Active Roll-outs in MDP with Irreversible Dynamics

Odalric-Ambrym Maillard, Timothy Mann, Ronald Ortner, Shie Mannor

Journal of Machine Learning Research x (xxxx) pp-pp Submitted date1; Published date2

Active Roll-outs in MDP with Irreversible Dynamics

Odalric-Ambrym Maillard[1] ODALRICAMBRYM.MAILLARD@INRIA.FR
Timothy A. Mann[2] MANN.TIMOTHY@GMAIL.COM
Ronald Ortner[3] RORTNER@UNILEOBEN.AC.AT
Shie Mannor[4] SHIE@EE.TECHNION.AC.IL
1. *Inria Saclay – Île de France*
Laboratoire de Recherche en Informatique
660 Claude Shannon
91190 Gif-sur-Yvette, France
2. *Google Deepmind*
London, United Kingdom
3. *Department für Mathematik und Informationstechnologie*
Lehrstuhl für Informationstechnologie
Montanuniversität Leoben
Franz-Josef-Straße 18
A-8700 Leoben, Austria
4. *The Technion – Israel institute of Technology*
Faculty of Electrical Engineering
Fishbach Building
32000 Haifa, Israel

Editor: ....

Abstract

In Reinforcement Learning (RL), regret guarantees scaling with the square root of the time horizon have been shown to hold only for communicating Markov decision processes (MDPs) where any two states are connected. This essentially means that an algorithm can eventually recover from any mistake. However, real-world tasks usually include situations where taking a single "bad" action can permanently trap a learner in a suboptimal region of the state-space. Since it is provably impossible to achieve sub-linear regret in general multi-chain MDPs, we assume a weak mechanism that allows the learner to request additional information. Our main contribution is to address: (i) how much external information is needed, (ii) how and when to use it, and (iii) how much regret is incurred. We design an algorithm that minimizes requests for external information in the form of rollouts of a policy specified by the learner by *actively* requesting it only when needed. The algorithm provably achieves $O(\sqrt{T})$ active regret after T steps in a large class of multi-chain MDPs, by only requesting $O(\log(T))$ rollout transitions. The superiority of our algorithm to standard algorithms such as **R-Max** and **UCRL** is demonstrated in experiments on some illustrative grid-world examples.

Keywords: Reinforcement learning, Regret analysis, Multi-chain, MDP, Recoverability.

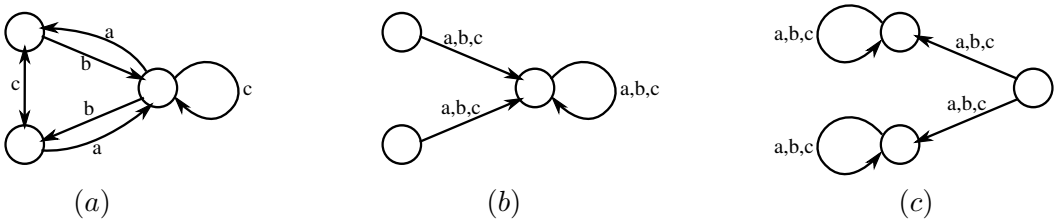


Figure 1: Example of (a) a communicating MDP, (b) a unichain MDP with a single recurrent class, and (c) a multi-chain MDP with two recurrent classes. The circles represent states while the labeled edges represent transitions due to executing actions $\{a, b, c\}$.

1. Introduction

In Reinforcement Learning (RL), an agent interacts with an environment by sequentially playing an action and then receiving an observation and a reward for that action. The goal of the agent is to maximize its accumulated reward, and the most popular model for representing RL problems is the Markov decision process (MDP) model.

In this paper, we consider an agent that interacts with a multi-chain MDP, in a single stream of observations-actions-rewards, with no reset. In a multi-chain MDP, there does not always exist a state-action path with positive probability between any two states. This is in stark contrast with the more common and restrictive assumption that the MDP is communicating (that is, between any two states, there always exist a finite path of positive probability that connects the two states). For illustration, we depict in Figure 1 an example of a communicating MDP, an example of a unichain MDP with one recurrent class, and an example of a multi-chain MDP with two recurrent classes. Note that without a special reset action that takes the learner back to an initial state, the learning problem is difficult. In fact, it is provably impossible to achieve sub-linear regret performance by only interacting with the environment and without prior information. This can be shown easily by looking at a so-called Heaven-or-Hell MDP depicted in Figure 2. In such an MDP, one action leads to an absorbing state giving reward 0 (Hell), and another action leads to an absorbing state giving reward 1 (Heaven). Without knowing which action leads to hell and which to heaven, an algorithm must incur linear regret in such MDPs (in a minimax sense). Thus, the only way to hope for sub-linear regret is to access an external source of information (in the sense that it does not result from an interaction between the learner and the MDP). We discuss later in the second next paragraph what type of external information does not trivialize the problem.

Motivation In this paper, we want to provide an answer to the open theoretical question: How to achieve near-optimal learning guarantees in this setting? Since learning without external information is not possible, the answer mainly deals with the external information used by the learner. That is, our contribution is three-fold, addressing:

(i) how much external information is needed,

(ii) how and when to use it, and

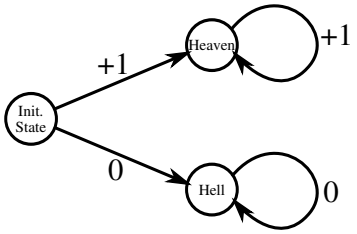


Figure 2: The Heaven-or-Hell MDP. The agent's choice from the initial state either sends it to the Heaven state where it receives a reward of +1 forever, or to the Hell state where it receives a reward of 0 forever. Once the agent is in either the Heaven or Hell state it cannot return to the initial state.

(iii) how much (active) regret is incurred.

More precisely, we introduce a setting where the learner is allowed to —instead of taking an action in the current state s_t— ask for a rollout (starting in s_t) of a policy under a stopping condition both to be specified by the learner. For this setting we propose **A-ROGUE** (for **A**ctive **R**oll**O**ut-**GU**ided **E**xploration), a learning algorithm that, using only $O(\log(T))$ rollout transitions, achieves a $O(\sqrt{T})$ regret bound in multi-chain MDPs under some modest technical assumptions, see Theorem 7 in Section 4. This regret is optimal in T, as it matches the known lower bounds for communicating MDPs (see Jaksch et al., 2010). The performance of the algorithm is supported by illustrative numerical experiments in Section 6.

Other types of external information In practice, one may sometimes have access to additional elements that help, or even trivialize the multi-chain MDP problem, such as a simulator, a harness, an expert, or a teacher. For instance, under the assumption that one has access to a harness designed by an expert that prevents a learner from accessing certain regions of the state space, then the initial multi-chain MDP can simply be reduced to a communicating MDP. Likewise, if one has a "rescue/repair" action that sets the learner back to an initial state, at the price of a large negative reward, then the initial MDP can again be reduced to an equivalent communicating MDP. Third, under the assumption that an inexpensive and accurate simulator is available, there is no need for learning in the real-world and the problem reduces to a planning problem. Thus, the multi-chain MDP formulation becomes potentially interesting when no harness, rescue/repair action is available, and when the only available simulators are either expensive or inaccurate. Finally, assuming that a good teacher is available, who provides examples of actions/policies to follow at various steps may also trivialize the problem. However, a good teacher has to know both the environment and what the learner has learned about the environment, and must monitor the learner to provide the right advice at the right moment. Note also that a teacher must be accurate. Such an assumption is however rarely met in practice, and designing a good teacher is difficult. Instead of assuming a teacher that constantly monitors the learner, we are interested in an *active* learner who can detect when it requires external information and can specify which kind of information she wants to get feedback on. Thus, we propose

a setting where the learner can actively ask for rollouts of a policy with a stopping time criterion that are both designed by the learner.

Setting We target the challenging goal of learning in a single stream of state-action-rewards with no reset in a possibly *non-recoverable* multi-chain MDP, in the strong sense of *regret minimization* by combining standard MDP algorithms with *active information retrieval*: The learner can decide at any time not to act in the MDP but instead to propose a policy and a stopping condition and ask for a rollout of the policy until the stopping condition is met. To keep it simple, each rollout transition counts the same as a time step in which the learner executes an action and gives no reward. We capture this with the convenient notion of *active regret* that combines both the regret and the amount of rollout transition asked; see Section 2 for details. However, since our proof actually bounds the standard regret term and the extra information term separately, it is straightforward to adapt our approach to handle the case when different costs for rollout transitions and time-steps are used.

Previous work Previous work has mostly considered two types of performance guarantees: (1) regret and (2) sample complexity. *Regret* measures the difference between the sum of rewards received by the learning algorithm and the sum of rewards received by the optimal policy over a finite or infinite time horizon (Jaksch et al., 2010). *Sample complexity* measures the number of time steps the learning algorithm acts suboptimally from its current state (Kakade, 2003). Mostly, it is assumed that the underlying MDP has no transient region (Jaksch et al., 2010; Maillard et al., 2013) or exactly one recurrent class (Bartlett and Tewari, 2009), meaning that their guarantees do not hold in general multi-chain MDPs (such as the one given in Figure 2). Sometimes alternative guarantees are given that have the unintuitive property that if the learner is trapped in a region with low rewards, it can still learn to act "optimally" with respect to that region of the state-space.

With the goal of preventing the learner to take irreversible actions, various notions of "safety"have been discussed in the literature. Thus, Moldovan and Abbeel (2012) constrain the class of policies to satisfy a safety constraint. Although they provide some theoretical support for the resulting algorithm, they do not compare to the optimal policy (which may be considered as unsafe), and assume a prior that essentially tells where it is unsafe to go. Also, their formulation of the safety constraint makes the learning problem NP-hard, contrary to our approach. Further, note that in contrast to notions of safety (Moldovan and Abbeel, 2012; García and Fernández, 2012), executing an irreversible action should not necessarily be avoided if it is the optimal thing to do. Indeed, if the region contains an optimal cycle the algorithm should enter it (but it first needs to be confident since the action is irreversible). Our approach is reminiscent of the earlier work of Clouse (1997), Hailu and Sommer (1998), Hans et al. (2008) and more recently of Geramifard et al. (2011) and García and Fernández (2012), in which the agent explicitly asks for advice when it encounters specific situations. There, advice is generally an action proposed by a teacher, assumed to be good, or (in some sense) safe. Chernova and Veloso (2009) further propose to ask for a demonstration (rollout) from a teacher, who also chooses which policy to demonstrate. Unlike this, in our setting the learner proposes policies by herself.

In the planning community, some recent related work has shown how to plan efficiently in goal-oriented MDPs with traps, see (Teichteil-Königsbuch et al., 2011; Kolobov et al.,

2012). However, none of these approaches provides any theoretical analysis but only gives experimental results. A notable exception, though in the different universal knowledge seeking framework, is (Orseau et al., 2013), which provides a Bayesian algorithm that, given a set of models of the environment, tries to identify the right one by acting so as to maximize the information gain at each step.

The intuitive difficulty of dealing with many recurrent classes Regarding the amount of extra information asked by the learner (rollout transitions), our findings show that *detecting* whether an action is irreversible (enters a recurrent class) is generally not too costly, while *deciding whether to execute* a irreversible action is much more expensive. The reason is that since an optimal cycle may consist only of a single state, in the worst case *every state-action pair* inside a recurrent region reachable from the current state must be visited sufficiently often to decide whether to enter this class. Also, rollouts must be sufficiently long to reach any state in the class, which means that dealing with large traps (recurrent regions with large diameter) is *intrinsically hard*. Fortunately, in most applications, recurrent classes correspond to "traps" that —almost by definition— only consist of a few states. They typically model a physical component failure, where no (sensible) action can be performed from that point.

Contribution and outline We consider finite (multi-chain) MDPs where some actions have (potentially) irreversible consequences that may cause high regret in the long-run. In Section 2, we introduce the convenient notion of *active regret* as well as two new quantities that capture the difficulty of learning optimal behavior in an MDP (as indicated in the previous paragraph): the *local diameter* $d^\star$ of an MDP intuitively measuring the "size" of a recurrent class and the *sharpness of action-gaps* $\gamma^\star$ that quantifies how bad an irreversible action can be. In Section 3, we introduce the **A-ROGUE** algorithm, that can handle multi-chain MDPs with possibly irreversible actions, by precisely deciding when and for which policy to ask for a rollout. **A-ROGUE** can be seen as a natural generalization of **UCRL** to the case of multi-chain MDPs, and indeed coincides with **UCRL** when one of its parameters (diameter guess) is set to ∞. Our main result Theorem 7 in Section 4 shows that **A-ROGUE** enjoys strong finite-time regret guarantees scaling as $\tilde{O}\Big(\big(\sqrt{\frac{d^\star}{\gamma^\star}}+d^\star\big)\sqrt{T}\Big)$, by only resorting to $O(\frac{1}{\gamma^{\star 5}}\log(T))$ rollout transitions. Section 5 provides a detailed analysis of the regret performance, including some innovations of independent interest. Thus, we demonstrate how to estimate the return time to some state and give respective upper and lower bounds. Further, we obtain a bound on the number of rollouts needed to cover a subset of states from a given reference state. In the final Section 6, we present the results of some experiments in illustrative gridworlds. They show that popular algorithms driven by the "optimism in the face of uncertainty" principle such as **UCRL** and **R-Max**, and which enjoy strong performance guarantees in MDPs with strong recurrence properties, are clearly outperformed by **A-ROGUE**.

Contents

2. Setup and Notations

Setting Let $\Gamma(\mathcal{X})$ denote the set of all probability distributions over a non-empty set $\mathcal{X}$. We consider an undiscounted Markov decision processes (MDP) $M = \langle \mathcal{S}, \mathcal{A}, p, \nu \rangle$ with horizon T, finite state space $\mathcal{S}$ of size S, finite action space $\mathcal{A}$ of size A, transition kernel $p : \mathcal{S} \times \mathcal{A} \to \Gamma(\mathcal{S})$ that maps each state-action pair to a distribution over next states, and reward function $\nu : \mathcal{S} \times \mathcal{A} \to \Gamma([0,1])$ that maps each state-action pair (s,a) to a real-value distribution over $[0,1]$ with mean $\mu(s,a)$. If the state and action spaces are clear from context and we are only interested in the mean rewards $\mu(s,a)$, we often write MDPs abbreviated as $M = (\mu, p)$.

In any state s of the MDP, a learner is allowed to either

1. execute some action a, then receive a reward $r(s,a) \sim \nu(s,a)$ and move to a next state according to $p(\cdot|s,a)$, or

2. ask for a rollout from the current state s for a policy π and a stopping criterion (both provided by the learner herself), until the stopping criterion is met.

All states visited and rewards generated during the rollout execution are reported to the learner, but the learner receives no rewards. After the rollout the learner is in the same state as before it, however each time-step in the rollout counts like a time-step in which an action is executed, cf. the definition of *active regret* below.

2.1 Regret Definition

We measure the regret of an algorithm $\mathbb{A}$ starting in state s_1 and interacting during $T = T_1 + T_2$ time steps with the MDP, where T_1 is the total number of actions executed in

the MDP and T_2 is the total number of transitions of all requested rollouts. The regret is defined as the difference between the rewards accumulated by an optimal T-step policy π^* run from s_1 during T time steps and the rewards accumulated by the algorithm $\mathbb{A}$ from s_1 (during the T_1 steps that $\mathbb{A}$ executed an action in the MDP). That is, we define the (expected) *active regret* by $\mathbb{E}\big[\mathfrak{R}_T\big]$, where

$$\mathfrak{R}_T = \sum_{t=1}^{T} r(s_t^\star, a_t^\star) - \sum_{t=1}^{T} r(s_t, a_t) \tag{1}$$

with $r(s,a) \sim \nu(s,a)$, and where $(s_t^\star, a_t^\star)$ is the state-action pair visited at time step t when following $\pi^\star$, and a_t is the action executed by the algorithm $\mathbb{A}$ in the state s_t visited by $\mathbb{A}$ at time step t. Note that counting the rollout transitions against the horizon is equivalent to putting an implicit cost on each rollout transition (the reward that an optimal policy would receive at the same time). Note also that the (expected) active regret is never smaller than the (expected) "traditional" regret (Jaksch et al., 2010).

Note that access to rollouts does not eliminate the exploration-exploitation dilemma. Time increases independent of whether the learner is acting in the MDP or rather observing one step in a simulated rollout. Since the algorithm only receives reward when it is acting in the MDP, it must balance the number of time steps spent on observing rollouts with acting in the environment to obtain maximize rewards. As already mentioned, for the sake of simplicity we use the simplifying assumption that observing a rollout transition has the same cost as executing an action, while extensions to different cost functions are straightforward.

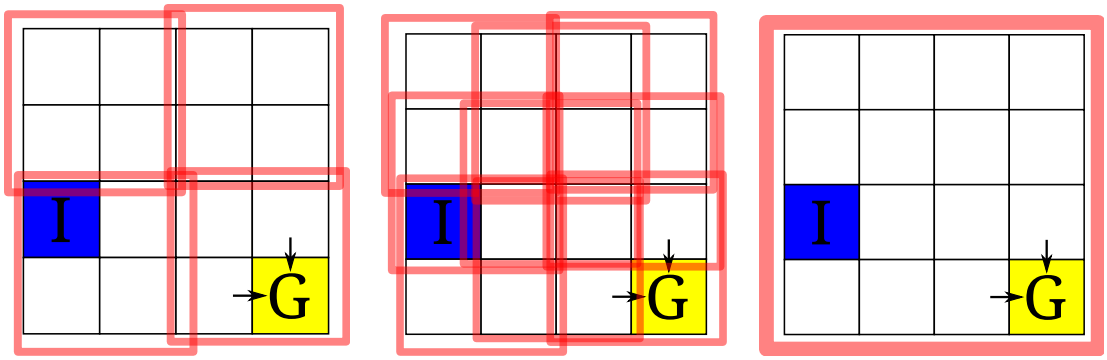
2.2 Measures of Complexity

For any two states $s, s' \in \mathcal{S}$ and a policy π, let $\mathcal{T}_\pi^M(s,s')$ be the (possibly infinite) *expected* number of steps needed to reach s' starting from state s when following policy π in the MDP M. Further, let $\mathcal{T}^M(s,s') = \min_\pi \mathcal{T}_\pi^M(s,s')$ be the minimal transition time between s and s'. Classically, the (global) *diameter* of an MDP M is defined to be $D^M = \max_{s,s'} \mathcal{T}^M(s,s')$ (Jaksch et al., 2010), however in multi-chain MDPs, this quantity may be infinite. For a subset $\mathfrak{S} \subset \mathcal{S}$, we also define $D_{\mathfrak{S}}^M = \max_{s,s' \in \mathfrak{S}} \mathcal{T}^M(s,s')$.

Another standard notion is the *action gap* (Farahmand, 2011). For horizon T, let $Q_{T-t}^M(s,a)$ be the optimal action-value of the state-action pair (s,a) at time t for the MDP M (thus following an optimal policy), see Szepesvári (2010). The action gap of (s,a) then is defined as

$$\Delta_{T-t}^M(s,a) = \max_{a' \in \mathcal{A}} Q_{T-t}^M(s,a') - Q_{T-t}^M(s,a)\,.$$

Note that $\Delta_{T-t}^M(s,a) > d$ implies that it is not possible to reach s from a successor state of (s,a) in d steps on average (since otherwise the difference of value would be less than d), that is $\mathcal{T}^M(s',s) > d$. Similarly, a (global) diameter $D^M \leqslant d$ as assumed in (Jaksch et al., 2010; Brafman and Tennenholtz, 2003) implies that for all states and time steps every action is d-optimal, that is, $\Delta_{T-t}^M(s,a) \leqslant d$.



(a) (b) (c)

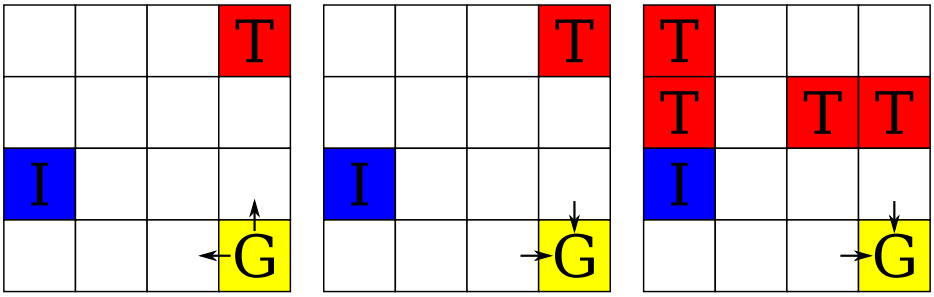
Figure 3: Coverings of a 4×4-gridworld MDP: (a) a non-maximal set of maximally large regions with diameter not larger than 2; (b) the unique 2-covering of the MDP; (c) the unique 4-covering of the MDP, with partition $\{\mathfrak{S}_1^\star\}$ where $\mathfrak{S}_1^\star = \mathcal{S}$.

Local diameter We now introduce the new notion of *local diameter*. First, the *d-covering* of an MDP M is a maximal set $\{\mathfrak{S}_i \subset \mathcal{S} | D_{\mathfrak{S}_i}^M \leqslant d\}$ of *maximally large* state regions $\{\mathfrak{S}_i\}_{i \leqslant I}$ with $\mathfrak{S}_i \subset \mathcal{S}$ that all satisfy $D_{\mathfrak{S}_i}^M \leqslant d$. That is, each $\mathfrak{S}_i$ cannot be extended without increasing its diameter beyond d. We provide in Figure 3 an illustration of three coverings of a 4×4-gridworld MDP: The first covering corresponds to a partition with sets of diameter not larger than 2, but it is not a 2-covering, since it is a *non-maximal* set of maximally large regions with diameter not larger than 2. The second and third coverings are the unique 2-covering and 4-covering of the MDP. The former one happens to be also a (trivial) partition of the state space. Note that if there is an overlap between $\mathfrak{S}_i$ and $\mathfrak{S}_{i'}$, then $D_{\mathfrak{S}_i \cup \mathfrak{S}_{i'}}^M \leqslant 2d$. In general, the $\mathfrak{S}_i$ of a d-covering will be multiply overlapping. Only in particular cases a d-covering will be a *partition* (non-overlapping covering) of the state space.

Definition 1 *The local diameter $d^\star$ is the smallest d such that its d-covering is a partition of the state space. We denote the corresponding partition by $\{\mathfrak{S}_i^\star\}_{i \leqslant I}$.*

For two sets $\mathfrak{S}_1^\star, \mathfrak{S}_2^\star$ of this partition, and states $s_1 \in \mathfrak{S}_1^\star, s_2 \in \mathfrak{S}_2^\star$, we cannot have simultaneously $\mathcal{T}(s_1, s_2) \leqslant d^\star$ and $\mathcal{T}(s_2, s_1) \leqslant d^\star$, since otherwise it would be possible to create a new set $\mathfrak{S}_{I+1}^\star$ with diameter less than $d^\star$ that overlaps the existing partition, which is not allowed by definition of $d^\star$. Likewise, it is obvious that $\{\mathfrak{S}_i^\star\}_{i \leqslant I}$ is unique. In the sequel, we will informally use the word "trap" to refer to regions $\mathfrak{S}_i^\star$ from which it is difficult to escape, that is, such that $\min_{s' \notin \mathfrak{S}_i^\star} \min_{s \in \mathfrak{S}_i^\star} \mathcal{T}(s, s')$ is "large".

Example 1 *Consider the gridworld MDPs depicted in Figure 4, and assume for simplicity that all transitions are deterministic. For the MDP (a), one can easily observe that the local diameter $d^\star = 6$ with two regions (the trap, and the other states). For the MDP (b), we have $d^\star = 5$ with three regions, and for (c), $d^\star = 6$ with six regions.*



(a) (b) (c)

Figure 4: Some gridworld environments. The letter I denotes the agent's initial state, T an absorbing trap, and G a goal state. Arrows leaving a goal state indicate that it is not absorbing. Arrows entering a goal state indicate that it is absorbing.

Sharpness of action-gaps In terms of rewards, inside a region $\mathfrak{S}_i^\star$, the action gaps cannot be too big: If $s \in \mathfrak{S}_i^\star$ and $a \in \mathcal{A}$ are such that all possible successor states (i.e., $\mathbf{supp}(p(\cdot|s,a))$) are contained in $\mathfrak{S}_i^\star$, then for all t, we must have $\Delta_{T-t}^M(s,a) \leqslant d^\star$. Thus, only if a leads to a different region $\mathfrak{S}_{i'}^\star$, the action gap can be bigger than $d^\star$. We capture the size of the action gap at such frontiers by the following definition.

Definition 2 *An MDP M with local diameter $d^\star$ is said to have $\gamma^\star$-sharp action-gaps if $\gamma^\star$ is the largest γ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, either $\Delta_{T-t}^M(s,a) \leqslant d^\star$ or $\Delta_{T-t}^M(s,a) > \gamma(T-t)$ for all $t \leqslant T$. In the latter case, we say action a is irreversible at state s.*

Recalling that inside a region $\mathfrak{S}_i$ one can reach any state from any other state in no more then $d^\star$ steps on average, this definition immediately implies the following observation.

Observation 1 *If $s \in \mathfrak{S}_i^\star$ and (s,a) leads to $\mathfrak{S}_{i'}^\star$ with $\Delta_{T-t}^M(s,a) > \gamma^\star(T-t)$, then the average number of steps needed to go from $\mathfrak{S}_{i'}^\star$ to $\mathfrak{S}_i^\star$ cannot be less than $\gamma^\star(T-t) - 2d^\star$.*

Thus, in an MDP with large $\gamma^\star$ recovering from an action that enters a new region takes more time (actions are less "recoverable"). One needs to be more cautious, but at the same time such actions are easier to detect. In an MDP with small $\gamma^\star$, recovering from an action that enters a new region takes less time. One need not be that cautious, but detecting such actions is more difficult.

To give further intuition about $\gamma^\star$, we now consider a simple illustrative example that shows how it can be bounded.

Example 2 *Consider an MDP with three disjoint regions A, B, C each with diameter $d^\star$. Starting in A, one can only reach C from a state-action pair (s_1, a_1) with deterministic transition, and B from (s_2, a_2) with a transition that goes back to s_2 with probability $1-p$, and such that s_2 is $d^\star$ expected steps from s_1. There are no other transitions between A, B and C. The learner always receives reward 0.5 in region A, 0.1 in B, and 0.9 in C. To compute $\gamma^\star$, we look at $\Delta_{T-t}^M(s_2, a_2)$. We have $V^\star(s_2) = 0.5d^\star + 0.9(T-t+1-d^\star)$, and $Q(s_2,a_2) \leqslant p(T-t)0.1 + (1-p)(0.5d^\star + 0.9(T-t-d^\star))$, that is, $\Delta_{T-t}^M(s_2,a_2) \geqslant 0.8p(T-t) - p(0.4d^\star - 0.9)$. Now, if $T-t \geqslant d^\star$, then $\Delta_{T-t}^M(s_2,a_2) > 0.4p(T-t)$, and*

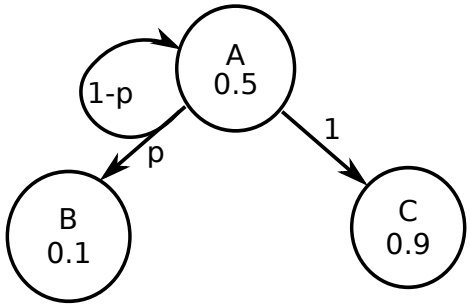


Figure 5: The MDP considered in Example 2.

otherwise $\Delta^M_{T-t}(s_2, a_2) \leqslant d^\star$. *Thus we deduce that* $\gamma^\star \geqslant 0.4p$. *In this example, the larger* p, *the less recoverable is action* a_2 *and we thus need to be confident when executing it. More generally, the larger* γ, *the less recoverable are actions and the more confident the learner needs to be.*

Mobility assumption In the sequel, we only consider MDPs with local diameter $d^\star$ and $\gamma^\star$-sharp action-gaps with non-trivial $\gamma^\star > 0$. This ensures that irreversible actions can be detected and thus that an algorithm can decide to execute it nor not. To further focus on interesting cases, we make the following "mobility" assumption:

Assumption 1 *There is a* $p_0 > 0$ *such that for all* i, i' *with* $p(\mathfrak{S}_i^\star | \mathfrak{S}_{i'}^\star, \mathcal{A}) > 0$ *there is a* $(s, a) \in \mathfrak{S}_{i'}^\star \times \mathcal{A}$ *such that* $p(\mathfrak{S}_i^\star | s, a) \geqslant p_0$ *and* $p(\mathfrak{S}_i^\star \cup \mathfrak{S}_{i'}^\star | s, a) = 1$.

That is, whenever it is possible to reach a region from another one, it is also possible to enter it easily (first condition) and without risking entering another region (second condition). This implies that there is always at least one action that does not lead to two regions at the same time. To give more intuition, an MDP that does not satisfy Assumption 1 contains a region such that *any* action leading to this region from another region also leads to a third one with positive probability. In such a situation, it is easy to design rewards such that no optimal policy can guarantee avoiding a "bad" region, see e.g. Figure 6. Thus Assumption 1 allows us to focus on cases where an optimal policy can achieve high cumulative reward. Note that this holds in most cases of practical interest, where "bad" regions are "well-separated" from "good" regions. We believe that Assumption 1 might be weakened further, for instance by assuming that the restriction only applies to regions $\mathfrak{S}_i^\star$ visited by an optimal policy. However, since the analysis under such a weakened assumption would become more tedious, we did not investigate this direction in detail.

3. The A-ROGUE Algorithm

We now present the **A-ROGUE** algorithm (for **A**ctive **R**oll**O**ut-**GU**ided **E**xploration) to handle the setting introduced in Section 2. We start with a high-level description as given in Algorithm 1, the full algorithm with all the technical details can be found below as Algorithm 2.

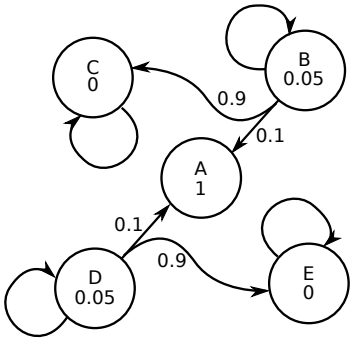


Figure 6: An example of MDP that does not satisfy Assumption 1. Here, when in region B or D, the optimal policy will try to reach region A, which risks entering the bad regions C and E with high probability.

A-ROGUE is an extension of **UCRL** (Jaksch et al., 2010) designed to handle irreversible actions. As **UCRL**, **A-ROGUE** computes an optimistic policy, but before being executed, each optimistic action is first *tested*. If necessary, **A-ROGUE** asks for external information in an *active* way via rollouts. **A-ROGUE** tries to minimize the number of rollouts used and is thus rollout efficient.

3.1 High-level Description of A-ROGUE

Episodes, plausible MDPs, and optimistic policy Similar to **UCRL**, **A-ROGUE** proceeds in episodes[1] $k = 1, \ldots, K$ of increasing length. At the start of each episode, a set of plausible MDPs $\mathcal{M}_k$ is determined (using confidence intervals for rewards and transition probabilities, respectively), cf. line 4. The set $\mathcal{M}_k$ is updated only when confidence intervals have changed a lot. In this case a new episode begins (line 6). Given the set $\mathcal{M}_k$, an optimistic policy π_k^+ that maximizes the reward over all plausible MDPs in $\mathcal{M}_k$ and all policies is computed (line 5).

Checking for irreversibility Unlike **UCRL**, **A-ROGUE** does not naively play the actions suggested by the optimistic policy. Instead, at each time step it is first tested whether the optimistic action recommended by π_k^+ is potentially irreversible. Thus, **A-ROGUE** estimates the time needed to recover from this action (line 9), that is, to come back to the current state on a shortest trajectory after executing the action. The respective *backup policy* computed uses a *pessimistic* estimate in order to make sure that we can indeed come back to the reference state. An optimistic policy would obviously underestimate the real time to reach the reference state and may not be able to avoid traps. If the estimate of the recovery time is small enough (line 10), **A-ROGUE** plays the optimistic action, as there is a reliable way to get back to the reference state.

1. Note that the word "episode" her comes from Jaksch et al. (2010) and only refers to internal stages of the algorithm. It has nothing to do with episodic learning, and does not involve any reset to an initial state. As already stated, our algorithm interacts with the environment in a single stream of state-action-rewards with no reset.

Testing for optimality If the estimate of the recovery time is large, the optimistic action is considered irreversible, and **A-ROGUE** evaluates whether the action is optimal (lines 12f). If **A-ROGUE** decides that the action is optimal (considering the optimality gap of the optimistic action), it is executed. In this case also the reference state is updated. Indeed, in the analysis it turns out to be important that the reference state is only updated after playing a near-optimal action with large return time. Indeed, updating the reference state at each time step to be the current state and thus computing a return path to the last visited state at each time step is not safe: Say each return path is guaranteed to be at most c-steps long, after executing m actions, one could only ensure a return path of at most cm-steps, that is linear in the number of executed steps. This is not enough to ensure sub-linear regret guarantee in a worst case scenario. In contrast, updating the reference state and computing a return path to the last visited state reached *after playing a near-optimal action with large return time* enables to get a controlled regret.

Asking for rollouts Otherwise, if the optimistic action could not be identified as optimal, the algorithm needs more information and asks for a rollout (line 16). The policy that the algorithm asks the rollout for is not always the same. Alternatingly, the rollout is either asked for the optimistic policy π_k^+ (which helps to decide whether the algorithm shall enter a transient region), or for the policy that plays the optimistic action suggested by π_k^+ followed by the pessimistic backup policy (which helps improving the estimate of the return time to the reference state). The length of the rollout is of the order of the local diameter. However, the rollout is also stopped when the reference state is reached (indicating that the return time needs to be updated), or when the episode termination criterion is met, that is, when confidence intervals have changed a lot.

Remark 3 *The choice of* **UCRL** *as base algorithm that is modified to work for multi-chain MDPs is to a certain extent arbitrary. We think that the presented methods could be adapted to work with different algorithms such as* **R-Max** *as well. However, for* **UCRL** *we have access to regret analysis that results in bounds nearly matching the known lower bounds. Thus,* **UCRL** *is a natural choice.*

3.2 Details of A-ROGUE

After giving the high-level description, we are now ready to give the full technical details of **A-ROGUE**, cf. Algorithm 2.

Parameters Beside the horizon T and a confidence parameter δ, the algorithm uses a parameter $\varepsilon > 0$ and a function $d : \mathbb{N} \to \mathbb{R}$ in order to control the length of rollouts. That is, the length of a rollout is bounded by $(1 + \frac{1}{\varepsilon})D_k$, where $D_k = d(t_k)$ is an estimate for an upper bound on the local diameter of the MDP. Obviously, ε needs to be sufficiently small to guarantee that rollouts are long enough to provide the necessary information, but not too small to avoid incurring large regret for the steps of the rollout.

The parameter γ controls how conservative the algorithm behaves. Thus, for $\gamma = 0$, **A-ROGUE** won't execute any irreversible action with high-probability. This can be interesting if we know that all irreversible actions are bad. As γ gets larger, **A-ROGUE** behaves more and more like **UCRL**.

Algorithm 1 A high-level description of **A-ROGUE**

Set current time step $t := 1$.
for episodes $k = 1, 2, \ldots$ **do**
  Let the reference backup state $\mathbf{s}_k$ be the current state s_t.
  Define set of plausible MDPs $\mathcal{M}_k$ based on past observations.
  Compute *optimistic policy* π_k^+ that maximizes the reward over all MDPs in $\mathcal{M}_k$.

  while confidence intervals haven't changed too much **do**
    Compute a *pessimistic backup* policy $\pi_k^{\circlearrowleft -}$ that tries to reach the reference state $\mathbf{s}_k$ after playing the optimistic action $a_t^+ = \pi_k^+(s_t)$.
    Compute the expected pessimistic recovery time to reach $\mathbf{s}_k$.
    if the expected *time to reach* $\mathbf{s}_k$ after playing a_t^+ is small enough **then**
      Phase **I***:* Execute a_t^+, get reward r_t and next state s_{t+1}. Set $t := t + 1$.
    else
      if the estimated *optimality gap* of a_t^+ is small enough **then**
        Phase **II***:* Execute a_t^+, get reward r_t and next state s_{t+1}.
        Update the reference state $\mathbf{s}_k = s_{t+1}$. Set $t := t + 1$.
      else
        Phase **III***:* Depending on the last rollout, ask either for a rollout of the optimistic policy π_k^+ (if the last rollout was for the backup policy), or for a rollout that first plays a_t^+ and then the backup policy $\pi_k^{\circlearrowleft -}$ (if the last rollout was for the optimistic policy). Stop when the confidence intervals have changed a lot, or $\mathbf{s}_k$ is reached, or the rollout is too long. Set $t := t + 1$ for each transition.
      end if
    end if
  end while
end for

Algorithm 2 The **A-ROGUE** algorithm

Require: Horizon T, confidence δ, lower bound γ on sharpness $\gamma^\star$, function $d : \mathbb{N} \to \mathbb{R}$ for guessing the diameter, parameter $\varepsilon \in (0,1)$ for control of length of rollouts, initial state s_1.

Set $t := 1$.
for $k = 1, 2, \ldots$ **do**
  Let $t_k := t$ (starting time of episode k) and set the reference backup state $\mathbf{s}_k = s_{t_k}$.
  Let $\mathcal{M}_k := \mathcal{M}_{t_k}(\delta/T^2)$.
  Compute $\widehat{M}_k^+ = (r_k^+, p_k^+)$ that satisfies $V_{T-t_k}^{M_k^+}(s_{t_k}) = \sup_{M \in \mathcal{M}_k} V_{T-t_k}^{M}(s_{t_k})$ as well as respective value Q_k^+ and policy π_k^+.
  Let $\mathtt{Stop}_k(t) := \{\exists s, a : N_t(s,a) \geqslant 2N_{t_k}(s,a) \vee 1\}$.
  while not $\mathtt{Stop}_k(t)$ **do**
    Compute backup MDP $M_k^{\circlearrowleft -} = (r_k^{\circlearrowleft}, p_k^{\circlearrowleft -})$ with reference state $\mathbf{s}_k$, horizon $\frac{2}{\gamma}(1 + \frac{1}{\varepsilon})D_k$, backup values $Q_k^{\circlearrowleft -}$ and policy $\pi_k^{\circlearrowleft -}$, and $D_k = d(t_k)$.
    Compute the expected pessimistic recovery time $\mathcal{T}_t^- = |Q_k^{\circlearrowleft -}(s_t, \pi_k^+(s_t))|$.
    if $\mathcal{T}_t^- < 2D_k$ **then**
      Phase **I***:* Execute $a_t^+ = \pi_k^+(s_t)$, get reward r_t and next state s_{t+1}. Set $t := t+1$.
    else
      if $\Delta_{T-t}^-(s_t, a_t^+) < \gamma(T-t)$ **then**
        Phase **II**: Execute a_t^+, get reward r_t and next state s_{t+1}.
        Update the reference backup state $\mathbf{s}_k = s_{t+1}$, $t := t+1$.
      else
        Phase **III**: Depending on the last rollout, ask either for a rollout of π_k^+ (if the last rollout was for a backup policy $\pi^{\circlearrowleft}$), or for a rollout that first plays a_t^+ and then the backup policy $\pi_k^{\circlearrowleft -}$ (if the last rollout was for an optimistic policy π^+). The rollout is stopped when either $\mathtt{Stop}_k(t)$ happens, or $\mathbf{s}_k$ is reached, or its length is more than $(1 + \frac{1}{\varepsilon})D_k$. Set $t := t+1$ for each observed transition.
      end if
    end if
  end while
end for

Plausible MDPs (line 4). In episode k starting at time t_k, **A-ROGUE** defines, using the past observations, the empirical reward function $\widehat{r}_{t_k}$ and the empirical transition kernel $\widehat{p}_{t_k}$. They are used to define the set $\mathcal{M}_k$ of plausible MDPs as follows:

Definition 4 *The set $\mathcal{M}_t(\delta)$ of δ-plausible MDPs at step t compatible with the empirical reward function $\widehat{r}_t$ and empirical transition kernel $\widehat{p}_t$ is the set of all MDPs $\langle \mathcal{S}, \mathcal{A}, \tilde{p}, \tilde{\nu} \rangle$ with mean reward function $\tilde{\mu} : \mathcal{S} \times \mathcal{A} \to [0,1]$ and transition kernel $\tilde{p} : \mathcal{S} \times \mathcal{A} \to \Gamma(\mathcal{S})$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|\widehat{r}_t(s,a) - \tilde{\mu}(s,a)\right| \;\leqslant\; C_t^{\mu}(s,a) := \sqrt{\tfrac{\log(2SA/\delta)}{2N_t(s,a)}}, \tag{2}$$

$$\left\|\widehat{p}_t(\cdot|s,a) - \tilde{p}(\cdot|s,a)\right\|_1 \;\leqslant\; C_t^{p}(s,a) := \sqrt{\tfrac{\log(2^{S+1}SA/\delta)}{2N_t(s,a)}}, \tag{3}$$

where $N_t(s,a) := \sum_{i=1}^{t} \mathbb{I}\{s_i = s, a_i = a\}$ is the number of visits of the pair (s,a) up to time t.

[2]

We prove later in Section 5.1.2 that the true MDP is plausible with high probability:

Lemma 5 *With probability higher than $1 - 2\delta/T$, the true model is plausible simultaneously for all episodes k. That is, $M^\star \in \mathcal{M}_{t_k}(\delta/T^2)$ for all k.*

Remark 6 *Both quantities $C_t^{\mu}(s,a)$ and $C_t^{p}(s,a)$ could be sharpened if some further knowledge is given about the reward and transitions. Also, following Filippi et al. (2010) we recommend to use KL-based confidence-bounds instead of* (3) *for practical applications.*

Extended Value Iteration (line 5). The so-called optimistic MDP $\widehat{M}_k^+ = (r_k^+, p_k^+) \in \mathcal{M}_k$ maximizes the optimal value $V_{T-t_k}^{M}(s_{t_k})$ among all plausible MDPs M and can be computed e.g. by Extended Value Iteration, see Jaksch et al. (2010) via an *augmented* MDP which we denote $\mathcal{M}_k^+$. Q_k^+ is the associated optimistic Q-function (with horizon $T - t_k$), and π_k^+ the respective optimistic policy (that is, the greedy policy w.r.t. Q_k^+). We also use Q_{T-t} to denote the Q-value function for the true MDP $M^\star$ and horizon $T - t$.

Estimating the return time (line 8–9). Before executing the optimistic action $a_t^+ = \pi_k^+(s)$ in some state s, **A-ROGUE** tests if there is a short path back to the reference backup state $\mathbf{s}_k$ after playing a_t^+. That is, it tests if the action is reversible. This is a simple stochastic shortest path (SSP) problem: Defining a deterministic reward function by $r_k^{\circlearrowleft}(s) = -1$ if $s \neq \mathbf{s}_k$, and 0 else, consider the set of δ/T^2-plausible backup MDPs $\mathcal{M}_k^{\circlearrowleft} = \{\tilde{M} = (r_k^{\circlearrowleft}, \tilde{p}) : \tilde{p} \text{ satisfies } (3)\}$. Any SSP algorithm (or also Extended Value Iteration) can be used to compute the pessimistic return policy $\pi_k^{\circlearrowleft -}$ to the backup state $\mathbf{s}_k$, together with the pessimistic backup values $Q_k^{\circlearrowleft -}$. Both quantities are defined similarly to Q_k^+ and π_k^+, but using the set of MDPs $\mathcal{M}_k^{\circlearrowleft}$ instead of $\mathcal{M}_k$. Since $r_k^{\circlearrowleft}$ is negative, the Q-values are also negative, and $\mathcal{T}_t^- = |Q_k^{\circlearrowleft -}(s_t, a_t^+)|$ represents the minimal expected return time from s_t

2. Further, in case it is known that each state has only L successor states, one can replace 2^S with 2^L in the definition of $C_t^p(s,a)$.

back to $\mathbf{s}_k$ after playing action $a_t^+ \stackrel{\text{def}}{=} \pi_k^+(s_t)$, in the least favorable plausible backup MDP. Importantly, $Q_k^{\circlearrowleft -}$ is computed with time horizon $\frac{2}{\gamma}(1+\frac{1}{\varepsilon})D_k$ (as opposed to $T-t_k$ for Q_k^+), which implies that $\mathcal{T}_t^-$ is also less than $\frac{2}{\gamma}(1+\frac{1}{\varepsilon})D_k$. Note that we do not want to run rollouts for too long since we incur a linear regret along each rollout.

Deciding whether to act or to ask for a rollout (lines 10–13). If the estimated return time $\mathcal{T}_t^-$ is $< 2D_k$,[3] where D_k is an estimate of the local diameter of the MDP, **A-ROGUE** executes the optimistic action $a_t^+ = \pi_k^+(s_t)$. Otherwise, if $\mathcal{T}_t^- \geqslant 2D_k$, a_t^+ is considered to be an irreversible action, since executing this action may lead to suffer a higher regret than the worst-case regret one would suffer for playing an action in a MDP with diameter D_k. Thus, in the following we will refer to the test in line 10 of the algorithm as the *recovery test* and say that the test is passed, if $\mathcal{T}_t^- < 2D_k$.

To decide if playing a_t^+ is good or not, **A-ROGUE** computes a pessimistic estimate for the *action gap*. Intuitively, we want to test whether the quantity $\max_a Q_{T-t}(s,a) - Q_{T-t}(s,a_t^+)$ is small. If it is smaller than the minimal action gap $\min_{a:\Delta_{T-t}(s,a)>0}\Delta_{T-t}(s,a)$ of s, then the action a_t^+ must be optimal. Since we do not have direct access to Q_{T-t}, the used optimality test uses the set of plausible models $\mathcal{M}_k$ instead to compare to the action gap in the worst case plausible MDP:

$$\Delta_{T-t}^-(s_t,a_t^+) \triangleq \sup_{M\in\mathcal{M}_k}\Big(\max_a Q_{T-t}^M(s_t,a) - Q_{T-t}^M(s_t,a_t^+)\Big)\,.$$

(The supremum is actually a maximum, as it is enough to optimize over the simplex.) When $\Delta_{T-t}^-(s_t,a_t^+) < \gamma(T-t)$, this means that the action gap is indeed small, thus action a_t^+ is considered to be optimal and hence played. In the following, we refer to the test in line 13 as the *optimality test* and say that an action a_t^+ passes the test, if $\Delta_{T-t}^-(s_t,a_t^+) < \gamma(T-t)$.

Asking for a rollout (line 17). In case a_t^+ does not pass the optimality test, **A-ROGUE** asks for an exploratory rollout from s_t. As already explained, the algorithm alternates the policy for which a rollout is requested. Either **A-ROGUE** requests a rollout of the optimistic policy π_k^+, or (if the last rollout has been requested for the optimistic policy) a rollout of the non-stationary policy that first executes $a_t^+ = \pi_k^+(s_t)$ and then policy $\pi_k^{\circlearrowleft -}$, the pessimistic policy that tries to come back as fast as possible to the backup state $\mathbf{s}_k$. The rollout is stopped if either (i) the number of visits in some state-action pair has doubled (when also the episode terminates), (ii) the backup state $\mathbf{s}_k$ is reached, or (iii) the length of the rollout exceeds $\big(1+\frac{1}{\varepsilon}\big)D_k$.

4. Performance Guarantees

The following theorem is, to the best of our knowledge, the first sublinear regret guarantee that holds for a large class of multi-chain MDPs with possibly infinite diameter. For convenience, it is stated and proved assuming that all regions are directly reachable from the region $\mathfrak{S}_{i_0}^\star$ where the learner starts. This assumption is relaxed below, see Remark 8.

3. To account for estimation errors of the return time, the algorithm tests if the return time is less than $2D_k$ (and not simply D_k).

Theorem 7 (Bound on active regret) *Let $M^\star$ be an MDP with local diameter $d^\star$ and $\gamma^\star$-sharp action-gaps that satisfies Assumption 1 with access probability $\geqslant p_0$. Assume also that all transition probabilities are either 0 or $\geqslant p_{min} > 0$. Given an upper bound $D \geqslant d^\star$, the active regret of* **A-ROGUE** *with parameters δ, $d(t) = D$, $\gamma \leqslant \gamma^\star$, and $\varepsilon < \min\left\{\frac{p_0 d^\star}{p_0 d^\star + p_0 + d^\star}, \frac{p_0}{1+p_0}\right\}$ after T steps is upper bounded by*

$$\begin{aligned}\mathbb{E}\big[\mathfrak{R}_T\big] \;\leqslant\; & \left(5+\tfrac{1}{p_0}\right)\sqrt{\tfrac{8d^\star}{\gamma^\star}\left(T+\tfrac{2d^\star}{\gamma^\star}\right)} + \left(4+\tfrac{1}{p_0}\right) d^\star SA \log_2\left(\tfrac{8T}{SA}\right) \\ & + 6(\sqrt{2}+1)\left(2+\tfrac{1}{p_0}+\tfrac{1}{p_{min}}\right) d^\star \sqrt{\log(2^{S/2+1}AST^2/\delta)}\sqrt{SAT} \\ & + \underbrace{16\frac{SD^3(1+\varepsilon)^4}{\gamma^2\varepsilon^3(1-\varepsilon)}\log(2^{S+1}SAT^2/\delta)}_{\textit{Bound on number of asked rollout transitions}} + \tfrac{1+\varepsilon}{\gamma\varepsilon}D + 2\delta\,.\end{aligned}$$

Simplified, the regret is of order

$$\mathbb{E}\big[\mathfrak{R}_T\big] = O\Bigg(\underbrace{\frac{d^\star S\sqrt{A\log(\mathbf{T}AS/\delta)} + \sqrt{\frac{d^\star}{\gamma^\star}}\sqrt{\mathbf{T}} + \frac{d^\star SA}{p_0}\log\left(\frac{\mathbf{T}}{SA}\right)}{p_0 \wedge p_{min}}}_{\text{Regret when acting}} + \underbrace{\frac{D^3S^2}{\gamma^2\varepsilon^3}\log\left(\frac{\mathbf{T}AS}{\delta}\right)}_{\text{Extra information asked}}\Bigg).$$

This bound captures the essential behavior of the algorithm: The first term is basically of the same order as the regret of **UCRL** (Jaksch et al., 2010). This term also gives the long-term regime of the algorithm (after it identified irreversible actions). The remaining term gives the number of transitions in rollouts the algorithm asked for. This happens in two situations: first when the algorithm wants to find out whether an action is irreversible, and second when it needs to know whether an action identified as irreversible is optimal. While the irreversibility of an action is rather cheap to identify and can be accomplished by rollouts of length $O(D_k)$, the second case needs much more information, as it depends on the action-gap of the current state. Indeed, before playing an irreversible action the algorithm needs an accurate estimate not only of the value *inside* the region it intends to enter, but also of the value of states *outside* the region. Note that the algorithm not only suffers big loss when entering a bad region, it also incurs some regret when *not* entering an optimal region.

Remark 8 *If no bound on $d^\star$ is available, one can use $d(t) = \log(t)$ instead. We get a similar bound, where one needs to replace D^2 in the first operand of the* max *by $e^{d^\star}$, and D in the second operand of the* max *by* $\log(T)$.

Remark 9 *If all regions except the initial one cannot be escaped a.s., then one can replace the term $\sqrt{\frac{8d^\star}{\gamma^\star}}\sqrt{T+\frac{2d^\star}{\gamma^\star}}$ in Theorem 7 with a constant. This term is indeed due to the possibility of escaping a region and re-entering it possibly several times, which in a worst case produces this additional regret term.*

Remark 10 *If not all regions are directly reachable from $\mathfrak{S}^\star_{i_0}$ but some are only reachable by crossing, say, ℓ intermediate regions, then the bound of Theorem 7 holds with $d^\star$ replaced with $(\ell+1)d^\star$ and D replaced with $(\ell+1)D$.*

Remark 11 *As a special case, the result holds as well in the case when the MDP is indeed weakly communicating and has small diameter D. Then $d^\star = D$, $\gamma^\star = \infty$, and $p_0 = 1$. Note that in this case* **A-ROGUE** *with parameter $\gamma = \infty$ coincides with* **UCRL**.

Overview of the proof Before providing a fully detailed proof in Section 5, let us give some high-level sketches in order to highlight the main steps of the proof. The proof is based on ideas similar to those for showing the regret bound for **UCRL** in Jaksch et al. (2010). However, we need to handle four additional difficulties: 1) the different behavior of **A-ROGUE** in each episode (cf. phases I, II, and III in lines 10, 13, and 15 of the algorithm), 2) the accurate estimation of the return time $\mathcal{T}_t^-$ to identify irreversible actions, 3) the decisions to play irreversible actions or not, and 4) the total number of roll-out transitions asked by the algorithm.
First, the case when no irreversible action is played by the algorithm nor the optimal policy can be handled by standard concentration bounds and using the local diameter $d^\star$, similar to the already mentioned proof in Jaksch et al. (2010).
When irreversible actions come into play, regret is incurred in two cases: when we play an irreversible action that leads to a bad region, or when we do not play an irreversible action that leads to a good region early enough. For the first case, we show that an action that passes the test of line 13 must be $d^\star$-optimal with high probability, and that the estimated return time $\mathcal{T}_t^-$ for an irreversible action must be larger than $2D_k$ with high probability. This prevents the algorithm from playing bad irreversible actions. For the second case, we show that any $d^\star$-optimal action will pass the test of line 13 for small enough confidence bounds, depending on the γ-sharpness of the MDP. The last part of the proof is to show that the confidence bounds indeed shrink fast enough, and to handle the number of roll-out transition requested by **A-ROGUE**. This is a tricky part of the proof, where the length of the roll-outs as well as the policy chosen to be executed must be chosen carefully. Here, we first show that when the optimal return time in the true MDP $\mathcal{T}_t = |Q_k^{\circlearrowleft\star}(s_t, a_t)|$is small, so is $\mathcal{T}_t^- = |Q_k^{\circlearrowleft -}(s_t, a_t)|$, so that the algorithm executes the optimistic policy (and does not ask for a roll-out). Then, we use the observation that by definition of $d^\star$, and by Assumption 1, each region can be explored from its boundary by $O(d^\star)$-long roll-outs. We show that a controlled number of roll-outs of length $(1 + \frac{1}{\varepsilon})D_k$ of the policies $\pi^{\circlearrowleft -}$ and π_k^+ is enough to shrink the confidence bounds fast enough, leading to the final sub-linear regret.

5. Regret Analysis (Proof of Theorem 7)

In this section, we now provide a complete proof of Theorem 7. As the proof is a bit long, it is divided into several parts. At the core the proof follows the lines of proof of the regret bound of the regret bound for **UCRL** (Jaksch et al., 2010) However, there are several additional difficulties we need to take care of: 1) the different behavior of **A-ROGUE** in each episode (lines 10, 13, and 15 of the algorithm), 2) the accurate estimation of the return time $\mathcal{T}_t^-$, 3) dealing with irreversible actions and 4) the total number of roll-out transitions asked by the algorithm. The proof provides a few innovations in order to handle these difficulties. Each of them may be of independent interest. We first give a brief overview of the single parts of the proof.

- Sections 5.1.1 and 5.4: Decomposing the cumulated reward / Bellman equation / confidence bounds.

  In these first two sections, we provide an initial decomposition of the cumulative reward, disregarding regret due to entering another region, so as to provide the main picture of the proof. We use a new decomposition in Section 5.1.1 and control the regret backward from the last episode to earlier episodes, as opposed to standard proofs. A second innovation is that, when introducing the Bellman equation, following the proof technique from Jaksch et al. (2010) (see Section 5.4), we replace the span of the optimistic value function by a more accurate local span over the immediate successor of the current state, and control it using the local diameter of Definition 1.

- Section **??**: Dealing with irreversible transitions.

  We modify the decomposition from Sections 5.1.1 and 5.4, to control what happens when the optimal policy plays an irreversible action. One specific difficulty is to handle the case when we need to play an irreversible action to enter an optimal region but we do not play it for lack of confidence, thus causing possibly linear regret during this time lapse. We show in Lemma 12 that an action that passes the test in l.13 must be $d^\star$-optimal (if the model $M^\star$ is plausible), and that every $d^\star$-optimal action will pass the test in l.13 when confidence intervals become small enough. Here the notion of local diameter and γ-sharpness play an important role in defining the tests used in **A-ROGUE**. This result enables us to ensure that, provided that some states are visited often enough (so that confidence intervals become small), we eventually enter the optimal region after a controlled number of steps, and thus can control the regret.

- Section 5.1.2: Confidence intervals and high probability events.

  We derive simple confidence bounds that characterize the uncertainty in our estimates of the reward and transition probability distributions as a function of the number of samples observed from each state-action pair and other relevant factors. In addition to showing how the learned model improves with samples (as is necessary in the regret analysis of **UCRL**), these confidence intervals also allow us to bound the minimum number of steps to return to the current backup state (which is unique to **A-ROGUE**). Specifically, we provide in Section 5.1.4 a bound on the estimated return time, by seeing it as a value function. We thus use the definition of the backup MDP and a Bellman decomposition similar to the one used for the MDP model in order to control the estimation error.

- Section 5.5: Total number of roll-outs.

  In Section 5.5, we bound the number of episodes and the number of simulated roll-outs requested by **A-ROGUE**. This is a tricky part of the proof, where we have to show that the length of the roll-outs as well as the roll-out policy are well chosen by the algorithm. This is an important innovation of the proof, where we study in detail the number of roll-outs of some length d from some policy π needed to ensure that a state-action that is reachable in d steps is visited a certain number of n times on average. We then use this to control the total number of roll-outs from such a policy π that are needed in order to visit often enough the set of all states reachable in d steps from

one point, and apply it to $\pi^{\circlearrowleft -}$. We believe that such results (as given in Section **??**) are of independent interest. We then show that the backup-policy satisfies the desired property, and thus get a control on the number of roll-out transitions needed so that we have an accurate estimation of the back-up time.

- Section 5.6: Regret.

  In this last section, we bound the regret of **A-ROGUE** first when the diameter is unknown and second when an upper bound for the diameter is known. This section, ties together the findings in the previous sections of the appendix and provides the main proof of Theorem 7.

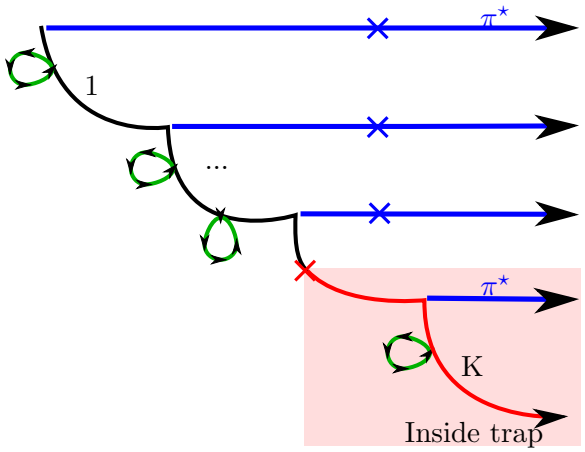


Figure 7: A high over-view of the algorithm and of the proof: In black, the trajectory followed by one execution of **A-ROGUE** divided into K episodes (internal to the algorithm). In blue, optimal policies starting from the starting point of each **A-ROGUE**-episode. We consider a case in which it is optimal to enter a trap: the blue cross indicates when the optimal policy enters it, the red cross when **A-ROGUE** enters it. Finally, the green loops symbolize parts of episodes when roll-outs are asked and during which we accumulate regret for not acting. The proof starts by capturing what happens in the last episode K (Section 5.2.1), and progressively compares the algorithm to an optimal policy run from earlier and earlier starting points (Section 5.2.2). The special point when **A-ROGUE** enters the trap is dealt with separately in section **??**. The total number of roll-out transitions (green parts) is handled in Section 5.5.

Let us start by introducing notations. The algorithm proceed in T time-steps $t \in \{1, \dots, T\}$. We denote this set $\mathbb{T}_T$, and introduce more generally for all $t \in \mathbb{N}$ the notation $\mathbb{T}_t = \{1, \dots, t\}$, with the convention that $\mathbb{T}_0 = \emptyset$. We make a distinction between the time-steps when an action is output, which is called a decision-step, and the time-steps when an roll-out transition is observed, which we call a roll-out step. We denote $\mathbb{T}_t^D \subset \mathbb{T}_t$ the subset of time-steps corresponding to decision-steps until time-step t, and $\mathbb{T}_t^R = \mathbb{T}_t \setminus \mathbb{T}_t^D$ the corresponding subset of roll-out steps.

The algorithm proceed in successive episodes. The number of episodes until time T is a random variable that we denote K_T. For $k \in \{1, \dots, K_T\}$, episode k starts at $t_k \in \mathcal{T}_T$. Thus $t_1 = 1$, $t_{k+1} - t_k$ is the length of episode k, and we denote by convention $t_{K_T+1} = T + 1$. Thus, the last episode starts at $t_{K_T} \in \mathbb{T}_T$, and stops at time $T = t_{K_T+1} - 1$. With this convention, we write $\mathbb{T}_{(k)} = \mathbb{T}_{t_{k+1}-1} \backslash \mathbb{T}_{t_k-1}$ the time-steps corresponding to episode k, for $k \in \{1, \dots, K_T\}$, and similarly, define $\mathbb{T}^D_{(k)} = \mathbb{T}^D_{t_{k+1}-1} \setminus \mathbb{T}^D_{t_k-1}$ and $\mathbb{T}^R_{(k)} = \mathbb{T}^R_{t_{k+1}-1} \setminus \mathbb{T}^R_{t_k-1}$. We also introduce $\mathbb{T}_{k,t} = \mathbb{T}_{(k)} \cap \mathbb{T}_t$, the set of time-steps between t_k and t (that is $\mathbb{T}_{k,t} = \{t_k, \dots, t\}$), as well as the corresponding set $\mathbb{T}^D_{k,t} = \mathbb{T}^D_{(k)} \cap \mathbb{T}^D_t$ of decision-steps and $\mathbb{T}^R_{k,t} = \mathbb{T}^R_{(k)} \cap \mathbb{T}^R_t$ of roll-out steps in between steps t_k and t. Finally, we use a specific notation for special time-steps in an episode: $\underline{t}^D_k = \operatorname{argmin}_{t \in \mathbb{T}^D_{(k)}} t$ and $\bar{t}^D_k = \operatorname{argmax}_{t \in \mathbb{T}^D_{(k)}} t$ are the first and last decision-steps in episode k. Likewise, $\underline{t}^R_k = \operatorname{argmin}_{t \in \mathbb{T}^R_{(k)}} t$ and $\bar{t}^R_k = \operatorname{argmax}_{t \in \mathbb{T}^R_{(k)}} t$ are the first and last roll-out-steps in episode k. We also introduce the first and last time-steps of $\mathbb{T}_{(k)}$, $\underline{t}_k = \operatorname{argmin}_{t \in \mathbb{T}_{(k)}} t = t_k$ and $\bar{t}_k = \operatorname{argmin}_{t \in \mathbb{T}_{(k)}} t = t_{k+1} - 1$, for coherence.

Regarding states, starting from the same initial state s_1, s_t denotes a state reached by **A-ROGUE** at time-step t, and $s^\star_t$ one reached by the optimal policy $\pi^\star_T$ with horizon T at the same time-step. We say that the algorithm enters a region $\mathfrak{S}^\star_i$ at time-step t if $s_{t-1} \notin \mathfrak{S}^\star_i$ and $s_t \in \mathfrak{S}^\star_i$. We say it leaves region $\mathfrak{S}^\star_i$ at time-step t if $s_{t-1} \in \mathfrak{S}^\star_i$ and $s_t \notin \mathfrak{S}^\star_i$. In the sequel, we denote by $\underline{\mathbf{t}}^m_i \in [-\infty, +\infty]$ the m-th smallest time-steps when the algorithm enters region $\mathfrak{S}^\star_i$, and $\overline{\mathbf{t}}^m_i \in [-\infty, +\infty]$ the m-th smallest time-steps when it leaves region $\mathfrak{S}^\star_i$. We also denote $\underline{\mathbf{t}}$ the first time-step when the algorithm leaves the initial region, and $\overline{\mathbf{t}}$ the last time-steps when the algorithm leaves any region. Finally, we denote $\overline{\mathbf{k}}$ the episode that contains $\overline{\mathbf{t}}$, that is such that $\overline{\mathbf{t}} \in \mathbb{T}_{(\overline{\mathbf{k}})}$.

5.1 Preliminary results

Before presenting the core of the proof of Theorem 7, we start by introducing a few useful results ans lemmas.

5.1.1 Decomposing the cumulated reward.

The expected active regret $\mathbb{E}\Big[\mathfrak{R}_T\Big]$ of **A-ROGUE** at time T can be decomposed using the cumulative rewards received during each episode $k \in \mathbb{N}$ until the last episode K_T.

In the sequel, if a (possibly non-stationary) policy $\pi = (\pi_t)_{t \geqslant 1}$ is executed in an MDP M at (not necessarily consecutive) times $t' \in \tilde{\mathbb{T}} = \{t_1, \dots, t_\tau\}$, starting from state s at time t_1 then we define its accumulated reward by $R^\pi_M(\tilde{\mathbb{T}}|s)$. Its is defined by

$$R^\pi_M(\tilde{\mathbb{T}}|s) = \sum_{t' \in \tilde{\mathbb{T}}} r(s_{t'}, \pi_{t'}(s_{t'}))\,,$$

where $\pi_{t'}$ refers to the policy π at time t'. Note that the time matters since the policy π needs not be stationary. Typically, an optimal strategy executes $\pi^\star$ at each time step in true MDP $M^\star$, whereas **A-ROGUE** executes an optimistic policy π_k in episode k in $M^\star$ only when it does not ask for roll-out transitions, that is when $t \in \tilde{\mathbb{T}} = \mathbb{T}^D_{(k)}$. Since **A-ROGUE** receives no rewards during the execution of a roll-out, the reward accumulated during episode k, from t_k up to $t_{k+1} - 1$, is

$$R^{\mathbb{A}}_{M^\star,k} = R^{\pi_k}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\,.$$

Then, using these notations, we can rewrite the regret as

$$\begin{aligned}
\mathbb{E}\Big[\mathfrak{R}_T\Big] &= \mathbb{E}\Big[\sum_{t=1}^{T} r(s_t^\star, a_t^\star)\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} r(s_t, a_t)\Big] \\
&= \mathbb{E}\Big[R_{M^\star}^{\pi^\star}(\mathbb{T}|s_1) - \sum_{k=1}^{K_T}\Big(\sum_{t\in\mathbb{T}_{(k)}^D} r(s_t,a_t) + \sum_{t\in\mathbb{T}_{(k)}^R} 0\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K_T}\Big(R_{M^\star}^{\pi^\star}(\mathbb{T}_{(k)}|s_{t_k}^\star) - R_{M^\star}^{\pi_k}(\mathbb{T}_{(k)}^D|s_{t_k})\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K_T}\Big(R_{M^\star}^{\pi^\star}(\mathbb{T}_{(k)}|s_{t_k}^\star) - R_{M^\star,k}^{\mathbb{A}}\Big)\Big]\,. \qquad (4)
\end{aligned}$$

5.1.2 Confidence intervals and high probability events

In this section, we derive confidence intervals for the reward and transition kernel estimates. Let us introduce

$$\widehat{r}_t(s,a) = \frac{1}{N_{t-1}(s,a)}\sum_{t'=1}^{t-1} r(s_{t'},a_{t'})\mathbb{I}[s_{t'}=s, a_{t'}=a]\,,$$

where we recall that $N_{t-1}(s,a) = \sum_{t'=1}^{t-1}\mathbb{I}[s_{t'}=s,a_{t'}=a]$.

Since the reward function $r_{M_k^+}$ is a plausible reward function according to Definition 4, $r_{M_k^+}(s,a)$ is close to $\widehat{r}_{t_k}(s,a)$ by construction. On the other hand, $\widehat{r}_{t_k}(s,a)$ is close to the mean $\mu(s,a)$ by concentration of measure. We use Hoeffding's inequality together with a union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all possible values of $N_{t_k}(s,a)\leqslant t_k\leqslant T$, and deduce that with probability higher than $1-\delta/T$, simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all episodes $k\leqslant K_T$,

$$\begin{aligned}
&|r_{M_k^+}(s,a) - \mu(s,a)| \\
&\quad= \big|r_{M_k^+}(s,a) - \widehat{r}_{t_k}(s,a)\big| + \big|\widehat{r}_{t_k}(s,a) - \mu(s,a)\big| \\
&\quad\leqslant 2\sqrt{\frac{\log(2AST^2/\delta)}{2N_{t_k}(s,a)}} = 2C_k^\mu(s,a)\,, \qquad (5)
\end{aligned}$$

provided that the set of plausible MDPs chosen in **A-ROGUE** is $\mathcal{M}_t(\delta/T^2)$.

Following a similar argument but with specific inequality for the $\|\cdot\|_1$ control (Weissman et al., 2003), we obtain that with probability higher than $1-\delta/T$, simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, and all episodes $k\leqslant K_T$,

$$\|p_{M_k^+}(\cdot|s,a) - p(\cdot|s,a)\|_1 \leqslant 2\sqrt{\frac{\log(2^{S+1}AST^2/\delta)}{2N_{t_k}(s,a)}} = 2C_k^p(s,a)\,. \qquad (6)$$

Thus, combining (5) and (6) together with a union bound over all episodes, we obtain

Lemma 5 *On an event Ω of probability higher than $1-2\delta/T$, the true model is plausible (that is $M^\star \in \mathcal{M}_{t_k}(\delta/T^2)$) simultaneously for all episodes k.*

This Lemma will help us show that **A-ROGUE** avoids entering a bad region with high probability. Now, on the complement of this event, Ω^c, the regret is uncontrolled, and **A-ROGUE** incurs a possible linear regret, leading to an error term $\frac{2\delta}{T}T = 2\delta$. In the sequel, we introduce the notation $\mathbb{E}_\Omega[\cdot] \stackrel{\text{def}}{=} \mathbb{E}[\cdot\mathbb{I}\{\Omega\}]$. Thus, by Lemma 5, it holds

$$\mathbb{E}\Big[\mathfrak{R}_T\Big] \leqslant \mathbb{E}\Big[\mathfrak{R}_T\mathbb{I}\{\Omega\}\Big] + \frac{2\delta}{T}T = \mathbb{E}_\Omega\Big[\mathfrak{R}_T\Big] + 2\delta\,. \tag{7}$$

5.1.3 Sub-optimality test

In this section we consider the test on the sub-optimality gaps (Line 13). Without such a test, with high probability the algorithm would never decide to play an irreversible action even if this is the optimal thing to do. Indeed, since with high probability all played actions would not be irreversible due to test of Line 10 (See section 5.1.4), the algorithm would not enter another region. The regret can be linear in difference between the time $\mathbf{t}^\star$ when the optimal policy would have entered an optimal region to the time $\mathbf{t}$ the algorithm enters the same region. The test of line 13 on sub-optimality gaps prevents $\mathbf{t} - \mathbf{t}^\star$ (and thus the regret) to be too large. Note that in case when it is optimal to say in the initial region, $\mathbf{t}^\star = \infty$ and the difference is trivially bounded.

Lemma 12 *On the event Ω that $M^\star$ is plausible, an action a_t that passes the sub-optimality test in l.13 is $d^\star$-optimal:*

$$Q_{T-t}(s,a^\star) - Q_{T-t}(s,a_t) \leqslant d^\star\,.$$

Further, if a_t is $d^\star$-optimal, then it passes the sub-optimality test in l.13 as soon as

$$\|C_k^\mu\|_\infty + \left(2 + \frac{1}{p_0}\right) d^\star \|C_k^p\|_\infty < \frac{\gamma}{4}\,. \tag{8}$$

Note that if the test in l.13 is passed for a_t, this implies that no roll-out is asked at that time.

Recall that Q_{T-t} denotes the value function for the true model $M^\star$ at time t. Intuitively, we want to test at time t whether the quantity $\max_{a:Q_{T-t}(s,a_t)\neq Q_{T-t}(s,a)} Q_{T-t}(s,a) - Q_{T-t}(s,a_t)$ is small enough. If it is smaller than $\min_{a:Q_{T-t}(s,a^\star)\neq Q_{T-t}(s,a)} \Delta_t(s,a)$, that is, the action gap of s, then the optimistic action must be actually optimal. Since we do not have access to Q_{T-t}, the test of line 13 uses the plausible models $\mathcal{M}_k$ in episode k instead.

Step 1: We first show that if a_t is a *bad action*, that is $\Delta_{T-t}(s,a_t) = Q_{T-t}(s,a^\star) - Q_{T-t}(s,a_t) > d^\star$ then $\Delta^-_{T-t}(s,a_t)$ is big. Indeed, in this case and under the event $M^\star \in \mathcal{M}_k$

we have

$$\begin{aligned}
\Delta^-_{T-t}(s,a_t) &= \sup_{M\in\mathcal{M}_k}\Big(\max_{a\neq a_t} Q^M_{T-t}(s,a) - Q^M_{T-t}(s,a_t)\Big)\\
&\geqslant \sup_{M\in\mathcal{M}_k}\Big(Q^M_{T-t}(s,a^\star) - Q_{T-t}(s,a^\star) + Q_{T-t}(s,a^\star)\\
&\qquad\qquad -Q_{T-t}(s,a_t) + Q_{T-t}(s,a_t) - Q^M_{T-t}(s,a_t)\Big)\\
&\geqslant \Delta_{T-t}(s,a_t) + \sup_{M\in\mathcal{M}_k}\Big(Q^M_{T-t}(s,a^\star) - Q_{T-t}(s,a^\star) + Q_{T-t}(s,a_t) - Q^M_{T-t}(s,a_t)\Big)\\
&\geqslant \Delta_{T-t}(s,a_t)\\
&\geqslant \gamma^\star(T-t)\,,
\end{aligned}$$

where in the third line, we used the fact that since $M^\star \in \mathcal{M}_k$ by assumption, then the supremum term is non-negative, and where in the last line, we used the property that since for all s,a either $\Delta_{T-t}(s,a) \leqslant d^\star$, or $\Delta_{T-t}(s,a) > \gamma(T-t)$, then $\Delta_{T-t}(s,a_t)$ must be larger than $\gamma(T-t)$. Thus, provided that $\gamma \leqslant \gamma^\star$, an action a_t that leads to a bad region must satisfy $\Delta^-_{T-t}(s,a_t) \geqslant \gamma(T-t)$. We deduce that it is enough to test if $\Delta^-_{T-t}(s,a_t^+) < \gamma(T-t)$ holds to ensure that an optimistic action a_t^+ entering a bad region is discarded (with high probability). Other actions that pass this test (satisfy $\Delta^-_{T-t}(s,a_t^+) < \gamma(T-t)$) cannot enter a bad region, and are thus $d^\star$-optimal. In other words, if we play an action that passes this test, it must satisfy $Q_{T-t}(s,a^\star) - Q_{T-t}(s,a_t^+) \leqslant d^\star$. This proves the first part of Lemma 12.

Step 2: Now, we show that when an optimistic action a_t is $d^\star$-optimal, it passes the test after not too many time steps. To that end, we derive the following inequalities

$$\begin{aligned}
&\sup_{M\in\mathcal{M}_k}\Big(\max_{a\neq a_t} Q^M_{T-t}(s_t,a) - Q^M_{T-t}(s_t,a_t)\Big)\\
&\quad= \sup_{M\in\mathcal{M}_k}\Big(Q^M_{T-t}(s_t,a_M) - Q^+_{T-t}(s_t,a_t) + Q^+_{T-t}(s_t,a_t) - Q^M_{T-t}(s_t,a_t)\Big)\\
&\quad\leqslant \sup_{M\in\mathcal{M}_k}\Big(Q^M_{T-t}(s_t,a_M) - Q^+_{T-t}(s_t,a_M) + Q^+_{T-t}(s_t,a_t) - Q^M_{T-t}(s_t,a_t)\Big)\\
&\quad\leqslant \sup_{M\in\mathcal{M}_k}\Big(Q^+_{T-t}(s_t,a_t) - Q^M_{T-t}(s_t,a_t)\Big)\,, \qquad (9)
\end{aligned}$$

where we introduced $a_M = \operatorname{argmax}_{a\neq a_t} Q^M_{T-t}(s_t,a)$. In the second line we used the fact that $Q^+_{T-t}(s_t,a_t) \geqslant Q^+_{T-t}(s_t,a_M)$ since a_t is optimistic. In the third line, we used the fact that Q^+ is the optimistic Q-function. Note that there is no reason that a_t be the best action for a plausible MDP M.

Now, we want to control the difference of value between the successor states of a_t. If $s_t \in \mathfrak{S}_{i_0}$ and a_t is at the frontier of a region $\mathfrak{S}_i$, then several situations may occur: First, if all the successor states of a_t are inside the same communicating region $\mathfrak{S}_i$, then the difference of value between the successor states must be less than the diameter $d^\star$ of the region. Second, in other cases two successor states of a_t belong to different regions, say $\mathfrak{S}_i$ and $\mathfrak{S}_j$. Since a_t is $d^\star$-optimal by assumption, the difference of values between its successor states cannot exceed $d^\star$ plus the difference of values between the successor states of $s_t, a^\star$, where $a^\star$ is an optimal action in state s_t. Now the difference of values between successor

states $s \in \mathfrak{S}_i, s' \in \mathfrak{S}_j$ of $s_t, a^\star$ cannot be larger than $d^\star/p_0 + d^\star = (1 + 1/p_0)d^\star$: Indeed by assumption 1, at least one state-action pair $(\tilde{s}, a)$ in the region $\mathfrak{S}_i$ satisfies $p(\mathfrak{S}_j|\tilde{s}, a) \geqslant p_0$ without risking entering a third region. Now, it requires at most $d^\star$ steps to reach $\tilde{s}$ from a point in $\mathfrak{S}_i$, and thus at most $d^\star/p_0$ steps to enter $\mathfrak{S}_j$. On the other hand, it requires at most $d^\star$ steps to reach this point from $s' \in \mathfrak{S}_j$, and thus the difference of value is at most $\max\{(1 + 1/p_0)d^\star, d^\star\}$. Thus, in this second case, the difference of values between the successor states of a_t is bounded by $(2 + \frac{1}{p_0})d^\star$. We can now resort to a simple Bellman propagation error analysis to get the following inequality for all $M \in \mathcal{M}_k$

$$\begin{aligned} Q^+_{T-t}(s_t, a_t) - Q^M_{T-t}(s_t, a_t) &\leqslant \sum_{j=1}^{T-t} 4\big\|\delta_s P^{a_t} P^{\pi^\star T-t-j}\big\|_1 \Big(\big\|C^\mu_k\big\|_\infty + (2 + \frac{1}{p_0})d^\star \big\|C^p_k\big\|_\infty\Big) \\ &\leqslant 4(T-t)\Big(\big\|C^\mu_k\big\|_\infty + (2 + \frac{1}{p_0})d^\star \big\|C^p_k\big\|_\infty\Big), \end{aligned} \tag{10}$$

where $C^\mu_k(s, a)$ and $C^p_k(s, a)$ are the confidence bounds on the reward (2) and transition kernel (3), respectively. Here P^a denotes the $S \times S$ transition matrix with (s, s') entry $p(s'|s, a)$ (from the true MDP $M^\star$), and δ_s the vector of size S with s^{th} component equal to 1 and others equal to 0. Equation (10) is obtained by first expanding the Q-values using the Bellman equation, second inserting the true $M^\star$, third controlling the error on the reward and on the probability distributions with the confidence bounds (5), (6) and fourth using the above discussion to control the range of the value function by $(2 + \frac{1}{p_0})d^\star$. Note that in case a leads to an absorbing state, $\|\delta_s P^a P^{\pi^\star T-t-j}\|_1$ may be equal to one for each j.

An action a_t passes the test if $\Delta^-_{T-t}(s_t, a_t) \leqslant \gamma(T - t)$. Now combining (9) and (10), we have shown that on the event that $M^\star$ is plausible, an optimistic action a_t that is $d^\star$ optimal must satisfy

$$\Delta^-_{T-t}(s_t, a_t) \leqslant 4(T-t)\Big(\big\|C^\mu_k\big\|_\infty + \big(2 + \frac{1}{p_0}\big)d^\star \big\|C^p_k\big\|_\infty\Big)$$

Thus, we deduce that in order that such an action passes the test of line 13 it is enough that

$$\big\|C^\mu_k\big\|_\infty + \left(2 + \frac{1}{p_0}\right) d^\star \big\|C^p_k\big\|_\infty < \frac{\gamma}{4}, \tag{11}$$

which concludes the proof of Lemma 12. □

5.1.4 Estimate of the return time

In this subsection, we relate $\mathcal{T}^-_{\mathbf{t}} = -Q^{\circlearrowleft -}_k(s_{\mathbf{t}}, a_{\mathbf{t}})$ to the return time $\mathcal{T}_{\mathbf{t}} = -Q^{\circlearrowleft \star}_k(s_{\mathbf{t}}, a_{\mathbf{t}})$ in the true MDP. We want to show that $\mathcal{T}^-_{\mathbf{t}}$ is not too small when $\mathcal{T}_{\mathbf{t}}$ is large (the action $a_{\mathbf{t}}$ is irreversible), to avoid executing actions that we are not sure of. We also want that $\mathcal{T}^-_{\mathbf{t}}$ is not much larger than $\mathcal{T}_{\mathbf{t}}$ when $\mathcal{T}_{\mathbf{t}}$ is small, to ensure that an optimistic action with low recovery time passes the test. Note that if $\mathcal{T}^-_{\mathbf{t}}$ is too large, we may ask for more roll-outs and thus incur a regret fro each additional roll-out.

We study the recovery test of line 10. First, we want to show that if a_t is a bad action, that is, $\Delta_{T-\mathbf{t}}(s_{\mathbf{t}}, a_{\mathbf{t}}) > \gamma^\star(T - \mathbf{t})$ (see Definition 2), then a_t does not pass the test. Indeed

in that case, it must be that $\mathcal{T}_{\mathbf{t}} > \gamma^\star(T-\mathbf{t})$ as well. Now $\mathcal{T}_{\mathbf{t}}^-$ is computed using the (pessimistic) extended value iteration algorithm. It does not use the horizon $T-\mathbf{t}$ but instead $c(1+\frac{1}{\varepsilon})D_k$, where the value $c=\frac{2}{\gamma}$ is chosen in the algorithm. Thus, for any not too large $\mathbf{t} \leqslant T - c(1+\frac{1}{\varepsilon})D_k$, provided that $M^\star$ is plausible it must be the case that

$$\mathcal{T}_{\mathbf{t}}^- \geqslant \gamma^\star c(1+\frac{1}{\varepsilon})D_k\,,$$

which is bigger than $2D_k$ if $\varepsilon < 1$ and $c \geqslant 1/\gamma^\star$ (We can further discard the case when $\mathbf{t} > T - c(1+\frac{1}{\varepsilon})D_k$ at the price of loosing at most a constant regret $c(1+\frac{1}{\varepsilon})D_k$). In this case, a non recoverable action does not pass the recovery test (on the event that $M^\star$ is plausible). Thus, it holds that

Lemma 13 *Assuming that $\mathcal{T}_{\mathbf{t}}^-$ is computed with horizon $c(1+\frac{1}{\varepsilon})D_k$, for $\mathbf{t} \leqslant T-c(1+\frac{1}{\varepsilon})D_k$, where $\varepsilon<1$ and $c\geqslant 1/\gamma^\star$, then on the event that $M^\star$ is plausible, then an action a_t such that $\Delta_{T-\mathbf{t}}(s_{\mathbf{t}},a_{\mathbf{t}}) > \gamma^\star(T-\mathbf{t})$ does not pass the recovery test, that is $\mathcal{T}_{\mathbf{t}}^- \geqslant 2D_k$.*

Now, two cases can happen when a_t is a good action $\Delta_{T-\mathbf{t}}(s_{\mathbf{t}},a_{\mathbf{t}}) \leqslant d^\star$: The recovery time $\mathcal{T}_{\mathbf{t}}$ is either small or big. It is not a problem if the return time is big but a_t still passes the test. We want to prevent that when the return time $\mathcal{T}_{\mathbf{t}}$ is small and a_t is not a bad action, then a_t does not pass the test. To that end, we show a bound on $\mathcal{T}_{\mathbf{t}}^-$ when $\mathcal{T}_{\mathbf{t}} \leqslant D_k$:

Lemma 14 *Under the event that $M^\star$ is plausible, if $\mathcal{T}_{\mathbf{t}} \leqslant D_k$, then it holds*

$$\mathcal{T}_{\mathbf{t}}^- \leqslant D_k\left[1+2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right)D_k\right]\,, \tag{12}$$

where

$$C_{k,\mathbf{t}}^p = \max_{s\in\mathcal{S}(s_{\mathbf{t}},D_k),a\in\mathcal{A}} C_k^p(s,a)\,,$$

and $\mathcal{S}(s,\ell)$ denotes the set of all states that are reachable from s in ℓ steps.

This result implies that if $2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right)D_k < 1$ then $\mathcal{T}_{\mathbf{t}}^- < 2D_k$ holds (on the event that $M^\star$ is plausible), that is the recovery test of line 10 is passed in this case. In Section 5.5, we bound the number of episodes that ensures that $2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right)D_k \leqslant \alpha$, for some $\alpha<1$.

Proof Let $p_k^{\circlearrowleft -}$ be the kernel corresponding to the worst plausible backup MDP, that is one associated to $Q_k^{\circlearrowleft -}$. Then let $\pi_k^{\circlearrowleft \star}$ be the optimal backup policy in the true MDP (that is with transition kernel p), and $\pi_k^{\circlearrowleft -}$ the optimal backup policy in the worst plausible backup MDP. By construction of the backup MDPs, it holds that $Q_k^{\circlearrowleft\star}(s,a)=0$ if $s=s_{t_k}$ and otherwise

$$Q_k^{\circlearrowleft\star}(s,a) = -1 + \sum_{s'\in\mathcal{S}} p(s'|s,a) Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s'))\,. \tag{13}$$

On the other hand, since $Q_k^{\circlearrowleft -}$ is computed by the algorithm with look-ahead horizon $c(1+\frac{1}{\varepsilon})D_k$, we get that $Q_k^{\circlearrowleft -}(s,a)=0$ if $s=s_{t_k}$ and otherwise

$$Q_k^{\circlearrowleft -}(s,a) \geqslant \max\Big\{-1+\sum_{s'\in\mathcal{S}} p_k^{\circlearrowleft -}(s'|s,a) Q_k^{\circlearrowleft -}(s',\pi_k^{\circlearrowleft -}(s')), -c\Big(1+\frac{1}{\varepsilon}\Big)D_k\Big\}\,.$$

Now let us compare these two values for the chosen action $a_{\mathbf{t}}$. It holds that

$$\begin{aligned} Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}}) - Q_k^{\circlearrowleft-}(s_{\mathbf{t}},a_{\mathbf{t}}) &= \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big(Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft-}(s'))\Big) \\ &\leqslant \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big(Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s'))\Big) \\ &\quad + \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big(Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft-}(s'))\Big) \\ &\quad + \sum_{s'\in\mathcal{S}} \Big(p(s'|s_{\mathbf{t}},a_{\mathbf{t}}) - p_k^{\circlearrowleft-}(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big)Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft-}(s'))\,. \end{aligned} \tag{14}$$

The second sum is negative since $Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s')) \leqslant Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft-}(s'))$ by construction of the pessimistic $\pi_k^{\circlearrowleft-}$, and the third sum is less than $\big\|p(\cdot|s_{\mathbf{t}},a_{\mathbf{t}}) - p_k^{\circlearrowleft-}(\cdot|s_{\mathbf{t}},a_{\mathbf{t}})\big\|_1 c(1+\frac{1}{\varepsilon})D_k$ since $|Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft-}(s'))| \leqslant c(1+\frac{1}{\varepsilon})D_k$ by construction. Thus, we make appear the confidence bound C_k^p from Lemma 5. More precisely, we consider the set $\mathcal{S}(s,\ell)$ of all states that are reachable from s in ℓ steps and introduce the quantity

$$C_{k,\mathbf{t}}^p = \max_{s\in\mathcal{S}(s_{\mathbf{t}},D_k),a\in\mathcal{A}} C_k^p(s,a)\,.$$

Indeed, we are interested specifically in states that are reachable in less than $\mathcal{T}_{\mathbf{t}} \leqslant D_k$ steps from $s_{\mathbf{t}}$ (when following $\pi_k^{\circlearrowleft\star}$). Using this notation, and since the pessimistic backup MDP and $M^\star$ are plausible by assumption, it holds that

$$\begin{aligned} &Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}}) - Q_k^{\circlearrowleft-}(s_{\mathbf{t}},a_{\mathbf{t}}) \leqslant \\ &\qquad \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big(Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s'))\Big) + 2C_k^p(s_{\mathbf{t}},a_{\mathbf{t}})c\Big(1+\frac{1}{\varepsilon}\Big)D_k \\ &\leqslant \quad 2C_{k,\mathbf{t}}^p c\Big(1+\frac{1}{\varepsilon}\Big)D_k + \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}},a_{\mathbf{t}})\Big(Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s'))\Big)\,. \end{aligned}$$

We now apply the same decomposition as equation (14) to the term $Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s'))$, that is, following the path of $\pi_k^{\circlearrowleft\star}$ starting from $s_{\mathbf{t}}$. Propagating this decomposition $j\in\mathbb{N}$ times we get

$$\begin{aligned} Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}}) - Q_k^{\circlearrowleft-}(s_{\mathbf{t}},a_{\mathbf{t}}) \leqslant \mathbb{E}\Bigg[&2C_{k,\mathbf{t}}^p c\Big(1+\frac{1}{\varepsilon}\Big)D_k j \\ &+ \sum_{s'\in\mathcal{S}} p(s'|s_{\mathbf{t}}^j,a_{\mathbf{t}}^j)\Big(Q_k^{\circlearrowleft\star}(s',\pi_k^{\circlearrowleft\star}(s')) - Q_k^{\circlearrowleft-}(s',\pi_k^{\circlearrowleft\star}(s'))\Big)\Bigg]\,, \end{aligned} \tag{15}$$

where the expectation is over the trajectory $\{s_{\mathbf{t}}^j,a_{\mathbf{t}}^j\}_j$, where $s_{\mathbf{t}}^j,a_{\mathbf{t}}^j$ denotes the state-action pair reached after j steps when following $\pi_k^{\circlearrowleft\star}$ from $s_{\mathbf{t}}$, and when we consider s_{t_k} to be absorbing, that is $s_{\mathbf{t}}^j$ is set to s_{t_k} for all $j \geqslant j_0$ where j_0 is the first (random) time when $s_{\mathbf{t}}^j = s_{t_k}$.

Now, since it holds that $Q_k^{\circlearrowleft\star}(s_{t_k},a_{t_k}) - Q_k^{\circlearrowleft -}(s_{t_k},a_{t_k}) = 0$ by construction, the term on the second line in (15) vanishes when $\pi_k^{\circlearrowleft\star}$ reaches s_{t_k}, which in the worst case happens after no more than $\mathcal{T}_{\mathbf{t}}$ steps on average over the trajectory $\{s_{\mathbf{t}}^j, a_{\mathbf{t}}^j\}_j$. Thus, we deduce that

$$\begin{aligned} Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}}) - Q_k^{\circlearrowleft -}(s_{\mathbf{t}},a_{\mathbf{t}}) &\leqslant 2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right) D_k \mathbb{E}[j_0] \\ &\leqslant 2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right) D_k \mathcal{T}_{\mathbf{t}}\,. \end{aligned}$$

Now, using the assumption that $\mathcal{T}_{\mathbf{t}} \leqslant D_k$again, we obtain

$$\begin{aligned} \mathcal{T}_{\mathbf{t}}^- &= \Big(- Q_k^{\circlearrowleft -}(s_{\mathbf{t}},a_{\mathbf{t}}) + Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}})\Big) - Q_k^{\circlearrowleft\star}(s_{\mathbf{t}},a_{\mathbf{t}}) \\ &\leqslant 2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right) D_k \mathcal{T}_{\mathbf{t}} + \mathcal{T}_{\mathbf{t}} \\ &\leqslant D_k\left[1 + 2C_{k,\mathbf{t}}^p c\left(1+\frac{1}{\varepsilon}\right) D_k\right], \end{aligned}$$

which concludes the proof of Lemma 14. ■

5.1.5 Number of episodes

To conclude this section of preliminary results, let us get a bound on on the number of episodes. Since the episodes stop only because of a doubling event, that is precisely as for **UCRL** we trivially have from Jaksch et al. (2010) the following standard control of the number of episodes

Lemma 15 *The total number of episodes is bounded as in Jaksch et al. (2010) by*

$$K_T \leqslant SA\log_2\left(\tfrac{8T}{SA}\right). \tag{16}$$

5.2 Dealing with non-irreversible actions

The control of the cumulated reward of **A-ROGUE** proceeds backward from the last to the first episodes. We first start in this section by handling what happens during the lasts episodes $k > \overline{\mathbf{k}}$ following the last entrance to a region $\mathfrak{S}_i^\star$. Then, we progressively unfold the control back to the first episode in the next sections. Since we want to apply Lemma 13, we accept to control the regret only up to $\tilde{T} = T - c(1+\frac{1}{\varepsilon})D_{K_T}$ at the price of loosing only a regret of $c(1+\frac{1}{\varepsilon})D_{K_T}$. For notational convenience however, we slightly abuse of notation and state all sequel results and proofs until section 5.6 with horizon T while they hold for $\tilde{T}$ instead.

We first show the following result, discarding what happens due to irreversible actions.

Lemma 16 *The regret between the strategy used by* **A-ROGUE** *from time $t_{\overline{\mathbf{k}}+1}$ (where it is in state $s_{t_{\overline{\mathbf{k}}+1}}$), and an optimal strategy starting at the same time in the same state, is*

$$\begin{aligned}
&\mathbb{E}\Big[R^{\pi^\star}_{M^\star}(\bigcup_{k>\overline{\mathbf{k}}}\mathbb{T}_{(k)}|s_{t_{\overline{\mathbf{k}}+1}}) - \sum_{k=\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] \\
&\quad\leqslant\ \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}\tau_+(\tilde{s}_{t_{k+1}},s^+_{t_{k+1}}) + \tau^{\pi^{\circlearrowleft}}(s_{t_{k+1}},s_{t_k}) + \tau^{\pi^{\curvearrowright\star}}(s_{t_k},\tilde{s}^\star_{t_{k+1}})\Big] \\
&\qquad + \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T}|\mathbb{T}^R_{(k)}|\Big] - \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}(23)_k + (30)_k\Big] - (17)\,,
\end{aligned}$$

where all quantities are introduced in Sections 5.2.1-5.2.4.

The proof is in several steps that are detailed below (Section 5.2.1) In all these steps, we assume that $K_T > \overline{\mathbf{k}}$ (otherwise the sums are trivial).

5.2.1 First step decomposition

Let us start with what happens for one episode $k > \overline{\mathbf{k}}$. We have

$$\begin{aligned}
\mathbb{E}\Big[R^{\mathbb{A}}_{M^\star,k}\Big] \ &= \ \mathbb{E}\Big[R^{\pi_k}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi_k}_{M^+_k}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] &(17)\\
&\quad + \mathbb{E}\Big[R^{\pi_k}_{M^+_k}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi_k}_{M^+_k}(\mathbb{T}_{(k)}|s_{t_k})\Big] &(18)\\
&\quad + \mathbb{E}\Big[R^{\pi_k}_{M^+_k}(\mathbb{T}_{(k)}|s_{t_k})\Big]\,. &(19)
\end{aligned}$$

In this first step, we apply this decomposition specially to $k = K_T$, for which $\mathbb{T}_{(K_T)} = [t_{K_T} : T]$. We handle (19) by noting that since π_{K_T} is the optimistic policy computed in episode K_T, in satisfies that

$$(19) \geqslant \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T}:T]|s_{t_{K_T}})\Big]\,.$$

Since the rewards are bounded in $[0,1]$, it holds that

$$(18) \geqslant -\mathbb{E}\Big[|\mathbb{T}_{(K_T)} \setminus \mathbb{T}^D_{(K_T)}|\Big] = -\mathbb{E}\Big[|\mathbb{T}^R_{(K_T)}|\Big]\,.$$

Thus, so far, combining these three equations together, we have shown that

$$\mathbb{E}\Big[R^{\mathbb{A}}_{M^\star,K_T}\Big] \geqslant \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T}:T]|s_{t_{K_T}})\Big] + (17) - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T)}|\Big]\,. \tag{20}$$

5.2.2 Preparing the unfolding step

Equation (20) relates the reward accumulated by the algorithm to that of a trajectory of an optimal policy starting from $s_{t_{K_T}}$, until the final time T. We now use this in order to progressively compare to the optimal policy starting from $s_{t_{K_T-1}}$ until time T, then from $s_{t_{K_T-2}}$ and so on, proceeding backward in the episodes, until comparing to the optimal

policy starting from $s_{t_{\overline{\mathbf{k}}+1}}$ up to time T. An important property that we use in the following is that for all episodes $k > \overline{\mathbf{k}}$, by definition, the algorithm stays in the same region $\mathfrak{S}_i^\star$, from $t_{\overline{\mathbf{k}}+1}$ until the final time T.

Considering that $K_T - 1 > \overline{\mathbf{k}}$ for the sake of showing how works the decomposition in the general case (some terms simplify if $K_T \leqslant \overline{\mathbf{k}} + 1$), we proceed as follows

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k>\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] &= \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-2} R^{\mathbb{A}}_{M^\star,k} + R^{\mathbb{A}}_{M^\star,K_T-1} + R^{\mathbb{A}}_{M^\star,K_T}\Big] && (21)\\
&\geqslant \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-2} R^{\mathbb{A}}_{M^\star,k}\Big] + (17) - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T)}|\Big]\\
&\quad + \mathbb{E}\Big[R^{\pi_{K_T-1}}_{M^\star}(\mathbb{T}^D_{(K_T-1)}|s_{t_{K_T-1}}) + R^{\pi^\star}_{M^\star}([t_{K_T}:T]|s_{t_{K_T}})\Big], && (22)
\end{aligned}$$

where for the first inequality, we applied (20) to control $R^{\mathbb{A}}_{M^\star,K_T-1}$ as well as the definition of $R^{\mathbb{A}}_{M^\star,K_T-1}$. Now, in order to control (22), we use the following decomposition

$$\begin{aligned}
(22) &\geqslant \mathbb{E}\Big[R^{\pi_{K_T-1}}_{M^\star}(\mathbb{T}^D_{(K_T-1)}|s_{t_{K_T-1}}) - R^{\pi_{K_T-1}}_{M^+_{K_T-1}}(\mathbb{T}^D_{(K_T-1)}|s_{t_{K_T-1}})\Big] && (23)\\
&\quad + \mathbb{E}\Big[R^{\pi_{K_T-1}}_{M^+_{K_T-1}}(\mathbb{T}^D_{(K_T-1)}|s_{t_{K_T-1}}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(K_T-1)}|s_{t_{K_T-1}})\Big] && (24)\\
&\quad + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}(\mathbb{T}_{(K_T-1)}|s_{t_{K_T-1}})\Big] - \mathbb{E}\Big[\mathbb{T}^R_{(K_T-1)}\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T}:T]|s_{t_{K_T}})\Big], && (25)
\end{aligned}$$

where we used ni the last line the fact that rewards are in $[0,1]$. Then, we use the fact that $\pi^\star$ is optimal in order to control the last term of (25)

$$\begin{aligned}
\mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T}:T]|s_{t_{K_T}})\Big] &\geqslant \mathbb{E}\Big[R^{\pi^{\circlearrowleft}}_{M^\star}([t_{K_T}:t_{K_T}+\tilde\tau-1]|s_{t_{K_T}})\\
&\quad + R^{\pi^{\curvearrowright\star}}_{M^\star}([t_{K_T}+\tilde\tau : t_{K_T}+\tilde\tau+\tau^\star-1]|s_{t_{K_T-1}})\\
&\quad + R^{\pi^\star_{K_T-1}}_{M^\star}([t_{K_T}+\tilde\tau+\tau^\star:T]|\tilde s^\star_{t_{K_T}})\Big]. && (26)
\end{aligned}$$

Here we first introduced the policy $\pi^{\circlearrowleft}$ that goes from $s_{t_{K_T}}$ back to the state $s_{t_{K_T-1}}$ visited at time t_{K_T-1} as fast as possible and the random time $\tilde\tau = \tau^{\pi^{\circlearrowleft}}(s_{t_{K_T}}, s_{t_{K_T-1}}) \in \mathbb{N}$ to reach it using this policy. Let us remind that for states $s, s' \in \mathcal{S}$ and some policy π, $\tau^\pi(s,s')$ denotes the random time corresponding to the first visit to state s' when following policy π from state s (in the current MDP). Then we introduced the state $\tilde s^\star_{t_{K_T}}$ that would have been reached at time t_{K_T} by the optimal policy $\pi^\star$ started at state $s_{t_{K_T-1}}$ and time t_{K_T-1}. We also introduced the policy $\pi^{\curvearrowright\star}$ that goes from $s_{t_{K_T-1}}$ to $\tilde s^\star_{t_{K_T}}$ visited at time t_{K_T-1} as fast as possible, together with its random time $\tau^\star = \tau^{\pi^{\circlearrowleft}\star}(s_{t_{K_T-1}}, \tilde s^\star_{t_{K_T}})$. In the last term of equation (26), we denoted $\pi^\star_{K_T-1}$ to insist on the fact that we use the optimal policy $\pi^\star$ started from state $s_{t_{K_T-1}}$ at time t_{K_T-1}, that we follow from state $\tilde s^\star_{t_{K_T}}$ (at the time it is reached) up to time T. Indeed, this is a priori different from the optimal policy that starts from $\tilde s^\star_{t_{K_T}}$ at time $t_{K_T} + \tilde\tau + \tau^\star$.

Regarding (26), since the rewards are all positive, on the one hand we have

$$R^{\pi^{\circlearrowleft}}_{M^\star}([t_{K_T} : t_{K_T} + \tilde{\tau} - 1]|s_{t_{K_T}}) + R^{\pi^{\curvearrowright\star}}_{M^\star}([t_{K_T} + \tilde{\tau} : t_{K_T} + \tilde{\tau} + \tau^\star - 1]|s_{t_{K_T-1}}) \geqslant 0\,,$$

and on the other hand, it holds that

$$\mathbb{E}\Big[R^{\pi^\star_{K_T-1}}_{M^\star}([t_{K_T} + \tilde{\tau} + \tau^\star : T]|\tilde{s}^\star_{t_{K_T}})\Big] \;\geqslant\; \mathbb{E}\Big[R^{\pi^\star_{K_T-1}}_{M^\star}([t_{K_T} : T]|\tilde{s}^\star_{t_{K_T}})\Big] - \mathbb{E}\Big[\tilde{\tau} + \tau^\star\Big]\,.$$

Thus, we deduce that (22) is bounded by

$$\begin{aligned}(22) \;&\geqslant\; (23) + (24) - \mathbb{E}\Big[\tilde{\tau} + \tau^\star\Big] \\ &\quad + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T-1} : t_{K_T} - 1]|s_{t_{K_T-1}})\Big] + \mathbb{E}\Big[R^{\pi^\star_{K_T-1}}_{M^\star}([t_{K_T} : T]|\tilde{s}^\star_{t_{K_T}})\Big] \\ &=\; (23) + (24) - \mathbb{E}\Big[\tilde{\tau} + \tau^\star\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T-1} : T]|s_{t_{K_T-1}})\Big]\,,\end{aligned}$$

where we used the definition of $\pi^\star_{K_T-1}$ and $\tilde{s}^\star_{t_{K_T}}$ in the second line.

Plugging in this back to the first step of this section, we deduce that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k>\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] \;&\geqslant\; \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-2} R^{\mathbb{A}}_{M^\star,k}\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T-1} : T]|s_{t_{K_T-1}})\Big] \\ &\quad + (17) - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T-1)}|\Big] - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T)}|\Big] \\ &\quad + (23)_{K_T-1} + (24)_{K_T-1} - \mathbb{E}\Big[\tau^{\pi^{\circlearrowleft}}(s_{t_{K_T}}, s_{t_{K_T-1}}) + \tau^{\pi^{\curvearrowright\star}}(s_{t_{K_T-1}}, \tilde{s}^\star_{t_{K_T}})\Big]\,,\end{aligned} \tag{27}$$

where we put the index of the episode in subscript of the equations for clarity purpose. The next step is now to handle the term $(24)_{K_T-1}$, and more generally $(24)_k$ for $k > \overline{\mathbf{k}}$.

5.2.3 Optimistic model

We now control, for $k > \overline{\mathbf{k}}$, the quantity

$$(24)_k = \mathbb{E}\Big[R^{\pi_k}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\Big]\,.$$

For that purpose, we use that π_k is the optimal policy for model M_k^+ with horizon T, and reason trajectory wise (that is, following a random trajectory from the policy, as opposed for instance to a distribution of states). Thus, it holds that

$$\begin{aligned}&\mathbb{E}\Big[R^{\pi_k}_{M_k^+}([t_k : T]|s_{t_k})\Big] \geqslant \max_\pi \mathbb{E}\Big[R^{\pi}_{M_k^+}([t_k : T]|s_{t_k})\Big] \\ &\geqslant\; \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) \\ &\quad + R^{\tilde{\pi}}_{M_k^+}([\bar{t}^D_k : \bar{t}^D_k + \tau_+(\tilde{s}_{\bar{t}^D_k}, s^+_{\bar{t}^D_k}) - 1]|\tilde{s}_{\bar{t}^D_k}) \\ &\quad + R^{\pi_k}_{M_k^+}([\bar{t}^D_k + \tau_+(\tilde{s}_{\bar{t}^D_k}, s^+) : T]|s^+_{\bar{t}^D_k})\Big] \\ &\geqslant\; \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] + \mathbb{E}\Big[R^{\pi_k}_{M_k^+}([\bar{t}^D_k : T]|s^+_{\bar{t}^D_k})\Big] - \mathbb{E}\Big[\tau_+(\tilde{s}_{\bar{t}^D_k}, s^+_{\bar{t}^D_k})\Big]\,,\end{aligned} \tag{28}$$

where we introduced $\tilde{s}_{\overline{t}_k^D}$ the random state reached by $\pi^\star_{t_k}$ after $|\mathbb{T}^D_{(k)}|$ steps when (virtually) acting in M_k^+ from s_{t_k}, and $s^+_{\overline{t}_k^D}$ the state reached by π_k after $|\mathbb{T}^D_{(k)}|$ steps when (virtually) acting in M_k^+ from s_{t_k}. In the second inequality, $\tilde{\pi}$ is the policy that minimizes in expectation, the time $\tau_+(\tilde{s}_{\overline{t}_k^D}, s^+_{\overline{t}_k^D})$ to go from $\tilde{s}_{\overline{t}_k^D}$ to $s^+_{\overline{t}_k^D}$ in the augmented MDP M_k^+.

Thus, since on the other hand, we have the decomposition

$$\mathbb{E}\Big[R^{\pi_k}_{M_k^+}([t_k:T]|s_{t_k})\Big] = \mathbb{E}\Big[R^{\pi_k}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] + \mathbb{E}\Big[R^{\pi_k}_{M_k^+}([\overline{t}_k^D:T]|s^+_{\overline{t}_k^D})\Big]\,,$$

we deduce from (28) the following bound

$$(24)_k \geqslant -\mathbb{E}\Big[\tau_+(\tilde{s}_{t_{k+1}}, s^+_{t_{k+1}})\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\Big]\,. \tag{29}$$

For convenience, let us number the second term that appears in the right hand side of this inequality;

$$\mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] \tag{30}$$

Plugging (29) back into (27), so far we have proved that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k>\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] \;\geqslant\; & \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-2} R^{\mathbb{A}}_{M^\star,k}\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T-1}:T]|s_{t_{K_T-1}})\Big] + (17) \\ & + (23)_{K_T-1} + (30)_{K_T-1} - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T-1)}|\Big] - \mathbb{E}\Big[|\mathbb{T}^R_{(K_T)}|\Big] \\ & - \mathbb{E}\Big[\tau_+(\tilde{s}_{t_{K_T}}, s^+_{t_{K_T}}) + \tau^{\pi^{\circlearrowleft}}(s_{t_{K_T}}, s_{t_{K_T-1}}) + \tau^{\pi^{\curvearrowright\star}}(s_{t_{K_T-1}}, \tilde{s}^\star_{t_{K_T}})\Big]\,,\end{aligned} \tag{31}$$

Now, it is not difficult to see that (17), $(23)_{K_T-1}$ and $(30)_{K_T-1}$ are similar. Indeed, we have

$$\begin{aligned}(23)_k &= \mathbb{E}\Big[R^{\pi_k}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi_k}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k})\Big]\,, \\ (30)_k &= \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\Big]\,, \\ (17) &= \mathbb{E}\Big[R^{\pi_{K_T}}_{M^\star}(\mathbb{T}^D_{(K_T)}|s_{t_{K_T}}) - R^{\pi_{K_T}}_{M^+_{K_T}}(\mathbb{T}^D_{(K_T)}|s_{t_{K_T}})\Big] = (23)_{K_T}\,.\end{aligned}$$

5.2.4 Unfolding step within one region

Now, one can recognize in the first two terms of equation (31), the exact same form as what happens in equation (22). Indeed, it holds (provided that $K_T - 2 \geqslant \overline{\mathbf{k}} + 1$) that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-2} R^{\mathbb{A}}_{M^\star,k}\Big] + \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{K_T-1}:T]|s_{t_{K_T-1}})\Big] &= \mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-3} R^{\mathbb{A}}_{M^\star,k}\Big] \\ + \mathbb{E}\Big[R^{\pi_{K_T-2}}_{M^\star}(\mathbb{T}_{(K_T-2)}|s_{t_{K_T-2}}) + R^{\pi^\star}_{M^\star}([t_{K_T-1}:T]|s_{t_{K_T-1}})\Big]\,.&\end{aligned}$$

and we recognize in the last line equation (22) for episode K_T-1 instead of K_T. Thus, we use the same decomposition (22)-(30) using $k=K_T-1$, K_T-2, etc. up to $k=\overline{\mathbf{k}}+1$, and we deduce that

$$\begin{aligned}\mathbb{E}\Big[\sum_{k>\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] \;\geqslant\; & \mathbb{E}\Big[R^{\pi^\star}_{M^\star}([t_{\overline{\mathbf{k}}+1}:T]|s_{t_{\overline{\mathbf{k}}+1}})\Big]+\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T}|\mathbb{T}^R_{(k)}|\Big]\\ &+\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}(23)_k+(30)_k\Big]+(23)_{K_T}\\ &-\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}\tau_+(\tilde{s}_{t_{k+1}},s^+_{t_{k+1}})+\tau^{\pi^{\circlearrowleft}}(s_{t_{k+1}},s_{t_k})+\tau^{\pi^{\curvearrowright\star}}(s_{t_k},\tilde{s}^\star_{t_{k+1}})\Big],\end{aligned}$$

That is, reorganizing the terms, we deduce that

$$\begin{aligned}&\mathbb{E}\Big[R^{\pi^\star}_{M^\star}(\bigcup_{k>\overline{\mathbf{k}}}\mathbb{T}_{(k)}|s_{t_{\overline{\mathbf{k}}+1}})-\sum_{k=\overline{\mathbf{k}}}R^{\mathbb{A}}_{M^\star,k}\Big]\\ &\leqslant\;\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}\tau_+(\tilde{s}_{t_{k+1}},s^+_{t_{k+1}})+\tau^{\pi^{\circlearrowleft}}(s_{t_{k+1}},s_{t_k})+\tau^{\pi^{\curvearrowright\star}}(s_{t_k},\tilde{s}^\star_{t_{k+1}})\Big]\\ &\quad+\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T}|\mathbb{T}^R_{(k)}|\Big]-\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1}(23)_k+(30)_k\Big]-(23)_{K_T}\,.\;\square\end{aligned}$$

5.3 Continuing the unfolding step

A-ROGUE stays in the same region during all the episodes $k>\overline{\mathbf{k}}$. Now, the episodes can be decomposed into two categories. Those during which the **A-ROGUE** stays in the same region, and those during which **A-ROGUE** plays an action that leads to another region (than the one at the starting time of that episode). We use a generic index k for the first category, and $\mathbf{k}$ for the second category. It is natural to first look at episode $\mathbf{k}=\overline{\mathbf{k}}$ where the last change of region occurs, before continuing backwards up to the first episode.

When entering a different region, the return time suddenly becomes large, and thus the term $\tau^{\pi^{\circlearrowleft}}(s_{t_{\mathbf{k}+1}},s_{t_{\mathbf{k}}})$ may not be controlled anymore. To handle this difficulty, we use a slightly different decomposition than the one used for the proof of Lemma 16. We introduce $\mathbf{t_k}\in[t_{\mathbf{k}},t_{\mathbf{k}+1})$ that corresponds to the time when **A-ROGUE** plays an irreversible action in episode $\mathbf{k}$. Further let us split $\mathbb{T}_{(\mathbf{k})}=\mathbb{T}_{(\mathbf{k}),1}\cup\mathbb{T}_{(\mathbf{k}),2}$ in two parts first corresponding to time-steps between $t_{\mathbf{k}}$ and $\mathbf{t_k}+1$ (excluded) and second to time-steps between $\mathbf{t_k}+1$ (included) and $t_{\mathbf{k}+1}$. Likewise, we introduce $\mathbb{T}^D_{(\mathbf{k}),1}$, $\mathbb{T}^D_{(\mathbf{k}),2}$, $\mathbb{T}^R_{(\mathbf{k}),1}$ and $\mathbb{T}^R_{(\mathbf{k}),2}$. Decomposing the regret before and after the time $\mathbf{t_k}$ is intuitively equivalent to introducing an additional episode starting at $\mathbf{t_k}$ (with the difference that $\mathbf{t_k}$ is unknown to the algorithm).

We start with $\mathbf{k} = \overline{\mathbf{k}}$ and handle the regret in $[\mathbf{t_k}, T]$ by applying the steps of Lemma 16 but replacing t_k with $\mathbf{t_k}$. That is, it holds

$$\begin{aligned}
&\mathbb{E}\Big[R^{\pi^\star}_{M^\star}(\mathbb{T}_{(\mathbf{k}),2} \cup \bigcup_{k>\overline{\mathbf{k}}} \mathbb{T}_{(k)} | s_{\mathbf{t_k}+1}) - R^{\pi_{\mathbf{k}}}_{M^\star}(\mathbb{T}_{(\mathbf{k}),2} | s_{\mathbf{t_k}+1}) - \sum_{k=\overline{\mathbf{k}}} R^{\mathbb{A}}_{M^\star,k}\Big] \\
\leqslant\ & \mathbb{E}\Big[\tau_+(\tilde{s}_{t_{\mathbf{k}+1}}, s^+_{t_{\mathbf{k}+1}}) + \tau^{\pi^{\circlearrowleft}}(s_{t_{\mathbf{k}+1}}, s_{\mathbf{t_k}+1}) + \tau^{\pi^{\curvearrowright\star}}(s_{\mathbf{t_k}+1}, \tilde{s}^\star_{t_{\mathbf{k}+1}})\Big] \\
&+\mathbb{E}\Big[\sum_{k=\overline{\mathbf{k}}+1}^{K_T-1} \tau_+(\tilde{s}_{t_{k+1}}, s^+_{t_{k+1}}) + \tau^{\pi^{\circlearrowleft}}(s_{t_{k+1}}, s_{t_k}) + \tau^{\pi^{\curvearrowright\star}}(s_{t_k}, \tilde{s}^\star_{t_{k+1}})\Big] \\
&+ \mathbb{E}\Big[|\mathbb{T}^R_{(\mathbf{k}),2}| + \sum_{k=\overline{\mathbf{k}}+1}^{K_T} |\mathbb{T}^R_{(k)}|\Big] - (23)_{K_T} \\
&-\mathbb{E}\Big[(23)_{\mathbf{k},2} + (30)_{\mathbf{k},2} + \sum_{k=\overline{\mathbf{k}}+1}^{K_T-1} (23)_k + (30)_k\Big],
\end{aligned}$$

where $(23)_{\mathbf{k},2}$ is similar to $(23)_k$ except is uses $\mathbb{T}^D_{(\mathbf{k}),2}$ instead of $\mathbb{T}^D_{(k)}$, and likewise $(30)_{\mathbf{k},2}$ is similar to $(30)_k$ in the same way. That is:

$$\begin{aligned}
(23)_{\mathbf{k},2} &= \mathbb{E}\Big[R^{\pi_{\mathbf{k}}}_{M^\star}(\mathbb{T}^D_{(\mathbf{k}),2}) | s_{\mathbf{t_k}+1}) - R^{\pi_{\mathbf{k}}}_{M^+_{\mathbf{k}}}(\mathbb{T}^D_{(\mathbf{k}),2}) | s_{\mathbf{t_k}+1})\Big] \\
(30)_{\mathbf{k},2} &= \mathbb{E}\Big[R^{\pi^\star}_{M^+_{\mathbf{k}}}(\mathbb{T}^D_{(\mathbf{k}),2}) | s_{\mathbf{t_k}+1}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(\mathbf{k}),2}) | s_{\mathbf{t_k}+1})\Big].
\end{aligned}$$

We focus on time $\mathbf{t_k}$ when **A-ROGUE** plays an irreversible action. Note that the action played must pass either the test of line 10 or of line 13. Under the event Ω, by Lemma 14, if the action passes the test of line 10, then it is not irreversible, thus the action must pass the test of line 13, and is thus $d^\star$-optimal by Lemma 12, that is , it holds

$$\mathbb{E}\Big[R^{a_{\mathbf{t}}}_{M^\star}([\mathbf{t} : \mathbf{t}+1] | s_{\mathbf{t}}) + R^{\pi^\star_{\mathbf{t}+1}}_{M^\star}([\mathbf{t}+1 : T] | s_{\mathbf{t}+1})\Big] \geqslant \mathbb{E}\Big[R^{\pi^\star_{\mathbf{t}}}_{M^\star}([\mathbf{t} : T] | s_{\mathbf{t}})\Big] - d^\star\,. \tag{32}$$

Now, assuming this was the only irreversible action played during the episode $\mathbf{k}$, we can handle the regret on previous steps of the episode $\mathbb{T}_{(\mathbf{k}),1} \setminus \{\mathbf{t_k}\}$ using a similar decomposition as that for Lemma 16. More generally, if m_k denotes the number of irreversible actions played **A-ROGUE** in episode k, at times $\{\mathbf{t}_{k,m}\}_{m\in[m_k]}$, then by equation (32), the regret during episode an k looses at most $m_k d^\star$ compared to an episode $k > \overline{\mathbf{k}}$ (for which $m_k = 0$). More formally, using for convenience the convention $\mathbf{t}_{k,m_k+1} \overset{\text{def}}{=} t_{k+1}$ and $\mathbf{t}_{k,0} \overset{\text{def}}{=} t_k - 1$, then we have proved that

Lemma 17 (Main Decomposition) *The regret of* **A-ROGUE** *is bounded as*

$$\begin{aligned}
&\mathbb{E}\Big[\mathfrak{R}_T\Big] = \mathbb{E}\Big[R^{\pi^\star}_{M^\star}(\bigcup_{k=1}^{K_T} \mathbb{T}_{(k)} | s_{t_k}) - \sum_{k=1}^{K_T} R^{\mathbb{A}}_{M^\star,k}\Big] \\
\leqslant\ & \mathbb{E}\Big[\sum_{k=1}^{K_T} m_k d^\star\Big] + \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{m=1}^{m_k+1} \tau_+(\tilde{s}_{\mathbf{t}_{k,m}}, s^+_{\mathbf{t}_{k,m}}) + \tau^{\pi^{\circlearrowleft}}(s_{\mathbf{t}_{k,m}}, s_{\mathbf{t}_{k,m-1}+1}) + \tau^{\pi^{\curvearrowright\star}}(s_{\mathbf{t}_{k,m-1}+1}, \tilde{s}^\star_{\mathbf{t}_{k,m}})\Big] \\
&+\mathbb{E}\Big[\sum_{k=1}^{K_T} |\mathbb{T}^R_{(k)}|\Big] - \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{m=1}^{m_k+1} (23)_{k,m} + (30)_{k,m}\Big],
\end{aligned} \tag{33}$$

where, introducing the sets $\mathbb{T}_{(k),m} = [\mathbf{t}_{k,m-1}+1 : \mathbf{t}_{k,m}]$ *(with corresponding* $\mathbb{T}^D_{(k),m}, \mathbb{T}^R_{(k),m}$*)*

$$\begin{aligned}(23)_{k,m} &= \mathbb{E}\Big[R^{\pi_k}_{M^\star}(\mathbb{T}^D_{(k),m})|s_{\mathbf{t}_{k,m-1}+1}) - R^{\pi_k}_{M_k^+}(\mathbb{T}^D_{(k),m})|s_{\mathbf{t}_{k,m-1}+1})\Big]\\ (30)_{k,m} &= \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k),m})|s_{\mathbf{t}_{k,m-1}+1}) - R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k),m})|s_{\mathbf{t}_{k,m-1}+1})\Big]\,.\end{aligned}$$

Note that equation (33) handles the general case, without assuming that $\overline{\mathbf{k}} < K_T$. Also $R^{\pi^\star}_{M^\star}(\bigcup_{k=1}^{K_T}\mathbb{T}_{(k)}|s_{t_k}) = R^{\pi^\star}_{M^\star}(\mathbb{T}|s_1)$ is the reward accumulated by an optimal policy $\pi^\star$ from the initial state s_1 between time 1 and T. Thus this lemma successfully enables us to compare our strategy to the optimal policy, and not simply, for instance, to a local optimal action.

Now, note that controlling the term $\mathbb{E}\Big[\sum_{k=1}^{K_T}|\mathbb{T}^R_{(k)}|\Big]$ is a priori not trivial: Indeed, at time $\mathbf{t}$ when **A-ROGUE** plays an irreversible action might a priori not be close to the time $\mathbf{t}^\star$ when the optimal policy plays an irreversible action, since the algorithm must decide that it is "good" to play an irreversible action and this may require many roll-outs. That is, in case $t_{k'} \leqslant \mathbf{t}^\star \leqslant t_{k'+1} \leqslant t_k \leqslant \mathbf{t}$ for some episode $k' < k$, the optimal policy enters a region earlier than **A-ROGUE**, and **A-ROGUE** may incur a linear regret between $\mathbf{t}^\star$ and $\mathbf{t}$ for not entering an optimal region. We control the number of roll-outs in Section 5.5 and show that **A-ROGUE** does not incur too much regret due to this phenomenon.

Bounding the number of irreversible actions We now bound $m = \sum_{k=1}^{K_T} m_k$. First of all, if all regions except the initial one kS_{i_0} are not escapable with probability 1, then we simply have $\mathbb{E}\Big[md^\star\Big] \leqslant 1$. In the other cases let us note that it must take at time t at least $\max\{\gamma^\star(T-t) - 2d^\star, d^\star\}$ steps on average to get from one region of the partition to a neighboring one. Thus in the worst case, we cannot cross more than m times on average, where m satisfies $\sum_{i=1}^{m}(\gamma^\star d^\star i - 2d^\star) \leqslant T$: Indeed, let t_i be the i-th time when a crossing occurs. Since the last (m-th) time, at least $d^\star$ steps are required to transit between regions, we deduce that $T - t_m \leqslant d^\star$ on average. Likewise, applying this argument backward from m to i, we get $T - t_i \leqslant (m-i+1)d^\star$, and thus $\gamma^\star(T-t_i) - 2d^\star \leqslant \gamma^\star(m-i+1)d^\star - 2d^\star$. We then deduce the (rough) bound $\gamma^\star d^\star m^2 + (\gamma^\star - 4)d^\star m \leqslant 2T$ from which it follows that

$$\begin{aligned}\mathbb{E}\Big[md^\star\Big] &\leqslant \sqrt{\frac{2Td^\star}{\gamma^\star} + \frac{(4-\gamma^\star)^2 d^{\star 2}}{4\gamma^{\star 2}}} + \frac{(4-\gamma^\star)d^\star}{2\gamma^\star}\\ &\leqslant \sqrt{\frac{2Td^\star}{\gamma^\star} + \frac{4d^{\star 2}}{\gamma^{\star 2}}} + \frac{2d^\star}{\gamma^\star}\\ &\leqslant 2\sqrt{\frac{2d^\star}{\gamma^\star}}\sqrt{T + \frac{2d^\star}{\gamma^\star}}\,. \qquad (34)\end{aligned}$$

5.4 Bellman equation and local diameter

In this section, we control the terms (17), $(23)_k$ and $(30)_k$, by resorting to the Bellman propagation equation as well as the definition of the confidence sets used by the **A-ROGUE** algorithm. Then, we make use of the local diameter $d^\star$ to handle the $_\tau$ terms appearing in Lemma 16. We exhibit in particular the role of the assumption $k > \overline{\mathbf{k}}$ and of the diameter of the local region $\mathfrak{S}^\star_i$, as well as of the computation of the return time by the algorithm.

5.4.1 Bellman Propagation equation

For any MDP $M = (r_M, p_M)$, and horizon $h \in \mathbb{N}$ it holds from the Bellman propagation equation that

$$\mathbb{E}[R_M^\pi([t:t+h-1] \cap \mathbb{T}_{(k)}^D | s_t)] = \mathbb{E}\Big[r_M(s_t, \pi(s_t)) + \sum_{s'} p_M(s'|s_t, \pi(s_t)) \mathbb{E}[R_M^\pi([t+1:t+h-1] \cap \mathbb{T}_{(k)}^D | s')] \Big]. \tag{35}$$

Now let us focus on $(23)_k$, for $k > \overline{\mathbf{k}}$. If we denote the episode length to be $h_k = t_{k+1} - t_k$, we note that $[t_k + 1 : t_k + h - 1] \cap \mathbb{T}_{(k)}^D = \mathbb{T}_{(k)}^D \setminus \{t_k\}$, and deduce by applying (35) to the two terms of $(23)_k$ that

$$\begin{aligned}(23)_k = \mathbb{E}\Big[& \mu(s_{t_k}, a_{t_k}) \\ & + \sum_{s'} p(s'|s_{t_k}, a_{t_k}) \mathbb{E}[R_{M^\star}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s')] \\ & \qquad - r_{M_k^+}(s_{t_k}, a_{t_k}) \\ & - \sum_{s' \in \mathcal{S}} p_{M_k^+}(s'|s_{t_k}, a_{t_k}) \mathbb{E}[R_{M_k^+}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s')] \Big],\end{aligned}$$

where $M_k^+ = (r_{M_k^+}, p_{M_k^+})$ is the augmented MDP and $M^\star = (\mu, p)$ the true MDP. Regrouping the terms to make appear the estimation error of the reward function and transition kernel, we obtain

$$\begin{aligned}(23)_k = \mathbb{E}\Big[& \mu(s_{t_k}, a_{t_k}) - r_{M_k^+}(s_{t_k}, a_{t_k}) \\ & + \sum_{s' \in \mathcal{S}} \Big(p_{M_k^+}(s'|s_{t_k}, a_{t_k}) - p(s'|s_{t_k}, a_{t_k}) \Big) \mathbb{E}[R_{M_k^+}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s')] \\ & + \sum_{s' \in \mathcal{S}} p(s'|s_{t_k}, a_{t_k}) \mathbb{E}\big[R_{M_k^+}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s') - R_{M^\star}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s') \big] \Big],\end{aligned} \tag{36}$$

where in the last line, the difference can be decomposed iteratively in the same way. Now, note that, in the second line, since $p_{M_k^+}$ and p both sum to 1, one can replace $V_{\mathbb{T}_{(k)}^D \setminus \{t_k\}}^+(s') \stackrel{\text{def}}{=} \mathbb{E}[R_{M_k^+}^{\pi_k}(\mathbb{T}_{(k)}^D \setminus \{t_k\} | s')]$ by $V_{\mathbb{T}_{(k)}^D \setminus \{t_k\}}^+(s') - c$ for any constant c. In particular, if we introduce $\mathbf{sp}(f) = \max_{s \in \mathcal{S}} f(s) - \min_{s \in \mathcal{S}} f(s)$, then for a specific choice of c we get $V_{\mathbb{T}_{(k)}^D \setminus \{t_k\}}^+(s') - c \leqslant \mathbf{sp}(V_{\mathbb{T}_{(k)}^D \setminus \{t_k\}}^+)/2$. Now, a standard way to get a bound on $(23)_k$ is by unfolding the above equality, reorganizing the resulting terms according to each state-action pair and bounding the probability to reach a specific state-action pair in less than $t_{k+1} - t_k$ steps by 1. This leads to the bound

$$\begin{aligned}(23)_k \quad \geqslant \quad -\mathbb{E}\Bigg[& \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} v_k(s,a) |r_{M_k^+}(s,a) - \mu(s,a)| \\ & + \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} v_k(s,a) \|p_{M_k^+}(\cdot|s,a) - p(\cdot|s,a)\|_1 \max_{t \in \mathbb{T}_{(k)}} \frac{\mathbf{sp}(V_{[t+1:t_{k+1}-1] \cap \mathbb{T}_{(k)}^D}^+)}{2} \Bigg],\end{aligned}$$

where we introduced the total number of visits in episode k (including roll-out transitions)

$$v_k(s,a) = \sum_{t \in \mathbb{T}_{(k)}} \mathbb{I}\{s_t = s, a_t = a\}\,. \tag{37}$$

Note that we can proceed in the exact same way in order to derive a bound on $(30)_k$ by replacing π_k with $\pi^\star$, and on (17) by replacing $\mathbb{T}^D_{(k)}$ with $\mathbb{T}^D_{(K_T)}$.

5.4.2 Span to Local span

However, apart from the special case when there is only one region $\mathfrak{S}_i^\star$, there is in general no reason that the term $\max_{t\in\mathbb{T}_{(k)}} \mathbf{sp}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})$ be small. Indeed, when it is not possible to reach a common state from two different states, their value can be arbitrarily different. Thus, we proceed with a different decomposition, that we detail in the next paragraph.

The second line of (36) satisfies the following, for $t = t_k$ and all $c \in \mathbb{R}$:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{s'\in\mathcal{S}}\Big(p_{M_k^+}(s'|s_t,a_t) - p(s'|s_t,a_t)\Big)V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s')\Big] \\
&\quad\geqslant\ \mathbb{E}\Big[\sum_{s'\in\mathcal{S}:p(s'|s_t,a_t)>0}\Big(p_{M_k^+}(s'|s_t,a_t) - p(s'|s_t,a_t)\Big)V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s')\Big] \\
&\quad\geqslant\ -\mathbb{E}\Big[||p_{M_k^+}(\cdot|s_t,a_t) - p(\cdot|s_t,a_t)||_1\Big(\max_{s'\in\mathcal{S}:p(s'|s_t,a_t)>0} V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s') - c\Big)\Big]\,,
\end{aligned}$$

where we used the fact that the value is non negative. Thus, we can restrict our attention to the set $\mathcal{S}_t = \{s' \in \mathcal{S} : p(s'|s_t,a_t) > 0\} \subset \mathcal{S}$. By choosing $c = \frac{1}{2}\Big(\max_{s'\in\mathcal{S}_t} V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s') + \min_{s'\in\mathcal{S}_t} V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s')\Big)$, and using the notation $\mathbf{sp}_S(f) = \max_S f - \min_S f$, we get

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{s'\in\mathcal{S}}\Big(p_{M_k^+}(s'|s_t,a_t) - p(s'|s_t,a_t)\Big)V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}}(s')\Big] \\
&\quad\geqslant\ -\mathbb{E}\Big[||p_{M_k^+}(\cdot|s_t,a_t) - p(\cdot|s_t,a_t)||_1 \frac{1}{2}\mathbf{sp}_{\mathcal{S}_t}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})\Big]\,.
\end{aligned}$$

Thus, the main difference compared to the standard derivation is that we get $\mathbf{sp}_{\mathcal{S}_t}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})$ instead of $\mathbf{sp}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})$. Proceeding now with the usual next steps that consist in rewriting the sums to make appear the number of visits to each state action pair, we thus get the following bounds:

Lemma 18 *For all $k > \overline{\mathbf{k}}$, it holds that*

$$\begin{aligned}
(23)_k &= \mathbb{E}\Big[R^{\pi_k}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k}) - R^{\pi_k}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] \\
&\geqslant\ -\mathbb{E}\Bigg[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_k(s,a)|r_{M_k^+}(s,a) - \mu(s,a)| \\
&\qquad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_k(s,a)\|p_{M_k^+}(\cdot|s,a) - p(\cdot|s,a)\|_1 \max_{t\in\mathbb{T}_{(k)}} \frac{\mathbf{sp}_{\mathcal{S}_t}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})}{2}\Bigg]\,.
\end{aligned}$$

where $V^{+}_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}} \stackrel{\text{def}}{=} \mathbb{E}[R^{\pi_k}_{M_k^+}([t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}|\cdot)]$.

$$\begin{aligned}(30)_k &= \mathbb{E}\Big[R^{\pi^\star}_{M_k^+}(\mathbb{T}^D_{(k)}|s_{t_k}) - \ R^{\pi^\star}_{M^\star}(\mathbb{T}^D_{(k)}|s_{t_k})\Big] \\ &\geqslant \ -\mathbb{E}\Bigg[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_k(s,a)|r_{M_k^+}(s,a)-\mu(s,a)| \\ &\qquad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_k(s,a)\|p_{M_k^+}(\cdot|s,a)-p(\cdot|s,a)\|_1 \max_{t\in\mathbb{T}_{(k)}} \frac{\mathbf{sp}_{\mathcal{S}_t}(V^{+,\star}_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}})}{2}\Bigg],\end{aligned}$$

where $V^{+,\star}_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}} \stackrel{\text{def}}{=} \mathbb{E}[R^{\pi^\star}_{M_k^+}([t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k)}|\cdot)]$.

The result of Lemma 18 shows that actually only the difference of values between two possible *successor states* of a visited state (under $M^\star$) matters, and not between two arbitrary states: This is especially important in the setting of multi-chain MDPs.

5.4.3 Diameter and local diameter

First, let us bound the quantity $\tau^{\pi^{\circlearrowleft}}(s_{\mathbf{t}_{k,m}}, s_{\mathbf{t}_{k,m-1}+1})$. By construction, all actions that have been played on the path from state $s_{\mathbf{t}_{k,m-1}+1}$ to state $s_{\mathbf{t}_{k,m}}$ stay in the same region. Thus, $s_{\mathbf{t}_{k,m-1}+1}$ and $s_{\mathbf{t}_{k,m}}$ belong to the same region, and thus it holds

$$\mathbb{E}\Big[\tau^{\pi^{\circlearrowleft}}(s_{\mathbf{t}_{k,m}}, s_{\mathbf{t}_{k,m-1}+1})\Big] \leqslant d^\star\,. \tag{38}$$

Second, let us focus on $\tau^{\pi^{\curvearrowright\star}}(s_{\mathbf{t}_{k,m-1}+1}, \tilde{s}^\star_{\mathbf{t}_{k,m}})$. In order that $\tilde{s}^\star_{\mathbf{t}_{k,m}} \in \mathfrak{S}_j$ is reached by the optimal policy from state $s_{\mathbf{t}_{k,m-1}+1} \in \mathfrak{S}_i$ (with possibly $i=j$), then it must be that $p(\mathfrak{S}_j|\mathfrak{S}_i) > 0$. Thus, by assumption 1, at least one state-action pair $(\tilde{s},a)$ in the region $\mathfrak{S}_i$ satisfies $p(\mathfrak{S}_j|\tilde{s},a) \geqslant p_0$ without risking entering a third region. Now, it requires at most $d^\star$ steps on average to reach $\tilde{s}$ from a point in $\mathfrak{S}_i$, and thus at most $d^\star/p_0$ steps to enter $\mathfrak{S}_j$. On the other hand, it requires at most $d^\star$ steps in $\mathfrak{S}_i$ to reach this point from $s_{\mathbf{t}_{k,m-1}+1}$, and then at most $d^\star$ steps in $\mathfrak{S}_j$ to reach $\tilde{s}^\star_{\mathbf{t}_{k,m}}$ from a successor of $\tilde{s}$ in $\mathfrak{S}_j$. Thus, we deduce that:

$$\mathbb{E}\Big[\tau^{\pi^{\curvearrowright\star}}(s_{\mathbf{t}_{k,m-1}+1}, \tilde{s}^\star_{\mathbf{t}_{k,m}})\Big] \leqslant (2+\frac{1}{p_0})d^\star\,. \tag{39}$$

Third, let us focus on the span of the optimistic value, restricted to successor states of the visited states. Once again by construction, all actions that have been played on the path $\{(s_t,a_t)\}_{t\in\mathbb{T}^D_{(k),m}}$ from state $s_{\mathbf{t}_{k,m-1}+1}$ to state $s_{\mathbf{t}_{k,m}}$ stay in the same region (say $\mathfrak{S}_i$) and thus satisfy $\mathcal{T}(s_t,a_t;s_0) \leqslant d^\star$ for some state $s_0 \in \mathfrak{S}_i$. For any successor state $s' \neq s_0$, we must then have $1 + p(s'|s_t,a_t)\mathcal{T}(s';s_0) \leqslant \mathcal{T}(s_t,a_t;s_0) \leqslant d^\star$, and thus $\mathcal{T}(s';s_0) \leqslant \frac{d^\star-1}{p_{min}}$, assuming that all transitions happen with probability either 0 or at least $p_{min} > 0$. Now, the expected value of $\pi^\star$ in state s' must be at least the value in state s_0 plus the value on the way to reach it (by optimality in $M^\star$):

$$V^+_t(s') \geqslant \mathbb{E}\Big[V^+_{[t:t+\tau(s',s_0)-1]}(s') + V^+_{t+\tau(s',s_0)}(s_0)\Big] \geqslant \mathbb{E}\Big[V^+_{t+\tau(s',s_0)}(s_0)\Big]\,,$$

where the expectation is over the random variable $\tau(s', s_0) \in \mathbb{N}$. Likewise, $V_t^+(s')$ can be upper bounded by

$$\mathbb{E}\Big[V^+_{t-\tau(s_0,s')}(s_0)\Big] \geqslant \mathbb{E}\Big[V^+_{[t-\tau(s_0,s'):t-1]}(s_0)\Big] + V_t^+(s') \geqslant V_t^+(s')\,.$$

Thus, we deduce that for any successor s', s'' of (s_t, a_t) it holds

$$\begin{aligned} V_t^+(s') - V_t^+(s'') &\leqslant \mathbb{E}\Big[V^+_{t-\tau(s_0,s')}(s_0) - V^+_{t+\tau(s'',s_0)}(s_0)\Big] \leqslant \mathbb{E}\Big[\tau(s_0,s') + \tau(s'',s_0)\Big] \\ &\leqslant (2 + \frac{1}{p_0})d^\star + \frac{d^\star - 1}{p_{min}}\,. \end{aligned}$$

Thus, the span restricted on $\mathcal{S}_t$ is not greater than $\Big(2 + \frac{1}{p_0} + \frac{1}{p_{\min}}\Big)d^\star$. Following Jaksch et al. (2010), the same argument applies to the span of the optimistic value (following the optimistic policy) as well, since the augmented MDP contains at least the transitions of the true MDP, provided that $M^\star$ is plausible. Thus, on the event Ω that $M^\star$ is plausible, it holds

$$\max_{t \in \mathbb{T}_{(k),m}} \mathbf{sp}_{\mathcal{S}_t}(V^+_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k),m}}) \leqslant \Big(2 + \frac{1}{p_0} + \frac{1}{p_{\min}}\Big)d^\star\,. \tag{40}$$

Now since the optimistic value of $\pi^\star$ in state s' is larger than the optimal value of $\pi^\star$, which must be at least the value in state s_0 plus the value on the way to reach it, we can follow a similar argument to show that:

$$\max_{t \in \mathbb{T}_{(k),m}} \mathbf{sp}_{\mathcal{S}_t}(V^{+,\star}_{[t+1:t_{k+1}-1]\cap\mathbb{T}^D_{(k),m}}) \leqslant \Big(2 + \frac{1}{p_0} + \frac{1}{p_{\min}}\Big)d^\star\,. \tag{41}$$

Finally, we focus on the term $\tau_+(\tilde{s}_{\mathbf{t}_{k,m}}, s^+_{\mathbf{t}_{k,m}})$ that correspond to a shortest visit time when navigating in the augmented MDP M_k^+. The fact that the algorithm does not execute irreversible actions in $M^\star$ between time $\mathbf{t}_{k,m-1}+1$ and $\mathbf{t}_{k,m}$ does not discard, a priori, the possibility that an optimal policy $\pi^\star$ started from $s_{\mathbf{t}_{k,m-1}+1}$ executes actions that are irreversible in the MDP M_k^+. Put differently, the time to go $\tau_+(\tilde{s}_{\mathbf{t}_{k,m}}, s_{\mathbf{t}_{k,m-1}+1})$ from $\tilde{s}_{\mathbf{t}_{k,m}}$ back to $s_{\mathbf{t}_{k,m-1}+1}$ in M_k^+ may be huge if $s_{\mathbf{t}_{k,m-1}+1} \in \mathfrak{S}_i$ and $\pi^\star$ enters a different region $\mathfrak{S}_j$ when acting in M_k^+ during $\mathbb{T}_{(k),m}$. Note that if during $\mathbb{T}_{(k),m}$, $\pi^\star$ stays in the same region in $M^\star$, then since **A-ROGUE** does the same, the expected time to reach the state reached by **A-ROGUE** from the one reached by $\pi^\star$ is bounded by $d^\star$. Since times are shorten in the augmented MDP, $\tau_+(\tilde{s}_{\mathbf{t}_{k,m}}, s^+_{\mathbf{t}_{k,m}}) \leqslant d^\star$ in this case. Thus, we deduce the following bound that holds when $\pi^\star$ stays in the same region on $\mathbb{T}_{(k),m}$:

$$\tau_+(\tilde{s}_{\mathbf{t}_{k,m}}, s^+_{\mathbf{t}_{k,m}}) \leqslant d^\star\,. \tag{42}$$

Now the regret for not entering an optimal region is due to lack of observations in this region, and possibly in the current region as well. Lemma 20 proves a worst-case bound that handles not only the maximum number of roll-outs asked by the **A-ROGUE** before it can decide to play an irreversible action, but actually the number of samples (not only coming from roll-outs) so that every reachable state is well estimated: If **A-ROGUE** acts

in the MDP and gathers information, then it may require less roll-outs. Thus, this enables to capture the maximum regret due to not making an irreversible action when $\pi^\star$ does. That is, the regret due to the case when $\pi^\star$ enters another region in $M^\star$ but **A-ROGUE** does not can be handled by the bound on the total number of roll-out transitions from Lemma 20, and thus we can simply use the same bound.

Combining (42), (38) and (39) together, and slightly abusing notations (the following only applies to (k,m) such that $\pi^\star$ stays in the same region on $\mathbb{T}_{(k),m}$), we deduce that

$$\begin{aligned}\mathbb{E}_\Omega\Big[\sum_{k=1}^{K_T}\sum_{m=1}^{m_k+1}\tau_+(\tilde{s}_{\mathbf{t}_{k,m}},s^+_{\mathbf{t}_{k,m}})+\tau^{\pi^{\circlearrowleft}}(s_{\mathbf{t}_{k,m}},s_{\mathbf{t}_{k,m-1}+1})+\tau^{\pi^{\frown\star}}(s_{\mathbf{t}_{k,m-1}+1},\tilde{s}^\star_{\mathbf{t}_{k,m}})\Big]&\\ \leqslant \mathbb{E}_\Omega\Big[\sum_{k=1}^{K_T}(m_k+1)\left(4+\frac{1}{p_0}\right)d^\star\Big]&\\ \leqslant \left(4+\frac{1}{p_0}\right)\mathbb{E}\Big[md^\star+K_Td^\star\Big]\,.&\end{aligned}$$

Combining the result Lemma 18 with equations (40) and (41) we get an upper bound that can be extended to all episodes. Then, plugging this in the decomposition of the regret from Lemma 17 and continuing the unfolding steps, we deduce the following

Corollary 19 *Let us introduce the constant* $c_p \overset{\text{def}}{=} 2+\frac{1}{p_0}+\frac{1}{p_{\min}}$. *Then, it holds that*

$$\begin{aligned}\mathbb{E}_\Omega\Big[\mathfrak{R}_T\Big] \;\leqslant\; & \mathbb{E}[md^\star]+\left(4+\frac{1}{p_0}\right)\mathbb{E}\Big[md^\star+K_Td^\star\Big]+\mathbb{E}\Big[\sum_{k'=1}^{K_T}|\mathbb{T}^R_{k'}|\mathbb{I}\{\Omega\}\Big]\\ &+3\mathbb{E}_\Omega\Big[\sum_{k'=k}^{K_T}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}v_k(s,a)\big|r_{M^+_{k'}}(s,a)-\mu(s,a)\big|\\ &+\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}v_k(s,a)\big\|p_{M^+_k}(\cdot|s,a)-p(\cdot|s,a)\big\|_1c_pd^\star\Big]\,,\end{aligned}$$

where m denotes the total number of actions played by **A-ROGUE** that exit the current region.

5.4.4 Cumulative estimation error

We can combine the confidence bounds (5) and (6) together with the cumulative estimation error over K episodes (the expectation term in the last term of (**??**)) and obtain by Lemma 5 that on an event Ω of probability higher than $1-2\delta/T$, it is controlled as

$$\begin{aligned}\sum_{k'=1}^{K_T}\Big(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}v_{k'}(s,a)\big|r_{M^+_{k'}}(s,a)-\mu(s,a)\big|+\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}v_{k'}(s,a)\big\|p_{M^+_k}(\cdot|s,a)-p(\cdot|s,a)\big\|_1c_pd^\star\Big)&\\ \leqslant \Big(2\sqrt{\log(2^{S/2+1}AST^2/\delta)}c_pd^\star\Big)\sum_{k'=1}^{K_T}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}v_{k'}(s,a)N^{-1/2}_{t_{k'}}(s,a)&\\ \leqslant 2\sqrt{\log(2^{S/2+1}AST^2/\delta)}c_pd^\star(\sqrt{2}+1)\sqrt{SAT}\,,&\end{aligned}\tag{43}$$

where we used in the second inequality the fact that $v_{k'}(s,a) \leqslant N_{t_{k'}}(s,a)$ due to the criterion used for stopping episodes, together with an application of equation (20) from Jaksch et al. (2010) that uses Jensen's inequality, namely

$$\sum_{k'=1}^{K_T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{v_{k'}(s,a)}{\sqrt{N_{t_{k'}}(s,a)}} \leqslant (\sqrt{2}+1)\sqrt{SAT}\,.$$

We now combine (43) and (34) together with (33), reorganize the terms and use the control on K_T from Lemma 15, and obtain that the regret is controlled on an event Ω of high probability by

$$\begin{aligned}\mathbb{E}_\Omega\big[\mathfrak{R}_T\big] \;\leqslant\; & \left(5+\frac{1}{p_0}\right)2\sqrt{\frac{2d^\star}{\gamma^\star}}\sqrt{T+\frac{2d^\star}{\gamma^\star}} + \left(4+\frac{1}{p_0}\right)d^\star SA\log_2\Big(\frac{8T}{SA}\Big) \\ & + \mathbb{E}\Big[\sum_{k'=1}^{K_T}|\mathbb{T}_{k'}^R|\mathbb{I}\{\Omega\}\Big] + 6\sqrt{\log(2^{S/2+1}AST^2/\delta)}c_p d^\star(\sqrt{2}+1)\sqrt{SAT}\,. \quad (44)\end{aligned}$$

The notation $\mathbb{E}_\Omega$ has been introduced in section 5.1.2. Once again, we slightly abuse the notations, as here in $\mathbb{E}\Big[\sum_{k'=1}^{K_T}|\mathbb{T}_{k'}^R|\mathbb{I}\{\Omega\}\Big]$ we incorporate the episodes such that $\pi^\star$ exists the current region.

5.5 Total number of roll-outs

In this section, our goal is to control the cumulative sum of roll-out transitions $\sum_{k=1}^{K_T}|\mathbb{T}_k^R|$ that remains to be controlled. Note that **A-ROGUE** stops asking for new roll-out at a time t such that at each time step $t' \geqslant t$, at least one of the test of Line 10 and of Line 13 is passed. We study these two tests in order to derive a round on the total number of roll-out transitions. We will show that

Lemma 20 *Under the assumption that the parameter $\varepsilon < 1$ of* **A-ROGUE** *is chosen such that*

$$\varepsilon < \min\left\{\frac{p_0 d^\star}{1+p_0+d^\star}, \frac{p_0}{1+p_0}\right\},$$

and if Ω denotes the event that $M^\star$ is plausible for in all episodes, then it holds that

$$\mathbb{E}\Big[\sum_{k=1}^{K_T}|\mathbb{T}_k^R|\mathbb{I}\{\Omega\}\Big] \;\leqslant\; 16S\frac{(1+\varepsilon)^3}{\gamma^2\varepsilon^2(1-\varepsilon)}\log(2^{S+1}SAT^2/\delta)\mathbb{E}\Big[\sum_{k:D_{k-1}\leqslant d^\star} D_k^2\Big].$$

Part of this bound is obtained by computing the number of roll-outs executed from the initial region $\mathfrak{S}_{i_0}^\star$ to cover the entire state-space sufficiently often. Actually the same bound when adding to the number of roll-outs the regret in episodes such that $\pi^\star$ exists the current region. The proof is divided in several steps:

Number of roll-outs and covering times: Our goal is to control the number of roll-outs needed before we can ensure that in each state $s_{\mathbf{t}}$, the pessimistic return time

satisfies $\mathcal{T}_{\mathbf{t}}^{-} \leqslant 2D_k$ when $\mathcal{T}_{\mathbf{t}} \leqslant D_k$. By the bound (12) of Lemma 14 in Section (5.1.4), this is linked to the number of times each state-action pair (s,a) with $s \in \mathcal{S}(s_{\mathbf{t}}, D_k)$ is visited in total, where $\mathcal{S}(s,\ell)$ is the set of all states that are reachable from s in ℓ steps. Thus let s be a state reachable in d steps from $s_{\mathbf{t}}$, and let us call a roll-out with maximal number of steps bounded by d a d-roll-out. We introduce $p^{\pi,d}_{s_{\mathbf{t}},s}$, the probability that visiting s for the first time, starting from $s_{\mathbf{t}}$ and following policy π after playing $a_{\mathbf{t}}$ takes no more than d steps (in total). It is not difficult to show that this probability is given by

$$p^{\pi,d}_{s_{\mathbf{t}},s} = \delta_{s_{\mathbf{t}}}^{\top}\left(P^{a_{\mathbf{t}}} + \sum_{j=0}^{d-2} P^{a_{\mathbf{t}}}_{L,-s}\left(P^{\pi}_{-s}\right)^{j} P^{\pi}_{R,-s}\right)\delta_s,$$

where $P^{a_{\mathbf{t}}}_{L,-s}$ (respectively $P^{\pi}_{R,-s}$) is the $S-1\times S$ transition matrix ($S\times S-1$) induced by action $a_{\mathbf{t}}$ (policy π) with line (row) corresponding to state s removed, and δ_s is the S vector with values 1 for state s and 0 else. Note that it also means that with one d-roll-out of a policy π, $s \in \mathcal{S}(s_{\mathbf{t}}, d)$ is visited at least $p^{\pi,d}_{s_{\mathbf{t}},s}$ times on average. Thus, we deduce that no more than $n/p^{\pi,d}_{s_{\mathbf{t}},s}$ number of d-roll-outs of the policy π is needed so that $s \in \mathcal{S}(s_{\mathbf{t}}, d)$ is visited at least n times on average. It is worth noticing at this point that we only target a control on $\sum_{k=1}^{K_T} |\mathbb{T}_k^R|$ *in expectation*, as opposed to a high probability bound.[4]

Now, **A-ROGUE** uses roll-outs of length $d = (1+\frac{1}{\varepsilon})D_k$ in episode k. Note that $2C^p_{k,\mathbf{t}} c\left(1+\frac{1}{\varepsilon}\right) D_k \leqslant \alpha$ happens ($\alpha = 1$ is used in the recovery test l.10) as soon as all corresponding state-action pairs are visited at least

$$n_0 = \frac{2c^2\left(1+\frac{1}{\varepsilon}\right)^2 D_k^2}{\alpha^2}\log(2^{S+1}SAT^2/\delta)$$

times, and for this to happen, no more than $\sum_{s\in S(s_{\mathbf{t}},D_k)} n_0/p^{\pi_k^{\circlearrowleft -},d}_{s_{\mathbf{t}},s}$ many $\left(1+\frac{1}{\varepsilon}\right) D_k$-roll-outs of $\pi_k^{\circlearrowleft -}$ (the backup policy to $\mathbf{s}_k$) are required on average. Thus, assuming we can show that $1/p^{\pi_k^{\circlearrowleft -},d}_{s_{\mathbf{t}},s} \leqslant \beta$ for states $s \in S(s_{\mathbf{t}}, D_k)$ this means that every state in $\mathcal{S}(s_{\mathbf{t}}, d)$ is visited enough when **A-ROGUE** has observed at most the following number of transitions from $\pi_k^{\circlearrowleft -}$

$$n(D_k) \stackrel{\text{def}}{=} 2c^2\Big(1+\frac{1}{\varepsilon}\Big)^3 D_k^3 \big|S(s_{\mathbf{t}}, \big(1+\frac{1}{\varepsilon}\big)D_k)\big| \frac{\beta}{\alpha^2}\log(2^{S+1}SAT^2/\delta)\,.$$

Known and unknown local diameter Before determining β, let us discuss the guessed diameter D_k a little bit. In case an upper bound D on the diameter $d^\star$ is known in advance $D_k = D$, and thus at most $n(D)$ many roll-out transitions from $\pi_k^{\circlearrowleft -}$ are observed before the test of line 10 is passed for all states with $\mathcal{T}_t \leqslant D$.

Now, when no upper bound on the diameter $d^\star$ is known in advance and $D_k < d^\star$, that is the guessed diameter is smaller than the true local diameter, the states in $\mathcal{S}(s_{\mathbf{t}}, d^\star)\backslash\mathcal{S}(s_{\mathbf{t}}, D_k)$ may not be visited often enough, and thus one cannot ensure accurate estimation. Since $D_k = \log(t_k)$ increases with episodes, we eventually get to visit states reachable in $\mathcal{S}(s_{\mathbf{t}}, d^\star)$.

4. To derive such a stronger result, we would need, with the current proof, to get a control on the number of visits of $s \in \mathcal{S}(s_{\mathbf{t}}, d)$ in high probability, and not only in expectation, which is more difficult and luckily not needed for our purpose here.

However visiting the D_{k+1}-reachable states may require to revisit all the D_k reachable states as well and not only the frontier of this set (and this is reward consuming). In all cases, states are visited enough after the total number of roll-out transitions from $\pi_k^{\circlearrowleft -}$, $k \geqslant 1$ is at most

$$\sum_{k:D_k \leqslant d^\star} n(D_k) + n(D_{\tilde{k}}) = \sum_{k:D_{k-1} \leqslant d^\star} n(D_k)\,,$$

on average, where $\tilde{k}$ is the smallest episode such that $D_k \geqslant d^\star$. If $k \to D_k$ does not reach $d^\star$ fast enough, this potentially causes a large waste of roll-outs (but we do not want that D_k becomes to large with respect to $d^\star$ either) and thus a a large active regret. Choosing $D_k = \log(t_k)$ ensures the loss is not too important.

Lower bound on the probability of first visit in d-steps at most The next step is to find β such that $\beta \geqslant 1/p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d}$. First, note that in the case when $\mathcal{T}_{\pi_k^{\circlearrowleft -}}(s_{\mathbf{t}}, s) \leqslant (1+\alpha)D_k$ it holds that $p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d} \overset{d\to\infty}{\to} 1$, and thus

$$\begin{aligned}
\mathcal{T}_{\pi_k^{\circlearrowleft -}}(s_{\mathbf{t}}, s) &= \sum_{d=0}^{\infty} \big(p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d} - p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d-1}\big) d \\
&\geqslant \sum_{d > \lfloor (1+\frac{1}{\varepsilon})D_k \rfloor} \big(p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d} - p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d-1}\big) \Big(\big\lfloor (1+1/\varepsilon) D_k \big\rfloor + 1 \Big) \\
&\geqslant \Big(1 - p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -}, \left\lfloor (1+\frac{1}{\varepsilon})D_k \right\rfloor}\Big) \Big(1 + \frac{1}{\varepsilon}\Big) D_k\,, \qquad (45)
\end{aligned}$$

where in the first inequality we used the fact that $p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -},d}$ is a not decreasing function of d, and in the second one a telescoping argument. Combining (45) together with $\mathcal{T}_{\pi_k^{\circlearrowleft -}}(s_{\mathbf{t}}, s) \leqslant (1+\alpha)D_k$ and reorganizing the terms, we deduce that

$$p_{s_{\mathbf{t}},s}^{\pi_k^{\circlearrowleft -}, \lfloor (1+\frac{1}{\varepsilon})D_k \rfloor} \geqslant \frac{(1+\frac{1}{\varepsilon})D_k - D_k(1+\alpha)}{(1+\frac{1}{\varepsilon})D_k} = \frac{1-\alpha\varepsilon}{1+\varepsilon}\,,$$

which means that, provided that $\alpha\varepsilon < 1$, we can use $\beta = \frac{1+\varepsilon}{1-\alpha\varepsilon}$.

Finally, **A-ROGUE** alternates between roll-outs of the optimistic policy π_k^+ and roll-outs of the pessimistic backup policy $\pi_k^{\circlearrowleft -}$. We focused here on controlling the number transitions generated by $\pi_k^{\circlearrowleft -}$ only (at the price of losing a factor 2). More precisely, since the roll-outs from π_k^+ and from $\pi_k^{\circlearrowleft -}$ are of the same size and we alternate at every roll-out, by discarding the transitions received using π_k^+ (as non-informative in the worst case) we at most double the number of transitions needed before the condition on the pessimistic return time is satisfied. Thus, we deduce that the backup time is accurately estimated when the total number of transitions $\sum_{k=1}^{K_T} |\mathbb{T}_k^R|$ is at most

$$N(d^\star) \overset{\text{def}}{=} 4\frac{1+\varepsilon}{1-\alpha\varepsilon}\frac{c^2\left(1+\frac{1}{\varepsilon}\right)^3}{\alpha^2}\log(2^{S+1}SAT^2/\delta) \sum_{k:D_{k-1}\leqslant d^\star} D_k^3 \Big| S(s_{\mathbf{t}}, (1+1/\varepsilon)D_k) \Big|\,. \qquad (46)$$

Thus, far, we have seen that after this number of transitions on average has been observed for all visited states, every return time following an action that is not irreversible is well estimated, and thus no roll-out is asked for these states, that is, roll-out transitions may only be asked in case the optimistic action is irreversible.

Sub-optimality test The set $\bigcup_{s\in\mathfrak{S}^\star_{i_0}} S(s,(1+1/\varepsilon)D_k)$ covers the entire state space $\mathcal{S}$ if it is possible to reach any state from $\mathfrak{S}^\star_{i_0}$ is no more than $(1+1/\varepsilon)D_k$ steps. Since from the boundary of a region, we need no more than $d^\star + d^\star/p_0$ steps to enter the region and then $d^\star$ to reach any state inside the region, this requires at most $(2+\frac{1}{p_0})d^\star$ steps, and thus, we want condition on ε so that $(1+\frac{1}{\varepsilon})D_k \geqslant (2+\frac{1}{p_0})d^\star$. If $D_k \geqslant d^\star$, then this inequality holds for $\varepsilon \leqslant \frac{p_0}{1+p_0}$. In this case, for the previous number of roll-out transitions (46), we deduce that each state-action pair is visited at least n_0 times and thus it holds

$$\|C_k^p\|_\infty \leqslant \frac{\alpha}{2c(1+\frac{1}{\varepsilon})D_k}\,. \tag{47}$$

This of course provides a control on $\|C_k^\mu\|_\infty$ as well (simply look at the expression for both terms). Now, we want to ensure that (8) holds so that a $d^\star$ optimal action does pass the test of line 13. Indeed in this case, no more roll-out is asked by **A-ROGUE**. Using (47) in order to control both $\|C_k^p\|_\infty$ and $\|C_k^\mu\|_\infty$, and reorganizing the terms, we obtain the inequality

$$\|C_k^\mu\|_\infty + \Big(2+\frac{1}{p_0}\Big)d^\star\|C_k^p\|_\infty \;\leqslant \frac{\alpha}{2c(1+\frac{1}{\varepsilon})D_k}\left(\Big(2+\frac{1}{p_0}\Big)d^\star + \sqrt{\frac{\log(2AST^2/\delta)}{\log(2^{S+1}AST^2/\delta)}}\right).$$

As a result, the condition (11) for a $d^\star$-optimal action to pass the sub-optimality test of **A-ROGUE** is satisfied when $D_k \geqslant d^\star$, $\varepsilon \leqslant \frac{p_0}{1+p_0}$ and

$$\frac{2\alpha}{c(1+\frac{1}{\varepsilon})d^\star}\Big((2+\frac{1}{p_0})d^\star + 1\Big) < \gamma\,.$$

since $\alpha = 1$ and $c = \frac{2}{\gamma}$, then this simplifies to

$$1 \leqslant \frac{1+\frac{1}{\varepsilon}}{2+\frac{1}{p_0}+\frac{1}{d^\star}} \text{ that is } \varepsilon \leqslant \frac{1}{1+1/p_0+1/d^\star}\,. \tag{48}$$

Thus, we deduce that for the choice

$$\varepsilon \leqslant \min\left\{\frac{p_0 d^\star}{p_0 d^\star + p_0 + d^\star}, \frac{p_0}{1+p_0}\right\},$$

then after at most $N(d^\star)$ roll-out transitions asked by **A-ROGUE** for the episodes such that k such that $D_k \geqslant d^\star$, the recovering time is well-estimated and $d^\star$-optimal actions do pass the test of line 13, that is at least one of the tests of line 10 and line 13 is passed. Thus the total number of roll-out transitions asked by the algorithm is upper-bounded by (let us remind that δ is also a parameter of the algorithm)

$$\mathbb{E}\Big[\sum_{k=1}^{K_T}|\mathbb{T}_k^R|\mathbb{I}\{\Omega\}\Big] \;\leqslant\; 4S\frac{(1+\varepsilon)^4}{\varepsilon^3(1-\varepsilon)}c^2\log(2^{S+1}SAT^2/\delta)\mathbb{E}\Big[\sum_{k:D_{k-1}\leqslant d^\star} D_k^3\Big].$$

This concludes the proof of Lemma 20. □

5.6 Upper bound on the regret

We now bound the regret of the **A-ROGUE** first in the case when the diameter $d^\star$ is unknown, and then in the case an upper bound $D \geqslant d^\star$ is known. We remind that the algorithm uses $c = 2/\gamma$ for the scaling of the planning horizon of backup-up MDPs and $\alpha = 1$ for the recovery test.

In this section, we stop using the abuse of notation introduced in section 5.2, that consists in using horizon T instead of horizon $\tilde{T} = T - c(1+\frac{1}{\varepsilon})D_{K_T} \leqslant T$, which was done at the price of loosing only a regret of $c(1+\frac{1}{\varepsilon})D_{K_T} = \frac{1+\varepsilon}{\gamma\varepsilon}D_{K_t}$.

5.6.1 Regret when an upper-bound on the diameter is given

In case when a bound D on the diameter is known, the algorithm can use $D_k = D$ for all episodes. In that case,the bound on the number of roll-out transitions simplifies into

$$\mathbb{E}\Big[\sum_{k=1}^{K_{\tilde{T}}} |\mathbb{T}_k^R| \mathbb{I}\{\Omega\}\Big] \quad \leqslant \quad 16S\frac{(1+\varepsilon)^4}{\gamma^2\varepsilon^3(1-\varepsilon)}\log(2^{S+1}SAT^2/\delta)D^3\,. \tag{49}$$

We plug the previous bound (49) in the decomposition of the regret (44) and of (7), and deduce (provided that $\gamma \leqslant \gamma^\star$ and ε is as in Lemma 20) that the expected regret is bounded by

$$\begin{aligned}\mathbb{E}\Big[\mathfrak{R}_T\Big] \quad \leqslant \quad & \Big(5+\frac{1}{p_0}\Big)2\sqrt{\frac{2d^\star}{\gamma^\star}}\sqrt{T+\frac{2d^\star}{\gamma^\star}} + \Big(4+\frac{1}{p_0}\Big)d^\star SA\log_2\Big(\frac{8T}{SA}\Big) \\ & +6\sqrt{\log(2^{S/2+1}AST^2/\delta)}\Big(2+\frac{1}{p_0}+\frac{1}{p_{\min}}\Big)d^\star(\sqrt{2}+1)\sqrt{SAT} \\ & +16S\frac{(1+\varepsilon)^4}{\gamma^2\varepsilon^3(1-\varepsilon)}\log(2^{S+1}SAT^2/\delta)D^3 + \frac{1+\varepsilon}{\gamma\varepsilon}D + 2\delta\,.\end{aligned}$$

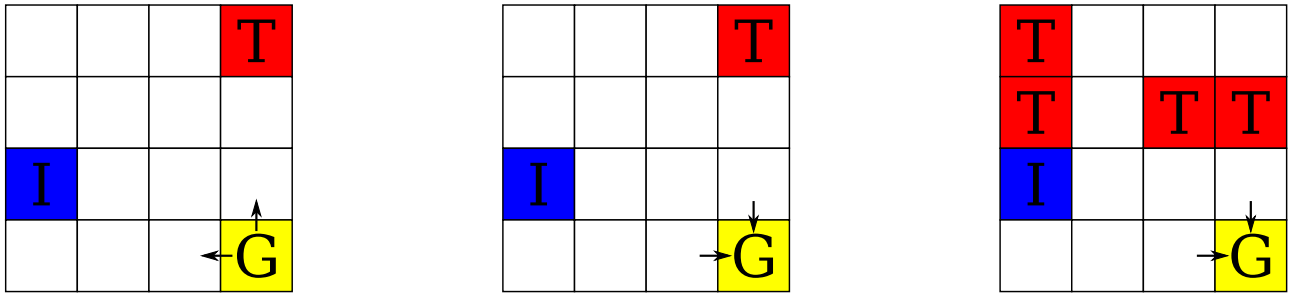
This concludes the first part of Theorem 7.

5.6.2 Regret when the diameter is unknown

In the case the diameter is unknown, a proxy for the diameter is given by D_k. Using a slowly increasing diameter $D_k = \log(t_k)$, it becomes eventually larger than $d^\star$. This results in a regret scaling with $\sqrt{T}$ with an additional constant term that is however exponential in the true diameter $d^\star$. One could make this dependency polynomial in $d^\star$ at the price of using a D_k polynomial in t_k leading to worst dependency in T. More formally, we get, using the crude lower bound $t_k \geqslant k$

$$\sum_{k:D_{k-1}\leqslant d^\star} D_k^3 \leqslant d^{\star 3}e^{d^\star}. \tag{50}$$

Thus, plugging in (50) into (49), and combining the resulting term with the decomposition of the regret (44) and (7), we deduce (provided that $\gamma \leqslant \gamma^\star$ and ε is as in Lemma 20)



(a) (b) (c)

Figure 8: Gridworld environments: I is the agent's initial state, T an absorbing trap, and G a goal state. Arrows entering (resp. leaving) a goal state indicate that it is absorbing (resp. not absorbing).

that the expected regret is controlled by

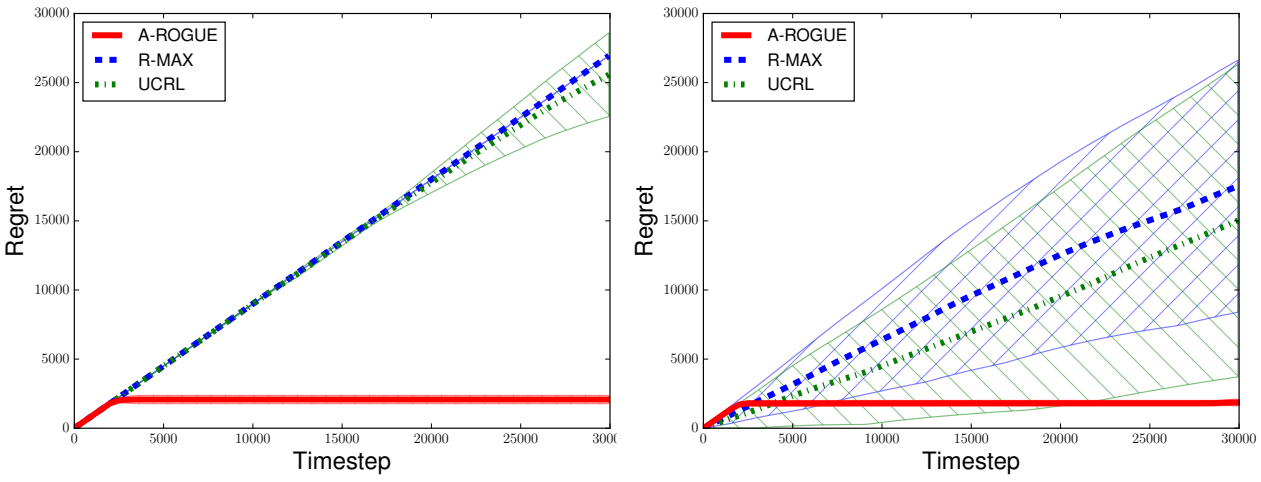
$$\begin{aligned}\mathbb{E}\left[\mathfrak{R}_T\right] \leqslant & \left(5+\frac{1}{p_0}\right)2\sqrt{\frac{2d^\star}{\gamma^\star}}\sqrt{T+\frac{2d^\star}{\gamma^\star}}+\left(4+\frac{1}{p_0}\right)d^\star SA\log_2\left(\frac{8T}{SA}\right)\\ &+6\sqrt{\log(2^{S/2+1}AST^2/\delta)}\left(2+\frac{1}{p_0}+\frac{1}{p_{min}}\right)d^\star(\sqrt{2}+1)\sqrt{SAT}\\ &16S\frac{(1+\varepsilon)^4}{\gamma^2\varepsilon^3(1-\varepsilon)}\log(2^{S+1}SAT^2/\delta)d^{\star 3}\min\{e^{d^\star},SA\log_2\left(\tfrac{8T}{SA}\right)\}\\ &+\frac{1+\varepsilon}{\gamma\varepsilon}\log(T)+2\delta\,.\end{aligned}$$

This concludes the proof of Theorem 7. □

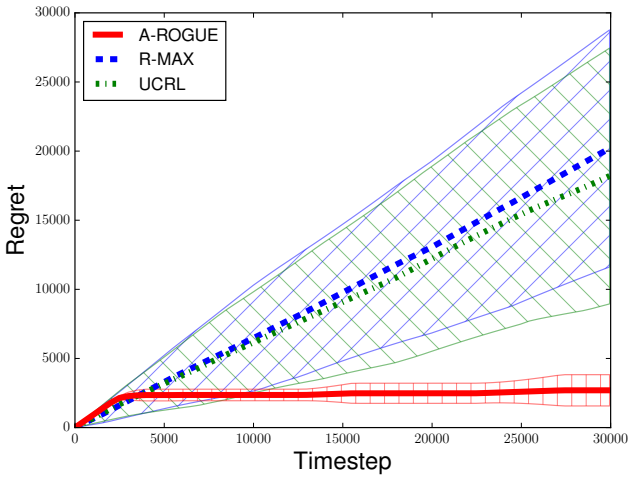
6. Experiments & Results

We compared **A-ROGUE** to **UCRL** Jaksch et al. (2010) and **R-Max** Brafman and Tennenholtz (2003) where **UCRL** enjoys strong regret guarantees for communicating MDPs[5] and **R-Max** has strong sample complexity guarantees. The reason for comparing only to such algorithms is because they are the only one we know that provably enjoy non-trivial performance guarantees in communicating MDPs. Thus they serve here as a benchmark. For illustration, we used three 4×4 gridworld MDPs (Figure 8) with stochastic transition probabilities and four actions $\{n,s,e,w\}$. Of course, the algorithms did not have knowledge about this structure and thus had to consider the full $S(S-1)A+SA=1024$-dimensional problem. Though these MDPs look extremely simple, they are actually tricky due to the presence of traps, and are enough to illustrate the main drawback of existing algorithms assuming fast recoverability. Given the agent's position (x,y), if the agent is not blocked by a wall or in an absorbing state, executing action n transfers the agent to state $(x,y+1)$ with probability 0.8 or to either $(x-1,y+1)$ or $(x+1,y+1)$ each with probability 0.1.

5. We do not compare against **REGAL**, as it is still an open question whether this algorithm can be implemented, see Bartlett and Tewari (2009).



(a) (b)



(c)

Figure 9: Comparison of average regret achieved by **A-ROGUE**, **UCRL**, and **R-Max** on gridworlds from (a) Figure 8a, (b) Figure 8b, and (c) Figure 8c. The shaded area represents one standard deviation for each algorithm.

Actions s, e, and w are defined similarly in the other three directions. Each gridworld has a single absorbing goal state, which gives a reward of 0.9. All three tasks contain one or more "trap" states, which are nearly-absorbing states where all actions give 0 reward and lead to the initial state with probability 10^{-4}. The reason for this low-probability transition is to ensure that the MDP has finite (though huge) diameter, and thus that the regret bound for **UCRL** holds. Putting the probability from 10^{-4} to 0 only favors our algorithm, but looks unfair with respect to other algorithms. Note also that this construction is different from imposing there is a reset action. The maximum possible immediate reward is $R_{\max} = 1$, which means that even if the goal state is discovered early, all state-action pairs need to be explored to determine an optimal policy. The horizon was set to $T = 30,000$. **R-Max**, was run with sampling parameter $m = 20$ (see Jong and Stone (2008) for explanations). For **UCRL**, we used $\delta = 0.05$. We loosely tuned **A-ROGUE** with $\delta = 0.05$, $\gamma = 0.35$, $\varepsilon = 1/3$, assuming that $p_0 \geqslant 1/2$, and we used the upper bound $D = 8$ on $d^\star$.

The first task (Figure 8a) contains a single trap and a non-absorbing goal state. The agent must take no irreversible actions to act optimally. **UCRL** and **R-Max** are expected to perform poorly here because even if they have discovered the goal state before the trap state, they keep exploring until all state-action pairs have been visited. Thus they get stuck in the trap state. **A-ROGUE**, on the other hand, can learn about the trap without entering it using roll-outs. As we can see in Figure 9a, **A-ROGUE** has small regret compared to the optimistic strategies **UCRL** and **R-Max**.

In the second task (Figure 8b), there is one trap state and one absorbing goal state. Here **UCRL** and **R-Max** explore the environment optimistically and discover the goal or the trap state with about equal probability, which leads to large regret and high variance, while **A-ROGUE** is able to use roll-outs to learn about the trap states and avoid them. Figure 9b shows that **UCRL** and **R-Max** incur here large regret and high variance. **A-ROGUE** achieves both small regret and low variance.

The third task (Figure 8c) contains four absorbing trap states and a single absorbing goal state. It is harder than the previous two because there are more "trap" states. Notice that again, **A-ROGUE** achieves low regret while **UCRL** and **R-Max** both have large regret and high variance (Figure 9c).

In all tasks, **A-ROGUE** successfully avoids entering bad traps with high probability. On the other hand, **UCRL** and **R-Max** accrue very large regret.

7. Conclusion

We obtained the first sublinear regret performance for multi-chain MDPs, as we are unaware of any work providing theoretical performance guarantees in the multichain MDP setting with no reset. The key ingredient for achieving this performance is to *actively* ask for external information. We captured the notion of external information here by giving the learner the option to ask for roll-outs. Most of the other possibilities require stronger assumptions and trivialize the problem by recasting it into a standard communicating MDP aith modified state-action space. Other possibilities, that are however restricted to specific situations include resorting to an expert, a simulator, or a teacher and can be seen as special cases of our setting. **A-ROGUE** is able to detect when it requires such external information and what experiment to design and ask for a respective roll-out. The heuristic idea behind

the algorithm is intuitive, which is a strength, but precisely quantifying the amount of external information needed, and how to use them in an efficient way is challenging and is the focus of this paper.

Trading-off planning costs and experimentation costs A key component in our model is that roll-outs are costly: We assume that a zero reward is given for each roll-out transition, thus incurring a maximal regret for each asked transition; this is a strong penalty. Instead, one could provide a small reward, especially when producing a roll-out step is much cheaper than a real execution. Alternatively, when roll-outs are numerical simulations, it makes sense to consider the numerical cost of running a simulation and the actual time it may take competing to executing the actual action. Since our analysis separates the bound on the total number of roll-out asked by the algorithm and the cumulative regret of the algorithm while acting, such modifications when using a different cost for roll-out transitions and executed actions can be done easily.

Perspective In the paper, we provide one answer to a critical question that appears in many realistic applications of Reinforcement Learning (that is, explicitly quantifying the amount of external information sufficient to guarantee a $O(\sqrt{T})$ learning regret when interacting with a multi-chain MDP). Obviously, when one wants to address a real-world problem, as one should pay attention to additional difficulties (such as inaccurate state model, trembling-hand phenomenon, robustness) that need to be addressed together with the multi-chain issue. In such situations, we do not claim to provide an end-to-end solution that can directly be used off-the-shelf but a component that must be combined with other mechanisms. Amongst the many questions one can ask, an interesting direction of research is how to deal with the case when the roll-outs requested by the learner only come from an approximate model, as opposed to the same model.

References

Peter L. Bartlett and Ambuj Tewari. REGAl: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, March 2003. ISSN 1532-4435.

Sonia Chernova and Manuela Veloso. Interactive policy learning through confidence-based autonomy. *J. Artif. Int. Res.*, 34(1):1–25, 2009. ISSN 1076-9757.

J. Clouse. On integrating apprentice learning and reinforcement learning. Technical report, Amherst, MA, USA, 1997.

Amir Massoud Farahmand. Action-gap phenomenon in reinforcement learning. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pages 172–180, Granada, Spain, 2011.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *Allerton Conference on Communication, Control, and Computing*, pages 115–122, 2010.

Javier García and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *J. Artif. Int. Res.*, 45(1):515–564, 2012. ISSN 1076-9757.

Alborz Geramifard, Joshua Redding, Nicholas Roy, and Jonathan P How. UAV cooperative control with stochastic risk models. In *American Control Conference (ACC), 2011*, pages 3393–3398. IEEE, 2011.

Getachew Hailu and Gerald Sommer. Learning by biasing. In *Robotics and Automation, 1998. Proceedings. 1998 IEEE International Conference on*, volume 3, pages 2168–2173. IEEE, 1998.

Alexander Hans, Daniel Schneega, Anton Maximilian Schäfer, and Steffen Udluft. Safe exploration for reinforcement learning. In *ESANN*, pages 143–148, 2008.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, August 2010. ISSN 1532-4435.

Nicholas K. Jong and Peter Stone. Hierarchical model-based reinforcement learning: Rmax + MAXQ. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning*, July 2008.

Sham Machandranath Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

Andrey Kolobov, Mausam, and Daniel S. Weld. A theory of goal-oriented MDPs with dead ends. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA*, pages 438–447, 2012.

O. Maillard, P. Nguyen, R. Ortner, and D. Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International conference on Machine Learning*, JMLR W&CP 28(1), pages 543–551, Atlanta, USA, 2013.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in Markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, Edinburgh, Scotland, UK, 2012.

Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In Sanjay Jain, Rmi Munos, Frank Stephan, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pages 158–172. Springer Berlin Heidelberg, 2013.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

Florent Teichteil-Königsbuch, Vincent Vidal, and Guillaume Infantes. Extending classical planning heuristics to probabilistic planning with dead-ends. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA*, 2011.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Technical Report HPL-2003-97*, 2003. URL `www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf`.