Exploration in Reward Machines with Low Regret

Anonymous Authors[1]

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 & De Giacomo, 2019; Littman et al., 2017) that are more appropriately modeled as *Non-Markovian Decision Processes (NMDPs)*. 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.

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 subclasses of NMDPs. In Regular Decision Processes (RDPs) (Brafman & 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), a Deterministic Finite-State Automaton (DFA) whose transitions are labeled with high-level events and whose states compactly encode the entire history of observations. 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 of the transition dynamics and associated reward functions. The goal of the agent is to minimize its regret, which calls for balancing exploration and exploitation. We focus on an intermediate setting where the underlying 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 2.1), 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 & Ji, 2019)), but they are tailored to leverage the structure in MDPRMs; see Section 3.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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` and `UCRL-RM-B` achieve a regret growing as $\widetilde{O}(\sqrt{\mathbf{k}_M OAT})$, 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. Furthermore, $\mathbf{k}_M$ is an MDP-dependent and algorithm-dependent factor, which reflects the contribution of diameter-like quantities to the regret. In fact, $\mathbf{k}_M$ is defined in terms of a notion of diameter, called the RM-restricted diameter, which reflects the connectivity in M *jointly* determined by the dynamics and the sparsity structure of the reward machine. The RM-restricted diameter is always smaller than D_{cp}, and interestingly, in some instances of MDPRM, it is proportional to D_{cp}/Q. Hence, this novel notion refines the conventional diameter (Jaksch et al., 2010), and could be of interest in other settings of reward machines. The presented regret bound exhibits a two-fold improvement over those of baselines: (i) It is independent of the number Q of states of the reward machine, whereas the state-of-the-art existing bounds depend on $\sqrt{Q}$; and (ii) existing bound necessarily depend on D_{cp} or $\sqrt{D_{\mathsf{cp}}}$, the diameter of the cross-product MDP, whereas ours depend (via $\mathbf{k}_M$) on RM-restricted diameters of the various states. In summary, our regret bounds sometimes improve over the state-of-the-art by a factor of $Q^{3/2}$ —see 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.

1.2. Related Work

In the case of Markovian rewards and dynamics, there is a rich and growing literature on average-reward RL in finite tabular MDPs, where several algorithms with theoretical regret guarantees are presented (e.g., (Bartlett & Tewari, 2009; Burnetas & Katehakis, 1997; Fruit et al., 2018b; Jaksch et al., 2010; Ouyang et al., 2017; QIAN et al., 2019; Talebi & Maillard, 2018; Tossou et al., 2019; Wei et al., 2020; Bourel et al., 2020; Zhang & Ji, 2019; Ok et al., 2018)). In the absence of structure assumptions, as established in (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 & Ji, 2019) reports 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, some papers study RL in episodic MDPs; see, e.g., (Dann et al., 2017; Gheshlaghi Azar et al., 2017; Osband et al., 2013).

The focus of this paper is RL for the class of MDPRMs under the average-reward criterion, in an intermediate setting where the underlying DFA is *known*, while the actual transition distributions are *unknown*. Several authors propose algorithms with polynomial sample complexity or sublinear regret for different classes of NMDPs (Lattimore et al., 2013; Maillard et al., 2013; Sunehag & Hutter, 2015). However, even though 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 in our work. The algorithm S3M (Abadi & Brafman, 2020) integrates RL with the logical formulas of RDPs, but does not admit polynomial sample complexity in the PAC setting. (Ronca & De Giacomo, 2021) presents 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 a large number of researchers to the field. Specifically, there is previous work for proving convergence guarantees of RL (Toro Icarte et al., 2018; 2022), for studying the relationship to linear temporal logic (Camacho et al., 2019) and for investigating how to learn DFAs from traces (De Giacomo et al., 2020; Furelos-Blanco et al., 2021; Gaon & Brafman, 2020; Hasanbeig et al., 2021; Toro Icarte et al., 2019; Xu et al., 2020). RL with linear temporal logic, strongly related to DFAs, has been successfully applied to robotics tasks with non-Markovian reward (Shah et al., 2020). (Clark & Thollard, 2004) studies the learnability of probabilistic DFAs (PDFAs) in the PAC setting. However, we are not aware of any previous work involving reward machines that establishes high-probability 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 PDFAs,

hidden Markov models (HMMs) and POMDPs has been previously studied (Dupont et al., 2005).

Notations. We introduce notations that will be used throughout. 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. 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 background on Markov decision processes and reward machines.

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.[1] $\mathcal{P}$ denotes a set of atomic propositions and $L : \mathcal{O}\times\mathcal{A}\times\mathcal{O} \to 2^{\mathcal{P}}$ denotes a labeling function such that L assigns, to each triplet (o, a, o'), a subset of $\mathcal{P}$.

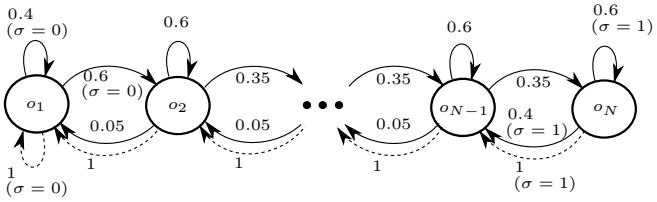
The notion of M above coincides with the conventional notion of average-reward MDPs except that (i) it assumes a *non-Markovian reward function* and (ii) it is equipped with a labeling mechanism (defined via L and $\mathcal{P}$). These labels describe high-level events associated to (o, a, o') triplets that can be detected from the environment. The interaction between the agent and the environment M proceeds as follows. Starting from some initial state $o_1 \in \mathcal{O}$ at time $t = 1$, 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 $h_t := (o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t)$. Upon executing a_t in o_t, the environment generates a next-state o_{t+1} sampled from $p(\cdot|o_t, a_t)$, and assigns a label $\sigma_t = L(o_t, a_t, o_{t+1})$ —note the dependence of h_t on $\sigma_1, \ldots, \sigma_{t-1}$ through the labeling function L. Then, the agent receives a random reward $r_t \sim \mathbf{R}(h_t, a_t)$ by the end of the current time step. The state of M then transits to o_{t+1} and a new decision step begins. As in conventional MDPs, after T steps of interactions, the agent's cumulative reward is $\sum_{t=1}^{T} r_t$.

[1] The bounded support $[0, 1]$ is with no loss of generality. This can be extended to σ-sub-Gaussian reward distributions with unbounded supports.

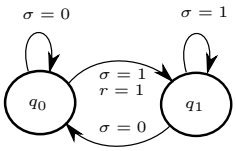
Reward Machines. In this paper we restrict attention to a class of non-Markovian reward functions that are encoded by Reward Machines (RMs) (Toro Icarte et al., 2018; 2022), whose definition coincides with conventional Deterministic Finite-State Automata (DFAs). An RM is a tuple $\mathcal{R} = (\mathcal{Q}, 2^{\mathcal{P}}, \tau, \nu)$, where $\mathcal{Q}$ is a finite set of states and $2^{\mathcal{P}}$ is an input alphabet. Furthermore, $\tau : \mathcal{Q}\times 2^{\mathcal{P}} \to \mathcal{Q}$ 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 $\mathrm{r} : \mathcal{O}\times\mathcal{A} \to \Delta_{[0,1]}$. This is very similar to the standard definition of reward machine (Toro Icarte et al., 2022), though in our case the set of terminal states is empty. In simple 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 $\mathrm{r}_t = \nu(q_t, \sigma_t)$, where $\mathrm{r}_t : \mathcal{O}\times\mathcal{A} \to \Delta_{[0,1]}$ only depends on the current state q_t and current label σ_t. Conditioned on (q_t, σ_t), r_t is independent of earlier labels and states $(q_1, \sigma_1, \ldots, q_{t-1}, \sigma_{t-1})$. 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 $\mathrm{r}_t = \nu(q_t, \sigma_t)$ which is a mapping $\mathrm{r}_t : \mathcal{O}\times\mathcal{A} \to \Delta_{[0,1]}$. Then, the agent receives a reward $r_t \sim \mathrm{r}_t(o_t, a_t)$ (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 time step begins.

Figure 1 illustrates the MDP and RM components, respectively, of an example MDPRM based on the popular *River-Swim* domain (Strehl & Littman, 2008). The labeled MDP has N observations $o_1, \ldots, o_N$, two actions `right` (solid arrows) and `left` (broken arrows), and two propositions $\sigma = 0$ and $\sigma = 1$. Action `left` always succeeds, while action `right` sometimes fails, and may even move backwards. Label $\{\sigma = 0\}$ is observed when applying any action in o_1, while label $\{\sigma = 1\}$ is similarly observed in o_N. The *patrol2* RM simulates a patrolling task with two



(a) The *RiverSwim* Labeled MDP.



(b) The *patrol2* RM.

Figure 1. The N-state labeled *RiverSwim* MDP, and the *patrol2* RM.

checkpoints, providing a stochastic reward each time the agent completes a cycle of $\{\sigma = 0\}$ and $\{\sigma = 1\}$, i.e. the agent has to repeatedly visit the left-most and right-most observations of the MDP. We remark that the current MDP observation is not sufficient to predict what to do next, and therefore has to be combined with the current RM state.

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

Lemma 2.1. *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 $M_{\mathsf{cp}} = (\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S} = \mathcal{Q} \times \mathcal{O}$, and where for $s = (q, o), s' = (q', o') \in \mathcal{S}$ and $a \in \mathcal{A}$,*

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

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

In view of equivalence between M and M_{cp}, one could apply any off-the-shelf algorithm to M_{cp}, as it perfectly adheres to the Markovian property. In fact, M_{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, RL in MDPRMs involves an agent who is seeking to maximize her cumulative reward, and her performance is measured in terms of regret with respect to an oracle algorithm being aware of a gain-optimal policy. To formally define regret, we introduce some necessary concepts. A stationary deterministic policy in an MDPRM M is a mapping $\pi : \mathcal{Q} \times \mathcal{O} \to \mathcal{A}$, such that for all pairs $(q, o) \in \mathcal{Q} \times \mathcal{O}$, it prescribes an action $\pi(q, o) \in \mathcal{A}$. Let Π be the set of all stationary deterministic 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 $r_t \sim \mathrm{r}_t(o_t, a_t)$ and $\mathrm{r}_t = \nu(q_t, L(o_t, \pi(q_t, o_t), o_{t+1}))$ 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 maximal 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 2.1), 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*[2] 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 $r_t \sim \mathrm{r}_t(o_t, a_t)$). 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 objective of the learner is to minimize regret, which entails balancing exploration and exploitation. 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, where the notion of diameter is defined in (Jaksch et al., 2010).[3] We assume that the associated M_{cp} is communicating. (M_{cp} is defined in Lemma 2.1).[4]

[2]This implies that both τ and ν are known. The assumption of knowing ν can be easily relaxed at the expense of a slightly larger multiplicative factor in our regret bounds; see the discussion in Appendix A.

[3]The diameter D of an MDP M is $D = \max_{s \neq s'} \min_\pi \mathbb{E}[T^\pi(s, s')]$, where $T^\pi(s, s')$ denotes the number of steps it takes to get to s' starting from s and following policy π (Jaksch et al., 2010).

[4]Assuming that $\mathcal{R}$ and M are both communicating is not sufficient to guarantee that M_{cp} is communicating.

n summary, we impose the following assumption on the considered MDPRM:

Assumption 2.2. We assume: (i) the RM $\mathcal{R}$ is known, and (ii) M_{cp} is communicating.

3. Learning Algorithms for MDPRM

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 & Ji, 2019).

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; Jaksch et al., 2010) and is defined as follows:

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

where for $n \in \mathbb{N}$,

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

By construction, this confidence set guarantees that uniformly for all t, $p \in C^1_{t,\delta/OA}$, with probability at least $1-\delta$, that is, $\mathbb{P}(\exists t \in \mathbb{N} : p \notin C^1_{t,\delta/OA}) \leq \delta$ (Lattimore & Szepesvári, 2020; Maillard, 2019). The second one is based on Bernstein's inequality (combined with a peeling technique) and is defined as follows:

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

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

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 $\mathrm{r} = \nu(q,\sigma)$. For all $s = (q,o)$, $s' = (q',o')$, and a,

$$\widehat{P}_t(s'|s,a) = \widehat{p}_t(o'|o,a)\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}},$$
$$\widehat{R}_t(s,a) = \sum_{o'} \widehat{p}_t(o'|o,a)\overline{\nu}(q, L(o,a,o')).$$

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 $t \geq 1$ and for a confidence parameter $\delta \in (0,1)$:

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

where $C = C^1_{t,\delta/OA}$ or $C = C^2_{t,\delta/O^2A}$. 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, $\mathbb{P}(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}) \leq 2\delta$, as formalized in Lemma C.1 in the Appendix C. This crucially 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 $p' \in C^2_{t,\delta/O^2A}$.

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 variant 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 variant when necessary.

UCRL−RM variants implement 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 $C^1_{t,\delta}$ (UCRL−RM−L1) or $C^2_{t,\delta}$ (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 speaking, 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 $\pi \in \Pi_{M'}$ in MDPRM M'. As in UCRL2 this optimization problem can be efficiently solved via a variant of the EVI algorithm of (Jaksch et al., 2010) (see Algorithm 2 in Appendix A). In each iteration of EVI, the algorithm has to solve, for each (q, o, a), the following problem:

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

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 $C(o,a) = \cap_{o'} C^2_{t,\delta/O^2A}(o,a)$. EVI (in Algorithm 2 in Appendix A) returns a policy π_k, which is guaranteed to be $\frac{1}{\sqrt{t_k}}$-optimal. The algorithm commits to π_k for $t \geq t_k$ until the number of observations on some pair (o,a) is doubled. In other words, the sequence $(t_k)_{k \geq 1}$ satisfies: $t_1 = 1$, and for $k \geq 1$,

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

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 $C^1_{t,\delta}$ (resp. $C^2_{t,\delta}$). 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 bounds in terms of regret, 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.

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

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

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 $\pi_k = \texttt{EVI}\big(C, \frac{1}{\sqrt{t_k}}\big)$ — see Algorithm 2 in Appendix A.
      (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 $a_t = \pi_k(q_t, o_t)$, and receive the next state o_{t+1} and reward $r_t(q_t, L(o_t, \pi_k(q_t, o_t), o_{t+1}))$
      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

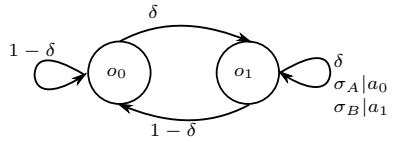
Intuitively, for a given $s = (q,o)$, $\mathcal{B}_s \subseteq \mathcal{Q}$ collects all possible next-states of the RM that can be reached via the *detectable events* in o. In the worst-case $\mathcal{B}_s = \mathcal{Q}$ for some state $s \in \mathcal{S}$. However, many high-level tasks in practice often admit RMs with sparse structures, in which there are states s for which $\mathcal{B}_s$ 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 4.1 (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, the *RM-restricted diameter* of MDPRM M for s, is defined as

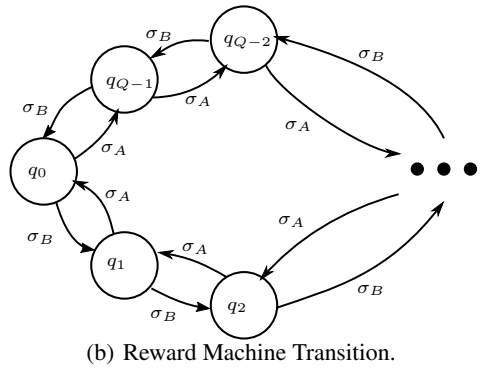
$$D_s := \max_{s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}} \min_\pi \mathbb{E}[T^\pi(s_1, s_2)].$$

Replacing $\mathcal{B}_s$ with $\mathcal{Q}$ in Definition 4.1, one recovers the diameter of M_{cp}, i.e., $D_s = D_{\mathsf{cp}}$. 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 small) subset of $\mathcal{Q}$, D_s is therefore a problem-dependent refinement of D_{cp}. We remark that a (cardinality-wise) small $\mathcal{B}_s$ does not necessarily imply that $D_s \ll D_{\mathsf{cp}}$ as D_s is determines 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.

Consider the MDPRM shown in Figure 4, 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



(a) The underlying labeled MDP.



(b) Reward Machine Transition.

Figure 2. Illustrating example for the RM-restricted diameter D_s.

counter-clockwise directions, respectively. As detailed in the Appendix E, we can show that: for all $q \in \mathcal{Q}$,

$$D_{o_1,q} = \tfrac{2}{\delta} + 1 + \tfrac{\delta}{1-\delta}, \quad D_{o_0,q} = \tfrac{1}{\delta},$$

whereas $D_{\mathsf{cp}} = \frac{\lfloor Q/2 \rfloor}{\delta} + 1 + \frac{\delta}{1-\delta}$ So, while D_{cp} grows with $\frac{Q}{\delta}$, D_s for all $s \in \mathcal{S}$ remain constant and only grow with $\frac{1}{\delta}$. In summary, we have $D_s \lesssim D_{\mathsf{cp}}/Q$. Another example with numerically computed D_s is provided in the Appendix E.

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 $C^1_{t,\delta}$:

Theorem 4.2 (Regret of `UCRL-RM-L1`)**.** *Under `UCRL-RM-L1`, uniformly over all $T \geq 2$, with probability higher than $1 - 4\delta$,*

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

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

In the following theorem, we present a regret bound for `UCRL-RM-B` (constructed using $C^2_{t,\delta}$). To this end, for $(o,a) \in \mathcal{S}$, we define $K_{o,a}$ as 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 4.3 (Regret of `UCRL-RM-B`)**.** *Under `UCRL-RM-B`, uniformly over all $T \geq 2$, with probability higher than $1 - 5\delta$,*

$$\mathfrak{R}(T) \leq O\Big(\sqrt{\mathbf{c}'_M T \log(O^2 A \log(\log(T))/\delta)} + D_{\mathsf{cp}} OA \log(T)\Big),$$

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

We remark that both regret bounds depend on D_{cp} *only* via a logarithmic term. 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[5]

$$O\Big(D_{\mathsf{cp}}\sqrt{T \log(\log(T)) \sum_{o,a} K_{o,a})}\Big).$$

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.

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 in (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.

[5]The regret of UCRL2-B can be improved to $O\Big(\log(T)\sqrt{TD_{\mathsf{cp}} \sum_{o,a} K_{o,a}}\Big)$ as reported in (Fruit et al., 2020), and the same improvement can carry over to `UCRL-RM-B`. We exclude comparisons to the regret of EBF (Zhang & Ji, 2019) growing as $O(\sqrt{D_{\mathsf{cp}}QOAT \log(T)})$, as it does not admit an efficient implementation.

In the following theorem, we present a regret lower bound that holds for any MDPRM M with a communicating cross-product M_{cp}.

Theorem 5.1. *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 observation states, A actions, Q RM states, and diameter D_{cp} of the associated M_{cp}, such that the regret of any algorithm $\mathbb{A}$ on these MDPRMs satisfies*

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

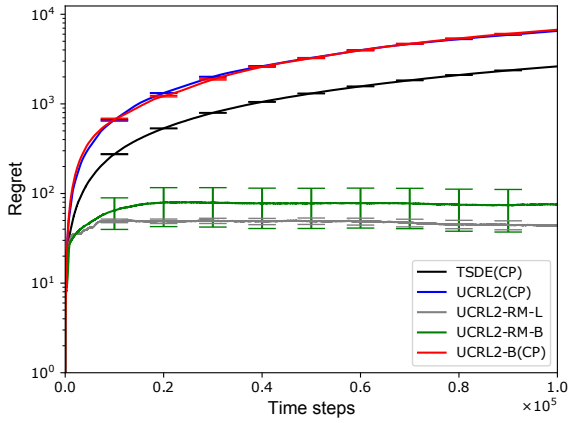
where $c_0 > 0$ is a universal constant.

This theorem asserts a *worst-case* regret lower bound growing as $\Omega(\sqrt{D_{\mathsf{cp}}OAT})$. To establish this result, we carefully construct an instance of MDPRM (with O observation states, A actions, and with an RM with Q states). 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, whose structure challenges exploration, whereas for p we inspire from the worst-case MDPs presented in (Jaksch et al., 2010). We finally remark that the lower bound does not contradict our regret bounds, as for the worst-case instances considered $\max_q D_{q,o} \simeq D_{\mathsf{cp}}$.

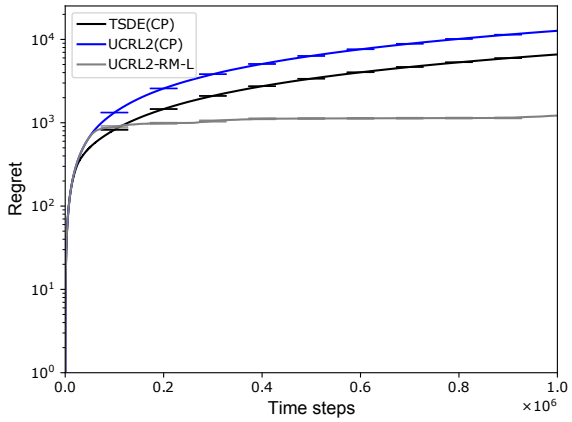
6. Experiments

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). To make the comparison fair, for UCRL2 and UCRL2B, we used improved confidence sets defined similarly to $C^1_{t,\delta}$ and $C^2_{t,\delta}$, respectively. Due to space constraint, we report the results for the *RiverSwim* MDPRM (Figure 1). All necessary details regarding the reported experimental results below together with additional results are provided in Appendix E.

Figure 3(a) shows the results of various algorithms in a 6-state RiverSwim MDPRM, where all results are averaged over 200 runs, shown together with 95% confidence intervals. Figure 3(b) displays the regret of various algorithms, but for a 20-state RiverSwim MDPRM, where all results are averaged over 100 runs. As the two figures reveal, both variants of `UCRL-RM` significantly outperform all the baselines, implying that the empirical gain in terms of regret due to exploiting the structure in RM is significant. 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), hence it was excluded.



(a) 6-state



(b) 20-state

Figure 3. Experimental results for a 6-state and 20-state *RiverSwim* with *patrol2* RM. (Note the logarithmic y-axis.)

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, `UCRL-RM-L1` and `UCRL-RM-B`, tailored to leverage the structure of MDPRMs, and analysed their regret. As demonstrated using the derived regret upper bounds and in numerical experiments, our 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.

References

Abadi, E. and Brafman, R. I. Learning and solving regular decision processes. In *29th International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 1948–1954. International Joint Conferences on Artificial Intelligence, 2020.

Bacchus, F., Boutilier, C., and Grove, A. Rewarding behaviors. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1160–1167, 1996.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 35–42, 2009.

Bourel, H., Maillard, O.-A., and Talebi, M. S. Tightening exploration in upper confidence reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1056–1066, 2020.

Brafman, R. I. and De Giacomo, G. Regular decision processes: A model for non-markovian domains. In *IJCAI*, pp. 5516–5522, 2019.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

Camacho, A., Toro Icarte, R., Klassen, T. Q., Valenzano, R. A., and McIlraith, S. A. LTL and beyond: Formal languages for reward function specification in reinforcement learning. In *IJCAI*, volume 19, pp. 6065–6073, 2019.

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

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pp. 5711–5721, 2017.

De Giacomo, G., Favorito, M., Iocchi, L., and Patrizi, F. Imitation learning over heterogeneous agents with restraining bolts. In *International Conference on Automated Planning and Scheduling*, pp. 517–521, 2020.

Dupont, P., Denis, F., and Esposito, Y. Links between probabilistic automata and hidden markov models: probability distributions, learning models and induction algorithms. *Pattern recognition*, 38(9):1349–1371, 2005.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems 31*, pp. 2994–3004, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1578–1586, 2018b.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of UCRL2 with empirical Bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Furelos-Blanco, D., Law, M., Jonsson, A., Broda, K., and Russo, A. Induction and exploitation of subgoal automata for reinforcement learning. *Journal of Artificial Intelligence Research*, 70:1031–1116, 2021.

Gaon, M. and Brafman, R. Reinforcement learning with non-Markovian rewards. In *AAAI Conference on Artificial Intelligence*, pp. 3980–3987, 2020.

Gheshlaghi Azar, M., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 263–272, 2017.

Hasanbeig, M., Jeppu, N., Abate, A., Melham, T., and Kroening, D. Deepsynth: Automata synthesis for automatic task segmentation in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 7647–7656, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Hutter, M., and Sunehag, P. The sample-complexity of general reinforcement learning. In *International Conference on Machine Learning*, pp. 28–36, 2013.

Littman, M. L., Topcu, U., Fu, J., Isbell, C., Wen, M., and MacGlashan, J. Environment-independent task specifications via gltl. *arXiv preprint arXiv:1704.04341*, 2017.

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

Maillard, O.-A., Nguyen, P., Ortner, R., and Ryabko, D. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International Conference on Machine Learning*, pp. 543–551, 2013.

Ok, J., Proutiere, A., and Tranos, D. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Osband, I., Russo, D., and Van Roy, B. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*, pp. 3003–3011, 2013.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown Markov decision processes: A Thompson Sampling approach. In *Advances in Neural Information Processing Systems 30*, pp. 1333–1342, 2017.

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

QIAN, J., Fruit, R., Pirotta, M., and Lazaric, A. Exploration bonus for regret minimization in discrete and continuous average reward MDPs. In *Advances in Neural Information Processing Systems 32*, pp. 4891–4900, 2019.

Ronca, A. and De Giacomo, G. Efficient pac reinforcement learning in regular decision processes. *arXiv preprint arXiv:2105.06784*, 2021.

Shah, A., Wadhwania, S., and Shah, J. Interactive robot training for non-Markov tasks. *arXiv preprint arXiv:2003.02232*, 2020.

Sondik, E. J. *The Optimal Control of Partially Observable Markov Processes*. PhD thesis, Stanford University, 1971.

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

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

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 770–805, 2018.

Toro Icarte, R., Klassen, T., Valenzano, R., and McIlraith, S. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pp. 2107–2116. PMLR, 2018.

Toro Icarte, R., Waldie, E., Klassen, T., Valenzano, R., Castro, M., and McIlraith, S. Learning reward machines for partially observable reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

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

Tossou, A., Basu, D., and Dimitrakakis, C. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pp. 10170–10180. PMLR, 2020.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Technical Report*, 2003.

Xu, Z., Gavran, I., Ahmad, Y., Majumdar, R., Neider, D., Topcu, U., and Wu, B. Joint inference of reward machines and policies for reinforcement learning. In *International Conference on Automated Planning and Scheduling*, pp. 590–598, 2020.

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

A. Further Algorithmic Details

A.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` in (Jaksch et al., 2010).

Algorithm 2 `EVI`(C, ε)

Initialize: $u^{(0)} \equiv 0, u^{(-1)} \equiv -\infty, n = 0$
while $\max_{s\in\mathcal{S}}(u^{(n)} - u^{(n-1)})(s) - \min_x(u^{(n)} - u^{(n-1)})(s) > \varepsilon$ **do**
  Get $\widetilde{p}^q$ for all $q \in \mathcal{Q}$ using `MAXP` (Algorithm 3 for `UCRL-RM-L1`, Algorithm 4 for `UCRL-RM-B`)
  For all $(q, o) \in \mathcal{S}$, update:

$$u^{(n+1)}(q, o) = \max_{a\in\mathcal{A}} \sum_{(q',o')\in\mathcal{S}} \widetilde{p}^q(o'|o, a)\Big[\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}(q', o')\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}}\Big]$$

  Set $n = n + 1$
end while
Output:

$$\pi_{n+1}(s) = \operatorname{argmax}_{a\in\mathcal{A}} \sum_{(q',o')\in\mathcal{S}} \widetilde{p}^q(o'|o, a)\Big[\overline{\nu}\big(q, L(o, a, o')\big) + u^{(n)}(q', o')\mathbb{I}_{\{q'=\tau(q,L(o,a,o'))\}}\Big]$$

`EVI` relies on solving the maximization problem in (3) in each round: Algorithm 3 find a solution to problem (3) for $C = C^1_{t,\delta}$ (i.e., for `UCRL-RM-L1`), whereas Algorithm 4 does it for $C = C^2_{t,\delta}$ (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 & Brunskill, 2015)).

For all (q, o, a), set $\widetilde{p}^q(\cdot|o, a)$ as any optimizer to the inner maximization of the `EVI`2, resolved for each triplet (q, o, a) by the algorithms 3 and 4. Here the superscript q in $\widetilde{p}^q(\cdot|s, a)$ signifies that the optimistic transition probability depends on q.

Algorithm 3 `MAXP` for `UCRL-RM-L1`

Initialize: $\forall o' \in \mathcal{O}, \widetilde{p}^q(o'|o, a) = \widehat{p}(o'|o, a)$
$\forall o' \in \mathcal{O}, u_{(q,o,a)}(o') = u(\tau(q, L(o, a, o')), o')$
$o_{max} = \operatorname{argmax}_{o'} \mu(q, L(o, a, o'))u_{(q,o,a)}(o')$
$\widetilde{p}^q(o_{max}|o, a) = \max\{1, \widehat{p}(o_{max}|o, a) + \frac{1}{2}\beta_{N_t(o,a)}(\frac{\delta}{OA})\}$
$\mathcal{L} = \operatorname{argsort}_{o'} \mu(q, L(o, a, o'))u_{(q,o,a)}(o')$ and $l = 0$
while $\sum_{o'\in\mathcal{O}} \widetilde{p}^q(o'|o, a) > 1$ **do**
  $\widetilde{p}^q(L_l|o, a) = \max(0, \widetilde{p}^q(L_l|o, a) + 1 - \sum_{o'\in\mathcal{S}} \widetilde{p}^q(o'|o, a))$
  Set $l = l + 1$
end while
Output: $\widetilde{p}^q(\cdot|o, a)$

Algorithm 4 `MAXP` for `UCRL-RM-B`

Initialize: $\forall o' \in \mathcal{O}, \widetilde{p}^q(o'|o, a) = \min\{p' \in C^2_{t,\delta/O^2A}(o, a, o')\}$
$\forall o' \in \mathcal{O}, u_{(q,o,a)}(o') = u(\tau(q, L(o, a, o')), o')$
$\mathcal{L} = \operatorname{argsort}_{o'} \mu(q, L(o, a, o'))u_{(q,o,a)}(o')$ and $l = OA - 1$
while $\sum_{o'\in\mathcal{O}} \widetilde{p}^q(o'|o, a) < 1$ **do**
  $\widetilde{p}^q(L_l|o, a) = \min(\max(C^2_{t,\delta/O^2A}(o, a, L_l)), 1 - \sum_{o'\in\mathcal{O}} \widetilde{p}^q(o'|o, a))$
  $l = l - 1$
end while
Output: $\widetilde{p}^q(\cdot|o, a)$

A.2. Unknown Mean Rewards

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

$$C_{t,\delta}^{\text{reward}}(q,\sigma) = \left\{\lambda \in [0,1] : |\widehat{\nu}_t(\cdot|q,\sigma) - \lambda| \leq \beta_{N_t(q,\sigma)}(\delta)\right\}, \quad C_{t,\delta}^{\text{reward}} = \cap_{q,\sigma} C_{t,\delta}^{\text{reward}}(q,\sigma),$$

where $\widehat{\nu}_t(\cdot|q,\sigma)$ denotes the empirical mean reward built using $N_t(q,\sigma)$ observations collected from the reward distribution $\nu(q,\sigma)$. Here, for $n \in \mathbb{N}$, $\beta_n(\delta) = \sqrt{\frac{1}{2n}\left(1+\frac{1}{n}\right)\log\left(\sqrt{n+1}/\delta\right)}$.

Then, it suffices to replace $\overline{\nu}$ with the upper confidence set, that is, to replace $\overline{\nu}(q,\sigma)$ with $\widehat{\nu}_t(\cdot|q,\sigma) + \beta_{N_t(q,\sigma)}(\delta)$. The penalty due to this is to increase the regret bound by an additive logarithmic factor.

B. The Cross-Product MDP: Proof of Lemma 2.1

Lemma 2.1 (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 $M_{\mathsf{cp}} = (\mathcal{S}, \mathcal{A}, P, R)$, 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 $h_t := (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$, 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 $r \in [0,1]$:

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

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

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

Moreover, for any $r \in [0,1]$, we have

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

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. □

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 C.1. *For all $\delta \in (0,1)$, we have:*

$$(i) \qquad \mathbb{P}(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}(C^1)) \leq \delta\,,$$
$$(ii) \qquad \mathbb{P}(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}(C^2)) \leq 2\delta\,.$$

The proof of Lemma C.1, presented below, builds on the concentration inequalities collected in Section C.4.

Proof (of Lemma C.1). Part (i). Using Lemma C.9 gives: For any (o,a),

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

For $\mathcal{M}_{t,\delta}$ constructed using $C^1_{t,\delta/OA}$, we thus have,

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

Part (ii). Lemma C.10 ensures that for any (o,a,o'),

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

Hence, for $\mathcal{M}_{t,\delta}$ built using $C^2_{t,\delta/O^2A}$, we have

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

□

C.1. Proof of Theorem 4.2

The proof follows quite similar lines as in the proof of (Jaksch et al., 2010, Theorem 2). Let $\delta \in (0,1)$. 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 C.8, we deduce that

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

with probability at least $1-\delta$.

Denoting $\mu(s,a) := \sum_{o'} p(o'|o,a)\overline{\nu}(q, L(o,a,o'))$ for $s=(q,o)$, we have

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

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 $M \in \mathcal{M}_k$ (that is, the set of plausible MDPRMs contains the true model), and *bad* otherwise.

Control of the regret due to bad episodes. By Lemma C.1, the set $\mathcal{M}_k$ contains the true MDPRMs 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 (hence, $M \in \mathcal{M}_k$). Since $M \in \mathcal{M}_k$, the choice of π_k and $\widetilde{M}_k = (\mathcal{S}, \mathcal{A}, \widetilde{p}, \mathcal{R}, \mathcal{P}, \mathcal{L})$, we have that $g_k := g_{\pi_k}(\widetilde{M}_k) \ge g^\star - \frac{1}{\sqrt{t_k}}$. Hence, the regret accumulated in episode k satisfies:

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

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

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

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

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

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

$$\widetilde{\mu}_k(s,a) := \sum_{s' \in \mathcal{S}} \widetilde{p}_k^q(o'|o,a)\overline{\nu}(q, L(o,a,o'))\,.$$

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

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

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:

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

Combining this with (4) yields

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

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:

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

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

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

Putting these three bounds together, we thus get

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

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

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

Upper bound on L_1. We have

$$
\begin{aligned}
v_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k &= \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} v_k(s,\pi_k(s))\Big(\widetilde{P}_k(s'|s,\pi_k(s)) - P(s'|s,\pi_k(s))\Big)w_k(s') \\
&\le \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s)) \sum_{o'\in\mathcal{O}}\sum_{q'\in\mathcal{Q}} \Big(\widetilde{p}_k^q(o'|o,\pi_k(s)) - p(o'|o,\pi_k(s))\Big)\mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}} w_k(q',o') \\
&\le \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s)) \sum_{o'\in\mathcal{O}} \Big|\widetilde{p}_k^q(o'|o,\pi_k(s)) - p(o'|o,\pi_k(s))\Big| \cdot \max_{s'\in\mathcal{B}_{q,o}\times\mathcal{O}} \big|w_k(q',o')\big| \underbrace{\sum_{q'\in\mathcal{Q}} \mathbb{I}_{\{q'=\tau(q,L(o,\pi_k(s),o'))\}}}_{=1} \\
&\le \sum_{s\in\mathcal{S}} v_k(s,\pi_k(s)) \big\|\big(\widetilde{p}_k^q - p\big)(\cdot|o,\pi_k(s))\big\|_1 \cdot \max_{s'\in\mathcal{B}_{q,o}\times\mathcal{O}} \big|w_k(q',o')\big| \\
&\le \sum_{o\in\mathcal{O}}\sum_{q\in\mathcal{Q}} v_k(q,o,\pi_k(q,o))\beta_{N_k(o,\pi_k(q,o))}\big(\tfrac{\delta}{OA}\big)\cdot D_{q,o} \qquad (8) \\
&\le \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\sum_{q\in\mathcal{Q}} v_k(q,o,a)\cdot\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big)\cdot D_{q,o} \\
&\le \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big)\cdot \max_{q\in\mathcal{Q}} D_{q,o} \sum_{q\in\mathcal{Q}} v_k(q,o,a) \\
&\le \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big)\cdot \max_{q\in\mathcal{Q}} D_{q,o}\cdot v_k(o,a)\,, \qquad (9)
\end{aligned}
$$

where (8) follows from Lemma C.2, stated and proven in Section C.3. Combining (9) with (7) and summing over all good episodes, we obtain:

$$
\begin{aligned}
\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M\in\mathcal{M}_k\}} &\le \sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I})w_k\mathbb{I}_{\{M\in\mathcal{M}_k\}} + 2\sum_{k=1}^{m(T)}\sum_{o,a}\frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
&\quad + \sum_{k=1}^{m(T)}\sum_{o,a} v_k(o,a)\beta_{N_k(o,a)}\big(\tfrac{\delta}{OA}\big)\Big(2+\max_{q\in\mathcal{Q}} D_{q,o}\Big) \\
&= \sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I})w_k\mathbb{I}_{\{M\in\mathcal{M}_k\}} + 2\sum_{k=1}^{m(T)}\sum_{o,a}\frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
&\quad + \sum_{k=1}^{m(T)}\sum_{o,a} v_k(o,a)\sqrt{\frac{2}{N_k(o,a)}\Big(1+\frac{1}{N_k(o,a)}\Big)\log\Big(\frac{OA(2^O-2)}{\delta}\sqrt{N_k(o,a)+1}\Big)}\Big(2+\max_{q\in\mathcal{Q}} D_{q,o}\Big) \\
&= \underbrace{\sum_{k=1}^{m(T)} v_k(\mathbf{P}_k - \mathbf{I})w_k\mathbb{I}_{\{M\in\mathcal{M}_k\}}}_{L_2} + 2\sum_{k=1}^{m(T)}\sum_{o,a}\frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
&\quad + 2\sqrt{O+\log\big(OA\sqrt{T+1}/\delta\big)}\sum_{k=1}^{m(T)}\sum_{o,a}\frac{v_k(o,a)}{\sqrt{N_k(o,a)}}\Big(\max_{q\in\mathcal{Q}} D_{q,o}+2\Big)\,. \qquad (10)
\end{aligned}
$$

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). However, we provide a finer of control of such a sequence. 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 k with $M \in \mathcal{M}_k$, we have:

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

where $\mathbf{e}_i$ denotes a vector with the i-th element being 1 and the others being zero. Hence, $L_2 \leq \sum_{t=1}^{T} Z_t + m(T)D_{\mathsf{cp}}$. Let $\mathrm{supp}(\cdot)$ denote the support set of a distribution. Since $s_{t+1} \in \mathrm{supp}(P(\cdot|s, a))$, we have

$$
\begin{aligned}
|Z_t| &\leq \Big|\sum_{s'} P(s'|s_t, a_t)w_{k(t)}(s') - w_{k(t)}(s_{t+1})\Big| \\
&\leq \sum_{s'} P(s'|s_t, a_t)\big|w_{k(t)}(s') - w_{k(t)}(s_{t+1})\big| \\
&\leq \sum_{s'} P(s'|s_t, a_t) \max_{s\in\mathcal{S}, s'\in\cup_{a\in\mathcal{A}}\mathrm{supp}(P(\cdot|s,a))} |w_{k(t)}(s') - w_{k(t)}(s)| \\
&= \max_{s\in\mathcal{S}, s'\in\cup_{a\in\mathcal{A}}\mathrm{supp}(P(\cdot|s,a))} |w_{k(t)}(s') - w_{k(t)}(s)|\,.
\end{aligned}
$$

Note that

$$
\mathrm{supp}(P(\cdot|s, a)) = \{(q', o') \in \mathcal{S} : p(o'|o, a) > 0 \,\text{and}\, q' = L(o, a, o')\} \subseteq \mathcal{B}_s \times \mathcal{O}\,.
$$

Hence,

$$
|Z_t| \leq \max_{s\in\mathcal{S}, s'\in\mathcal{B}_s\times\mathcal{O}} |w(s') - w(s)| \leq \max_s D_s
$$

So, Z_t is bounded by $\frac{1}{2}\max_s D_s$, and also $\mathbb{E}[Z_t|s_1, a_1, \ldots, s_t, a_t] = 0$, so that $(Z_t)_{t\geq 1}$ is martingale difference sequence. Therefore, by Corollary C.8, we get:

$$
\mathbb{P}\left(\exists T : \sum_{t=1}^{T} Z_t \geq \max_s D_s\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\right) \leq \delta\,.
$$

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

$$
\begin{aligned}
L_2 &\leq \max_s D_s\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + m(T)D_{\mathsf{cp}} \\
&\leq \max_s D_s\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}}OA\log_2\left(\tfrac{8T}{OA}\right), \qquad (11)
\end{aligned}
$$

where we used Lemma C.5 to upper bound $m(T)$.

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

$$
\begin{aligned}
\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M\in\mathcal{M}_k\}} &\leq 2\sqrt{O + \log\left(OA\sqrt{T+1}/\delta\right)} \sum_{k=1}^{m(T)} \sum_{o,a} \left(\max_{q\in\mathcal{Q}} D_{q,o} + 2\right) \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \\
&\quad + 2\sum_{k=1}^{m(T)} \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} + \max_s D_s\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}}OA\log_2\left(\tfrac{8T}{OA}\right), \qquad (12)
\end{aligned}
$$

with probability higher than $1-\delta$ and uniformly over all $T \in \mathbb{N}$. Applying Lemma C.4 and using the Cauchy-Schwarz inequality:

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

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

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

Combining this with (12), and putting together, we have that with probability at least $1-4\delta$,

$$\begin{aligned}\mathfrak{R}(T) \leq\ & 2(\sqrt{2}+1)\sqrt{O + \log\big(OA\sqrt{T+1}/\delta\big)}\big(\sqrt{\mathbf{c}_M}+2\big)\sqrt{AT} + 2(\sqrt{2}+1)\sqrt{OAT} \\ & + (\max_s D_s + 1)\sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)} + D_{\mathsf{cp}} OA \log_2\big(\tfrac{8T}{OA}\big)\,,\end{aligned}$$

thus proving the theorem. □

C.2. Proof of Theorem 4.3

Let $\delta \in (0,1)$. Following the same steps as in the proof of Theorem 4.2, we have

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

with probability at least $1-\delta$, where Δ_k is defined similarly to the proof of Theorem 4.2. Furthermore, By Lemma C.1, with probability at least $1-2\delta$, $\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}_{\{M\notin\mathcal{M}_k\}=0}$.

Let's now focus on good episodes, i.e., episodes k where $M \in \mathcal{M}_k$. Similarly to the proof of Theorem 4.2, we have that

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

The first and third terms are bounded as in the proof of Theorem 4.2. However, the second term in the right-hand side is

bounded as follows:

$$
\begin{aligned}
\widetilde{\mu}_k(s,a) - \mu(s,a) &= \sum_{o' \in \mathcal{O}} \big(\widetilde{p}^q(o'|o,a) - p(o'|o,a)\big)\overline{\nu}(q, L(o,a,o')) \\
&\leq \sum_{o' \in \mathcal{O}} \big|\widetilde{p}^q(o'|o,a) - p(o'|o,a)\big| \\
&\leq \sum_{o' \in \mathcal{O}} \sqrt{\frac{2\widetilde{p}^q(o'|o,a)(1-\widetilde{p}^q(o'|o,a))}{N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{O^2 A}\big)} \\
&+ \sum_{o' \in \mathcal{O}} \sqrt{\frac{2p(o'|o,a)(1-p(o'|o,a))}{N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{O^2 A}\big)} + \frac{2}{3N_k(o,a)}\beta'_{N_k(o,a)}\big(\tfrac{\delta}{O^2 A}\big) \\
&\overset{\text{(a)}}{\leq} \sqrt{\beta'_T\big(\tfrac{\delta}{O^2 A}\big)} \sum_{o' \in \mathcal{O}} \sqrt{\frac{2\widehat{p}_k(o'|o,a)(1-\widehat{p}_k(o'|o,a))}{N_k(o,a)}} \\
&+ \sqrt{\beta'_T\big(\tfrac{\delta}{O^2 A}\big)} \sum_{o' \in \mathcal{O}} \sqrt{\frac{2p(o'|o,a)(1-p(o'|o,a))}{N_k(o,a)}} + \frac{4}{N_k(o,a)}\beta'_T\big(\tfrac{\delta}{O^2 A}\big) \\
&\overset{\text{(b)}}{\leq} \sqrt{8\beta'_T\big(\tfrac{\delta}{O^2 A}\big)\frac{K_{o,a}}{N_k(o,a)}} + \frac{4}{N_k(o,a)}\beta'_T\big(\tfrac{\delta}{O^2 A}\big) \qquad (13)
\end{aligned}
$$

where (a) follows from Lemma C.3, 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 4.2, we have

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

where w_k is defined as in the proof of Theorem 4.2.

Upper Bound on L_1. We have

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

where the last inequality follows from Lemma C.2.

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 (13), we obtain

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

The rest of the proof follows similar lines as in the proof of Theorem 4.2. The only difference is that the final bound holds with probability $1 - 5\delta$. □

C.3. Technical Lemmas

Lemma C.2. *For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have:*

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

Proof. We first show that for all $s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}$, we have $u_k^{(i)}(s_1) - u_k^{(i)}(s_2) \leq D_s$, which further implies

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

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 $u_k^{(i)}(s_1) \geq u_k^{(i)}(s_2) - D_s$, 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 C.3. *Consider x and y satisfying $|x - y| \leq \sqrt{2y(1-y)\zeta} + \zeta/3$. Then,*

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

Proof. By Taylor's expansion, we have

$$
\begin{aligned}
y(1-y) &= x(1-x) + (1-2x)(y-x) - (y-x)^2 \\
&= x(1-x) + (1-x-y)(y-x) \\
&\leq x(1-x) + |1-x-y| \left(\sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta \right) \\
&\leq x(1-x) + \sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta \,.
\end{aligned}
$$

Using the fact that $a \leq b\sqrt{a} + c$ implies $a \leq b^2 + b\sqrt{c} + c$ for nonnegative numbers a, b, and c, we get

$$
\begin{aligned}
y(1-y) &\leq x(1-x) + \tfrac{1}{3}\zeta + \sqrt{2\zeta\left(x(1-x) + \tfrac{1}{3}\zeta\right)} + 2\zeta \\
&\leq x(1-x) + \sqrt{2\zeta x(1-x)} + 3.15\zeta \\
&= \left(\sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta} \right)^2 + 2.65\zeta \,, \qquad (14)
\end{aligned}
$$

where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ valid for all $a, b \geq 0$. Taking square-root from both sides and using the latter inequality give the desired result:

$$\sqrt{y(1-y)} \leq \sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta} + \sqrt{2.65\zeta} \leq \sqrt{x(1-x)} + 2.4\sqrt{\zeta}\,.$$

□

Lemma C.4 ((Jaksch et al., 2010, Lemma 19),(Talebi & Maillard, 2018, Lemma 24))**.** *For any sequence of numbers $z_1, z_2, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$, it holds*

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

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

Lemma C.5 ((Jaksch et al., 2010, Proposition 18))**.** *The number $m(T)$ of episodes up to time $T \geq OA$ satisfies*

$$m(T) \leq OA \log_2\left(\tfrac{8T}{OA}\right).$$

C.4. Concentration Inequalities

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

We begin with the following definition:

Definition C.6 (Sub-Gaussian Observation Noise)**.** A sequence $(Y_t)_t$ has conditionally σ-sub-Gaussian noise if

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

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

Lemma C.7 (Time-Uniform Laplace Concentration for Sub-Gaussian Distributions)**.** *Let $Y_1, \ldots, Y_n$ 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*

$$\mathbb{P}\left(\exists n \in \mathbb{N}, \quad |\widehat{\mu}_n - \mu| \geq \sigma\sqrt{\left(1+\frac{1}{n}\right)\frac{2\ln\left(\sqrt{n+1}/\delta\right)}{n}}\right) \quad \leq \quad \delta\,.$$

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 C.7:

Corollary C.8 (Time-Uniform Azuma-Hoeffding Concentration Using Laplace)**.** *Let $(X_t)_{t\geq 1}$ 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*

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

Lemma C.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 C.9 (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 $(X_t)_{t\geq 1}$ 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 $\delta \in (0,1)$,*

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

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

Lemma C.10 (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 $\delta \in (0,1)$,*

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

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

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 C.10 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 $\delta \in (0,1)$,

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

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

D. Regret Lower Bound

In this section, we prove Theorem 5.1. 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 & 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 & Szepesvári, 2020, Chapter 38.7).

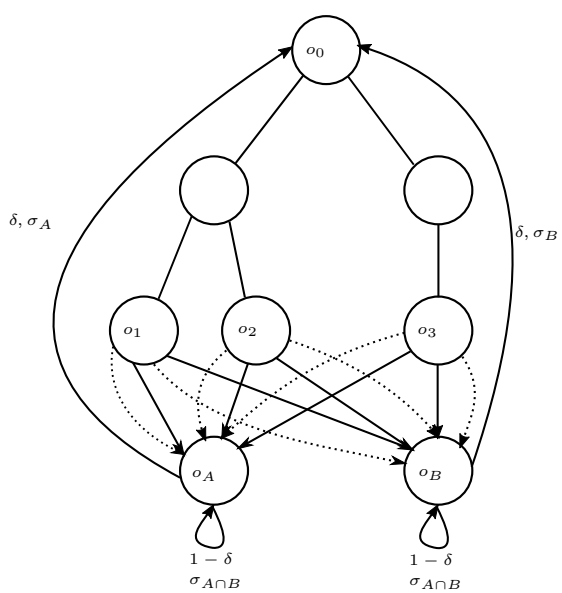


Figure 4. Construction of the underlying labeled MDP for the LB with $A = 2$ and $O = 8$, based on (Lattimore & Szepesvári, 2020, Chapter 38.7).

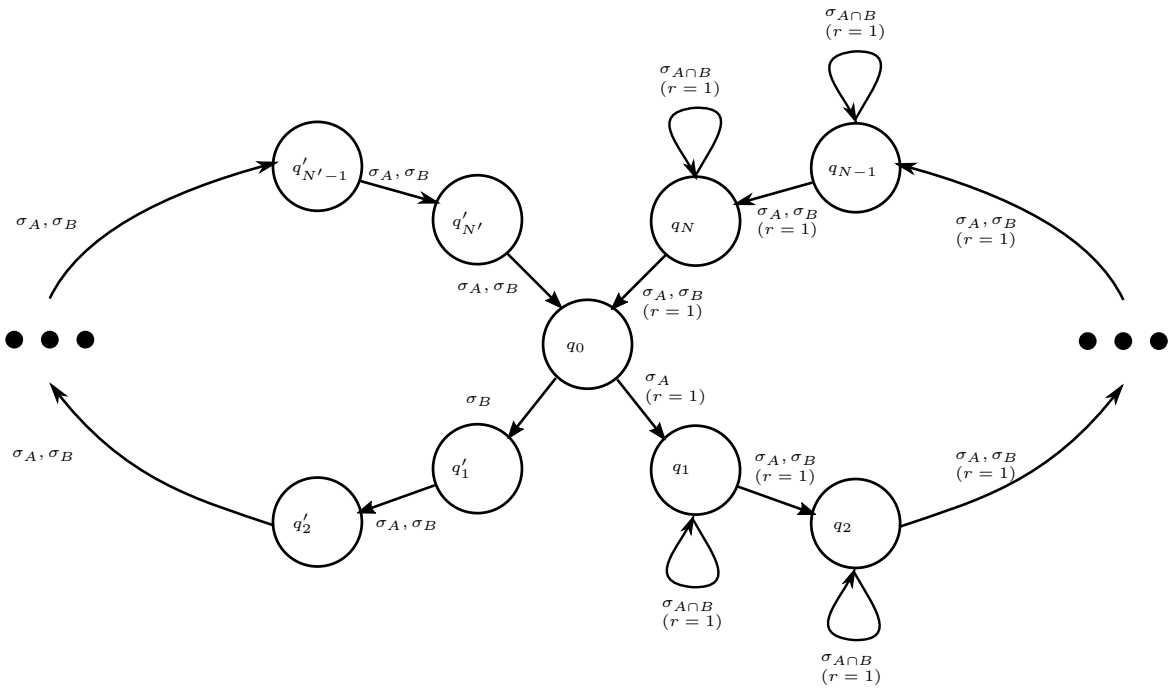


Figure 5. Construction of the underlying RM for the LB with a double-cyclic a 'good' cycle giving rewards and 'bad' cycle of similar length giving no reward.

Proof (of Theorem 5.1). 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 5 and 4 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 $o_1, o_2, \ldots, o_L$. 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 cross-product diameter of 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 $a \in \mathcal{A}$,

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

To build the RM, we let $N = \lceil (Q-1)/2 \rceil$ and $N' = \lfloor (Q-1)/2 \rfloor$. The idea is to divide the RM into 2 cycles of length N and N'. To this effect, we have q_0 the origin, then a set $\{q_i\}_{i=1}^{N}$ of RM states for the good cycle, and a set $\{q'_j\}_{j=1}^{N'}$ for the bad cycle. Then, we build the following RM transitions function τ and reward function r, for all $i \in 1, N$ and all $j \in 1, N'$

(see Figure 4):

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

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 progress on 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 on 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 $\mathcal{L}^M$:

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

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 & 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, to do so we first build upon (Lattimore & Szepesvári, 2020, Claim 38.9) that shows that the diameter of the underlying MDP of M_j, that will be written $D(M_j)$, 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 & 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 an 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 "episode" where a new episode start whenever reaching (q_0, o_0). This yields immediately, by construction and using our knowledge of the optimal policy:

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

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 allow 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 & 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 $\Delta \leq 1/4$, then using the entropy inequalities from (Lattimore & Szepesvári, 2020, Equation 14.16), we have:

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

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

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

Summing over j and applying Cauchy-Schwarz give us

$$\begin{aligned}\sum_{j=1}^{k} \mathbb{E}_j[T_{\text{tot}} - T_j] &\geq \sum_{j=1}^{k} \mathbb{E}_0[T_{\text{tot}} - T_j] - \frac{T\Delta}{D_{\mathsf{cp}}}\sum_{j=1}^{k}\sqrt{2\mathbb{E}_0[T_j]} \\ &\geq (k-1)\mathbb{E}_0[T_{\text{tot}}] - \frac{T\Delta}{D_{\mathsf{cp}}}\sqrt{2k\mathbb{E}_0[T_{\text{tot}}]} \\ &\geq \frac{c_1 T(k-1)}{D_{\mathsf{cp}}} - \frac{T\Delta}{D_{\mathsf{cp}}}\sqrt{\frac{2c_2 Tk}{D_{\mathsf{cp}}}}.\end{aligned}$$

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

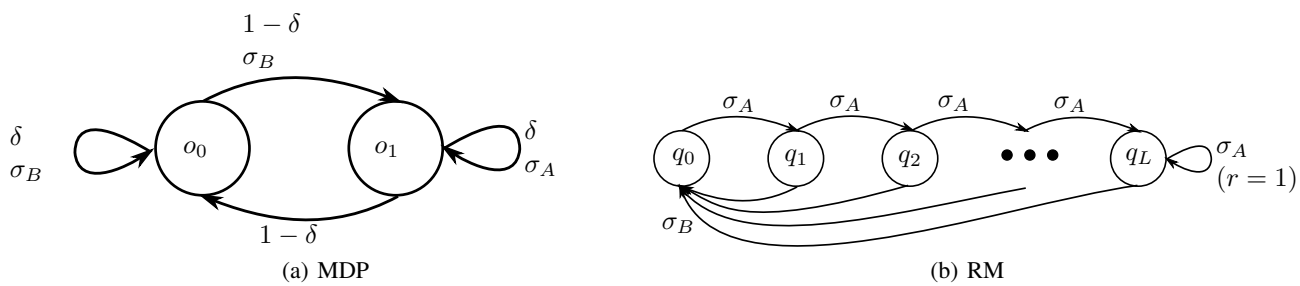
$$\sum_{j=1}^{k} \mathbb{E}_j[T_{\text{tot}} - T_j] \geq \frac{c_1 T(k-1)}{2kD_{\mathsf{cp}}}.$$

This implies that there exists j such that $\mathbb{E}_j[T_{\text{tot}} - T_j] \geq \frac{c_1 T(k-1)}{2kD_{\mathsf{cp}}}$, which leads to the final result using the previous lower bound on the regret

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

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

E. Details of Diameter Computations



(a) MDP (b) RM

Figure 6. Example $D_{\mathsf{cp}} \to \infty$ when $D \to 1$

Diameters Computation for the MDPRPM in Figure 4. 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, 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. We denote this number of step N then we give ourselves D_N the diameter over a communicating subset of $\mathcal{Q}$ with maximum number of steps between 2 steps being Nn we can then observe that:

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

then a simple recurrence show 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 conclude the analysis.

Diameters computation for example of Figure 6. 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 7 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 Figures 9(a) and 9(b) the same results as in section 6 but without the log-scale, which could be useful to better compare the standard deviation of various algorithms.

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 UCRL2B (excluding the exact universal constants and T). Table 1 presents these values for different *RiverSwim* MDPRMs with progressive difficulty levels. As the table 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 8 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.

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 our domains, rewards are deterministic for which TSDE exhibits a very unstable behaviour, which in turn would increase the realized regret significantly. In other words, we did so to attain a better empirical for TSDE.

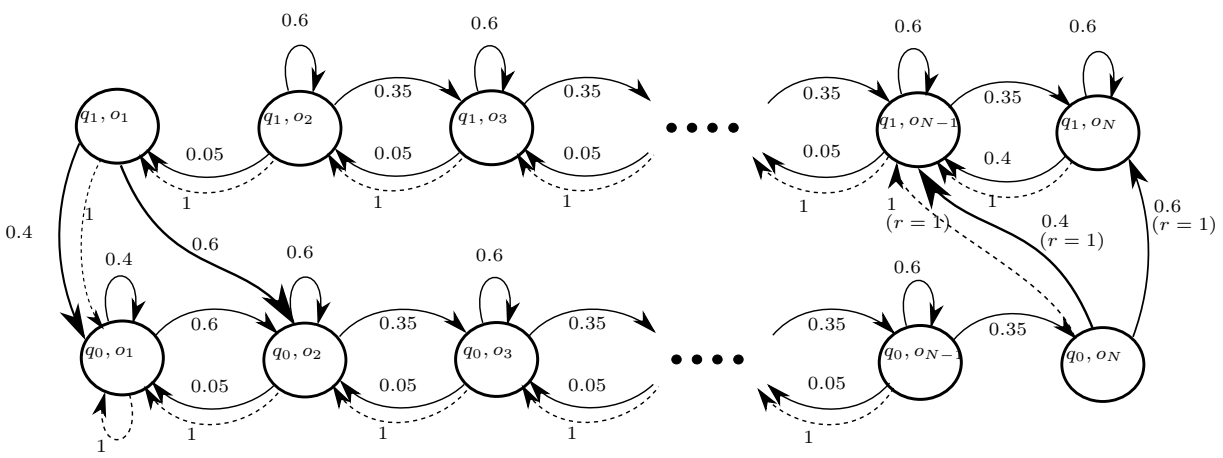


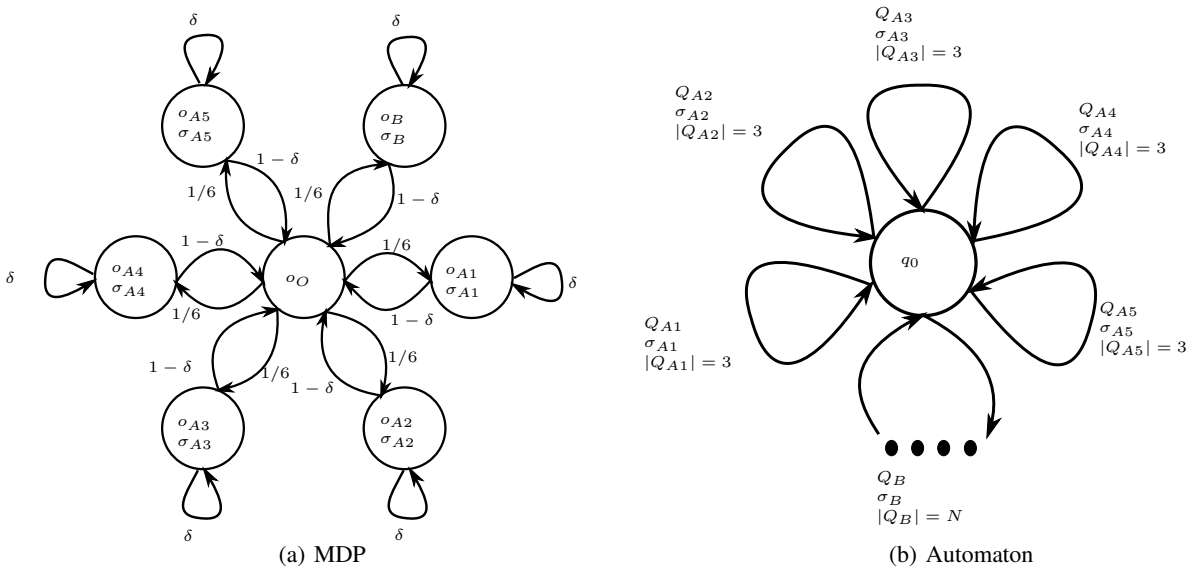
Figure 7. The cross-product MDP associated to the N-state *RiverSwim* MDP with the *patrol2* RM

O	$\sqrt{OA\mathbf{c}_M}$	$\sqrt{\mathbf{c}'_M}$	$D_{\mathsf{cp}}\sqrt{\sum_{q,o,a} K_{(q,o),a}}$	$D_{\mathsf{cp}}QO\sqrt{A}$
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).

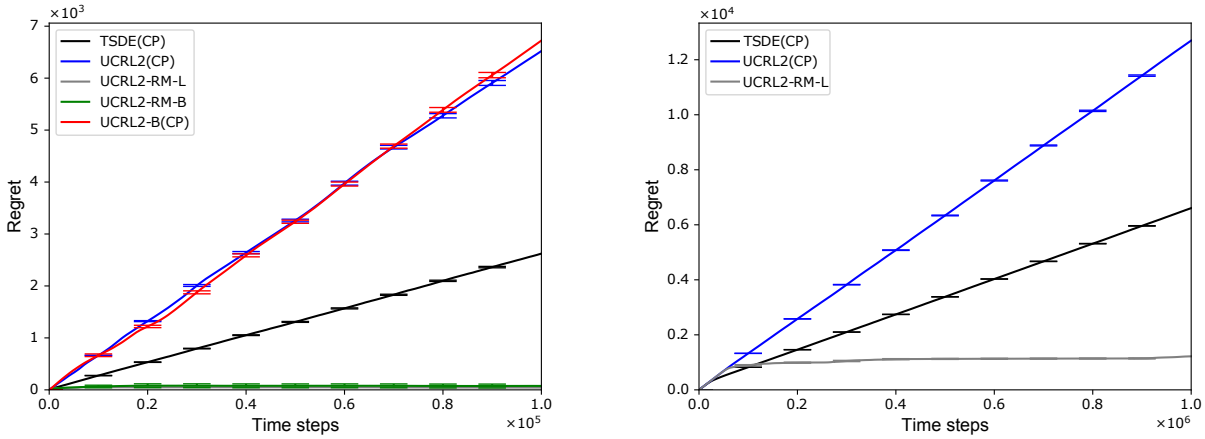
N	$\sqrt{OA\mathbf{c}_M}$	$\sqrt{\mathbf{c}'_M}$	$D_{\mathsf{cp}}\sqrt{\sum_{q,o,a} K_{(q,o),a}}$	$D_{\mathsf{cp}}QO\sqrt{A}$
4	468.0	272.4	3032.0	20339.3
5	468.1	272.4	3360.6	23100.5
6	468.2	272.5	3699.7	26029.4
8	504.2	293.2	4407.0	32384.4
10	550.1	319.5	5151.9	39404.6
12	600.2	348.3	5932.8	47090.0

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).



(a) MDP (b) Automaton

Figure 8. The *Multitask* MDPRM.



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

Figure 9. Regret in 6-state and 20-state *RiverSwim*