Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes

Ronan Fruit
Sequel Team - Inria Lille
ronan.fruit@inria.fr

Matteo Pirotta
Sequel Team - Inria Lille
matteo.pirotta@inria.fr

Alessandro Lazaric
Facebook AI Research
lazaric@fb.com

Abstract

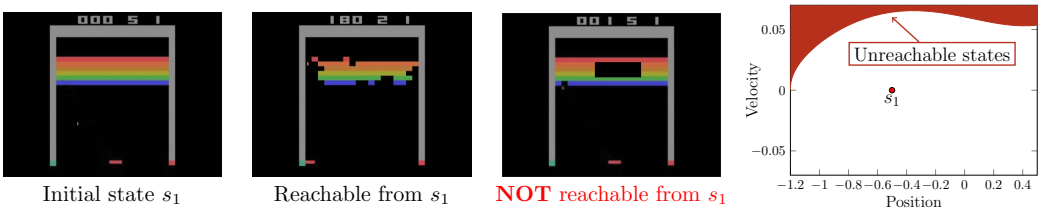
While designing the state space of an MDP, it is common to include states that are transient or not reachable by any policy (e.g., in mountain car, the product space of speed and position contains configurations that are not physically reachable). This results in weakly-communicating or multi-chain MDPs. In this paper, we introduce TUCRL, the first algorithm able to perform efficient exploration-exploitation in any finite Markov Decision Process (MDP) without requiring any form of prior knowledge. In particular, for any MDP with S^{C} communicating states, A actions and $\Gamma^{\mathsf{C}} \leq S^{\mathsf{C}}$ possible communicating next states, we derive a $\widetilde{O}(D^{\mathsf{C}}\sqrt{\Gamma^{\mathsf{C}} S^{\mathsf{C}} AT})$ regret bound, where D^{C} is the diameter (i.e., the length of the longest shortest path between any two states) of the communicating part of the MDP. This is in contrast with existing optimistic algorithms (e.g., UCRL, Optimistic PSRL) that suffer linear regret in weakly-communicating MDPs, as well as posterior sampling or regularised algorithms (e.g., REGAL), which require prior knowledge on the bias span of the optimal policy to achieve sub-linear regret. We also prove that in weakly-communicating MDPs, no algorithm can ever achieve a logarithmic growth of the regret without first suffering a linear regret for a number of steps that is exponential in the parameters of the MDP. Finally, we report numerical simulations supporting our theoretical findings and showing how TUCRL overcomes the limitations of the state-of-the-art.

1 Introduction

Reinforcement learning (RL) [1] studies the problem of learning in sequential decision-making problems where the dynamics of the environment is unknown, but can be learnt by performing actions and observing their outcome in an online fashion. A sample-efficient RL agent must trade off the *exploration* needed to collect information about the environment, and the *exploitation* of the experience gathered so far to gain as much reward as possible. In this paper, we focus on the regret framework in *infinite-horizon average-reward* problems [2], where the exploration-exploitation performance is evaluated by comparing the rewards accumulated by the learning agent and an optimal policy. Jaksch et al. [2] showed that it is possible to efficiently solve the exploration-exploitation dilemma using the *optimism in face of uncertainty* (OFU) principle. OFU methods build confidence intervals on the dynamics and reward (i.e., construct a set of plausible MDPs), and execute the optimal policy of the "best" MDP in the confidence region [e.g., 2, 3, 4, 5, 6]. An alternative approach is posterior sampling (PS) [7], which maintains a posterior distribution over MDPs and, at each step, samples an MDP and executes the corresponding optimal policy [e.g., 8, 9, 10, 11, 12].

Weakly-communicating MDPs and misspecified states. One of the main limitations of UCRL [2] and optimistic PSRL [12] is that they require the MDP to be communicating so that its diameter D (i.e., the length of the longest path among all shortest paths between any pair of states) is finite. While assuming that all states are reachable may seem a reasonable assumption, it is rarely verified *in practice*. In fact, it requires a designer to carefully define a state space $\mathcal{S}$ that contains all reachable

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) Breakout (b) Mountain Car

Figure 1: Examples of non-communicating domains. Fig. b represents a phase plane plot of the Mountain car domain $(x, \dot{x}) \in [-1.2, 0.6] \times [-0.07, 0.07]$. The initial state is $(-0.5, 0)$ and the red area corresponds to non-reachable states from the initial state. Other non-reachable states may exist. Fig. a shows the initial state, one reachable state (*middle*) and an unreachable one (*right*).

states (otherwise it may not be possible to learn the optimal policy), but it excludes unreachable states (otherwise the resulting MDP would be non-communicating). This requires a considerable amount of prior knowledge about the environment. Consider a problem where we learn from images e.g., the Atari Breakout game [13]. The state space is the set of "plausible" configurations of the brick wall, ball and paddle positions. The situation in which the wall has an hole in the middle is a valid state (e.g., as an initial state) but it cannot be observed/reached starting from a dense wall (see Fig. 1a). As such, it should be removed to obtain a "well-designed" state space. While it may be possible to design a suitable set of "reachable" states that define a communicating MDP, this is often a difficult and tedious task, sometimes even impossible. Now consider a continuous domain e.g., the Mountain Car problem [14]. The state is decribed by the position x and velocity $\dot{x}$ along the x-axis. The state space of this domain is usually defined as the cartesian product $[-1.2, 0.6] \times [-0.07, 0.07]$. Unfortunately, this set contains configurations that are not physically reachable as shown on Fig. 1b. The *dynamics* of the system is constrained by the *evolution equations*. Therefore, the car can not go arbitrarily fast. On the leftmost position ($x = -1.2$) the speed $\dot{x}$ cannot exceed 0 due to the fact that such position can be reached only with velocity $\dot{x} \leq 0$. To have a higher velocity, the car would need to acquire momentum from further left (i.e., $x < -1.2$) which is impossible by design (-1.2 is the left-boundary of the position domain). The maximal speed reachable for $x > -1.2$ can be attained by applying the maximum acceleration at any time step starting from the state $(x, \dot{x}) = (-1.2, 0)$. This identifies the curve reported in the Fig. 1b which denotes the boundary of the unreachable region. Note that other states may not be reachable. Whenever the state space is *misspecified* or the MDP is weakly communicating (i.e., $D = +\infty$), OFU-based algorithms (e.g.,UCRL) optimistically attribute large reward and non-zero probability to reach states that have never been observed, and thus they tend to repeatedly attempt to *explore* unreachable states. This results in poor performance and linear regret. A first attempt to overcome this major limitation is REGAL.C [3] (Fruit et al. [6] recently proposed SCAL, an implementable efficient version of REGAL.C), which requires prior knowledge of an upper-bound H to the span (i.e., range) of the optimal bias function h^*. The optimism of UCRL is then "constrained" to policies whose bias has span smaller than H. This implicitly "removes" non-reachable states, whose large optimistic reward would cause the span to become too large. Unfortunately, an accurate knowledge of the bias span may not be easier to obtain than designing a well-specified state space. Bartlett and Tewari [3] proposed an alternative algorithm – REGAL.D– that leverages on the *doubling trick* [15] to avoid any prior knowledge on the span. Nonetheless, we recently noticed a major flaw in the proof of [3, Theorem 3] that questions the validity of the algorithm (see App. A for further details). PS-based algorithms also suffer from similar issues.[1] To the best of our knowledge, the only regret guarantees available in the literature for this setting are [17, 18, 19]. However, the counter-example of Osband and Roy [20] seems to invalidate the result of Abbasi-Yadkori and Szepesvári [17]. On the other hand, Ouyang et al. [18] and Theocharous et al. [19] present PS algorithms with expected *Bayesian* regret scaling linearly with H, where H is an upper-bound on the optimal bias spans of all the MDPs that can be drawn from the prior distribution ([18, Asm. 1] and [19, Sec. 5]). In [18, Remark 1], the authors claim that their algorithm does not require the knowledge of H to derive the regret bound. However, in App. B we show on a very simple

[1] We notice that the problem of weakly-communicating MDPs and misspecified states does not hold in the more restrictive setting of finite horizon [e.g., 8] since exploration is directly tailored to the states that are reachable *within* the known horizon, or under the assumption of the existence of a recurrent state [e.g., 16].

example that for most continuous prior distributions (e.g., uninformative priors like Dirichlet), it is very likely that $H = +\infty$ implying that the regret bound may not hold (similarly for [19]). As a result, similarly to REGAL.C, the prior distribution should contain prior knowledge on the bias span to avoid poor performance.

In this paper, we present TUCRL, an algorithm designed to trade-off exploration and exploitation in weakly-communicating and multi-chain MDPs (e.g., MDPs with misspecified states) without any prior knowledge and under the only assumption that the agent starts from a state in a communicating subset of the MDP (Sec. 3). In communicating MDPs, TUCRL eventually (after a finite number of steps) performs as UCRL, thus achieving problem-dependent logarithmic regret. When the true MDP is weakly-communicating, we prove that TUCRL achieves a $\widetilde{O}(\sqrt{T})$ regret that with polynomial dependency on the MDP parameters. We also show that it is not possible to design an algorithm achieving logarithmic regret in weakly-communicating MDPs without having an exponential dependence on the MDP parameters (see Sec. 5). TUCRL is the first computationally tractable algorithm in the OFU literature that is able to adapt to the MDP nature without any prior knowledge. The theoretical findings are supported by experiments on several domains (see Sec. 4).

2 Preliminaries

We consider a finite *weakly-communicating* Markov decision process [21, Sec. 8.3] $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle$ with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$. Each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}_s$ is characterized by a reward distribution with mean $r(s, a)$ and support in $[0, r_{\max}]$ as well as a transition probability distribution $p(\cdot|s, a)$ over next states. In a weakly-communicating MDP, the state-space $\mathcal{S}$ can be *partioned* into two subspaces [21, Section 8.3.1]: a *communicating* set of states (denoted $\mathcal{S}^{\mathtt{C}}$ in the rest of the paper) with each state in $S^{\mathtt{C}}$ accessible –with non-zero probability– from any other state in $S^{\mathtt{C}}$ under some stationary deterministic policy, and a –possibly empty– set of states that are *transient* under all policies (denoted $\mathcal{S}^{\mathtt{T}}$). We also denote by $S = |\mathcal{S}|$, $S^{\mathtt{C}} = |\mathcal{S}^{\mathtt{C}}|$ and $A = \max_{s \in \mathcal{S}} |\mathcal{A}_s|$ the number of states and actions, and by $\Gamma^{\mathtt{C}} = \max_{s \in \mathcal{S}^{\mathtt{C}}, a \in \mathcal{A}} \|p(\cdot|s, a)\|_0$ the maximum support of all transition probabilities $p(\cdot|s, a)$ with $s \in \mathcal{S}^{\mathtt{C}}$. The sets $\mathcal{S}^{\mathtt{C}}$ and $\mathcal{S}^{\mathtt{T}}$ form a partition of $\mathcal{S}$ i.e., $\mathcal{S}^{\mathtt{C}} \cap \mathcal{S}^{\mathtt{T}} = \emptyset$ and $\mathcal{S}^{\mathtt{C}} \cup \mathcal{S}^{\mathtt{T}} = \mathcal{S}$. A deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ maps states to actions and it has an associated *long-term average reward* (or *gain*) and a *bias function* defined as

$$g_M^\pi(s) := \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} r\big(s_t, \pi(s_t)\big)\right]; \quad h_M^\pi(s) := \underset{T \to \infty}{C\text{-}\lim}\, \mathbb{E}\left[\sum_{t=1}^{T} \big(r(s_t, \pi(s_t)) - g_M^\pi(s_t)\big)\right],$$

where the bias $h_M^\pi(s)$ measures the expected total difference between the rewards accumulated by π starting from s and the stationary reward in *Cesaro-limit*[2] (denoted $C\text{-}\lim$). Accordingly, the difference of bias values $h_M^\pi(s) - h_M^\pi(s')$ quantifies the (dis-)advantage of starting in state s rather than s'. In the following, we drop the dependency on M whenever clear from the context and denote by $sp_{\mathcal{S}}\{h^\pi\} := \max_{s \in \mathcal{S}} h^\pi(s) - \min_{s \in \mathcal{S}} h^\pi(s)$ the *span* of the bias function. In weakly communicating MDPs, any optimal policy $\pi^* \in \arg\max_\pi g^\pi(s)$ has *constant* gain, i.e., $g^{\pi^*}(s) = g^*$ for all $s \in \mathcal{S}$. Finally, we denote by D, resp. $D^{\mathtt{C}}$, the diameter of M, resp. the diameter of the communicating part of M (i.e., restricted to the set $\mathcal{S}^{\mathtt{C}}$):

$$D := \max_{(s,s') \in \mathcal{S} \times \mathcal{S}, s \neq s'} \{\tau_M(s \to s')\}, \qquad D^{\mathtt{C}} := \max_{(s,s') \in \mathcal{S}^{\mathtt{C}} \times \mathcal{S}^{\mathtt{C}}, s \neq s'} \{\tau_M(s \to s')\}, \tag{1}$$

where $\tau_M(s \to s')$ is the expected time of the shortest path from s to s' in M.

Learning problem. Let M^* be the true (*unknown*) weakly-communicating MDP. We consider the learning problem where $\mathcal{S}$, $\mathcal{A}$ and $r_{\max}$ are *known*, while sets $\mathcal{S}^{\mathtt{C}}$ and $\mathcal{S}^{\mathtt{T}}$, rewards r and transition probabilities p are *unknown* and need to be estimated on-line. We evaluate the performance of a learning algorithm $\mathfrak{A}$ after T time steps by its cumulative *regret* $\Delta(\mathfrak{A}, T) = Tg^* - \sum_{t=1}^{T} r_t(s_t, a_t)$. Furthermore, we state the following assumption.

Assumption 1. *The initial state s_1 belongs to the communicating set of states $\mathcal{S}^{\mathtt{C}}$.*

While this assumption somehow restricts the scenario we consider, it is fairly common in practice. For example, all the domains that are characterized by the presence of a resetting distribution (e.g., episodic problems) satisfy this assumption (e.g., mountain car, cart pole, Atari games, taxi, etc.).

[2] For policies whose associated Markov chain is aperiodic, the standard limit exists.

Multi-chain MDPs. While we consider weakly-communicating MDPs for ease of notation, all our results extend to the more general case of multi-chain MDPs.[3] In this case, there may be multiple communicating and transient sets of states and the optimal gain g^* is different in each communicating subset. In this case we define S^{C} as the set of states that are accessible –with non-zero probability– from s_1 (s_1 included) under some stationary deterministic policy. $\mathcal{S}^{\mathsf{T}}$ is defined as the complement of S^{C} in $\mathcal{S}$ i.e., $\mathcal{S}^{\mathsf{T}} := \mathcal{S} \setminus S^{\mathsf{C}}$. With these new definitions of S^{C} and $\mathcal{S}^{\mathsf{T}}$, Asm. 1 needs to be reformulated as follows:

Assumption 1 for Multi-chain MDPs. *The initial state s_1 is accessible –with non-zero probability– from any other state in $\mathcal{S}^{\mathcal{C}}$ under some stationary deterministic policy. Equivalently, $\mathcal{S}^{\mathcal{C}}$ is a communicating set of states.*

Note that the states belonging to $\mathcal{S}^{\mathsf{T}}$ can either be transient or belong to other communicating subsets of the MDP disjoint from $\mathcal{S}^{\mathsf{C}}$. It does not really matter because the states in $\mathcal{S}^{\mathsf{T}}$ will never be visited by definition. As a result, the regret is still defined as before, where the learning performance is compared to the optimal gain $g^*(s_1)$ related to the communicating set of states $\mathcal{S}^{\mathsf{C}} \ni s_1$.

3 Truncated Upper-Confidence for Reinforcement Learning (TUCRL)

In this section we introduce Truncated Upper-Confidence for Reinforcement Learning (TUCRL), an optimistic online RL algorithm that efficiently balances exploration and exploitation to learn in non-communicating MDPs without prior knowledge (Fig. 2).

Similar to UCRL, at the beginning of each episode k, TUCRL constructs confidence intervals for the reward and the dynamics of the MDP. Formally, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ we define

$$B_{p,k}(s,a) = \left\{ \widetilde{p}(\cdot|s,a) \in \mathcal{C} : \ \forall s' \in \mathcal{S}, |\widetilde{p}(s'|s,a) - \widehat{p}(s'|s,a)| \le \beta_{p,k}^{sas'} \right\}, \tag{2}$$

$$B_{r,k}(s,a) := [\widehat{r}_k(s,a) - \beta_{r,k}^{sa}, \widehat{r}_k(s,a) + \beta_{r,k}^{sa}] \cap [0, r_{\max}], \tag{3}$$

where $\mathcal{C} = \{p \in \mathbb{R}^S | \forall s', \ p(s') \ge 0 \wedge \sum_{s'} p(s') = 1\}$ is the $(S-1)$-probability simplex, while the size of the confidence intervals is constructed using the empirical Bernstein's inequality [22, 23] as

$$\beta_{r,k}^{sa} := \sqrt{\frac{14\widehat{\sigma}_{r,k}^2(s,a) b_{k,\delta}}{N_k^+(s,a)}} + \frac{\frac{49}{3} r_{\max} b_{k,\delta}}{N_k^{\pm}(s,a)}, \qquad \beta_{p,k}^{sas'} := \sqrt{\frac{14\widehat{\sigma}_{p,k}^2(s'|s,a) b_{k,\delta}}{N_k^+(s,a)}} + \frac{\frac{49}{3} b_{k,\delta}}{N_k^{\pm}(s,a)},$$

where $N_k(s,a)$ is the number of visits in (s,a) before episode k, $N_k^+(s,a) := \max\{1, N_k(s,a)\}$, $N_k^{\pm}(s,a) := \max\{1, N_k(s,a) - 1\}$, $\widehat{\sigma}_{r,k}^2(s,a)$ and $\widehat{\sigma}_{p,k}^2(s'|s,a)$ are the empirical variances of $r(s,a)$ and $p(s'|s,a)$ and $b_{k,\delta} = \ln(2SAt_k/\delta)$. The set of plausible MDPs associated with the confidence intervals is then $\mathcal{M}_k = \left\{ M = (\mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p}) : \ \widetilde{r}(s,a) \in B_{r,k}(s,a), \ \widetilde{p}(\cdot|s,a) \in B_{p,k}(s,a) \right\}$. UCRL is optimistic w.r.t. the confidence intervals so that for all states s that have never been visited the optimistic reward $\widetilde{r}(s,a)$ is set to $r_{\max}$, while all transitions to s (i.e., $\widetilde{p}(s|\cdot,\cdot)$) are set to the largest value compatible with $B_{p,k}(\cdot,\cdot)$. Unfortunately, some of the states with $N_k(s,a) = 0$ may be actually unreachable (i.e., $s \in \mathcal{S}^{\mathsf{T}}$) and UCRL would uniformly explore the policy space with the hope that at least one policy reaches those (optimistically desirable) states. TUCRL addresses this issue by first constructing empirical estimates of $\mathcal{S}^{\mathsf{C}}$ and $\mathcal{S}^{\mathsf{T}}$ (i.e., the set of communicating and transient states in M^*) using the states that have been visited so far, that is $\mathcal{S}_k^{\mathsf{C}} := \left\{ s \in \mathcal{S} \ \middle| \ \sum_{a \in \mathcal{A}_s} N_k(s,a) > 0 \right\} \cup \{s_{t_k}\}$ and $\mathcal{S}_k^{\mathsf{T}} := \mathcal{S} \setminus \mathcal{S}_k^{\mathsf{C}}$, where t_k is the starting time of episode k.

In order to avoid optimistic exploration attempts to unreachable states, we could simply execute UCRL on $\mathcal{S}_k^{\mathsf{C}}$, which is guaranteed to contain only states in the communicating set (since $s_1 \in \mathcal{S}^{\mathsf{C}}$ by Asm. 1, we have that $\mathcal{S}_k^{\mathsf{C}} \subseteq \mathcal{S}^{\mathsf{C}}$). Nonetheless, this algorithm could *under-explore* state-action pairs that would allow discovering other states in $\mathcal{S}^{\mathsf{C}}$, thus getting stuck in a subset of the communicating states of the MDP and suffering linear regret. While the states in $\mathcal{S}_k^{\mathsf{C}}$ are guaranteed to be in the communicating subset, it is not possible to know whether states in $\mathcal{S}_k^{\mathsf{T}}$ are actually reachable from $\mathcal{S}_k^{\mathsf{C}}$ or not. Then TUCRL first "guesses" a lower bound on the probability of transition from states $s \in \mathcal{S}_k^{\mathsf{C}}$ to $s' \in \mathcal{S}_k^{\mathsf{T}}$ and whenever the maximum transition probability from s to s' compatible with the confidence intervals (i.e., $\widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'}$) is below the lower bound, it assumes that such transition

[3] In the case of misspecified states, we implicitly define a multi-chain MDP, where each non-reachable state has a self-loop dynamics and it defines a "singleton" communicating subset.

Input: Confidence $\delta \in]0,1[$, $r_{\max}$, $\mathcal{S}$, $\mathcal{A}$
Initialization: Set $N_0(s,a) := 0$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $t := 1$ and observe s_1.
For episodes $k = 1, 2, ...$ **do**
1. Set $t_k = t$ and episode counters $\nu_k(s,a) = 0$
2. Compute estimates $\widehat{p}_k(s'|s,a)$, $\widehat{r}_k(s,a)$ and a set $\overline{\mathcal{M}}_k$
3. Compute an $r_{\max}/\sqrt{t_k}$-approximation $\widetilde{\pi}_k$ of Eq. 5
4. **While** $t_k == t$ **or** $\left(\sum_{a \in \mathcal{A}_{s_t}} N_k(s_t,a) > 0 \right.$ **and** $\left. \nu_k(s_t, \widetilde{\pi}_k(s_t)) \leq \max\{1, N_k(s_t, \widetilde{\pi}_k(s_t))\} \right)$ **do**
   (a) Execute $a_t = \widetilde{\pi}_k(s_t)$, obtain reward r_t, and observe s_{t+1}
   (b) Set $\nu_k(s_t, a_t) \mathrel{+}= 1$ and set $t \mathrel{+}= 1$
5. Set $N_{k+1}(s,a) = N_k(s,a) + \nu_k(s,a)$

Figure 2: TUCRL algorithm.

is not possible. This strategy is based on the intuition that a transition either does not exist or it should have a sufficiently "big" mass. However, these transitions should be periodically reconsidered in order to avoid *under-exploration* issues. More formally, let $(\rho_t)_{t \in \mathbb{N}}$ be a non-increasing sequence to be defined later, for all $s' \in \mathcal{S}_k^{\mathsf{T}}$, $s \in \mathcal{S}_k^{\mathsf{C}}$ and $a \in \mathcal{A}_s$, the empirical mean $\widehat{p}_k(s'|s,a)$ and variance $\widehat{\sigma}^2_{p,k}(s'|s,a)$ are zero (i.e., this transition has never been observed so far), so the largest probability (most optimistic) of transition from s to s' through any action a is $\widetilde{p}_k^+(s'|s,a) = \frac{49}{3} \frac{b_{k,\delta}}{N_k^{\pm}(s,a)}$. TUCRL compares $\widetilde{p}_k^+(s'|s,a)$ to ρ_{t_k} and forces all transition probabilities below the threshold to zero, while the confidence intervals of transitions to states that have already been explored (i.e., in $\mathcal{S}_k^{\mathsf{C}}$) are preserved unchanged. This corresponds to constructing the alternative confidence interval

$$\overline{B}_{p,k}(s,a) = B_{p,k}(s,a) \cap \left\{ \widetilde{p}(\cdot|s,a) \in \mathcal{C} : \forall s' \in \mathcal{S}_k^{\mathsf{T}} \text{ and } \widetilde{p}_k^+(s'|s,a) < \rho_{t_k}, \widetilde{p}(s'|s,a) = 0 \right\}. \quad (4)$$

Given $\overline{B}_{p,k}$, TUCRL (implicitly) constructs the corresponding set of plausible MDPs $\overline{\mathcal{M}}_k$ and then solves the optimistic optimization problem

$$(\widetilde{M}_k, \widetilde{\pi}_k) = \arg\max_{M \in \overline{\mathcal{M}}_k, \pi} \{g_M^{\pi}\}. \quad (5)$$

The resulting algorithm follows the same structure as UCRL and it is shown in Fig. 2. The episode stopping condition at line 4 is slightly modified w.r.t. UCRL. In fact, it guarantees that one action is always executed and it forces an episode to terminate as soon as a state previously in $\mathcal{S}_k^{\mathsf{T}}$ is visited (i.e., $N_k(s_t,a) = 0$). This minor change guarantees that $N_{k+1}(s,a) = 0$ for all the states $s \in \mathcal{S}_k^{\mathsf{T}}$ that were not reachable at the beginning of the episode. The algorithm also needs minor modifications to the extended value iteration (EVI) algorithm used to solve (5) to guarantee both efficiency and convergence. All technical details are reported in App. C.

In practice, we set $\rho_t = \frac{49 b_{t,\delta}}{3} \sqrt{\frac{SA}{t}}$, so that the *condition to remove transition* reduces to $N_k^{\pm}(s,a) > \sqrt{t_k/SA}$. This shows that only transitions from state-action pairs that have been poorly visited so far are enabled, while if the state-action pair has already been tried often and yet no transition to $s' \in \mathcal{S}_k^{\mathsf{T}}$ is observed, then it is assumed that s' is not reachable from s, a. When the number of visits in (s,a) is big, the transitions to "unvisited" states should be discarded because if the transition actually exists, it is most likely extremely small and so it is worth exploring other parts of the MDP first. Symmetrically, when the number of visits in (s,a) is small, the transitions to "unvisited" states should be enabled because the transitions are quite plausible and the algorithm should try to explore the outcome of taking action a in s and possibly reach states in $\mathcal{S}_k^{\mathsf{T}}$. We denote the set of state-action pairs that are not sufficiently explored by $\mathcal{K}_k = \left\{ (s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} : N_k^{\pm}(s,a) \leq \sqrt{t_k/SA} \right\}$.

3.1 Analysis of TUCRL

We prove that the regret of TUCRL is bounded as follows.

Theorem 1. *For any weakly communicating MDP M, with probability at least $1-\delta$ it holds that for any $T > 1$, the regret of* TUCRL *is bounded as*

$$\Delta(\text{TUCRL}, T) = O\left(r_{\max} D^{\mathcal{C}} \sqrt{\Gamma^{\mathcal{C}} S^{\mathcal{C}} A T \ln\left(\frac{SAT}{\delta}\right)} + r_{\max} \left(D^{\mathcal{C}}\right)^2 S^3 A \ln^2\left(\frac{SAT}{\delta}\right) \right).$$

The first term in the regret shows the ability of TUCRL to adapt to the communicating part of the true MDP M^* by scaling with the *communicating* diameter D^{C} and MDP parameters S^{C} and Γ^{C}. The second term corresponds to the regret incurred in the early stage where the regret grows linearly. When M^* is communicating, we match the square-root term of UCRL (first term), while the second term is just slightly bigger than the one appearing in UCRL by a multiplicative factor S^2/Γ^{C} (ignoring logarithmic terms, see Sec. 5).

We now provide a sketch of the proof of Thm. 1 (the full proof is reported in App. D). In order to preserve readability, all following inequalities should be interpreted up to minor approximations and in high probability.

Let $\Delta_k := \sum_{s,a} \nu_k(s,a)(g^* - r(s,a))$ be the regret incurred in episode k, where $\nu_k(s,a)$ is the number of visits to s,a in episode k. We decompose the regret as

$$\Delta(\text{TUCRL}, T) \lesssim \sum_{k=1}^{m} \Delta_k \cdot \mathbb{1}\{M^* \in \mathcal{M}_k\} \lesssim \sum_{k=1}^{m} \Delta_k \cdot \mathbb{1}\{t_k < C(k)\} + \sum_{k=1}^{m} \Delta_k \cdot \mathbb{1}\{t_k \geq C(k)\}$$

where $C(k) = O\left((D^{\mathsf{C}})^2 S^3 A \ln^2(2SAt_k/\delta)\right)$ defines the length of a full exploratory phase, where the agent may suffer *linear regret*.

Optimism. The first technical difficulty is that whenever some transitions are disabled, the plausible set of MDPs $\overline{\mathcal{M}}_k$ may actually be *biased* and not contain the true MDP M^*. This requires to prove that TUCRL (i.e., the gain of the solution returned by EVI) is always optimistic despite "wrong" confidence intervals. The following lemma helps to identify the possible scenarios that TUCRL can produce (see App. D.2).[4]

Lemma 1. *Let episode k be such that $M^* \in \mathcal{M}_k$, $\mathcal{S}_k^{\mathsf{T}} \neq \emptyset$ and $t_k \geq C(k)$. Then, either $\mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$ (case I) or $\mathcal{K}_k \neq \emptyset$, i.e., $\exists (s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}$ for which transitions to $\mathcal{S}_k^{\mathsf{T}}$ are allowed (case II).*

This result basically excludes the case where $\mathcal{S}_k^{\mathsf{T}} \supset \mathcal{S}^{\mathsf{T}}$ (i.e., some states have not been reached) and yet no transition from $\mathcal{S}_k^{\mathsf{C}}$ to them is enabled. We start noticing that when $\mathcal{S}_k^{\mathsf{T}} = \emptyset$, the true MDP $M^* \in \mathcal{M}_k = \overline{\mathcal{M}}_k$ w.h.p. by construction of the confidence intervals. Similarly, if $\mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$ then $M^* \in \overline{\mathcal{M}}_k$ w.h.p., since TUCRL only truncates transitions that are indeed forbidden in M^* itself. In both cases, we can use the same arguments in [2] to prove optimism. In *case II* the gain of any state $s' \in \mathcal{S}_k^{\mathsf{T}}$ is set to $r_{\max}$ and, since there exists a path from $\mathcal{S}_k^{\mathsf{C}}$ to $\mathcal{S}_k^{\mathsf{T}}$, the gain of the solution returned by EVI is $r_{\max}$, which makes it trivially optimistic. As a result we can conclude that $\widetilde{g}_k \gtrsim g^*$ (up to the precision of EVI).

Per-episode regret. After bounding the optimistic reward $\widetilde{r}_k(s,a)$ w.r.t. $r(s,a)$, the only part left to bound the per-episode regret Δ_k is the term $\widetilde{\Delta}_k = \sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a))$. Similar to UCRL, we could use the (optimistic) optimality equation and rewrite $\widetilde{\Delta}_k$ as

$$\widetilde{\Delta}_k = \sum_{s \in \mathcal{S}} \nu_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) \widetilde{h}_k(s') - \widetilde{h}_k(s) \right) = \nu_k' \left(\widetilde{P}_k - I \right) w_k \tag{6}$$

where $w_k := \widetilde{h}_k - \min_{s \in \mathcal{S}}\{\widetilde{h}_k\} e$ is a shifted version of the vector $\widetilde{h}_k$ returned by EVI at episode k, and then proceed by bounding the difference between $\widetilde{P}_k$ and P_k using standard concentration inequalities. Nonetheless, we would be left with the problem of bounding the ℓ_∞ norm of w_k (i.e., the range of the optimistic vector $\widetilde{h}_k$) over the whole state space, i.e., $\|w_k\|_\infty = sp_{\mathcal{S}}\{\widetilde{h}_k\} = \max_{s \in \mathcal{S}} \widetilde{h}_k(s) - \min_{s \in \mathcal{S}} \widetilde{h}_k(s)$. While in communicating MDPs, it is possible to bound this quantity by the diameter of the MDP as $sp_{\mathcal{S}}\{h_k\} \leq D$ [2, Sec. 4.3], in weakly-communicating MDPs $D = +\infty$, thus making this result uninformative. As a result, we need to restrict our attention to the subset of communicating states $\mathcal{S}^{\mathsf{C}}$, where the diameter is finite. We then split the per-step regret over states depending on whether they are explored enough or not as $\Delta_k \lesssim \sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a)) \mathbb{1}\{(s,a) \notin \mathcal{K}_k\} + r_{\max} \sum_{s,a} \nu_k(s,a) \mathbb{1}\{(s,a) \in \mathcal{K}_k\}$. We start focusing on the poorly visited state-action pairs, i.e., $(s,a) \in \mathcal{K}_k$. In this case TUCRL may suffer the maximum per-step regret $r_{\max}$ but the number of times this event happen is cumulatively "small" (App. D.4.1):

[4] Notice that $M^* \in \mathcal{M}_k$ is true w.h.p. since $\mathcal{M}_k$ is obtained using non-truncated confidence intervals.

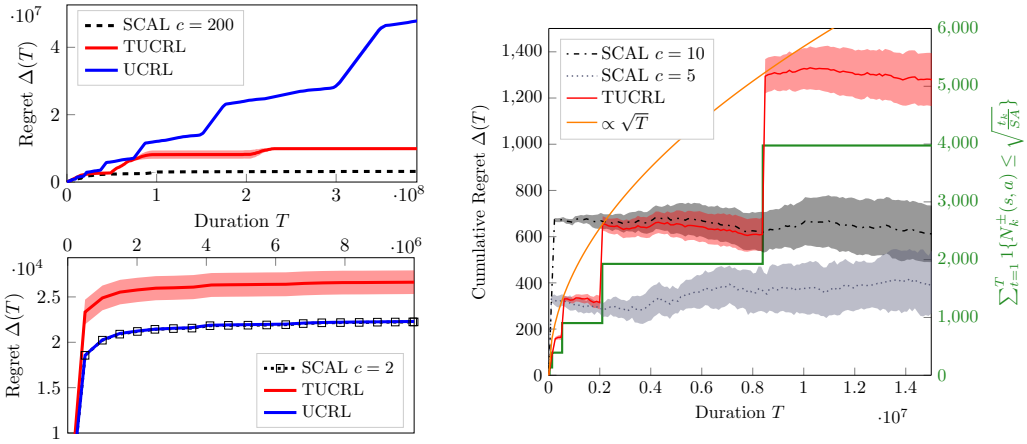


Figure 3: Cumulative regret in the taxi with misspecified states *(left-top)* and in the communicating taxi *(left-bottom)*, and in the weakly communicating three-states domain with $D = +\infty$ *(right)*. Confidence intervals $\beta_{r,k}$ and $\beta_{p,k}$ are shrunk by a factor 0.05 and 0.01 for the three-states domain and taxi, respectively. Results are averaged over 20 runs and 95% confidence intervals are reported.

Lemma 2. *For any $T \ge 1$ and any sequence of states and actions $\{s_1, a_1, \ldots\ldots s_T, a_T\}$ we have:*

$$\sum_{k=1}^{m} \sum_{s,a} \nu_k(s,a) \mathbb{1}\underbrace{\{N_k^{\pm}(s,a) \le \sqrt{t_k/SA}\}}_{(s,a)\in\mathcal{K}_k} \le \sum_{t=1}^{T} \mathbb{1}\left\{N_{k_t}^{\pm}(s_t,a_t) \le \sqrt{t/SA}\right\} \le 2\left(\sqrt{S^c AT} + S^c A\right)$$

When $(s,a) \notin \mathcal{K}_k$ (i.e., $N_k^{\pm}(s,a) > \sqrt{t_k/SA}$ holds), $\sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a)) \cdot \mathbb{1}\{(s,a) \notin \mathcal{K}_k\}$ can be bounded as in Eq. 6 but now restricted on $\mathcal{S}_k^{\mathsf{C}}$, so that,

$$\nu_k(\widetilde{P}_k - I)\widetilde{h}_k = \sum_{s\in\mathcal{S}_k^{\mathsf{C}}} \nu_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s'\in\mathcal{S}_k^{\mathsf{C}}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \right).$$

Since the stopping condition guarantees that $\nu_k(s, \widetilde{\pi}_k(s)) = 0$ for all $s \in \mathcal{S}_k^{\mathsf{T}}$, we can first restrict the outer summation to states in $\mathcal{S}^{\mathsf{C}}$. Furthermore, all state-action pairs $(s,a) \notin \mathcal{K}_k$ are such that the optimistic transition probability $\widetilde{p}_k(s'|s,a)$ is forced to zero for all $s' \in \mathcal{S}_k^{\mathsf{T}}$, thus reducing the inner summation. We are then left with providing a bound for the range of w_k *restricted* to the states in $\mathcal{S}_k^{\mathsf{C}}$, i.e., $sp_{\mathcal{S}_k^{\mathsf{C}}}\{w_k\} = \max_{s\in\mathcal{S}_k^{\mathsf{C}}}\{w_k\}$. We recall that EVI run on a set of plausible MDPs $\overline{\mathcal{M}}_k$ returns a function $\widetilde{h}_k$ such that $\widetilde{h}_k(s') - \widetilde{h}_k(s) \le r_{\max} \cdot \tau_{\overline{\mathcal{M}}_k}(s \to s')$, for any pair $s, s' \in \mathcal{S}$, where $\tau_{\overline{\mathcal{M}}_k}(s \to s')$ is the expected shortest path in the extended MDP $\overline{\mathcal{M}}_k$. Furthermore, since $M^* \in \mathcal{M}_k$, for all $s, s' \in \mathcal{S}_k^{\mathsf{C}}$, $\tau_{\mathcal{M}_k}(s \to s') \le D^{\mathsf{C}}$. Unfortunately, since M^* may not belong to $\overline{\mathcal{M}}_k$, the bound on the shortest path in $\mathcal{M}_k$ (i.e., $\tau_{\mathcal{M}_k}(s \to s')$) may not directly translate into a bound for the shortest path in $\overline{\mathcal{M}}_k$, thus preventing from bounding the range of $\widetilde{h}_k$ even on the subset of states in $\mathcal{S}_k^{\mathsf{C}}$. Nonetheless, in App. E we show that a minor modification to the confidence intervals of $\overline{\mathcal{M}}_k$ makes the shortest paths between any two states $s, s' \in \mathcal{S}_k^{\mathsf{C}}$ equivalent in both sets of plausible MDPs, thus providing the bound $sp_{\mathcal{S}_k^{\mathsf{C}}}\{w_k\} \le D^{\mathsf{C}}$. The final regret in Thm. 1 is then obtained by combining all different terms.

4 Experiments

In this section, we present experiments to validate the theoretical findings of Sec. 3. We compare TUCRL against UCRL and SCAL.[5] We first consider the taxi problem [24] implemented in OpenAI Gym [25].[6] Even such a simple domain contains *misspecified states*, since the state space is constructed as the outer product of the taxi position, the passenger position and the destination. This

[5] To the best of out knowledge, there exists no implementable algorithm to solve the optimization step of REGAL and REGAL.D.

[6] The code is available on GitHub.

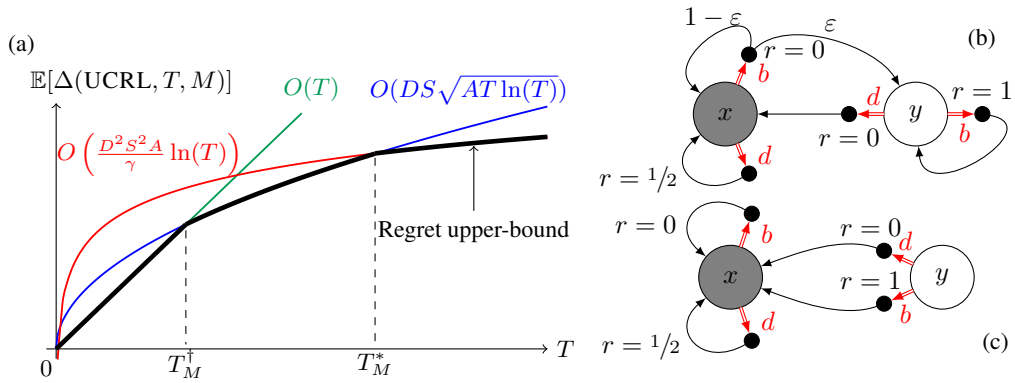


Figure 4: 4a Expected regret of UCRL (with known horizon T given as input) as a function of T. 4b 4c Toy example illustrating the difficulty of learning non-communicating MDPs. We represent a family of possible MDPs $\mathcal{M} = (M_\varepsilon)_{\varepsilon\in[0,1]}$ where the probability ε to go from x to y lies in $[0, 1]$.

leads to states that cannot be reached from any possible starting configuration (all the starting states belong to $\mathcal{S}^{\text{C}}$). More precisely, out of 500 states in $\mathcal{S}$, 100 are non-reachable. On Fig. 3*(left)* we compare the regret of UCRL, SCAL and TUCRL when the misspecified states are present *(top)* and when they are removed *(bottom)*. In the presence of misspecified states *(top)*, the regret of UCRL clearly grows linearly with T while TUCRL is able to *learn* as expected. On the other hand, when the MDP is communicating *(bottom)* TUCRL performs similarly to UCRL. The small loss in performance is most likely due to the initial exploration phase during which the confidence intervals on the transition probabilities used by UCRL (see definition of $\overline{\mathcal{M}}_k$) are tighter than those used by TUCRL (see definition of $\overline{\mathcal{M}}_k^+$). TUCRL uses a "loose" bound on the ℓ_1-norm while UCRL uses S different bounds, one for every possible next state. Finally, SCAL outperforms TUCRL by exploiting prior knowledge on the bias span.

We further study TUCRL regret in the simple three-state domain introduced in [6] (see App. G for details) with different reward distributions (uniform instead of Bernouilli). The environment is composed of only three states (s_0, s_1 and s_2) and one action per state, except in s_2 where two actions are available. As a result, the agent only has the choice between two possible policies. Fig. 3*(left)* shows the cumulative regret achieved by TUCRL and SCAL (with different upper-bounds on the bias span) when the diameter is *infinite* i.e., $\mathcal{S}^{\text{C}} = \{s_0, s_2\}$ and $\mathcal{S}^{\text{T}} = \{s_1\}$ (we omit UCRL, since it suffers linear regret). Both SCAL and TUCRL quickly achieve sub-linear regret as predicted by theory. However, SCAL and TUCRL seem to achieve different growth rates in regret: while SCAL appears to reach a logarithmic growth, the regret of TUCRL seems to grow as $\sqrt{T}$ with *periodic* "jumps" that are increasingly distant (in time) from each other. This can be explained by the way the algorithm works: while most of the time TUCRL is optimistic on the restricted state space $\mathcal{S}^{\text{C}}$ (i.e., $\mathcal{S}_k^{\text{C}} = \mathcal{S}^{\text{C}}$), it *periodically* allows transitions to the set $\mathcal{S}^{\text{T}}$ (i.e., $\mathcal{S}_k^{\text{C}} = \mathcal{S}$), which is indeed not reachable. Enabling these transitions triggers aggressive *exploration* during an entire episode. The policy played is then sub-optimal creating a "jump" in the regret. At the end of this *exploratory episode*, $\mathcal{S}_k^{\text{C}}$ will be set again to $\mathcal{S}^{\text{C}}$ and the regret will stop increasing until the condition $N_k^{\pm} \leq \sqrt{t_k/SA}$ occurs again (the time between two consecutive exploratory episodes grows quadratically). The cumulative regret incurred during exploratory episodes can be bounded by the term plotted in green on Fig. 3*(left)*. In Lem. 2 we proved that this term is always bounded by $O(\sqrt{S^{\text{C}}AT})$. Therefore, it is not surprising to observe a $\sqrt{T}$ increase of both the green and red curves. Unfortunately, the growth rate of the regret will keep increasing as $\sqrt{T}$ and will never become logarithmic unlike SCAL (or UCRL when the MDP is communicating). This is because the condition $N_k^{\pm} \leq \sqrt{t_k/SA}$ will always be triggered $\Theta(\sqrt{T})$ times for any T. In Sec. 5 we show that this is not just a drawback specific to TUCRL, but it is rather an *intrinsic limitation* of learning in weakly-communicating MDPs.

5 Exploration-exploitation dilemma with infinite diameter

In this section we further investigate the empirical difference between SCAL and TUCRL and prove an impossibility result *characterising* the *exploration-exploitation dilemma* when the diameter is allowed to be *infinite* and *no prior knowledge* on the optimal bias span is available.

We first recall that the expected regret $\mathbb{E}[\Delta(\text{UCRL}, M, T)]$ of UCRL (with input parameter $\delta = 1/3T$) after $T \geq 1$ time steps and for any finite MDP M can be bounded in several ways:

$$\mathbb{E}[\Delta(\text{UCRL}, M, T)] \leq \begin{cases} r_{\max} T \text{ (by definition)} \\ C_1 \cdot r_{\max} D \sqrt{\Gamma S A T \ln(3T^2)} + \frac{1}{3} \text{ [2, Theorem 2]} \\ C_2 \cdot r_{\max} \frac{D^2 \Gamma S A}{\gamma} \ln(T) + C_3(M) \text{ [2, Theorem 4]} \end{cases} \tag{7}$$

where $\gamma = g^*_M - \max_{s,\pi}\{g^\pi_M(s) : \ g^\pi_M(s) < g^*_M\}$ is the gap in gain, $C_1 := 34$ and $C_2 := 34^2$ are numerical constants independent of M, and $C_3(M) := O(\max_{\pi:\pi(s)=a} T_\pi)$ with T_π a measure of the "mixing time" of policy π. The three different bounds lead to three different *growth rates* for the function $T \longmapsto \mathbb{E}[\Delta(\text{UCRL}, M, T)]$ (see Fig. 4a): 1) for $T^\dagger_M \geq T \geq 0$, the expected regret is linear in T, 2) for $T^*_M \geq T \geq T^\dagger_M$ the expected regret grows as $\sqrt{T}$, 3) finally for $T \geq T^*_M$, the increase in regret is only logarithmic in T. These different *"regimes"* can be observed empirically (see [6, Fig. 5, 12]). Using (7), it is easy to show that the time it takes for UCRL to achieve sub-linear regret is at most $T^\dagger_M = \widetilde{O}(D^2 \Gamma S A)$. We say that an algorithm is *efficient* when it achieves sublinear regret after a number of steps that is polynomial in the parameters of the MDP (i.e., UCRL is then *efficient*). We now show with an example that *without prior knowledge*, any *efficient* learning algorithm must satisfy $T^*_M = +\infty$ when M has *infinite diameter* (i.e., it cannot achieve logarithmic regret).

Example 1. *We consider a family of weakly-communicating MDPs $\mathcal{M} = (M_\varepsilon)_{\varepsilon \in [0,1]}$ represented on Fig. 4*(right)*. Every MDP instance in $\mathcal{M}$ is characterised by a specific value of $\varepsilon \in [0,1]$ which corresponds to the probability to go from x to y. For $\varepsilon > 0$ (Fig. 4b), the optimal policy of M_ε is such that $\pi^*(x) = b$ and the optimal gain is $g^*_\varepsilon = 1$ while for $\varepsilon = 0$ (Fig. 4c) the optimal policy is such that $\pi^*(x) = d$ and the optimal gain is $g^*_0 = 1/2$. We assume that the learning agent knows that the true MDP M^* belongs to $\mathcal{M}$ but does not know the value ε^* associated to $M^* = M_{\varepsilon^*}$. We assume that all rewards are deterministic and that the agent starts in state x (coloured in grey).*

Lemma 3. *Let $C_1, C_2, \alpha, \beta > 0$ be positive real numbers and f a function defined for all $\varepsilon \in]0,1]$ by $f(\varepsilon) = C_1(1/\varepsilon)^\alpha$. There exists no learning algorithm $\mathfrak{A}_T$ (with known horizon T) satisfying both*
1. for all $\varepsilon \in]0,1]$, there exists $T^\dagger_\varepsilon \leq f(\varepsilon)$ such that $\mathbb{E}[\Delta(\mathfrak{A}_T, M_\varepsilon, x, T)] < 1/6 \cdot T$ for all $T \geq T^\dagger_\varepsilon$,
2. and there exists $T^*_0 < +\infty$ such that $\mathbb{E}[\Delta(\mathfrak{A}_T, M_0, x, T)] \leq C_2(\ln(T))^\beta$ for all $T \geq T^*_0$.

Note that point *1* in Lem. 3 formalizes the concept of *"efficient learnability"* introduced by Sutton and Barto [26, Section 11.6] i.e., "learnable within a polynomial rather than exponential number of time steps". All the MDPs in $\mathcal{M}$ share the same number of states $S = 2 \geq \Gamma$, number of actions $A = 2$, and gap in average reward $\gamma = 1/2$. As a result, any function of S, Γ, A and γ will be considered as constant. For $\varepsilon > 0$, the diameter coincides with the optimal bias span of the MDP and $D = sp_{\mathcal{S}}\{h^*\} = 1/\varepsilon < +\infty$, while for $\varepsilon = 0$, $D = +\infty$ but $sp_{\mathcal{S}}\{h^*\} = 1/2$. As shown in Eq. 7 and Thm. 1, UCRL and TUCRL satisfy property *1.* of Lem. 3 with $\alpha = 2$ and $C_1 = O(S^2 A)$ but do not satisfy *2.* On the other hand, SCAL satisfies *2.* with $\beta = 1$ and $C_2 = O(H^2 SA/\gamma)$ (although this result is not available in the literature, it is straightforward to adapt the proof of UCRL [2, Theorem 4] to SCAL) but since [6, Theorem 12] holds only when $H \geq sp_{\mathcal{S}}\{h^*\}$, SCAL only satisfies *1.* for $\varepsilon \geq 1/H$ and $\varepsilon = 0$ (not for $\varepsilon \in]0, 1/H[$). Lem. 3 proves that no algorithm can actually achieve both *1.* and *2.* As a result, since TUCRL satisfies *1.*, it cannot satisfy *2.* This matches the empirical results presented in Sec. 4 where we observed that when the diameter is infinite, the growth rates of the regret of SCAL and TUCRL were respectively logarithmic and of order $\Theta(\sqrt{T})$. An algorithm that does not satisfy *1.* could potentially satisfy *2.* but, by definition of *1.*, it would suffer linear regret for a number of steps that is more than *polynomial* in the parameters of the MDP (more precisely, $e^{D^{1/\beta}}$). This is not a very desirable property and we claim that an *efficient* learning algorithm should always prefer *finite time guarantees* (*1.*) over *asymptotic guarantees* (*2.*) when they cannot be accommodated.

6 Conclusion

We introduced TUCRL, an algorithm that efficiently balances exploration and exploitation in weakly-communicating and multi-chain MDPs, when the starting state s_1 belongs to a communicating set (Asm. 1). We showed that TUCRL achieves a square-root regret bound and that, in the general case, it is not possible to design algorithm with logarithmic regret and polynomial dependence on the MDP parameters. Several questions remain open: **1)** relaxing Asm. 1 by considering a transient initial state (i.e., $s_1 \in \mathcal{S}^{\text{T}}$), **2)** refining the lower bound of Jaksch et al. [2] to finally understand whether it is possible to scale with $sp_{\mathcal{S}}\{h^*\}$ (at least in communicating MDPs) instead of D without any prior knowledge (the flaw in REGAL.D may suggest it is indeed impossible).

Acknowledgments

This research was supported in part by French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and French National Research Agency (ANR) under project ExTra-Learn (n.ANR-14-CE24-0010-01).

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[2] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[3] Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*, pages 35–42. AUAI Press, 2009.

[4] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in mdps with options without prior knowledge. In *NIPS*, pages 3169–3179, 2017.

[5] Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 2018.

[6] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *CoRR*, abs/1802.04020, 2018.

[7] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[8] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

[9] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. AUAI Press, 2015.

[10] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 2017.

[11] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

[12] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[14] Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

[15] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, Oct 1995. doi: 10.1109/SFCS.1995.492488.

[16] Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org, 2015.

[17] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pages 2–11, Arlington, Virginia, United States, 2015. AUAI Press. ISBN 978-0-9966431-0-8.

[18] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems 30*, pages 1333–1342. Curran Associates, Inc., 2017.

[19] Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *CoRR*, abs/1711.07979, 2017.

[20] Ian Osband and Benjamin Van Roy. Posterior sampling for reinforcement learning without episodes. *CoRR*, abs/1608.02731, 2016.

[21] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

[22] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[23] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

[24] Thomas G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.*, 13:227–303, 2000.

[25] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016.

[26] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. Adaptive computation and machine learning. MIT Press, second edition, 2018. ISBN 9780262039246.

[27] Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 543–551, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

A Mistake in the regret bound of REGAL.D

A.1 Regularized optimistic RL (REGAL)

In weakly communicating MDPs, to avoid the over-optimism of UCRL, Bartlett and Tewari [3] proposed to penalise the optimism on g^* by the optimal bias span $sp_{\mathcal{S}}\{h^*\}$. Formally, at each episode k, their algorithm –REGAL– solves the following optimization problem:

$$\widetilde{M}_k = \arg\max_{M \in \mathcal{M}_k} \{g^*_M - C_k \cdot sp_{\mathcal{S}}\{h^*_M\}\} \tag{8}$$

where $C_k \geq 0$ is a regularisation coefficient. Note that such optimization requires to first compute the optimal policy for a given MDP $M \in \mathcal{M}_k$ and then evaluate the regularized gain. Implicitly, this defines the optimistic policy $\widetilde{\pi}_k = \arg\max_{\pi \in \Pi^{\text{SD}}} \{g^{\pi}_{\widetilde{M}_k}\}$. The term $sp_{\mathcal{S}}\{h^*\}$ can be interpreted as a measure of the *complexity* of the environment: the bigger $sp_{\mathcal{S}}\{h^*\}$, the more difficult it is to achieve the stationary reward g^* by following the optimal policy. In *supervised learning*, regularisation is often used to penalise the objective function by a measure of the complexity of the model so as to avoid *overfitting*. It is thus reasonable to expect that *over-optimism* in online RL can also be avoided through regularisation.

The regret bound of REGAL holds only when C_k is set to $\Theta(1/\sum_{s,a} \nu_k(s,a))$. This means that REGAL requires the knowledge of (future) visit counts $\nu_k(s,a)$ before episode k begins in order to tune the regularisation coefficient C_k. Unfortunately, an episode stops when the number of visits in a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ has doubled and it is not possible to predict the future sequence of states of a given policy for two reasons: 1) the true MDP M^* is unknown and 2) what is observed is a random *sampled* trajectory (as opposed to *expected*). As a result, REGAL is not *implementable*. Bartlett and Tewari [3] proposed an alternative algorithm –REGAL.D– that leverages on the *doubling trick* to guess the length of episode k (i.e., $\sum_{s,a} \nu_k(s,a)$) and proved a slightly worse regret bound than for REGAL. REGAL.D divides an episode k into sub-iterations where it applies the doubling trick techniques. At each sub-iteration j, REGAL.D guesses that the length of the episode will be at most 2^j and it solves problem (8) with $C_{k,j} \propto 1/\sqrt{2^j}$. Then, it executes the optimistic policy $\widetilde{\pi}_{k,j}$ on the true MDP until the UCRL stopping condition is reached or 2^j steps are performed. In the first case the episode k ends since the guess was correct, while, in the second case, a new sub-iteration $j+1$ is started. This implies that for any k, j:

$$\sum_{s,a} \nu_{k,j}(s,a) \leq 2^j, \tag{9}$$

where $\nu_{k,j}(s,a)$ denotes the number of visits to (s,a) during episode k and sub-iteration j.

A.2 The doubling trick issue

The mistake in REGAL.D is located in the proof of the regret [3, Theorem 3] (see Sec. 6.3). Let $\widetilde{h}_{k,j}$ denote the optimistic bias span at episode k and sub-iteration j induced by the doubling trick. At a high level, the mistake comes from the attempt to upper-bound the term $x \cdot \sum_{s,a} \nu_{k,j}(s,a)$ by $x \cdot 2^j$ (for a given x) using the fact the $\sum_{s,a} \nu_{k,j}(s,a) \leq 2^j$. Unfortunately, this is possible only under the assumption that $x \geq 0$ that does not hold in the case of REGAL.D.

Formally, while bounding $\sum_{k \in G} \Delta_k$, the authors have to deal with the term (derived by the combination of [3, Eq. 15] and [3, Lem. 11] with [3, Eq. 14]):

$$U := \sum_{k \in G} \sum_{j} sp_{\mathcal{S}}\left\{\widetilde{h}_{k,j}\right\} \left(c\sqrt{\sum_{s,a} \nu_{k,j}(s,a)} - C_{k,j} \sum_{s,a} \nu_{k,j}(s,a) \right)$$

where $c := 2S\sqrt{12\ln(2AT/\delta)} + \sqrt{2\ln(1/\delta)} \geq 0$ and recall that $\sum_{s,a} \nu_{k,j}(s,a)$ denotes the *actual* length of the episode k at sub-iteration j. In the REGAL.D proof the authors directly replaced the actual length of the episode with the guessed length $2^j := \ell_{k,j}$ showing that the first term can be upper-bounded by $c \cdot \sqrt{\sum_{s,a} \nu_{k,j}(s,a)} \leq c \cdot \sqrt{2^j}$ (due to Eq. 9). Concerning the second term, they write $-C_{k,j} \sum_{s,a} \nu_{k,j}(s,a) \leq -C_{k,j} 2^j$. Since $-C_{k,j} := -c/\sqrt{2^j} \leq 0$ is negative, this last inequality is

not true and the reverse inequality holds instead (using Eq. 9): $-C_{k,j}\sum_{s,a}\nu_{k,j}(s,a) \geq -C_{k,j}2^j$. Therefore, it is not possible to guarantee that $U \leq 0$ as claimed by Bartlett and Tewari [3] (the authors probably didn't pay attention to the sign). To do this, we would need to *lower-bound* $\sum_{s,a}\nu_{k,j}(s,a)$. Unfortunately, the only lower bound with probability 1 available for that term is $\min_{s,a}\{N_k(s,a)\}+2$. This is not big enough to cancel the term $c\sqrt{\sum_{s,a}\nu_{k,j}(s,a)}$ and $C_{k,j}$ needs to be increased. As a result, the term $sp_{\mathcal{S}}\{h^\star\}\sum_{k\in G}\sum_j C_{k,j}\sqrt{\sum_{s,a}\nu_{k,j}(s,a)}$ becomes too big and all the proof collapses.

Notice that a similar mistake is contained in the work by Maillard et al. [27] where they use a regularized approach to learn a state representation in online settings. Similarly to [3], the authors have to bound the term $\sum_{s,a}\nu_{k,j}(s,a)(g^*-\widetilde{g}_{k,j})$. By exploiting the fact that $g^*-\widetilde{g}_{k,j}\leq\alpha$ (we omit the full expression of α for sake of clarity) [27, Eq. 17 Sec. 5.2] the authors derived the bound $\sum_{s,a}\nu_{k,j}(s,a)(g^*-\widetilde{g}_{k,j})\leq 2^j\cdot\alpha$ [27, Eq. 18]. The difference $g^*-\widetilde{g}_{k,j}$ might be negative in which case the result does not hold. Actually for the case in which there is no regularization $C_{k,j}=0$, $g^*\leq\widetilde{g}_{k,j}$ which is what is used in the regret proof of UCRL. Therefore, it is very likely that the sign of $g^*-\widetilde{g}_{k,j}$ can sometimes be negative.

In conclusion, it seems unavoidable to use a *lower-bound* (and not an upper-bound) on $\sum_{s,a}\nu_{k,j}(s,a)$ to derive a correct regret bound for REGAL.D. As already mentioned, given the current stopping condition of an episode, the only reasonable lower bound is $\min_{s,a}\{N_k(s,a)\}+2$ and it does not seem sufficient to derive a sensible regret bound. Another research direction could be to change the stopping condition. However, one of the terms in the regret bound of REGAL (and of REGAL.D) scales as $m\sqrt{T}\log_2(T)$ where m is the number of episodes. The term m is highly sensitive to the stopping condition and there is very little margin if we want to avoid $m\sqrt{T}\log_2(T)$ to become the leading term in the regret bound. All the efforts we put in this direction were unsuccessful. We conjecture that regularising by the optimal bias span might not allow to learn MDPs with infinite diameter.

B Unbounded optimal bias span with continuous Bayesian priors/posteriors

Recently, Ouyang et al. [18] and Theocharous et al. [19] proposed posterior sampling algorithms and proved bounds on the expected Bayesian regret. The regret bounds that they derive scale linearly with H, where H is the highest optimal bias span of all the MDPs that can be drawn from the prior/posterior distribution. Formally, let $f(\theta)$ be the density function of the prior/posterior distribution over the family of MDPs (M_θ) parametrised by θ. Then:

$$H := \sup_{\theta: f(\theta)>0}\{sp_{\mathcal{S}}\{h^*_\theta\}\}.$$

In this section we present an example where H is *infinite* and argue that it is probably the case for most priors/posteriors used in practice.

Example 2 (Unbounded optimal bias span with continuous prior/posterior)**.** *Consider the example of Fig. 5. There is only one action in every state and so one optimal policy. The (unique) action that can be played in state s_0 loops on s_0 with probability $1-\theta$ and goes to s_1 with probability θ. The reward associated to this action is 0. Symmetrically, the (unique) action that can be played in state s_1 loops on s_1 with probability $1-\theta$ and goes to s_1 with probability θ. The reward associated to this action is 1. This MDP is characterised by the parameter θ and we denote it by M_θ. For any $\theta\in[0,1]$, we denote by g^*_θ (resp. h^*_θ) the optimal gain (resp. bias) of M_θ. Observe that when $\theta>0$, M_θ is ergodic*

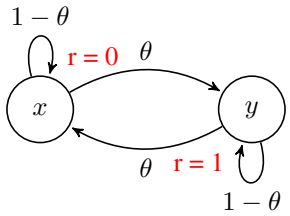


Figure 5: Toy example of a parametrised MDP M_θ with a single policy (one action per state).

and therefore the optimal gain $g^*_\theta = 1/2$ is state-independent whereas when $\theta = 0$, M_θ is multichain and the optimal gain does depend on the initial state: $g^*_0(x) = 0 < 1 = g^*_0(y)$.

Let's assume that the prior/posterior distribution we use on M_θ is characterised by a probability density function f satisfying $f(\theta) > 0$ for all $\theta > 0$ and $f(0) = 0$. Note that this assumption does not constrain the "smoothness" of f e.g., f can have continuous derivatives of all orders. Under this assumption, f is non-zero only for ergodic MDPs. It goes without saying that for all $\theta \in [0, 1]$ (0 included), $sp_{\mathcal{S}}\{h^*_\theta\} < +\infty$ by definition (the optimal bias span is always finite). More precisely we have:

$$g^*_\theta = \begin{cases} [1/2, 1/2]^\intercal & \text{if } \theta > 0 \\ [0, 1]^\intercal & \text{if } \theta = 0 \end{cases} \qquad \text{and} \qquad sp_{\mathcal{S}}\{h^*_\theta\} = \begin{cases} \frac{1}{2\theta} & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

As a result, although $sp_{\mathcal{S}}\{h^*_\theta\}$ is always *finite*, i.e., $\forall \theta \in [0, 1],\ sp_{\mathcal{S}}\{h^*_\theta\} < +\infty$, it is *unbounded* on the set of plausible MDPs $\theta \in]0, 1]$ satisfying $f(\theta) > 0$, i.e.,

$$H := \sup_{\theta \in]0,1]} \{sp_{\mathcal{S}}\{h^*_\theta\}\} = \lim_{\theta \to 0^+} \frac{1}{2\theta} = +\infty$$

Therefore, the regret bound $\widetilde{O}(HS\sqrt{AT})$ proved by Ouyang et al. [18], Theocharous et al. [19] does not hold with prior/posterior f since $H = +\infty$. One might argue that the proofs in [18, 19] could be fixed by showing that H is bounded with probability 1. Unfortunately, for any $C \in [0, +\infty[$, the probability $\mathbb{P}(sp_{\mathcal{S}}\{h^*_\theta\} \geq C) = \int_{\theta=0}^{\frac{1}{2C}} f(\theta) d\theta > 0$ of sampling an MDP with $sp_{\mathcal{S}}\{h^*_\theta\} \geq C$ is strictly positive. We therefore conjecture that for this specific choice of priors/posteriors, the regret proof in [18, 19] cannot be fixed without major changes and new arguments. More generally, let's imagine that we have a prior/posterior distribution f satisfying:

- there exists θ_0 such that M_{θ_0} has non-constant gain i.e., $sp_{\mathcal{S}}\left\{g^*_{\theta_0}\right\} > 0$,
- there exists an open neighbourhood of θ_0 denoted Θ_0 such that $\forall \theta \in \Theta_0$, M_θ has constant gain (e.g., M_θ is weakly-communicating) and $f(\theta) > 0$.

In this case we will face the same problem as in Ex. 2 i.e.,

$$\sup_{\theta:\ f(\theta)>0} \{sp_{\mathcal{S}}\{h^*_\theta\}\} = +\infty \ \text{ and } \ \forall C \in [0, +\infty[,\ \mathbb{P}(sp_{\mathcal{S}}\{h^*_\theta\} \geq C) > 0$$

When the set of plausible MDPs is *finite*, this problem cannot occur. But most priors/posteriors used in practice are *continuous* distributions. For instance, a Dirichlet distribution will most likely satisfy the above assumptions.

C Algorithmic Details

For technical reasons (see App. E), we consider a slight *relaxation of the optimization problem* (5) in which $\overline{\mathcal{M}}_k$ is replaced by a relaxed extended MDP $\overline{\mathcal{M}}^+_k \supseteq \overline{\mathcal{M}}_k$ defined by using ℓ_1-norm concentration inequalities for $p(\cdot|s, a)$. Let $B^+_{p,k}(s, a) = \{\widetilde{p}(\cdot|s, a) \in \mathcal{C} :\ \|\widetilde{p}(\cdot|s, a) - \widehat{p}(\cdot|s, a)\|_1 \leq \sum_{s'} \beta^{sas'}_{p,k}\}$ (resp. $\overline{B}^+_{p,k}$) be the relaxed confidence interval, then $\mathcal{M}^+_k$ (resp. $\overline{\mathcal{M}}^+_k$) is the corresponding (relaxed) set of plausible MDPs. This relaxed optimistic optimization problem is solved by running extended value iteration (EVI) on $\overline{\mathcal{M}}^+_k$ (up to accuracy $\epsilon_k = r_{\max}/\sqrt{t_k}$). Technically, we restrict EVI to work on the set of states $\mathcal{S}^{\text{EVI}}_k$ that are optimistically reachable from the communicating set $\mathcal{S}^{\text{C}}_k$. In practice, $\mathcal{S}^{\text{EVI}}_k = \mathcal{S}^{\text{C}}_k$ when $\mathcal{K}_k = \emptyset$ since all the transitions to $\mathcal{S}^{\text{T}}_k$ are forbidden, otherwise $\mathcal{S}^{\text{EVI}}_k = \mathcal{S}$. Alg. 1 shows this variation of EVI that we name *Truncated EVI*. Then, at each episode k, TUCRL runs TEVI with the following parameters: $(\widetilde{g}_k, \widetilde{h}_k, \widetilde{\pi}_k) = \text{TEVI}(\mathbf{0}, \overline{\mathcal{M}}^+_k, \mathcal{S}^{\text{EVI}}_k, \epsilon_k)$. Starting from an initial vector $v_0 = 0$, TEVI iteratively applies (on a subset $\mathcal{S}^{\text{EVI}}_k$ of states) the optimal Bellman operator $\widetilde{L}_{\overline{\mathcal{M}}^+_k}$ associated to the (extended) MDP $\overline{\mathcal{M}}^+_k$ defined as

$$\forall v \in \mathbb{R}^S, \quad \widetilde{L}_{\overline{\mathcal{M}}^+_k} v(s) := \max_{a \in \mathcal{A}_s} \left\{ \max_{\widetilde{r} \in B_{r,k}(s,a)} \widetilde{r} + (\widetilde{p}^{sa})^\intercal v \right\}, \tag{10}$$

where $\widetilde{p}^{sa} = \arg\max_{\widetilde{p} \in \overline{B}^+_{p,k}(s,a)} \{\widetilde{p}^T v\}$ can be solved using [2, Fig. 2], except for $(s, a) \notin \mathcal{K}_k$ for which we force $\widetilde{p}^{sa}(s') := 0$ for any $s' \in \mathcal{S}^{\text{T}}_k$ (see Alg. 2). If TEVI is stopped when

Algorithm 1 TRUNCATED EXTENDED VALUE ITERATION (TEVI)

Input: value vector v_0, extended MDP $\mathcal{M}$, set of states $\overline{\mathcal{S}}$, accuracy ϵ
Output: g_n, v_n, π_n
$n := 0$
$v_1(s) := \widetilde{L}_{\mathcal{M}} v_0(s) := \max_{a\in\mathcal{A}_s} \left\{\max_{\widetilde{r}\in B_r(s,a)} \widetilde{r} + \max_{\widetilde{p}\in B_p(s,a)} \widetilde{p}^{\mathsf{T}} v_0\right\}, \forall s \in \overline{\mathcal{S}}$ (see App. C)
while $\max_{s\in\overline{\mathcal{S}}} \{v_{n+1}(s) - v_n(s)\} - \min_{s\in\overline{\mathcal{S}}} \{v_{n+1}(s) - v_n(s)\} > \epsilon$ **do**
  $n := n + 1$
  $v_{n+1}(s) := \widetilde{L}_{\mathcal{M}} v_n(s), \forall s \in \overline{\mathcal{S}}$
end while
$g_n := \frac{1}{2}\left(\max_{s\in\overline{\mathcal{S}}} \{v_{n+1}(s) - v_n(s)\} + \min_{s\in\overline{\mathcal{S}}} \{v_{n+1}(s) - v_n(s)\}\right)$
$\pi_n(s) \in \arg\max_{a\in\mathcal{A}_s} \left\{\max_{\widetilde{r}\in B_r(s,a)} \widetilde{r} + \max_{\widetilde{p}\in B_p(s,a)} \widetilde{p}^{\mathsf{T}} v_n\right\}, \forall s \in \overline{\mathcal{S}}$

$sp_{\mathcal{S}_k^{\text{EVI}}}\{v_{n+1} - v_n\} \le \epsilon_k$ and the true MDP is sufficiently explored, then the greedy policy $\widetilde{\pi}_k := \pi_n$ w.r.t. v_n is ϵ_k-optimistic, i.e., $\widetilde{g}_k := g_n \ge g^*_{M^*} - \epsilon_k$ (see Sec. 3.1 for details). The policy $\widetilde{\pi}_k$ is then executed until the number of visits to a state-action pair is doubled or a new state is "discovered" (i.e., $s_t \in \mathcal{S}^{\mathsf{T}}_{k_t}$). Note that the condition $sp_{\mathcal{S}_k^{\text{EVI}}}\{v_{n+1} - v_n\} \le \epsilon_k$ is always met after a *finite* number of steps since the extended MDP $\overline{\mathcal{M}}_k^+$ is communicating on the restricted state space $\mathcal{S}_k^{\text{EVI}}$. Finally, notice that when the true MDP M^* is communicating, there exists an episode $\overline{k}$ s.t. for all $k \ge \overline{k}$, $\mathcal{S}_k^{\mathsf{T}} = \emptyset$ and TUCRL *can be reduced to* UCRL by considering $\mathcal{M}_k$ in place of $\overline{\mathcal{M}}_k^+$.

D Regret of TUCRL

We follow the proof structure of Jaksch et al. [2], Fruit et al. [6] and use similar notations. Nonetheless, several parts of the proof significantly differ from [2, 6]:

- in Sec. D.2 we prove that after a finite number of steps, TUCRL is *gain-optimistic* (which is not as straightforward as in the case of UCRL),
- in Sec. D.3 we show that the sums taken over the whole state space $\mathcal{S}$ that appear in the main term of the regret decomposition of UCRL can be restricted to sums over $\mathcal{S}_k^{\mathsf{C}}$ thanks to the new stopping condition used for episodes and the use of the condition $N_k^{\pm}(s,a) > \sqrt{t_k/SA}$ (see (18)),
- in Sec. D.4.1, we bound the number of time steps spent in "bad" state-action pairs (s,a) satisfying $N_k^{\pm}(s,a) \le \sqrt{t_k/SA}$,
- in Sec. D.4.3, we bound the number of episodes with the new stopping condition.

D.1 Splitting into episodes

The regret of TUCRL after T time steps is defined as: $\Delta(\text{TUCRL}, T) := Tg^* - \sum_{t=1}^T r_t(s_t, a_t)$. Defining $\Delta_k = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nu_k(s,a)\,(g^* - r(s,a))$ and using the same arguments as in [2, 6], it holds with probability $1 - \frac{\delta}{12T^{4/5}}$ that:

$$\Delta(\text{TUCRL}, T) \le \sum_{k=1}^{m} \Delta_k + r_{\max}\sqrt{\frac{5}{2} T \ln\left(\frac{8T}{\delta}\right)} \tag{11}$$

D.2 Episodes with $M^* \in \mathcal{M}_k$

We now assume that $M^* \in \mathcal{M}_k$. As done in App. C, let's denote by $\widetilde{g}_k$, $\widetilde{h}_k$ and $\widetilde{\pi}_k$ the outputs of TEVI$(\mathbf{0}, \mathcal{M}_k^{\diamond}, \mathcal{S}_k^{\text{EVI}}, \varepsilon_k)$ (see Alg. 1) where $\varepsilon_k := r_{\max}/\sqrt{t_k}$ and

$$\mathcal{S}_k^{\text{EVI}} = \begin{cases} \mathcal{S}_k^{\mathsf{C}} & \text{if } \mathcal{K}_k = \emptyset \\ \mathcal{S} & \text{otherwise} \end{cases}, \qquad \mathcal{M}_k^{\diamond} = \begin{cases} \mathcal{M}_k = \overline{\mathcal{M}}_k & \text{if } \mathcal{S}_k^{\mathsf{T}} = \emptyset \\ \overline{\mathcal{M}}_k^+ & \text{otherwise} \end{cases}. \tag{12}$$

TEVI returns an approximate solution of a slightly modified version of Problem 5:

$$(\widetilde{M}_k, \widetilde{\pi}_k) = \arg\max_{M \in \mathcal{M}_k^\diamond, \pi} \{g_M^\pi\}.$$

In order to bound Δ_k we first show that $\widetilde{g}_k \gtrsim g^*$ (up to $r_{\max}/\sqrt{t_k}$-accuracy). If $\mathcal{S}_k^{\mathsf{T}} = \emptyset$ then by definition $\mathcal{M}_k^\diamond = \overline{\mathcal{M}}_k = \mathcal{M}_k \ni M^*$ and so we can use the same argument as in [2, Sec. 4.3 & Thm. 7]. If $\mathcal{S}_k^{\mathsf{T}} \neq \emptyset$, the true MDP M^* might not be "included" in the extended MDP $\overline{\mathcal{M}}_k^+$ considered by EVI and we cannot use the same argument. To overcome this problem we first assume that t_k is big enough which allows us to prove a useful lemma (Lem. 4):

$$t_k \geq \frac{2401}{9}\left(D^{\mathsf{C}}\right)^2 SA\left(S_k^{\mathsf{T}} \ln\left(\frac{2SAt_k}{\delta}\right)\right)^2 := C(k) \tag{13}$$

where $S_k^{\mathsf{T}} := |\mathcal{S}_k^{\mathsf{T}}|$ is the cardinal of $\mathcal{S}_k^{\mathsf{T}}$.

Lemma 4. *Let episode k be such that $M^* \in \mathcal{M}_k$, $\mathcal{S}_k^{\mathsf{T}} \neq \emptyset$ and* (13) *holds. Then,*

$$\left(\forall (s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}, N_k^{\pm}(s,a) > \sqrt{\frac{t_k}{SA}}\right) \implies \mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$$

Proof. Assume that episode k is such that (13) holds and that for any state-action pair $(s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}$

$$N_k^{\pm}(s,a) > \sqrt{\frac{t_k}{SA}} \geq \frac{49}{3} D^{\mathsf{C}} S_k^{\mathsf{T}} \ln\left(\frac{2SAt_k}{\delta}\right)$$

Since $\mathcal{S}_k^{\mathsf{T}} \neq \emptyset$ and $M^* \in \mathcal{M}_k$, for any $(s,a,s') \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} \times \mathcal{S}_k^{\mathsf{T}}$

$$\begin{aligned}
\underbrace{p(s'|s,a)}_{\text{transition probability in } M^*} &\leq \underbrace{\widehat{p}_k(s'|s,a)}_{=0} + \beta_k^{sas'} = \underbrace{\sqrt{\frac{14\widehat{\sigma}_{p,k}^2(s'|s,a)\ln(2SAt_k/\delta)}{N_k^+(s,a)}}}_{=0} + \frac{49\ln(2SAt_k/\delta)}{3N_k^{\pm}(s,a)} \\
&\leq \frac{49\ln(2SAt_k/\delta)}{3N_k^{\pm}(s,a)} < \frac{1}{D^{\mathsf{C}} S_k^{\mathsf{T}}}
\end{aligned}$$

where we have exploited the fact that $\widehat{p}(s'|s,a) = 0$ and $\widehat{\sigma}_{p,k}^2(s'|s,a) = 0$ for any state $s' \in \mathcal{S}_k^{\mathsf{T}}$ (remember that $N_k(s,a,s') = 0$).

We denote by $\tau_{M^*}(s \to s')$ the shortest path between any pair of states $(s,s') \in \mathcal{S} \times \mathcal{S}$ in the true MDP M^*. Fix an arbitrary target state $\overline{s} \in \mathcal{S}_k^{\mathsf{T}}$ and denote by $\tau(s) := \tau_{M^*}(s \to \overline{s})$ and $\tau_{\min} := \min_{s \in \mathcal{S}_k^{\mathsf{C}}}\{\tau(s)\}$. We have

$$\begin{aligned}
&\tau(\overline{s}) = 0 \\
\forall s \in \mathcal{S}_k^{\mathsf{C}} \quad &\tau(s) = 1 + \min_{a \in \mathcal{A}_s}\left\{\sum_{s' \in \mathcal{S}} \underbrace{p(s'|s,a)\tau(s')}_{\geq 0}\right\} \geq 1 + \min_{a \in \mathcal{A}_s}\left\{\sum_{s' \in \mathcal{S}_k^{\mathsf{C}}} p(s'|s,a) \underbrace{\tau(s')}_{\geq \tau_{\min}}\right\} \\
&\geq 1 + \tau_{\min} \cdot \min_{a \in \mathcal{A}}\left\{\sum_{s' \in \mathcal{S}_k^{\mathsf{C}}} p(s'|s,a)\right\} = 1 + \tau_{\min} \cdot \min_{a \in \mathcal{A}}\left\{1 - \sum_{s' \in \mathcal{S}_k^{\mathsf{T}}} \underbrace{p(s'|s,a)}_{< \frac{1}{D^{\mathsf{C}} S_k^{\mathsf{T}}}}\right\} \\
&> 1 + \tau_{\min}\left(1 - \sum_{s' \in \mathcal{S}_k^{\mathsf{T}}} \frac{1}{D^{\mathsf{C}} S_k^{\mathsf{T}}}\right) = 1 + \tau_{\min}\left(1 - \frac{1}{D^{\mathsf{C}}}\right)
\end{aligned}$$

Applying the above inequality to $\widetilde{s} \in \mathcal{S}_k^{\mathsf{C}}$ achieving $\tau(\widetilde{s}) = \tau_{\min}$ yields $\tau_{\min} > D^{\mathsf{C}}$. This implies that the shortest path in M^* between any state $s \in \mathcal{S}_k^{\mathsf{C}} \subseteq \mathcal{S}^{\mathsf{C}}$ and any state in $\overline{s} \in \mathcal{S}_k^{\mathsf{T}}$ is strictly bigger than D^{C} but by definition D^{C} is the longest shortest path between any pair of states in $\mathcal{S}^{\mathsf{C}}$. Therefore, $\overline{s} \in \mathcal{S}^{\mathsf{T}}$. Since $\overline{s} \in \mathcal{S}_k^{\mathsf{T}}$ was chosen arbitrarily, then $\mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$. □

As a consequence of Lem. 4, under the assumptions that $M^* \in \mathcal{M}_k$, $\mathcal{S}_k^{\mathsf{T}} \neq \emptyset$ and (13) holds, there are only two possible cases:

1. Either $\mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$,
2. or $\exists (s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} \ : \ N_k^{\pm}(s,a) \leq \sqrt{\frac{t_k}{SA}}$.

Case 1: $\mathcal{S}_k^{\mathsf{T}} = \mathcal{S}^{\mathsf{T}}$ implies that $M^* \in \overline{\mathcal{M}}_k^+$. This is because for any $(s,a,s') \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} \times \mathcal{S}_k^{\mathsf{T}}$ we have $p(s'|s,a) = \widetilde{p}_k(s'|s,a) = 0$ and for any $(s,a,s') \notin \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} \times \mathcal{S}_k^{\mathsf{T}}$ we have $|p(s'|s,a) - \widehat{p}_k(s'|s,a)| \leq \beta_{p,k}^{sas'}$ and so $p(\cdot|s,a) \in \overline{B}_{p,k}^+(s,a)$. Since $M^* \in \overline{\mathcal{M}}_k^+$, we can use the same argument as Jaksch et al. [2, Sec. 4.3 & Theorem 7] to prove $\widetilde{g}_k \geq g^* - \frac{r_{\max}}{\sqrt{t_k}}$.

Case 2: For any $(s,a) \in \mathcal{S}_k^{\mathsf{T}} \times \mathcal{A}$, $\overline{B}_{p,k}^+(s,a) = \mathcal{C}$ is the $(S-1)$-simplex denoting the maximal uncertainty about the transition probabilities, and $B_{r,k}(s,a) = [0, r_{\max}]$. We will now construct an MDP $M' \in \overline{\mathcal{M}}_k^+$ with optimal gain $r_{\max}$. For all $(s,a) \in \mathcal{S}_k^{\mathsf{T}} \times \mathcal{A}$, we set the transitions to $p_{M'}(s|s,a) = 1$ and rewards to $r_{M'}(s,a) = r_{\max}$. Let $(\overline{s}, \overline{a}) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}$ such that $N_k^{\pm}(\overline{s}, \overline{a}) \leq \sqrt{\frac{t_k}{SA}}$ (which exists by assumption). We set $p_{M'}(s'|\overline{s}, \overline{a}) > 0$ for all $s' \in \mathcal{S}_k^{\mathsf{T}}$. This is possible because by definition of $\overline{\mathcal{M}}_k^+$, the support of $p(\cdot|\overline{s}, \overline{a})$ is not restricted to $\mathcal{S}_k^{\mathsf{C}}$. Finally, for all state-action pairs $(s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}$, we set $p_{M'}(\overline{s}|s,a) > 0$. This is possible because by definition of $\overline{\mathcal{M}}_k^+$, the support of $p(\cdot|s,a)$ is only restricted to $\mathcal{S}_k^{\mathsf{C}}$ and $\overline{s} \in \mathcal{S}_k^{\mathsf{C}}$. In M', for all policies, all states in $\mathcal{S}_k^{\mathsf{T}}$ are absorbing states (i.e., loop on themselves with probability 1) with maximal reward $r_{\max}$ and all other states $s \in \mathcal{S}_k^{\mathsf{C}}$ are transient. The optimal gain of M' is thus $r_{\max}$ and since $M' \in \overline{\mathcal{M}}_k^+$ we conclude that $\widetilde{g}_k \geq r_{\max} - \frac{r_{\max}}{\sqrt{t_k}} \geq g^* - \frac{r_{\max}}{\sqrt{t_k}}$.

In conclusion, TEVI is always returning an *optimistic* policy when the assumptions of Lem. 4 hold. The regret Δ_k accumulated in episode k can thus be upper-bounded as:

$$\Delta_k = \sum_{s,a} \nu_k(s,a)(g^* - r(s,a)) = \sum_{s,a} \nu_k(s,a)(\underbrace{g^*}_{\leq \widetilde{g}_k + \frac{r_{\max}}{\sqrt{t_k}}} - \widetilde{r}_k(s,a)) + \sum_{s,a} \nu_k(s,a)(\widetilde{r}_k(s,a) - r(s,a))$$

$$\leq \underbrace{\sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a))}_{:= \widetilde{\Delta}_k} + \sum_{s,a} \nu_k(s,a)(\widetilde{r}_k(s,a) - r(s,a)) + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}$$

To bound the difference between the optimistic reward $\widetilde{r}_k$ and the true reward r we introduce the estimated reward $\widehat{r}_k$:

$$\forall s,a \in \mathcal{S} \times \mathcal{A}, \ \ \widetilde{r}_k(s,a) - r(s,a) = \underbrace{\widetilde{r}_k(s,a) - \widehat{r}_k(s,a)}_{\leq \beta_{r,k}^{sa} \text{ by construction}} + \underbrace{\widehat{r}_k(s,a) - r(s,a)}_{\leq \beta_{r,k}^{sa} \text{ since } M \in \mathcal{M}_k} \leq 2\beta_{r,k}^{sa}$$

and so in conclusion:

$$\Delta_k \leq \widetilde{\Delta}_k + \underbrace{2\sum_{s,a} \nu_k(s,a)\beta_{r,k}^{sa} + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}}_{:= U_k^1} \tag{14}$$

D.3 Bounding $\widetilde{\Delta}_k$

The goal of this section is to bound the term $\widetilde{\Delta}_k := \sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a))$. We start by discarding the state-action pairs $(s,a) \in \mathcal{K}_k$ that have been poorly visited so far:

$$\widetilde{\Delta}_k = \sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a)) \underbrace{\mathbb{1}\{(s,a) \notin \mathcal{K}_k\}}_{:= \mathbb{1}_k(s,a)} + \sum_{s,a} \nu_k(s,a) \underbrace{(\widetilde{g}_k - \widetilde{r}_k(s,a))}_{\leq r_{\max}} \mathbb{1}\{(s,a) \in \mathcal{K}_k\}$$

$$\leq \underbrace{\sum_{s,a} \nu_k(s,a)(\widetilde{g}_k - \widetilde{r}_k(s,a)) \mathbb{1}_k(s,a)}_{:= \widetilde{\Delta}'_k} + r_{\max} \sum_{s,a} \nu_k(s,a) \mathbb{1}\{(s,a) \in \mathcal{K}_k\} \tag{15}$$

We will now bound the term $\widetilde{\Delta}'_k = \sum_s \nu_k(s, \widetilde{\pi}_k(s))(\widetilde{g}_k - \widetilde{r}_k(s, \widetilde{\pi}_k(s)))\mathbb{1}_k(s, \widetilde{\pi}_k(s))$. We recall that the policy $\widetilde{\pi}_k$ is obtained by executing TEVI$(0, \mathcal{M}_k^\diamond, \mathcal{S}_k^{\mathrm{EVI}}, \varepsilon_k)$ (see Alg. 1) where $\varepsilon_k := r_{\max}/\sqrt{t_k}$ and $\mathcal{S}_k^{\mathrm{EVI}}$ and $\mathcal{M}_k^\diamond$ are defined in (12). In all possible cases for both $\mathcal{S}_k^{\mathrm{EVI}}$ and $\mathcal{M}_k^\diamond$, this amounts to applying value iteration to a *communicating* MDP with finite state space $\mathcal{S}_k^{\mathrm{EVI}}$ and *compact* action space. By [21, Thm. 8.5.6], since the convergence criterion of value iteration is met we have:

$$\forall s \in \mathcal{S}_k^{\mathrm{EVI}}, \quad \left| \widetilde{h}_k(s) + \widetilde{g}_k - \widetilde{r}_k(s, \widetilde{\pi}_k(s)) - \sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) \widetilde{h}_k(s') \right| \leq \frac{r_{\max}}{\sqrt{t_k}} \tag{16}$$

For all $s \notin \mathcal{S}_k^{\mathsf{C}}$, $\nu_k(s, \widetilde{\pi}_k(s)) = 0$ due to the stopping condition of episode k. Therefore we can plug (16) in $\widetilde{\Delta}'_k$ and derive an upper bound restricted to the set $\mathcal{S}_k^{\mathsf{C}} \subseteq \mathcal{S}_k^{\mathrm{EVI}}$. Before to do that, we further decompose $\widetilde{\Delta}'_k$ as:

$$\begin{aligned} \widetilde{\Delta}'_k &\leq \sum_s \nu_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) \widetilde{h}_k(s') - \widetilde{h}_k(s) + \frac{r_{\max}}{\sqrt{t_k}} \right) \mathbb{1}_k(s, \widetilde{\pi}_k(s)) \\ &= \nu'_k \left(\widetilde{P}_k - I \right) \widetilde{h}_k + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}} \mathbb{1}_k(s,a) \end{aligned} \tag{17}$$

where $\nu'_k = (\nu_k(s, \widetilde{\pi}_k(s))\mathbb{1}_k(s, \widetilde{\pi}_k(s)))_{s \in \mathcal{S}}$ is the vector of visit counts for each state and the corresponding action chosen by $\widetilde{\pi}_k$ multiplied by the indicator function $\mathbb{1}_k$, $\widetilde{P}_k = (\widetilde{P}_k(s'|s, \widetilde{\pi}_k(s)))_{s,s' \in \mathcal{S}}$ is transition matrix associated to $\widetilde{\pi}_k$ in $\overline{\mathcal{M}}_k^+$ and I is the identity matrix. We now focus on the term $\nu'_k(\widetilde{P}_k - I)\widetilde{h}_k$. Since the rows of $\widetilde{P}_k$ sum to 1, $\forall \lambda \in \mathbb{R}$, $\big(\widetilde{P}_k - I\big)\widetilde{h}_k = \big(\widetilde{P}_k - I\big)\big(\widetilde{h}_k + \lambda e\big)$ where $e = (1, \dots 1)^\intercal$ is the vector of all ones. Let's take $\lambda := -\min_{s \in \mathcal{S}_k^{\mathsf{c}}} \{\widetilde{h}_k(s)\}$ and define $w_k := \widetilde{h}_k + \lambda e$ so that for all $s \in \mathcal{S}_k^{\mathsf{C}}$, $w_k(s) \geq 0$ and $\min_{s \in \mathcal{S}_k^{\mathsf{c}}} \{w_k(s)\} = 0$. We have:

$$\nu'_k(\widetilde{P}_k - I)\widetilde{h}_k = \sum_{s \in \mathcal{S}} \nu_k(s, \widetilde{\pi}_k(s))\mathbb{1}_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \right)$$

We denote by $k_t := \sup\{k \geq 1 : \ t_k \leq t\}$ the current episode at time t. Whenever $s_t \in \mathcal{S}_{k_t}^{\mathsf{T}}$, episode k_t stops before executing any action (see the stopping condition of TUCRL in Fig. 2) implying that $\forall s \in \mathcal{S}_k^{\mathsf{T}}$, $\nu_k(s, \widetilde{\pi}_k(s)) = 0$. Therefore we have:

$$\nu'_k(\widetilde{P}_k - I)\widetilde{h}_k = \sum_{s \in \mathcal{S}_k^{\mathsf{c}}} \nu_k(s, \widetilde{\pi}_k(s))\mathbb{1}_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \right)$$

For all states s such that $\mathbb{1}_k(s, \widetilde{\pi}_k(s)) = 1$, i.e., satisfying $N_k^{\pm}(s, \widetilde{\pi}_k(s)) > \sqrt{t_k/SA}$, we force TEVI to set $\widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) = 0$, $\forall s' \in \mathcal{S}_k^{\mathsf{T}}$, by construction of $\overline{\mathcal{M}}_k^+$ so that:

$$\nu'_k(\widetilde{P}_k - I)\widetilde{h}_k = \sum_{s \in \mathcal{S}_k^{\mathsf{c}}} \nu_k(s, \widetilde{\pi}_k(s))\mathbb{1}_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}_k^{\mathsf{c}}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \right) \tag{18}$$

We can now introduce p:

$$\sum_{s' \in \mathcal{S}_k^{\mathsf{c}}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) = \sum_{s' \in \mathcal{S}_k^{\mathsf{c}}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) w_k(s') - p(s'|s, \widetilde{\pi}_k(s)) w_k(s') \tag{19}$$

$$+ \left(\sum_{s' \in \mathcal{S}_k^{\mathsf{c}}} p(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \right) \tag{20}$$

By definition $\mathcal{S}_k^{\mathsf{C}} \subseteq \mathcal{S}^{\mathsf{C}}$ and using $(1, \infty)$-Hölder's inequality , the term (19) can be bounded as (19) $\leq \|\widetilde{p}_k(\cdot|s, \widetilde{\pi}_k(s)) - p(\cdot|s, \widetilde{\pi}_k(s))\|_{1, \mathcal{S}^{\mathsf{c}}} \cdot \max_{s' \in \mathcal{S}_k^{\mathsf{c}}} \{w_k(s')\}$ where for any vector $v \in \mathbb{R}^{\mathcal{S}}$, $\|v\|_{1, \mathcal{S}^{\mathsf{c}}} := \sum_{s \in \mathcal{S}^{\mathsf{c}}} |v(s)|$. Define $\overline{s} \in \arg\max_{s \in \mathcal{S}_k^{\mathsf{c}}} \{w_k(s)\}$ and $\widetilde{s} \in \arg\min_{s \in \mathcal{S}_k^{\mathsf{c}}} \{w_k(s)\}$. By definition $\overline{s}, \widetilde{s} \in \mathcal{S}_k^{\mathsf{C}}$ and $w_k(\widetilde{s}) = \min_{s \in \mathcal{S}_k^{\mathsf{c}}} \{w_k(s)\} = 0$. By Lem. 7, we know that for all $s, s' \in \mathcal{S}_k^{\mathsf{C}}$,

the difference $w_k(s') - w_k(s) = \widetilde{h}_k(s') - \widetilde{h}_k(s)$ is upper bounded by $r_{\max} \cdot \tau_{\overline{\mathcal{M}}_k^+}(s \to s')$. We also know by Lem. 6 that for all $s, s' \in \mathcal{S}_k^{\mathsf{C}}$, $\tau_{\mathcal{M}_k^+}(s \to s') = \tau_{\overline{\mathcal{M}}_k^+}(s \to s')$. Since $M^* \in \mathcal{M}_k^+$ (M^* is the true MDP), we also have that for all $s, s' \in \mathcal{S}_k^{\mathsf{C}} \subseteq \mathcal{S}^{\mathsf{C}}$, $\tau_{\mathcal{M}_k^+}(s \to s') \leq \tau_{M^*}(s \to s') \leq D^{\mathsf{C}}$. In conclusion, $\forall s, s' \in \mathcal{S}_k^{\mathsf{C}}$, $w_k(s') - w_k(s) \leq r_{\max} D^{\mathsf{C}}$ and in particular $\max_{s' \in \mathcal{S}_k^{\mathsf{C}}}\{w_k(s')\} = w_k(\overline{s}) = w_k(\overline{s}) - w_k(\widetilde{s}) \leq r_{\max} D^{\mathsf{C}}$. Similarly to what we did to bound $|\widetilde{r}_k - r|$ (14), we bound the distance in ℓ_1-norm between $\widetilde{p}_k$ and p by introducing $\widehat{p}_k$:

$$\|\widetilde{p}_k - p\|_{1,\mathcal{S}^{\mathsf{C}}} \leq \|\widetilde{p}_k - \widehat{p}_k\|_{1,\mathcal{S}^{\mathsf{C}}} + \|\widehat{p}_k - p\|_{1,\mathcal{S}^{\mathsf{C}}} \leq 2 \left(\sum_{s' \in \mathcal{S}^{\mathsf{C}}} \beta_{p,k}^{s\widetilde{\pi}_k(s)s'} \right) \tag{21}$$

We now bound the contribution of the term (20). Jaksch et al. [2] decompose this term into a martingale difference sequence and a telescopic sum but due to the indicator function $\mathbb{1}_k$, in our case the sum is not telescopic anymore and an additional term appears.

$$\begin{aligned}
(20) &= \sum_{s \in \mathcal{S}} \nu_k(s, \widetilde{\pi}_k(s)) \mathbb{1}_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}_k^{\mathsf{C}}} \underbrace{p(s'|s, \widetilde{\pi}_k(s)) w_k(s')}_{\geq 0,\ \forall s' \in \mathcal{S}_k^{\mathsf{C}}} \mathbb{1}\{s \in \mathcal{S}_k^{\mathsf{C}}\} - w_k(s) \mathbb{1}\{s \in \mathcal{S}_k^{\mathsf{C}}\} \right) \\
&\leq \sum_{s \in \mathcal{S}} \nu_k(s, \widetilde{\pi}_k(s)) \mathbb{1}_k(s, \widetilde{\pi}_k(s)) \left(\sum_{s' \in \mathcal{S}_k^{\mathsf{C}}} p(s'|s, \widetilde{\pi}_k(s)) w_k(s') - w_k(s) \mathbb{1}\{s \in \mathcal{S}_k^{\mathsf{C}}\} \right) \\
&= \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{s' \in \mathcal{S}_k^1} p(s'|s_t, \widetilde{\pi}_k(s_t)) w_k(s') - w_k(s_t) \mathbb{1}\{s_t \in \mathcal{S}_k^{\mathsf{C}}\} \right) \mathbb{1}_k(s_t, \widetilde{\pi}_k(s_t)) \\
&= \sum_{t=t_k}^{t_{k+1}-1} \underbrace{\left(\sum_{s' \in \mathcal{S}_k^1} p(s'|s_t, \widetilde{\pi}_k(s_t)) w_k(s') - w_k(s_{t+1}) \mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{C}}\} \right) \mathbb{1}_k(s_t, \widetilde{\pi}_k(s_t))}_{:= X_t} \qquad (22) \\
&\quad + \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \left(w_k(s_{t+1}) \mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{C}}\} - w_k(s_t) \mathbb{1}\{s_t \in \mathcal{S}_k^{\mathsf{C}}\} \right) \mathbb{1}_k(s_t, \widetilde{\pi}_k(s_t))}_{\text{not telescopic due to } \mathbb{1}_k!} \qquad (23)
\end{aligned}$$

Define the filtration $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \ldots, s_{t+1})$. Since k_t is $\mathcal{F}_{t-1}$-measurable:

$$\mathbb{E}\left[w_{k_t}(s_{t+1}) \mathbb{1}\{s_{t+1} \in \mathcal{S}_{k_t}^{\mathsf{C}}\} \mathbb{1}_{k_t}(s_t, \widetilde{\pi}_{k_t}(s_t)) | \mathcal{F}_{t-1} \right] = \underbrace{\sum_{s' \in \mathcal{S}_{k_t}^{\mathsf{C}}} p(s'|s_t, \widetilde{\pi}_{k_t}(s_t)) w_{k_t}(s') \mathbb{1}_{k_t}(s_t, \widetilde{\pi}_{k_t}(s_t))}_{\mathcal{F}_{t-1}-\text{measurable}}$$

implying $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$ and so $(X_t, \mathcal{F}_t)_{t \geq 1}$ is a martingale difference sequence (MDS) with $|X_t| \leq r_{\max} D^{\mathsf{C}}$. We will bound (22) in the next section (Sec. D.4) using Azuma's inequality. Using the fact that $\mathbb{1}_k(s_t, \widetilde{\pi}_k(s_t)) = \mathbb{1}\{(s_t, \widetilde{\pi}_k(s_t)) \notin \mathcal{K}_k\} = 1 - \mathbb{1}\{(s_t, \widetilde{\pi}_k(s_t)) \in \mathcal{K}_k\}$ we can make a telescopic sum appear and rewrite (23) as:

$$\begin{aligned}
(23) &= \underbrace{\sum_{t=t_k}^{t_{k+1}-1} w_k(s_{t+1}) \mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{C}}\} - w_k(s_t) \mathbb{1}\{s_t \in \mathcal{S}_k^{\mathsf{C}}\}}_{= w_k(s_{t_{k+1}}) \mathbb{1}\{s_{t_{k+1}} \in \mathcal{S}_k^{\mathsf{C}}\} - w_k(s_{t_k}) \mathbb{1}\{s_{t_k} \in \mathcal{S}_k^{\mathsf{C}}\} \leq r_{\max} D^{\mathsf{C}}} \quad \text{(telescopic sum)} \\
&\quad + \sum_{t=t_k}^{t_{k+1}-1} \underbrace{\left(w_k(s_t) \mathbb{1}\{s_t \in \mathcal{S}_k^{\mathsf{C}}\} - w_k(s_{t+1}) \mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{C}}\} \right)}_{\leq r_{\max} D^{\mathsf{C}}} \mathbb{1}\{(s_t, \widetilde{\pi}_k(s_t)) \in \mathcal{K}_k\} \\
&\leq r_{\max} D^{\mathsf{C}} + r_{\max} D^{\mathsf{C}} \sum_{s,a} \nu_k(s,a) \mathbb{1}\{(s_t, \widetilde{\pi}_k(s_t)) \in \mathcal{K}_k\} \qquad (24)
\end{aligned}$$

By gathering (15), (17), (21), (22) and (24) we obtain the following bound for $\widetilde{\Delta}_k$:

$$\begin{aligned}
\widetilde{\Delta}_k &\leq 2r_{\max}D^{\complement}\sum_{s,a}\sum_{s'\in\mathcal{S}^{\complement}}\underbrace{\mathbb{1}_k(s,a)}_{\leq 1}\underbrace{\nu_k(s,a)\beta_{p,k}^{sas'}}_{\geq 0}+\sum_{t=t_k}^{t_{k+1}-1}X_t+r_{\max}D^{\complement} \\
&\quad + r_{\max}(D^{\complement}+1)\sum_{s,a}\nu_k(s,a)\mathbb{1}\{N_k^{\pm}(s,a)\leq\sqrt{t_k/SA}\}+r_{\max}\sum_{s,a}\underbrace{\frac{\nu_k(s,a)}{\sqrt{t_k}}}_{\geq 0}\underbrace{\mathbb{1}_k(s,a)}_{\leq 1} \\
&\leq 2r_{\max}D^{\complement}\sum_{s,a}\sum_{s'\in\mathcal{S}^{\complement}}\nu_k(s,a)\beta_{p,k}^{sas'}+r_{\max}(D^{\complement}+1)\sum_{s,a}\nu_k(s,a)\mathbb{1}\{(s,a)\in\mathcal{K}_k\} \\
&\quad + \sum_{t=t_k}^{t_{k+1}-1}X_t+r_{\max}D^{\complement}+r_{\max}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}} := U_k^2
\end{aligned} \tag{25}$$

D.4 Summing over episodes with $M^*\in\mathcal{M}_k$ and $t_k\geq C(k)$

Denote by $\mathbb{1}(k) := \mathbb{1}\{t_k\geq C(k)\}\cdot\mathbb{1}\{M^*\in\mathcal{M}_k\}$ the indicator function taking value 1 only when both $M^*\in\mathcal{M}_k$ and $t_k\geq C(k)$. By gathering (14) and (25) we obtain:

$$\sum_{k=1}^{m}\Delta_k\cdot\mathbb{1}(k)\leq\sum_{k=1}^{m}\underbrace{U_k^1}_{\geq 0}\cdot\underbrace{\mathbb{1}(k)}_{\leq 1}+\sum_{k=1}^{m}\underbrace{U_k^2}_{\geq 0}\cdot\underbrace{\mathbb{1}(k)}_{\leq 1}\leq\sum_{k=1}^{m}U_k^1+U_k^2 \tag{26}$$

and so

$$\begin{aligned}
\sum_{k=1}^{m}\Delta_k\cdot\mathbb{1}(k) \leq & 2\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\left(r_{\max}D^{\complement}\sum_{s'\in\mathcal{S}^{\complement}}\beta_{p,k}^{sas'}+\beta_{r,k}^{s,a}\right)+2r_{\max}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}} \\
& + r_{\max}(D^{\complement}+1)\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\mathbb{1}\{(s,a)\in\mathcal{K}_k\}+\sum_{t=1}^{T}X_t\mathbb{1}(k_t)+r_{\max}mD^{\complement}
\end{aligned} \tag{27}$$

We will now upper-bound the terms appearing in (27). The main novelty of (27) compared to UCRL is the term $\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\mathbb{1}\{(s,a)\in\mathcal{K}_k\}$ which is not present in the proof of Jaksch et al. [2]. We will show in the next section that this term is bounded by $O(\sqrt{\mathcal{S}^{\complement}AT})$. All the other terms are similar to those found in UCRL.

D.4.1 Poorly visited state-action pairs

We first notice that by definition $t_{k_t}\leq t$ where $k_t := \sup\{k\geq 1:\ \ t_k\leq t\}$ is the current episode at time t. As a result,

$$\mathbb{1}\left\{(s,a)\in\mathcal{K}_{k_t}\right\} := \mathbb{1}\left\{N_{k_t}^{\pm}(s_t,a_t)\leq\sqrt{t_{k_t}/SA}\right\}\leq\mathbb{1}\left\{N_{k_t}^{\pm}(s_t,a_t)\leq\sqrt{t/SA}\right\}$$

Instead of directly bounding $\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\mathbb{1}\{(s,a)\in\mathcal{K}_k\}$ we will bound the number of visits Z_T in state-action pairs that have been visited less than $\sqrt{t/SA}$ times

$$Z_T := \sum_{t=1}^{T}\mathbb{1}\left\{N_{k_t}^{\pm}(s_t,a_t)\leq\sqrt{t/SA}\right\}\geq\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\mathbb{1}\left\{(s,a)\in\mathcal{K}_k\right\}$$

Note that the quantity $N_k(s,a)$ is updated only after the end of episode k and the stopping condition of episodes used by TUCRL implies that (see Fig. 2):

$$\forall k\geq 1,\ \forall(s,a)\in\mathcal{S}\times\mathcal{A},\ \ \nu_k(s,a)\leq N_k^{+}(s,a) \tag{28}$$

Moreover, for all $(s,a)\notin\mathcal{S}^{\complement}\times\mathcal{A}$, $\nu_k(s,a)=0$ implying that only the states $s\in S^{\complement}$ should be considered in the above sums. Using (28), we prove the following lemma:

Lemma 5. *For any $T \geq 1$ and any sequence of states and actions $\{s_1, a_1, \ldots\ldots s_T, a_T\}$ we have:*

$$Z_T \leq 2\sqrt{S^{\mathsf{C}} A T} + 2S^{\mathsf{C}} A.$$

Proof. For any episode k starting at time t_k, and for any state-action pair (s, a) we recall that $N_k(s, a)$ denotes the number of visits in (s, a) prior to episode k (k not included) and by $\nu_k(s, a)$ the number of visits in (s, a) during episode k:

$$N_k(s,a) := \sum_{t=1}^{t_k - 1} \mathbb{1}\{(s_t, a_t) = (s, a)\} \text{ and } \nu_k(s,a) := \sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}\{(s_t, a_t) = (s, a)\}$$

and so ${}_k(s, a) = \sum_{i=1}^{k-1} \nu_i(s, a)$. By convention, we denote by $N_{k_T+1}(s, a) := \sum_{t=1}^{T} \mathbb{1}\{(s_t, a_t) = (s, a)\}$ the total number of visits in (s, a) after T time steps (T included). We first decompose Z_T as:

$$Z_T := \sum_{s,a} \sum_{t=1}^{T} \mathbb{1}\Big\{ \max\{1, N_{k_t}(s,a) - 1\} \leq \sqrt{t/SA} \Big\} \cdot \mathbb{1}\Big\{(s_t, a_t) = (s, a)\Big\} = \sum_{s \in S^{\mathsf{C}}} \sum_{a} Z_T(s, a)$$

$$\text{where } Z_T(s,a) := \sum_{t=1}^{T} \mathbb{1}\Big\{ \max\{1, N_{k_t}(s,a) - 1\} \leq \sqrt{t/SA} \Big\} \cdot \mathbb{1}\Big\{(s_t, a_t) = (s, a)\Big\}$$

Using the fact that for all $t \geq 1$, $t_{k_t} \leq t \leq t_{k_t+1} - 1$ we have:

$$\begin{aligned}
\forall T \geq \tau \geq 1, \;\; Z_\tau(s,a) &= \sum_{t=1}^{\tau} \underbrace{\mathbb{1}\Big\{ \max\{1, N_{k_t}(s,a) - 1\} \leq \sqrt{t/SA} \Big\}}_{\leq 1} \cdot \underbrace{\mathbb{1}\{(s_t, a_t) = (s, a)\}}_{\geq 0} \\
&\leq \sum_{t=1}^{\tau} \mathbb{1}\{(s_t, a_t) = (s, a)\} \leq \sum_{t=1}^{t_{k_\tau+1}-1} \mathbb{1}\{(s_t, a_t) = (s, a)\} \\
&= N_{k_\tau+1}(s,a) \qquad (29)
\end{aligned}$$

Let's define $t_{s,a}$ as the last time that $Z_t(s, a)$ was incremented by 1:

$$\begin{aligned}
t_{s,a} &:= \max\Big\{ T \geq t \geq 1 : \max\{1, N_{k_t}(s,a) - 1\} \leq \sqrt{t/SA} \text{ and } (s_t, a_t) = (s, a) \Big\} \\
&= \min\Big\{ T \geq t \geq 1 : Z_t(s,a) = Z_T(s,a) \Big\}
\end{aligned}$$

We denote by $m_{s,a} := k_{t_{s,a}}$ the corresponding episode. By definition,

$$Z_T(s,a) = Z_{t_{s,a}}(s,a) \qquad (30)$$

and

$$\max\{1, N_{m_{s,a}}(s,a) - 1\} \leq \sqrt{t_{s,a}/SA} \qquad (31)$$

Using (29) with $\tau = t_{s,a}$ we obtain:

$$Z_{t_{s,a}} \leq N_{m_{s,a}+1}(s,a) \qquad (32)$$

Moreover, by definition of $N_k(s, a)$ and (28):

$$\begin{aligned}
N_{m_{s,a}+1}(s,a) &= N_{m_{s,a}}(s,a) + \underbrace{\nu_{m_{s,a}}(s,a)}_{\leq N^+_{m_{s,a}}(s,a)} \leq 2 \underbrace{\max\{1, N_{m_{s,a}}(s,a)\}}_{\leq \max\{1, N_{m_{s,a}}(s,a) - 1\} + 1} \\
\implies N_{m_{s,a}+1}(s,a) &\leq 2 \cdot \max\{1, N_{m_{s,a}}(s,a) - 1\} + 2 \qquad (33)
\end{aligned}$$

Gathering (30), (31), (32), and (33) we obtain:

$$\begin{aligned}
Z_T(s,a) = Z_{t_{s,a}}(s,a) &\leq \max\{1, N_{m_{s,a}+1}(s,a) - 1\} + 1 \leq 2 \cdot \max\{1, N_{m_{s,a}}(s,a) - 1\} + 2 \\
&\leq 2\sqrt{t_{s,a}/SA} + 2 \\
&\leq 2\sqrt{T/SA} + 2 \\
\implies Z_T = \sum_{s \in S^{\mathsf{C}}} \sum_{a} Z_T(s,a) &\leq 2\sqrt{S^{\mathsf{C}} A T} + 2S^{\mathsf{C}} A
\end{aligned}$$

where for the last inequality we used the fact that $S^{\mathsf{C}} \leq S$ (by definition) implying $S^{\mathsf{C}}/\sqrt{S} = \sqrt{S^{\mathsf{C}}/S} \cdot \sqrt{S^{\mathsf{C}}} \leq \sqrt{S^{\mathsf{C}}}$. This concludes the proof. □

As a consequence of Lem. 5:

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\mathbb{1}\{N_k^{\pm}(s,a)\leq\sqrt{t_k/S^{\mathtt{C}}A}\}\leq Z_T\leq 2\sqrt{S^{\mathtt{C}}AT}+2S^{\mathtt{C}}A \tag{34}$$

D.4.2 Confidence bounds $\beta_{r,k}^{sa}$ and $\beta_{p,k}^{sas'}$

Since (28) holds, Lemma 19 of Jaksch et al. [2] can still be applied. Moreover, exploiting again the fact that for all $(s,a)\notin\mathcal{S}^{\mathtt{C}}\times\mathcal{A}$, $\nu_k(s,a)=0$ we obtain

$$\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}}\leq\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{N_k^{+}(s,a)}}\leq\left(\sqrt{2}+1\right)\sqrt{S^{\mathtt{C}}AT} \tag{35}$$

and as shown in [6, Appendix F.7] (with the difference that S is restricted to $S^{\mathtt{C}}$) we have:

$$\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{N_k^{\pm}(s,a)}\leq 6S^{\mathtt{C}}A+2S^{\mathtt{C}}A\ln(T) \tag{36}$$

The terms $\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa}$ and $\sum_{k=1}^{m}\sum_{s,a,s'\in\mathcal{S}^{\mathtt{C}}}\nu_k(s,a)\beta_{p,k}^{sas'}$ can then be bounded exactly as in [6, App. F.7] with S replaced by $S^{\mathtt{C}}$ (except in the logarithm).

D.4.3 Number of episodes

The stopping condition of episodes used by TUCRL (see Fig. 2) combines the original stopping condition of UCRL with the condition $s_t\in\mathcal{S}_{k_t}^{\mathtt{T}}$. Using only inequality (28), Jaksch et al. [2, Figure 1] proved that for any any sequence $\{s_1,a_1,\ldots,s_T,a_T\}$, the number of episodes is bounded by $1+2SA+SA\log_2\left(\frac{T}{SA}\right)$. Since (28) also holds in our case, the total number of episodes m after T time steps can be bounded by the same quantity (with S replaced by $S^{\mathtt{C}}$ since sates in $\mathcal{S}^{\mathtt{T}}$ will never be visited) plus the number of times the event $s_t\in\mathcal{S}_{k_t}^{\mathtt{T}}$ occurs. Since whenever $s_t\in\mathcal{S}_{k_t}^{\mathtt{T}}$ state s_t is removed from $\mathcal{S}_{k_t+1}^{\mathtt{T}}$ and s_t necessarily belongs to $\mathcal{S}^{\mathtt{C}}$ (by definition), this event can happen at most $\mathcal{S}^{\mathtt{C}}$ times. By Proposition 18 in [2] we thus have:

$$m\leq 1+2S^{\mathtt{C}}A+S^{\mathtt{C}}A\log_2\left(\frac{T}{S^{\mathtt{C}}A}\right)+S^{\mathtt{C}} \tag{37}$$

D.4.4 Martingale Difference Sequence $X_t\cdot\mathbb{1}(k_t)$

In Sec. D.3 we already proved that $(X_t,\mathcal{F}_t)_{t\geq1}$ is an MDS i.e., for all $t\geq1$, $\mathbb{E}[X_t|\mathcal{F}_{t-1}]=0$. Since k_t is $\mathcal{F}_{t-1}$-measurable, we also have $\mathbb{E}[X_t\mathbb{1}(k_t)|\mathcal{F}_{t-1}]=\mathbb{1}(k_t)\cdot\mathbb{E}[X_t|\mathcal{F}_{t-1}]=0$ with $|X_t\mathbb{1}(k_t)|\leq r_{\max}D^{\mathtt{C}}$. Therefore, $(X_t\mathbb{1}(k_t),\mathcal{F}_t)_{t\geq1}$ is also an MDS. By Azuma's inequality (see for example [2, Lemma 10]):

$$\sum_{t=1}^{T}X_t\mathbb{1}(k_t)\leq r_{\max}D^{\mathtt{C}}\sqrt{\frac{5}{2}T\ln\left(\frac{8T}{\delta}\right)}\ \text{ w.p.}\geq 1-\frac{\delta}{12T^{5/4}} \tag{38}$$

D.5 Completing the regret bound

By gathering (27), (34), (35), (36), (38) and (37) we conclude that with probability at least $1-\frac{\delta}{12T^{5/4}}$:

$$\begin{aligned}\sum_{k=1}^{m}\Delta_k\cdot\mathbb{1}(k)\leq\;& 2\left(\sqrt{28}+\sqrt{14}\right)r_{\max}\sqrt{S^{\mathsf{C}}AT\ln\left(\frac{2SAT}{\delta}\right)}\left(D^{\mathsf{C}}\sqrt{(\Gamma^{\mathsf{C}}-1)}+1\right)\\&+\frac{196}{3}r_{\max}S^{\mathsf{C}}A\ln\left(\frac{2SAT}{\delta}\right)(3+\ln(T))\left(D^{\mathsf{C}}S^{\mathsf{C}}+1\right)\\&+2r_{\max}(D^{\mathsf{C}}+1)(\sqrt{S^{\mathsf{C}}AT}+S^{\mathsf{C}}A)\\&+r_{\max}D^{\mathsf{C}}\sqrt{\frac{5}{2}T\ln\left(\frac{8T}{\delta}\right)}+2\left(\sqrt{2}+1\right)r_{\max}\sqrt{S^{\mathsf{C}}AT}\\&+r_{\max}D^{\mathsf{C}}\left(1+2S^{\mathsf{C}}A+S^{\mathsf{C}}A\log_2\left(\frac{T}{SA}\right)+S^{\mathsf{C}}\right)\\\leq& C\cdot\left(r_{\max}D^{\mathsf{C}}\sqrt{\Gamma^{\mathsf{C}}S^{\mathsf{C}}AT\ln\left(\frac{SAT}{\delta}\right)}+r_{\max}D^{\mathsf{C}}\left(S^{\mathsf{C}}\right)^2A\ln^2\left(\frac{SAT}{\delta}\right)\right)\end{aligned}\tag{39}$$

where C is a numerical constant independent of the MDP instance.

From (11), with probability at least $1-\frac{\delta}{12T^{5/4}}$:

$$\begin{aligned}\Delta(\text{TUCRL},T)&\leq\sum_{k=1}^{m}\Delta_k+r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{8T}{\delta}\right)}\\&=\underbrace{\sum_{k=1}^{m}\Delta_k\mathbb{1}(k)}_{\text{see (39)}}+\sum_{k=1}^{m}\Delta_k\cdot(1-\mathbb{1}(k))+r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{8T}{\delta}\right)}\end{aligned}$$

where $1-\mathbb{1}(k)$ is the complement of $\mathbb{1}(k)$ i.e., takes value 1 only when either $t_k<C(k)$ (see (13) for the definition of $C(k)$) or $M^*\notin\mathcal{M}_k$. As is proved in Appendix F.2 of [6], since both (28) and Theorem 1 of Fruit et al. [6] hold, we have that with probability at least $1-\frac{\delta}{20T^{5/4}}\geq 1-\frac{\delta}{12T^{5/4}}$:

$$\sum_{k=1}^{m}\Delta_k\mathbb{1}\{M^*\notin\mathcal{M}_k\}\leq r_{\max}\sqrt{T}\tag{40}$$

As a consequence of (28) $t_{k+1}\leq 2t_k$. Thus, by definition of the condition $t_k<C(k)$ we have

$$\sum_{k=1}^{m}\Delta_k\cdot\underbrace{\mathbb{1}\{t_k<C(k)\}}_{\geq 0}\leq 2r_{\max}C(k)\leq\frac{4802}{9}r_{\max}\left(D^{\mathsf{C}}\right)^2S^3A\ln^2\left(\frac{2SAT}{\delta}\right)\tag{41}$$

Finally, by Boole's inequality: $1-\mathbb{1}(k)\leq\mathbb{1}\{M^*\notin\mathcal{M}_k\}+\mathbb{1}\{t_k<C(k)\}$ and so

$$\sum_{k=1}^{m}\Delta_k\cdot(1-\mathbb{1}(k))\leq\underbrace{\sum_{k=1}^{m}\Delta_k\cdot\mathbb{1}\{M^*\notin\mathcal{M}_k\}}_{\text{see (40)}}+\underbrace{\sum_{k=1}^{m}\Delta_k\cdot\mathbb{1}\{t_k<C(k)\}}_{\text{see (41)}}$$

In conclusion, there exists a numerical constant C independent of the MDP instance such that for any MDP and any $T>1$, with probability at least $1-\frac{\delta}{12T^{5/4}}-\frac{\delta}{12T^{5/4}}-\frac{\delta}{12T^{5/4}}=1-\frac{\delta}{4T^{5/4}}$ we have:

$$\Delta(\text{TUCRL},T)\leq C\cdot\left(r_{\max}D^{\mathsf{C}}\sqrt{\Gamma S^{\mathsf{C}}AT\ln\left(\frac{SAT}{\delta}\right)}+r_{\max}\left(D^{\mathsf{C}}\right)^2S^3A\ln^2\left(\frac{SAT}{\delta}\right)\right)\tag{42}$$

Since $\sum_{T=2}^{+\infty}\frac{\delta}{4T^{5/4}}=\delta$, by taking a union bound we have that the regret bound (42) holds with probability at least $1-\delta$ for all $T>1$.

Algorithm 2 OPTIMISTIC TRANSITION PROBABILITIES (OTP) [2]

Input: Probability estimate $\widehat{p} \in \mathbb{R}^n$, confidence interval $\beta \in \mathbb{R}$, value vector $v \in \mathbb{R}^n$, subset of states $\mathcal{I} \subseteq \{s_1, \ldots, s_m\}$, $m \leq n$, such that $\sum_{s \in \mathcal{I}} \widehat{p}(s) = 1$
Output: Optimistic probabilities $\widetilde{p} \in \mathbb{R}^n$

Let $\mathcal{I} = \{s_1, s_2, \ldots, s_m\}$ such that $v(s_1) \geq v(s_2) \geq \ldots \geq v(s_m)$
$\widetilde{p}_1(s_1) = \min\left\{1, \widehat{p}(s_1) + \frac{\beta}{2}\right\}$
$\widetilde{p}_1(s_j) = \widehat{p}(s_j), \quad \forall 1 < j \leq m$
$j = m$
$i = 1$
while $\sum_{s \in \mathcal{I}} \widetilde{p}_i(s) > 1$ **do**
  $i = i + 1$
  $\widetilde{p}_i(s) = \widetilde{p}_{i-1}(s), \quad \forall s \neq s_j$
  $\widetilde{p}_i(s_j) = \max\left\{0, 1 - \sum_{s \in \mathcal{I} \setminus \{s_j\}} \widetilde{p}_{i-1}(s)\right\}$
  $j = j - 1$
end while
$\widetilde{p}_i(s) := 0, \ \forall s \in \mathcal{S} \setminus \mathcal{I}$
$\widetilde{p} := \widetilde{p}_i$

E Shortest Path Analysis

We are interesting in comparing the shortest path of any pair $(s, \overline{s}) \in \mathcal{S} \times \mathcal{S}_k^{\mathsf{C}}$ in $\mathcal{M}_k^+$ and $\overline{\mathcal{M}}_k^+$. Formally, given a target state $\overline{s}$, the stochastic shortest path $\tau_M(s) := \tau_M(s \to \overline{s})$ of an (extended) MDP M is the (negation) solution of the following Bellman equation

$$\begin{aligned} \tau_M(s) &= -1 + \max_{a \in \mathcal{A}_s, p \in B_p(s,a)} \left\{p^{\mathsf{T}} \tau_M\right\}, \qquad \forall s \neq \overline{s} \\ \tau_M(\overline{s}) &= 0 \end{aligned} \tag{43}$$

E.1 Equivalence of Shortest Path in $\mathcal{M}_k^+$ and $\overline{\mathcal{M}}_k^+$

We start by proving the following.

Lemma 6. *For any pair* $(s, \overline{s}) \in \mathcal{S} \times \mathcal{S}_k^{\mathsf{C}}$, $\tau_{\mathcal{M}_k^+}(s \to \overline{s}) = \tau_{\overline{\mathcal{M}}_k^+}(s \to \overline{s})$.

In order to analyse the properties of the stochastic shortest path we need to investigative the maximization over the confidence interval $B_p(s,a)$ either in $\mathcal{M}_k^+$ or $\overline{\mathcal{M}}_k^+$. This problem can be solved using Alg. 2. For any state-action pair (s,a), we define $\widetilde{p}_{\mathcal{M}_k^+}(\cdot|s,a) = \text{OTP}(\widehat{p}(\cdot|s,a), B_{p,k}^+(s,a), \tau, \mathcal{S})$ and $\widetilde{p}_{\overline{\mathcal{M}}_k^+}(\cdot|s,a) = \text{OTP}(\widehat{p}(\cdot|s,a), \overline{B}_{p,k}^+(s,a), \tau, \mathcal{S})$. It is easy to notice that the optimistic probability vectors built by Alg. 2 satisfy (either in $\mathcal{M}_k^+$ or in $\overline{\mathcal{M}}_k^+$)

$$\begin{aligned} \forall i \in \{1, \ldots, n\}, \qquad & \widetilde{p}_i(s_1) \geq \widehat{p}(s_1) \\ \forall i \in \{2, \ldots, n\}, \forall l \in \{n-i+2, n\}, \qquad & \widetilde{p}_i(s_l) = \max\left\{0, 1 - \sum_{s' \neq s_l} \widetilde{p}_{i-1}(s')\right\} \\ & = \max\left\{0, \widehat{p}(s_l) - \left(\sum_{s'} \widetilde{p}_{i-1}(s') - 1\right)\right\} \\ & \leq \widehat{p}(s_l) \end{aligned}$$

where $s_1, \ldots, s_n$ are such that $\tau(s_1) \geq \ldots \geq \tau(s_n)$. The algorithm may stop before n iterations but this means that the states not processed are kept at $\widehat{p}$.

We start considering the case in which $(s,a) \in \mathcal{K}_k$. Recall that $\forall s' \in \mathcal{S}_k^{\mathsf{T}}$, $\widehat{p}(s'|s,a) = 0$ by definition since s' is not reachable from $\mathcal{S}_k^{\mathsf{C}}$ (i.e., $N_k(s,a,s') = 0$) and that $\mathcal{M}_k^+$ and $\overline{\mathcal{M}}_k^+$ consider

the same empirical average for the transition probabilities (i.e., $\widehat{p}$). The shortest path to $\overline{s}$ is such that $\max_s\{\tau(s)\} = \tau(\overline{s}) = 0$ and $\tau(s) \leq -1$ for any state $s \in \mathcal{S} \setminus \{\overline{s}\}$ (either in $\mathcal{M}_k^+$ or $\overline{\mathcal{M}}_k^+$). As a consequence, $s_1 = \overline{s}$ and for any $s' \in \mathcal{S}_k^{\mathrm{T}}$,

$$\widetilde{p}_{\mathcal{M}_k^+}(s') \leq \widehat{p}(s') = 0, \text{ and } \widetilde{p}_{\overline{\mathcal{M}}_k^+}(s') \leq \widehat{p}(s') = 0$$

which ensures that $\forall (s,a) \in \mathcal{K}_k$ the constraints in $\overline{\mathcal{M}}_k^+$ hold. This results is independent from the vector v provided to OTP. Then, for any vector $v \in V = \{v \in \mathbb{R}^S | v(\overline{s}) = 0 \wedge v(s) \leq -1, \ \forall s \in \mathcal{S} \setminus \{\overline{s}\}\}$, we have that $\mathcal{I}^1 = \mathcal{I}^2$, since $\beta^1 = \beta^2$ and $\widetilde{p}_{\mathcal{M}_k^+}(s) = \widetilde{p}_{\overline{\mathcal{M}}_k^+}(s) = 0$ for any $s \in \mathcal{S}_k^2$ then: $\widetilde{p}_{\mathcal{M}_k^+}(s') = \widetilde{p}_{\overline{\mathcal{M}}_k^+}(s'), \ \forall s' \in \mathcal{S}$. Finally, $\forall (s,a) \in (\mathcal{S} \times \mathcal{A}) \setminus \mathcal{K}_k$ it is trivial to notice that: $\forall s' \in \mathcal{S}$, $\forall v \in V$, $\widetilde{p}_{\mathcal{M}_k^+}(s') = \widetilde{p}_{\overline{\mathcal{M}}_k^+}(s')$ since $B_{p,k}^+(s,a) = \overline{B}_{p,k}^+(s,a)$.

The proof follows by noticing that $\tau_{\mathcal{M}_k^+} \in V$ and $\tau_{\overline{\mathcal{M}}_k^+} \in V$.

E.2 Bounding the bias span

Lemma 7. *Consider an (extended) MDP M and define L_M as the associated optimal (extended) Bellman operator. Given $h_0 = \mathbf{0}$, and $h_i = (L_M)^i h_0$ we have that*

$$\forall s, s' \in \mathcal{S}, \ \ h_i(s') - h_i(s) \leq r_{\max} \tau_M(s \to s')$$

where $\tau_M(s \to s')$ is the minimum expected shortest path from s to s' in M.

Proof. The proof follows from the application of the argument in [2, Sec. 4.3.1]. □

F Proof of Lem. 3

We prove the statement by contradiction: we assume that there exists a learning algorithm denoted $\mathfrak{A}_T$ satisfying

1. for all $\varepsilon \in]0,1]$, there exists $T_\varepsilon^\dagger \leq f(\varepsilon)$ such that $\mathbb{E}[\Delta(\mathfrak{A}_T, M_\varepsilon, x, T)] < 1/6 \cdot T$ for all $T \geq T_\varepsilon^\dagger$,
2. there exists $T_0^* < +\infty$ such that $\mathbb{E}[\Delta(\mathfrak{A}_T, M_0, x, T)] \leq C_2(\ln(T))^\beta$ for all $T \geq T_0^*$.

Any randomised strategy for choosing an action at time t is equivalent to an (a priori) random choice from the set of all deterministic strategies. Thus, it is sufficient to show a contradiction when the action played by $\mathfrak{A}_T$ at any time t is a deterministic function of the past trajectory $h_t := \{s_1, a_1, r_1, \ldots, s_t\}$. In the rest of the proof we assume that $\mathfrak{A}_T$ maps any sequence of observations $h_t = \{s_1, a_1, r_1, \ldots, s_t\}$ to a (single) action a_t.

By trivial induction it is easy to see that as long as state y has not been visited, the history h_t is independent of ε ($\mathfrak{A}_T$ can not distinguish between different values of ε and plays exactly the same action when the past history is the same).

Let's define $N_T^0(x,b) := \sum_{t=1}^T \mathbb{1}\{(s_t, a_t) = (x,b)\}$ the number of visits in (x,b) with $a_t = \mathfrak{A}_T(h_t)$ and $\varepsilon = 0$. Note that $N_T^0(x,b)$ is not random since when $\varepsilon = 0$ both action b and action d loop on x with probability 1. For any $\varepsilon \in [0,1]$ and any horizon T define the event:

$$F(T, \varepsilon) := \bigcap_{1 \leq t \leq T} \{s_t \neq y\}$$

where the sequence of states s_t is obtained by executing $\mathfrak{A}_T$ on MDP M_ε. We will denote by $\overline{F(T,\varepsilon)}$ the complement of $F(T,\varepsilon)$.

For any horizon T, and independently of ε, there is only one possible trajectory $h_T = \{s_1, a_1, r_1, \ldots, s_T\}$ that never goes to y and which corresponds to the trajectory observed when $\varepsilon = 0$. When $\varepsilon = 0$, the probability of this trajectory is 1 and so $\mathbb{P}\left(F(T,0)\right) = 1$ (recall that everything is deterministic in this case) while in general we have:

$$\forall T \geq 1, \ \forall \varepsilon \in [0,1], \ \ \mathbb{P}\left(F(T,\varepsilon)\right) = (1-\varepsilon)^{N_T^0(x,b)} \tag{44}$$

We now prove by contradiction that

$$\lim_{T\to+\infty} N_T^0(x,b) = +\infty \tag{45}$$

Let's assume that $C := \max\left\{10, \max_{T\geq 1}\{N_T^0(x,b)\}\right\} < +\infty$. Taking $\varepsilon = 1/C$ and applying the law of total expectation we obtain:

$$\begin{aligned}\forall T \geq 1,\ \ \mathbb{E}[\Delta(\mathfrak{A}_T, M_{1/C}, x, T)] &= \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_T, M_{1/C}, x, T)|F(T,1/C)\right]}_{=T/2+1/2\cdot N_T^0(x,b)\geq T/2} \cdot \underbrace{\mathbb{P}\left(F(T,1/C)\right)}_{=(1-1/C)^{N_T^0(x,b)}} \\ &\quad + \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_T, M_{1/C}, x, T)|\overline{F(T,1/C)}\right] \cdot \mathbb{P}\left(\overline{F(T,1/C)}\right)}_{\geq 0} \\ &\geq \frac{T}{2}\cdot\left(1-\frac{1}{C}\right)^{N_T^0(x,b)} \geq \frac{T}{2}\cdot\underbrace{\left(1-\frac{1}{C}\right)^{C}}_{\geq 1/3 \ \text{ by Lem. 8}} \geq \frac{T}{6}\end{aligned}$$

where we used the fact that

- $N_T^0(x,b) \leq C$ and $(1-1/C) \in [0,1]$ by definition, implying $\left(1-\frac{1}{C}\right)^{N_T^0(x,b)} \leq \left(1-\frac{1}{C}\right)^C$,
- since $C \geq 10$ we have $\left(1-\frac{1}{C}\right)^C \geq 1/3$ by Lem. 8 applied to $x = 1/C$,
- and finally under event $F(T,1/C)$, the regret incurred is exactly $T/2 + 1/2\cdot N_T^0(x,b) \geq T/2$.

This contradicts our assumption that there exists $T_{1/C}^\dagger < +\infty$ such that for all $T \geq T_{1/C}^\dagger$, $\mathbb{E}[\Delta(\mathfrak{A}_T, M_{1/C}, x, T)] < T/6$ and so (45) holds.

Since $\lim_{T\to+\infty} N_T^0(x,b) = +\infty$, it is possible to construct a strictly increasing sequence $(T_n)_{n\in\mathbb{N}}$ such that:

$\forall n \in \mathbb{N},\ N_{T_{n+1}}^0(x,b) > N_{T_n}^0(x,b),\ \ T_0 = T_0^*,\ \ T_1 \geq C_2,\ \ T_1 \geq C_2(\ln(T_1))^\beta$ and $N_{T_1}^0(x,b) \geq 10$

We also define the (strictly decreasing) sequence: $\varepsilon_n := 1/N_{T_n}^0(x,b),\ \forall n \geq 1$. By the law of total expectation:

$$\begin{aligned}\mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)] &= \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)|F(T_n,\varepsilon_n)\right]}_{\geq T_n/2} \cdot \underbrace{\mathbb{P}\left(F(T_n,\varepsilon_n)\right)}_{=(1-\varepsilon_n)^{N_{T_n}^0(x,b)}} \\ &\quad + \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)|\overline{F(T_n,\varepsilon_n)}\right] \cdot \mathbb{P}\left(\overline{F(T_n,\varepsilon_n)}\right)}_{\geq 0} \\ &\geq \frac{T_n}{2}\cdot(1-\varepsilon_n)^{N_{T_n}^0(x,b)} = \frac{T_n}{2}\cdot\underbrace{(1-\varepsilon_n)^{1/\varepsilon_n}}_{\geq 1/3 \ \text{ by Lem. 8}} \geq \frac{T_n}{6}\end{aligned} \tag{46}$$

where we applied Lem. 8 to $x = \varepsilon_n \leq 1/10$ since $N_{T_n}^0(x,b) \geq 10$ for all $n \geq 1$. Moreover, since by construction for all $n \geq 1$, $T_n > T_0 = T_0^*$ we have by assumption that

$$\begin{aligned}\forall n \geq 1,\ \ \mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_0, x, T_n)] &= \frac{1}{2}N_{T_n}^0(x,b) = \frac{1}{2\varepsilon_n} \leq C_2(\ln(T_n))^\beta \\ \implies T_n &\geq \exp\left(\frac{1}{(2C_2\cdot\varepsilon_n)^{1/\beta}}\right)\end{aligned}$$

Since $\lim_{n\to+\infty} 1/\varepsilon_n = +\infty$ and $\lim_{x\to+\infty} \exp\left(x^{1/\beta}\right)/x^\alpha = +\infty$ there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $T_n \geq f(\varepsilon_n)$. By assumption, for all $n \geq N$,

$$\mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)] < \frac{T_n}{6}$$

which contradicts (46) therefore concluding the proof.

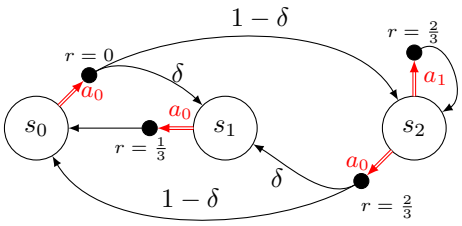


Figure 6: Three-state domain introduced in [6]

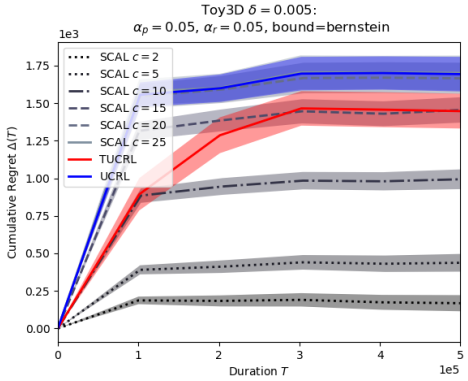


Figure 7: Communicating three-state domain ($\delta = 0.005$)

Lemma 8. *For all* $x \in]0, 1/10]$*, we have* $(1-x)^{1/x} \geq 1/3$.

Proof. It is easy to verify that the derivative of $x \longmapsto (1-x)^{1/x}$ is:

$$\forall x \in]0, 1/10], \quad \frac{d}{dx}\left((1-x)^{1/x}\right) = -\underbrace{\frac{(1-x)^{1/x-1}}{x^2}}_{\geq 0} \cdot ((1-x)\ln(1-x) + x)$$

It is well known that for all $x \in]0, 1[$, $x < -\ln(1-x) < \frac{x}{1-x}$ implying that $(1-x)\ln(1-x) + x$ is positive. Therefore, $\frac{d}{dx}\left((1-x)^{1/x}\right)$ is negative on $]0, 1/10]$ implying that $x \longmapsto (1-x)^{1/x}$ is decreasing. As a result: $\forall x \in]0, 1/10], \quad (1-x)^{1/x} \geq 0.9^{10} > 1/3$. □

G Experiments - Three-State Domain

This domain was introduced in [6] in order to show the inability of UCRL to learn in weakly communicating MDPs. The graphical representation of the domain is reported in Fig. 6. We keep the same means for the rewards (reported on Fig. 6) but we change the distributions: uniform distributions with range $1/5$ instead of Bernouillis. In the main paper we showed how the algorithms behave when $\delta = 0$. Here we consider the case the MDP is communicating by defining $\delta = 0.005$. Fig. 7 shows that, as expected, TUCRL behaves similarly to UCRL. In this example it is able to outperform UCRL since the preliminary phase in which transitions to non-observed states are forbidden leads to a more conservative exploration that, due to the structure of the problem (s_1 is difficult to reach but it is also non-optimal), results in a smaller regret.