Exploration in Reward Machines with Low Regret

Hippolyte Bourel
Department of Computer Science
University of Copenhagen

Anders Jonsson
Dept. Information and Communication Technologies
Universitat Pompeu Fabra

Odalric-Ambrym Maillard
Univ. Lille, Inria, CNRS, Centrale Lille
UMR 9189 – CRIStAL

Mohammad Sadegh Talebi
Department of Computer Science
University of Copenhagen

Abstract

We study reinforcement learning (RL) for decision processes with non-Markovian reward, in which high-level knowledge in the form of reward machines is available to the learner. Specifically, we investigate the efficiency of RL under the average-reward criterion, in the regret minimization setting. We propose two model-based RL algorithms that each exploits the structure of the reward machines, and show that our algorithms achieve regret bounds that improve over those of baselines by a multiplicative factor proportional to the number of states in the underlying reward machine. To the best of our knowledge, the proposed algorithms and associated regret bounds are the first to tailor the analysis specifically to reward machines, either in the episodic or average-reward settings. We also present a regret lower bound for the studied setting, which indicates that the proposed algorithms achieve a near-optimal regret. Finally, we report numerical experiments that demonstrate the superiority of the proposed algorithms over existing baselines in practice.

1 INTRODUCTION

Most state-of-the-art reinforcement learning (RL) algorithms assume that the underlying decision process has Markovian reward and dynamics, i.e. that future observations depend only on the current state-action of the system. In this case, the Markov Decision Process (MDP) is a suitable mathematical model for representing the task to be solved (Puterman, 2014). However, there are many application scenarios with non-Markovian reward and/or dynamics (Bacchus et al., 1996; Brafman and De Giacomo, 2019; Littman et al., 2017) that are more appropriately modeled as *Non-Markovian Decision Processes (NMDPs)*. NMDPs capture environments in which the optimal action depends on events that occurred in the past, implying that the learning agent has to remember parts of the history. For example, a robot may receive a reward for delivering an item only if the item was previously requested, and a self-driving car is more likely to skid and lose control if it previously rained. Consider a mobile robot that has to track an object which is no longer in the robot's field of view. By remembering where the object was last seen, the robot has a better chance of discovering the object again. An even more precise estimation is given by the sequence of last observations (which also capture direction of movement). This can be formalized by defining high-level events that correspond to past observations.

In general, the future observations of an NMDP can depend on an infinite history or trace, preventing efficient learning. Consequently, recent research has focused on tractable sub-classes of NMDPs. In Regular Decision Processes (RDPs) (Brafman and De Giacomo, 2019), the reward function and next state distribution are conditioned on logical formulas, making RDPs fully observable. Another popular formalism is the *Reward Machine* (RM) (Toro Icarte et al., 2018, 2022), which is a Deterministic Finite-State Automaton (DFA) providing a compact representation of history that compresses the entire sequence of past events into a single state, which can be combined with the current observation to determine the best action. Hence, the current state of the reward machine is sufficient to fully specify the reward function.

In this paper, we investigate RL in Markov decision processes with reward machines (MDPRMs) under the average-reward criterion, where the agent performance is measured through the notion of regret with respect to an oracle aware

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

of the transition dynamics and associated reward functions. The goal of the agent is to minimize its regret, which entails balancing exploration and exploitation. We focus on an intermediate setting where the underlying a DFA is *known*, while the actual transition distributions are *unknown*. For a given MDPRM, it is possible to formulate an *equivalent cross-product* MDP (adhering to the Markov property) as discussed in the literature (Toro Icarte et al., 2018) – see Lemma 1 here – and apply provably efficient off-the-shelf algorithms *obliviously* to the structure induced by the MDPRM. However, this would lead to large regret, both empirically and theoretically, as the associated cross-product MDP usually has a large state-space. Therefore, sample-efficient learning of near-optimal policies entails exploiting the intrinsic structure of MDPRMs in an efficient manner.

1.1 Outline and Contributions

We formalize regret minimization in average-reward MDPRMs (Section 2), and establish a first, to the best of our knowledge, regret lower bound for MDPRMs (Section 5). We introduce two algorithms, `UCRL-RM-L1` and `UCRL-RM-B`, whose designs are inspired by the celebrated UCRL2 algorithm (Jaksch et al., 2010) and its variants (e.g., (Fruit et al., 2018b, 2020; Zhang and Ji, 2019)), but they are tailored to leverage the structure in MDPRMs; see Section 3. The two algorithms admit a similar design and mainly differ in the choice of confidence sets used. Nonetheless, they attain different performance in terms of empirical and theoretical regret. We present numerical experiments (in Section 6) demonstrating that both `UCRL-RM-L1` and `UCRL-RM-B` significantly improve over existing tabular RL baselines when directly applied to the associated cross-product MDP. They also attain smaller regret bounds than these baselines as detailed in Section 4. Specifically, `UCRL-RM-L1` (resp. `UCRL-RM-B`) achieves a regret growing as $\widetilde{O}(\sqrt{\mathbf{c}_M OAT})$ (resp. $\widetilde{O}(\sqrt{\mathbf{c}'_M OAT})$) in an MDPRM M, where OA is the size of its observation-action space, T is the number of time steps, and $\widetilde{O}(\cdot)$ hides logarithmic and constant terms (Section 4). Furthermore, $\mathbf{c}_M$ and $\mathbf{c}'_M$ are MDP-dependent quantities. Specifically, $\mathbf{c}_M$ and $\mathbf{c}'_M$ are defined in terms of a novel notion of connectivity in MDPRMs, which we call the RM-restricted diameter, that is a problem-dependent refinement of the diameter D_{cp} of the cross-product MDP associated to M. [1] The RM-restricted diameter of M reflects the connectivity in M *jointly* determined by the dynamics and the sparsity structure of the reward machine, and we believe it could be of interest in other settings of reward machines. The RM-restricted diameter is *always* smaller than D_{cp}, and in some MDPRM instances, it is proportional to D_{cp}/Q, where Q denotes the number of states of the reward machine. The presented regret bounds exhibit a two-fold improvement over those of baselines: (i) They are independent of Q, whereas the existing bounds depend on $\sqrt{Q}$; and (ii) existing bound necessarily depend on D_{cp} or $\sqrt{D_{\mathsf{cp}}}$, whereas ours depend (via $\mathbf{c}_M$) on RM-restricted diameters of the various states. In summary, our regret bounds improve over the state-of-the-art by a factor between $Q^{1/2}$ and $Q^{3/2}$, depending on how large T is and on the sparsity in RM (Section 4). To the best of our knowledge, this work is the first studying regret minimization in average-reward MDPRMs, and the proposed algorithms constitute the first attempt to tailor and analyse regret specifically for MDPRMs or MDPs with associated DFAs.

Regarding the assumption of known RM, we note that in most practical applications of RL, a human expert specifies the reward. RMs provide an intuitive way to specify reward in terms of high-level events without knowledge of the task dynamics. Labels enable a human expert to express precedence, e.g., that event A should take place before B. Such precedence information exists in many applications of RL, and we therefore believe that prior knowledge of the RM is not an unreasonable assumption. The case of unknown RM, while admitting broader applications, turns the problem into a POMDP which admits weaker learning guarantees. Achieving a sublinear regret in such POMDPs may require additional assumptions (e.g., uniqueness of RM) that render strong in some applications.

1.2 Related Work

In the case of Markovian rewards and dynamics, there is a rich and growing literature on average-reward RL, where several algorithms with theoretical regret guarantees are presented; see, e.g., (Bartlett and Tewari, 2009; Burnetas and Katehakis, 1997; Jaksch et al., 2010; Ouyang et al., 2017; Fruit et al., 2018a; Talebi and Maillard, 2018; Tossou et al., 2019; Wei et al., 2020; Bourel et al., 2020; QIAN et al., 2019; Zhang and Ji, 2019; Pesquerel and Maillard, 2022)). In the absence of structure assumptions, as established by Jaksch et al. (2010), no algorithm can have a regret lower than $\Omega(\sqrt{DSAT})$ in a communicating MDP with S states, A actions, diameter D, and after T steps of interactions. The best available regret bounds, achievable by computationally implementable algorithms, grow as $O(\sqrt{DSAKT}\log(T))$ (Fruit et al., 2020) or as $O(D\sqrt{KSAT\log(T)})$ (Fruit et al., 2018a), where K denotes the maximal number of next-states under any state-action pair in the MDP. (We note that Zhang and Ji (2019) report a regret of $O(\sqrt{DSAT\log(T)})$, but the presented algorithm does not admit a computationally efficient implementation.) Besides this growing line of research, there is a rich literature on RL in episodic MDPs; see, e.g., (Dann et al., 2017; Gheshlaghi Azar et al., 2017).

The focus of this paper is RL for the class of MDPRMs under the average-reward criterion, in an intermediate setting where the underlying RM is *known*. Several authors propose

[1] The diameter of a finite MDP M is defined as $D = \max_{s\neq s'} \min_\pi \mathbb{E}[T^\pi(s,s')]$, where $T^\pi(s,s')$ is the number of steps it takes to reach s' starting from s and following policy π (Jaksch et al., 2010). For an MDPRM M, we denote its associated cross-product MDP by M_{cp}, and its diameter by D_{cp}.

algorithms with polynomial sample complexity or sublinear regret for different classes of NMDPs (Lattimore et al., 2013; Maillard et al., 2013; Sunehag and Hutter, 2015). Although these algorithms could be applied to MDPRMs, they do not exploit the particular structure of the DFAs, and hence the resulting theoretical bounds are not as tight as ours. The S3M algorithm of Abadi and Brafman (2020) integrates RL with the logical formulas of RDPs, but does not admit polynomial sample complexity in the PAC setting. Ronca and De Giacomo (2021) present the first RL algorithm for RDPs whose PAC sample complexity grows polynomially in terms of the underlying parameters, though the sample complexity bound is not very tight and could not be used to derive a high-probability regret bound.

Research on reward machines is relatively recent, but has grown quickly in popularity and already attracted many researchers to the field. Initial research focused on proving convergence guarantees for RL algorithms specifically devised for RMs (Toro Icarte et al., 2018, 2022). There is also a rich literature on RL with temporal specifications expressed in Linear Temporal Logic (LTL) (Cai et al., 2022; Camacho et al., 2019; Hamilton et al., 2022; Kazemi et al., 2022; Xu and Fekri, 2022). Because of the equivalence between LTL and Büchi automata, LTL specifications are often translated to DFAs similar to RMs, and sometimes combined with hierarchical RL (den Hengst et al., 2022). More recently, many researchers have investigated how to learn RMs or similar DFAs from experience in the form of traces (Abate et al., 2022; De Giacomo et al., 2020; Furelos-Blanco et al., 2021; Gaon and Brafman, 2020; Hasanbeig et al., 2021; Saqur, 2022; Toro Icarte et al., 2019; Verginis et al., 2022; Xu et al., 2020), and extensions to stochastic and probabilistic RMs in which either the rewards or the transitions are non-deterministic exist (Corazza et al., 2022; Dohmen et al., 2022). Another recent extension is to learn entire hierarchies of RMs (Furelos-Blanco et al., 2022).

RL with RMs or LTL specifications has been successfully applied to complex robotic tasks with non-Markovian rewards (Camacho et al., 2021; Mo et al., 2022; Shah et al., 2020). RMs have also been used in combination with multiagent RL (Dann et al., 2022; Neary et al., 2021) and approaches for zero-shot learning based on compositionality (Tasse et al., 2022; Zheng et al., 2022). Clark and Thollard (2004) study the learnability of Probabilistic DFAs in the PAC setting. However, we are not aware of any previous work involving RMs that report regret bounds in the episodic or average-reward setting.

NMDPs are related to Partially-Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998; Sondik, 1971), in which the current agent observation is not sufficient to predict the future. Two common approaches for POMDPs are 1) maintaining a finite history of observations; or 2) maintaining a belief state. However, a finite history of observations yields a history space whose size is exponential in the history length, while maintaining and updating a belief state is worst-case exponential in the size of the original observation space. The relationship between Probabilistic DFAs, hidden Markov models (HMMs) and POMDPs has been previously studied by Dupont et al. (2005).

Finally, we mention that MDPRMs might be viewed as non-stationary MDPs, where rewards vary over time. Algorithms for non-stationary MDPs (e.g., (Wei and Luo, 2021) and references therein) crucially rely on the number of reward changes to be sublinear (in T) in order to achieve a sublinear regret. However, the number of changes in MDPRMs could grow linearly in T. As a result, directly applying such algorithms may yield a linear regret in MDPRMs.

Notations. Given a set A, Δ_A denotes the simplex of probability distributions over A. With a slight abuse of notation, we use $\Delta_{X,A}$ to denote the set of mappings of the form $X \to \Delta_A$. A^* denotes (possibly empty) sequences of elements from A, and A^+ denotes non-empty sequences. $\mathbb{I}_A$ denotes the indicator function of event A.

2 PROBLEM FORMULATION

2.1 MDPRMs: Average-Reward Markov Decision Processes with Reward Machines

We begin with introducing some necessary background.

Labeled Markov Decision Processes. A *labeled average-reward MDP* (Xu et al., 2020) is a tuple $M = (\mathcal{O}, \mathcal{A}, p, \mathbf{R}, \mathcal{P}, L)$, where $\mathcal{O}$ is a finite set of (observation) states with cardinality O, $\mathcal{A}$ is a finite set of actions available at each state with cardinality A, $p : \mathcal{O}\times\mathcal{A} \to \Delta_{\mathcal{O}}$ is the transition function such that $p(o'|o,a)$ denotes the probability of transiting to state $o' \in \mathcal{O}$, when executing action $a \in \mathcal{A}$ in state $o \in \mathcal{O}$. $\mathbf{R} : (\mathcal{O}\times\mathcal{A})^+ \to \Delta_{[0,1]}$ denotes a history-dependent reward function such that for every history $h \in (\mathcal{O}\times\mathcal{A})^* \times \mathcal{O}$ and action $a \in \mathcal{A}$, $\mathbf{R}(h,a)$ defines a reward distribution.[2] $\mathcal{P}$ denotes a set of atomic propositions and $L : \mathcal{O}\times\mathcal{A}\times\mathcal{O} \to 2^{\mathcal{P}}$ denotes a labeling function assigning a subset of $\mathcal{P}$ to each (o,a,o'). These labels describe high-level events associated to (o,a,o') triplets that can be detected from the environment. The agent interacts with M as follows. At each time step $t \in \mathbb{N}$, the agent is in state $o_t \in \mathcal{O}$ and chooses an action $a_t \in \mathcal{A}$ based on $h_t := (o_1, a_1, \dots, o_{t-1}, a_{t-1}, o_t)$. Upon executing a_t in o_t, M generates a next-state $o_{t+1} \sim p(\cdot|o_t,a_t)$ and assigns a label $\sigma_t = L(o_t, a_t, o_{t+1})$. Then, the agent receives a reward $r_t \sim \mathbf{R}(h_t, a_t)$. Then, the state transits to o_{t+1} and a new decision step begins. As in MDPs, after T steps of interactions, the agent's cumulative reward is $\sum_{t=1}^T r_t$.

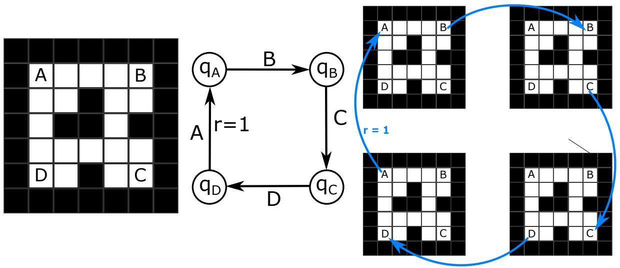
Reward Machines (RMs). We restrict attention to a class of non-Markovian reward functions that are encoded by

[2] This can be straightforwardly extended to σ-sub-Gaussian reward distributions with unbounded supports.

RMs (Toro Icarte et al., 2018, 2022), whose definition coincides with conventional DFAs. An RM is a tuple $\mathcal{R} = (\mathcal{Q}, 2^{\mathcal{P}}, \tau, \nu)$, where $\mathcal{Q}$ is a finite set of states and $2^{\mathcal{P}}$ is an input alphabet. $\tau : \mathcal{Q}\times 2^{\mathcal{P}} \to \mathcal{Q}$ denotes a deterministic transition function such that $q' = \tau(q, \sigma)$ denotes the next-state of $\mathcal{R}$ when an input σ is received in state q, with the convention that $\tau(q, \emptyset) = q$. Finally, $\nu : \mathcal{Q}\times 2^{\mathcal{P}} \to \Delta_{\mathcal{O}\times\mathcal{A},[0,1]}$ denotes the output function of $\mathcal{R}$, which returns a reward function $\mathbf{r} : \mathcal{O} \times \mathcal{A} \to \Delta_{[0,1]}$.[3] In words, the RM $\mathcal{R}$ converts a (sequentially received) sequence of labels to a sequence of Markovian reward functions such that the output reward function at time t is $\mathbf{r}_t = \nu(q_t, \sigma_t)$, where $\mathbf{r}_t : \mathcal{O} \times \mathcal{A} \to \Delta_{[0,1]}$ only depends on the current state q_t and current label σ_t. Conditioned on (q_t, σ_t), $\mathbf{r}_t$ is independent of $(q_1, \sigma_1, \ldots, q_{t-1}, \sigma_{t-1})$. Thus, RMs provide a compact representation for a class of non-Markovian reward functions that can depend on the entire history.

Average-Reward MDPs with Reward Machines. Restricting the generic history-dependent reward function $\mathbf{R}$ to RMs leads to MDPs with RMs. Formally, an average-reward MDP with RM (MDPRM) is a tuple $M = (\mathcal{O}, \mathcal{A}, p, \mathcal{R}, \mathcal{P}, L)$, where $\mathcal{O}, \mathcal{A}, p, \mathcal{P}$, and L are defined as in (labeled) average-reward MDPs, and where $\mathcal{R}$ is an RM, which generates reward functions. The agent's interaction with an MDPRM M proceeds as follows. At each time $t \in \mathbb{N}$, the agent observes $o_t \in \mathcal{O}$ and $q_t \in \mathcal{Q}$, and chooses an action $a_t \in \mathcal{A}$ based on o_t and q_t as well as (potentially) her past decisions and observations. The environment generates a next-state $o_{t+1} \sim p(\cdot|o_t, a_t)$ and reveals an event $\sigma_t = L(o_t, a_t, o_{t+1})$. The RM $\mathcal{R}$, being in state q_t, receives σ_t and outputs a reward function $\mathbf{r}_t = \nu(q_t, \sigma_t)$ which is a mapping $\mathbf{r}_t : \mathcal{O} \times \mathcal{A} \to \Delta_{[0,1]}$. Then, the agent receives a reward $r_t \sim \mathbf{r}_t(o_t, a_t)$ (at the end of the current time step). Then, the environment and RM states transit to their next states o_{t+1} and $q_{t+1} = \tau(q_t, \sigma_t)$, and a new step begins.

An example MDPRM is illustrated in Figure 1, which consists of a labeled 4-room gridworld (Figure 1(a)) and an RM with 4 states (Figure 1(b)). It corresponds to a patrolling task in a gridworld consisting in repeatedly visiting the specific locations A, B, C, and D, in that order, similar to OfficeWorld (Toro Icarte et al., 2018). The agent has 4 actions in each location (observation state) of the gridworld, corresponding to going up, down, right, and left. (We assume that walls act as reflectors.) All actions lead to stochastic transitions: A given action moves the agent in the intended direction with probability (w.p.) 0.7, and in each perpendicular direction w.p. 0.15. When visiting a corner of the gridworld and performing any action, the agent observes the respective events A, B, C, or D. Hence, we set $\mathcal{P} = \{A, B, C, D\}$.[4] The RM requires observing these events in the fixed order (A → B → C → D) to produce a fixed reward of 1. We remark that the current MDP observation (i.e., location) is not sufficient to predict what to do next, and therefore has to be combined with the current RM state. However, the task can be represented by a Markovian reward using the cross-product MDP shown in Figure 1(c).



(a) Labeled-MDP (b) RM (c) Equivalent MDP

Figure 1: An MDPRM consisting of a 4-room gridworld and an RM, corresponding to a patrolling task

For a given MDPRM, one can derive an *equivalent tabular MDP* (with a Markovian reward function), whose state-space is $\mathcal{S} := \mathcal{Q} \times \mathcal{O}$. Hence, this associated MDP is often called the *cross-product MDP* of M. We shall use $M_{\textsf{cp}}$ to denote the associated cross-product MDP to M. The following lemma characterizes $M_{\textsf{cp}}$. Variants of this result appeared in, e.g., (Toro Icarte et al., 2022); we state it for completeness and slightly extend it to hold for reward distributions. Proof is in Appendix A.

Lemma 1 *Let $M = (\mathcal{O}, \mathcal{A}, p, \mathcal{R}, \mathcal{P}, L)$ be a finite MDPRM. Then, an associated cross-product MDP to M is $M_{\textsf{cp}} = (\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S} = \mathcal{Q} \times \mathcal{O}$, and where for $s = (q, o)$, $s' = (q', o') \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$P(s'|s, a) = p(o'|o, a)\mathbb{I}_{\{q' = \tau(q, L(o,a,o'))\}}\,, \tag{1}$$

$$R(s, a) = \sum_{o' \in \mathcal{O}} p(o'|o, a)\nu(q, L(o, a, o')). \tag{2}$$

The equivalence between M and $M_{\textsf{cp}}$ implies that one could apply any off-the-shelf algorithm to $M_{\textsf{cp}}$, as it perfectly adheres to the Markovian property. In fact, $M_{\textsf{cp}}$ can be used as a proxy to develop learning algorithms for MDPRM.

2.2 Regret Minimization in MDPRMs

We are now ready to formalize RL in MDPRMs in the regret minimization setting, which is the main focus of this paper. As in tabular RL, it involves an agent who is seeking to maximize its cumulative reward, and its performance is measured in terms of regret with respect to an oracle algorithm who knows and always applies a gain-optimal policy.

[3] This is very similar to the standard definition of RM by Toro Icarte et al. (2022), though in our case the set of terminal states is empty. It is worth noting that DFAs and RMs admit identical definitions except that RMs output reward functions.

[4] $\mathcal{P}$ could be extended, e.g., by adding F (denoting a hypothetical 'furniture') so that one can introduce logical formulas using F with office locations to indicate whether the furniture is broken.

To formally define regret, we introduce some necessary concepts. A stationary deterministic policy in an MDPRM M is a mapping $\pi : \mathcal{Q}\times\mathcal{O}\to\mathcal{A}$ prescribing an action $\pi(q,o)\in\mathcal{A}$ for all $(q,o)\in\mathcal{Q}\times\mathcal{O}$. Let Π be the set of all such policies in M. The long-term average-reward (or gain) of policy $\pi\in\Pi$, when starting in (q,o), is defined as:

$$g^\pi(q,o) = \liminf_{T\to\infty}\frac{1}{T}\mathbb{E}_\pi\Big[\sum_{t=1}^{T} r_t\Big| q_1 = q, o_1 = o\Big]$$

where $r_t\sim\mathbf{r}_t(o_t,a_t)$ and $\mathbf{r}_t=\nu(q_t, L(o_t,\pi(q_t,o_t),o_{t+1}))$ for all t. Here the expectation is taken with respect to randomness in r_t and over all possible histories h_t (which implicitly depend on generated events too). Let $g^\star = \max_\pi g^\pi$ denote the optimal gain over all (possibly history dependent) policies. Any policy achieving $g^\star$ is an optimal policy. Following the same arguments as in tabular MDPs together with the equivalence between M and its M_{cp} (Lemma 1), it is guaranteed that there exists at least one optimal policy in Π. *We assume that the transition function p is initially unknown, but that the RM $\mathcal{R}$ is* known. The agent interacts with M for T steps according to the protocol specified in the previous subsection (i.e., observing (o_t,q_t), choosing a_t based on past experience, observing the event $\sigma_t = L(o_t,a_t,o_{t+1})$ and receiving the reward $r_t\sim\mathbf{r}_t(o_t,a_t)$). We define the regret of an agent (or learning algorithm) $\mathbb{A}$ as

$$\mathfrak{R}(\mathbb{A},T) := Tg^\star - \sum_{t=1}^{T} r_t.$$

Alternatively, the agent's objective is to minimize regret, which entails balancing exploration and exploitation. We stress that regret $\mathfrak{R}(\mathbb{A},T)$ compares the T-step reward collected by $\mathbb{A}$ against an oracle that uses the *same reward machine* $\mathcal{R}$ as the agent.[5] In order to achieve a regret sublinearly growing with T, we need some notion of connectivity in the MDPRM, as in tabular MDPs. We first recall that a tabular MDP is communicating if it is possible to reach any state from any other state under some stationary deterministic policy. Alternatively, an MDP is communicating if and only if its diameter is finite (Jaksch et al., 2010) (see the footnote on page 2). In summary, we impose the following assumptions:

Assumption 1 *We assume: (i) the RM $\mathcal{R}$ is known, and (ii) the associated MDP M_{cp} is communicating.* [6]

3 LEARNING ALGORITHMS FOR MDPRMS

In this section, we present algorithms for learning in MDPRMs, which follow a model-based approach, similar to UCRL2 (Jaksch et al., 2010) and its variants (Bourel et al., 2020; Fruit et al., 2020, 2018b; QIAN et al., 2019; Zhang and Ji, 2019). To simplify exposition, we assume that the reward distributions $\nu(\cdot,\cdot)$ of the RM are known. This assumption can be easily relaxed at the expense of a slightly increased regret. We discuss in Appendix B how to tailor the algorithms to the case of unknown rewards.

3.1 Confidence Sets

We begin with introducing empirical estimates and confidence sets used by the algorithms. We first present confidence sets for the observation dynamics p, and then show how they yield confidence sets for the transition and reward functions of the cross-product MDP M_{cp}.

Confidence Sets for Observation Dynamics p. Formally, under a given algorithm, let $N_t(o,a,o')$ denote the number of times a visit to (o,a) was followed by a visit to o', up to time t: $N_t(o,a,o') := \sum_{i=1}^{t-1}\mathbb{I}_{\{(o_i,a_i,o'_{i+1})=(o,a,o')\}}$. Further, $N_t(o,a) := \max\{1,\sum_{o'}N_t(o,a,o')\}$. Using the observations collected up to $t\geq 1$, we define the empirical estimate $\widehat{p}_t(o'|o,a) = \frac{N_t(o,a,o')}{N_t(o,a)}$ for $p(o'|o,a)$, for any $o,o'\in\mathcal{O}$ and $a\in\mathcal{A}$. We consider two confidence sets for p. The first one uses a *time-uniform* variant of Weissman's concentration inequality (Weissman et al., 2003) and is defined as follows (Asadi et al., 2019):

$$C^1_{t,\delta}(o,a) = \Big\{p'\in\Delta_{\mathcal{O}} : \|\widehat{p}_t(\cdot|o,a)-p'\|_1 \leq \beta_{N_t(o,a)}(\delta)\Big\}$$

and $C^1_{t,\delta} = \cap_{o,a}C^1_{t,\delta}(o,a)$, where for $n\in\mathbb{N}$,

$$\beta_n(\delta) := \sqrt{\tfrac{2}{n}\big(1+\tfrac{1}{n}\big)\log\big(\sqrt{n+1}\tfrac{2^O-2}{\delta}\big)}.$$

By construction, it guarantees that uniformly for all t, $p\in C^1_{t,\delta/OA}$, with probability at least $1-\delta$, that is, $\mathbb{P}(\exists t\in\mathbb{N} : p\notin C^1_{t,\delta/OA})\leq\delta$.

The second confidence set is based on Bernstein's inequality (combined with a peeling technique) and is defined as follows (Maillard, 2019):

$$C^2_{t,\delta}(o,a,o') = \Bigg\{u\in[0,1] : |\widehat{p}_t(o'|o,a)-u| \leq \sqrt{\tfrac{2u(1-u)}{N_t(o,a)}\beta'_{N_t(o,a)}(\delta)} + \tfrac{\beta'_{N_t(o,a)}(\delta)}{3N_t(o,a)}\Bigg\},$$

and $C^2_{t,\delta} = \cap_{o,a,o'}C^2_{t,\delta}(o,a,o')$, where for $n\in\mathbb{N}$ and $\delta\in(0,1)$, $\beta'_n(\delta) := \eta\log\Big(\frac{\log(n+1)\log(n\eta)}{\delta\log^2(\eta)}\Big)$, where $\eta>1$ is an arbitrary choice. (We set $\eta=1.12$, as suggested by Maillard (2019), to get a small bound.) By construction, $C^2_{t,\delta}$ traps p with high probability, uniformly for all t: $\mathbb{P}(\exists t\in\mathbb{N} : p\notin C^2_{t,\delta/2O^2A})\leq\delta$.

[5] Our regret bounds can be extended straightforwardly to hold for regret defined as $\sum_{t=1}^{T}(r_t^\star - r_t)$, where $r_t^\star$ denotes the reward obtained by the oracle at time t. In fact, by applying Azuma-Hoeffding, the two notions of regret are related at the expense of an additive term $B\sqrt{T\log(T/\delta)}$ with B denoting the span of the optimal bias function in $M_{\rm cp}$, which always satisfies $B\leq D_{\rm cp}$; we refer to a more thorough discussion in (Talebi and Maillard, 2018).

[6] Assuming that $\mathcal{R}$ and M are both communicating is not sufficient to guarantee that M_{cp} is communicating. We demonstrate this using a simple (albeit pathological) example in Appendix E.

Confidence Sets for M_{cp}. We show that $C^1_{t,\delta}$ and $C^2_{t,\delta}$ yield confidence sets for the transition function P and reward function R of M_{cp}. To this effect, let us define the empirical estimates for P and R as follows. By a slight abuse of notation, let $\widehat{R}$ denote the empirical mean of distribution R, and let $\overline{\nu}$ denote the mean of the reward function $\mathbf{r} = \nu(q,\sigma)$.[7] For all $s=(q,o)$, $s'=(q',o')$, and a,

$$\widehat{P}_t\big(s'|s,a\big) = \widehat{p}_t(o'|o,a)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}},$$

$$\widehat{R}_t\big(s,a\big) = \sum_{o'} \widehat{p}_t(o'|o,a)\overline{\nu}\big(q,L(o,a,o')\big).$$

Now, the collection of all $p \in C^1_{t,\delta}$ (resp. $p \in C^2_{t,\delta}$) defines a confidence set for P (centered at $\widehat{P}_t$) and for R (centered at $\widehat{R}_t$) with similar probabilistic guarantees as for $C^1_{t,\delta}$ (resp. $C^2_{t,\delta}$). More concretely, we leverage this observation to introduce the following set of MDPRMs, which are plausible with the collected data up to time $t \ge 1$ and for a confidence parameter $\delta \in (0,1)$:

$$\mathcal{M}_{t,\delta} := \{M' = (\mathcal{O},\mathcal{A},p',\mathcal{R},\mathcal{P},L) : p' \in C\}\ ,$$

where $C = C^1_{t,\delta/OA}$ or $C = C^2_{t,\delta/2O^2A}$. This construction ensures that the true MDPRM M belongs to $\mathcal{M}_{t,\delta}$ with high probability, uniformly for all t. More precisely, for all $\delta \in (0,1)$, and for either choice of C,

$$\mathbb{P}(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}) \le \delta,$$

as formalized in Lemma 2 in Appendix C. This relies on the equivalence between any candidate MDPRM $M' \in \mathcal{M}_{t,\delta}$ and its associated cross-product MDP $M'_{\mathsf{cp}} = (\mathcal{S},\mathcal{A},P',R')$ where P' and R' are defined similarly to (1), but with the true p replaced by $p' \in C^1_{t,\delta/OA}$ or $p' \in C^2_{t,\delta/2O^2A}$.

3.2 From Confidence Sets to Algorithms: `UCRL-RM-L1` and `UCRL-RM-B`

We present an algorithm, called `UCRL-RM`, using the confidence sets presented above. We consider two variants of `UCRL-RM` depending on which confidence set is used: The variant using $C^1_{t,\delta}$, called `UCRL-RM-L1`, can be seen as an extension of UCRL2 (Jaksch et al., 2010) to MDPRMs. Whereas the one built using $C^2_{t,\delta}$, which we call `UCRL-RM-B`, extends UCRL2-style algorithms with Bernstein's confidence sets (in, e.g., (Bourel et al., 2020; Fruit et al., 2020, 2018b)) to MDPRMs. Both variants have a very similar design, and differ only in the choice of the confidence sets and an internal procedure used in the policy computation —however, they achieve different regret bounds and empirical performance. In the sequel, we shall use `UCRL-RM` to refer to both variants, but will make specific pointers to each when necessary.

`UCRL-RM` implements a form of the *optimism in the face of uncertainty* principle, but in an efficient manner for MDPRMs. Similarly to many model-based approaches developed based on this principle, they proceed in internal episodes (indexed by $k \in \mathbb{N}$) of varying lengths, where within each episode the policy is kept unchanged. Specifically, letting t_k denote the first step of episode k, `UCRL-RM` considers the set of plausible MDPs, $\mathcal{M}_{t_k,\delta}$, built using $C^1_{t,\delta/OA}$ (`UCRL-RM-L1`) or $C^2_{t,\delta/2O^2A}$ (`UCRL-RM-B`), and seeks a policy $\pi_k : \mathcal{S} \to \mathcal{A}$ that has the largest gain over all possible deterministic policies in all MDPRMs in $\mathcal{M}_{t_k,\delta}$. Practically, as in UCRL2, it suffices to find any $\frac{1}{\sqrt{t_k}}$-optimal solution to the following optimization problem:

$$\max_{M' \in \mathcal{M}_{t,\delta}, \pi \in \Pi_{M'}} g^\pi(M'),$$

where $g^\pi(M')$ denotes the gain of policy π in MDPRM M'. This optimization problem can be efficiently solved via a variant of the `EVI` algorithm of Jaksch et al. (2010). (Due to space constraints, we present it in Algorithm 2 in Appendix B.) In the case of MDPRMs, each iteration of `EVI` involves solving, for each (q,o,a):

$$\max_{z \in C(o,a)} \sum\nolimits_{(q',o') \in \mathcal{S}} \Big[\overline{\nu}\big(q,L(o,a,o')\big) + u(q',o')\Big] \\ \times \mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}} z(o'),$$

where u is the value function at the current iteration of `EVI`, and where $C(o,a) = C^1_{t,\delta/OA}(o,a)$ or $C(o,a) = \cap_{o'} C^2_{t,\delta/2O^2A}(o,a,o')$. `EVI` returns a policy π_k, which is guaranteed to be $\frac{1}{\sqrt{t_k}}$-optimal. `UCRL-RM` commits to π_k for $t \ge t_k$ until the number of observations on some pair (o,a) is doubled.[8] More precisely, the sequence $(t_k)_{k \ge 1}$ satisfies: $t_1 = 1$, and for $k \ge 1$,

$$t_k = \min\Big\{t > t_{k-1} : \max_{o,a} \frac{\sum_{t'=t_{k-1}}^{t} \mathbb{I}_{\{(o_{t'},a_{t'})=(o,a)\}}}{N_{t_{k-1}}(o,a)} \ge 1\Big\}.$$

The pseudo-code of `UCRL-RM` is presented in Algorithm 1. We recover `UCRL-RM-L1` (resp. `UCRL-RM-B`) if $\mathcal{M}_{t,\delta}$ is constructed using C^1 (resp. C^2). Both algorithms receive the RM $\mathcal{R}$ as well as a confidence parameter $\delta \in (0,1)$ as input. Despite their similar design, they achieve different performance both theoretically and empirically.

4 REGRET BOUNDS

In this section, we present finite-time regret bounds for the two variants of `UCRL-RM` that hold with high probability. Both regret bounds depend on a problem-dependent quantity that, just as the diameter in tabular MDPs, reflects a measure of connectivity in MDPRMs.

[7] We recall that ν is assumed known to the agent as mentioned earlier. This will be relaxed in Appendix B.2.

[8] This is quite similar to the stopping criterion in UCRL2.

Algorithm 1 UCRL-RM

```
Require: $\mathcal{O}, \mathcal{A}, \mathcal{R}, \delta$
  Initialize: For all $(o,a,o')$, set $N_0(o,a) = 0$, $N_0(o,a,o') = 0$ and $v_0(o,a) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state $s_1 = (q_1, o_1)$
  for episodes $k \geq 1$ do
    Set $t_k = t$
    Set $N_{t_k}(o,a) = N_{t_{k-1}}(o,a) + v_k(o,a)$ for all $(o,a)$
    Set $v_k(o,a) = 0$ for all $(o,a)$;
    Compute empirical estimates $\widehat{p}_{t_k}(\cdot|o,a)$ for all $(o,a)$
    Compute $\pi_k = \mathtt{EVI}\big(C, \frac{1}{\sqrt{t_k}}\big)$ —see Algorithm 2 in Appendix B.
      (Set $C = C^1_{t_k,\delta/OA}$ for UCRL-RM-L1, and $C = C^2_{t_k,\delta/O^2A}$ for UCRL-RM-B.)
    while $v_k(o_t, \pi_k(q_t,o_t)) < \max\{1, N_{t_k}(o_t, \pi_k(q_t,o_t))\}$ do
      Play action $a_t = \pi_k(q_t,o_t)$
      Receive next-state $o_{t+1} \sim p(\cdot|o_t,a_t)$
      Receive reward $r_t \sim \nu(q_t, L(o_t,a_t,o_{t+1}))$
      Set $N_{t+1}(o_t,a_t,o_{t+1}) = N_t(o_t,a_t,o_{t+1}) + 1$
      Set $v_k(o_t,a_t) = v_k(o_t,a_t) + 1$
      Set $t = t+1$
    end while
  end for
```

We begin with formalizing this notion of connectivity. For $s = (q,o) \in \mathcal{S}$, define

$$\mathcal{B}_s := \bigcup_{a,o'} \big\{q' \in \mathcal{Q} : q' = \tau\big(q, L(o,a,o')\big)\big\}.$$

Intuitively, for a given $s = (q,o)$, $\mathcal{B}_s \subseteq \mathcal{Q}$ collects all possible next-states of the RM that can be reached from q via the *detectable events* in o. In the worst-case $\mathcal{B}_s = \mathcal{Q}$ for some state $s \in \mathcal{S}$. However, many high-level tasks in practice often admit RMs with sparse structures, where for some s, $\mathcal{B}_s$ is a small subset of $\mathcal{Q}$. Using $\mathcal{B}_s$, we define a notion of *RM-restricted diameter* for s, which, as we shall see, proves relevant for MDPRMs:

Definition 1 (RM-Restricted Diameter) *Consider state $s = (q,o) \in \mathcal{S}$. For $s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}$ with $s_1 \neq s_2$, let $T^\pi(s_1,s_2)$ denote the number of steps it takes to get to s_2 starting from s_1 and following policy π. Then, we define the* RM-restricted diameter *of MDPRM M for state s as*

$$D_s := \max_{s_1,s_2 \in \mathcal{B}_s \times \mathcal{O}} \min_\pi \mathbb{E}[T^\pi(s_1,s_2)].$$

Replacing $\mathcal{B}_s$ with $\mathcal{Q}$ in Definition 1, one recovers D_{cp}, the diameter of M_{cp}. In view of $\mathcal{B}_s \subseteq \mathcal{Q}$, $D_s \leq D_{\mathsf{cp}}$ for all $s \in \mathcal{S}$. Since $\mathcal{B}_s$ could be a proper (and possibly cardinality-wise small) subset of $\mathcal{Q}$, D_s is therefore a problem-dependent refinement of D_{cp}. We remark that a small $\mathcal{B}_s$ does not necessarily imply that $D_s \ll D_{\mathsf{cp}}$ as D_s is determined by both $\mathcal{B}_s$ and the transition function P of M_{cp}. Interestingly, however, there exist cases where $D_s \lesssim D_{\mathsf{cp}}/Q$, as we illustrate below.

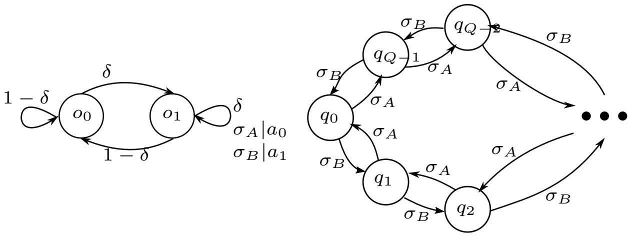


Figure 2: An example where RM-restricted diameter $D_s \lesssim D_{\mathsf{cp}}/Q$. The labeled MDP in left, and the RM in right.

Consider the MDPRM shown in Figure 2, where there are two observation states o_0 and o_1, with identical transition probabilities parameterized by $\delta \in (0, \frac{1}{2})$. In o_0, there is one action, but no event. In o_1, there are two actions: a_0 (that results in detecting σ_A) and a_1 (which leads to detecting σ_B). The RM has Q states arranged in a cycle, such that σ_A and σ_B yield transitions in the clockwise and counter-clockwise directions, respectively. As detailed in Appendix E, we can show that: For all $q \in \mathcal{Q}$, $D_{o_1,q} = \frac{2}{\delta} + 1 + \frac{\delta}{1-\delta}$ and $D_{o_0,q} = \frac{1}{\delta}$, whereas $D_{\mathsf{cp}} = \frac{\lfloor Q/2 \rfloor}{\delta} + 1 + \frac{\delta}{1-\delta}$. So, while D_{cp} grows as $\frac{Q}{\delta}$, D_s for all $s \in \mathcal{S}$ will be $\frac{1}{\delta}$. In summary, we have $D_s \lesssim D_{\mathsf{cp}}/Q$. Another example with numerically computed D_s is provided in Appendix F.

Regret Bounds. We are now ready to present the regret bounds. The following theorem provides a regret bound for UCRL-RM-L1, which was constructed using C^1:

Theorem 1 *Under UCRL-RM-L1, uniformly over all $T \geq 2$, with probability higher than $1 - 5\delta$,*

$$\mathfrak{R}(T) \leq O\Big(\sqrt{\mathbf{c}_M A T (O + \log(\sqrt{T}/\delta))} + D_{\mathsf{cp}}\sqrt{T\log(\sqrt{T}/\delta)}\Big)\,,$$

where $\mathbf{c}_M = \sum_{o\in\mathcal{O}} \max_{q\in\mathcal{Q}} D^2_{q,o}$.

To present a regret bound for UCRL-RM-B (constructed using C^2), for $(o,a) \in \mathcal{S}$ we let $K_{o,a}$ be the number of possible next-states in $\mathcal{O}$ under (o,a), that is, $K_{o,a} := |\{o' \in \mathcal{O} : p(o'|o,a) > 0\}|$.

Theorem 2 *Under UCRL-RM-B, uniformly over all $T \geq 2$, with probability higher than $1 - 5\delta$,*

$$\mathfrak{R}(T) \leq O\Big(\sqrt{\mathbf{c}'_M T \log(\log(T)/\delta)} + D_{\mathsf{cp}}\sqrt{T\log(\log(T)/\delta)}\Big),$$

where $\mathbf{c}'_M = \sum_{o\in\mathcal{O},a\in\mathcal{A}} K_{o,a} \max_{q\in\mathcal{Q}} D^2_{q,o}$.

The problem-dependent quantities $\mathbf{c}_M$ and $\mathbf{c}'_M$ reflect the (weighted) contribution of RM-restricted diameters to the regret. In the worst-case, $\mathbf{c}_M \leq OD^2_{\mathsf{cp}}$ and $\mathbf{c}'_M = D^2_{\mathsf{cp}} \sum_{o,a} K_{o,a}$, but in view of the example earlier, there are problem instances in which $\mathbf{c}_M \lesssim OD^2_{\mathsf{cp}}/Q^2$ and $\mathbf{c}'_M \lesssim D^2_{\mathsf{cp}}/Q^2 \sum_{o,a} K_{o,a}$.

Comparison with Tabular RL Algorithms for M_{cp}. Any algorithm available for tabular RL could be directly applied to M_{cp}, obliviously to the RM. In doing so, UCRL2 *(with improved confidence sets used here)* achieves a regret of $O(D_{\mathsf{cp}}\sqrt{AOQT(OQ+\log T)})$ whereas UCRL2-B achieves a regret of $O(D_{\mathsf{cp}}\sqrt{T\log(\log(T))Q\sum_{o,a}K_{o,a}}))$.[9] In comparison with these bounds, for moderate time-horizons T, we obtain an improvement in the regret bound by a multiplicative factor of at least Q, but in some examples this can be as large as Q^2. For large horizons (relative to O), the respective gains over UCRL2 are $\sqrt{Q}$ and $Q^{3/2}$. We also achieve a similar gain over UCRL2-B —we present a more detailed comparison in Appendix G.

5 REGRET LOWER BOUND

We also present a regret lower bound for learning MDPRMs. For communicating tabular MDPs with S states, A actions, and diameter D, a regret lower bound of $\Omega(\sqrt{DSAT})$ is presented by Jaksch et al. (2010), which relies on a carefully constructed family of worst-case MDPs. *However, this does not translate to a lower bound of $\Omega(\sqrt{D_{\mathsf{cp}}QOAT})$ for the cross-product* M_{cp} associated to a given MDPRM M. This is due to the fact that the transition function of the aforementioned worst-case MDPs does not satisfy (1). In other words, *there exist no MDPRMs* for which those worst-case MDPs become their associated cross-product MDPs. In the following theorem, we present a regret lower bound that holds for any MDPRM M with a communicating cross-product M_{cp}.

Theorem 3 *For any $O\geq 3$, $A\geq 2$, $Q\geq 2$, and $D_{\mathsf{cp}}\geq Q(6+2\log_A(O))$, $T\geq D_{\mathsf{cp}}OA$ and $|\mathcal{P}|\geq 2$, there exists a family of MDPRMs with O observations states, A actions, Q RM states, and diameter D_{cp} of the associated M_{cp}, in which the regret of any algorithm $\mathbb{A}$ satisfies*

$$\mathbb{E}[\mathfrak{R}(\mathbb{A},T)]\geq c_0\sqrt{D_{\mathsf{cp}}OAT},$$

where $c_0>0$ is a universal constant.

This theorem asserts a *worst-case* regret lower bound growing as $\Omega(\sqrt{D_{\mathsf{cp}}OAT})$ and is proven in Appendix D. To establish this result, we carefully construct an instance of MDPRM. In order to make it a worst-case instance, both p and $\mathcal{R}$ have to be chosen in a way to challenge exploration. To this end, we construct an RM with a non-trivial structure, whereas for p, we take inspiration from the worst-case MDPs presented by Jaksch et al. (2010), so that on the resulting MDPRM, the regret of any algorithm grows as $\Omega(\sqrt{D_{\mathsf{cp}}OAT})$ even when the RM and associated rewards are known to the learner. We finally remark that the lower bound does not contradict our regret bounds, as for the worst-case instances considered, $\max_q D_{q,o}\simeq D_{\mathsf{cp}}$.

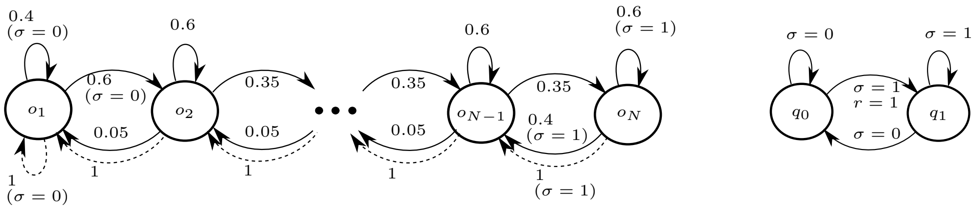
6 EXPERIMENTS

In this section, we present a set of experiments comparing the empirical performance of our algorithms with those of state-of-the-art baselines (applied to the cross-product MDP). As baselines, we consider UCRL2 (Jaksch et al., 2010), UCRL2B (Fruit et al., 2020), and TSDE (Ouyang et al., 2017), all provided with the knowledge of the reward function. To make the comparison fair, for UCRL2 and UCRL2B we used improved confidence sets defined similarly to C^1 and C^2, respectively. All codes are made publicly available at `https://github.com/HippolyteBourel/UCRL-RM`.

The presented experiments consider patrolling tasks, which are motivated by the tasks that consist of repetitively visiting multiple key locations in a given environment. Such tasks arise in many applications in transportation and robotics. Despite their concise definitions, they render challenging when exploration of an unknown environment is required. The considered MDPRMs are built using standard domains *RiverSwim* and gridworlds. The *RiverSwim* domain, shown in Figure 3, is combined with the *patrol2* RM that requires to patrol the two extreme locations in *RiverSwim* (i.e., o_1 and o_N) to output a fixed reward. We consider two variants of gridworlds: a 2-room with a task of patrolling 3 corners (Figure 5) and a 4-room with a patrol of 4 corners presented in Figure 1. In these gridworlds, the agent can perform 4 actions corresponding to going up, right, down, and left. Each action leads to moving the agent in the intended direction (w.p. 0.7), in each perpendicular direction (w.p. 0.1), or no move (w.p. 0.1). Any transition going into a wall has the effect of staying in place.

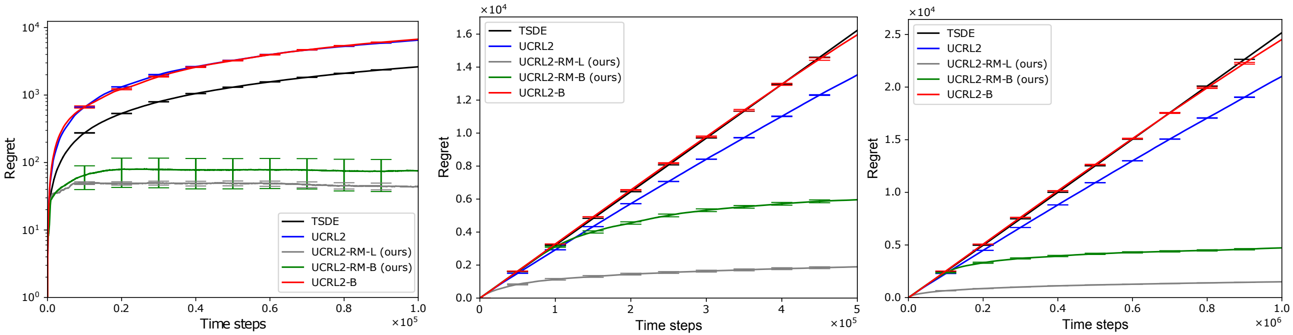
Figures 4(a)-4(c) show the regret over time together with 95% confidence intervals. Figure 4(a) depicts the results in a 6-state RiverSwim MDPRM, where all results are averaged over 200 runs. (Note the logarithmic y-axis.) Figure 4(b) shows the regret, averaged over 100 runs, for the MDPRM with the 2-room domain. Finally, Figure 4(c) presents the regret, averaged over 100 runs, in the MDPRM with the 4-room domain. As these figures reveal, both variants of `UCRL-RM` significantly outperform all the baselines. Furthermore, `UCRL-RM-L1` yields better performance than `UCRL-RM-B` in gridworlds. In view of the definition of regret, these results corroborate that, in terms of collected rewards, the benefit of exploiting the structure in MDPRMs could be significant. This is further verified by the corresponding empirical gain $\frac{1}{t}\sum_{t'=1}^{t}r_{t'}$ shown in Figure 6 for various algorithms in the 4-room MDPRM, together with 95% confidence intervals. The horizontal line (in magenta) shows the optimal gain $g^\star$ achieved by the oracle. In particu-

[9] A factor $\sqrt{D_{\mathsf{cp}}/\log(T)}$ can be shaved off the regret of UCRL2-B as reported by Fruit et al. (2020), and the same improvement may carry over to `UCRL-RM-B`. We exclude comparisons to EBF introduced by Zhang and Ji (2019) as it does not admit an efficient implementation.



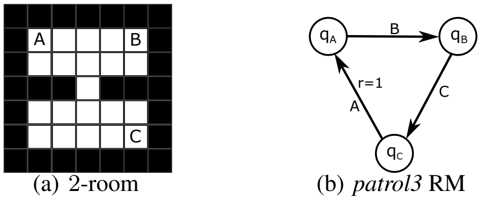
(a) Labeled *RiverSwim* MDP (b) The *Patrol2* RM

Figure 3: The N-observation labeled *RiverSwim* MDP (Strehl and Littman, 2008), and the *patrol2* RM



(a) *RiverSwim* ($N = 6$), logarithmic y-axis (b) 2-room (c) 4-room

Figure 4: Experimental results displaying the regret over time in various environments



(a) 2-room (b) *patrol3* RM

Figure 5: A 2-room gridworld with *patrol3* RM

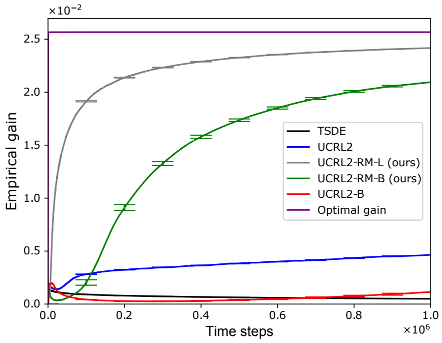


Figure 6: Empirical gain in the 4-room with *patrol4*

lar, Figure 6 indicates that the empirical gain (i.e., empirical per-step reward) under UCRL-RM-L1 quickly approaches $g^\star$ compared to the rest. (The corresponding figures for other environments are presented in Appendix F.)

Finally, we report the empirical running times of the algorithms in Table 3 in Appendix F. We remark that UCRL-RM-B is computationally more expensive than UCRL-RM-L1 (the same comparison holds between UCRL2 and UCRL2B in terms of involved computations). The additional cost arises from the instability of the convergence of EVI with element-wise confidence bounds. We chose to arbitrarily stop EVI after 100 iterations (with the exceptions of the results in Figure 4(a)) in order to impose a computational cost of the same order of magnitude for all algorithms. This has a slight negative impact on the results of both UCRL-RM-B and UCRL2B.

7 CONCLUSION

We studied reinforcement learning in average-reward Markov decision processes with reward machines (MDPRMs), in the regret minimization setting, under the assumption of a known reward machine (RM) but unknown dynamics. We introduced two algorithms tailored to leverage the structure of MDPRMs, and analysed their regret. Both algorithms significantly outperform existing baselines, both in theory and in practice. We also presented a regret lower bound for MDPRMs, establishing that the reported regret bounds are near-optimal. An interesting future work direction is to devise efficient algorithms for MDPRMs when the state of the RM is not observed. Another interesting future work is to consider RMs with stochastic transitions, where the resulting regret bounds may depend on Q. The more interesting, yet very challenging, question is to improve the lower bound (Theorem 3) to potentially make appear such a dependence on Q.

Acknowledgements

The authors would like to thank anonymous reviewers for their comments. Hippolyte Bourel and Mohammad Sadegh

Talebi are partially supported by the Independent Research Fund Denmark, grant number 1026-00397B. Anders Jonsson is partially supported by the Spanish grant PID2019-108141GB-I00 and the European project TAILOR (H2020, GA 952215). Odalric-Ambrym Maillard is supported by the French Ministry of Higher Education and Research, Inria, Scool, the Hauts-de-France region, the MEL and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF.

References

Eden Abadi and Ronen I Brafman. Learning and solving regular decision processes. In *International Joint Conference on Artificial Intelligence*, 2020.

Alessandro Abate, Yousif Almulla, James Fox, David Hyland, and Michael Wooldridge. Learning task automata for reinforcement learning using hidden Markov models. *arXiv preprint arXiv:2208.11838*, 2022.

Mahsa Asadi, Mohammad Sadegh Talebi, Hippolyte Bourel, and Odalric-Ambrym Maillard. Model-based reinforcement learning exploiting state-action equivalence. *arXiv preprint arXiv:1910.04077*, 2019.

Fahiem Bacchus, Craig Boutilier, and Adam Grove. Rewarding behaviors. In *National Conference on Artificial Intelligence*, 1996.

Peter L Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence*, 2009.

Hippolyte Bourel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, 2020.

Ronen I Brafman and Giuseppe De Giacomo. Regular decision processes: A model for non-Markovian domains. In *International Joint Conference on Artificial Intelligence*, 2019.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Mingyu Cai, Erfan Aasi, Calin Belta, and Cristian-Ioan Vasile. Overcoming exploration: Deep reinforcement learning in complex environments from temporal logic specifications. *arXiv preprint arXiv:2201.12231*, 2022.

Alberto Camacho, Rodrigo Toro Icarte, Toryn Q Klassen, Richard Anthony Valenzano, and Sheila A McIlraith. LTL and beyond: Formal languages for reward function specification in reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2019.

Alberto Camacho, Jacob Varley, Andy Zeng, Deepali Jain, Atil Iscen, and Dmitry Kalashnikov. Reward machines for vision-based robotic manipulation. In *International Conference on Robotics and Automation*, 2021.

Alexander Clark and Franck Thollard. PAC-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research*, 5(May):473–497, 2004.

Jan Corazza, Ivan Gavran, and Daniel Neider. Reinforcement learning with stochastic reward machines. In *AAAI Conference on Artificial Intelligence*, 2022.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, 2017.

Michael Dann, Yuan Yao, Natasha Alechina, Brian Logan, and John Thangarajah. Multi-agent intention progression with reward machines. In *International Joint Conference on Artificial Intelligence*, 2022.

Giuseppe De Giacomo, Marco Favorito, Luca Iocchi, and Fabio Patrizi. Imitation learning over heterogeneous agents with restraining bolts. In *International Conference on Automated Planning and Scheduling*, 2020.

Floris den Hengst, Vincent Francois-Lavet, Mark Hoogendoorn, and Frank van Harmelen. Reinforcement learning with option machines. In *International Joint Conference on Artificial Intelligence*, 2022.

Taylor Dohmen, Noah Topper, George Atia, Andre Beckus, Ashutosh Trivedi, and Alvaro Velasquez. Inferring probabilistic reward machines from non-Markovian reward signals for reinforcement learning. In *International Conference on Automated Planning and Scheduling*, 2022.

Pierre Dupont, François Denis, and Yann Esposito. Links between probabilistic automata and hidden Markov models: Probability distributions, learning models and induction algorithms. *Pattern Recognition*, 38(9):1349–1371, 2005.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems 31*, 2018a.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, 2018b.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of UCRL2 with empirical Bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Daniel Furelos-Blanco, Mark Law, Anders Jonsson, Krysia Broda, and Alessandra Russo. Induction and exploitation of subgoal automata for reinforcement learning. *Journal of Artificial Intelligence Research*, 70:1031–1116, 2021.

Daniel Furelos-Blanco, Mark Law, Anders Jonsson, Krysia Broda, and Alessandra Russo. Hierarchies of reward machines. *arXiv preprint arXiv:2205.15752*, 2022.

Maor Gaon and Ronen Brafman. Reinforcement learning with non-Markovian rewards. In *AAAI Conference on Artificial Intelligence*, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Nathaniel Hamilton, Preston K Robinette, and Taylor T Johnson. Training agents to satisfy timed and untimed signal temporal logic specifications with reinforcement learning. In *International Conference on Software Engineering and Formal Methods*, 2022.

Mohammadhosein Hasanbeig, Natasha Yogananda Jeppu, Alessandro Abate, Tom Melham, and Daniel Kroening. Deepsynth: Automata synthesis for automatic task segmentation in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2021.

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

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Milad Kazemi, Mateo Perez, Fabio Somenzi, Sadegh Soudjani, Ashutosh Trivedi, and Alvaro Velasquez. Translating omega-regular specifications to average objectives for model-free reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Marcus Hutter, and Peter Sunehag. The sample-complexity of general reinforcement learning. In *International Conference on Machine Learning*, 2013.

Michael L Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. Environment-independent task specifications via GLTL. *arXiv preprint arXiv:1704.04341*, 2017.

Odalric-Ambrym Maillard. Mathematics of statistical sequential decision making. *Habilitation à Diriger des Recherches*, 2019.

Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International Conference on Machine Learning*, 2013.

Ya-Wen Mo, ChiKai Ho, and Chung-Ta King. Managing shaping complexity in reinforcement learning with state machines - using robotic tasks with unspecified repetition as an example. In *International Conference on Mechatronics and Automation*, 2022.

Cyrus Neary, Zhe Xu, Bo Wu, and Ufuk Topcu. Reward machines for cooperative multi-agent reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 2021.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson Sampling approach. In *Advances in Neural Information Processing Systems 30*, 2017.

Fabien Pesquerel and Odalric-Ambrym Maillard. IMED-RL: Regret optimal learning of ergodic Markov decision processes. In *Advances in Neural Information Processing Systems 35*, 2022.

Martin L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Jian QIAN, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward MDPs. In *Advances in Neural Information Processing Systems 32*, 2019.

Alessandro Ronca and Giuseppe De Giacomo. Efficient PAC reinforcement learning in regular decision processes. In *International Joint Conference on Artificial Intelligence*, 2021.

Raeid Saqur. Reward learning using structural motifs in inverse reinforcement learning. *arXiv preprint arXiv:2209.13489*, 2022.

Ankit Shah, Samir Wadhwania, and Julie Shah. Interactive robot training for non-Markov tasks. *arXiv preprint arXiv:2003.02232*, 2020.

Edward Jay Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, 1971.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74 (8):1309–1331, 2008.

Peter Sunehag and Marcus Hutter. Rationality, optimism and guarantees in general reinforcement learning. *Journal of Machine Learning Research*, 16:1345–1390, 2015.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *International Conference on Algorithmic Learning Theory*, 2018.

Geraud Nangue Tasse, Devon Jarvis, Steven James, and Benjamin Rosman. Skill machines: Temporal logic composition in reinforcement learning. *arXiv preprint arXiv:2205.12532*, 2022.

Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, 2018.

Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. Learning reward machines for partially observable reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical Bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

Christos Verginis, Cevahir Koprulu, Sandeep Chinchali, and Ufuk Topcu. Joint learning of reward machines and policies in environments with partially known semantics. *arXiv preprint arXiv:2204.11833*, 2022.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, 2021.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Technical Report*, 2003.

Duo Xu and Faramarz Fekri. Integrating symbolic planning and reinforcement learning for following temporal logic specifications. In *International Joint Conference on Neural Networks*, 2022.

Zhe Xu, Ivan Gavran, Yousef Ahmad, Rupak Majumdar, Daniel Neider, Ufuk Topcu, and Bo Wu. Joint inference of reward machines and policies for reinforcement learning. In *International Conference on Automated Planning and Scheduling*, 2020.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems 32*, pages 2823–2832, 2019.

Xuejing Zheng, Chao Yu, and Minjie Zhang. Lifelong reinforcement learning with temporal logic formulas and reward machines. *Knowledge-Based Systems*, 257:109650, 2022.

Appendix

Table of Contents

A THE CROSS-PRODUCT MDP: PROOF OF LEMMA 1 **14**

B FURTHER ALGORITHMIC DETAILS **14**

B.1 Extended Value Iteration for MDPRMs 14

B.2 Unknown Mean Rewards 16

C REGRET ANALYSIS OF `UCRL-RM` **16**

C.1 Proof of Theorem 1 17

C.2 Proof of Theorem 2 21

C.3 Technical Lemmas 23

C.4 Concentration Inequalities 23

D REGRET LOWER BOUND **25**

E DETAILS OF DIAMETER COMPUTATIONS **28**

F DETAILS OF EXPERIMENTS AND FURTHER EXPERIMENTS **29**

G COMPARISON OF REGRET BOUNDS **30**

A THE CROSS-PRODUCT MDP: PROOF OF LEMMA 1

Lemma 1 (restated) *Let $M = (\mathcal{O}, \mathcal{A}, p, \mathcal{R}, \mathcal{P}, L)$ be a finite MDPRM. Then, an associated cross-product MDP to M is $M_{\textbf{\textit{cp}}} = (\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S} = \mathcal{Q} \times \mathcal{O}$, and where for $s = (q, o)$, $s' = (q', o') \in \mathcal{S}$, and $a \in \mathcal{A}$,*

$$P(s'|s, a) = p(o'|o, a)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}} , \quad R(s, a) = \sum_{o'\in\mathcal{O}} p(o'|o, a)\nu(q, L(o, a, o')).$$

Proof. Let $M = (\mathcal{O}, \mathcal{A}, p, \mathcal{R}, \mathcal{P}, L)$ and $\mathcal{S} = \mathcal{Q} \times \mathcal{O}$. For any $t \in \mathbb{N}$, let $h_t := (s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t)$, where $s_{t'} := (q_{t'}, o_{t'})$. We show that for any $h \in (\mathcal{S} \times \mathcal{A})^{t-1} \times \mathcal{S}$, $s' = (q', o') \in \mathcal{S}$, $a \in \mathcal{A}$, and $B \subseteq [0, 1]$:

$$\begin{aligned}
\mathbb{P}(s_{t+1} = s'|h_t = h, a_t = a) &= p(o'|o, a)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}} , \\
\mathbb{P}(r_t \in B|h_t = h, a_t = a) &= \sum_{o'\in\mathcal{O}} p(o'|o, a)\nu\big(q, L(o, a, o')\big)(B) ,
\end{aligned}$$

thus implying that the state and reward dynamics are fully determined by (s_t, a_t). For any $(q', o') \in \mathcal{S}$, we have

$$\begin{aligned}
\mathbb{P}\Big(s_{t+1} = (q', o')\Big|h_t = h, a_t = a\Big) &= \mathbb{P}\Big(o_{t+1} = o'\Big|h_t = h, a_t = a\Big)\mathbb{P}\Big(q_{t+1} = q'\Big|h_t = h, o_{t+1} = o', a_t = a\Big) \\
&= \mathbb{P}\Big(o_{t+1} = o'\Big|o_t = o, a_t = a\Big)\mathbb{P}\Big(q_{t+1} = q'\Big|s_t = (q, o), o_{t+1} = o', a_t = a\Big) \\
&= p(o'|o, a)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}} ,
\end{aligned}$$

where the second line follows from the fact that observation dynamics are Markovian and from the definition of RMs.

Moreover, for any set $B \subseteq [0, 1]$, we have

$$\begin{aligned}
\mathbb{P}\Big(r_t \in B\Big|h_t = h, a_t = a\Big) &= \sum_{o'\in\mathcal{O}} \mathbb{P}\Big(o_{t+1} = o'\Big|h_t = h, a_t = a\Big)\mathbb{P}\Big(r_t = r\Big|h_t = h, o_{t+1} = o', a_t = a\Big) \\
&= \sum_{o'\in\mathcal{O}} p(o'|o, a)\mathbb{P}\Big(r_t \in B\Big|s_t = (q, o), o_{t+1} = o', a_t = a\Big) \\
&= \sum_{o'\in\mathcal{O}} p(o'|o, a)\nu\big(q, L(o, a, o')\big)(B) ,
\end{aligned}$$

thus verifying the two claims. Now letting P and R be defined as in the lemma, we have that $(\mathcal{S}, \mathcal{A}, P, R)$ constitutes an MDP. □

B FURTHER ALGORITHMIC DETAILS

B.1 Extended Value Iteration for MDPRMs

We present the complete specification of Extended Value Iteration (`EVI`) used as a subroutine in `UCRL-RM`. Algorithm 2 presents the pseudo-code of `EVI`, which closely follows the design of EVI of Jaksch et al. (2010).

`EVI` relies on solving the following maximization problem in each round, and for any (q, o, a):

$$\max_{z\in C(o,a)} \sum_{(q',o')\in\mathcal{S}} \Big[\overline{\nu}\big(q, L(o, a, o')\big) + u(q', o')\Big]\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}} z(o'), \tag{3}$$

where u is the value function at the current iteration of `EVI`, and where $C(o, a) = C^1_{t,\delta/OA}(o, a)$ or $C(o, a) = \cap_{o'} C^2_{t,\delta/2O^2A}(o, a, o')$. Algorithm 3 finds a solution to problem (3) for $C = C^1$ (i.e., for `UCRL-RM-L1`), whereas Algorithm 4 does so for $C = C^2$ (i.e., for `UCRL-RM-B`). Algorithm 3 is quite similar to the one used in UCRL2 (Jaksch et al., 2010), whereas Algorithm 4 is used in UCRL2B and similar (e.g., in (Dann and Brunskill, 2015)).

For $(q, o, a) \in \mathcal{Q} \times \mathcal{O} \times \mathcal{A}$, let $\widetilde{p}^q(\cdot|o, a)$ be any optimal solution to problem (3) —namely, we denote the optimal z by $\widetilde{p}^q(\cdot|o, a)$. Here, the superscript q in $\widetilde{p}^q(\cdot|s, a)$ signifies that the optimistic transition probability depends on q.

Algorithm 2 EVI(C, ε)

```
Initialize: $u^{(0)} \equiv 0, u^{(-1)} \equiv -\infty, n = 0$
while $\max_{s \in \mathcal{S}}(u^{(n)} - u^{(n-1)})(s) - \min_{s \in \mathcal{S}}(u^{(n)} - u^{(n-1)})(s) > \varepsilon$ do
    Get $\widetilde{p}^q$ for all $q \in \mathcal{Q}$ using MAXP-L1 (Algorithm 3, for UCRL-RM-L1) or MAXP-B (Algorithm 4, for UCRL-RM-B)
    For all $(q, o) \in \mathcal{S}$, update:
        $$u^{(n+1)}(q, o) = \max_{a \in \mathcal{A}} \sum_{(q', o') \in \mathcal{S}} \widetilde{p}^q(o'|o, a)\Big[\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}(q', o')\Big]\mathbb{I}_{\{q' = \tau(q, L(o, a, o'))\}}$$
    Set $n = n + 1$
end while
Output:
    $$\pi_{n+1}(s) = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{(q', o') \in \mathcal{S}} \widetilde{p}^q(o'|o, a)\Big[\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}(q', o')\Big]\mathbb{I}_{\{q' = \tau(q, L(o, a, o'))\}}$$
```

Algorithm 3 MAXP-L1

```
For all $o' \in \mathcal{O}$, set $p(o') = \widehat{p}(o'|o, a)$
$o_{\max} = \operatorname{argmax}_{o' \in \mathcal{O}} \Big\{\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}\big(\tau(q, L(o, a, o')), o'\big)\Big\}$
$p(o_{\max}) = \max\big\{1,\, p(o_{\max}) + \frac{1}{2}\beta_{N_t(o,a)}(\frac{\delta}{OA})\big\}$
$\mathcal{L} = \operatorname{argsort}_{o'} \Big\{\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}\big(\tau(q, L(o, a, o')), o'\big),\, o' \in \mathcal{O}\Big\}$
$\ell = 0$
while $\sum_{o' \in \mathcal{O}} p(o') > 1$ do
    $p(\mathcal{L}_\ell) = \max\Big\{0,\, p(\mathcal{L}_\ell) + 1 - \sum_{o' \in \mathcal{O}} p(o')\Big\}$
    Set $\ell = \ell + 1$
end while
Output: $\widetilde{p}^q(\cdot|o, a) = p$
```

Algorithm 4 MAXP-B

```
For all $o' \in \mathcal{O}$, set $p(o') = \min\big\{p' \in C^2_{t, \delta/2O^2A}(o, a, o')\big\}$
$\mathcal{L} = \operatorname{argsort}_{o'} \Big\{\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}\big(\tau(q, L(o, a, o')), o'\big),\, o' \in \mathcal{O}\Big\}$
$\ell = OA - 1$
while $\sum_{o' \in \mathcal{O}} p(o') < 1$ do
    Set
        $$p(\mathcal{L}_\ell) = \min\bigg\{\max\Big\{z \in C^2_{t, \delta/2O^2A}\big(o, a, \mathcal{L}_\ell\big)\Big\}, 1 - \sum_{o' \in \mathcal{O}} p(o')\bigg\}$$
    $\ell = \ell - 1$
end while
Output: $\widetilde{p}^q(\cdot|o, a) = p$
```

B.2 Unknown Mean Rewards

Now we discuss the case of unknown mean rewards, i.e., when the agent has no prior knowledge about $\overline{\nu}$. To accommodate this situation, the agent maintains confidence sets for the various mean rewards as follows. For $(q,\sigma) \in \mathcal{Q} \times 2^{\mathcal{P}}$, define

$$C^{\text{reward}}_{t,\delta}(q,\sigma) = \Big\{\lambda \in [0,1] : \big|\widehat{\nu}_t(q,\sigma) - \lambda\big| \le \beta_{N_t(q,\sigma)}(\delta)\Big\}\,, \quad C^{\text{reward}}_{t,\delta} = \cap_{q,\sigma} C^{\text{reward}}_{t,\delta}(q,\sigma),$$

where $\widehat{\nu}_t(q,\sigma)$ denotes the empirical mean reward built using $N_t(q,\sigma)$ observations collected from the reward distribution $\nu(q,\sigma)$. Here, for $n \in \mathbb{N}$, $\beta_n(\delta) = \sqrt{\frac{1}{2n}\big(1+\frac{1}{n}\big)\log\big(\sqrt{n+1}/\delta\big)}$. Then, it suffices to replace $\overline{\nu}$ with its upper confidence set, that is, to replace $\overline{\nu}(q,\sigma)$, in problem (3), with

$$\widehat{\nu}_t(q,\sigma) + \beta_{N_t(q,\sigma)}\big(\tfrac{\delta}{Q|2^{\mathcal{P}}|}\big)\,.$$

The penalty due to this is an increase in the regret bound by an additive term that is independent of any diameter-like quantity (i.e., D_{cp} or $D_{q,o}$). The regret bound will depend on $\sqrt{\log Q}$, which will however be dominated by existing $\sqrt{\log T}$ terms.

C REGRET ANALYSIS OF `UCRL-RM`

In this section, we provide regret analyses of the two variants of `UCRL-RM`.

We first present a lemma, which formally states that the set of plausible MDPRMs maintained by `UCRL-RM-L1` and `UCRL-RM-B` contain the true MDPRM with high probability and uniformly over time:

Lemma 2 *For all $\delta \in (0,1)$, we have:*

$$\begin{aligned} &(i) \qquad \mathbb{P}\big(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}(C^1)\big) \le \delta\,, \\ &(ii) \qquad \mathbb{P}\big(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}(C^2)\big) \le \delta\,, \end{aligned}$$

where $\mathcal{M}_{t,\delta}(C^1)$ and $\mathcal{M}_{t,\delta}(C^1)$ denote the set of MDPRMs built using $C^1_{t,\delta/OA}$ and $C^2_{t,\delta/2O^2A}$, respectively.

A proof of Lemma 2, presented below, builds on the concentration inequalities collected in Section C.4.

Proof (of Lemma 2). Part (i). Using Lemma 8, we have for any (o,a),

$$\mathbb{P}\Big(\exists t \in \mathbb{N},\, p(\cdot|o,a) \notin C^1_{t,\delta/OA}(o,a)\Big) \le \frac{\delta}{OA}\,.$$

For $\mathcal{M}_{t,\delta}(C^1)$, we thus have

$$\begin{aligned} \mathbb{P}\big(\exists t \in \mathbb{N},\, M \notin \mathcal{M}_{t,\delta}(C^1)\big) &= \mathbb{P}\Big(\exists t \in \mathbb{N}, \exists p \notin C^1_{t,\delta/OA}\Big) \\ &= \mathbb{P}\Big(\exists t \in \mathbb{N}, \exists (o,a) \in \mathcal{O} \times \mathcal{A}, p(\cdot|o,a) \notin C^1_{t,\delta/OA}(o,a)\Big) \\ &\le \sum_{o\in\mathcal{O}, a\in\mathcal{A}} \mathbb{P}\Big(\exists t \in \mathbb{N}, p(\cdot|o,a) \notin C^1_{t,\delta/OA}(o,a)\Big) \\ &\le \sum_{o\in\mathcal{O}, a\in\mathcal{A}} \frac{\delta}{OA} = \delta\,. \end{aligned}$$

Part (ii). Lemma 9 ensures that for any (o,a,o'),

$$\mathbb{P}\Big(\exists t \in \mathbb{N},\, p(o'|o,a) \notin C^2_{t,\delta/2O^2A}(o,a,o')\Big) \le \frac{\delta}{O^2A}\,.$$

Hence, for $\mathcal{M}_{t,\delta}(C^2)$, we have

$$\begin{aligned}
\mathbb{P}\big(\exists t \in \mathbb{N},\, M \notin \mathcal{M}_{t,\delta}(C^2)\big) &= \mathbb{P}\Big(\exists t \in \mathbb{N}, \exists p \notin C^2_{t,\delta/2O^2A}\Big) \\
&= \mathbb{P}\Big(\exists t \in \mathbb{N}, \exists (o,a,o') \in \mathcal{O}\times\mathcal{A}\times\mathcal{O}, p(o'|o,a) \notin C^2_{t,\delta/2O^2A}(o,a,o')\Big) \\
&\le \sum_{o,o'\in\mathcal{O}, a\in\mathcal{A}} \mathbb{P}\Big(\exists t\in\mathbb{N}, p(o'|o,a) \notin C^2_{t,\delta/2O^2A}(o,a,o')\Big) \\
&\le \sum_{o,o'\in\mathcal{O}, a\in\mathcal{A}} \frac{\delta}{O^2A} = \delta\,.
\end{aligned}$$

□

C.1 Proof of Theorem 1

As in most regret analyses for model-based algorithms that work based on the optimism principle, the proof builds on the regret analysis by Jaksch et al. (2010), but it includes novel steps due to the structure of MDPRMs.

Let $\delta \in (0,1)$. We closely follow the notations used by Jaksch et al. (2010). To simplify notations, we define the short-hand $J_k := J_{t_k}$ for various random variables that are fixed within a given episode k and omit their dependence on δ (for example $\mathcal{M}_k := \mathcal{M}_{t_k,\delta}$). We let $m(T)$ denote the number of episodes initiated by the algorithm up to time T.

Observe that $\mathbb{E}[r_t|s_t,a_t] = \sum_{o'} p(o'|o_t,a_t)\overline{\nu}(q_t, L(o_t,a_t,o'))$. Hence, by applying Corollary 1, we deduce that

$$\begin{aligned}
\mathfrak{R}(T) &= \sum_{t=1}^T g^\star - \sum_{t=1}^T r_t \\
&\le \sum_{t=1}^T \sum_{o,q,a}\Big(g^\star - \sum_{o'} p(o'|o,a)\overline{\nu}(q,L(o,a,o'))\Big)\mathbb{I}_{\{(q_t,o_t,a_t)=(q,o,a)\}} + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} \\
&= \sum_{o,q,a}\Big(g^\star - \sum_{o'} p(o'|o,a)\overline{\nu}(q,L(o,a,o'))\Big)N_{m(T)}(q,o,a) + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,
\end{aligned}$$

with probability at least $1-\delta$.

For $s=(q,o)$, define

$$\mu(s,a) := \sum_{o'} p(o'|o,a)\overline{\nu}(q,L(o,a,o')) = \sum_{q',o'} p(o'|o,a)\overline{\nu}(q,L(o,a,o'))\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}}$$

Hence, the first term in the previous inequality reads

$$\begin{aligned}
\sum_{s,a}(g^\star - \mu(s,a))N_{m(T)}(s,a) &= \sum_{k=1}^{m(T)}\sum_{s,a}\underbrace{\sum_{t=t_k}^{t_{k+1}-1}\mathbb{I}_{\{s_t=s,a_t=a\}}}_{:=v_k(s,a)}\big(g^\star - \mu(s,a)\big) \\
&= \sum_{k=1}^{m(T)}\sum_{s,a} v_k(s,a)\big(g^\star - \mu(s,a)\big)\,.
\end{aligned}$$

Introducing $\Delta_k := \sum_{s,a} v_k(s,a)\big(g^\star - \mu(s,a)\big)$ for $1 \le k \le m(T)$, we get

$$\mathfrak{R}(T) \le \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,$$

with probability at least $1-\delta$.

A given episode k is called *good* if $M \in \mathcal{M}_k$, and *bad* otherwise.

Control of the regret due to bad episodes. By Lemma 2, the set $\mathcal{M}_k$ contains the true MDPRM with probability higher than $1-\delta$ uniformly for all T, and for all episodes $k = 1, \dots, m(T)$. As a consequence, with probability at least $1-\delta$, $\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M \notin \mathcal{M}_k\}} = 0$.

Control of the regret due to good episodes. To upper bound regret in good episodes, we closely follow (Jaksch et al., 2010) and decompose the regret to control the transition and reward functions. Consider a good episode k. Since $M \in \mathcal{M}_k$, the choice of π_k and $\widetilde{M}_k = (\mathcal{S}, \mathcal{A}, \widetilde{p}, \mathcal{R}, \mathcal{P}, L)$ satisfy $g_k := g^{\pi_k}(\widetilde{M}_k) \geq g^\star - \frac{1}{\sqrt{t_k}}$. Hence, the regret accumulated in episode k satisfies:

$$\Delta_k \leq \sum_{s,a} v_k(s,a)\big(g_k - \mu(s,a)\big) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \,. \tag{4}$$

It is a direct consequence of (Puterman, 2014, Theorem 8.5.6) that when the convergence criterion holds at iterate i in `EVI`, then

$$|u_k^{(i+1)}(s) - u_k^{(i)}(s) - g_k| \leq \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,. \tag{5}$$

By the design of `EVI`, note that for all $s \in \mathcal{S}$,

$$u_k^{(i+1)}(s) = \sum_{s' \in \mathcal{S}} \widetilde{p}_k^q(o'|o, \pi_k(s))\Big[\overline{\nu}(q, L(o, \pi_k(s), o')) + u_k^{(i)}(s')\Big]\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}\,,$$

where we recall that $\widetilde{p}_k^q(\cdot|o, \pi_k(q,o))$ is the transition probability distribution of the optimistic MDPRM $\widetilde{M}_k$ in $s = (q,o)$. For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, define

$$\widetilde{\mu}_k(s,a) := \sum_{q',o'} \widetilde{p}_k^q(o'|o,a)\overline{\nu}(q, L(o,a,o'))\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}\,.$$

Then, (5) gives, for all $s \in \mathcal{S}$,

$$\Big|g_k - \widetilde{\mu}_k(s, \pi_k(s)) - \Big(\sum_{s'} \widetilde{p}_k^q(o'|o,\pi_k(s))\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}} u_k^{(i)}(s') - u_k^{(i)}(s)\Big)\Big| \leq \frac{1}{\sqrt{t_k}}\,.$$

Defining $\mathbf{g}_k = g_k\mathbf{1}$, $\widetilde{\boldsymbol{\mu}}_k := \big(\widetilde{\mu}_k(s,\pi_k(s))\big)_s$, $\widetilde{\mathbf{P}}_k := \Big(\widetilde{p}_k^q\big(o'|o,\pi_k(s)\big)\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}\Big)_{s,s'}$ and $v_k := \big(v_k\big(s,\pi_k(s)\big)\big)_s$, we can rewrite the above inequality as:

$$\Big|\mathbf{g}_k - \widetilde{\boldsymbol{\mu}}_k - (\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)}\Big| \leq \frac{1}{\sqrt{t_k}}\mathbf{1}\,.$$

Also, we can rewrite (4) as

$$\Delta_k \leq \sum_{s,a} v_k(s,a)\big(g_k - \widetilde{\mu}_k(s,a)\big) + \sum_{s,a} v_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}\,. \tag{6}$$

The first term in the right-hand side of (6) is bounded by $v_k(\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)} + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}$. The second term is bounded as follows:

$$\begin{aligned}
\sum_{s,a} v_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) &= \sum_{s,a} v_k(s,a) \sum_{o' \in \mathcal{O}} \big(\widetilde{p}_k^q(o'|o,a) - p(o'|o,a)\big)\overline{\nu}(q, L(o,a,o')) \underbrace{\sum_{q'} \mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}}_{=1} \\
&\leq \sum_{s,a} v_k(s,a)\big\|\widetilde{p}_k^q(\cdot|o,a) - p(\cdot|o,a)\big\|_1 \\
&\leq 2\sum_{s,a} v_k(s,a)\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) \\
&= 2\sum_{o,a} \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) \underbrace{\sum_{q} v_k(q,o,a)}_{=v_k(o,a)} \\
&= 2\sum_{o,a} v_k(o,a)\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big)\,,
\end{aligned}$$

where we used that $\overline{\nu}(q, L(o,a,o')) \le 1$. Moreover, since $t_k \ge \max_{o,a} N_k(o,a)$, we can bound the third term in the right-hand side of (6) as:

$$\sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \le \sum_{o,a} \frac{1}{\sqrt{N_k(o,a)}} \sum_q v_k(q,o,a) = \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \,.$$

Putting these three bounds together, we thus get

$$\Delta_k \le v_k(\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)} + 2\sum_{o,a} v_k(o,a)\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) + 2\sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \,.$$

Let us define, for all $s \in \mathcal{S}$,

$$w_k(s) := u_k^{(i)}(s) - \frac{1}{2}\Big(\min_{s' \in \mathcal{B}_s \times \mathcal{O}} u_k^{(i)}(s') + \max_{s' \in \mathcal{B}_s \times \mathcal{O}} u_k^{(i)}(s')\Big) \,.$$

In view of the fact that $\widetilde{\mathbf{P}}_k$ is row-stochastic (i.e., its rows sum to one), we obtain

$$\Delta_k \le v_k(\mathbf{P}_k - \mathbf{I})w_k + \underbrace{v_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k}_{L_1} + 2\sum_{o,a} v_k(o,a)\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) + 2\sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \,. \tag{7}$$

Upper bound on L_1. We have

$$\begin{aligned}
v_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k &= \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} v_k(s,\pi_k(s))\Big(\widetilde{P}_k(s'|s,\pi_k(s)) - P(s'|s,\pi_k(s))\Big)w_k(s') \\
&= \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s)) \sum_{o'\in\mathcal{O}} \sum_{q'\in\mathcal{Q}} \Big(\widetilde{p}_k^q(o'|o,\pi_k(s)) - p(o'|o,\pi_k(s))\Big)\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}} w_k(q',o') \\
&\le \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s)) \sum_{o'\in\mathcal{O}} \Big|\widetilde{p}_k^q(o'|o,\pi_k(s)) - p(o'|o,\pi_k(s))\Big| \cdot \max_{s'\in\mathcal{B}_{q,o}\times\mathcal{O}} \big|w_k(q',o')\big| \underbrace{\sum_{q'\in\mathcal{Q}} \mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}}_{=1}
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\le \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s))\big\|\big(\widetilde{p}_k^q - p\big)(\cdot|o,\pi_k(s))\big\|_1 \cdot \max_{s'\in\mathcal{B}_{q,o}\times\mathcal{O}} \big|w_k(q',o')\big| \\
&\le \sum_{o\in\mathcal{O}} \sum_{q\in\mathcal{Q}} v_k(q,o,\pi_k(q,o))\beta_{N_k(o,\pi_k(q,o))}\big(\tfrac{\delta}{OA}\big) \cdot D_{q,o} && (9) \\
&\le \sum_{a\in\mathcal{A}} \sum_{o\in\mathcal{O}} \sum_{q\in\mathcal{Q}} v_k(q,o,a) \cdot \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) \cdot D_{q,o} \\
&\le \sum_{a\in\mathcal{A}} \sum_{o\in\mathcal{O}} \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) \cdot \max_{q\in\mathcal{Q}} D_{q,o} \sum_{q\in\mathcal{Q}} v_k(q,o,a) \\
&\le \sum_{a\in\mathcal{A}} \sum_{o\in\mathcal{O}} \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big) \cdot \max_{q\in\mathcal{Q}} D_{q,o} \cdot v_k(o,a) \,, && (10)
\end{aligned}$$

where (8) we used the definition of $\mathcal{B}_s$, and where (9) follows from Lemma 3, stated and proven in Section C.3. Combining (10) with (7) and summing over all good episodes, we obtain:

$$\begin{aligned}
\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M\in\mathcal{M}_k\}} \le \sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I}) w_k \mathbb{I}_{\{M\in\mathcal{M}_k\}} + 2\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
+ \sum_{k=1}^{m(T)} \sum_{o,a} v_k(o,a) \beta_{N_k(o,a)}\left(\tfrac{\delta}{OA}\right)\Big(2 + \max_{q\in\mathcal{Q}} D_{q,o}\Big) \\
= \sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I}) w_k \mathbb{I}_{\{M\in\mathcal{M}_k\}} + 2\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
+ \sum_{k=1}^{m(T)} \sum_{o,a} v_k(o,a) \sqrt{\frac{2}{N_k(o,a)}\Big(1+\frac{1}{N_k(o,a)}\Big) \log\Big(\tfrac{OA(2^O-2)}{\delta}\sqrt{N_k(o,a)+1}\Big)}\Big(2 + \max_{q\in\mathcal{Q}} D_{q,o}\Big) \\
= \underbrace{\sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I}) w_k \mathbb{I}_{\{M\in\mathcal{M}_k\}}}_{L_2} + 2\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
+ 2\sqrt{O + \log\big(OA\sqrt{T+1}/\delta\big)} \sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \Big(\max_{q\in\mathcal{Q}} D_{q,o} + 2\Big)\,, \qquad (11)
\end{aligned}$$

where we use the trivial inequality $1 \le N_k(o,a) \le T$.

Upper bound on L_2. To upper bound L_2, we define a martingale difference sequence similarly to the proof of Theorem 2 in (Jaksch et al., 2010). Let $(Z_t)_{t\ge1}$ be a sequence with

$$Z_t := (P(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}}) w_{k(t)} \mathbb{I}_{\{M\in\mathcal{M}_{k(t)}\}},$$

for all t, where $k(t)$ denotes the episode containing time step t. For any good episode k, we have:

$$\begin{aligned}
v_k(\mathbf{P}_k - \mathbf{I}) w_k &= \sum_{t=t_k}^{t_{k+1}-1} (P(\cdot|s_t,a_t) - \mathbf{e}_{s_t}) w_k \\
&= \sum_{t=t_k}^{t_{k+1}-1} \Big(P(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}} + \mathbf{e}_{s_{t+1}} - \mathbf{e}_{s_t}\Big) w_k \\
&= \sum_{t=t_k}^{t_{k+1}-1} Z_t + w_k(s_{t_{k+1}}) - w_k(s_{t_k}) \le \sum_{t=t_k}^{t_{k+1}-1} Z_t + D_{\mathsf{cp}}\,,
\end{aligned}$$

where $\mathbf{e}_i$ denotes a vector with the i-th element being 1 and the others being zero. Hence, $L_2 \le \sum_{t=1}^{T} Z_t + m(T) D_{\mathsf{cp}}$. As established in (Jaksch et al., 2010), $|Z_t| \le \|P(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}}\|_1 \|w_{k(t)}\|_\infty \le D_{\mathsf{cp}}$ and $\mathbb{E}[Z_t|s_1,a_1,\dots,s_t,a_t] = 0$, so that $(Z_t)_{t\ge1}$ is martingale difference sequence. Therefore, by Corollary 1, we get:

$$\mathbb{P}\bigg(\exists T : \sum_{t=1}^{T} Z_t \ge D_{\mathsf{cp}}\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\bigg) \le \delta\,.$$

Thus, for all T, with probability at least $1-\delta$, it holds

$$\begin{aligned}
L_2 &\le D_{\mathsf{cp}}\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + m(T) D_{\mathsf{cp}} \\
&\le D_{\mathsf{cp}}\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}} OA \log_2\left(\tfrac{8T}{OA}\right), \qquad (12)
\end{aligned}$$

where we used Lemma 6 to upper bound $m(T)$.

The Final Bound. For the regret built during the good episodes, we have

$$\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M \in \mathcal{M}_k\}} \le 2\sqrt{O + \log\big(OA\sqrt{T+1}/\delta\big)} \sum_{k=1}^{m(T)} \sum_{o,a} \Big(\max_{q\in\mathcal{Q}} D_{q,o} + 2\Big) \frac{v_k(o,a)}{\sqrt{N_k(o,a)}}$$
$$+ 2\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} + D_{\mathsf{cp}}\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}} OA \log_2\big(\tfrac{8T}{OA}\big)\,, \quad (13)$$

with probability higher than $1-\delta$ and uniformly over all $T \in \mathbb{N}$. Applying Lemma 5 and using the Cauchy-Schwarz inequality:

$$\begin{aligned}\sum_{k=1}^{m(T)} \sum_{o,a} \max_{q\in\mathcal{Q}} D_{q,o} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} &\le (\sqrt{2}+1)\sum_{o,a} \max_{q\in\mathcal{Q}} D_{q,o}\sqrt{N_T(o,a)} \\ &\le (\sqrt{2}+1)\sqrt{\sum_{o,a} \max_{q\in\mathcal{Q}} D_{q,o}^2 \cdot \sum_{o,a} N_T(o,a)} \\ &= (\sqrt{2}+1)\sqrt{T\sum_{o,a} \max_{q\in\mathcal{Q}} D_{q,o}^2} = (\sqrt{2}+1)\sqrt{\mathbf{c}_M AT}\,,\end{aligned}$$

where, with a slight abuse of notation, we used $N_T(o,a)$ to denote the number of visits to (o,a) after T rounds. Similarly, we have

$$\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \le (\sqrt{2}+1)\sum_{o,a}\sqrt{N_T(o,a)} \le (\sqrt{2}+1)\sqrt{OA\sum\nolimits_{o,a} N_T(o,a)} = (\sqrt{2}+1)\sqrt{OAT}\,.$$

Combining this with (13), and putting together, we have that with probability at least $1-4\delta$,

$$\mathfrak{R}(T) \le 2(\sqrt{2}+1)\sqrt{O+\log\big(OA\sqrt{T+1}/\delta\big)}\big(\sqrt{\mathbf{c}_M}+2\big)\sqrt{AT} + 2(\sqrt{2}+1)\sqrt{OAT}$$
$$+ (D_{\mathsf{cp}}+1)\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}} OA\log_2\big(\tfrac{8T}{OA}\big)\,,$$

thus proving the theorem. □

C.2 Proof of Theorem 2

Let $\delta \in (0,1)$. Following the same steps as in the proof of Theorem 1, we have

$$\mathfrak{R}(T) \le \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,$$

with probability at least $1-\delta$, where Δ_k is defined similarly to the proof of Theorem 1. Furthermore, By Lemma 2, with probability at least $1-\delta$, $\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M\notin\mathcal{M}_k\}} = 0$.

Let's now focus on good episodes, i.e., episodes k where $M \in \mathcal{M}_k$. Similarly to the proof of Theorem 1, we have that

$$\Delta_k \le \sum_{s,a} v_k(s,a)\big(g_k - \widetilde{\mu}(s,a)\big) + \sum_{s,a} v_k(s,a)\big(\widetilde{\mu}_k(s,a) - \mu(s,a)\big) + \sum_{s,a}\frac{v_k(s,a)}{\sqrt{t_k}}\,.$$

The first and third terms are bounded as in the proof of Theorem 1. However, the second term in the right-hand side is bounded as follows:

$$
\begin{aligned}
\widetilde{\mu}_k(s,a)-\mu(s,a) &= \sum_{o'\in\mathcal{O}} \big(\widetilde{p}_k^q(o'|o,a)-p(o'|o,a)\big)\overline{\nu}(q,L(o,a,o')) \\
&\le \sum_{o'\in\mathcal{O}} \big|\widetilde{p}_k^q(o'|o,a)-p(o'|o,a)\big| \\
&\le \sum_{o'\in\mathcal{O}} \sqrt{\frac{2\widetilde{p}_k^q(o'|o,a)(1-\widetilde{p}_k^q(o'|o,a))}{N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{2O^2A}\big)} \\
&+ \sum_{o'\in\mathcal{O}} \sqrt{\frac{2p(o'|o,a)(1-p(o'|o,a))}{N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{2O^2A}\big)} + \frac{2}{3N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{2O^2A}\big) \\
&\overset{\text{(a)}}{\le} \sqrt{\beta'_T\big(\tfrac{\delta}{2O^2A}\big)}\sum_{o'\in\mathcal{O}} \sqrt{\frac{2\widehat{p}_k(o'|o,a)(1-\widehat{p}_k(o'|o,a))}{N_k(o,a)}} \\
&+ \sqrt{\beta'_T\big(\tfrac{\delta}{2O^2A}\big)}\sum_{o'\in\mathcal{O}} \sqrt{\frac{2p(o'|o,a)(1-p(o'|o,a))}{N_k(o,a)}} + \frac{4}{N_k(o,a)}\beta'_T\big(\tfrac{\delta}{2O^2A}\big) \\
&\overset{\text{(b)}}{\le} \sqrt{8\beta'_T\big(\tfrac{\delta}{2O^2A}\big)\frac{K_{o,a}}{N_k(o,a)}} + \frac{4}{N_k(o,a)}\beta'_T\big(\tfrac{\delta}{2O^2A}\big) \qquad (14)
\end{aligned}
$$

where (a) follows from Lemma 4, and where (b) uses the fact that for a distribution $p\in\Delta_{\mathcal{O}}$ with K non-zero elements, we have

$$
\sum_{o\in\mathcal{O}}\sqrt{p(o)(1-p(o))} = \sum_{o:p(o)>0}\sqrt{p(o)(1-p(o))}\sqrt{\sum_{o:p(o)>0}p(o)\sum_{o:p(o)>0}(1-p(o))} = \sqrt{K-1}\,.
$$

Hence, using the bounds derived in the proof of Theorem 1, we have

$$
\begin{aligned}
\Delta_k &\le v_k(\mathbf{P}_k-\mathbf{I})w_k + \underbrace{v_k(\widetilde{\mathbf{P}}_k-\mathbf{P}_k)w_k}_{L_1} + \sqrt{8\beta'_T\big(\tfrac{\delta}{2O^2A}\big)}\sum_{o,a}v_k(o,a)\sqrt{\frac{K_{o,a}}{N_k(o,a)}} \\
&+ 4\beta'_T\big(\tfrac{\delta}{2O^2A}\big)\sum_{o,a}\frac{v_k(o,a)}{N_k(o,a)} + 2\sum_{o,a}\frac{v_k(o,a)}{N_k(o,a)}\,.
\end{aligned}
$$

where w_k is the same as in the proof of Theorem 1.

Upper Bound on L_1. We have

$$
\begin{aligned}
&v_k(\widetilde{\mathbf{P}}_k-\mathbf{P}_k)w_k \\
&\quad= \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s))\sum_{o'\in\mathcal{O}}\sum_{q'\in\mathcal{Q}}\Big(\widetilde{p}_k^q(o'|o,\pi_k(s))-p(o'|o,\pi_k(s))\Big)\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}w_k(q',o') \\
&\quad\le \sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}} v_k(s,a)\sum_{o'\in\mathcal{O}}\sum_{q'\in\mathcal{Q}}\Big(\widetilde{p}_k^q(o'|o,a)-p(o'|o,a)\Big)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}}w_k(q',o') \\
&\quad\le \sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}} v_k(s,a)\sum_{o'\in\mathcal{O}}\Big|\widetilde{p}_k^q(o'|o,a)-p(o'|o,a)\Big|\cdot\max_{s'\in\mathcal{B}_{q,o}\times\mathcal{O}}\big|w_k(q',o')\big|\underbrace{\sum_{q'\in\mathcal{Q}}\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}}}_{=1} \\
&\quad\le \frac{1}{2}\sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}} D_s v_k(s,a)\sum_{o'\in\mathcal{O}}\Big|\widetilde{p}_k^q(o'|o,a)-p(o'|o,a)\Big|\,,
\end{aligned}
$$

where the last inequality follows from Lemma 3.

Now plugging in the bound derived for $\sum_{o'\in\mathcal{O}} \left|\widetilde{p}_k^q(o'|o,a) - p(o'|o,a)\right|$ in (14), we obtain

$$
\begin{aligned}
&v_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k \\
&\qquad \le \sqrt{8\beta'_T\big(\tfrac{\delta}{2O^2A}\big)}\sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\max_{q\in\mathcal{Q}} D_{q,o}\cdot v_k(o,a)\sqrt{\frac{K_{o,a}}{N_k(o,a)}} + 4D_{\mathsf{cp}}\beta'_T\big(\tfrac{\delta}{2O^2A}\big)\sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\frac{v_k(o,a)}{N_k(o,a)}\,.
\end{aligned}
$$

The rest of the proof follows similar lines as in the proof of Theorem 1. □

C.3 Technical Lemmas

Lemma 3 *For all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have:*

$$
\max_{s'\in\mathcal{B}_s\times\mathcal{O}} |w_k(s')| \le \frac{D_s}{2}\,, \qquad \|w_k\|_\infty \le \frac{D_{\mathsf{cp}}}{2}\,.
$$

Proof. The proof is quite similar to the one of Lemma 8 in (Bourel et al., 2020). We first show that for all $s_1, s_2 \in \mathcal{B}_s\times\mathcal{O}$, we have $u_k^{(i)}(s_1) - u_k^{(i)}(s_2) \le D_s$, which further implies

$$
\max_{x\in\mathcal{B}_s\times\mathcal{O}} |w_k(x)| \le \tfrac{D_s}{2}.
$$

To prove this, recall that similarly to (Jaksch et al., 2010), we can combine all MDPRMs in $\mathcal{M}_k$ to form a single MDPRM $\widetilde{\mathcal{M}}_k$ with continuous action space $\mathcal{A}'$. In this extended MDPRM, in any $s=(q,o)\in\mathcal{S}$, and for each $a\in\mathcal{A}$, there is an action in $\mathcal{A}'$ with mean $\widetilde{\mu}_k(s,a)$ and transition probability $\widetilde{P}_k(\cdot|s,a)$ (of the associated M_{cp}) belonging to the maintained confidence sets. Similarly to (Jaksch et al., 2010), we note that $u_k^{(i)}(s)$ amounts to the total expected i-step reward of an optimal non-stationary i-step policy starting in state s on the MDPRM $\widetilde{\mathcal{M}}_k$ with the extended action set. The RM-restricted diameter of state s of this extended MDPRM is at most D_s, since by assumption k is a good episode and hence $\mathcal{M}_k$ contains the true MDPRM M, and therefore, the actions of the true MDPRM are contained in the continuous action set of $\widetilde{\mathcal{M}}_k$. Now, if there were states $s_1, s_2\in\mathcal{B}_s\times\mathcal{O}$ with $u_k^{(i)}(s_1) - u_k^{(i)}(s_2) > D_s$, then an improved value for $u_k^{(i)}(s_1)$ could be achieved by the following non-stationary policy: First follow a policy that moves from s_1 to s_2 most quickly, which takes at most D_s steps on average. Then follow the optimal i-step policy for s_2. We thus have $u_k^{(i)}(s_1) \ge u_k^{(i)}(s_2) - D_s$, since at most D_s of the i rewards of the policy for s_2 are missed. This is a contradiction, and so the claim follows. The second bound directly follows from the same arguments as in (Jaksch et al., 2010). □

Lemma 4 ((Bourel et al., 2020, Lemma 11)) *Consider x and y satisfying $|x-y| \le \sqrt{2y(1-y)\zeta} + \zeta/3$. Then,*

$$
\sqrt{y(1-y)} \le \sqrt{x(1-x)} + 2.4\sqrt{\zeta}\,.
$$

Lemma 5 ((Jaksch et al., 2010, Lemma 19),(Talebi and Maillard, 2018, Lemma 24)) *For any sequence of numbers $z_1, z_2, \dots, z_n$ with $0\le z_k \le Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$, it holds*

$$
(i) \qquad \sum_{k=1}^n \frac{z_k}{\sqrt{Z_{k-1}}} \le \big(\sqrt{2}+1\big)\sqrt{Z_n}\,.
$$

$$
(ii) \qquad \sum_{k=1}^n \frac{z_k}{Z_{k-1}} \le 2\log(Z_n) + 1\,.
$$

Lemma 6 ((Jaksch et al., 2010, Proposition 18)) *The number $m(T)$ of episodes up to time $T\ge OA$ satisfies*

$$
m(T) \le OA\log_2\big(\tfrac{8T}{OA}\big)\,.
$$

C.4 Concentration Inequalities

In this subsection, we collect a few useful concentration inequalities. They can be found in, e.g., (Maillard, 2019; Lattimore and Szepesvári, 2020; Dann et al., 2017; Bourel et al., 2020).

We begin with the following definition:

Definition 2 (Sub-Gaussian Observation Noise) *A sequence $(Y_t)_t$ has conditionally σ-sub-Gaussian noise if*

$$\forall t, \forall \lambda \in \mathbb{R}, \quad \log \mathbb{E}[\exp\big(\lambda(Y_t - \mathbb{E}[Y_t|\mathcal{F}_{t-1}])\big)\big|\mathcal{F}_{t-1}] \le \frac{\lambda^2\sigma^2}{2}\,,$$

where $\mathcal{F}_{t-1}$ denotes the σ-algebra generated by $Y_1, \dots, Y_{t-1}$.

Lemma 7 (Time-Uniform Laplace Concentration for Sub-Gaussian Distributions) *Let $Y_1, \dots, Y_n$ be a sequence of n i.i.d. real-valued random variables with mean μ, such that $Y_n - \mu$ is σ-sub-Gaussian. Let $\widehat{\mu}_n = \frac{1}{n}\sum_{s=1}^n Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\bigg(\exists n \in \mathbb{N}, \quad |\widehat{\mu}_n - \mu| \ge \sigma\sqrt{\Big(1+\frac{1}{n}\Big)\frac{2\ln\big(\sqrt{n+1}/\delta\big)}{n}}\bigg) \quad \le \quad \delta\,.$$

The "Laplace" method refers to using the Laplace method of integration for optimization. We recall that random variables bounded in $[0,1]$ are $\frac{1}{2}$-sub-Gaussian. The following corollary is an immediate consequence of Lemma 7:

Corollary 1 (Time-Uniform Azuma-Hoeffding Concentration Using Laplace) *Let $(X_t)_{t\ge1}$ be a martingale difference sequence such that for all t, $X_t \in [a,b]$ almost surely for some $a, b \in \mathbb{R}$. Then, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\bigg(\exists T \in \mathbb{N} : \sum_{t=1}^T X_t \ge (b-a)\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\bigg) \le \delta\,.$$

Lemma 7 can be used to provide a time-uniform variant of Weissman's concentration inequality (Weissman et al., 2003) using the method of mixture (a.k.a. the Laplace method):

Lemma 8 (Time-Uniform L1-Deviation Bound for Categorical Distributions Using Laplace) *Consider a finite alphabet $\mathcal{X}$ and let P be a probability distribution over $\mathcal{X}$. Let $(X_t)_{t\ge1}$ be a sequence of i.i.d. random variables distributed according to P, and let $\widehat{P}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i = x\}}$ for all $x \in \mathcal{X}$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \|P - \widehat{P}_n\|_1 \ge \sqrt{\frac{2}{n}\Big(1+\frac{1}{n}\Big)\log\left(\sqrt{n+1}\frac{2^{|\mathcal{X}|}-2}{\delta}\right)}\right) \le \delta\,.$$

The following lemma provides a time-uniform Bernstein-type concentration inequality for bounded random variables:

Lemma 9 (Time-Uniform Bernstein for Bounded Random Variables Using Peeling) *Let $Z = (Z_t)_{t\in\mathbb{N}}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F} = (\mathcal{F}_t)_t$ be its natural filtration. Assume for all $t \in \mathbb{N}$, $|Z_t| \le b$ and $\mathbb{E}[Z_s^2|\mathcal{F}_{s-1}] \le v$ for some positive numbers v and b. Let n be an integer-valued (and possibly unbounded) random variable that is $\mathcal{F}$-measurable. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\bigg(\exists n \in \mathbb{N},\ \frac{1}{n}\sum_{t=1}^n Z_t \ge \sqrt{\frac{2\ell_n(\delta)v}{n}} + \frac{\ell_n(\delta)b}{3n}\bigg) \le \delta\,,$$

$$\mathbb{P}\bigg(\exists n \in \mathbb{N},\ \frac{1}{n}\sum_{t=1}^n Z_t \le -\sqrt{\frac{2\ell_n(\delta)v}{n}} - \frac{\ell_n(\delta)b}{3n}\bigg) \le \delta\,,$$

where $\ell_n(\delta) := \eta\log\left(\frac{\log(n)\log(\eta n)}{\delta\log^2(\eta)}\right)$, with $\eta > 1$ being an arbitrary parameter.

Lemma 9 is derived from Lemma 2.4 in (Maillard, 2019). We note that any $\eta > 1$ is valid here, but numerically optimizing the bound shows that $\eta = 1.12$ seems to be a good choice and yields a small bound. For example, when $(X_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. Bernoulli random variables with mean μ, we have, for all $\delta \in (0,1)$,

$$\mathbb{P}\bigg(\exists n \in \mathbb{N},\ \mu - \frac{1}{n}\sum_{t=1}^n X_t \ge \sqrt{\frac{2\ell_n(\delta)\mu(1-\mu)}{n}} + \frac{\ell_n(\delta)}{3n}\bigg) \le \delta\,,$$

$$\mathbb{P}\bigg(\exists n \in \mathbb{N},\ \mu - \frac{1}{n}\sum_{t=1}^n X_t \le -\sqrt{\frac{2\ell_n(\delta)\mu(1-\mu)}{n}} - \frac{\ell_n(\delta)}{3n}\bigg) \le \delta\,,$$

D REGRET LOWER BOUND

In this section, we prove Theorem 3. Our proof uses the machinery of establishing a minimax regret lower bound in Jaksch et al. (2010) for tabular MDPs. (We also refer to (Lattimore and Szepesvári, 2020, Chapter 38.7).) This machinery for tabular MDPs consists in crafting a worst-case MDP and showing that the regret under any algorithm on the MDP is lower bounded. We take a similar approach here but stress that constructing a worst-case MDPRM entails constructing a worst-case reward machine and a labeled MDP simultaneously. In terms of notations and presentation, we closely follow (Lattimore and Szepesvári, 2020, Chapter 38.7).

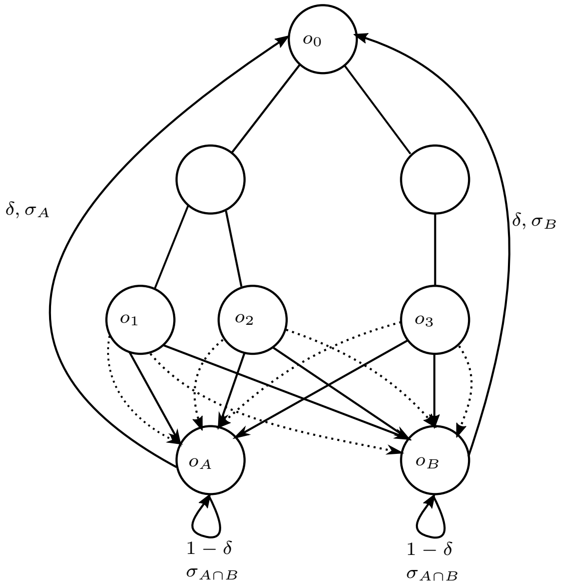


Figure 7: Construction of the underlying labeled MDP for the LB with $A = 2$ and $O = 8$, based on the worst-case MDP in (Lattimore and Szepesvári, 2020, Chapter 38.7).

Proof (of Theorem 3). To prove the theorem, we construct a worst-case MDPRM, which can be seen as an MDPRM that models a bandit problem with approximately OA arms, such that obtaining the reward requires to pick the 'good arm' Q times. Figures 8 and 7 show the construction, given O and A: We build a tree of minimum depth with at most A children for each node using exactly $O - 2$ observations. The root of the tree is denoted o_0 and transitions within the tree are deterministic. So, in a node of the tree the agent can simply select the child to transition to. Let L be the number of leaves, and let us index observations as $o_1, o_2, \dots, o_L$. The last two observations are o_A and o_B where events are given as detailed later. Then, for each $i \in [\![1, L]\!]$ the agent can choose any action $a \in \mathcal{A}$ and transitions to either o_A or o_B according to:

$$p(o_A|o_i, a) = \frac{1}{2} + \varepsilon(a, i) \quad \textit{and} \quad p(o_B|o_i, a) = \frac{1}{2} - \varepsilon(a, i),$$

where $\varepsilon(a, i) = 0$ for all (a, i) pairs except for one particular pair, for which $\varepsilon(a, i) = \Delta > 0$. ($\Delta$ will be chosen later in the proof.) The transition probabilities at o_A and o_B under any $a \in \mathcal{A}$ satisfy:

$$p(o|o, a) = 1 - \delta, \quad p(o_0|o, a) = \delta, \quad o \in \{o_A, o_B\}.$$

Let us choose $\delta = \frac{6Q}{D_{\mathsf{cp}}}$. Note that by the assumptions of the theorem, $\delta \in (0, 1]$. Furthermore, this choice ensures that the diameter of the cross-product MDP associate to the described MDPRM is at most D_{cp}, regardless of the value of Δ. Also, for the diameter of the labeled MDP, D, we will have $D = \frac{4}{\delta}$.

The labelling function is defined as follows. Since we assume $|\mathcal{P}| \geq 2$, we can consider three events $\sigma_A, \sigma_B, \sigma_{A\cap B}$ and define labelling function as follows: For all action $a \in \mathcal{A}$,

$$\begin{aligned} L(o_A, a, o_0) &= \sigma_A, \qquad & L(o_A, a, o_A) &= \sigma_{A\cap B}, \\ L(o_B, a, o_0) &= \sigma_B, \qquad & L(o_B, a, o_B) &= \sigma_{A\cap B}. \end{aligned}$$

To build the RM, we let $N = \lceil (Q-1)/2 \rceil$ and $N' = \lfloor (Q-1)/2 \rfloor$ so that $N + N' = Q - 1$. The idea is to arrange the Q many nodes of the RM into 2 cycles of lengths N and N'; see Figure 7. To this effect, we let q_0 be the origin. Then, the set

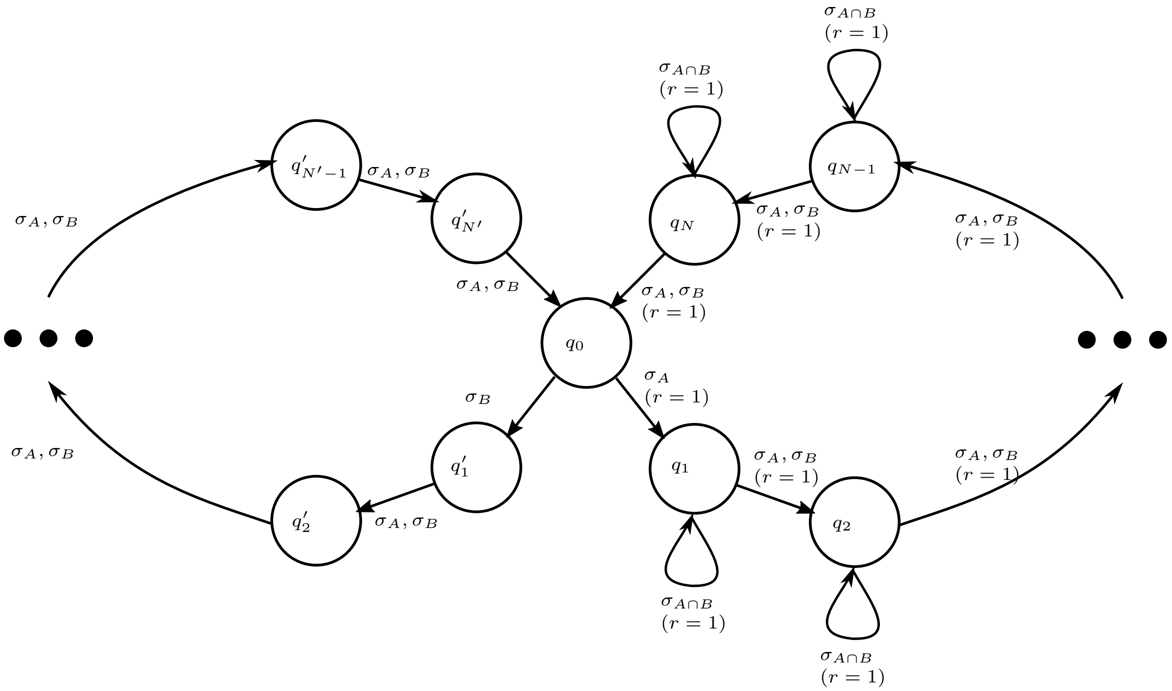

$\sigma_{A\cap B}$ $(r=1)$
$\sigma_{A\cap B}$ $(r=1)$
$q'_{N'-1}$
σ_A, σ_B
$q'_{N'}$
q_N
q_{N-1}
σ_A, σ_B $(r=1)$
σ_A, σ_B
σ_A, σ_B $(r=1)$
σ_A, σ_B
σ_A, σ_B $(r=1)$
q_0
σ_B
σ_A $(r=1)$
σ_A, σ_B
q'_1
q_1
σ_A, σ_B $(r=1)$
σ_A, σ_B $(r=1)$
q'_2
σ_A, σ_B
q_2
$\sigma_{A\cap B}$ $(r=1)$
$\sigma_{A\cap B}$ $(r=1)$

Figure 8: Construction of the underlying RM for the lower bound with a double-cyclic a ‘good’ cycle giving rewards and ‘bad’ cycle of similar length giving no reward.

$\{q_i\}_{i=0}^{N}$ of RM states defines the ‘good’ cycle, whereas the set $\{q'_j\}_{j=1}^{N'} \cup \{q_0\}$ define the ‘bad’ cycle. Then, we build the RM transition function τ and reward function r, for all $i \in [\![1, N]\!]$ and all $j \in [\![1, N']\!]$:

$$\begin{aligned}
\tau(q_0, \sigma_A) &= q_1, & r(q_0, \sigma_A) &= 1,\\
\tau(q_0, \sigma_B) &= q'_1, & r(q_0, \sigma_B) &= 0,\\
\tau(q_i, \sigma_A) &= q_{i+1}, & r(q_i, \sigma_A) &= 1,\\
\tau(q_i, \sigma_B) &= q_{i+1}, & r(q_i, \sigma_B) &= 1,\\
\tau(q_i, \sigma_{A\cap B}) &= q_i, & r(q_i, \sigma_{A\cap B}) &= 1,\\
\tau(q_N, \sigma_A) &= q_0, & r(q_N, \sigma_A) &= 1,\\
\tau(q_N, \sigma_B) &= q_0, & r(q_N, \sigma_B) &= 1,\\
\tau(q_N, \sigma_{A\cap B}) &= q_N, & r(q_N, \sigma_{A\cap B}) &= 1,\\
\tau(q'_j, \sigma_A) &= q'_{j+1}, & r(q'_j, \sigma_A) &= 0,\\
\tau(q'_j, \sigma_B) &= q'_{j+1}, & r(q'_j, \sigma_B) &= 0,\\
\tau(q'_{N'}, \sigma_A) &= q_0, & r(q'_{N'}, \sigma_A) &= 0,\\
\tau(q'_{N'}, \sigma_B) &= q_0, & r(q'_{N'}, \sigma_B) &= 0,
\end{aligned}$$

where all non-specified transitions imply no change of state, and where all non-specified rewards are zero. This means that in q_0, the agent needs to realize the event σ_A to initiate a rotation of the ‘good’ cycle, where in all states the agent will get a reward when staying in either o_A or o_B and progresses one step forward in the cycle when leaving one of both RM-states. On the other hand, if the agent is in q_0, she receives the event σ_B and then initiates a rotation of the ‘bad’ cycle, without any reward but similar length and transitions as for the ‘good’ cycle.

In summary, each time the agent arrives in $s_0 = (o_0, q_0)$, she selects which leaf to visit and then chooses an action from that leaf. This corresponds to choosing one of $k = LA = \Omega(OA)$ meta actions. The optimal policy is to select the meta action with the largest probability of transitioning to the observation o_A. The choice of δ ensures that the agent expects to stay at state o_A or o_B for approximately D rounds. Since all choices are equivalent when $q \neq q_0$, the agent expects to make about $\frac{2T}{DQ}$ decisions and the rewards are roughly in $[0, \frac{DQ}{8}]$, or $3DQ = 2D_{\mathsf{cp}}$, so we should expect the regret to be $\Omega(D_{\mathsf{cp}}\sqrt{kT/D_{\mathsf{cp}}}) = \Omega(\sqrt{TD_{\mathsf{cp}}OA})$.

Characterisation of the MDPRM. Using the introduced notations, we introduce $\mathcal{L}$ and $\mathcal{L}^M$:

$$\mathcal{L} = \{(q_0, o, a) : a \in \mathcal{A} \text{ and } o \text{ is a leaf of the tree}\},$$
$$\mathcal{L}^M = \{(o, a) : a \in \mathcal{A} \text{ and } o \text{ is a leaf of the tree}\}.$$

By definition, both have k elements. Then, let M_0 be the MDPRM with $\varepsilon(o, a) = 0$ for all pairs in $\mathcal{L}^M$. Then let M_j be the MDPRM with $\varepsilon(o, a) = \Delta$ for the j-th observation-action pair in the set $\mathcal{L}^M$. Similarly to (Lattimore and Szepesvári, 2020), we define the stopping time T_{stop} as the first time when the number of visits of (q_0, s_0) is at least $T/D_{\mathsf{cp}} - 1$, or T if the state (q_0, s_0) is not visited enough:

$$T_{\text{stop}} = \min\left\{T, \min\left\{t : \sum_{t'=1}^{t} \mathbb{I}_{\{s_{t'}=(q_0,o_0)\}} \geq \frac{T}{D_{\mathsf{cp}}} - 1\right\}\right\}.$$

Also, let T_j be the number of visits to the j-th triplet of $\mathcal{L}$ until T_{stop} and $T_{\text{tot}} = \sum_{j=1}^{k} T_j$. We also let $P_j, 0 \leq j \leq k$ denote the probability distribution of $T_1, \ldots, T_k$ induced by the interaction of π and M_j and let $\mathbb{E}_j[\cdot]$ be the expectation with respect to P_j.

Now, we study the characteristics of the MDPRM. In doing so, we first build upon (Lattimore and Szepesvári, 2020, Claim 38.9) that establishes that the diameter of the underlying MDP of M_j, denoted by $D(M_j)$, is bounded by D for all $j \in [\![1, k]\!]$. Then, we have for $D_{\mathsf{cp}}(M_j)$ cross-product diameter of the MDPRM M_j:

$$D_{\mathsf{cp}}(M_j) \leq DN + DN \sum_{i=1}^{\infty} \frac{1}{2^i} + DN' \leq \frac{3}{2} QD = D_{\mathsf{cp}}.$$

The first inequality can be interpreted as the fact that the cross-product diameter is smaller that completing the 2 loops of the RM plus accounting the probability to have a transition to the "wrong" loop when in q_0. The rest follows by construction and we note that we can ignore Δ due to the fact that it can only reduce the diameters.

Following the same arguments as in Claim 38.10 of (Lattimore and Szepesvári, 2020), there exist universal constants $0 < c_1 < c_2 < \infty$ such that $D_{\mathsf{cp}} \mathbb{E}_0[T_\sigma]/T \in [c_1, c_2]$. By construction, we have

$$\frac{D_{\mathsf{cp}} \mathbb{E}_0[T_{\text{tot}}]}{T} \leq \frac{\mathbb{E}[T_{\text{tot}}]}{OA} \leq \frac{T}{DN'OA} \leq c_2$$

Similarly,

$$\frac{D_{\mathsf{cp}} \mathbb{E}[T_{\text{tot}}]}{T} \geq \frac{\mathbb{E}_0[T_{\text{tot}}]}{OA} \geq \frac{T}{DNOA} \geq c_1.$$

Finally, we write $\mathbb{E}[\mathfrak{R}_j(T)]$ the expected regret of policy π in the MDPRM M_j over T steps and prove that there exists a universal constant $c_3 > 0$ such that:

$$\mathbb{E}[\mathfrak{R}_j(T)] \geq c_3 \Delta D_{\mathsf{cp}} \mathbb{E}[T_{\text{tot}} - T_j]\,.$$

To prove this result, we first write the definition of the expected regret:

$$\mathbb{E}[\mathfrak{R}_j(T)] = \sum_{t=1}^{T} \mathbb{E}_j^\star[r_t] - \sum_{t=1}^{T} \mathbb{E}_j[r_t],$$

where $\mathbb{E}_j^\star$ is the expectation in MDPRM M_j when following the optimal policy, which mean always choosing the j-th element of $\mathcal{L}$ when in (q_0, o_0). Now, we can decompose the cumulative reward by "episodes", where a new episode start whenever reaching (q_0, o_0). By construction and using our knowledge of the optimal policy, this yields:

$$\begin{aligned}\mathbb{E}[\mathfrak{R}_j(T)] &\geq \mathbb{E}_j[T_{\text{tot}}]\big(\frac{1}{2} + \Delta\big)\frac{DN}{4} - \mathbb{E}[T_{\text{tot}} - T_j]\frac{DN}{8} - \mathbb{E}_j[T_j]\big(\frac{1}{2} + \Delta\big)\frac{DN}{4} \\ &= \mathbb{E}_j[T_{\text{tot}} - T_j]\Delta\frac{DN}{4},\end{aligned}$$

or by definition of D and N there exists a universal constant $c_3 > 0$ such that $c_3 D_{\mathsf{cp}} \geq \frac{DN}{4}$, which allows us to conclude.

The Final Lower Bound. Let $\mathrm{D}(P, Q)$ denote the Kullback-Leibler divergence between two probability distributions P and Q. Similarly to (Lattimore and Szepesvári, 2020, Chapter 38.7) and (Jaksch et al., 2010) (as well as lower bound proofs for bandit problems), we have $\mathrm{D}(P_0, P_j) = \mathbb{E}_0[T_j]d(1/2, 1/2 + \Delta)$, where $d(p, q)$ is the relative entropy between Bernoulli distributions with respective means p and q. Now the conclusion of the proof is exactly the same as for MDPs (Jaksch et al., 2010): We assume that the chosen Δ will satisfy $\Delta \le 1/4$, then using the entropy inequalities from (Lattimore and Szepesvári, 2020, Equation 14.16), we have:

$$\mathrm{D}(P_0, P_j) \le 4\Delta^2 \mathbb{E}_0[T_j].$$

Then following the same steps as in (Lattimore and Szepesvári, 2020, Chapter 38.7) and using Pinsker's inequality, and using the fact that $0 \le T_{\text{tot}} - T_j \le T_{\text{tot}} \le T/D_{\textsf{cp}}$, we have

$$\mathbb{E}_j[T_{\text{tot}} - T_j] \ge \mathbb{E}[T_{\text{tot}} - T_j] - \frac{T}{D_{\textsf{cp}}}\sqrt{\frac{\mathrm{D}(P_0, P_j)}{2}} \ge \mathbb{E}_0[T_{\text{tot}} - T_j] - \frac{T\Delta}{D_{\textsf{cp}}}\sqrt{2\mathbb{E}_0[T_j]}.$$

Summing over j and applying Cauchy-Schwarz give us

$$\begin{aligned}\sum_{j=1}^{k} \mathbb{E}_j[T_{\text{tot}} - T_j] &\ge \sum_{j=1}^{k} \mathbb{E}_0[T_{\text{tot}} - T_j] - \frac{T\Delta}{D_{\textsf{cp}}}\sum_{j=1}^{k}\sqrt{2\mathbb{E}_0[T_j]} \\ &\ge (k-1)\mathbb{E}_0[T_{\text{tot}}] - \frac{T\Delta}{D_{\textsf{cp}}}\sqrt{2k\mathbb{E}_0[T_{\text{tot}}]} \\ &\ge \frac{c_1 T(k-1)}{D_{\textsf{cp}}} - \frac{T\Delta}{D_{\textsf{cp}}}\sqrt{\frac{2c_2Tk}{D_{\textsf{cp}}}}\,.\end{aligned}$$

Now choosing $\Delta = \frac{c_1(k-1)}{2}\sqrt{\frac{D_{\textsf{cp}}}{2c_2Tk}}$ yields

$$\sum_{j=1}^{k} \mathbb{E}_j[T_{\text{tot}} - T_j] \ge \frac{c_1T(k-1)}{2kD_{\textsf{cp}}}\,.$$

This implies that there exists j such that $\mathbb{E}_j[T_{\text{tot}} - T_j] \ge \frac{c_1T(k-1)}{2kD_{\textsf{cp}}}$, which leads to the final result using the previous lower bound on the regret

$$\mathbb{E}[\mathfrak{R}_j(T)] \ge c_3 D_{\textsf{cp}}\Delta\mathbb{E}_j[T_{\text{tot}} - T_j] \ge \frac{c_1^2c_3T(k-1)^2}{4k}\sqrt{\frac{D_{\textsf{cp}}}{2c_2Tk}} = c_0\sqrt{D_{\textsf{cp}}OAT},$$

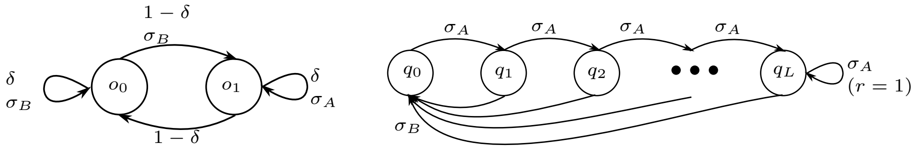
with $c_0 > 0$ being a universal constant. □

E DETAILS OF DIAMETER COMPUTATIONS

Numerical Computation of Diameters. The RM-restricted diameter D_s can be numerically computed via solving the corresponding reachability problems of conventional MDPs. The (global) diameter of a given tabular MDP M can be computed as follows. For a given s, we modify M to another MDP M', where M' has zero rewards everywhere except for s_1 in which the reward is 1 under all actions. Also, M' has the same transition as M, except for s_1 in which $p(s_1|s_1, a) = 1$ under all actions in s_1, so as to make s_1 absorbing. Now the optimal bias function in M', denoted by $b^\star$ (found via Value Iteration), can be interpreted as follows: $b^\star(s_1) - b^\star(s_2)$ denotes the amount of steps (in expectation) needed to reach s_1 from s_2. Hence, the *farthest state from* s_1 has a path length of $\max_{s_2}(b^\star(s_1) - b^\star(s_2))$, and hence, by definition of diameter, $D = \max_{s_1,s_2\in\mathcal{S}}(b^\star(s_1) - b^\star(s_2))$.

We modify the above procedure slightly to compute D_s as follows —the procedure is implemented in `diameter.py`, in our uploaded codes. In view of the equivalence between an MDPRM and its associated cross-product MDP, we will be working with the latter. To find D_s for a given s, we restrict to $s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}$.

Diameters Computation for the MDPRM in Figure 2. In this MDPRM, we notice that for $\delta \in (0, \frac{1}{2})$, the RM restricted diameter from o_0 for any $q \in \mathcal{Q}$ coincides with the diameter of the underlying MDP. Hence, $D_{o_0,q} = \frac{1}{\delta}$ for all $q \in \mathcal{Q}$. Now,



(a) MDP (b) RM

Figure 9: Example $D_{\mathsf{cp}} \to \infty$ when $D \to 1$

observe that the diameter for both D_{cp} and the restricted diameters from o_1 will be the expected number of steps for a trajectory from (q, o_0) and (q', o_0) where q and q' are two RM-states with the maximum number of steps possible between them. Let N denote this number of steps. Then, we give ourselves D_N the diameter over a communicating subset of $\mathcal{Q}$ with maximum number of steps between 2 steps being N, which yields:

$$D_1 = \frac{1}{\delta} + (1-\delta) + \delta\Big(1 + \frac{1}{1-\delta}\Big) = \frac{1}{\delta} + 1 + \frac{\delta}{1-\delta}.$$

Then a simple recurrence shows that for all N:

$$D_N = \frac{N}{\delta} + 1 + \frac{\delta}{1-\delta},$$

or, we have $N = 2$ for D_{q,o_1} and $N = \lfloor Q/2 \rfloor$ for D_{cp}, which concludes the analysis.

Diameters Computation for the MDPRM in Figure 9. This additional example shows the absence of correlation in general between the diameter D of the underlying MDP and the diameter D_{cp} of the cross product. Indeed, if assume $\delta \in (0,1)$ and $L \geq 2$ then we immediately have $D = \frac{1}{1-\delta}$, and the construction of MDPRM ensures that $D_{\mathsf{cp}} > \frac{1}{\delta^L}$. Thus when $\delta \to 0$ we can immediately conclude that $D \to 1$ and $D_{\mathsf{cp}} \to \infty$.

This example illustrates the difficulty of MDPRM when the events are "dense", which can lead in extreme cases to unsolvable problems (non-communicating cross-product) despite a simple underlying MDP. Nonetheless, we remark that in a practical use of MDPRM, events would be expected to be scarce thus leading to MDPRM where $D_{\mathsf{cp}} \leq DN$ where N is the longest path within the RM. The previous example represents such a case.

F DETAILS OF EXPERIMENTS AND FURTHER EXPERIMENTS

In this section, we provide further details about the experiments reported in Section 6 and present additional experimental results. All our experiments are implemented in python3, the environments being based on a framework from the package *gym* (see (Brockman et al., 2016)).

Figure 10 shows the cross-product MDP M_{cp} associated to *RiverSwim-patrol2* MDPRM. In fact, this is the MDP to which the baseline algorithms in the experiments are applied. We also present in Figure 12(a) the same results for the 6-state *RiverSwim* MDPRM as in Section 6 but without the log-scale, and in Figure 12(b) results in a similar 20-state environment. The results in the MDPRM based on the 20-state *RiverSwim* differ from the other environments due to the global under-performance of the algorithms based on Bernstein-type confidence bounds. As explained in Section 6, Bernstein-type confidence sets lead to excessive computational cost due to problems of convergence in `EVI`. Hence, for algorithms using such confidence sets, we chose to limit the number of iterations in `EVI` to 100. This proved necessary for practical constraints. However, it results in a fairer comparison between the algorithms as it ensures having similar computational cost for all tested algorithms (by forcefully lowering the computational time of Bernstein variants). We remark that this limitation comes at the expense of worsening the performance of the algorithms. The results in the 20-state *RiverSwim* MDPRM are an extreme example of the consequences of this choice. This environment is indeed extremely challenging to any exploration strategy, a challenge that may lead to longer burn-in phase in algorithms based on Bernstein-style confidence sets.

In Table 3, we report the realized running times (in seconds) of the various algorithms for a time horizon of 10^5 steps. Note that the constraint in `EVI` for Bernstein-type algorithms is applied, implying that the running times are forcefully reduced for these algorithms. To illustrate the consequences of this choice, we use TSDE, which also suffers occasionally from problems of convergence in the VI. These problems are negligible in all environments except in the 6-state *RiverSwim-patrol2* —in our other experiments TSDE is loosely constrained to a maximum of 1000 iterations of VI, which in practice is almost never necessary. We display the running time of TSDE for the 6-state *RiverSwim-patrol2* with the constraint of 100 iterations and in parenthesis with a constraint of 1000 iterations. The resulting increase of a full order of magnitude is thus expected, which is similar to the behaviour observed with the Bernstein-style algorithms. Finally, we stress that our implementations are not optimised (even for python3). We believe more efficient implementations of these algorithms that would enjoy significantly reduced running times are possible.

Figure 13 displays the empirical gain (defined as $\frac{1}{t}\sum_{t'=1}^{t} r_{t'}$) of the various algorithms in the *RiverSwim* domains and 2-room MDPRMs, together with 95% confidence intervals. The horizontal line (in magenta) shows the optimal gain $g^\star$ achieved by the oracle. Overall, Figure 13 shows that the empirical gain (i.e., empirical per-step reward) under UCRL-RM-L1 quickly approaches $g^\star$ compared to the rest in all environments. These figures propose alternative representations to the various regret plots presented in this paper, but are based on the same experiments. Figure 13(a) demonstrates the superiority of the approach and the limitation of the baselines that failed to exhibit any significant learning curve in the allotted time horizon. Figure 13(b) shows once again the limitation of the Bernstein-type approach under practical constraints (as made explicit in the previous paragraph), where it is worth noting that the UCRL2-B baseline fails to gather any considerable reward until the end.

Through tables, we illustrate the practical values of the diameters and the associated leading terms of regret bounds of UCRL-RM-L1, UCRL-RM-B, UCRL2, and UCRL2-B(excluding the exact universal constants). Table 1 presents these values for different *RiverSwim* MDPRMs with progressive difficulty levels. As it shows, there is not a big difference between the RM-restricted diameter and D_{cp} due to the specific structure of *RiverSwim*. On the other hand, Table 2 shows similar values associated to the MDPRM shown in Figure 11 for various lengths N of the abnormal sub-task. Note that 2 actions are available in this MDPRM, both with the same transitions but one yielding no event. It is a relevant example in matter of diameters as it represents a simplification (for computational and illustrative purpose) of a situation where multiple sub-tasks are part of the RM, each with their own rewards. To compute the diameters used in these tables, the procedure detailed in Appendix E is applied.

We note that in all the reported experiments, we ran TSDE without using the knowledge of the mean rewards, contrary to the other algorithms. This is because in the case of deterministic rewards, TSDE exhibits a very unstable behaviour, which in turn would increase the realized regret significantly. In other words, ignoring the knowledge on mean reward rendered more beneficial for TSDE and we did so to attain a better empirical regret for it.

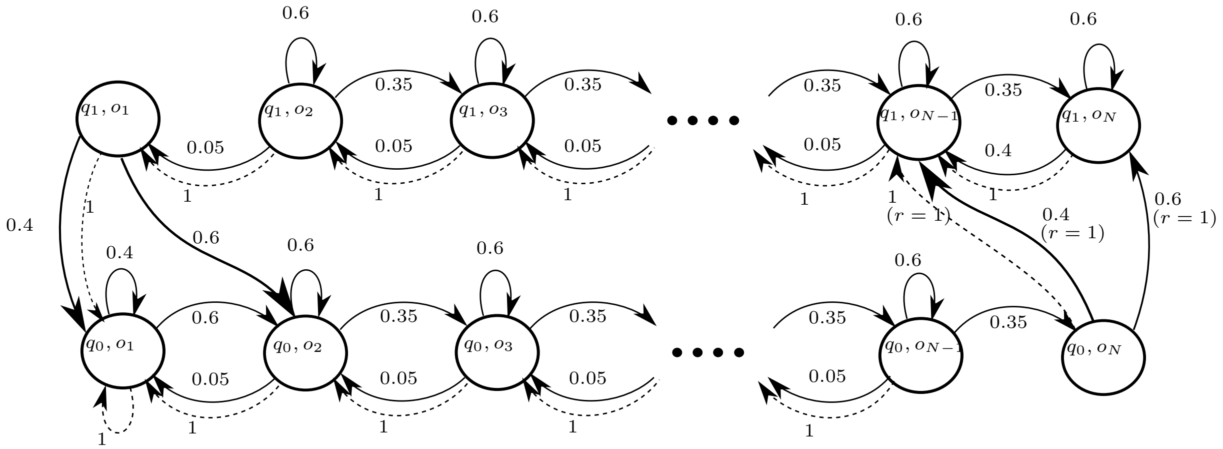


Figure 10: The cross-product MDP associated to the N-observation *RiverSwim* MDP with the *patrol2* RM

G COMPARISON OF REGRET BOUNDS

In this section, we present a more detailed comparison between the regret bounds of UCRL-RM variants and those of UCRL2 and UCRL2-B applied to M_{cp}, obliviously to the structure of the MDPRM. We exclude comparison to EBF introduced by Zhang and Ji (2019), whose regret matches the lower bound in tabular MDPs, as it does not admit an efficient implementation to the best of our knowledge.

O	$\sqrt{OA\mathbf{c}_M}$	$\sqrt{\mathbf{c}'_M}$	$D_{\textsf{cp}}\sqrt{\sum_{q,o,a} K_{(q,o),a}}$	$D_{\textsf{cp}}QO\sqrt{A}$
6	93.8	54.0	133.6	334.3
12	319.3	130.28	443.1	1551.1
20	726.0	229.6	1009.4	4542.5
40	2130.5	476.4	2978.0	18893.9
70	5005.4	846.1	7013.4	58783.2
100	8595.8	1215.6	12044.3	120745.6

Table 1: Various quantities related to the regret bounds for *RiverSwim* with *patrol2* RM with various number of observation states: Column 2 (UCRL-RM-L1), Column 3 (UCRL-RM-B), Column 4 (UCRL2B), Column 5 (UCRL2)

N	$\sqrt{OA\mathbf{c}_M}$	$\sqrt{\mathbf{c}'_M}$	$D_{\textsf{cp}}\sqrt{\sum_{q,o,a} K_{(q,o),a}}$	$D_{\textsf{cp}}QO\sqrt{A}$
4	468.0	272.4	3032.0	20339.3
5	468.1	272.4	3360.6	23100.5
6	468.2	272.5	3699.7	26029.4
8	504.2	293.2	4407.0	32384.4
10	550.1	319.5	5151.9	39404.6
12	600.2	348.3	5932.8	47090.0

Table 2: Various quantities related to the regret bounds for the *Multitask* MDPRM with various length N of the abnormal sub-task: Column 2 (UCRL-RM-L1), Column 3 (UCRL-RM-B), Column 4 (UCRL2B), Column 5 (UCRL2)

Algorithm	riverSwim6-patrol2	riverSwim20-patrol2
TSDE	3.9 (relaxed constraints: 113.1)	58.5
UCRL2	2.0	136.9
UCRL-RM-L1	1.7	48.05
UCRL-RM-B	1.8	39.8
UCRL2-B	3.0	45.6

Table 3: Empirical running times (in seconds) of various algorithms for a fixed time horizon of 10^5 steps

Regret Bounds of UCRL2 and UCRL2-B on the Cross-product MDP. UCRL2 (Jaksch et al., 2010) attains the following regret bound on $M_{\textsf{cp}}$ that holds with high probability:

$$\mathfrak{R}(\text{UCRL2}, T) = O\Big(D_{\textsf{cp}} OQ\sqrt{AT\log T}\Big)\,.$$

One may use improved confidence sets in UCRL2, similar to those used in UCRL-RM-L1. This improved variant of UCRL2 achieves a regret bound growing as

$$\mathfrak{R}(\text{UCRL2}_{\text{improved}}, T) = O\Big(D_{\textsf{cp}}\sqrt{AOQT(OQ+\log T)}\Big)\,.$$

UCRL2-B achieves the following regret bound on $M_{\textsf{cp}}$:

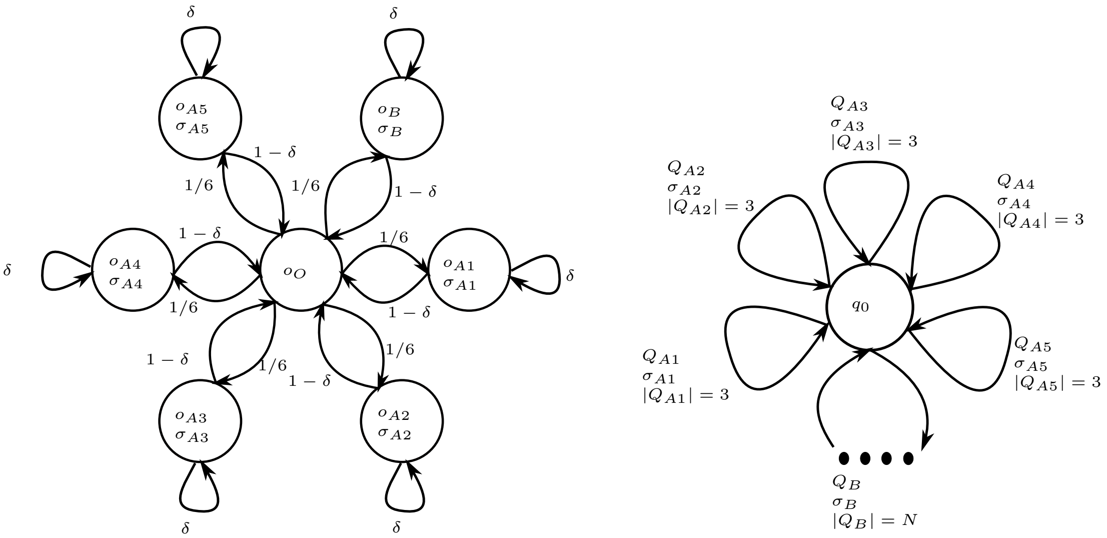
$$\mathfrak{R}(\text{UCRL2-B}, T) = O\Big(D_{\textsf{cp}}\sqrt{QT\log T\sum\nolimits_{o,a} K_{o,a}}\Big)$$

Regret under UCRL-RM-L1. By Theorem 1, UCRL-RM-L1 achieves the following regret bound on MDPRM M:

$$\mathfrak{R}(\texttt{UCRL-RM-L1}, T) = O\Big(\sqrt{\mathbf{c}_M AT\big(O+\log T\big)}\Big)$$

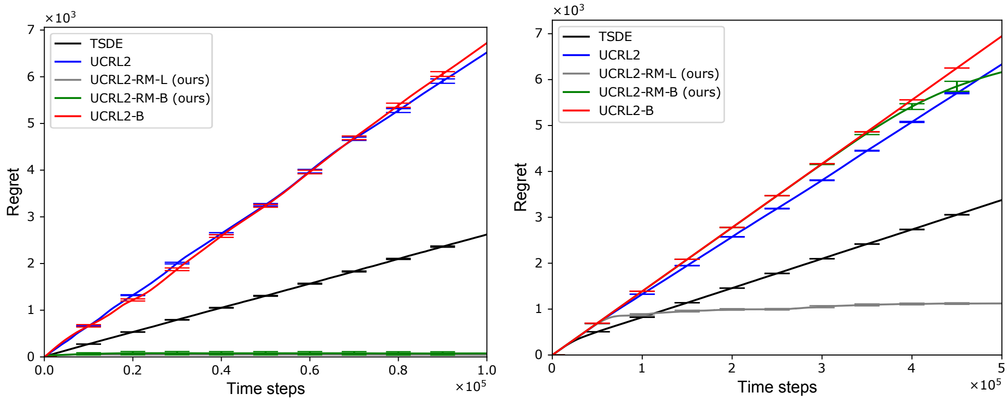
where $\mathbf{c}_M = \sum\limits_{o\in\mathcal{O}} \max\limits_{q\in\mathcal{Q}} D^2_{q,o}$. In view of $D_s \le D_{\textsf{cp}}$, $\mathbf{c}_M \le OD^2_{\textsf{cp}}$, the regret of UCRL-RM-L1, in the worst case grows as:

$$\mathfrak{R}(\texttt{UCRL-RM-L1}, T) = O\Big(D_{\textsf{cp}}\sqrt{OAT\big(O+\log T\big)}\Big)\,.$$



(a) MDP (b) RM

Figure 11: The *Multitask* MDPRM



(a) 6-state (b) 20-state

Figure 12: Regret in 6-state and 20-state *RiverSwim*

However, in some specific instances, we have $D_s \lesssim D_{\textsf{cp}}/Q$ for all s, so that $\mathbf{c}_M \sim O(D_{\textsf{cp}}/Q)^2$ for such M. On such MDPRMs, we have

$$\mathfrak{R}(\texttt{UCRL-RM-L1}, T) = O\Big(\tfrac{D_{\textsf{cp}}}{Q}\sqrt{OAT(O + \log T)}\Big)\,.$$

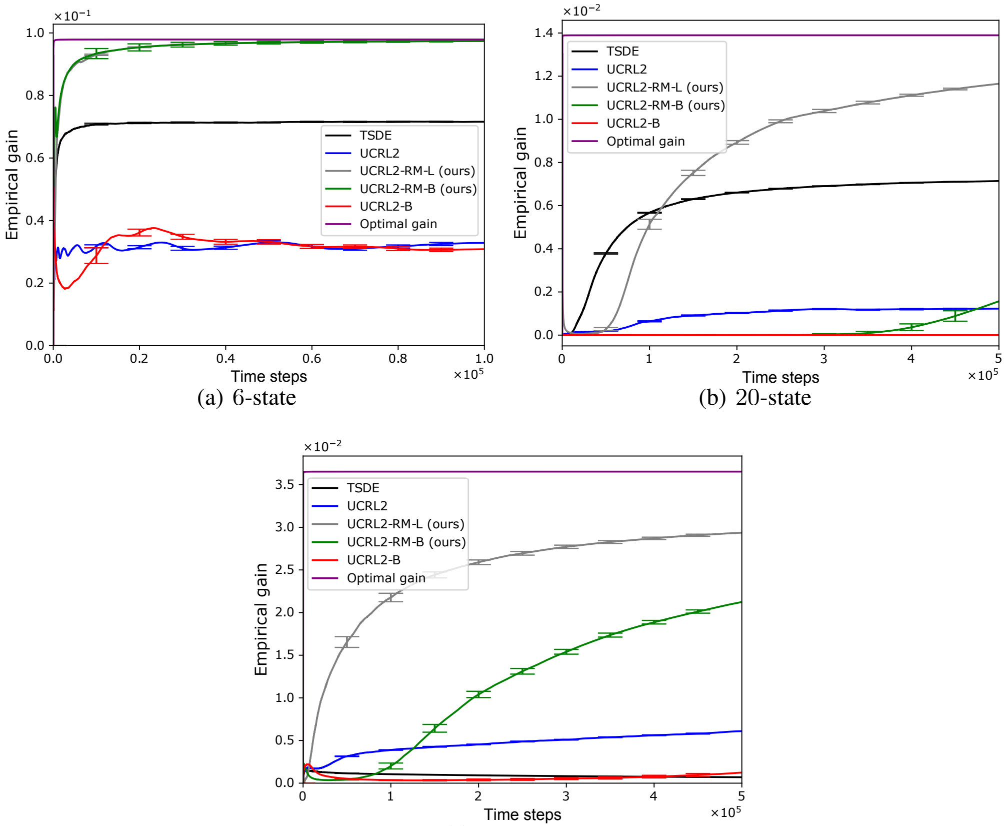
In comparison with the *improved* regret bounds of UCRL2, `UCRL-RM-L1` achieves a regret that is smaller by a factor of, at least, $\sqrt{Q}$. However, in instances where $D_s \lesssim D_{\textsf{cp}}/Q$, `UCRL-RM-L1` improves the regret bound of UCRL2 by a factor of $Q^{3/2}$. (The improvements are higher had we used the classical regret bound of UCRL2.)

Regret under `UCRL-RM-B`. By Theorem 2, `UCRL-RM-B` achieves the following regret bound on MDPRM M:

$$\mathfrak{R}(\texttt{UCRL-RM-B}, T) = O\Big(\sqrt{\mathbf{c}'_M T \log\log T}\Big),$$

where $\mathbf{c}'_M = \sum_{o\in\mathcal{O}, a\in\mathcal{A}} K_{o,a} \max_{q\in\mathcal{Q}} D^2_{q,o}$. In view of $D_s \le D_{\textsf{cp}}$, $\mathbf{c}'_M \le D^2_{\textsf{cp}} \sum_{o,a} K_{o,a}$. Hence, the regret of `UCRL-RM-L1`, in the worst case grows as:

$$\mathfrak{R}(\texttt{UCRL-RM-B}, T) = O\Big(D_{\textsf{cp}}\sqrt{\sum\nolimits_{o,a} K_{o,a} T \log\log T}\Big)\,.$$



(a) 6-state (b) 20-state

(c) 2-room

Figure 13: Empirical gain in 6-state and 20-state *RiverSwim* and the 2-room MDPRM

However, in some specific instances, we have $D_s \lesssim D_{\mathsf{cp}}/Q$ for all s (e.g., the one in Section 4, Figure 2), so that $\mathbf{c}'_M \lesssim (D_{\mathsf{cp}}/Q)^2 \sum_{o,a} K_{o,a}$ for such M. On such MDPRMs, we have

$$\mathfrak{R}(\texttt{UCRL-RM-B}, T) = O\Big(\frac{D_{\mathsf{cp}}}{Q}\sqrt{\sum\nolimits_{o,a} K_{o,a} T \log\log T}\Big)\,.$$

In comparison with UCRL2-B, `UCRL-RM-B` achieves a regret that is smaller by a factor of, at least, $\sqrt{Q}$. Moreover, in instances where $D_s \lesssim D_{\mathsf{cp}}/Q$, `UCRL-RM-B` achieves an improvement over UCRL2-B by a factor of $Q^{3/2}$.