4HL}{3n_{i,j}}\right) \\ &\leq 2\sqrt{L}\sqrt{\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^*_{i,j+1}}_{(a)}}\sqrt{\underbrace{\sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ}})\sum_{j=h}^{H-1}\frac{1}{n_{i,j}}}_{(b)}} + \sum_{i=1}^{k}\mathbb{I}(i \in [k]_{\text{typ}})\sum_{h=j}^{H-1}\frac{4HL}{3n_{i,j}} \end{aligned} \tag{56}$$

We now prove bounds on (a) and (b) respectively

$$(a) = \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{\pi}_{i,j+1}}_{(c)} + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^*_{i,j+1} - \mathbb{V}^{\pi}_{i,j+1}}_{(d)}. \tag{57}$$

(c) and (d) can be bounded under the events $\mathcal{E}$ and $\Omega_{k,h}$ using the results of Lem. 8 and Lem.9. We then deduce

$$\begin{aligned}(a) &\leq HT_k + 2H^2U_{k,1} + 6H^2\sqrt{T_kL} + \frac{4H^2L}{3} \\ &\leq 2HT_k + 2H^2U_{k,1},\end{aligned}$$

where the last line follows by the fact that for the typical episodes $T_k \geq 250H^2S^2AL^2$. Thus if $T_k \leq H^2L$ the term $C_{k,h}$ trivially equals to 0 otherwise the higher order terms are bounded by $O(HT_k)$.

We now bound (b) using a pigeon-hole argument

$$(b) \leq 2\sum_{(x,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_k(x,a)}\frac{1}{n} \leq 2SA\sum_{n=1}^{T}\frac{1}{n} \leq 2SA\ln(3T).$$

Plugging the bound on (a) and (b) into Eq. 56 and taking in to account that for the typical episodes $[k]_{\text{typ}}$ we have that $T \geq H^2L$ completes the proof.

□

We now bound $B_{k,h}$:

Lemma 12. *Let $k \in [K]$ and $h \in [H]$. Let the bonus is defined according to Algo. 4. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$,*

$$B_{k,h} \leq 11L\sqrt{T_kHSA} + 12\sqrt{H^2SAL^2U_{k,1}} + 570H^2S^2AL^2, \tag{58}$$

$$B_{k,h,x} \leq 11L\sqrt{N'_{k,h}(x)HSA} + 12\sqrt{H^2SAL^2U_{k,h,x}} + 570H^2S^2AL^2, \tag{59}$$

Proof. Here we only prove the bound on Eq. 58. The proof for the bound of Eq. 59 can be done in a very similar manner, as it is shown in the previous lemmas (the only difference is that $HN'_{k,h}(x)$ and $U_{k,h,x}$ replace T_k and $U_{k,1}$, respectively). We first notice that the following holds:

$$B_{k,h} \leq \underbrace{\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\sqrt{\frac{8\widehat{\mathbb{V}}_{i,j+1}L}{n_{i,j}}}}_{(a)} + \underbrace{L\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\left(\sqrt{\frac{8}{n_{i,j}}\sum_{y\in\mathcal{S}}\widehat{p}_{i,j}(y)\min\left(\frac{100^2S^2H^2AL^2}{N'_{i,j+1}(y)}, H^2\right)}\right)}_{(b)}.$$

We first note that the bound on $B_{k,h}$ is similar to the bound on $C_{k,h}$. The main difference (beside the difference in H.O.Ts) is that here $\mathbb{V}^*_{h+1}$ is replaced by $\widehat{\mathbb{V}}_{i,j+1}$. So in our proof we first focus on dealing with this difference.

The Cauchy–Schwarz inequality leads to:

$$(a) \leq \sqrt{8L}\sqrt{\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1}(x,a)}_{(c)}}\sqrt{\underbrace{\sum_{i=1}^{k}\mathbb{I}(k\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\frac{1}{n_{i,j}}}_{(d)}},$$

The bound on (d) is identical to the corresponding bound in Lem. 11. So we only focus on bounding (c):

$$(c) = \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\mathbb{V}^{\pi}_{i,j+1}}_{(e)} + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\widehat{\mathbb{V}}_{i,j+1} - \mathbb{V}^{\pi}_{i,j+1}}_{(f)}. \tag{60}$$

(e) and (f) can be bounded in high probability using the results of Lem. 8 and Lem.10. This implies

$$\begin{aligned}(c) &\leq HT_k + 3H^2U_{k,1} + 15H^2S\sqrt{AT_kL} + \frac{4H^2L}{3}\\ &\leq 2HT_k + 3H^2U_{k,1},\end{aligned}$$

where the last line follows by the fact that for the typical episodes $T_k \geq 250H^2S^2AL$. Thus if $T_k \leq 250H^2S^2L$ then $B_{k,h}$ trivially equals to 0 otherwise the higher order terms are bounded by $O(HT)$. Combining the bound on (b) and (c) leads to the following bound on (a):

$$(a) \leq 8L\sqrt{HSAT_k} + 12HL\sqrt{SAU_{k,1}}.$$

To bound (b) we make use of Cauchy-Schwarz inequality again.

$$(b) \leq \sqrt{\underbrace{8\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}\sum_{y\in\mathcal{S}}\widehat{p}_{i,j}(y)b'_{i,j+1}(y)}_{(g)}\underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}\frac{\mathbb{I}(i\in[k]_{\text{typ}})}{n_{i,j}}}_{(h)}}.$$

The term (h) bounded by $2SAL$ using a pigeon-hole argument (see Lem. 11). We proceed by bounding (g):

$$(g) \leq \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}(\widehat{p}_{i,j}-p_{i,j})b'_{i,j+1}}_{(i)} + \underbrace{\sum_{i=1}^{k}\sum_{j=h}^{H-1}(p_{i,j}b'_{\mathbb{V}} - b'_{i,j+1}(x_{i,j+1}))}_{(j)} + \underbrace{\sum_{i=1}^{k}\mathbb{I}(i\in[k]_{\text{typ}})\sum_{j=h}^{H-1}b'_{i,j+1}(x_{i,j+1})}_{k}.$$

Given that the event $\mathcal{E}$ holds the term (i) bounded by $2\sqrt{2}H^2S\sqrt{ALT_k}$ by using the pigeon-hole argument. Under the event $\mathcal{E}$ the event $\mathcal{E}_{az}(\mathcal{F}_{b',k,h}, H^2)$ holds. This implies that the term (j) is also bounded by $2H^2\sqrt{T_kL}$ as it is sum of the martingale differences. The term (k) is also bounded by $20000H^3S^3A^3L^3$ using the pigeon-hole argument. Combining all these bounds together leads to the following bound on (b)

$$(b) \leq \sqrt{32\sqrt{2}H^2S^2\sqrt{T_kAL^3} + 32H^2SA\sqrt{T_kL^3} + 320000S^4H^4A^2L^3}.$$

Combining this with the bound on (a) and taking into account the fact that we only bound the $B_{k,h}$ for the typical episodes, in which $T_k \geq 250H^2S^2AL^2$, completes the proof.

□

Lemma 13. *Let the bonus is defined according to Algo. 4. Then under the events $\mathcal{E}$ and $\Omega_{K,1}$ the following hold*

$$\textit{Regret}(K) \leq \widetilde{\textit{Regret}}(K) \leq U_{K,1} \leq 15L\sqrt{HSAT} + 16HL\sqrt{SAU_{k,1}} + 820H^2S^2A^2L + 2H\sqrt{TL}. \tag{61}$$

Proof. We first notice that Regret(K) and Regret(K) are bounded by $U_{k,1}$ due to Lem.6. To bound $U_{k,1}$ we sum up the regret due to $B_{k,h}$ and $C_{k,h}$ from Lem. 11 and Lem. 12. We also bound the sum $\sum_{k=1}^{K}\sum_{h=1}^{H} c_{4,k,h}$ by $2HSAL$ using a pigeon hole argument. We also note that $B_{k,h}$ and $C_{k,h}$ only account for the regret of typical episodes in which $T \geq H^2S^2A^2L$. The regret of those episodes which do not belong to the typical set $[k]_{\text{typ}}$, can be bounded by $O(H^2S^2A^2L^2)$, trivially. □

The following lemma establishes an explicit bound on the regret:

Lemma 14. *Let the bonus is defined according to Algo. 4. Then under the events $\mathcal{E}$ and $\Omega_{K,1}$ the following hold*

$$\textit{Regret}(K) \leq \widetilde{\textit{Regret}}(K) \leq U_{K,1} \leq 30L\sqrt{HSAT} + 2500H^2S^2AL^2 + 4H\sqrt{TL}. \tag{62}$$

Proof. The proof follows by solving the bound of Lem. 13 in terms of $U_{k,1}$. which only contributes to the additional regret of $O(H^2L^2SA)$. □

Lemma 15. *Let the bonus is defined according to Algo. 3. Then under the events $\mathcal{E}$ and $\Omega_{K,1}$ the following holds*

$$\textit{Regret}(K) \leq \widetilde{\textit{Regret}}(K) \leq U_{K,1} \leq 20HL\sqrt{SAT} + 250H^2S^2AL^2. \tag{63}$$

Proof. The proof up to Lem. 11 is identical to the proof of Lem. 14. The main difference is to prove bound on $C_{k,h}$ and $B_{k,h}$ here we use a loose bound of $O(H\sqrt{\frac{SAL}{n_{k,h}}})$ for both exploration bonus $b_{k,h}$ and the confidence interval $c_{1,k,h}$ and then sum these terms using a pigeon-hole argument (The proof is provided in Jaksch et al., 2010) which leads to a bound of $O(H\sqrt{SATL})$ on both $B_{K,1}$ and $C_{K,1}$. Plugging these results into the bound of Lem. 7 combined with the regret of non-typical episodes complete the proof

□

Lemma 16. *Let the bonus is defined according to Algo. 4. Let $k \in [K]$ and $h \in [H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$,*

$$\begin{aligned}\textit{Regret}(k,x,h) &\leq \widetilde{\textit{Regret}}(k,x,h) \leq 30HL\sqrt{SAN'_{k,h}(x)} + 2500H^2S^2AL^2 + 4H^{1.5}\sqrt{N'_{k,h}(x)L} \\ &\leq 100H^{1.5}SL\sqrt{AN'_{k,h}(s)}.\end{aligned}$$

Proof. The proof is similar to the proof of total regret. Here also we use Lem. 12, Lem. 11 and a pigeon-hole argument to bound the regrets due to $B_{k,h}$, $C_{k,h}$ and $c_{4,k,h}$. We then incorporate these terms into Lem.6 to bound the regret in terms of $U_{k,h,x}$. The result follows by solving the bound w.r.t. the upper bound $U_{k,h,x}$. □

Lemma 17. *Let the bonus b is defined according to Algo. 4. Let $k \in [K]$ and $h \in [H]$. Then under the events $\mathcal{E}$ and $\Omega_{k,h}$ the following hold for every $x \in \mathcal{S}$*

$$V_{k,h}(x) - V^*_h(x) \leq 100\sqrt{\frac{H^3S^2AL^2}{N'_{k,h}(s)}}.$$

Proof. From Lem. 16 we have that

$$
\begin{aligned}
& 100H^{1.5}SL\sqrt{AN'_{k,h}(s)} \\
\geq & \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x)(V_{i,h}(x) - V_h^{\pi_i}(x)) \\
\geq & (V_{k,h}(x) - V_h^*(x)) \sum_{i=1}^{k} \mathbb{I}(x_{i,h} = x) = N'_{k,h}(x)(V_{k,h}(x) - V_h^*(x)),
\end{aligned}
$$

where the last inequality holds due to the fact that $V_{k,h}$ by definition is monotonically non-increasing in k. The proof then follows by collecting terms.

□

Lemma 18. *Let the bonus b is defined according to Algo. 3. Then under the event $\mathcal{E}$ the set of events $\{\Omega_{k,h}\}_{k\in[K],h\in H}$ hold.*

Proof. We prove this result by induction. First we notice that for $h = H$ by definition $V_{k,h} = V_h^*$ thus the inequality $V_{k,h} \geq V_h^*$ trivially holds. Thus to prove this result for $h < H$ we only need to show that if the inequality $V_{k,h} \geq V_h^*$ holds for h it also holds for $h-1$ for every $h < H$:

$$
\begin{aligned}
V_{k,h}(x) - V_h^*(x) &= \mathcal{T}_k V_{k,h-1}(x) - \mathcal{T}V^*(x) \geq b_k(x, \pi_h^*(x)) + \widehat{P}_{k,h}^{\pi^*} V_{k,h+1}(x) - P_h^{\pi^*} V_{k,h+1}(x) \\
&= b_k(x, \pi_h^*(x)) + \widehat{P}_{k,h}^{\pi^*}(V_{k,h+1} - V_{h+1}^*)(x) + (\widehat{P}_{k,h}^{\pi^*} - P_h^{\pi^*})V_{h+1}^*(x) \\
&\geq b_k(x, \pi_h^*(x)) + (\widehat{P}_{k,h}^{\pi^*} - P_h^{\pi^*})V_{h+1}^*(x),
\end{aligned}
$$

where the last line follows by the induction condition that $V_{k,h+1} \geq V_{h+1}^*$. The fact that the event $\mathcal{E}$ hols implies that $(P_h^{\pi^*} - \widehat{P}_{k,h}^{\pi^*})V_{h+1}^*(x) \leq c_1(N_k(x, \pi_h^*(x))) \leq b_k(x, \pi_h^*(x))$, which completes the proof.

□

Lemma 19. *Let the bonus b is defined according to Algo. 4. Then under the event $\mathcal{E}$ the set of events $\{\Omega_{k,h}\}_{k\in[K],h\in H}$ hold.*

Proof. We prove this result by induction. We first notice that in the case of the first episode $V_{1,h} = H \geq V_h^*$.

To prove this result by induction in the case of $1 < k \in [K]$ we need to show that in the case of $h \in [H-1]$ if $\Omega_{k,h+1}$ holds then $\Omega_{k,h}$ also holds.

If $\Omega_{k,h-1}$ holds then $V_{i,j} \geq V_j^*$ for every $(i,j) \in [k,h]_{\text{hist}}$. We can then invoke the result of Lem. 17 which implies

$$
V_{k,h+1}(x) - V_{h+1}^*(x) \leq \frac{100H^{1.5}SL\sqrt{A}}{\sqrt{N'_{k,h+1}(x)}}.
$$

Using this result which guarantees that $V_{k,h+1}$ is close to V_{h+1}^* we prove that $V_{k,h} - V_h^* \geq 0$, that is the event $\Omega_{k,h}$ holds.

$$
V_{k,h} - V_h^* = \min(V_{k-1,h}, \mathcal{T}_{k,h} V_{i,j+1}, H) - V_h^*
$$

If $V_{k-1,h} \leq \mathcal{T}_{k,h} V_{i,j+1}$ the result $V_{k,h} - V_h^* = V_{k-1,h} - V_h^* \geq 0$ holds trivially. Also if $V_{k-1,h} \geq H$ the result trivially holds. So we only need to consider the case that $\mathcal{T}_{k,h} V_{i,j+1} \leq V_{k-1,h} \leq H$ in that case we have w

$$\begin{aligned} V_{k,h}(x) - V_h^*(x) &\geq \mathcal{T}_{k,h} V_{i,j+1}(x) - \mathcal{T} V_{h+1}^*(x) \\ &\overset{(I)}{\geq} b_{k,h}(x, \pi^*(x,h)) + \widehat{P}_h^{\pi^*} V_{i,j+1}(x) - P_h^{\pi^*} V_{h+1}^*(x) \\ &= b_{k,h}(x, \pi^*(x,h)) + (\widehat{P}_h^{\pi^*} - P_h^{\pi^*}) V_{h+1}^*(x) + \widehat{P}_h^{\pi^*}(V_{i,j+1} - V_{h+1}^*)(x) \\ &\overset{(II)}{\geq} b_{k,h}(x, \pi^*(x,h)) + (\widehat{P}_h^{\pi^*} - P_h^{\pi^*}) V_{h+1}^*(x), \end{aligned}$$

where in (I) we rely on the fact that $\pi_{k,h}$ is the greedy policy w.r.t. $V_{k,h}$. Thus

$$b_{k,h}(x, \pi^*(x,h)) + \widehat{P}_h^{\pi^*} V_{i,j+1}(x) \quad \leq \quad b_{k,h}(x, \pi_k(x,h)) + \widehat{P}_h^{\pi_k} V_{i,j+1}(x).$$

Also (II) follows from the induction assumption. Under the event $\mathcal{E}$ we have

$$\begin{aligned} V_{k,h} - V_h^* &\geq b_{k,h} - c_1(\widehat{\mathbb{V}}_h^*, N_k) \\ &\geq \underbrace{\sqrt{\frac{8\widehat{\mathbb{V}}_{k,h} L}{N_k}} - 2\sqrt{\frac{\widehat{\mathbb{V}}_h^* L}{N_k}}}_{(a)} - \frac{14L}{3N_k} \\ &\quad + \sqrt{\frac{\widehat{P}_k\left[8\min\left(\frac{100^2 H^3 S^2 A^2 L^2}{N'_{k,h+1}}, H^2\right)\right]}{N_k}} + \frac{14L}{3N_k}. \end{aligned}$$

We now prove a lower bound on (a):

$$(a) \geq \begin{cases} -\sqrt{\frac{4\widehat{\mathbb{V}}_{k,h}^* - 8\widehat{\mathbb{V}}_{k,h}}{N_k}} & \widehat{\mathbb{V}}_{k,h} \leq \mathbb{V}^*, \\ 0 & \text{otherwise.} \end{cases}$$

We proceed by bounding $\widehat{\mathbb{V}}_{k,h}^*$ in terms of $\widehat{\mathbb{V}}_{k,h}$ from above:

$$\widehat{\mathbb{V}}_{k,h}^* \overset{(I)}{\leq} 2\widehat{\mathbb{V}}_{k,h} + 2\mathrm{Var}_{y\sim\widehat{P}_k}(V_{k,h+1}(y) - V_{h+1}^*(y)) \leq 2\widehat{\mathbb{V}}_{k,h} + 2\underbrace{\widehat{P}_k(V_{k,h+1} - V_{h+1}^*)^2}_{(b)},$$

where (I) is an application of Lem. 2. We now bound (b). Combining this result with the result of Eq. 64 leads to the following bound on (a)

$$(a) \geq \begin{cases} -\sqrt{\frac{8\widehat{P}_k\left[\min\left(\frac{100^2 H^3 S^2 A L^2}{N'_{k,h+1}}, H^2\right)\right]}{N_k}} & \widehat{\mathbb{V}}_{k,h} \leq \widehat{\mathbb{V}}^*, \\ 0 & \text{otherwise,} \end{cases}$$

where the last inequality holds under the event $\mathcal{E}$. The proof is completed by plugging (a) and (b) into Eq. 64 which proves that $V_{k,h} \geq V_h^*$ thus the event $\Omega_{k,h} holds$.

□