JMLR: Workshop and Conference Proceedings vol 23 (2012) 40.1–40.24 25th Annual Conference on Learning Theory

Autonomous Exploration For Navigating In MDPs

Shiau Hong Lim SHONGLIM@GMAIL.COM
Peter Auer AUER@UNILEOBEN.AC.AT
Montanuniversität Leoben, Franz-Josef-Straße 18, 8700 Leoben, Austria.

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

While intrinsically motivated learning agents hold considerable promise to overcome limitations of more supervised learning systems, quantitative evaluation and theoretical analysis of such agents are difficult. We propose to consider a restricted setting for autonomous learning where systematic evaluation of learning performance is possible. In this setting the agent needs to learn to navigate in a Markov Decision Process where extrinsic rewards are not present or are ignored. We present a learning algorithm for this scenario and evaluate it by the amount of exploration it uses to learn the environment.

Keywords: autonomous exploration, reinforcement learning, optimism in the face of uncertainty, computational learning theory

1. Introduction

Learning agents with intrinsic motivation have been studied in many different settings (Schmidhuber, 1991, 2010; Oudeyer et al., 2007; Oudeyer and Kaplan, 2007; Baranes and Oudeyer, 2009; Singh et al., 2004, 2010). However, there is no well-defined performance measure or systematic theoretical analysis of actual, implementable autonomous learning agents. In the most general setting this is certainly a formidable task. We make a step in this direction by considering a restricted yet reasonably rich class of environments for which we can give tractable algorithms with well-defined performance guarantee.

The learning scenario and the learning algorithm that we analyze have some of the properties encountered also by biological learners: the environment is potentially infinite and the learner observes only a small part of it by autonomous exploration; the learner selects exploration paths which are the most promising to reveal new relevant information; the learner checks if its current knowledge of the environment is sufficiently accurate; the learner incrementally extends its knowledge about the environment by moving from simpler intrinsic goals to more complicated ones.

2. Problem Definition

We consider learning a discrete-time Markov Decision Process (MDP) $\mathcal{M}$ with no external rewards. We assume a countable, possibly infinite state space $\mathcal{S}$, a finite set of actions $\mathcal{A}$ with $|\mathcal{A}| = A$, and an unknown transition function P where $P(s'|s,a)$ gives the probability of reaching next state s' if action a is executed in state s. A learner can gain knowledge about its environment only by choosing an action in its current state and following the resulting trajectory.

The learner is expected to explore this environment autonomously. Such autonomously motivated exploration, however, should ideally be as efficient as possible, wasting no time or resources

on parts of the world that are already known to the agent. As suggested in (Schmidhuber, 2010), a quantitative measure of exploration progress is needed. We believe that a natural measure is in terms of the set of states that can be confidently reached by the agent from a designated reference or "neutral" starting state.

Definition 1 (Reachability) *Let s_0 be the designated neutral starting state. For any (possibly non-stationary) policy π, let $\tau(s|\pi)$ be the expected number of steps before reaching s for the first time, when executing policy π starting from s_0.*

Trying to reach a specific state s can be viewed as a subtask whose complexity is measured by the average number of steps necessary to reach s. Grouping states according to their reachability, we can consider the states reachable in L steps,[1]

$$\mathcal{S}_L = \{s \in \mathcal{S} : \min_\pi \tau(s|\pi) \leq L\}.$$

A natural question is: how many exploration steps are necessary (or sufficient) to learn, for every $s \in \mathcal{S}_L$, a policy π_s with $\tau(s|\pi_s) \leq (1+\varepsilon)L$ for some $\varepsilon > 0$?

Since the state space can be infinite, it is possible that a learner wanders off in some direction or gets stuck, without the ability to return to the starting state. To address this issue, we make the following assumption.

Assumption 1 *In every state there is a designated* RESET *action available, that will transition back to the starting state s_0 with probability* 1.[2]

The motivation for this learning scenario is that it captures interesting properties of an undirected exploration problem. In contrast to a standard MDP learning problem, where the goal is to find a good policy for a single problem, in the proposed scenario the learner has to find good policies for a set of problems: it needs to find a good policy for each reachable state. Still, what is learned by finding a good policy for one state, might be used to find a good policy for another state. The learner gets to know its environment in an efficient way.

We believe that this scenario cannot be solved efficiently by a simple adaptation of existing reinforcement learning algorithms. One problem is to deal with a potentially infinite state space and to concentrate exploration on a reachable subset of states. Our algorithm solves this problem by incrementally increasing the search area. Another problem is to generate appropriate reward functions that can be fed to a reinforcement learner. Our algorithm solves this by implicitly using a different reward function for each state that is to be reached. Still the question is when to switch from one reward function to a new one.

2.1. Negative Results

Ideally, one would look for a learning algorithm where the total number of exploration steps is polynomial in $|\mathcal{S}_{(1+\varepsilon)L}|$, A, L, and $1/\varepsilon$. The reason that the larger $|\mathcal{S}_{(1+\varepsilon)L}|$ instead of $|\mathcal{S}_L|$ is expected to appear in the exploration bound, is that the learner might be unable to distinguish between states

1. In all the following we assume that the starting state s_0 is fixed and we omit it from any notation.
2. It is possible to relax this requirement in several ways, such as a RESET that requires multiple actions or needs to be learned, but this mostly clouds the main ideas with unnecessary details.

reachable in L steps and those reachable in $(1+\varepsilon)L$ steps (given a reasonable amount of exploration). Thus the learner might learn also policies to reach states in $\mathcal{S}_{(1+\varepsilon)L}$, which needs to be reflected in the bound. Nevertheless, the following result shows that even with the larger $|\mathcal{S}_{(1+\varepsilon)L}|$ efficient learning is not possible in general.

Proposition 2 *For any learning algorithm, any $L \geq 2$, and any $T \geq 0$, there is an MDP $\mathcal{M}$ with actions $\mathcal{A} = \{a_0, a_1, \mathrm{RESET}\}$ and the following properties:*

- *there is only one state s_L (besides s_0) that is reachable in L steps, $\mathcal{S}_L = \{s_0, s_L\}$, and there is no other state reachable in $\frac{3}{2}L$ steps,*
- *after T exploration steps, the probability that the learning algorithm outputs a policy for reaching s_L in $\frac{3}{2}L$ steps is less than 2^{1-L}.*

The proof (given in the Appendix) uses a construction such that s_L is reachable in L steps only by a policy that touches upon an arbitrarily large number of intermediate states. To find this policy the intermediate states need to be explored, which takes time proportional to the number of intermediate states. Thus for a large number of intermediate states, s_L is reached only by luckily choosing the right L actions, which happens with exponentially small probability.

To avoid the problem with non-reachable intermediate states, one could attempt to learn only policies for states in the subset $\mathcal{S}_L^\circ \subseteq \mathcal{S}_L$, where each state in $\mathcal{S}_L^\circ$ can be reached in L steps by a policy *staying* in $\mathcal{S}_L^\circ$.

Definition 3 ($\mathcal{S}_L^\circ$) *A* policy π on $\mathcal{S}'$ *is a policy with $\pi(s) = \mathrm{RESET}$ for any $s \notin \mathcal{S}'$. Let $\mathcal{S}_L^\circ$ be the largest set such that all states in $\mathcal{S}_L^\circ$ are reachable in L steps by policies on $\mathcal{S}_L^\circ$.*

Unfortunately, learning $\mathcal{S}_L^\circ$ may still require an exponential number of exploration steps.

Proposition 4 *For any learning algorithm, any $L \geq 14$, and any $T \leq \frac{2^L}{8L^2}$, there is an MDP $\mathcal{M}$ with actions $\mathcal{A} = \{a_0, a_1, \mathrm{RESET}\}$ and the following properties:*

- *$|\mathcal{S}_L^\circ| \leq 4L^2$, and there are no other states reachable in $\frac{3}{2}L$ steps,*
- *there exists $g_0 \in \mathcal{S}_L^\circ$ such that after T exploration steps, the probability that the learning algorithm outputs a policy for reaching g_0 in $\frac{3}{2}L$ steps is less than $1/L$.*

The proof (given in the Appendix) exploits the difficulty in distinguishing reachable states (which need to be explored) from unreachable states (which should be ignored). The construction of the proof yields an exponential number of unreachable states that are indistinguishable from reachable states (without extensive exploration). Thus the learning algorithm spends an exponential number of exploration steps on unreachable states.

2.2. Positive Results

By concentrating on MDPs that allow for an incremental discovery of reachable states, we arrive at a definition for a subset of states $\mathcal{S}_L^\rightarrow$ that can indeed be learned efficiently.

Definition 5 ($\mathcal{S}_L^\rightarrow$) *Let $\prec$ be a partial order on $\mathcal{S}$. The set $\mathcal{S}_L^\prec$ of states* reachable in L steps incrementally in respect to $\prec$*, is defined as follows:*

- $s_0 \in \mathcal{S}_L^{\prec}$,
- *if there is a policy* π *on* $\{s' \in \mathcal{S}_L^{\prec} : s' \prec s\}$ *and* $\tau(s|\pi) \leq L$, *then* $s \in \mathcal{S}_L^{\prec}$.

The set $\mathcal{S}_L^{\rightarrow}$ *of states* reachable in L steps in respect to some partial order *is given by*

$$\mathcal{S}_L^{\rightarrow} = \bigcup_{\prec} \mathcal{S}_L^{\prec}.$$

Proposition 6 *The set* $\mathcal{S}_L^{\rightarrow}$ *is finite for any* L. *Furthermore, there exists a partial order* $\prec$ *with* $\mathcal{S}_L^{\rightarrow} = \mathcal{S}_L^{\prec}$.

The proof is given in the appendix.

Remark 7 *The definition of* $\mathcal{S}_L^{\rightarrow}$ *captures some intuition about the ability to incrementally learn navigation tasks. A partial order — unknown to the learner — ranks the difficulty of the navigation tasks, such that a learner that is adaptive to this difficulty might incrementally extend its skill to navigate in the environment.*

Our main result bounds the number of exploration steps necessary to learn policies for the states in $\mathcal{S}_L^{\rightarrow}$, using our new algorithm UcbExplore presented in the next section.

Theorem 8 *When algorithm* UcbExplore *is run with inputs* s_0, $\mathcal{A}$, $L \geq 1$, $\varepsilon > 0$, *and* $\delta \in (0,1)$, *then with probability* $1 - \delta$

- *it terminates after* $O\left(\frac{SAL^3}{\varepsilon^3}\left(\log\frac{SAL}{\varepsilon\delta}\right)^3\right)$ *exploration steps,*
- *discovers a set of states* $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$,
- *and for each* $s \in \mathcal{K}$ *outputs a policy* π_s *with* $\tau(s|\pi_s) \leq (1+\varepsilon)L$,

where $S = |\mathcal{K}| \leq |\mathcal{S}_{(1+\varepsilon)L}^{\rightarrow}|$.

By running algorithm UcbExplore with increasing $L = 1, (1+\varepsilon), (1+\varepsilon)^2, \ldots$, nearly optimal policies (in respect to some partial order) will be discovered for all reachable states.

Corollary 9 *If* UcbExplore *is run with* $L_k = (1+\varepsilon)^k$ *and* $\delta_k = \frac{\delta}{2(k+1)^2}$ *for* $k = 0, 1, 2, \ldots$, *then with probability* $1 - \delta$, *for any* $L \geq 1$ *and any* $s \in \mathcal{S}_L^{\rightarrow}$, *the algorithm will discover a policy* π_s *with* $\tau(s|\pi_s) \leq (1+\varepsilon)^2 L$ *after* $O\left(\frac{SAL^3}{\varepsilon^4}\left(\log\frac{SAL}{\varepsilon\delta}\right)^3\right)$ *exploration steps where* $S = |\mathcal{S}_{(1+\varepsilon)^2 L}^{\rightarrow}|$.

3. Algorithm UcbExplore

Figure 1 presents our algorithm UcbExplore. The main idea of the algorithm is to consider reaching a particular state as a task, and to choose an "optimistic" task in each iteration of the algorithm. By optimistic we mean choosing the easiest state to reach – the one that seems to be reachable in the shortest number of steps from s_0, based on information collected so far. The environment will then be explored using an optimistic policy for this optimistic task.

The algorithm keeps a set $\mathcal{K}$ of "known" states and a set $\mathcal{U}$ of "unknown" states. A state is "known" when a $(1+\varepsilon)L$-step policy for that state has been found (with high confidence). The

Input: A confidence parameter $\delta \in (0,1)$, an error threshold $\varepsilon > 0$, $L \geq 1$, $\mathcal{A}$ and s_0.
Output: A set of known reachable states $\mathcal{K}$ and corresponding policies π_s for all $s \in \mathcal{K}$.

1. Set $\epsilon \leftarrow \frac{\min(\varepsilon,1)}{8}$ and initialize $k \leftarrow 1, \mathcal{U} \leftarrow \{\}, \mathcal{K} \leftarrow \{\}$, and $s_{\text{new}} \leftarrow s_0$.

2. For each round,

   (a) **State Discovery**

   If $s_{\text{new}} \notin \mathcal{K}$, add s_{new} to $\mathcal{K}$, then sample each action $a \in \mathcal{A}$ in s_{new} $\left\lceil L \log \frac{8AL|\mathcal{K}|^2}{\delta} \right\rceil$ times, adding any newly discovered states into $\mathcal{U}$.

   Stop the algorithm if $\mathcal{U}$ is empty.

   (b) **Compute Optimistic Policy**

   For each $s \in \mathcal{U}$, compute a Γ-step optimistic policy $\tilde{\pi}_s$ and its corresponding value $\tilde{u}_\Gamma(s_0|\tilde{\pi}_s, \mathcal{M}_s)$, where $\Gamma = \lceil (1+\frac{1}{\epsilon})L \rceil$. Let

   $$\tilde{u}^* = \max_{s \in \mathcal{U}} \tilde{u}_\Gamma(s_0|\tilde{\pi}_s, \mathcal{M}_s).$$

   Stop the algorithm if $\tilde{u}^* < \Gamma - L$. Otherwise choose a state $\tilde{s} \in \mathcal{U}$ as the target state satisfying $\tilde{u}_\Gamma(s_0|\tilde{\pi}_{\tilde{s}}, \mathcal{M}_{\tilde{s}}) = \tilde{u}^*$ and set the policy $\tilde{\pi} \leftarrow \tilde{\pi}_{\tilde{s}}$.

   (c) **Policy Evaluation**

   Run $\tilde{\pi}$ for up to $\lambda = \left\lceil \frac{6}{\epsilon^3} \log\left(\frac{16|\mathcal{K}|^2}{\delta}\right) \right\rceil$ episodes. Each episode begins at s_0 and ends either when $\tilde{s}$ is reached or Γ steps have been executed. The average number of steps $\hat{\tau}$ to reach $\tilde{s}$, and the fraction $\hat{p}$ of episodes that failed to reach $\tilde{s}$, are updated after each episode. Policy evaluation is terminated before finishing all λ episodes, if one of the following happens:

   - If $\frac{\hat{\tau}+\epsilon L+\hat{p}+\epsilon}{1-(\hat{p}+\epsilon)} > (1+8\epsilon)L$ after any episode, then $k \leftarrow k+1$ and a new round is started (the current round has been a *failure round*).
   - For any state-action pair (s,a), $a \neq \text{RESET}$, let $N(s,a)$ be the total number of times (s,a) has been executed in previous rounds, and let $v(s,a)$ be the number of times (s,a) has been executed in the current round. If $v(s,a) \geq \max\{1, N(s,a)\}$, then start a new round (the current round has been a *skipped round*).

   If the current round is neither a failure nor a skipped round (it is a *success round*), then remove $\tilde{s}$ from $\mathcal{U}$, set $s_{\text{new}} \leftarrow \tilde{s}$ and output $\pi_{s_{\text{new}}} \leftarrow \tilde{\pi}$.

Figure 1: Algorithm UcbExplore

states in $\mathcal{U}$ are states that have been discovered as potential members of $\mathcal{S}_L^{\rightarrow}$, but the algorithm has yet to produce a $(1+\varepsilon)L$-step policy for any of them. All observed state transitions are recorded, to be used as samples for computing future policies.

For any state s^*, we define an induced MDP $\mathcal{M}_{s^*}$ such that all actions in the state s^* in $\mathcal{M}_{s^*}$ give reward 1 and transition back to s^* with probability 1. All other states and actions in M_{s^*} behave

exactly as in $\mathcal{M}$, and give zero rewards. Thus maximizing the total rewards in $\mathcal{M}_{s^*}$ is equivalent to minimizing the number of steps to reach s^*.[3]

Each major iteration of the algorithm is referred to as a "round". In each round, a new optimistic target state $\tilde{s} \in \mathcal{U}$ is chosen. The optimistic policy for $\mathcal{M}_{\tilde{s}}$ will then be executed for a number of episodes where each episode can be up to $\Gamma = \lceil (1 + \frac{1}{\epsilon})L \rceil$ steps. The outcome of a round can be either a success or a failure. If it is a success, $\tilde{s}$ will become "known". The algorithm stops when it is highly likely that all states in $\mathcal{S}_L^{\rightarrow}$ are already known. Otherwise a new round will begin.

The following subsections describe each of the three major steps in the algorithm in more details.

3.1. State Discovery

Since the state space is unknown and possibly infinite, there is a need for state discovery. Whenever there is a new "known" state s_{new}, this step is performed in order to discover any states reachable from s_{new}. By definition of a known state, the algorithm has found a policy $\pi_{s_{\text{new}}}$ that can reach s_{new} in $(1+\varepsilon)L$ steps. Using this policy, it is possible to sample any action $a \in \mathcal{A}$ in s_{new} by first resetting to s_0 and then executing $\pi_{s_{\text{new}}}$ until s_{new} is reached. Each sample requires on average at most $(1+\varepsilon)L + 1$ steps.

3.2. Computing Optimistic Policies

Let $\Gamma = \lceil (1 + \frac{1}{\epsilon})L \rceil$ and let π be a policy with horizon Γ. For $i \in \{0, \ldots, \Gamma\}$, let $u_i(s|\pi, \mathcal{M}_{s^*})$ be the expected total i-step reward if π is followed for i steps beginning at state s in $\mathcal{M}_{s^*}$. Let

$$u_i^*(s|\mathcal{M}_{s^*}) = \max_{\pi \text{ on } \mathcal{K}} u_i(s|\pi, \mathcal{M}_{s^*})$$

be the expected i-step total reward for an optimal policy (restricted to the known states).

Central to the algorithm is the computation of an "optimistic" policy, which is an optimal policy with respect to an optimistic estimation of $u_i^*(s|\mathcal{M}_{s^*})$. The optimistic i-step reward $\tilde{u}_i(s|\mathcal{M}_{s^*})$ is an upper confidence bound of $u_i^*(s|\mathcal{M}_{s^*})$ computed based on an approximate transition function $\hat{P}_i(\cdot|s,a)$ using transitions observed in the past. Figure 2 gives the algorithm for computing an optimistic policy.

Note that one of the inputs to the algorithm in Fig. 2 is a round index k. As explained in the next section, the index k is incremented only after a "failure" round.

3.3. Policy Evaluation

In each round, once an optimistic target state $\tilde{s}$ is chosen, the corresponding optimistic policy is evaluated. The evaluation is performed in a number of episodes. In each episode, the policy is executed, starting at s_0, until either $\tilde{s}$ is reached or Γ steps have been executed. If $\tilde{s}$ is reached the episode is considered a success, otherwise it is a failure.

After each episode, the average number of steps per episode (regardless of success) is lower bounded by

$$\hat{\tau} = \frac{\sum_{j=1}^{n} \hat{\tau}_j}{\lambda}$$

3. Losses instead of rewards can be used by assigning loss 1 to any action taken in states other than s^*, the results would be mathematically equivalent. Although losses might be more natural in this setting, we use rewards for compatibility with UCRL2 (Jaksch et al., 2010), from which we borrow some ideas.

Input: $\mathcal{A}$, $\mathcal{K}$, $\mathcal{U}$, L, Γ, ϵ, δ, s_0, target state $s^* \in \mathcal{U}$, an index $k \in \{1, 2, \ldots\}$ and for each $s \in \mathcal{K}$,$a \in \mathcal{A}$, a set of $N(s,a) \geq 0$ independent transitions.
Output: An optimistic policy $\tilde{\pi}$ for $\mathcal{M}_{s^*}$ and its corresponding Γ-step value $\tilde{u}_\Gamma(s_0)$.

1. For each $s \in \mathcal{K}$, $a \in \mathcal{A}$, divide the set of $N(s,a)$ transitions into Γ disjoint sets, each with at least $\lfloor \frac{N(s,a)}{\Gamma} \rfloor$ transition samples. Let $N_i(s,a)$ be the total number of transitions in the i-th sample set and $v_i(s'|s,a)$ be the number of transitions that end up in state s' in the i-th sample set. If $N_i(s,a) > 0$, let $\hat{P}_i(s'|s,a) = \frac{v_i(s'|s,a)}{N_i(s,a)}$ for all s'. Otherwise let $\hat{P}_i(s^*|s,a) = 1$ and $\hat{P}_i(s'|s,a) = 0$ for all other states s'.
2. Let $\tilde{u}_0(s) = 0$ for all s.
3. Let $\tilde{u}_i(s^*) = i$ for $i = 1, \ldots, \Gamma$.
4. For each $i = 1, \ldots, \Gamma$
   - For each $s \in \mathcal{K}$ and each $a \in \mathcal{A}$, let
   $$\tilde{q}_i(s,a) = \min\left\{i, \left(\hat{P}_i(\cdot|s,a)\tilde{u}_{i-1}(\cdot) + \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N(s,a)\}}}\right)\right\}$$
   and
   $$\tilde{u}_i(s) = \max_a \tilde{q}_i(s,a)$$
   where $\sigma_k = \sqrt{\log \frac{2A|\mathcal{U}||\mathcal{K}|\Gamma k^5}{\epsilon^4 \delta}}$.
   - For all $s \notin \mathcal{K}$, $s \neq s^*$, let $\tilde{u}_i(s) = \tilde{u}_{i-1}(s_0)$.
5. For $s \in \mathcal{K}$ and $i = 0, \ldots, \Gamma - 1$, let $\tilde{\pi}_i(s) = \arg\max_a \tilde{q}_{\Gamma - i}(s,a)$.
   (After executing Γ steps the policy resets to s_0 and restarts.)
   For each $s \notin \mathcal{K}$, let $\tilde{\pi}_i(s) =$ RESET.

Figure 2: Algorithm for computing an optimistic policy

where $\lambda = \left\lceil \frac{6}{\epsilon^3} \log\left(\frac{16|\mathcal{K}|^2}{\delta}\right)\right\rceil$, $n \leq \lambda$ is the number of episodes so far, and $\hat{\tau}_j$ is the actual number of steps taken before the episode ends. Also, the failure rate is lower bounded by

$$\hat{p} = \frac{f}{\lambda}$$

where f is the number of episodes (out of n) that have failed to reach $\tilde{s}$ so far.

At the end of each episode, a performance check is carried out. If

$$\frac{\hat{\tau} + \epsilon L + \hat{p} + \epsilon}{1 - (\hat{p} + \epsilon)} > (1 + 8\epsilon)L$$

then the round is considered a "failure" and a new round will begin, with k incremented. If λ episodes have been executed without failing the performance check, then this is a successful round and the target $\tilde{s}$ will become a new "known" state.

To prevent potentially bad state-action pairs from getting executed too many times in a single round, a round is terminated early when this happens (resulting in a skipped round, see Step 2c in Fig. 1).

4. Analysis of Algorithm UcbExplore and Proof of Theorem 8

We can classify each round in the main algorithm into 3 types based on its outcome:

1. A successful round where a new known state is removed from $\mathcal{U}$ and gets added to $\mathcal{K}$.
2. A failure round, which is terminated due to a failed performance check (Section 3.3).
3. A "skipped" round, which is terminated due to frequent visits to a possibly bad state (Section 3.3).

Note that the index k in the algorithm only gets incremented in a failure round. It will be shown that the number of rounds for the other two types can be easily bounded. When we mention "round k" it will mean failure round k.

For most of the lemmas in the subsequent sections we give a "proof sketch", which is a short, intuitive version of the proof, while the full proof is given in the appendix.

4.1. Bound on the Number of Steps in Each Round

We first give a useful bound on the maximum number of steps that can actually be executed in each round of the algorithm (Lemma 10). A consequence of this is that we can then bound the number of "skipped" rounds (Lemma 11).

Lemma 10 *Assuming $\epsilon \in (0, \frac{1}{8}]$, the number of actual steps executed in any round is at most $2L\lambda$ where $\lambda = \left\lceil \frac{6}{\epsilon^3} \log \frac{16|\mathcal{K}|^2}{\delta} \right\rceil$.*

Proof Sketch For a successful round, since it passed the performance check (to verify that it can reach the target in $L + O(\epsilon L)$ steps) in all λ episodes, the average steps per episode must be at most $L + O(\epsilon L)$ steps. For a failure or "skipped" round, suppose a policy failed the performance check after some $n \leq \lambda$ episodes, it means that it passed the check in all the previous $n - 1$ episodes, and the last episode only adds at most $O(\frac{L}{\epsilon}) = O(\epsilon L \lambda)$ steps to the total. ■

Lemma 11 *At any round, there can be at most $\log_2 4L\lambda$ previously skipped rounds due to any particular state-action pair, where $\lambda = \left\lceil \frac{6}{\epsilon^3} \log \frac{16|\mathcal{K}|^2}{\delta} \right\rceil$.*

Proof By Lemma 10 the total number of steps in any round is at most $2L\lambda$. In order for a round to be skipped due to a particular state-action pair (s, a), it must be that $N(s, a) \leq v(s, a) \leq 2L\lambda$, which means that its total previous visits must be at most $2L\lambda$.

Since, after every "skipped" round, the total number of visits for this state-action pair will be doubled, it follows that this can only happen at most $\log_2 4L\lambda$ times. ■

4.2. State Discovery

Since $\mathcal{S}_L^{\rightarrow}$ is unknown, an important aspect of the algorithm is to ensure, with high probability, that none of the states in $\mathcal{S}_L^{\rightarrow}$ are "missed". In particular, we need to ensure that in every round, unless all of $\mathcal{S}_L^{\rightarrow}$ is already known, at least one of the states in $\mathcal{S}_L^{\rightarrow}$ is also in $\mathcal{U}$ and that it is reachable in L steps with a policy restricted to $\mathcal{K}$ (RESET will be performed in all other states). Lemma 12 provides the necessary guarantee.

Given any state $s \in \mathcal{S}_L^{\rightarrow}$, we define $\mathcal{S}_{\prec s} = \{s' \in \mathcal{S}_L^{\rightarrow} : s' \prec s\}$ with respect to a partial order $\prec$ such that there is a π_s^* on $\mathcal{S}_{\prec s}$ with $\tau(s|\pi_s^*) \leq L$. Proposition 6 guarantees that there is at least one such partial order on $\mathcal{S}_L^{\rightarrow}$.

Lemma 12 *With probability at least $1-\frac{\delta}{4}$, at any round, either $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$ or there exists $s^* \in \mathcal{S}_L^{\rightarrow} \setminus \mathcal{K}$ such that $s^* \in \mathcal{U}$ and $\mathcal{S}_{\prec s^*} \subseteq \mathcal{K}$.*

Proof In any round such that $\mathcal{S}_L^{\rightarrow} \setminus \mathcal{K}$ is not empty, there is an $s^* \in \mathcal{S}_L^{\rightarrow} \setminus \mathcal{K}$ with $\mathcal{S}_{\prec s^*} \subseteq \mathcal{K}$. To show that $s^* \in \mathcal{U}$, first note that since $\mathcal{S}_{\prec s^*} \subseteq \mathcal{K}$, there exists a policy on $\mathcal{K}$ that can reach s^* in L steps. Thus there is an $s \in \mathcal{K}$ and an action a with $P(s^*|s,a) \geq \frac{1}{L}$, and hence with high probability s^* is found during the state discovery phase for s (see Step 2a in Fig. 1). This is made formal by Lemma 17 in the appendix. ■

4.3. Policy Evaluation and Performance Bound

In each round a new optimistic policy $\tilde{\pi}$ for an optimistic target state $\tilde{s}$ is computed and then evaluated on $\mathcal{M}$. If $\tilde{\pi}$ passes the performance check (Section 3.3) then $\tilde{s}$ will become known. The following lemma ensures that with high probability, all policies that pass the performance check can reach the target state in $(1+8\epsilon)L$ steps.

Lemma 13 *Let $\pi_{s_1}, \pi_{s_2}, \ldots, \pi_{s_n}$ be any sequence of policies output by the algorithm for the corresponding target states $s_1, s_2, \ldots, s_n$. For any $0 < \epsilon \leq \frac{1}{8}$, with probability at least $1 - \frac{\delta}{4}$, $\tau(s_i|\pi_{s_i}) \leq (1+8\epsilon)L$ for all policies π_{s_i} in the sequence.*

Proof Sketch Let π be a policy output by the algorithm and let T_Γ be a random variable denoting the total number of steps that it takes before reaching either the target state or the end of an episode (Γ steps). Since π must have passed the performance check, it's empirical performance satisfies $\frac{\hat{\tau}+\epsilon L+\hat{p}+\epsilon}{1-(\hat{p}+\epsilon)} \leq (1+8\epsilon)L$ where $\hat{\tau}$ is the empirical average of T_Γ and $\hat{p}$ is the empirical failure rate (of reaching the target state). Since T_Γ is bounded between 0 and Γ, it is possible to bound its variance, and to show by applying Bernstein's inequality that passing the performance check implies $\mathrm{E}(T_\Gamma) \leq \hat{\tau} + \epsilon L$ and the true failure rate $p \leq \hat{p} + \epsilon$ with high probability.

It is then possible to show that a non-stationary, infinite-horizon policy can be derived from π by simply performing the RESET action (and repeat the same policy) whenever the target is not reached after Γ steps, and the expected number of steps to reach the target state with the resulting policy will be at most $(1+8\epsilon)L$. ■

4.4. Optimistic Policy

The following lemma shows that the value estimate using the empirical transition probabilities $\hat{P}_i(\cdot|s,a)\tilde{u}_{i-1}(\cdot)$ is close to $P(\cdot|s,a)\tilde{u}_{i-1}(\cdot)$ for large $N(s,a)$. Note that we omit the target state s^* in the notation whenever it is clear based on the context, where $\tilde{u}_i(s)$ means $\tilde{u}_i(s|\mathcal{M}_{s^*})$ and $u_i^*(s)$ means $u_i^*(s|\mathcal{M}_{s^*})$. One consequence of this lemma is that the policy computed by the algorithm in Fig. 2 is optimistic with high probability.

Lemma 14 *At round k, with probability at least $1 - \frac{\epsilon^4\delta}{k^5}$, for every target $s^* \in \mathcal{U}$ such that $u_\Gamma^*(s_0|\mathcal{M}_{s^*}) \geq \Gamma - L$, for every state $s \in \mathcal{K}$, every action $a \in \mathcal{A}$ and every $i \in \{1,\ldots,\Gamma\}$,*

$$\left|\hat{P}_i(\cdot|s,a)\tilde{u}_{i-1}(\cdot) - P(\cdot|s,a)\tilde{u}_{i-1}(\cdot)\right| \leq \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1,N(s,a)\}}} \tag{1}$$

and

$$\tilde{u}_i(s) \geq u_i^*(s) \qquad . \tag{2}$$

Proof Sketch The proof makes use of two key ideas:

1. First, by construction, $\hat{P}_i(\cdot|s,a)$ and $\tilde{u}_{i-1}(\cdot)$ are independent for any $i \in \{1,\ldots,\Gamma\}$ since they are computed using separate samples of past transitions (see Step 1 in Fig. 2).
2. Second, due to the RESET action, the value of $\tilde{u}_i(s)$ for any s is lower bounded by $\tilde{u}_{i-1}(s_0)$. Since $\tilde{u}_i$ is also upper bounded by i, and assuming that $\tilde{u}_{i-1}(s_0)$ is optimistic, the range of values for $\tilde{u}_i(s)$ is at most $L+1$.

With the above, we can obtain inequality (1) by applying Hoeffding's inequality. By using the same reasoning inductively for $i = 1,\ldots,\Gamma$, (2) can be easily derived and a union bound can be applied to obtain the total failure probability. ■

4.5. Regret Bounds

We make use of regret bounds to bound the number of failure rounds. The regret in a Γ-step episode is defined as

$$\Gamma - L - \sum_{i=0}^{\Gamma-1} r_i$$

where r_i is the reward received at step i in the episode. We consider the regret only in failure rounds. First we establish that the total regret of a failure round is at least some Δ_{failure} – this is provided by Lemma 15. Next we upper-bound the total regret after m failure rounds by some Δ – this is provided by Lemma 16. It follows that the number of failure rounds is bounded as $m \leq \frac{\Delta}{\Delta_{\text{failure}}}$.

Lemma 15 *In any failure round, the total regret is at least $\lambda\epsilon L$ where $\lambda = \left\lceil \frac{6}{\epsilon^3}\log\frac{16|\mathcal{K}|^2}{\delta}\right\rceil$.*

Proof Sketch This follows from a failed performance check (see Section 3.3), implying that the target state is reached only after $(1+\epsilon)L$ steps on average. ■

Lemma 16 *Let $S = |\mathcal{S}^{\rightarrow}_{(1+8\epsilon)L}|$. Then, with probability at least $1-\delta$, the average per-round regret in m failure rounds is at most*

$$173\frac{L^2\sqrt{SA}}{\epsilon^2\sqrt{m}}\log\frac{m}{\epsilon\delta}$$

for all $m \geq SAL$.

Proof Sketch Lemma 14 provides the key inequality needed to bound the regret. The rest of the proof borrows ideas from Jaksch et al. (2010). Please refer to the appendix for details. ■

4.6. Proof of Theorem 8

Proof First, we count the total steps in successful rounds. By Lemma 13, with probability at least $1-\delta/4$ all the policies output by the algorithm have $\tau(s|\pi_s) \leq (1+8\epsilon)L$. Therefore there are most $S = |\mathcal{S}^{\rightarrow}_{(1+8\epsilon)L}|$ successful rounds, each taking at most $O(\frac{L}{\epsilon^3}\log\frac{S}{\delta})$ steps (by Lemma 10). Additionally, for each new known state and each action, there are $O(L^2\log\frac{ALS}{\delta})$ discovery steps. We therefore have a total of $O(\frac{SAL^2}{\epsilon^3}\log\frac{SAL}{\delta})$ steps.

Next, we look at the number of "skipped" rounds. Since for each state-action pair there are at most $O(\log\frac{LS}{\epsilon\delta})$ skipped rounds (Lemma 11), each with at most $O(\frac{L}{\epsilon^3}\log\frac{S}{\delta})$ steps (Lemma 10), the total number of steps is $O(\frac{SAL}{\epsilon^3}\log\frac{SL}{\epsilon\delta})$.

We now focus on the number of failure rounds. We show that

$$M = 9\cdot 15^2 SAL^2\left(\log\frac{15^2SAL^2}{\epsilon\delta}\right)^2$$

is an upper bound on the number of failure rounds. Let $\alpha = 15^2SAL^2$. Using the fact that $x > 3\log x$ for all $x > 5$, it follows that

$$M = 9\alpha\left[\log\frac{\alpha}{\epsilon\delta}\right]^2 = \alpha\left[\log\frac{\alpha}{\epsilon\delta}\cdot\left(\frac{\alpha}{\epsilon\delta}\right)^2\right]^2 > \alpha\left[\log\frac{\alpha}{\epsilon\delta}\cdot 9\left(\log\frac{\alpha}{\epsilon\delta}\right)^2\right]^2 = \alpha\left(\log\frac{M}{\epsilon\delta}\right)^2 .$$

By Lemma 16, the per-round total regret after $m \geq M$ failure rounds is bounded by

$$173\frac{L^2\sqrt{SA}}{\epsilon^2\sqrt{m}}\left(\log\frac{m}{\epsilon\delta}\right) < 173\frac{L^2\sqrt{SA}}{\epsilon^2\sqrt{\alpha}} < 12\frac{L}{\epsilon^2} < \lambda_1\epsilon L\,,$$

where $\lambda_1 = \left\lceil\frac{6}{\epsilon^3}\log\frac{16}{\delta}\right\rceil$. Hence there is a failure round $k \leq m$ with total per-round regret less than $\lambda_1\epsilon L$. This, however, contradicts Lemma 15, which states that such a round cannot be a failure round. We conclude that the number of failure rounds is at most $M = O(SAL^2\left(\log\frac{SAL}{\epsilon\delta}\right)^2)$. Thus, by Lemma 10, the total number of steps in failure rounds is $O(\frac{SAL^3}{\epsilon^3}\left(\log\frac{SAL}{\epsilon\delta}\right)^3)$.

Finally, Lemma 12 guarantees that if the algorithm stops, all states in $\mathcal{S}^{\rightarrow}_L$ are known, and Lemma 13 gives the performance guarantee for all policies output by the algorithm, using the fact that $\epsilon = \frac{\min(1,\varepsilon)}{8}$. ■

4.7. Discussion

Our algorithm employs the idea of optimism under uncertainty, which underlies many PAC-MDP algorithms (Kearns and Singh, 1998; Brafman and Tennenholtz, 2002; Kakade, 2003; Strehl et al., 2006; Szita and Szepesvári, 2010). A particular point that we need to clarify is regarding the notion of "known" states. The meaning of a "known" state in UcbExplore is very different from that in the R-MAX algorithm (Brafman and Tennenholtz, 2002; Kakade, 2003). In UcbExplore, a state is "known" if we have learned a good policy to reach it. On the other hand, in R-MAX, a state is "known" if we have sampled its actions sufficiently often.

It remains an open question if the exploration bound for our algorithm is optimal. One would expect that the bounds can be improved to $\tilde{O}\left(\frac{SAL^2}{\epsilon^2}\right)$, but this has not been achieved. New methods will be necessary to obtain such an improvement.

Acknowledgements

We thank the anonymous reviewers for their very valuable comments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231495 (CompLACS) and n° 216886 (PASCAL2).

References

A. Baranes and P.-Y. Oudeyer. R-IAC: Robust Intrinsically Motivated Exploration and Active Learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169, Oct. 2009. ISSN 1943-0604. doi: 10.1109/TAMD.2009.2037513.

R. I. Brafman and M. Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 99:1563–1600, August 2010. ISSN 1532-4435.

S. M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, Gatsby Computationel Neuroscience Unit, University College London, 2003.

M. J. Kearns and S. P. Singh. Near-optimal reinforcement learning in polynominal time. In *ICML*, pages 260–268, 1998.

P.-Y. Oudeyer and F. Kaplan. What is Intrinsic Motivation? A Typology of Computational Approaches. *Frontiers in neurorobotics*, 1(November):6, Jan. 2007. ISSN 1662-5218. doi: 10.3389/neuro.12.006.2007.

P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11:265–286, 2007.

J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats*, pages 222–227, Cambridge, MA, USA, 1991. MIT Press. ISBN 0-262-63138-5.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (19902010). *Autonomous Mental Development, IEEE Transactions on*, 2(3):230–247, 2010.

S. P. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *NIPS*, 2004.

S. P. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE T. Autonomous Mental Development*, 2(2):70–82, 2010.

A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. Pac model-free reinforcement learning. In *ICML*, pages 881–888, 2006.

I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, pages 1031–1038, 2010.

Appendix A. Proofs for Section 2

A.1. Proof of Proposition 2

Proof Let $n \gg T$ and $n \gg L$, and consider the state space

$$\mathcal{S} = \{s_0, s_1^{(1)}, \ldots, s_1^{(n)}, s_2^{(1)}, \ldots, s_2^{(n)}, \ldots, s_{L-1}^{(1)}, \ldots, s_{L-1}^{(n)}, s_L\}$$

For the transition function, let

- $P(s_1^{(k)}|s_0, a_i) = 1/n$ for $k \in \{1, \ldots, n\}$ and $i \in \{0, 1\}$,
- and for each $l \in \{1, \ldots, L-1\}$ and $k \in \{1, \ldots, n\}$ there is an index $I(l,k) \in \{0,1\}$ with $P(s_{l+1}^{(k)}|s_l^{(k)}, a_{I(l,k)}) = 1$ and $P(s_0|s_l^{(k)}, a_{1-I(l,k)}) = 1$ (for notational convenience let $s_L^{(1)} = \ldots = s_L^{(n)} = s_L$).

We consider an independent and uniform random choice of the indices $I(l,k)$. After T exploration steps, any learning algorithm has explored at most T of the n paths $s_0 \to s_1^{(k)} \to \cdots \to s_{L-1}^{(k)} \to s_L$. The probability of hitting such an explored path by the random transition from s_0 is at most T/n. The probability — in respect to the random indices $I(l,k)$ — of reaching s_L on an unexplored path is 2^{-L}. Thus for large n the probability of reaching s_L in $\frac{3}{2}L$ steps is at most 2^{1-L}. Hence for any learning algorithm there is a fixed choice of indices $I(l,k)$, such that the probability — in respect to the randomization of the learning algorithm — of reaching s_L in $\frac{3}{2}L$ steps is less than 2^{1-L}.

For any fixed choice of indices $I(l,k)$ state s_L can be reached in L steps, and for large n no other state can be reached in $\frac{3}{2}L$ steps. ■

A.2. Proof of Proposition 4

Proof Consider an infinite state space $\mathcal{S}$ that is partitioned into a binary tree where each node is a disjoint subset of $\mathcal{S}$. Let $\mathcal{S}_0$ be the root node. Let $\mathcal{S}^{(0)}$ and $\mathcal{S}^{(1)}$ be the two child nodes of $\mathcal{S}_0$. Similarly, let $\mathcal{S}^{(p0)}$ and $\mathcal{S}^{(p1)}$ be the child nodes of $\mathcal{S}^{(p)}$ where p encodes the binary "path" leading to $\mathcal{S}^{(p)}$. So, an example node at depth 5 would be $\mathcal{S}^{(01101)}$.

Let $\mathcal{S}_0 = \{s_0\}$. For any other nodes, let $\mathcal{S}^{(p)} = \{s_1^{(p)}, \ldots, s_{2L}^{(p)}\}$. For the transition function, let

- $P(s_k^{(i)}|s_0, a_i) = \frac{1}{2L}$ for $k = 1, \ldots, 2L$ and $i = 0, 1$.
- $P(s_k^{(pi)}|s_j^{(p)}, a_i) = \frac{1}{2L}$ for $j = 1, \ldots, 2L$, $k = 1, \ldots, 2L$ and $i = 0, 1$.

Note that it is possible to deterministically reach any specific node at depth d within d steps from s_0, but not any particular state within the node.

Now we expand $\mathcal{S}$ by adding an additional set of states $\mathcal{T}$. The states in $\mathcal{T}$ are organized in a binary tree of depth $\log_2(2L^2)$ such that each node contains exactly one state. Let g_0 be the root node of $\mathcal{T}$ and $t^{(q)}$ be a node encoded by its path q. All transitions within $\mathcal{T}$ are deterministic such that $P(t^{(qi)}|t^{(q)}, a_i) = 1$ for $i = 0, 1$. There are $2^{\log_2(2L^2)} = 2L^2$ leaf nodes in $\mathcal{T}$.

Consider a uniform random choice over all paths of length $L - 1 - \log_2(2L^2)$ (in the original $\mathcal{S}$) starting from $\mathcal{S}_0$. Let p^* be a particular choice. Let $\mathcal{S}^*$ be the set of all states in every node along this path up to node $\mathcal{S}^{(p^*)}$. For each state in $\mathcal{S}^{(p^*)}$, modify the transition function such that $P(g_0|s_j^{(p^*)}, a_i) = 1$ for $j = 1, \ldots, 2L$ and $i = 0, 1$. Since each node from $\mathcal{S}$ contains at most $2L$ states, the total number of states in $\mathcal{S}^*$ is at most $2L^2$. Let each of these states be a unique leaf node of $\mathcal{T}$.

We now show that in this modified state space, all states in $\mathcal{S}^*$ can be reached in L steps. To see this, first note that it is possible to reach any one of the states in $\mathcal{S}^{(p^*)}$ in $L - 1 - \log_2(2L^2)$ steps. After that, any action will deterministically transition to g_0, and from g_0 it only takes $\log_2(2L^2)$ steps to reach any leaf node – each corresponds to a node in $\mathcal{S}^*$. Adding up, the total number of steps is L.

We therefore have $\mathcal{S}_L^\circ = \mathcal{S}^* \cup \mathcal{T}$ since it contains all the states that are reachable in L steps with a policy on $\mathcal{S}^* \cup \mathcal{T}$. Furthermore, $|\mathcal{S}_L^\circ| \leq 4L^2$. All other states require at least $2L$ steps (on average) to reach.

Suppose a given learning algorithm stops and outputs π before reaching the node $\mathcal{S}^{(p^*)}$. Since p^* is unknown to the algorithm, it is possible to choose p^* such that π has the probability of at most $(\frac{1}{2})^{L-1-\log_2(2L^2)} = \frac{4L^2}{2^L}$ of reaching $\mathcal{S}^{(p^*)}$. It follows that to output an $\frac{3}{2}L$-step policy the node $\mathcal{S}^{(p^*)}$ must be visited at least once.

There are $n = 2^{L-1-\log_2(2L^2)} = \frac{2^L}{4L^2}$ nodes at depth $L - 1 - \log_2(2L^2)$, each requires $> L/2$ steps to reach from s_0. For any learning algorithm there is a choice of p^*, such that the probability – in respect to the randomization of the algorithm – of reaching g_0 after $T \leq \frac{2^L}{8L^2}$ exploration steps is at most $1/L$. ■

A.3. Proof of Proposition 6

Proof Since $\mathcal{S}_L^\prec \subseteq \mathcal{S}_L$ for any partial order $\prec$, we have $\mathcal{S}_L^\rightarrow \subseteq \mathcal{S}_L$. Thus for finiteness of $\mathcal{S}_L^\rightarrow$ we show that $\mathcal{S}_L$ is finite.

Let $p_k(s)$ be the probability of being in state s after exactly k steps by following the policy that maximizes this probability. Then

$$\begin{aligned} p_{k+1}(s') &\leq \sum_s p_k(s) \max_a P(s'|s, a) \\ &\leq \sum_s p_k(s) \sum_a P(s'|s, a) \end{aligned}$$

and

$$\begin{aligned}\sum_{s'} p_{k+1}(s') &\leq \sum_{s'}\sum_{s} p_k(s)\sum_{a} P(s'|s,a) \\ &= \sum_{s} p_k(s)\sum_{a} 1 \\ &\leq A\sum_{s} p_k(s).\end{aligned}$$

Since $\sum_s p_0(s) = 1$, we have $\sum_s p_k(s) \leq A^k$. By Markov's inequality, for any state $s \in \mathcal{S}_L$, the probability that it is reached by its optimal policy within $2L$ steps is at least $1/2$, therefore $\sum_{k=0}^{2L} p_k(s) \geq 1/2$. Consequently,

$$2A^{2L} \geq \sum_{k=0}^{2L} A^k \geq \sum_{k=0}^{2L}\sum_{s} p_k(s) \geq \sum_{s\in\mathcal{S}_L}\sum_{k=0}^{2L} p_k(s) \geq \frac{1}{2}|\mathcal{S}_L|$$

which implies $|\mathcal{S}_L| \leq 4A^{2L}$.

Since $\mathcal{S}_L^{\rightarrow}$ is finite, it can be represented as the union of finitely many $\mathcal{S}_L^{\prec}$. Thus it is sufficient to show that for any partial orders $\prec_\alpha$ and $\prec_\beta$ there is a partial order $\prec_\gamma$ with $\mathcal{S}_L^{\prec_\alpha} \cup \mathcal{S}_L^{\prec_\beta} \subseteq \mathcal{S}_L^{\prec_\gamma}$. We define $\prec_\gamma$ on $\mathcal{S}_L^{\prec_\alpha} \cup \mathcal{S}_L^{\prec_\beta}$ as the transitive closure of

$$\{s' \prec_\gamma s\} = \{s' \prec_\alpha s : s', s \in \mathcal{S}_L^{\prec_\alpha}\} \cup \{s' \prec_\beta s : s \notin \mathcal{S}_L^{\prec_\alpha}\}.$$

Since $\prec_\gamma$ extends $\prec_\alpha$ by adding only relations $s' \prec_\gamma s$ with $s \notin \mathcal{S}_L^{\prec_\alpha}$, $\prec_\gamma$ is a partial order on $\mathcal{S}_L^{\prec_\alpha} \cup \mathcal{S}_L^{\prec_\beta}$. Since $\{s' : s' \prec_\alpha s\} = \{s' : s' \prec_\gamma s\}$ for $s \in \mathcal{S}_L^{\prec_\alpha}$ and $\{s' : s' \prec_\beta s\} \subseteq \{s' : s' \prec_\gamma s\}$ for $s \notin \mathcal{S}_L^{\prec_\alpha}$, we get that $\mathcal{S}_L^{\prec_\alpha} \cup \mathcal{S}_L^{\prec_\beta} \subseteq \mathcal{S}_L^{\prec_\gamma}$. ■

A.4. Proof of Corollary 9

Proof The corollary follows from Theorem 8 by observing that for any $L \geq 1$ there is a K with $L_K/(1+\epsilon) \leq L \leq L_K$, and that

$$\sum_{k=0}^{K} L_k^3 \left(\log\frac{SAL_k}{\epsilon\delta_k}\right)^3 = O\left(\frac{L_K^3}{\epsilon}\left(\log\frac{SAL_K}{\epsilon\delta}\right)^3\right).$$

Furthermore, $\sum_{k\geq 0}\delta_k \leq \delta$. ■

Appendix B. Proofs for Section 4

B.1. Proof of Lemma 10

Proof Let t be the total number of steps executed and $\hat{\tau} = \frac{t}{\lambda}$.

If the round is a success, then it passes the failure check

$$\frac{\hat{\tau} + \epsilon L + \hat{p}}{1 - (\hat{p} + \epsilon)} \leq (1 + 8\epsilon)L$$

for all $n = \lambda$ episodes, which gives $\hat{\tau} \leq (1 + 7\epsilon)L$.

If the round is a failure or "skipped" after $n \leq \lambda$ episodes, it must pass the failure check in the first $n - 1$ episodes and therefore the total steps in the first $n - 1$ episodes is at most $(1 + 7\epsilon)L\lambda$. The last episode adds at most $\Gamma = \lceil (\frac{1}{\epsilon} + 1)L \rceil < \epsilon L\lambda$ steps for a total of less than $(1 + 8\epsilon)L\lambda$ steps. Since $\epsilon \in (0, 1/8]$ the proof is complete. ■

B.2. Proof of Lemma 12

The following Lemma is used in proving Lemma 12.

Lemma 17 *Let $s_1, s_2, \ldots$ be any sequence of distinct states. Suppose that for every s_i in the sequence, each action $a \in \mathcal{A}$ is executed $\lceil L\phi_i \rceil$ times, where $\phi_i = \log \frac{8ALi^2}{\delta}$. Let $\mathcal{S}'_{s_i,a}$ be the set of all next states visited during the $\lceil L\phi_i \rceil$ executions of (s_i, a). Then*

$$\Pr\left(\exists i, s, a : P(s|s_i, a) \geq \frac{1}{L} \wedge s \notin \mathcal{S}'_{s_i,a}\right) \leq \frac{\delta}{4}.$$

Proof The probability that a particular (s_i, a) fails to discover a particular s with $P(s|s_i, a) \geq \frac{1}{L}$ is at most

$$\left(1 - \frac{1}{L}\right)^{\lceil L\phi_i \rceil} \leq \left\{\left(1 - \frac{1}{L}\right)^{L}\right\}^{\phi_i} \leq \left(\frac{1}{e}\right)^{\phi_i} = \frac{\delta}{8ALi^2} \quad .$$

Note that for any (s_i, a), there can be at most L next states with probability at least $\frac{1}{L}$ (since the probabilities must sum up to 1). Taking the union bound over all s_i in the sequence, all actions $a \in \mathcal{A}$ and all valid next states s, the probability of any failure is at most

$$\sum_{i=1}^{\infty} \sum_{a} \sum_{\substack{s \in \mathcal{S}_L^{\rightarrow}, \\ P(s|s_i,a) \geq \frac{1}{L}}} \frac{\delta}{8ALi^2} \leq \sum_{i=1}^{\infty} \frac{\delta}{8i^2} = \frac{\delta}{8} \sum_{i=1}^{\infty} \frac{1}{i^2} \leq \frac{\delta}{4} \,.$$

■

B.3. Proof of Lemma 13

Proof We use i as the index for the sequence of successful rounds, each with an associated target state s_i and policy π_{s_i}.

In any given round where the chosen target is s_i, let $\hat{\tau}_j$ be the total number of of steps in the j-th episode of that round (before reaching either the target or $\Gamma = \left(\frac{1}{\epsilon} + 1\right) L$ steps). Let f_i be the

total number of failed episodes in that round. Recall that for the algorithm to output a policy π_{s_i}, its empirical performance after λ_i episodes must satisfy the following:

$$\frac{\hat{\tau} + \epsilon L + \hat{p} + \epsilon}{1 - (\hat{p} + \epsilon)} \leq (1 + 8\epsilon)L \tag{3}$$

where

$$\hat{\tau} = \frac{\sum_{j=1}^{\lambda_i} \hat{\tau}_j}{\lambda_i} \quad \text{and} \quad \hat{p} = \frac{f_i}{\lambda_i} \qquad .$$

This trivially implies that for a successful round it must be that $\hat{\tau} < (1 + 8\epsilon)L$.

Let T_Γ be the random variable denoting the total number of steps before reaching either the target s_i or Γ steps, when we run π_{s_i} for one episode. Clearly the range of T_Γ is between 0 and Γ. Note that $T_\Gamma < \Gamma$ implies a success while $T_\Gamma = \Gamma$ can mean either successfully reaching s_i after Γ steps or a failure episode. Let $\mathrm{E}(T_\Gamma)$ and $\mathrm{Var}(T_\Gamma)$ denote the expectation and variance of T_Γ respectively.

Let

$$\alpha = \frac{\epsilon^3\Gamma + \sqrt{(\epsilon^3\Gamma)^2 + 48\epsilon^3\mathrm{Var}(T_\Gamma)}}{12} \qquad .$$

By Bernstein's inequality, we have that

$$\Pr\left(\mathrm{E}(T_\Gamma) > \hat{\tau} + \alpha\right) < \exp\left(-\frac{\lambda_i\alpha^2}{2\mathrm{Var}(T_\Gamma) + \Gamma\alpha}\right) = \frac{\delta}{16i^2}$$

where we use α as defined above and

$$\lambda_i = \frac{6}{\epsilon^3}\log\frac{16i^2}{\delta} \qquad .$$

Therefore, with probability at least $1 - \frac{\delta}{16i^2}$, we have

$$\mathrm{E}(T_\Gamma) \leq \hat{\tau} + \alpha \qquad .$$

Now, note that

$$\mathrm{E}(T_\Gamma^2) = \sum_{t=0}^{\Gamma} \Pr(T_\Gamma = t)t^2 \leq \Gamma \sum_{t=0}^{\Gamma} \Pr(T_\Gamma = t)t = \Gamma\mathrm{E}(T_\Gamma)$$

and therefore

$$\mathrm{Var}(T_\Gamma) = \mathrm{E}(T_\Gamma^2) - [\mathrm{E}(T_\Gamma)]^2 \leq \Gamma\mathrm{E}(T_\Gamma) \qquad .$$

Using this bound for $\mathrm{Var}(T_\Gamma)$ and the fact that for a successful round $\hat{\tau} < (1 + 8\epsilon)L$, it is straightforward to show that $\mathrm{E}(T_\Gamma) \leq \hat{\tau} + \alpha$ implies $\mathrm{E}(T_\Gamma) \leq \hat{\tau} + \epsilon L$.

Let p be the true probability of failure to reach s_i within Γ steps. With Hoeffding's inequality it is straighforward to show that with probability at least $1 - \frac{\delta}{16i^2}$, $p \leq \hat{p} + \epsilon$.

We therefore have that when the algorithm outputs π_{s_i}, with probability at least $1 - \frac{\delta}{8i^2}$, $\mathrm{E}(T_\Gamma) \leq \hat{\tau} + \epsilon L$ and $p \leq \hat{p} + \epsilon$. Let T be the (random) number of steps to reach s_i with π_{s_i}. Note that π_{s_i} executes the RESET action whenever s_i is not reached after Γ steps and repeats the exact same policy until s_i is reached. We therefore have that for any $t \geq 0$,

$$\Pr(T = \Gamma + 1 + t) = p\Pr(T = t) \qquad .$$

The expected number of steps to reach s_i is therefore

$$\begin{aligned}
\tau(s_i|\pi_{s_i}) &= \sum_{t=0}^{\infty} \Pr(T=t)t \\
&= \left(\sum_{t=0}^{\Gamma} \Pr(T=t)t\right) + \sum_{t=\Gamma+1}^{\infty} \Pr(T=t)t \\
&= \left(\sum_{t=0}^{\Gamma} \Pr(T=t)t\right) + \sum_{t=0}^{\infty} \Pr(T=\Gamma+1+t)(\Gamma+1+t) \\
&= \left(\sum_{t=0}^{\Gamma} \Pr(T=t)t\right) + \sum_{t=0}^{\infty} p\Pr(T=t)(\Gamma+1+t) \\
&= \left(\sum_{t=0}^{\Gamma} \Pr(T=t)t\right) + p\Gamma + p\Big(1+\tau(s_i|\pi_{s_i})\Big) \\
&= \mathrm{E}(T_\Gamma) + p\Big(1+\tau(s_i|\pi_{s_i})\Big)
\end{aligned}$$

Rearranging, we have

$$\tau(s_i|\pi_{s_i}) = \frac{\mathrm{E}(T_\Gamma)+p}{1-p} \leq \frac{\hat{\tau}+\epsilon L+\hat{p}+\epsilon}{1-(\hat{p}+\epsilon)} \leq (1+8\epsilon)L \qquad . \tag{4}$$

Applying the union bound over $i=1,\ldots,n$, the total probability of failure is at most

$$\sum_{i=1}^{n} \frac{\delta}{8i^2} = \frac{\delta}{8}\sum_{i=1}^{n}\frac{1}{i^2} \leq \frac{\delta}{4} \qquad .$$

■

B.4. Proof of Lemma 14

Proof We first prove that for a fixed target $s^* \in \mathcal{U}$ equations (1) and (2) hold for all s,a,i with probability at least $1-\frac{\epsilon^4\delta}{|\mathcal{U}|k^5}$. Since there can be at most $|\mathcal{U}|$ possible targets, taking the union bound will complete the proof.

Fix the number of previous visits $N(s,a)$ for every state-action pair and fix a target $s^* \in \mathcal{U}$. This proof relies on the fact that for any particular state-action pair (s,a) and $i \in \{1,\ldots,\Gamma\}$, $\hat{P}_i(\cdot|s,a)$ is independent of $\tilde{u}_{i-1}(\cdot)$ since they use different, independent past transitions by construction.

The case for $i=1$ is trivially true. We will prove by induction that it is true for all $i>1$ up to Γ. Assume now that (1) and (2) hold for $i=1,\ldots,l$ for some $1 \leq l < \Gamma$ with probability $1-\frac{l\epsilon^4\delta}{|\mathcal{U}|\Gamma k^5}$, we need to show that it is true for $i=l+1$ with probability of failure at most $\frac{\epsilon^4\delta}{|\mathcal{U}|\Gamma k^5}$.

Fix a state-action pair (s,a). Recall that $\hat{P}_{l+1}(\cdot|s,a)$ is based on $n = \lfloor \frac{N(s,a)}{\Gamma} \rfloor$ independent past transitions from (s,a). Let $s'_1,\ldots,s'_n$ be the corresponding next states in these transitions. Let $\tilde{w}(\cdot) = \tilde{u}_l(\cdot) - (l-1-L)$. Let $\zeta_j = P(\cdot|s,a)\tilde{w}(\cdot) - \tilde{w}(s'_j)$.

Due to the RESET action and the fact that $u^*_\Gamma(s_0) \geq \Gamma - L$, we have that for any s'_j,

$$\tilde{u}_l(s'_j) \geq \tilde{q}_l(s'_j, \text{RESET}) = \tilde{u}_{l-1}(s_0) \geq l - 1 - L \qquad .$$

Thus

$$\begin{aligned} & (l-1-L) \leq \tilde{u}_l(s'_j) \leq l \\ \Rightarrow \quad & 0 \leq \tilde{w}(s'_j) \leq l - (l-1-L) = L+1 \end{aligned}$$

and therefore $|\zeta_j| \leq L+1$.

In other words, $\zeta_1, \ldots, \zeta_n$ are independent bounded random variables with expected value

$$\begin{aligned} \mathrm{E}(\zeta_j) &= \mathrm{E}[P(\cdot|s,a)\tilde{w}(\cdot) - \tilde{w}(s'_j)] \\ &= \mathrm{E}[P(\cdot|s,a)\tilde{w}(\cdot) - I_{s'_j}(\cdot)\tilde{w}(\cdot)] \\ &= \mathrm{E}[P(\cdot|s,a) - I_{s'_j}(\cdot)]\mathrm{E}[\tilde{w}(\cdot)] \qquad (5) \\ &= 0 \end{aligned}$$

where I is an indicator vector. Note that equation (5) holds due to the fact that $\tilde{w}(\cdot)$ is independent of $I_{s'_j}(\cdot)$. Let $\alpha = \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1,N(s,a)\}}}$. By Hoeffding's inequality,

$$\Pr\left(\left|\frac{\sum_{j=1}^n \zeta_j}{n}\right| > \alpha\right) < 2\exp\left(-\frac{2n\alpha^2}{(L+1)^2}\right) \leq \frac{\epsilon^4\delta}{A|\mathcal{U}||\mathcal{K}|\Gamma k^5}$$

where we make use of the fact that $n = \lfloor \frac{N(s,a)}{\Gamma} \rfloor \geq \frac{1}{2}\frac{\max\{1,N(s,a)\}}{\Gamma}$. It follows that with probability at least $1 - \frac{\epsilon^4\delta}{A|\mathcal{U}||\mathcal{K}|\Gamma k^5}$,

$$\left|\hat{P}_{l+1}(\cdot|s,a)\tilde{u}_l(\cdot) - P(\cdot|s,a)\tilde{u}_l(\cdot)\right| = \left|\frac{\sum_{j=1}^n \zeta_j}{n}\right| \leq \alpha = \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1,N(s,a)\}}} \qquad .$$

Taking the union bound over all states in $\mathcal{K}$ and all actions, we have that the probability of failure is at most $\frac{\epsilon^4\delta}{|\mathcal{U}|\Gamma k^5}$.

For (2), note that for any state-action pair (s,a),

$$\begin{aligned} \tilde{q}_{l+1}(s,a) &= r(s,a) + \hat{P}_{l+1}(\cdot|s,a)\tilde{u}_l(\cdot) + \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1,N(s,a)\}}} \\ &\geq r(s,a) + P(\cdot|s,a)\tilde{u}_l(\cdot) \\ &\geq r(s,a) + P(\cdot|s,a)u^*_l(\cdot) \\ &= q^*_{l+1}(s,a) \end{aligned}$$

and therefore

$$\tilde{u}_{l+1}(s) = \max_a \tilde{q}_{l+1}(s,a) \geq \max_a q^*_{l+1}(s,a) = u^*_{l+1}(s) \qquad .$$

■

B.5. Proof of Lemma 15

Proof Let $\hat{\tau}_j$ be the actual number of steps executed in episode j, for $j = 1, \ldots, n$ where $n \leq \lambda$ is the actual number of episodes executed in this round. Recall the failure check in the main algorithm (Section 3.3), which is based on the following empirical performance measure

$$\hat{\tau} = \frac{\sum_{j=1}^{n} \hat{\tau}_j}{\lambda} \quad \text{and} \quad \hat{p} = \frac{f}{\lambda}$$

where f is the number of failed episodes (where Γ steps have been executed without reaching the target state $\tilde{s}$).

Suppose $\hat{\tau} \leq (1+\epsilon)L$. Then $\frac{\hat{\tau}}{\epsilon} = \left(1 + \frac{1}{\epsilon}\right) L \leq \Gamma$. Since the total number of steps must be at least $f\Gamma$, we have

$$\hat{p}\lambda\Gamma = f\Gamma \leq \sum_{j=1}^{n} \hat{\tau}_j = \hat{\tau}\lambda$$

and therefore

$$\hat{p} \leq \frac{\hat{\tau}}{\Gamma} \leq \epsilon \qquad .$$

It is straightforward to verifty that this implies

$$\frac{\hat{\tau} + \epsilon L + \hat{p} + \epsilon}{1 - (\hat{p} + \epsilon)} \leq (1 + 8\epsilon)L$$

which means that this must be a successful round.

Therefore if this is a failure round, $\hat{\tau} > (1+\epsilon)L$. Let $\hat{u}_j = \Gamma - \hat{\tau}_j$ be the total rewards that would have been received in episode j if the optimistic policy is run in $\mathcal{M}_{\tilde{s}}$. For a failure round we therefore have

$$\sum_{j=1}^{n} \hat{u}_j = \sum_{j=1}^{n} \Gamma - \hat{\tau}_j = n\Gamma - \lambda\hat{\tau} < n\Gamma - \lambda(1+\epsilon)L.$$

The total regret in a failure round is therefore

$$\sum_{j=1}^{n} \Gamma - L - \hat{u}_j > n(\Gamma - L) - \Big(n\Gamma - \lambda(1+\epsilon)L\Big) \geq \lambda\epsilon L.$$

■

B.6. Proof of Lemma 16

Lemma 16 makes use of the following two lemmas:

Lemma 18 *Let*

$$X_i^{j,k} = P(\cdot|s_i^{j,k}, a_i^{j,k})\tilde{u}_{\Gamma-i-1}(\cdot) - \tilde{u}_{\Gamma-i-1}(s_{i+1}^{j,k})$$

where $s_i^{j,k}$ denotes the actual state visited at step i in episode j of round k and $a_i^{j,k}$ denotes the actual action taken at step i in episode j of round k (according to the policy in round k). Let n_k be

the actual number of episodes executed in round k and $\hat{\tau}^{j,k}$ be the actual number of steps executed in episode j of round k. Then

$$\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sum_{i=0}^{\hat{\tau}^{j,k}-1} X_i^{j,k} \leq 2(L+1)\sqrt{L\lambda m \log\frac{4m^2}{\delta}}$$

with probability at least $1-\frac{\delta}{4m^2}$.

Proof For any state s and any $l \in \{1,\ldots,\Gamma\}$, let $\tilde{w}_l(s) = \tilde{u}_l(s) - (l-1-L)$. It is easy to see that

$$X_i^{j,k} = P(\cdot|s_i^{j,k}, a_i^{j,k})\tilde{w}_{\Gamma-i-1}(\cdot) - \tilde{w}_{\Gamma-i-1}(s_{i+1}^{j,k}).$$

Similar to the reasoning in the proof for Lemma 14, we have that $|X_i^{j,k}| \leq L+1$.

Note that $\mathbb{E}[X_i^{j,k}|s_0^{1,1}, a_0^{1,1}, \ldots, s_i^{j,k}, a_i^{j,k}] = 0$ and therefore $X_i^{j,k}$ is a martingale difference sequence where $|X_i^{j,k}| \leq (L+1)$ for all k, j and i. Let $T = \sum_{k=1}^{m}\sum_{j=1}^{n_k}\hat{\tau}^{j,k}$ be the total number of steps, and $\alpha = 2(L+1)\sqrt{L\lambda m \log\frac{4m^2}{\delta}}$. By the Azuma-Hoeffding inequality,

$$\begin{aligned}
\Pr\left(\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sum_{i=0}^{\Gamma_{k,j}-1} X_{k,j}^{i} \geq \alpha\right) &\leq \exp\left(-\frac{\alpha^2}{2T(L+1)^2}\right) \\
&= \exp\left(-\frac{2L\lambda m}{T}\log\frac{4m^2}{\delta}\right) \\
&\leq \frac{\delta}{4m^2}
\end{aligned}$$

where we use the fact that $T \leq 2L\lambda m$ from Lemma 10. ■

Lemma 19 *Let $v_k(s,a)$ be the actual number of times state-action pair (s,a) is executed in round k and $N_k(s,a)$ be the total number of times (s,a) is executed before round k. After m failure rounds,*

$$\sum_{s\in\mathcal{K},a\in\mathcal{A}}\sum_{k=1}^{m}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} \leq (\sqrt{2}+1)\sqrt{2|\mathcal{K}|AL\lambda m}$$

Proof For a fixed $s \in \mathcal{K}$ and $a \in \mathcal{A}$. Following the idea from Lemma 19 in Jaksch et al. (2010), we prove, by induction, that the following statement holds for all $m \geq 1$

$$\sum_{k=1}^{m}\frac{v_k(s,a)}{\sqrt{V_k(s,a)}} \leq (\sqrt{2}+1)\sqrt{\sum_{k=1}^{m} v_k(s,a)}$$

where $V_k(s,a) = \max\{1, N_k(s,a)\} \geq \sum_{i=1}^{k-1} v_i(s,a)$.

We make use of the fact that $v_k(s,a) < V_k(s,a)$ for all k since otherwise it would result in a "skipped" round. This immediately implies that the statement holds for $m = 1$. Now, assume that the statement holds for some $n \geq 1$. Then

$$\sum_{k=1}^{n+1}\frac{v_k(s,a)}{\sqrt{V_k(s,a)}}$$

$$
\begin{aligned}
&\leq (\sqrt{2}+1)\sqrt{\sum_{k=1}^{n} v_k(s,a)} + \frac{v_{n+1}(s,a)}{\sqrt{V_{n+1}(s,a)}} \\
&= \sqrt{(\sqrt{2}+1)^2\left(\sum_{k=1}^{n} v_k(s,a)\right) + \frac{v_{n+1}(s,a)^2}{V_{n+1}(s,a)} + 2(\sqrt{2}+1)\sqrt{\sum_{k=1}^{n} v_k(s,a)}\frac{v_{n+1}(s,a)}{\sqrt{V_{n+1}(s,a)}}} \\
&\leq \sqrt{(\sqrt{2}+1)^2\left(\sum_{k=1}^{n} v_k(s,a)\right) + v_{n+1}(s,a) + 2(\sqrt{2}+1)v_{n+1}(s,a)} \\
&= (\sqrt{2}+1)\sqrt{\sum_{k=1}^{n+1} v_k(s,a)}
\end{aligned}
$$

and therefore it also holds for $n+1$.

Summing up over all (s,a), we have

$$
\begin{aligned}
\sum_{s,a}\sum_{k=1}^{m} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} &\leq (\sqrt{2}+1)\sum_{s,a}\sqrt{\sum_{k=1}^{m} v_k(s,a)} \\
&\leq (\sqrt{2}+1)\sqrt{2|\mathcal{K}|AL\lambda m}
\end{aligned}
$$

where we apply Jensen's inequality in the last inequality using the fact from Lemma 10 that

$$
\sum_{s,a}\sum_{k=1}^{m} v_k(s,a) \leq 2|\mathcal{K}|AL\lambda m.
$$

■

Proof (Lemma 16)

We first consider the total regret in the j-th episode of round k. Let $\tilde{s}^k$ be the chosen optimistic target. Let $\hat{\tau}^{j,k}$ be the actual number of steps executed in this episode and $s_0^{j,k}, s_1^{j,k}, \ldots, s_{\hat{\tau}_j}^{j,k}$ be the actual states visited. If $\hat{\tau}^{j,k} < \Gamma$ then it must be the case that $s_{\hat{\tau}_j}^{j,k} = \tilde{s}^k$ (reaching $\tilde{s}^k$ is the only way to stop before Γ steps). Note that we simulate running the optimistic policy in $\mathcal{M}_{\tilde{s}^k}$ where the state $\tilde{s}^k$ is absorbing with reward 1. Let $r_0, r_1, \ldots, r_{\Gamma-1}$ be the rewards that would have been collected in this episode. It follows that $r_0 = r_1 = \ldots = r_{\hat{\tau}^{j,k}-1} = 0$ and $r_{\hat{\tau}^{j,k}} = \ldots = r_{\Gamma-1} = 1$. Also $\sum_{i=\hat{\tau}^{j,k}}^{\Gamma-1} r_i = \Gamma - \hat{\tau}^{j,k} = \tilde{u}_{\Gamma-\hat{\tau}^{j,k}}(\tilde{s}^k)$. The regret is then given by

$$
\begin{aligned}
\Delta^{j,k} &= \Gamma - L - \sum_{i=0}^{\Gamma-1} r_i \\
&\leq \tilde{u}_\Gamma(s_0^{j,k}) - \sum_{i=0}^{\Gamma-1} r_i \\
&= \tilde{u}_\Gamma(s_0^{j,k}) - \tilde{u}_{\Gamma-\tau^{\hat{j},k}}(s_{\hat{\tau}^{j,k}}^{j,k}) \\
&= \hat{P}_\Gamma[\cdot|s_0^{j,k}, \tilde{\pi}_0(s_0^{j,k})]\tilde{u}_{\Gamma-1}(\cdot) + \frac{\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N_k[s_0^{j,k}, \tilde{\pi}_0(s_0^{j,k})]\}}} - \tilde{u}_{\Gamma-\tau^{\hat{j},k}}(s_{\hat{\tau}^{j,k}}^{j,k})
\end{aligned}
$$

$$
\begin{aligned}
&\leq P[\cdot|s_0^{j,k}, \tilde{\pi}_0(s_0^{j,k})]\tilde{u}_{\Gamma-1}(\cdot) + \frac{2\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N_k[s_0^{j,k}, \tilde{\pi}_0(s_0^{j,k})]\}}} \\
&\quad -\tilde{u}_{\Gamma-\tau^{\hat{j},k}}(s_{\hat{\tau}^{j,k}}^{j,k}) \qquad (6) \\
&= X_0^{j,k} + \tilde{u}_{\Gamma-1}(s_1^{j,k}) + \frac{2\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N_k[s_0^{j,k}, \tilde{\pi}_0(s_0^{j,k})]\}}} - \tilde{u}_{\Gamma-\tau^{\hat{j},k}}(s_{\hat{\tau}^{j,k}}^{j,k}) \qquad (7) \\
&\leq \sum_{i=0}^{\hat{\tau}^{j,k}-1} \left(X_i^{j,k} + I[s_i^{j,k} \in \mathcal{K}] \frac{2\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N_k[s_i^{j,k}, \tilde{\pi}_i(s_i^{j,k})]\}}} \right). \qquad (8)
\end{aligned}
$$

Equation 6 is due to Lemma 14 (we will consider the failure probability later). In equation 7, we use the definition for $X_i^{j,k}$ as in Lemma 18. In equation 8, we recursively apply the same set of arguments to $\tilde{u}_{\Gamma-i}(s_i^{j,k})$ for $i = 1, 2, \ldots, \hat{\tau}^{j,k} - 1$, at which point the last term gets canceled. The extra term $I[s_i^{j,k} \in \mathcal{K}]$ is an indicator function that is 1 if $s_i^{j,k} \in \mathcal{K}$ and 0 otherwise. The reason for this is that $\tilde{\pi}_i(s_i^{j,k}) = \text{RESET}$ for all states outside $\mathcal{K}$ and its transition is always to s_0 with probability 1 (see also Fig. 2).

Let n_k be the actual number of episodes run in round k. The total regret in round k is obtained by adding $\Delta^{j,k}$ for $j = 1, \ldots, n_k$

$$
\begin{aligned}
\Delta^k &= \sum_{j=1}^{n_k} \Delta^{j,k} \\
&\leq \sum_{j=1}^{n_k} \sum_{i=0}^{\hat{\tau}^{j,k}-1} \left(X_i^{j,k} + I[s_i^{j,k} \in \mathcal{K}^k] \frac{2\sigma_k(L+1)\sqrt{\Gamma}}{\sqrt{\max\{1, N_k[s_i^{j,k}, \tilde{\pi}_i(s_i^{j,k})]\}}} \right) \\
&= \left\{ \sum_{j=1}^{n_k} \sum_{i=0}^{\hat{\tau}^{j,k}-1} X_i^{j,k} \right\} + \left\{ 2\sigma_k(L+1)\sqrt{\Gamma} \sum_{s\in\mathcal{K}, a\in\mathcal{A}} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \right\}
\end{aligned}
$$

where in the last equation we regroup the steps into number of visits for each state-action pair.

Finally, the total regret after m failure rounds is given by

$$
\begin{aligned}
\Delta &= \sum_{k=1}^{m} \Delta^k \\
&\leq 2L\lambda\sqrt{\epsilon^2 m} + \sum_{k=\epsilon\sqrt{m}+1}^{m} \Delta^k \\
&\leq 2L\lambda\epsilon\sqrt{m} + \left\{ \sum_{k=\epsilon\sqrt{m}+1}^{m} \sum_{j=1}^{n_k} \sum_{i=0}^{\hat{\tau}^{j,k}-1} X_i^{j,k} \right\} + \\
&\quad \left\{ 2\sigma_m(L+1)\sqrt{\Gamma} \sum_{s,a} \sum_{k=\epsilon\sqrt{m}+1}^{m} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \right\} \\
&\leq 2L\lambda\epsilon\sqrt{m} + 2(L+1)\sqrt{L\lambda m \log\frac{4m^2}{\delta}} +
\end{aligned}
$$

$$
\begin{aligned}
& (2\sigma_m(L+1)\sqrt{\Gamma})(\sqrt{2}+1)\sqrt{2SAL\lambda m} \qquad (9)\\
\leq\ & m\left(173\frac{L^2\sqrt{SA}}{\epsilon^2\sqrt{m}}\log\frac{m}{\epsilon\delta}\right)
\end{aligned}
$$

where we apply Lemma 18 and Lemma 19 in equation 9. In the last inequality we simplify the terms by using the fact that $m \geq SAL$, $\epsilon \leq \frac{1}{8}$, $\lambda \leq \left\lceil \frac{6}{\epsilon^3}\log\frac{16S^2}{\delta}\right\rceil \leq \frac{12}{\epsilon^3}\log\frac{m}{\epsilon\delta}$ and $\sigma_m \leq \sqrt{\log\frac{2A|\mathcal{U}|S\Gamma m^5}{\epsilon^4\delta}} \leq 3\sqrt{\log\frac{m}{\epsilon\delta}}$.

We now bound the probability that some of the inequalities may fail.

By Lemma 14, the probability that inequality (6) fails for a particular k is bounded by $\frac{\epsilon^4\delta}{k^5}$. Taking the union bound over all $k = (\epsilon\sqrt{m}+1), \ldots, m$, the total failure probability is at most

$$
\sum_{k=\epsilon\sqrt{m}+1}^{m}\frac{\epsilon^4\delta}{k^5} \leq \int_{\epsilon\sqrt{m}}^{m}\frac{\epsilon^4\delta}{x^5}dx \leq \frac{\delta}{4m^2} \qquad .
$$

The inequality (9) fails with probability at most $\frac{\delta}{4m^2}$ by Lemma 18.

For any $m \geq SAL$, the whole inequality therefore holds with probability at least $1-\frac{\delta}{2m^2}$. Since $\sum_{m=2}^{\infty}\frac{1}{m^2} < 1$, taking the union bound over all $m \geq SAL$, the total failure probability is at most $\delta/2$.

Finally, $|\mathcal{K}| \leq S$ if all the known states are indeed reachable in $(1+8\epsilon)L$ steps. By Lemma 13 the probability of failure is at most $\delta/4$. Adding up all the failure probabilities completes the proof. ■