A. Algorithm

Algorithm 3 HOEFFDING-TYPE CONFIDENCE BOUNDS

```
input: state space $S$, action space $A$ and confidence parameter $\delta$.
initialization: arbitrary policy $\tilde{\pi}$, $m \leftarrow 1, \widetilde{B} \leftarrow c_{\min}, C_1 \leftarrow 0, \forall (s,a,s') \in S \times A \times S: \quad N(s,a,s') \leftarrow 0, N(s,a) \leftarrow 0$.
for $k = 1, 2, \dots$ do
  set $s \leftarrow s_{\text{init}}$.
  while $s \neq g$ do
    follow optimistic optimal policy: $a \leftarrow \tilde{\pi}(s)$.
    suffer cost: $C_m \leftarrow C_m + c(s,a)$.
    observe next state $s' \sim P(\cdot \mid s,a)$.
    update visit counters: $N(s,a,s') \leftarrow N(s,a,s') + 1, N(s,a) \leftarrow N(s,a) + 1$.
    if $N(s', \tilde{\pi}(s')) \leq \frac{5000\widetilde{B}^2|S|}{c_{\min}^2} \log \frac{\widetilde{B}|S||A|}{\delta c_{\min}}$ or $s' = g$ or $C_m \geq 24\widetilde{B} \log \frac{4m}{\delta}$ then
      # start new interval
      if $C_m \geq 24\widetilde{B} \log \frac{4m}{\delta}$ then
        update $B_\star$ estimate: $\widetilde{B} \leftarrow 2\widetilde{B}$.
      end if
      advance intervals counter: $m \leftarrow m + 1$.
      initialize cost suffered in interval: $C_m \leftarrow 0$.
      compute empirical transition function $\bar{P}$ for every $(s,a,s') \in S \times A \times S$:
          $\bar{P}(s' \mid s,a) = \frac{N(s,a,s')}{\max\{N(s,a),1\}}$.
      compute policy $\tilde{\pi}$ that minimizes the expected cost with respect to a transition function $\widetilde{P}$, such that for every
      $(s,a) \in S \times A$:
          $\|\widetilde{P}(\cdot \mid s,a) - \bar{P}(\cdot \mid s,a)\|_1 \leq 5\sqrt{\frac{|S| \log\left(|S||A|N_+(s,a)/\delta\right)}{N_+(s,a)}}$.
    end if
    set $s \leftarrow s'$.
  end while
end for
```

B. Proofs

B.1. Proofs for Section 3.1

B.1.1. PROOF OF LEMMA 3.3

Lemma (restatement of Lemma 3.3)**.** *With probability at least* $1 - \delta/2$, Ω^m *holds and* $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \leq 24 B_\star \log \frac{4m}{\delta}$ *for all intervals* m *simultaneously. This implies that the total number of steps of the algorithm is*

$$T = O\left(\frac{KB_\star}{c_{min}} \log \frac{KB_\star |S||A|}{\delta c_{min}} + \frac{B_\star^3 |S|^2 |A|}{c_{min}^3} \log^2 \frac{KB_\star |S||A|}{\delta c_{min}} \right).$$

Lemma B.1. *The event* Ω^m *holds for all intervals* m *simultaneously with probability at least* $1 - \delta/4$.

Proof. Fix a state s and an action a. Consider an infinite sequence $\{Z_i\}_{i=1}^{\infty}$ of draws from the distribution $P(\cdot \mid s,a)$. By Theorem D.2 we get that for a prefix of length t of this sequence (that is $\{Z_i\}_{i=1}^{t}$)

$$\|P(\cdot \mid s,a) - \bar{P}_{\{Z_i\}_{i=1}^t}(\cdot \mid s,a)\|_1 \leq 2\sqrt{\frac{|S| \log(\delta_t^{-1})}{t}} \,,$$

holds with probability $1 - \delta_t$, where $\bar{P}_{\{Z_i\}_{i=1}^t}(\cdot \mid s, a)$ is the empirical distribution defined by the draws $\{Z_i\}_{i=1}^t$. We repeat this argument for every prefix $\{Z_i\}_{i=1}^t$ of $\{Z_i\}_{i=1}^\infty$ and for every state-action pair, with $\delta_t = \delta/8|S||A|t^2$. Then from the union bound we get that Ω^m holds for all intervals m simultaneously with probability at least $1 - \delta/4$. □

Lemma B.2. *Let m be an interval. If Ω^m holds then $\widetilde{J}^m(s) \leq J^{\pi^\star}(s) \leq B_\star$ for every $s \in S$.*

Proof. Tarbouriech et al. (2020) show that all the transition functions in the confidence set of Eq. (4) can be combined into a single augmented MDP. The optimal policy of the augmented MDP can be found efficiently, e.g., with Extended Value Iteration. The optimistic policy is the optimal policy in the augmented MDP. It minimizes $\widetilde{J}^m(s)$ over all policies and feasible transition functions, for all states $s \in S$ simultaneously (following Bertsekas & Tsitsiklis, 1991). Since Ω^m holds, it follows that the real transition function is in the confidence set therefore it is also considered in the minimization. Thus $\widetilde{J}^m(s) \leq J^{\pi^\star}(s)$ for all $s \in S$. Finally, $J^{\pi^\star}(s) \leq B_\star$ by the definition of $B_\star$. □

Lemma B.3. *Let m be an interval and (s, a) be a known state-action pair. If Ω^m holds then*

$$\|\widetilde{P}_m(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1 \leq \frac{c(s, a)}{2B_\star} .$$

Proof. By the definition of the confidence set

$$\|\widetilde{P}_m(\cdot \mid s, a) - \bar{P}_m(\cdot \mid s, a)\|_1 \leq 5\sqrt{\frac{|S| \log\bigl(|S||A|N_+^m(s, a)/\delta\bigr)}{N_+^m(s, a)}} \leq \frac{c(s, a)}{4B_\star},$$

where the last inequality follows because $\log(x)/x$ is decreasing, and $N_+^m(s, a) \geq \frac{5000 B_\star^2 |S|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}$ since (s, a) is known. Similarly, since Ω^m holds we also have that

$$\|P(\cdot \mid s, a) - \bar{P}_m(\cdot \mid s, a)\|_1 \leq 5\sqrt{\frac{|S| \log\bigl(|S||A|N_+^m(s, a)/\delta\bigr)}{N_+^m(s, a)}} \leq \frac{c(s, a)}{4B_\star},$$

and the lemma follows by the triangle inequality. □

Lemma B.4. *Let $\tilde{\pi}$ be a policy and $\widetilde{P}$ be a transition function. Denote the cost-to-go of $\tilde{\pi}$ with respect to $\widetilde{P}$ by $\widetilde{J}$. Assume that for every $s \in S$, $\widetilde{J}(s) \leq B_\star$ and that*

$$\|\widetilde{P}(\cdot \mid s, \tilde{\pi}(s)) - P(\cdot \mid s, \tilde{\pi}(s))\|_1 \leq \frac{c(s, \tilde{\pi}(s))}{2B_\star}.$$

Then, $\tilde{\pi}$ is proper (with respect to P), and it holds that $J^{\tilde{\pi}}(s) \leq 2B_\star$ for every $s \in S$.

Proof. Consider the Bellman equations of $\tilde{\pi}$ with respect to transition function $\widetilde{P}$ at some state $s \in S$ (see Lemma 2.1), defined as

$$\begin{aligned}\widetilde{J}(s) &= c(s, \tilde{\pi}(s)) + \sum_{s' \in S} \widetilde{P}(s' \mid s, \tilde{\pi}(s)) \widetilde{J}(s') \\ &= c(s, \tilde{\pi}(s)) + \sum_{s' \in S} P(s' \mid s, \tilde{\pi}(s)) \widetilde{J}(s') + \sum_{s' \in S} \widetilde{J}(s') \Bigl(\widetilde{P}(s' \mid s, \tilde{\pi}(s)) - P(s' \mid s, \tilde{\pi}(s))\Bigr) . \qquad (8)\end{aligned}$$

Notice that by our assumptions and using Hölder inequality,

$$\begin{aligned}\left|\sum_{s' \in S} \widetilde{J}(s') \Bigl(\widetilde{P}(s' \mid s, \tilde{\pi}(s)) - P(s' \mid s, \tilde{\pi}(s))\Bigr)\right| &\leq \|\widetilde{P}(\cdot \mid s, \tilde{\pi}(s)) - P(\cdot \mid s, \tilde{\pi}(s))\|_1 \cdot \|\widetilde{J}\|_\infty \\ &\leq \frac{c(s, \tilde{\pi}(s))}{2B_\star} \cdot B_\star = \frac{c(s, \tilde{\pi}(s))}{2} .\end{aligned}$$

Plugging this into Eq. (8), we obtain

$$\widetilde{J}(s) \geq c(s,\tilde{\pi}(s)) + \sum_{s' \in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{J}(s') - \frac{c(s,\tilde{\pi}(s))}{2} = \frac{c(s,\tilde{\pi}(s))}{2} + \sum_{s' \in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{J}(s').$$

Therefore, defining $J' = 2\widetilde{J}$, then $J'(s) \geq c(s,\tilde{\pi}(s)) + \sum_{s' \in S} P(s' \mid s,\tilde{\pi}(s))J'(s')$ for all $s \in S$. The statement now follows by Lemma 2.1. □

Lemma B.5. *Let π be a proper policy such that for some $v > 0$, $J^{\pi}(s) \leq v$ for every $s \in S$. Then, the probability that the cost of π to reach the goal state from any state s is more than m, is at most $2e^{-m/4v}$ for all $m \geq 0$. Note that a cost of at most m implies that the number of steps is at most $\frac{m}{c_{min}}$.*

Proof. By Markov inequality, the probability that π accumulates cost of more than $2v$ before reaching the goal state is at most 1/2. Iterating this argument, we get that the probability that π accumulates cost of more than $2kv$ before reaching the goal state is at most 2^{-k} for every integer $k \geq 0$. In general, for any $m \geq 0$, the probability that π suffers a cost of more than m is at most $2^{-\lfloor m/2v \rfloor} \leq 2 \cdot 2^{-m/2v} \leq 2e^{-m/4v}$. □

For the next lemma we will need the following definitions. The trajectory visited in interval m is denoted by $U^m = (s_1^m, a_1^m, \ldots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$ where a_h^m is the action taken in s_h^m, and H^m is the length of the interval. In addition, the concatenation of the trajectories in the intervals up to and including interval m is denoted by $\bar{U}^m = \cup_{m'=1}^{m} U^{m'}$.

Lemma B.6. *Let m be an interval. For all $r \geq 0$, we have that*

$$\mathbb{P}\left[\sum_{h=1}^{H^m} c\left(s_h^m, a_h^m\right)\mathbb{I}\{\Omega^m\} > r \mid \bar{U}^{m-1}\right] \leq 3e^{-r/8B_\star}.$$

Proof. Note that Ω^m is determined given $\bar{U}^{m-1}$, and suppose that Ω^m holds otherwise $\sum_{h=1}^{H^m} c\left(s_h^m, a_h^m\right)\mathbb{I}\{\Omega^m\}$ is 0. Also assume that $r \geq 8B_\star$ or else the statement holds trivially.

Define the MDP $M^{\text{know}} = (S^{\text{know}}, A, P^{\text{know}}, c, s_{\text{init}})$ in which every state $s \in S$ such that $(s, \tilde{\pi}^m(s))$ is unknown is contracted into the goal state. Let P^{know} be the transition function induced in M^{know} by P, and let J^m_{know} be the cost-to-go of $\tilde{\pi}^m$ in M^{know} with respect to P^{know}. Similarly, define $\widetilde{P}^{\text{know}}_m$ as the transition function induced in M^{know} by $\widetilde{P}_m$, and $\widetilde{J}^m_{\text{know}}$ as the cost-to-go of $\tilde{\pi}^m$ in M^{know} with respect to $\widetilde{P}^{\text{know}}_m$. It is clear that $\widetilde{J}^m_{\text{know}}(s) \leq \widetilde{J}^m(s)$ for every $s \in S$, so by Lemma B.2, $\widetilde{J}^m_{\text{know}}(s) \leq B_\star$. Moreover, since all the states $s \in S$ for which $(s, \tilde{\pi}^m(s))$ is unknown were contracted to the goal state, we can use Lemma B.3 to obtain for all $s \in S^{\text{know}}$:

$$\|\widetilde{P}^{\text{know}}_m(\cdot \mid s, \tilde{\pi}^m(s)) - P^{\text{know}}(\cdot \mid s, \tilde{\pi}^m(s))\|_1 \leq \|\widetilde{P}_m(\cdot \mid s, \tilde{\pi}^m(s)) - P(\cdot \mid s, \tilde{\pi}^m(s))\|_1 \leq \frac{c(s, \tilde{\pi}^m(s))}{2B_\star}. \tag{9}$$

We can apply Lemma B.4 in M^{know} and obtain that $J^m_{\text{know}}(s) \leq 2B_\star$ for every $s \in S^{\text{know}}$. Notice that reaching the goal state in M^{know} is equivalent to reaching the goal state or an unknown state-action pair in M, and also recall that all state-action pairs in the interval are known except for the first one. Thus, from Lemma B.5,

$$\mathbb{P}\left[\sum_{h=2}^{H^m} c\left(s_h^m, a_h^m\right)\mathbb{I}\{\Omega^m\} > r - B_\star \mid \bar{U}^{m-1}\right] \leq 2e^{-(r-B_\star)/8B_\star} \leq 3e^{-r/8B_\star}.$$

Since $\widetilde{J}^m \leq B_\star$, our algorithm will never select an action whose instantaneous cost is larger than $B_\star$. Since the first state-action in the interval might not be known, its cost is at most $B_\star$, and therefore

$$\mathbb{P}\left[\sum_{h=1}^{H^m} c\left(s_h^m, a_h^m\right)\mathbb{I}\{\Omega^m\} > r \mid \bar{U}^{m-1}\right] \leq \mathbb{P}\left[\sum_{h=2}^{H^m} c\left(s_h^m, a_h^m\right)\mathbb{I}\{\Omega^m\} > r - B_\star \mid \bar{U}^{m-1}\right] \leq 3e^{-r/8B_\star}. \quad □$$

Proof of Lemma 3.3. From Lemma B.6, with probability at least $1-\delta/16m^2$, $\sum_{h=1}^{H^m} c\left(s_h^m, a_h^m\right) \leq 24 B_\star \log \frac{4m}{\delta}$, and by the union bound this holds for all intervals m simultaneously with probability at least $1-\delta/4$. By Lemma B.1, with probability $1-\delta/4$, Ω^m holds for all intervals m. Combining these two facts again by a union bound, we get that both Ω^m holds and the cost of interval m is at most $24 B_\star \log \frac{4m}{\delta}$ simultaneously to all intervals m with probability at least $1-\delta/2$.

If the cost of all intervals is bounded (and therefore so is the length of the interval), we can use the bound on the number of intervals in Observation 3.2 to conclude that

$$\begin{aligned} T &= O\left(\frac{B_\star}{c_{\min}} \log \frac{M}{\delta} \cdot \left(K + \frac{B_\star^2 |S|^2 |A|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}\right)\right) \\ &= O\left(\frac{K B_\star}{c_{\min}} \log \frac{K B_\star |S||A|}{\delta c_{\min}} + \frac{B_\star^3 |S|^2 |A|}{c_{\min}^3} \log^2 \frac{K B_\star |S||A|}{\delta c_{\min}}\right). \end{aligned}$$ □

B.1.2. Proof of Lemma 3.4

Lemma (restatement of Lemma 3.4)**.** *With probability at least* $1-\delta/2$*, we have*

$$\widetilde{R}_K \leq \frac{5000 B_\star^3 |S|^2 |A|}{c_{min}^2} \log \frac{B_\star |S||A|}{c_{min}\delta} + B_\star \sqrt{T \log \frac{4T}{\delta}} + 10 B_\star \sqrt{|S| \log \frac{|S||A|T}{\delta}} \sum_{s,a} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.$$

To analyze $\widetilde{R}_K$, we begin by plugging in the Bellman optimality equation of $\tilde{\pi}^m$ with respect to $\widetilde{P}_m$ into $\widetilde{R}_K$. This allows us to decompose $\widetilde{R}_K$ into three terms as follows.

$$\begin{aligned} \widetilde{R}_K &= \sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{J}^m(s_h^m) - \sum_{s' \in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m) \widetilde{J}^m(s')\right) \mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{\text{init}}) \\ &= \sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m)\right) \mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{\text{init}}) && (10) \\ &\quad + \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{s' \in S} \widetilde{J}^m(s') \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right) \mathbb{I}\{\Omega^m\} && (11) \\ &\quad + \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} P(s' \mid s_h^m, a_h^m) \widetilde{J}^m(s')\right) \mathbb{I}\{\Omega^m\}. && (12) \end{aligned}$$

Eq. (10) is a bound on the cost suffered from switching policies each time we visit an unknown state-action pair and is bounded by the following lemma.

Lemma B.7. $\sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m)\right) \mathbb{I}\{\Omega^m\} \leq B_\star |S||A| \cdot \frac{5000 B_\star^2 |S|}{c_{min}^2} \log \frac{B_\star |S||A|}{\delta c_{min}} + K \cdot J^{\pi^\star}(s_{init})$.

Proof. Note that per interval $\sum_{h=1}^{H^m} (\widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m))$ is a telescopic sum which equals $\widetilde{J}^m(s_1^m) - \widetilde{J}^m(s_{H^m+1}^m)$. Furthermore, for every two consecutive intervals $m, m+1$ one of the following occurs:

(i) If interval m ended in the goal state then $\widetilde{J}^m(s_{H^m+1}^m) = \widetilde{J}^m(g) = 0$ and $\widetilde{J}^{m+1}(s_1^{m+1}) = \widetilde{J}^{m+1}(s_{\text{init}})$. Thus, using Lemma B.2 for the last inequality,

$$\widetilde{J}^{m+1}(s_1^{m+1}) \mathbb{I}\{\Omega^{m+1}\} - \widetilde{J}^m(s_{H^m+1}^m) \mathbb{I}\{\Omega^m\} = \widetilde{J}^{m+1}(s_{\text{init}}) \mathbb{I}\{\Omega^{m+1}\} \leq J^{\pi^\star}(s_{\text{init}}).$$

This happens at most K times.

(ii) If interval m ended in an unknown state then

$$\widetilde{J}^{m+1}(s_1^{m+1}) \mathbb{I}\{\Omega^{m+1}\} - \widetilde{J}^m(s_{H^m+1}^m) \mathbb{I}\{\Omega^m\} \leq \widetilde{J}^{m+1}(s_1^{m+1}) \mathbb{I}\{\Omega^{m+1}\} \leq B_\star.$$

This happens at most $|S||A| \cdot \frac{5000 B_\star^2 |S|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}$ times. □

Lemma B.8 bounds Eq. (11) using techniques borrowed from Jaksch et al. (2010).

Lemma B.8. *It holds that*

$$\sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in S}\widetilde{J}^m(s')\left(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\right)\mathbb{I}\{\Omega^m\}\le 10B_\star\sqrt{|S|\log\frac{|S||A|T}{\delta}}\sum_{s,a}\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.$$

Proof. Using the definition of the confidence sets we obtain

$$\begin{aligned}\sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in S}\widetilde{J}^m(s')\left(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\right)\mathbb{I}\{\Omega^m\}\le&\\ \le B_\star\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}n_m(s,a)\|P(\cdot\mid s,a)-\widetilde{P}_m(\cdot\mid s,a)\|_1\mathbb{I}\{\Omega^m\}&\\ \le 10B_\star\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}n_m(s,a)\sqrt{\frac{|S|\log\left(|S||A|N_+^m(s,a)/\delta\right)}{N_+^m(s,a)}}&\\ \le 10B_\star\sqrt{|S|\log\frac{|S||A|T}{\delta}}\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.&\end{aligned}$$

where the first inequality follows from Hölder inequality and Lemma B.2, and the second because $\widetilde{P}_m$ and P are both in the confidence set of Eq. (4) when Ω^m holds. The third inequality follows because $N_+^m(s,a)\le T$. □

Lemma B.9 bounds the term in Eq. (12) using Azuma's concentration inequality.

Lemma B.9. *With probability at least* $1-\delta/2$,

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}P(s'\mid s_h^m,a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\}\le B_\star\sqrt{T\log\frac{4T}{\delta}}.$$

Proof. Consider the infinite sequence of random variables

$$X_t=\left(\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}P(s'\mid s_h^m,\tilde{\pi}^m(s_h^m))\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\},$$

where m is the interval containing time t, and h is the index of time step t within interval m. Notice that this is a martingale difference sequence, and $|X_t|\le B_\star$ by Lemma B.2. Now, we apply anytime Azuma's inequality (Theorem D.1) to any prefix of the sequence $\{X_t\}_{t=1}^{\infty}$. Thus, with probability at least $1-\delta/2$, for every T:

$$\sum_{t=1}^{T}X_t\le B_\star\sqrt{T\log\frac{4T}{\delta}}.$$

□

B.1.3. Proof of Theorem 3.1

Theorem (restatement of Theorem 3.1)**.** *Suppose that Assumption 2 holds. With probability at least* $1-\delta$ *the regret of Algorithm 1 is bounded as follows:*

$$R_K=O\left(\sqrt{\frac{B_\star^3|S|^2|A|K}{c_{min}}}\log\frac{KB_\star|S||A|}{\delta c_{min}}+\frac{B_\star^3|S|^2|A|}{c_{min}^2}\log^{3/2}\frac{KB_\star|S||A|}{\delta c_{min}}\right).$$

Lemma B.10. *Assume that the number of steps in every interval is is at most* $\frac{24B_\star}{c_{min}}\log\frac{4m}{\delta}$. *Then for every* $s\in S$ *and* $a\in A$,

$$\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}\le 3\sqrt{N_{M+1}(s,a)}\,.$$

Proof. We claim that, by the assumption of the lemma, for every interval m we have that $n_m(s,a) \le N_+^m(s,a)$. Indeed, if (s,a) is unknown then $n_m(s,a) = 1$ and since $N_+^m(s,a) \ge 1$ the claim follows. If (s,a) is known then $N_+^m(s,a) \ge \frac{5000 B_\star^2 |S|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}$ and by our assumption the length of the interval, and in particular $n_m(s,a)$, is at most $\frac{24 B_\star}{c_{\min}} \log \frac{4m}{\delta}$. Our statement then follows by Jaksch et al. (2010, Lemma 19). □

Proof of Theorem 3.1. With probability at least $1-\delta$, both Lemmas 3.3 and B.9 hold. Lemma 3.3 states that the length of every interval is at most $\frac{24 B_\star}{c_{\min}} \log \frac{4m}{\delta}$, and Lemma B.10 obtains

$$\sum_{s \in S} \sum_{a \in A} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}} \le 3 \sum_{(s,a) \in S \times A} \sqrt{N_{M+1}(s,a)} \le 3\sqrt{|S||A|T}, \tag{13}$$

where the last inequality follows from Jensen's inequality and the fact that $\sum_{(s,a)\in S\times A} N_{M+1}(s,a) \le T$. Next, we sum the bounds of Lemmas B.7 to B.9 and use Eq. (13) to obtain

$$R_K \le 5000 \frac{B_\star^3 |S|^2 |A|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}} + 30 B_\star |S| \sqrt{|A| T \log \frac{|S||A|T}{\delta}} + B_\star \sqrt{T \log \frac{4T}{\delta}}.$$

To finish the proof use Lemma 3.3 to bound T. □

B.2. Proofs for Section 4.1

B.2.1. Proof of Lemma 4.2

Lemma (restatement of Lemma 4.2). *With probability at least $1-\delta/2$, Ω^m holds for all intervals m simultaneously.*

Proof. Fix a triplet $(s,a,s') \in S \times A \times S^+$. Consider an infinite sequence $(Z_i)_{i=1}^\infty$ of draws from the distribution $P(\cdot \mid s,a)$ and let $X_i = \mathbb{I}\{Z_i = s'\}$. We apply Eq. (25) of Theorem D.3 with $\delta_t = \frac{\delta}{4|S|^2|A|t^2}$ to a prefix of length t of the sequence $(X_i)_{i=1}^\infty$. Then divide Eq. (25) by t and obtain that, after simplifying using the assumptions that $|S| \ge 2$ and $|A| \ge 2$, Eq. (6) holds with probability $1-\delta_t$. We repeat this argument for every prefix $(Z_i)_{i=1}^t$ of $(Z_i)_{i=1}^\infty$ and for every state-action-state triplet. Then from the union bound we get that Ω^m holds for all intervals m simultaneously with probability at least $1-\delta/2$. □

B.2.2. Proof of Lemma 4.3

Lemma (restatement of Lemma 4.3). *It holds that*

$$\begin{aligned}\widetilde{R}_M = &\sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m) \right) \mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{init}) \\ &+ \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m) \widetilde{J}^m(s') \right) \mathbb{I}\{\Omega^m\}.\end{aligned}$$

Lemma B.11. *Let m be an interval. If Ω^m holds then $\tilde{\pi}^m$ satisfies the Bellman equations in the optimistic model:*

$$\widetilde{J}^m(s) = c(s, \tilde{\pi}^m(s)) + \sum_{s' \in S} \widetilde{P}_m(s' \mid s, \tilde{\pi}^i(s)) \widetilde{J}^m(s'), \quad \forall s \in S.$$

Proof. Note that the Bellman equations hold in the optimistic model since as we defined this model, there is a nonzero probability of transition to the goal state by any action from every state. Thus in the optimistic model every policy is a proper policy and in particular Lemma 2.2 holds. □

Proof of Lemma 4.3. By Lemma B.11, we can use the Bellman equations in the optimistic model to have the following interpretation of the costs for every interval m and time h:

$$\begin{aligned} c(s_h^m, a_h^m)\mathbb{I}\{\Omega^m\} &= \Big(\widetilde{J}^m(s_h^m) - \sum_{s' \in S} \widetilde{P}_i(s' \mid s_h^m, a_h^m) \widetilde{J}^m(s') \Big) \mathbb{I}\{\Omega^m\} \\ &= \Big(\widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m) \Big) \mathbb{I}\{\Omega^m\} + \Big(\widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} \widetilde{P}_i(s' \mid s_h^m, a_h^m) \widetilde{J}^m(s') \Big) \mathbb{I}\{\Omega^m\}. \end{aligned} \tag{14}$$

We now write $\widetilde{R}_M = \sum_{m=1}^M \sum_{h=1}^{H^m} c(s_h^m, a_h^m)\mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{\text{init}})$, and substitute for each cost using Eq. (14) to get the lemma. □

B.2.3. PROOF OF LEMMA 4.4

Lemma (restatement of Lemma 4.4)**.** $\sum_{m=1}^M \big(\sum_{h=1}^{H^m} \widetilde{J}^m(s_h^m) - \widetilde{J}^m(s_{h+1}^m) \big) \mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{init}) \le 2B_\star |S||A| \log T$.

Lemma B.12. *Let m be an interval. If Ω^m holds then $\widetilde{J}^m(s) \le J^{\pi^\star}(s) \le B_\star$ for every $s \in S$.*

Proof. Denote by $\widetilde{P}$ the transition function computed by Algorithm 2 at the beginning of epoch $i(m)$, and by $\widetilde{J}$ the cost-to-go with respect to $\widetilde{P}$. We claim that for every proper policy π and state $s \in S$, $\widetilde{J}^\pi(s) \le J^\pi(s)$. Then, the lemma follows easily since $\widetilde{J}^m(s) \le \widetilde{J}^{\pi^\star}(s) \le J^{\pi^\star}(s) \le B_\star$.

Indeed, let $s \in S$ and consider the Bellman equations of π with respect to P:

$$J^\pi(s) = c(s, \pi(s)) + \sum_{s' \in S} P(s' \mid s, \pi(s)) J^\pi(s') \ge c(s, \pi(s)) + \sum_{s' \in S} \widetilde{P}(s' \mid s, \pi(s)) J^\pi(s'),$$

where the inequality follows because $\widetilde{P}(s' \mid s, a) \le P(s' \mid s, a)$ for every $(s, a, s') \in S \times A \times S$. This holds since P is in the confidence set of Eq. (6) (as Ω^m holds), and by the way $\widetilde{P}$ is computed in *Algorithm* 2. Therefore, by Lemma 2.1 we obtain that $J^\pi(s) \ge \widetilde{J}^\pi(s)$ for every $s \in S$ as required. □

Proof of Lemma 4.4. For every two consecutive intervals $m, m+1$, denoting $i = i(m)$, we have one of the following:

(i) If interval m ended in the goal state then $\widetilde{J}^{i(m)}(s_{H^m+1}^m) = \widetilde{J}^{i(m)}(g) = 0$ and $\widetilde{J}^{i(m+1)}(s_1^{m+1}) = \widetilde{J}^{i(m+1)}(s_{\text{init}})$. Therefore, by Lemma B.12,

$$\widetilde{J}^{i(m+1)}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{J}^{i(m)}(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} = \widetilde{J}^{i(m+1)}(s_{\text{init}})\mathbb{I}\{\Omega^{m+1}\} \le J^{\pi^\star}(s_{\text{init}}).$$

This happens at most K times.

(ii) If interval m ended in an unknown state-action pair or since the cost reached $B_\star$, and we stay in the same epoch, then $i(m) = i(m+1) = i$ and $s_1^{m+1} = s_{H^m+1}^m$. Thus

$$\widetilde{J}^{i(m+1)}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{J}^{i(m)}(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} = \widetilde{J}^{i}(s_1^{m+1})\mathbb{I}\{\Omega^m\} - \widetilde{J}^{i}(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} = 0.$$

(iii) If interval m ended by doubling the visit count to some state-action pair, then we start a new epoch. Thus by Lemma B.12,

$$\widetilde{J}^{i(m+1)}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{J}^{i(m)}(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} \le \widetilde{J}^{i+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} \le B_\star,$$

This happens at most $2|S||A| \log T$ times.

To conclude, we have

$$\begin{aligned} \sum_{m=1}^M \Big(\sum_{h=1}^{H^m} \widetilde{J}^{i(m)}(s_h^m) - \widetilde{J}^{i(m)}(s_{h+1}^m) \Big) \mathbb{I}\{\Omega^m\} - KJ^{\pi^\star}(s_{\text{init}}) &\le KJ^{\pi^\star}(s_{\text{init}}) + 2B_\star |S||A| \log T - KJ^{\pi^\star}(s_{\text{init}}) \\ &= 2B_\star |S||A| \log T. \end{aligned}$$

□

B.2.4. Proof of Lemma 4.5

Lemma (restatement of Lemma 4.5)**.** *With probability at least* $1-\delta/4$*, the following holds for all* $M = 1, 2, \ldots$ *simultaneously.*

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}\widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\}$$
$$\leq \sum_{m=1}^{M}\mathbb{E}\left[\left(\sum_{h=1}^{H^m}\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}\widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] + 3B_\star\sqrt{M\log\frac{8M}{\delta}}.$$

Proof. Consider the following martingale difference sequence $(X^m)_{m=1}^{\infty}$ defined by

$$X^m = \sum_{h=1}^{H^m}\Big(\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}\widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\Big)\mathbb{I}\{\Omega^m\}.$$

The Bellman optimality equations of $\tilde{\pi}^m$ with respect to $\widetilde{P}_m$ (Lemma B.11) obtain

$$|X^m| = \left|\left(\underbrace{\widetilde{J}^m(s_{H^m+1}^m)-\widetilde{J}^m(s_1^m)}_{\leq B_\star}+\underbrace{\sum_{h=1}^{H^m}c(s_h^m, a_h^m)}_{\leq 2B_\star}\right)\mathbb{I}\{\Omega^m\}\right| \leq 3B_\star,$$

where the inequality follows from Lemma B.12 and the fact that the total cost within each interval at most $2B_\star$ by construction. Therefore, we use anytime Azuma's inequality (Theorem D.1) to obtain that with probability at least $1-\delta/4$:

$$\sum_{m=1}^{M}X^m \leq \sum_{m=1}^{M}\mathbb{E}\big[X^m \mid \bar{U}^{m-1}\big] + 3B_\star\sqrt{M\log\frac{8M}{\delta}}. \quad \square$$

B.2.5. Proof of Lemma 4.6

Lemma (restatement of Lemma 4.6)**.** *For every interval m and time h, denote* $A_h^m = \frac{\log(|S||A|N_+^m(s_h^m,a_h^m)/\delta)}{N_+^m(s_h^m,a_h^m)}$*. Then,*

$$\mathbb{E}\left[\left(\sum_{h=1}^{H^m}\widetilde{J}^m(s_{h+1}^m)-\sum_{s'\in S}\widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\leq 16\cdot\mathbb{E}\left[\sum_{h=1}^{H^m}\sqrt{|S|\mathbb{V}_h^m A_h^m}\,\mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right] + 272\cdot\mathbb{E}\left[\sum_{h=1}^{H^m}B_\star|S|A_h^m\mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right],$$

where $\mathbb{V}_h^m$ *is the empirical variance defined as*

$$\mathbb{V}_h^m = \sum_{s'\in S^+}P(s' \mid s_h^m, a_h^m)\left(\widetilde{J}^m(s')-\sum_{s''\in S^+}P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\right)^2.$$

The next lemma gives a different interpretation to the confidence bounds of Eq. (6), and will be useful in the proofs that follow.

Lemma B.13. *Denote* $A_h^m = \log(|S||A|N_+^m(s,a)/\delta)/N_+^m(s,a)$*. When* Ω^m *holds we have for any* $(s, a, s') \in S \times A \times S^+$*:*

$$|P(s' \mid s, a) - \widetilde{P}_m(s' \mid s, a)| \leq 8\sqrt{P(s' \mid s, a)A_h^m} + 136A_h^m.$$

Proof. Since Ω^m holds we have for all $(s, a, s') \in S \times A \times S^+$ that

$$\bar{P}_m(s' \mid s, a) - P(s' \mid s, a) \leq 4\sqrt{\bar{P}_m(s' \mid s, a)A_h^m} + 28A_h^m.$$

This is a quadratic inequality in $\sqrt{\bar{P}_m(s' \mid s,a)}$. Using the fact that $x^2 \leq a \cdot x + b$ implies $x \leq a + \sqrt{b}$ with $a = 4\sqrt{A_h^m}$ and $b = P(s' \mid s,a) + 28A_h^m$, we have

$$\sqrt{\bar{P}_m(s' \mid s,a)} \leq 4\sqrt{A_h^m} + \sqrt{P(s' \mid s,a) + 28A_h^m} \leq \sqrt{P(s' \mid s,a)} + 10\sqrt{A_h^m},$$

where we used the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ that holds for any $x \geq 0$ and $y \geq 0$. Substituting back into Eq. (6) obtains

$$|P(s' \mid s,a) - \bar{P}_m(s' \mid s,a)| \leq 4\sqrt{P(s' \mid s,a)A_h^m} + 68A_h^m.$$

From a similar argument

$$|\widetilde{P}_m(s' \mid s,a) - \bar{P}_m(s' \mid s,a)| \leq 4\sqrt{P(s' \mid s,a)A_h^m} + 68A_h^m.$$

Using the triangle inequality finishes the proof. □

Proof of Lemma 4.6. Denote $X^m = \big(\sum_{h=1}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\big)\mathbb{I}\{\Omega^m\}$, and $Z_h^m = \big(\widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} P(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\big)\mathbb{I}\{\Omega^m\}$. Think of the interval as an infinite stochastic process, and note that, conditioned on $\bar{U}^{m-1}$, $\big(Z_h^m\big)_{h=1}^{\infty}$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^{\infty}$, where U^h is the trajectory of the learner from the beginning of the interval and up to and including time h. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval. Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^{\infty}$ which is bounded by $2B_\star/c_{\min}$. Hence by the optional stopping theorem $\mathbb{E}[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1}] = 0$, which gets us

$$\begin{aligned}
\mathbb{E}[X^m \mid \bar{U}^{m-1}] &= \mathbb{E}\left[\sum_{h=1}^{H^m}\left(\widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1}\right] + \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in S}\big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in S}\big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right].
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in S}\big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad = \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in S^+}\left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\left(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad \leq \mathbb{E}\left[8\sum_{h=1}^{H^m}\sum_{s' \in S^+}\sqrt{A_h^m P(s' \mid s_h^m, a_h^m)\left(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\right)^2}\,\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\qquad + \mathbb{E}\left[136\sum_{h=1}^{H^m}\sum_{s' \in S^+} A_h^m\left|\widetilde{J}^m(s') - \sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\right|\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad \leq \mathbb{E}\left[16\sum_{h=1}^{H^m}\sqrt{|S|\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} + 272|S|B_\star A_h^m\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right],
\end{aligned}$$

where the first equality follows since $\widetilde{J}^m(g) = 0$, and $P(\cdot \mid s_h^m, a_h^m)$ and $\widetilde{P}_i(\cdot \mid s_h^m, a_h^m)$ are probability distributions over S^+ whence $\sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')$ does not depend on s'. The first inequality follows from Lemma B.13, and the second inequality from Jensen's inequality, Lemma B.12, $|S^+| \leq 2|S|$, and the definition of $\mathbb{V}_h^m$. □

B.2.6. PROOF OF LEMMA 4.7

Lemma (restatement of Lemma 4.7)**.** *For any interval m, $\mathbb{E}\big[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \leq 44 B_\star^2$.*

Lemma B.14. *Let m be an interval and (s,a) be a known state-action pair. If Ω^m holds then for every $s' \in S^+$*

$$|\widetilde{P}_m\big(s' \mid s,a\big) - P\big(s' \mid s,a\big)| \leq \frac{1}{8}\sqrt{\frac{c_{min} \cdot P\big(s' \mid s,a\big)}{|S| B_\star}} + \frac{c_{min}}{4|S|B_\star}.$$

Proof. By Lemma B.13 we have that

$$|\widetilde{P}_m\big(s' \mid s,a\big) - P\big(s' \mid s,a\big)| \leq 8\sqrt{\frac{P(s' \mid s,a)\log\big(|S||A|N_+^m(s,a)/\delta\big)}{N_+^m(s,a)}} + \frac{136\log\big(|S||A|N_+^m(s,a)/\delta\big)}{N_+^m(s,a)}$$

which gives the required bound because $\log(x)/x$ is decreasing, and (s,a) is a known state-action pair so $N_+^m(s,a) \geq 30000 \cdot \frac{B_\star |S|}{c_{\min}} \log \frac{B_\star |S||A|}{\delta c_{\min}}$. □

Proof of Lemma 4.7. Note that the first state-action pair in the subinterval, (s_1^m, a_1^m), might be unknown and that all state-action pairs that appear afterwards are known. Thus, we bound

$$\mathbb{E}\Bigg[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mid \bar{U}^{m-1}\Bigg] = \mathbb{E}\Bigg[\mathbb{V}_1^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\Bigg] + \mathbb{E}\Bigg[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\Bigg].$$

The first summand is trivially bounded by $B_\star^2$ (Lemma B.12). We now upper bound $\mathbb{E}\big[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big]$. Denote $Z_h^m = \big(\widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} P(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\big)\mathbb{I}\{\Omega^m\}$, and think of the interval as an infinite stochastic process. Note that, conditioned on $\bar{U}^{m-1}$, $\big(Z_h^m\big)_{h=1}^\infty$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^\infty$, where U^h is the trajectory of the learner from the beginning of the interval and up to time h and including. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval. Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^\infty$ which is bounded by $2B_\star/c_{\min}$. Therefore, applying Lemma B.15 found below obtains

$$\mathbb{E}\Bigg[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\Bigg] = \mathbb{E}\Bigg[\Bigg(\sum_{h=2}^{H^m} Z_h^m \mathbb{I}\{\Omega^m\}\Bigg)^2 \mid \bar{U}^{m-1}\Bigg]. \tag{15}$$

We now proceed by bounding $|\sum_{h=1}^{H^m} Z_h^m|$ when Ω^m occurs. Therefore,

$$\begin{align}
\left|\sum_{h=2}^{H^m} Z_h^m\right| &= \left|\sum_{h=2}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \sum_{s' \in S} P(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right| \notag \\
&\leq \left|\sum_{h=2}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \widetilde{J}^m(s_h^m)\right| \tag{16} \\
&\quad + \left|\sum_{h=2}^{H^m} \widetilde{J}^m(s_h^m) - \sum_{s' \in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right| \tag{17} \\
&\quad + \left|\sum_{h=2}^{H^m} \sum_{s' \in S^+} \Big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\Big)\Big(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\Big)\right|, \tag{18}
\end{align}$$

where Eq. (18) is given as $P(\cdot \mid s_h^m, a_h^m)$ and $\widetilde{P}_i(\cdot \mid s_h^m, a_h^m)$ are probability distributions over S^+, $\sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')$ is constant w.r.t s', and $\widetilde{J}^m(g) = 0$.

We now bound each of the three terms above individually. Eq. (16) is a telescopic sum that is at most $B_\star$ on Ω^m (Lemma B.12). For Eq. (17), we use the Bellman equations for $\tilde{\pi}^m$ on the optimistic model defined by the transitions $\widetilde{P}_m$ (Lemma B.11) thus it is at most $\sum_{h=2}^{H^m} c\big(s_h^m, a_h^m\big) \le 2B_\star$ (see text following Lemma 4.5). For Eq. (18), recall that all states-action pairs at times $h = 2, \dots, H^m$ are known by definition of H^m. Hence by Lemma B.14,

$$
\begin{aligned}
&\left| \sum_{s' \in S^+} \Big(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P\big(s'' \mid s_h^m, a_h^m\big) \widetilde{J}^m(s'') \Big) \Big(\widetilde{P}_m\big(s' \mid s_h^m, a_h^m\big) - P\big(s' \mid s_h^m, a_h^m\big) \Big) \right| \\
&\quad \le \frac{1}{8} \sum_{s' \in S^+} \sqrt{\frac{c_{\min} \cdot P\big(s' \mid s_h^m, a_h^m\big) \Big(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P\big(s'' \mid s_h^m, a_h^m\big)\widetilde{J}^m(s'')\Big)^2}{|S| B_\star}} \\
&\qquad + \sum_{s' \in S^+} \frac{c_{\min}}{4|S|B_\star} \cdot \underbrace{\left| \widetilde{J}^m(s') - \sum_{s'' \in S^+} P\big(s'' \mid s_h^m, a_h^m\big) \widetilde{J}^m(s'') \right|}_{\le B_\star \text{ by Lemma B.12}} \\
&\quad \le \frac{1}{4} \sqrt{\frac{c_{\min} \cdot \mathbb{V}_h^m}{B_\star}} + \frac{c\big(s_h^m, a_h^m\big)}{2}, \qquad \text{(by Jensen's inequality, } c_{\min} \le c(s_h^m, a_h^m), |S^+| \le 2|S|)
\end{aligned}
$$

and again by Jensen's inequality and that the total cost throughout the interval is at most $2B_\star$, we have on Ω^m

$$
\begin{aligned}
\sum_{h=2}^{H^m} \frac{1}{4} \sqrt{\frac{c_{\min} \cdot \mathbb{V}_h^m}{B_\star}} + \frac{c\big(s_h^m, a_h^m\big)}{2} &\le \frac{1}{4} \sqrt{\underbrace{H^m}_{\le 2B_\star / c_{\min}} \cdot \sum_{h=2}^{H^m} \frac{c_{\min} \cdot \mathbb{V}_h^m}{B_\star}} + \frac{1}{2} \underbrace{\sum_{h=2}^{H^m} c\big(s_h^m, a_h^m\big)}_{\le 2B_\star} \qquad \text{(Jensen's inequality)} \\
&\le \frac{1}{4} \sqrt{2 \sum_{h=2}^{H^m} \mathbb{V}_h^m} + B_\star.
\end{aligned}
$$

Plugging these bounds back into Eq. (15) gets us

$$
\begin{aligned}
\mathbb{E}\left[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1} \right] &\le \mathbb{E}\left[\left(4B_\star + \frac{1}{4} \sqrt{2 \sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\}} \right)^2 \;\middle|\; \bar{U}^{m-1} \right] \\
&\le 32 B_\star^2 + \frac{1}{4} \mathbb{E}\left[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1} \right],
\end{aligned}
$$

where the last inequality is by the elementary inequality $(a+b)^2 \le 2(a^2+b^2)$. Rearranging gets us $\mathbb{E}\big[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \le 43 B_\star^2$, and the lemma follows. □

Lemma B.15. *Let $(X_t)_{t=1}^\infty$ be a martingale difference sequence adapted to the filtration $(\mathcal{F}_t)_{t=0}^\infty$. Let $Y_n = (\sum_{t=1}^n X_t)^2 - \sum_{t=1}^n \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}]$. Then $(Y_n)_{n=0}^\infty$ is a martingale, and in particular if τ is a stopping time such that $\tau \le c$ almost surely, then $\mathbb{E}[Y_\tau] = 0$.*

Proof. We first show that $(Y_n)_{n=1}^\infty$ is a martingale. Indeed,

$$
\begin{aligned}
\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] &= \mathbb{E}\left[\left(\sum_{t=1}^n X_t \right)^2 - \sum_{t=1}^n \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{n-1} \right] \\
&= \mathbb{E}\left[\left(\sum_{t=1}^{n-1} X_t \right)^2 - 2 \left(\sum_{t=1}^{n-1} X_t \right) X_n + X_n^2 - \sum_{t=1}^n \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{n-1} \right] \\
&= \left(\sum_{t=1}^{n-1} X_t \right)^2 - 2 \left(\sum_{t=1}^{n-1} X_t \right) \cdot 0 + \mathbb{E}[X_n^2 \mid \mathcal{F}_{n-1}] - \sum_{t=1}^n \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] \qquad (\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = 0) \\
&= \left(\sum_{t=1}^{n-1} X_t \right)^2 - \sum_{t=1}^{n-1} \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] = Y_{n-1}.
\end{aligned}
$$

We would now like to show that $\mathbb{E}[Y_\tau] = \mathbb{E}[Y_1] = 0$ using the optional stopping theorem. The latter holds since $\tau \leq c$ almost surely and $\mathbb{E}[Y_1] = \mathbb{E}[X_1^2 - \mathbb{E}[X_1^2 \mid \mathcal{F}_0]] = 0$. □

B.2.7. Proof of Lemma 4.8

Lemma (restatement of Lemma 4.8)**.** *With probability at least* $1 - \delta/4$*,*

$$\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in S} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\leq 614 B_\star \sqrt{M|S|^2|A| \log^2 \frac{T|S||A|}{\delta}} + 8160 B_\star |S|^2|A| \log^2 \frac{T|S||A|}{\delta}.$$

Proof. Recall the following definitions:

$$A_h^m = \frac{\log(|S||A|N_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)}. \qquad \mathbb{V}_h^m = \sum_{s' \in S^+} P(s' \mid s_h^m, a_h^m)\left(\widetilde{J}^m(s') - \sum_{s'' \in S^+} P(s'' \mid s_h^m, a_h^m)\widetilde{J}^m(s'')\right)^2.$$

From Lemma 4.6 we have that

$$\mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in S} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\leq \mathbb{E}\left[16\sqrt{|S|} \sum_{h=1}^{H^m} \sqrt{\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} + 272 B_\star |S| A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right].$$

Moreover, by applying the Cauchy-Schwartz inequality twice, we get that

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=1}^{H^m} \sqrt{\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right] &\leq \mathbb{E}\left[\sqrt{\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\}} \cdot \sqrt{\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\}} \;\middle|\; \bar{U}^{m-1}\right] \\
&\leq \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right]} \\
&\leq 7 B_\star \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right]}. \qquad \text{(Lemma 4.7)}
\end{aligned}$$

We sum over all intervals to obtain

$$\begin{aligned}
&\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in S} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{J}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \leq \\
&\leq 112 B_\star \sum_{m=1}^{M} \sqrt{|S| \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]} + 272 B_\star |S| \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\leq 112 B_\star \sqrt{M|S| \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]} + 272 B_\star |S| \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right],
\end{aligned}$$

where the last inequality follows from Jensen's inequality. We finish the proof using Lemma B.16 below. □

Lemma B.16. *With probability at least* $1 - \delta/4$*, the following holds for* $M = 1, 2, \ldots$ *simultaneously.*

$$\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \leq O\left(|S||A| \log^2 \frac{T|S||A|}{\delta}\right).$$

Proof. Define the infinite sequence of random variables: $X^m = \sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\}$ for which $|X^m| \leq 3\log(|S||A|/\delta)$ due to Lemma B.17 below. We apply Eq. (26) of Lemma D.4 to obtain with probability at least $1 - \delta/4$, for all $M = 1, 2, \ldots$ simultaneously

$$\sum_{m=1}^{M} \mathbb{E}\left[X^m \mid \bar{U}^{m-1}\right] \leq 2\sum_{m=1}^{M} X^m + 12\log\left(\frac{|S||A|}{\delta}\right)\log\left(\frac{8M}{\delta}\right).$$

Now, we bound the sum over X^m by rewriting it as a sum over epochs:

$$\sum_{m=1}^{M} X^m \leq \sum_{m=1}^{M}\sum_{h=1}^{H^m} \frac{\log(|S||A|N_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} \leq \log\frac{|S||A|T}{\delta}\sum_{s\in S}\sum_{a\in A}\sum_{i=1}^{E}\frac{n_i(s,a)}{N_+^i(s,a)},$$

where E is the last epoch. Finally, from Lemma B.18 below we have that for every $(s, a) \in S \times A$,

$$\sum_{i=1}^{E}\frac{n_i(s,a)}{N_+^i(s,a)} \leq 2\log N_{E+1}(s,a) \leq 2\log T.$$

We now plugin the resulting bound for $\sum_{m=1}^{M} X^m$ and simplify the acquired expression by using $M \leq T$. □

Lemma B.17. *For any interval m, $|\sum_{h=1}^{H^m} A_h^m| \leq 3\log(|S||A|/\delta)$.*

Proof. Note that all state-action pairs (s_h^m, a_h^m) (except the first one (s_1^m, a_1^m)) are known. Hence, for $h \geq 2$, $N_+^m(s_h^m, a_h^m) \geq 30000 \cdot \frac{B_\star|S|}{c_{\min}} \log\frac{B_\star|S||A|}{\delta c_{\min}}$. Therefore, since $\log(x)/x$ is decreasing and since $|S| \geq 2$ and $|A| \geq 2$ by assumption,

$$\begin{aligned}
\sum_{h=1}^{H^m}\frac{\log(|S||A|N_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} &\leq \frac{\log(|S||A|N_+^m(s_1^m, a_1^m)/\delta)}{N_+^m(s_1^m, a_1^m)} + \sum_{h=2}^{H^m}\frac{\log(|S||A|N_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} \\
&\leq \log(|S||A|/\delta) + \frac{c_{\min}H^m}{B_\star} \\
&\leq \log(|S||A|/\delta) + 2 && (H^m \leq \tfrac{2B_\star}{c_{\min}} \text{ by definition.}) \\
&\leq 3\log(|S||A|/\delta).
\end{aligned}$$

□

Lemma B.18. *For any sequence of integers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ and $Z_0 = 1$, it holds that*

$$\sum_{k=1}^{n}\frac{z_k}{Z_{k-1}} \leq 2\log Z_n.$$

Proof. We use the inequality $x \leq 2\log(1 + x)$ for every $0 \leq x \leq 1$ to obtain

$$\sum_{k=1}^{n}\frac{z_k}{Z_{k-1}} \leq 2\sum_{k=1}^{n}\log\left(1 + \frac{z_k}{Z_{k-1}}\right) = 2\sum_{k=1}^{n}\log\frac{Z_{k-1} + z_k}{Z_{k-1}} = 2\sum_{k=1}^{n}\log\frac{Z_k}{Z_{k-1}} = 2\log\prod_{k=1}^{n}\frac{Z_k}{Z_{k-1}} = 2\log Z_n.$$

□

B.2.8. Proof of Theorem 2.4

Theorem (restatement of Theorem 2.4)**.** *Assume that Assumption 2 holds. With probability at least $1 - \delta$ the regret of Algorithm 2 is bounded as follows:*

$$R_K = O\left(B_\star|S|\sqrt{|A|K}\log\frac{KB_\star|S||A|}{\delta c_{min}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{min}}}\log^2\frac{KB_\star|S||A|}{\delta c_{min}}\right).$$

Proof. Let C_M denote the cost of the learner after M intervals. First, with probability at least $1-\delta$, we have Lemmas 4.2, 4.5 and 4.8 via a union bound. Now, as Ω^m holds for all intervals, we have $\widetilde{R}_M = R_M$ for any number of intervals M. Plugging in the bounds of Lemmas 4.4, 4.5 and 4.8 into Lemma 4.3, we have that for any number of intervals M:

$$C_M = O\left(K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{M|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + B_\star|S|^2|A|\log^2\frac{T|S||A|}{\delta}\right).$$

We now plug in the bounds on M and T from Observation 4.1 into the bound above. First, we plug in the bound on M. As long as the K episodes have not elapsed we have that $M \leq O\big(C_M/B_\star + K + 2|S||A|\log T + \frac{B_\star|S|^2|A|}{c_{\min}}\log\frac{B_\star|S||A|}{\delta c_{\min}}\big)$. This gets after using the subadditivity of the square root to simplify the resulting expression,

$$\begin{aligned}C_M = O\Bigg(&K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}}\\ &+ \sqrt{B_\star C_M|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}\Bigg).\end{aligned}$$

From which, by solving for C_M (using that $x \leq a\sqrt{x}+b$ implies $x \leq (a+\sqrt{b})^2$ for $a \geq 0$ and $b \geq 0$), and simplifying the resulting expression by applying $J^{\pi^\star}(s_{\text{init}}) \leq B_\star$ and our assumptions that $K \geq |S|^2|A|$, $|S| \geq 2$, $|A| \geq 2$, we get that

$$\begin{aligned}C_M &= O\Bigg(\Bigg(\sqrt{B_\star|S|^2|A|\log^2\frac{T|S||A|}{\delta}}\\ &\qquad + \sqrt{K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}}\Bigg)^2\Bigg)\\ &= O\Bigg(B_\star|S|^2|A|\log^2\frac{T|S||A|}{\delta}\\ &\qquad + \sqrt{B_\star|S|^2|A|\log^2\frac{T|S||A|}{\delta}} \cdot \sqrt{K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}}\\ &\qquad + K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}\Bigg)\\ &= O\Bigg(B_\star|S|^2|A|\log^2\frac{T|S||A|}{\delta} + B_\star\sqrt{K^{1/4}|S|^3|A|^{3/2}\log^3\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^{5/2}|S|^4|A|^2}{c_{\min}^{1/2}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}\\ &\qquad + K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}\Bigg)\\ &= O\Bigg(K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{K|S|^2|A|\log^2\frac{T|S||A|}{\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{TB_\star|S||A|}{c_{\min}\delta}}\Bigg). \qquad (19)\end{aligned}$$

Note that in particular, by simplifying the bound above, we have $C_M = O\Big(\sqrt{B_\star^3|S|^4|A|^2KT/c_{\min}\delta}\Big)$. Next we combine this with the fact, stated in Observation 4.1 that $T \leq C_M/c_{\min}$. Isolating T gets $T = O\Big(\frac{B_\star^3|S|^4|A|^2K}{c_{\min}^3\delta}\Big)$, and plugging this bound back into Eq. (19) and simplifying gets us

$$C_M = O\left(K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star|S|\sqrt{|A|K\log^2\frac{KB_\star|S||A|}{c_{\min}\delta}} + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{\min}}\log^4\frac{KB_\star|S||A|}{c_{\min}\delta}}\right).$$

Finally, we note that the bound above holds for any number of intervals M as long as K episodes do not elapse. As the instantaneous costs in the model are positive, this means that the learner must eventually finish the K episodes from which we derive the bound for R_K claimed by the theroem. □

C. Lower Bound

In this section we prove Theorem 2.7. At first glance, it is tempting to try and use the lower bound of Jaksch et al. (2010, Theorem 5) on the regret suffered against learning average-reward MDPs by reducing any problem instance from an average-reward MDP to an instance of SSP. However, it is unclear to us if such a reduction is possible, and if it is, how to perform it.[2] We consequently prove the theorem here directly.

By Yao's minimax principle, in order to derive a lower bound on the learner's regret, it suffices to show a distribution over MDP instances that forces any deterministic learner to suffer a regret of $\Omega(B_\star\sqrt{|S||A|K})$ in expectation.

To simplify our arguments, let us first consider the following simpler problem before considering the problem in its full generality. Think of a simple MDP with two states: the initial state and a goal state. The set of actions A has a special action $a^\star$ chosen uniformly at random a-priori. Upon choosing the special action, the learner transitions to the goal state with probability $\approx 1/B_\star$ and remains at s_{init} with the remaining probability. Concretely $P(g \mid a^\star) = 1/B_\star$ and $P(s_{\text{init}} \mid a^\star) = 1 - 1/B_\star$, and for any other action $a \neq a^\star$ we have $P(g \mid a) = (1-\epsilon)/B_\star$ and $P(s_{\text{init}} \mid a) = 1 - (1-\epsilon)/B_\star$ for some $\epsilon \in (0, 1/8)$.[3] The costs of all actions equal 1; i.e., $c(s_{\text{init}}, a) = 1$ for all $a \in A$. Clearly, the optimal policy constantly plays $a^\star$ and therefore $J^{\pi^\star}(s_{\text{init}}) = B_\star$.

Fix any deterministic learning algorithm, we shall now quantify the regret of the learner during a single episode in terms of the number of times that it chooses $a^\star$. Let N_k denote the number of steps that the learner spends in s_{init} during episode k, and let $N_k^\star$ be the number of times the learner plays $a^\star$ at s_{init} during the episode. Note that N_k is also the total cost that the learning algorithm suffered during episode k. We have the following lemma.

Lemma C.1. $\mathbb{E}\left[N_k\right] - J^{\pi^\star}(s_{init}) = \epsilon \cdot \mathbb{E}\left[N_k - N_k^\star\right]$.

Proof. Let us denote by $s_1, s_2, \ldots$ and $a_1, a_2, \ldots$ the sequences of states and actions observed by the learner during the episode. We have,

$$\begin{aligned}
\mathbb{E}[N_k] &= \sum_{t=1}^{\infty} \mathbb{P}[s_t = s_{\text{init}}] \\
&= 1 + \sum_{t=2}^{\infty} \mathbb{P}[s_t = s_{\text{init}}] \\
&= 1 + \sum_{t=2}^{\infty} \mathbb{P}[s_t = s_{\text{init}} \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star]\mathbb{P}[s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] \\
&\quad + \sum_{t=2}^{\infty} \mathbb{P}[s_t = s_{\text{init}} \mid s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star]\mathbb{P}[s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star] \\
&= 1 + \sum_{t=2}^{\infty}\left(1 - \frac{1}{B_\star}\right)\mathbb{P}[s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] + \sum_{t=2}^{\infty}\left(1 - \frac{1-\epsilon}{B_\star}\right)\mathbb{P}[s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star] \\
&= 1 + \left(1 - \frac{1}{B_\star}\right)\sum_{t=1}^{\infty}\mathbb{P}[s_t = s_{\text{init}}, a_t = a^\star] + \left(1 - \frac{1-\epsilon}{B_\star}\right)\sum_{t=1}^{\infty}\mathbb{P}[s_t = s_{\text{init}}, a_t \neq a^\star] \\
&= 1 + \left(1 - \frac{1}{B_\star}\right)\mathbb{E}[N_k^\star] + \left(1 - \frac{1-\epsilon}{B_\star}\right)\mathbb{E}[N_k - N_k^\star].
\end{aligned}$$

Rearranging using $J^{\pi^\star}(s_{\text{init}}) = B_\star$ gives the Lemma's statement. □

[2] Even though a reduction in the reverse direction is fairly straight-forward in the unit-cost case (Tarbouriech et al., 2020).

[3] For ease of notation and since there is only one state other than g, we do not write this state as the origin state in the definition of the transition function.

By Lemma C.1 the overall regret of the learner over K episodes is: $\mathbb{E}[R_K] = \epsilon \cdot \mathbb{E}\big[N - N^\star\big]$, where $N = \sum_{k=1}^K N_k$ and $N^\star = \sum_{k=1}^K N_k^\star$. In words, the regret of the learner is ϵ times the expected number of visits to s_{init} in which the learner did not play $a^\star$.

In the remainder of the proof we lower bound N in expectation and upper bound the expected value of $N^\star$. To upper bound $N^\star$, we use standard techniques from lower bounds of multi-armed bandits (Auer et al., 2002) that bound the total variation distance between the distribution of the sequence of states traversed by the learner in the original MDP and that generated in a "uniform MDP" in which all actions are identical. However, we cannot apply this argument directly since it requires $N^\star$ to be bounded almost surely, yet here $N^\star$ depends on the total length of all K episodes which is unbounded in general. We fix this issue by looking only on the first T steps (where T is to be determined) and showing that the regret is large even in these T steps.

Formally, we view the run of the K episodes as a continuous process in which when the learner reaches the goal state we transfer it to s_{init} (at no cost) and let it restart from there. Furthermore, we *cap* the learning process to consist of exactly T steps as follows. If the K episodes are completed before T steps are elapsed, the learner remains in g (until completing T steps) without suffering any additional cost, and otherwise we stop the learner after T steps before it completes its K episodes. In this capped process, we denote the number of visits in s_{init} by N_- and the number of times the learner played $a^\star$ in s_{init} by $N_-^\star$. We have

$$\mathbb{E}[R_K] \geq \epsilon \cdot \big(\mathbb{E}\big[N_-\big] - \mathbb{E}\big[N_-^\star\big]\big). \tag{20}$$

The number of visits to s_{init} under this capping is lower bounded by the following lemma.

Lemma C.2. *For any deterministic learner, if $T \geq 2KB_\star$ then we have that $\mathbb{E}\big[N_-\big] \geq KB_\star/4$.*

Proof. If the capped learner finished its K episodes then $N_- = N$. Otherwise, it visits the goal state less than K times and therefore $N_- \geq T - K$. Hence $\mathbb{E}\big[N_-\big] \geq \mathbb{E}\big[\min\{T - K, N\}\big] \geq \sum_{k=1}^K \mathbb{E}\big[\min\{T/K - 1, N_k\}\big]$. Since $T \geq 2KB_\star$, the lemma will follow if we show that $N_k \geq B_\star$ with probability at least 1/4. We lower bound the probability that $N_k \geq B_\star$ by the probability of staying at s_{init} for $B_\star$ steps and picking $a^\star$ in the first $B_\star - 1$ steps. Indeed, using $(1 - 1/x)^{x-1} \geq 1/e$ for $x \geq 1$, we get that $\mathbb{P}[N_k \geq B_\star] \geq \big(1 - \frac{1}{B_\star}\big)^{B_\star - 1} \geq \frac{1}{4}$. □

We now introduce an additional distribution of the transitions which call $\mathbb{P}_{\text{unif}}$. $\mathbb{P}_{\text{unif}}$ is identical to $\mathbb{P}$ as defined above, except that $P(g \mid a) = (1 - \epsilon)/B_\star$ for all actions a. We denote expectations over $\mathbb{P}_{\text{unif}}$ by $\mathbb{E}_{\text{unif}}$. The following lemma uses standard lower bound techniques used for multi-armed bandits (see, e.g., Jaksch et al., 2010, Theorem 13) to bound the difference in the expectation of $N_-^\star$ when the learner plays in $\mathbb{P}$ compared to when it plays in $\mathbb{P}_{\text{unif}}$.

Lemma C.3. *For any deterministic learner we have that $\mathbb{E}\big[N_-^\star\big] \leq \mathbb{E}_{unif}\big[N_-^\star\big] + \epsilon T \sqrt{\mathbb{E}_{unif}[N_-^\star]/B}$.*

Proof. Fix any deterministic learner. Let us denote by $s^{(t)}$ the sequence of states observed by the learner up to time t and including. Now, as $N_-^\star \leq T$ and the fact that $N_-^\star$ is a function of $s^{(T)}$, $\mathbb{E}\big[N_-^\star\big] \leq \mathbb{E}_{\text{unif}}\big[N_-^\star\big] + T \cdot \text{TV}(\mathbb{P}_{\text{unif}}[s^{(T)}], \mathbb{P}[s^{(T)}])$, and Pinsker's inequality yields

$$\text{TV}(\mathbb{P}_{\text{unif}}[s^{(T)}], \mathbb{P}[s^{(T)}]) \leq \sqrt{\frac{1}{2}\text{KL}(\mathbb{P}_{\text{unif}}[s^{(T)}] \parallel \mathbb{P}[s^{(T)}])}. \tag{21}$$

Next, the chain rule of the KL divergence obtains

$$\text{KL}(\mathbb{P}_{\text{unif}}[s^{(T)}] \parallel \mathbb{P}[s^{(T)}]) = \sum_{t=1}^T \sum_{s^{(t-1)}} \mathbb{P}_{\text{unif}}[s^{(t-1)}] \cdot \text{KL}(\mathbb{P}_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \mathbb{P}[s_t \mid s^{(t-1)}]).$$

Observe that at any time, since the learning algorithm is deterministic, the learner chooses an action given $s^{(t-1)}$ regardless of whether $s^{(t-1)}$ was generated under $\mathbb{P}$ or under $\mathbb{P}_{\text{unif}}$. Thus, the $\text{KL}(\mathbb{P}_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \mathbb{P}[s_t \mid s^{(t-1)}])$ is zero if $a_{t-1} \neq a_\star$, and otherwise

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \mathbb{P}[s_t \mid s^{(t-1)}]) &= \sum_{s \in S} \mathbb{P}_{\text{unif}}[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] \log \frac{\mathbb{P}_{\text{unif}}[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star]}{\mathbb{P}[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star]} \\
&= \frac{1-\epsilon}{B_\star} \cdot \log(1-\epsilon) + \left(1 - \frac{1-\epsilon}{B_\star}\right) \log\left(1 + \frac{\epsilon}{B_\star - 1}\right) \\
&\leq \frac{\epsilon^2}{B_\star - 1}. && \text{(using } \log(1+x) \leq x \text{ for all } x > 0)
\end{aligned}$$

Plugging the above back into Eq. (21) and using $B_\star \geq 2$ gives the lemma. □

In the following result, we combine the lemma above with standard techniques from lower bounds of multi-armed bandits (see Jaksch et al., 2010, Thm. 5 for example).

Theorem C.4. *Suppose that $B_\star \geq 2$, $\epsilon \in (0, \frac{1}{8})$ and $|A| \geq 16$. For the problem described above we have that*

$$\mathbb{E}[R_K] \geq \epsilon K B_\star \left(\frac{1}{8} - 2\epsilon\sqrt{\frac{2K}{|A|}}\right).$$

Proof of Theorem C.4. Note that as under $\mathbb{P}_{\text{unif}}$ the transition distributions are identical for all actions, we have that

$$\sum_{a \in A | a^\star = a} \mathbb{E}_{\text{unif}}\left[N_-^\star\right] = \mathbb{E}_{\text{unif}}\left[\sum_{a \in A | a^\star = a} N_-^\star\right] = \mathbb{E}_{\text{unif}}\left[N_-\right] \leq T. \tag{22}$$

Suppose that $a^\star$ is sampled uniformly at random before the game starts. Denote the probability and expectation with respect to the distribution induced by a specific choice of $a^\star = a$ by $\mathbb{P}_a$ and $\mathbb{E}_a$ respectively. Then for $T = 2KB_\star$,

$$\begin{aligned}
\mathbb{E}[R_K] &= \frac{1}{|A|} \sum_{a \in A} \mathbb{E}_a[R_K] \\
&\geq \frac{1}{|A|} \sum_{a \in A} \mathbb{E}_a[N_- - N_-^\star] && \text{(Eq. (20))} \\
&\geq \frac{1}{|A|} \sum_{a \in A | a_\star = a} \left(\frac{KB_\star}{4} - \mathbb{E}_{\text{unif}}[N_-^\star] - \epsilon T \sqrt{\frac{\mathbb{E}_{\text{unif}}[N_-^\star]}{B_\star}}\right) && \text{(Lemmas C.2 and C.3)} \\
&\geq \frac{KB_\star}{4} - \frac{1}{|A|} \sum_{a \in A | a_\star = a} \mathbb{E}_{\text{unif}}[N_-^\star] - \epsilon T \sqrt{\frac{1}{B_\star} \cdot \frac{1}{|A|} \sum_{a \in A | a_\star = a} \mathbb{E}_{\text{unif}}[N_-^\star]} && \text{(Jensen's inequality)} \\
&\geq \frac{KB_\star}{4} - \frac{T}{|A|} - \epsilon T \sqrt{\frac{T}{B_\star |A|}} && \text{(Eq. (22))} \\
&= \epsilon\left(\frac{KB_\star}{4} - \frac{2KB_\star}{|A|} - 2\epsilon K B_\star \sqrt{\frac{2KB_\star}{|A| B_\star}}\right) \\
&= \epsilon K B_\star \left(\frac{1}{4} - \frac{2}{|A|} - 2\epsilon\sqrt{\frac{2K}{|A|}}\right).
\end{aligned}$$

The theorem follows from $|A| \geq 16$ and by rearranging. □

Proof of Theorem 2.7. Consider the following MDP. Let S be the set of states disregarding g. The initial state is sampled uniformly at random from S. Each $s \in S$ has its own special action $a_s^\star$. The transition distributions are defined $P(g \mid a_s^\star, s) = 1/B_\star$, $P(s \mid a_s^\star, s) = 1 - 1/B_\star$, and $P(g \mid a, s) = (1-\epsilon)/B_\star$, $P(s \mid a, s) = 1 - (1-\epsilon)/B_\star$ for any other action $a \in A \backslash \{a_s^\star\}$.

Note that for each $s \in S$, the learner is faced with a simple problem as the one described above from which it cannot learn about from other states $s' \neq s$. Therefore, we can apply Theorem C.4 for each $s \in S$ separately and lower bound the learner's expected regret the sum of the regrets suffered at each $s \in S$, which would depend on the number of times $s \in S$ is drawn as the initial state. Since the states are chosen uniformly at random there are many states (constant fraction) that are chosen $\Theta(K/|S|)$ times. Summing the regret bounds of Theorem C.4 over only these states and choosing ϵ appropriately gives the sought-after bound.

Denote by K_s the number of episodes that start in each state $s \in S$.

$$\mathbb{E}[R_K] \geq \sum_{s\in S} \mathbb{E}\left[\epsilon K_s B_\star \left(\frac{1}{8} - 2\epsilon\sqrt{\frac{2K_s}{|A|}}\right)\right] = \frac{\epsilon K B_\star}{8} - 2\epsilon^2 B_\star \sqrt{\frac{2}{|A|}} \sum_{s\in S} \mathbb{E}[K_s^{3/2}]. \tag{23}$$

Taking expectation over the initial states and applying Cauchy-Schwartz inequality gives

$$\sum_{s\in S} \mathbb{E}\left[K_s^{3/2}\right] \leq \sum_{s\in S} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s^2]} = \sum_{s\in S} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s]^2 + \mathbb{V}[K_s]} = \sum_{s\in S} \sqrt{\frac{K}{|S|}}\sqrt{\frac{K^2}{|S|^2} + \frac{K(|S|-1)}{|S|^2}} \leq K\sqrt{\frac{2K}{|S|}},$$

where we have used the expectation and variance formulas of the Binomial distribution. The lower bound is now given by applying the inequality above in Eq. (23) and choosing $\epsilon = \frac{1}{64}\sqrt{|A||S|/K}$. □

D. Concentration inequalities

Theorem D.1 (Anytime Azuma). *Let $(X_n)_{n=1}^\infty$ be a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)_{n=0}^\infty$ such that $|X_n| \leq B$ almost surely. Then with probability at least $1-\delta$,*

$$\left|\sum_{i=1}^n X_i\right| \leq B\sqrt{n \log \frac{2n}{\delta}}, \qquad \forall n \geq 1.$$

Theorem D.2 (Weissman et al., 2003). *Let $p(\cdot)$ be a distribution over m elements, and let $\bar{p}_t(\cdot)$ be the empirical distribution defined by t iid samples from $p(\cdot)$. Then, with probability at least $1-\delta$,*

$$\|\bar{p}_t(\cdot) - p(\cdot)\|_1 \leq 2\sqrt{\frac{m \log \frac{1}{\delta}}{t}}.$$

Theorem D.3 (Anytime Bernstein). *Let $(X_n)_{n=1}^\infty$ be a sequence of i.i.d. random variables with expectation μ. Suppose that $0 \leq X_n \leq B$ almost surely. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\mu n \log \frac{2n}{\delta}} + B \log \frac{2n}{\delta}. \tag{24}$$

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\sum_{i=1}^n X_i \log \frac{2n}{\delta}} + 7B \log \frac{2n}{\delta}. \tag{25}$$

Proof. Fix some $n \geq 1$. By Bernstein's concentration inequality (see for example, Cesa-Bianchi & Lugosi, 2006, Corollary A.3), we have with probability at least $1 - \frac{\delta}{2n^2}$ that Eq. (24) holds. By a union bound, the inequality holds with probability at least $1-\delta$ for all $n \geq 1$ simultaneously.

To show Eq. (25), note that in particular we have

$$\mu \cdot n - \sum_{i=1}^n X_i \leq 2\sqrt{B\mu n \log \frac{2n}{\delta}} + B \log \frac{2n}{\delta}$$

that is a quadratic inequality in μ. This implies that

$$\sqrt{\mu} \leq \sqrt{\frac{1}{n}\sum_{i=1}^n X_i} + 3\sqrt{\frac{B \log \frac{2n}{\delta}}{n}}.$$

Plugging this inequality back into the RHS of Eq. (24) gets us Eq. (25). □

Lemma D.4. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables with expectation adapted to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$. Suppose that $0 \le X_n \le B$ almost surely. Then with probability at least $1-\delta$, the following holds for all $n \ge 1$ simultaneously:*

$$\sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \le 2\sum_{i=1}^{n} X_i + 4B \log \frac{2n}{\delta}. \tag{26}$$

Proof. For all $n \ge 1$, we have

$$\begin{aligned}
\mathbb{E}[e^{-X_n/B} \mid \mathcal{F}_{n-1}] &\le \mathbb{E}\left[1 - \frac{X_n}{B} + \frac{X_n^2}{2B^2} \;\middle|\; \mathcal{F}_{n-1}\right] && (e^{-x} \le 1 - x + \tfrac{x^2}{2} \text{ for all } x \ge 0)\\
&\le 1 - \frac{\mathbb{E}[X_n \mid \mathcal{F}_{n-1}]}{B} + \frac{\mathbb{E}[X_n \mid \mathcal{F}_{n-1}]}{2B} && (X_n \le B)\\
&= 1 - \frac{\mathbb{E}[X_n \mid \mathcal{F}_{n-1}]}{2B}\\
&\le e^{-\mathbb{E}[X_n \mid \mathcal{F}_{n-1}]/2B}. && (1 - x \le e^{-x} \text{ for all } x)
\end{aligned}$$

Hence, fix some $n \ge 1$, then

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)\right)\right]\\
&= \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n-1}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)\right) \cdot \underbrace{\mathbb{E}\left[\exp\left(\frac{1}{B}\left(\frac{1}{2}\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] - X_n\right)\right) \;\middle|\; \mathcal{F}_{n-1}\right]}_{\le 1}\right]\\
&\le \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n-1}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)\right)\right]\\
&\le 1. && \text{(by repeating the last argument inductively.)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right) > 2B\log\frac{2n}{\delta}\right] &\le \mathbb{P}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)\right) > \frac{2n^2}{\delta}\right]\\
&\le \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)\right)\right] \cdot \frac{\delta}{2n^2} && \text{(Markov inequality)}\\
&\le \frac{\delta}{2n^2}.
\end{aligned}$$

Hence the above holds for all $n \ge 1$ via a union bound which provides the lemma. □