Exploration in Reward Machines with Low Regret

Hippolyte Bourel 1  Anders Jonsson 2  Odalric-Ambrym Maillard 3  Mohammad Sadegh Talebi 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 (Sec-
tion 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. 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 $O(\sqrt{K_{\mathcal{M}}OAT})$, in an MDPRM $M$, where $OA$ is the size of its observation-action space, $T$ is the number of time steps, and $O(\cdot)$ hides logarithmic and constant terms. Furthermore, $k_{\mathcal{M}}$ is an MDP-dependent and algorithm-dependent factor, which reflects the contribution of diameter-like quantities to the regret. In fact, $k_{\mathcal{M}}$ 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 $\sqrt{D_{cp}}$, the diameter of the cross-product MDP, whereas ours depend (via $k_{\mathcal{M}}$) 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 $O(\sqrt{DSAT} T \log(T))$ (Fruit et al., 2020) or as $O(D\sqrt{K} SAT \log(T))$ (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 $O(\sqrt{DSAT} T \log(T))$, 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 $3M$ (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 $A$, $\Delta_A$ 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 $X \rightarrow \Delta_A$. $A^*$ 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 $M = (\mathcal{O}, A, p, R, P, L)$, where $\mathcal{O}$ is a finite set of (observation) states with cardinality $\mathcal{O}$, $A$ is a finite set of (action) available at each state with cardinality $A$, $p : \mathcal{O} \times A \rightarrow \Delta(\mathcal{O})$ is the transition function such that $p(o'|o, a)$ denotes the probability of transitioning to state $o' \in \mathcal{O}$, when executing action $a \in A$ in state $o \in \mathcal{O}$. $R : (\mathcal{O} \times A)^+ \rightarrow \Delta_{[0,1]}$ denotes a history-dependent reward function such that for every history $h \in (\mathcal{O} \times A)^* \times \mathcal{O}$ and action $a \in A$, $R(h, a)$ defines a reward distribution. $P$ denotes a set of atomic propositions and $L : \mathcal{O} \times A \times \mathcal{O} \rightarrow 2^P$ denotes a labeling function such that $L$ assigns, to each triplet $(o, a, o')$, a subset of $P$.

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 $P$). 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 $a_t \in A$ 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 $r_t \sim \mathcal{R}(h_t, a_t)$ 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 $\sum_{t=1}^T r_t$.

**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 $R = (\mathcal{Q}, 2^\mathcal{P}, \tau, \nu)$, where $\mathcal{Q}$ is a finite set of states and $2^\mathcal{P}$ is an input alphabet. Furthermore, $\tau : \mathcal{Q} \times 2^\mathcal{P} \rightarrow \mathcal{Q}$ denotes a deterministic transition function such that $\tau(q, \sigma)$ denotes the next-state of $\mathcal{R}$ when an input $\sigma$ is received in state $q$, with the convention that $\tau(q, \emptyset) = q$. Finally, $\nu : \mathcal{Q} \times 2^\mathcal{P} \rightarrow \Delta_{\mathcal{O} \times A[0,1]}$ denotes the output function of $\mathcal{R}$, which returns a reward function $\nu : \mathcal{O} \times A \rightarrow \Delta_{[0,1]}$. 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 $r_t = \nu(q_t, \sigma_t)$, where $r_t : \mathcal{O} \times A \rightarrow \Delta_{[0,1]}$ only depends on the current state $q_t$ and current label $\sigma_t$. Conditioned on $(q_t, \sigma_t)$, $r_t$ 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 $R$ to RMs leads to MDPs with RMs. Formally, an average-reward MDP with RM (MDPRM) is a tuple $M = (\mathcal{O}, A, p, \mathcal{R}, P, L)$, where $\mathcal{O}, A, p, \mathcal{R},$ 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 $a_t \in A$ 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 $\sigma_t$ and outputs a reward function $r_t = \nu(q_t, \sigma_t)$ which is a mapping $r_t : \mathcal{O} \times A \rightarrow \Delta_{[0,1]}$. Then, the agent receives a reward $r_t \sim r_t(q_t, \sigma_t)$ (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 *RiverSwim* domain (Strehl & Littman, 2008). The labeled MDP has $N$ observations $o_1, \ldots, o_N$, two actions right (solid
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 $\pi : Q \times O \to A$, such that for all pairs $(q, o) \in Q \times O$, it prescribes an action $\pi(q, o) \in A$. Let $\Pi$ 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:

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

where $r_t \sim r_t(o_t, a_t)$ and $r_1 = \nu(q_1, L(o_1, \pi(q_1, o_1), o_{t+1}))$ 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 $g^* = \max_{\pi} g^\pi$ denote the maximal gain over all (possibly history dependent) policies. Any policy achieving $g^*$ is an optimal policy.

Following the same arguments as in tabular MDPs together with the equivalence between $M$ and $M_{cp}$ (Lemma 2.1), it is guaranteed that there exists at least one optimal policy in $\Pi$. We assume that the transition function $p$ is initially unknown, but that the RM $R$ 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 $r_t \sim r_t(o_t, a_t)$). We define the regret of an agent (or learning algorithm) $A$ as

$$\mathcal{R}(A, T) := T g^* - \sum_{t=1}^{T} r_t.$$  

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

\(^2\)This implies that both $\pi$ and $\nu$ are known. The assumption of knowing $\nu$ 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
n summary, we impose the following assumption on the considered MDP:

**Assumption 2.2.** We assume: (i) the RM \( R \) 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 \( t \):

\[
N_t(o, a, o') := \sum_{i=1}^{t-1} 1_{\{(o, a, o'_i) = (o, a, o')\}}.
\]

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 \( \hat{p}_t(o'|o, a) = \frac{N_t(o, a, o')}{N_t(o, a)} \) for \( p(o'|o, a) \), for any \( o, a, o' \in O \) and \( a \in A \). We consider two confidence sets for \( p \).

The first one uses a *time-uniform* variant of Weissman’s concentration inequality (Weissman et al., 2003), as introduced in (Asadi et al., 2019; Maillard, 2019):

\[
C^1_{t,\delta}(o, a) = \left\{ p' \in \Delta_O : \|\hat{p}_t(o'|o, a) - p'\|_1 \leq \beta_{N_t(o, a)}(\delta) \right\},
\]

\[
C^2_{t,\delta} = \cap_{o,a} C^1_{t,\delta}(o, a),
\]

where for \( n \in \mathbb{N} \),

\[
\beta_n(\delta) = \sqrt{\frac{2}{n} \left(1 + \frac{1}{n}\right) \log \left( \sqrt{n + 1} + \frac{\delta^2}{\delta^2 - 2} \right)}.
\]

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, \( \mathbb{P}(\exists t \in \mathbb{N} : p \notin C^1_{t,\delta/OA}) \leq \delta \). The second one is based on Bernstein’s inequality (combined with a peeling technique) and is defined as follows:

\[
C^2_{t,\delta}(o, a, o') = \left\{ u \in [0, 1] : |\hat{p}_t(o'|o, a) - u| \leq \sqrt{\frac{2u(1-u)}{N_t(o, a)} \beta_{N_t(o, a)}(\delta) + \beta_{N_t(o, a)}(\delta)} \right\}.
\]

and \( C^2_{t,\delta} = \cap_{o,a,o'} C^2_{t,\delta}(o, a, o') \), where for \( n \in \mathbb{N} \) and \( \delta \in (0, 1) \), \( \beta_{N_t(o, a)}(\delta) = \eta \log \left( \frac{\log(n+1) \log(n\delta)}{\delta^2} \right) \), 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 \( R_{cp} \). To this effect, let us define the empirical estimates for \( P \) and \( R \) as follows. By a slight abuse of notation, let \( \hat{R} \) denote the empirical mean of distribution \( R \), and let \( \hat{p} \) denote the mean of the reward function \( r = \nu(q, \sigma) \).

For all \( s = (q, o) \), \( s' = (q', o') \), and \( a \),

\[
\hat{P}_t(s'|s, a) = \hat{p}_t(o'|o, a) \Pi_{t} = \sigma(q, l(q, o, a', o', a)), \quad \hat{R}_t(s, a) = \sum_{o'} \hat{p}_t(o'|o, a) \sigma(q, l(q, o, a, o', a')).
\]

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 \( \hat{P} \)) and for \( R \) (centered at \( \hat{R} \)) 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 plausibly with the collected data up to time \( t \geq 1 \) and for a confidence parameter \( \delta \in (0, 1) \):

\[
M_{t,\delta} := M_{t,\delta}(C) := \left\{ M' = (O, A, p', R, \mathcal{P}, L) : p' \in C \right\},
\]

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 \( M_{t,\delta} \) with high probability, uniformly for all \( t \). More precisely, for all \( \delta \in (0, 1) \), and for each choice of \( C \), \( \mathbb{P}(\exists t \in \mathbb{N} : M \notin M_{t,\delta}) \leq 2\delta \), as formalized in Lemma C.1 in the Appendix C. This crucially relies on the equivalence between any candidate MDPRM \( M' \in M_{t,\delta} \) and its associated cross-product MDP \( M'_{cp} = (S, A, P, R) \) 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 \( C_{t, \delta}^1 \) (UCRL-RM-L1) or \( C_{t, \delta}^2 \) (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 MDPs 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:

\[
\max_{M' \in \mathcal{M}_{t_k, \delta}, \pi \in \Pi_{M'}} q^\pi(M') \quad \text{where} \quad q^\pi(M') = \mathbb{E}_v\left[\sum_{t=0}^{N_1-1} r_t + r_{N_1}\right],
\]

where \( u \) is the value function at the current iteration of EVI, and where \( C(o,a) = C_{1, \delta}^1/o_A(o,a) \) or \( C(o,a) = \bigcap_{o} C_{1, \delta}^2/o_A(o,a) \). EVI (in Algorithm 2 in Appendix A) returns a policy \( \pi_k \) which is guaranteed to be \( \frac{1}{\sqrt{t_k}} \)-optimal. The algorithm commits to \( \pi_k \) 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 \( k \geq 1 \),

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

The pseudo-code of UCRL-RM is presented in Algorithm 1.

We recover UCRL-RM-L1 (resp. UCRL-RM-B) if \( M \) is constructed using \( C_{t, \delta}^1 \) (resp. \( C_{t, \delta}^2 \)). Both algorithms receive the RM \( \mathcal{R} \) as well as a confidence parameter \( \delta \in (0, 1) \) as input. Despite their similar design, they achieve different performance 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.
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 $\sigma_A$) and $a_1$ (which leads to detecting $\sigma_B$). The RM has $Q$ states arranged in a cycle, such that $\sigma_A$ and $\sigma_B$ yield transitions in the clockwise and counter-clockwise directions, respectively. As detailed in the Appendix E, we can show that: for all $q \in Q$,
\[
D_{o_1,q} = \frac{2}{\delta} + 1 + \frac{\delta}{1-\delta}, \quad D_{o_0,q} = \frac{1}{\delta},
\]
whereas $D_{cp} = \frac{Q/2}{\delta} + 1 + \frac{\delta}{1-\delta}$. So, while $D_{cp}$ grows with $\frac{Q}{\delta}$, $D_s$ for all $s \in S$ remain constant and only grow with $\frac{1}{\delta}$. In summary, we have $D_s \lesssim D_{cp}/Q$. 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 $C_{t,\delta}^1$:

**Theorem 4.2 (Regret of UCRL-RM-L1).** Under UCRL-RM-L1, uniformly over all $T \geq 2$, with probability higher than $1 - 4\delta$,
\[
\mathcal{R}(T) \leq O\left(\sqrt{c_M AT(O + \log(OA\sqrt{T}/\delta))} + D_{cp} O A \log(T)\right),
\]
where $c_M = \sum_{o \in O} \max_{q \in Q} D_{q,o}^2$.

In the following theorem, we present a regret bound for UCRL-RM-B (constructed using $C_{t,\delta}^2$). To this end, for $(o,a) \in S$, we define $K_{o,a}$ as the number of possible next-states in $O$ under $(o,a)$, that is, $K_{o,a} := |\{o' \in O : p(o'|o,a) > 0\}|$.

**Theorem 4.3 (Regret of UCRL-RM-B).** Under UCRL-RM-B, uniformly over all $T \geq 2$, with probability higher than $1 - 5\delta$,
\[
\mathcal{R}(T) \leq O\left(\sqrt{c_M' T \log(O^2 A \log(\log(T))/\delta)} + D_{cp} O A \log(T)\right),
\]
where $c_M' = \sum_{o \in O, a \in A} K_{o,a} \max_{q \in Q} D_{q,o}^2$.

We remark that both regret bounds depend on $D_{cp}$ only via a logarithmic term. The problem-dependent quantities $c_M$ and $c_M'$ reflect the (weighted) contribution of RM-restricted diameters to the regret. In the worst-case, $c_M \leq O D_{cp}^2$ and $c_M' = D_{cp}^2 \sum_{o,a} K_{o,a}$, but in view of the example earlier, there are problem instances in which $c_M \lesssim O D_{cp}^2/Q^2$ and $c_M' \lesssim D_{cp}^2/Q^2 \sum_{o,a} K_{o,a}$.

**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 $O(D_{cp} \sqrt{AOQT(O + \log T)})$ whereas UCRL2-B achieves a regret of
\[
O\left(D_{cp} \sqrt{T \log(\log(T)) \sum_{o,a} K_{o,a}}\right).
\]

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 MDPMs. 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 $\Omega(\sqrt{D_{cp}QOAT})$ 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

\[^5\text{The regret of UCRL2-B can be improved to } O(\log(T) / \sqrt{T D_{cp} \sum_{o,a} K_{o,a}}) \text{ as reported in (Fruit et al., 2020), and the same improvement can carry over to UCRL-RM-B. We exclude comparisons to the regret of EBF (Zhang & Ji, 2019) growing as } O(\sqrt{D_{cp} Q O A T \log(T)}) \text{, as it does not admit an efficient implementation.}\]
worst-case MDPs become their associated cross-product MDPs.

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

**Theorem 5.1.** For any $O \geq 3$, $A \geq 2$, $Q \geq 2$, and $D_{cp} \geq Q(6 + 2 \log_A(O))$, $T \geq D_{cp}OAT$ and $|P| \geq 2$, 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 $\pi$ on these MDPRMs satisfies

$$E[\mathcal{R}(\pi, T)] \geq c_0 \sqrt{D_{cp}OAT},$$

where $c_0 > 0$ is a universal constant.

This theorem asserts a worst-case regret lower bound growing as $\Omega(\sqrt{D_{cp}OAT})$. 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 $R$ 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 $\max_q D_{q,o} \simeq D_{cp}$.

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 $C_{t, \delta}^1$ and $C_{t, \delta}^2$, 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.

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.
Acknowledgements

The authors would like to thank anonymous reviewers for their comments. Hippolyte Bourel and Mohammad Sadegh Talebi are partially supported by the Independent Research Fund Denmark, grant number 1026-00397B. Anders Jonsson is partially supported by Spanish grants PID2019-108141GB-I00 and PCIN-2017-082. Odalric-Ambrym Maillard is supported by CPER Nord-Pas-de-Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria, Inria Scool, the French Agence Nationale de la Recherche (ANR) under grant ANR-16-CE40-0002 (the BADASS project), the MEL, the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF.

References


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) and UCRL2B (Algorithm 4) in (Jaksch et al., 2010), whereas 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 $(q, a, o)$, set $\tilde{p}^q(\cdot|s,a)$ as any optimizer to the inner maximization of the EVI2, resolved for each triplet $(q, a, o)$ by the algorithms 3 and 4. Here the superscript $q$ in $\tilde{p}^q(\cdot|s,a)$ signifies that the optimistic transition probability depends on $q$.

### Algorithm 2 EVI ($C, \varepsilon$)

Initialize: $u^{(0)} \equiv 0$, $u^{(-1)} \equiv -\infty$, $n = 0$
while max$_{s \in S} (u^{(n)} - u^{(n-1)})(s) - \min_{s \in S} (u^{(n)} - u^{(n-1)})(s) > \varepsilon$ do
  Get $\tilde{p}^q$ for all $q \in Q$ using MAXP (Algorithm 3 for UCRL-RM-L1, Algorithm 4 for UCRL-RM-B)
  For all $(q, a, o) \in S$, update:
  
  $$u^{(n+1)}(q, a) = \max_{a' \in A} \sum_{(q', o') \in S} \tilde{p}^q(o'|q, a) \left[ \nu(q, L(o, a, o')) + u^{(n)}(q', o') I_{q'} = \tau(q, L(o, a, o')) \right]$$

  Set $n = n + 1$
end while
Output: $\pi_{n+1}(s) = \arg\max_{a \in A} \sum_{(q', o') \in S} \tilde{p}^q(o'|q, a) \left[ \nu(q, L(o, a, o')) + u^{(n)}(q', o') I_{q'} = \tau(q, L(o, a, o')) \right]$

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

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

Initialize: $\forall o' \in O, \tilde{p}^q(o'|s,a) = \tilde{p}^q(o'|q, a)$
$\forall o' \in O, u(q, o, a)(o') = u(\tau(q, L(o, a, o'), o')$
$o_{\text{max}} = \arg\max_{o'} \mu(q, L(o, a, o')) u(q, o, a)(o')$
$\tilde{p}^q(o_{\text{max}}|q, a) = \max(1, \tilde{p}^q(o_{\text{max}}|q, a) + \frac{1}{2} \beta N_{1(o, a)}(\frac{\delta}{\|\pi\|^2}))$
$\mathcal{L} = \arg\min_{o', \nu} \mu(q, L(o, a, o')) u(q, o, a)(o')$ and $l = 0$
while $\sum_{o' \in O} \tilde{p}^q(o'|s,a) > 1$ do
  $\tilde{p}^q(L_1|o, a) = \max(0, \tilde{p}^q(L_1|o, a) + 1 - \sum_{o' \in O} \tilde{p}^q(o'|s,a))$
  Set $l = l + 1$
end while
Output: $\tilde{p}^q(\cdot|o, a)$

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

Initialize: $\forall o' \in O, \tilde{p}^q(o'|s,a) = \min\{p' \in C^2_{\varepsilon, \delta/O^2}(o, a, o')\}$
$\forall o' \in O, u(q, o, a)(o') = u(\tau(q, L(o, a, o'), o')$
$\mathcal{L} = \arg\min_{o', \nu} \mu(q, L(o, a, o')) u(q, o, a)(o')$ and $l = OA - 1$
while $\sum_{o' \in O} \tilde{p}^q(o'|s,a) < 1$ do
  $\tilde{p}^q(L_1|o, a) = \min\{\max(C^2_{\varepsilon, \delta/O^2}(o, a, L_1)), 1 - \sum_{o' \in O} \tilde{p}^q(o'|s,a)\}$
  $l = l + 1$
end while
Output: $\tilde{p}^q(\cdot|o, a)$
A.2. Unknown Mean Rewards

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

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

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

Then, it suffices to replace $\beta$ with the upper confidence set, that is, to replace $\beta(q, \sigma)$ with $\tilde{\nu}_t(\cdot | q, \sigma) + \beta_{N_t(q, \sigma)}(\delta)$. 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 (repeated)** Let $M = (Q, A, P, R, L)$ be a finite MDP. Then, an associated cross-product MDP to $M$ is $M_{cp} = (S, A, P, R)$, where $S = Q \times O$, and where for $s = (q, o), s' = (q', o') \in S$, and $a \in A,$

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

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

$$P(s_t+1 = s' | h_t = h, a_t = a) = p(o' | q, a)I_{(q' = \tau(q, L(o, a, o')))},$$

$$P(r_t = r | h_t = h, a_t = a) = \sum_{o' \in O} p(o' | q, a)\nu(q, L(o, a, o')),$$

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

$$P(s_{t+1} = (q', o') | h_t = h, a_t = a)$$

$$= P(\, o_{t+1} = o'| h_t = h, a_t = a)P(\, q_{t+1} = q' | h_t = h, a_{t+1} = o', a_t = a)$$

$$= P(\, o_{t+1} = o'| o_t = o, a_t = a)P(\, q_{t+1} = q' | s_t = (q, o), a_{t+1} = o', a_t = a)$$

$$= p(o' | q, a)I_{(q' = \tau(q, L(o, a, o')))}.$$

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

$$P(r_t = r | h_t = h, a_t = a)$$

$$= \sum_{o' \in O} p(o' | q, a)P(r_t = r | h_t = h, o_{t+1} = o', a_t = a)$$

$$= \sum_{o' \in O} p(o' | q, a)\nu(q, L(o, a, o')),$$

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

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:

\begin{align*}
(i) & \quad P(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t, \delta}(C^1)) \leq \delta, \\
(ii) & \quad P(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t, \delta}(C^2)) \leq 2\delta.
\end{align*}

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 $(o, a)$,

$$P\left(\exists t \in \mathbb{N}, p(o|a) \notin C_{t, \delta/OA}(o, a)\right) \leq \frac{\delta}{O.A}.$$  

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

$$P\left(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t, \delta}\right) = P\left(\exists t \in \mathbb{N}, \exists p \notin C_{t, \delta/OA}\right)$$

$$= \sum_{o \in \mathcal{O}, a \in \mathcal{A}} P\left(\exists t \in \mathbb{N}, p(o|a) \notin C_{t, \delta/OA}(o, a)\right)$$

$$\leq \sum_{o \in \mathcal{O}, a \in \mathcal{A}} \frac{\delta}{O.A} = \delta.$$  

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

$$P\left(\exists t \in \mathbb{N}, p(o'|o, a) \notin C_{t, \delta/O2A}(o, a, o')\right) \leq \frac{2\delta}{O^2A}.$$  

Hence, for $\mathcal{M}_{t, \delta}$ built using $C_{t, \delta/O2A}$, we have

$$P\left(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t, \delta}\right) = P\left(\exists t \in \mathbb{N}, \exists p \notin C_{t, \delta/O2A}\right)$$

$$= \sum_{o, o' \in \mathcal{O}, a \in \mathcal{A}} P\left(\exists t \in \mathbb{N}, p(o'|o, a) \notin C_{t, \delta/O2A}(o, a, o')\right)$$

$$\leq \sum_{o, o' \in \mathcal{O}, a \in \mathcal{A}} \frac{2\delta}{O^2A} = 2\delta.$$  

\hfill \Box

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 $\delta$ (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 $\mathbb{E}[r_t|s_t, a_t] = \sum_{o'} p(o'|o, a)\mathbb{E}[q_t|L(o, a, o')]$. Hence, by applying Corollary C.8, we deduce that

$$\mathfrak{R}(T) = \sum_{t=1}^{T} g^* - \sum_{t=1}^{T} r_t$$

$$\leq \sum_{t=1}^{T} \sum_{o, q, a} \left(g^* - \sum_{o'} p(o'|o, a)\mathbb{E}[q_t|L(o, a, o')]\right) \mathbb{1}_{(q_t, a_t) = (q, o, a)} + \sqrt{\frac{1}{2}(T + 1) \log(\sqrt{T + 1/\delta})}$$

$$= \sum_{o, q, a} \left(g^* - \sum_{o'} p(o'|o, a)\mathbb{E}[q_t|L(o, a, o')]\right) N_{m(T)}(q, o, a) + \sqrt{\frac{1}{2}(T + 1) \log(\sqrt{T + 1/\delta})},$$
with probability at least $1 - \delta$. 

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

$$
\sum_{s,a} N_{m(T)}(s,a)(g^* - \mu(s,a)) = \sum_{k=1}^{m(T)} \sum_{s,a} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}_{[s',a',a]}(g^* - \mu(s,a)) \\
= \sum_{k=1}^{m(T)} \sum_{s,a} v_k(s,a)(g^* - \mu(s,a)). 
$$

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

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

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 $k = 1, \ldots, m(T)$. 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 $\pi_k$ and $\tilde{M}_k = (\mathcal{S}, \mathcal{A}, \tilde{p}, \tilde{r}, \tilde{P}, \mathcal{L})$, we have that $g_k := g_{\pi_k}(\tilde{M}_k) \geq g^* - \frac{1}{\sqrt{t_k}}$. Hence, the regret accumulated in episode $k$ satisfies:

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

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

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

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

$$
u_k^{(i+1)}(s) = \sum_{s' \in \mathcal{S}} \tilde{p}_k(o'|o, \pi_k(s)) \mathbb{P}(q, L(o, \pi_k(s), o')) + \mathbb{I}_{\{q' = \tau(q, L(o, \pi_k(s), o'))\}} u_k^{(i)}(s'),
$$

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

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

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

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

Defining $g_k = g_k 1$, $\tilde{\mu}_k := (\tilde{\mu}_k(s, \pi_k(s)))_s$, $\tilde{P}_k := \left( \tilde{p}_k(o'|o, \pi_k(s)) \mathbb{I}_{\{q' = \tau(q, L(o, \pi_k(s), o'))\}} \right)_{s, s'}$, and $v_k := (v_k(s, \pi_k(s)))_s$, we can rewrite the above inequality as:

$$
|g_k - \tilde{\mu}_k - (\tilde{P}_k - I) u_k^{(i)}| \leq \frac{1}{\sqrt{t_k}} 1.
$$
Combining this with (4) yields

\[ \Delta_k \leq \sum_{s,a} v_k(s,a)(g_k - \mu(s,a)) + \sum_{s,a} v_k(s,a)(\hat{\mu}_k(s,a) - \mu(s,a)) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}. \]  

(6)

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

\[
\sum_{s,a} v_k(s,a)(\hat{\mu}_k(s,a) - \mu(s,a)) = \sum_{s,a} v_k(s,a) \sum_{o' \in O} (\hat{\mu}^{(o)}(o'|o,a) - p(o'|o,a))\pi(q, L(o, a, o')) \\
\leq \sum_{s,a} v_k(s,a) \| \hat{\mu}^{(o)}(\cdot|o,a) - p(\cdot|o,a) \|_1 \\
\leq 2 \sum_{s,a} v_k(s,a) \beta_{N_k(o,a)} \left( \frac{\delta}{\| A \|} \right) \\
= 2 \sum_{o,a} \beta_{N_k(o,a)} \left( \frac{\delta}{\| A \|} \right) \sum_{q = v_k(o,a)} v_k(q, o, a) \\
= 2 \sum_{o,a} v_k(o,a) \beta_{N_k(o,a)} \left( \frac{\delta}{\| A \|} \right),
\]

where we used that \( \pi(q, L(o, a, o')) \leq 1 \). 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

\[ \Delta_k \leq v_k(\bar{P}_k - I)u_k^{(i)} + 2 \sum_{s,a} v_k(o,a) \beta_{N_k(o,a)} \left( \frac{\delta}{\| A \|} \right) + 2 \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}. \]

Let us define, for all \( s \in S \),

\[ w_k(s) := u_k^{(i)}(s) - \frac{1}{2} \left( \min_{s' \in B, X \in O} u_k^{(i)}(s') + \max_{s' \in B, X \in O} u_k^{(i)}(s') \right). \]

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

\[ \Delta_k \leq v_k(P_k - I)w_k + v_k(\bar{P}_k - P_k)w_k + 2 \sum_{o,a} v_k(o,a) \beta_{N_k(o,a)} \left( \frac{\delta}{\| A \|} \right) + 2 \sum_{o,a} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}}. \]  

(7)
Upper bound on $L_1$. We have

\[
v_k(\tilde{P}_k - P_k)w_k = \sum_{s \in S} \sum_{s' \in S} v_k(s, \pi_k(s)) \left( \tilde{P}_k(s'|s, \pi_k(s)) - P(s'|s, \pi_k(s)) \right) w_k(s') \leq \sum_{s \in S} v_k(s, \pi_k(s)) \sum_{o' \in O} \sum_{q' \in Q} \left( \tilde{p}_k(o'|o, \pi_k(s)) - p(o'|o, \pi_k(s)) \right) I_{q' = \tau(q, L(o, \pi_k(s), o'))} w_k(q', o') \leq \sum_{s \in S} v_k(s, \pi_k(s)) \sum_{o' \in O} \left[ \tilde{p}_k(o'|o, \pi_k(s)) - p(o'|o, \pi_k(s)) \right] \max_{s' \in B_{q', o \times O}} |w_k(q', o')| \sum_{q' \in Q} I_{q' = \tau(q, L(o, \pi_k(s), o'))} \]

\[
\leq \sum_{s \in S} v_k(s, \pi_k(s)) \| (\tilde{p}_k - p)(o, \pi_k(s)) \|_1 \max_{s' \in B_{q, o \times O}} |w_k(q', o')| \leq \sum_{o \in O} \sum_{q \in Q} v_k(q, o, \pi_k(q, o)) \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \cdot D_{q, o} \leq \sum_{a \in A} \sum_{o \in O} \sum_{q \in Q} v_k(q, o, a) \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \cdot D_{q, o} \sum_{q \in Q} v_k(q, o, a) \leq \sum_{a \in A} \sum_{o \in O} \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \max_{q \in Q} D_{q, o} \cdot v_k(o, a), \tag{8} \]

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:

\[
\sum_{k=1}^{m(T)} \Delta_k I_{\{M \in M_k\}} \leq \sum_{k=1}^{m(T)} v_k(\tilde{P}_k - I)w_k I_{\{M \in M_k\}} + 2 \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \sqrt{N_k(o, a)} + \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \left( 2 + \max_{q \in Q} D_{q, o} \right) \leq \sum_{k=1}^{m(T)} v_k(\tilde{P}_k - I)w_k I_{\{M \in M_k\}} + 2 \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \sqrt{N_k(o, a)} + \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \left( 2 + \max_{q \in Q} D_{q, o} \right) \leq \sum_{k=1}^{m(T)} v_k(\tilde{P}_k - I)w_k I_{\{M \in M_k\}} + 2 \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \sqrt{N_k(o, a)} + \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \beta_{\pi_k(q, o)} \left( \frac{\delta}{\| A \|} \right) \left( 2 + \max_{q \in Q} D_{q, o} \right) + 2 \sqrt{O + \log (OA\sqrt{T + 1/\delta})} \sum_{k=1}^{m(T)} \sum_{o, a} v_k(o, a) \left( \max_{q \in Q} D_{q, o} + 2 \right), \tag{10} \]

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 $(Z_t)_{t \geq 1}$ be a sequence with

\[
Z_t := (P(\cdot|s_t, a_t) - e_{s_{t+1}})w_k(s_t) I_{\{M \in M_t(s_t)\}},
\]
for all \(t\), where \(k(t)\) denotes the episode containing time step \(t\). For any \(k\) with \(M \in \mathcal{M}_k\), we have:

\[
v_k(P_k - I)w_k = \sum_{t=k}^{t_k+1-1} (P(\cdot|s_t, a_t) - e_{s_t})w_k
\]

\[
= \sum_{t=k}^{t_k+1-1} (P(\cdot|s_t, a_t) - e_{s_{t+1}} + e_{s_{t+1}} - e_{s_t})w_k
\]

\[
= \sum_{t=k}^{t_k+1-1} Z_t + w_k(s_{t+1}) - w_k(s_k) \leq \sum_{t=k}^{t_k+1-1} Z_t + D_{cp},
\]

where \(e_i\) denotes a vector with the \(i\)-th element being 1 and the others being zero. Hence, \(L_2 \leq \sum_{t=1}^{T} Z_t + m(T)D_{cp}\).

Let \(\text{supp}(\cdot)\) denote the support set of a distribution. Since \(s_{t+1} \in \text{supp}(P(\cdot|s, a))\), we have

\[
|Z_t| \leq \left| \sum_{s'} P(s'|s_t, a_t)w_k(t)(s') - w_k(t)(s_{t+1}) \right|
\]

\[
\leq \sum_{s'} P(s'|s_t, a_t)\left| w_k(t)(s') - w_k(t)(s_{t+1}) \right|
\]

\[
\leq \sum_{s'} P(s'|s_t, a_t) \max_{s, s' \in \text{supp}(P(\cdot|s, a))} \left| w_k(t)(s') - w_k(t)(s) \right|
\]

\[
= \max_{s, s' \in \text{supp}(P(\cdot|s, a))} \left| w_k(t)(s') - w_k(t)(s) \right|.
\]

Note that

\[
\text{supp}(P(\cdot|s, a)) = \{(q', o') \in S : p(a'|o, a) > 0 \text{ and } q' = L(o, a, o')\} \subseteq B_s \times O.
\]

Hence,

\[
|Z_t| \leq \max_{s, s' \in B_s \times O} \left| w(s') - w(s) \right| \leq D_s
\]

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

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

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

\[
L_2 \leq \max_s D_s \sqrt{\frac{1}{2}(T + 1) \log(\sqrt{T + 1}/\delta)} + m(T)D_{cp}
\]

\[
\leq \max_s D_s \sqrt{\frac{1}{2}(T + 1) \log(\sqrt{T + 1}/\delta)} + D_{cp}OA\log_2 \left( \frac{8T}{OA} \right),
\]

(11)

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

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

\[
\sum_{k=1}^{m(T)} \Delta_k \mathbb{1}_{\{M \in \mathcal{M}_k\}} \leq 2\sqrt{O + \log \left( OA\sqrt{\frac{T}{\delta}} \right)} \sum_{k=1}^{m(T)} \max_{q, o} \left( \frac{m(T)}{\max_{q, o} D_{q,o}} + 2 \right) \frac{v_k(o, a)}{\sqrt{N_k(o, a)}}
\]

\[
+ 2 \sum_{k=1}^{m(T)} \max_{o, a} \frac{v_k(o, a)}{\sqrt{N_k(o, a)}} + \max_s D_s \sqrt{\frac{1}{2}(T + 1) \log(\sqrt{T + 1}/\delta)} + D_{cp}OA\log_2 \left( \frac{8T}{OA} \right),
\]

(12)
with probability higher than $1 - \delta$ and uniformly over all $T \in \mathbb{N}$. Applying Lemma C.4 and using the Cauchy-Schwarz inequality:

$$\sum_{k=1}^{m(T)} \sum_{o,a} \max_{q \in \mathcal{Q}} D_{q,o} \frac{v_k(o,a)}{\sqrt{N_k(o,a)}} \leq (\sqrt{2} + 1) \sum_{o,a} \max_{q \in \mathcal{Q}} D_{q,o} \sqrt{N_T(o,a)}$$

$$\leq (\sqrt{2} + 1) \sqrt{\sum_{o,a} \max_{q \in \mathcal{Q}} D_{q,o}^2 \sum_{o,a} N_T(o,a)}$$

$$= (\sqrt{2} + 1) \sqrt{T \sum_{o,a} \max_{q \in \mathcal{Q}} D_{q,o}^2} = (\sqrt{2} + 1) \sqrt{c_M AT};$$

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

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

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

$$\mathfrak{R}(T) \leq 2(\sqrt{2} + 1) \sqrt{O + \log (OA \sqrt{T + 1}/\delta)(\sqrt{c_M} + 2) \sqrt{AT} + 2(\sqrt{2} + 1) \sqrt{OAT}}$$

$$+ (\max_s D_s + 1) \sqrt{\frac{1}{2} (T + 1) \log (\sqrt{T + 1}/\delta) + D_{cp} OA \log_2 \left( \frac{8T}{\delta A} \right)},$$

thus proving the theorem. \hfill \Box

### 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

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

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

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

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

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:
\[
\tilde{\mu}_K(s, a) = \mu(s, a) = \sum_{o' \in \mathcal{O}} (\tilde{p}^{o'}(o, a) - p(o'|o, a)) \mathbb{P}(q, L(o, a, o'))
\]
\[
\leq \sum_{o' \in \mathcal{O}} \left| \tilde{p}^{o'}(o, a) - p(o'|o, a) \right|
\]
\[
\leq \sum_{o' \in \mathcal{O}} \sqrt{\frac{2 \tilde{p}^{o'}(o, a)(1 - \tilde{p}^{o'}(o, a))}{N_k(o, a)} \beta'_{N_k(o, a)}(\frac{\delta}{\delta^2 A})}
\]
\[
+ \sum_{o' \in \mathcal{O}} \sqrt{\frac{2p(o'|o, a)(1 - p(o'|o, a))}{N_k(o, a)} \beta'_{N_k(o, a)}(\frac{\delta}{\delta^2 A}) + \frac{2}{3N_k(o, a)} \beta'_{N_k(o, a)}(\frac{\delta}{\delta^2 A})}
\]
\[
\leq ^{(a)} \sqrt{\beta'_T(\frac{\delta}{\delta^2 A}) \sum_{o' \in \mathcal{O}} \frac{2 \tilde{p}_k(o'|o, a)(1 - \tilde{p}_k(o'|o, a))}{N_k(o, a)}}
\]
\[
+ \sqrt{\beta'_T(\frac{\delta}{\delta^2 A}) \sum_{o' \in \mathcal{O}} \frac{2p(o'|o, a)(1 - p(o'|o, a))}{N_k(o, a)}} + \frac{4}{N_k(o, a)} \beta'_T(\frac{\delta}{\delta^2 A})
\]
\[
\leq ^{(b)} \sqrt{8 \beta'_T(\frac{\delta}{\delta^2 A}) \frac{K_{o,a}}{N_k(o, a)} + \frac{4}{N_k(o, a)} \beta'_T(\frac{\delta}{\delta^2 A})}
\]
(13)

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
\[
\Delta_k \leq v_k(\mathbf{P}_k - \mathbf{I})w_k + \sqrt{\frac{8 \beta'_T(\frac{\delta}{\delta^2 A})}{L_1} \sum_{o, a} v_k(o, a) \sqrt{\frac{K_{o,a}}{N_k(o, a)}}}
\]
\[
+ \frac{4}{N_k(o, a)} \beta'_T(\frac{\delta}{\delta^2 A}) \sum_{o, a} v_k(o, a) N_k(o, a) + 2 \sum_{o, a} v_k(o, a) N_k(o, a).
\]

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

**Upper Bound on \( L_1 \).** We have
\[
v_k(\mathbf{P}_k - \mathbf{P}_k)w_k
\]
\[
\leq \sum_{s \in \mathcal{S}} v_k(s, \pi_k(s)) \sum_{o' \in \mathcal{O}} \sum_{q' \in \mathcal{Q}} \left( \tilde{p}_k^s(o'|o, \pi_k(s)) - p(o'|o, \pi_k(s)) \right) \mathbb{I}_{(q' = \tau(q, L(o, \pi_k(s), o')))} w_k(q', o')
\]
\[
\leq \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} v_k(s, a) \sum_{o' \in \mathcal{O}} \sum_{q' \in \mathcal{Q}} \left( \tilde{p}_k^s(o'|o, a) - p(o'|o, a) \right) \mathbb{I}_{(q' = \tau(q, L(o,a,o')))} w_k(q', o')
\]
\[
\leq \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} v_k(s, a) \sum_{o' \in \mathcal{O}} \left| \tilde{p}_k^s(o'|o, a) - p(o'|o, a) \right| \cdot \max_{s' \in \mathcal{B}_{q', o} \times \mathcal{O}} \left| w_k(q', o') \right| \sum_{q' \in \mathcal{Q}} \mathbb{I}_{(q' = \tau(q, L(o,a,o')))}
\]
\[
\leq \frac{1}{2} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} D_s v_k(s, a) \sum_{o' \in \mathcal{O}} \left| \tilde{p}_k^s(o'|o, a) - p(o'|o, a) \right| ,
\]
where the last inequality follows from Lemma C.2.
Now plugging in the bound derived for \( \sum_{o' \in \mathcal{O}} |\tilde{P}_k(o'|a, a) - p(o'|a, a)| \) in (13), we obtain
\[
v_k(\bar{P}_k - P_k)w_k \leq \sqrt{8\beta'_T N_k(o, a)} \sum_{k \in \mathcal{O}} \max_{q \in \mathcal{Q}} D_{q,a} \cdot v_k(o, a) \sqrt{K_{o,a} N_k(o, a)} + 4D_{cp} \beta'_T N_k(o, a),
\]

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:
\[
\max_{s' \in \mathcal{B}_s \times \mathcal{O}} |w_k(s')| \leq \frac{D_s}{2}, \quad ||w_k||_{\infty} \leq \frac{D_{cp}}{2}.
\]

Proof. We first show that for all \(s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}\), we have \(u^{(i)}_k(s_1) - u^{(i)}_k(s_2) \leq D_s\), which further implies
\[
\max_{x \in \mathcal{B}_s \times \mathcal{O}} |w_k(x)| \leq \frac{D_s}{2}.
\]

To prove this, recall that similarly to (Jaksch et al., 2010), we can combine all MDPRMs in \(\hat{M}_k\) to form a single MDPRM \(\hat{M}_k\) with continuous action space \(\mathcal{A}'\). In this extended MDPRM, in any \(s = (q, a) \in \mathcal{S}\), and for each \(a \in \mathcal{A}\), there is an action in \(\mathcal{A}'\) with mean \(\hat{m}_k(s, a)\) and transition probability \(\hat{F}_k(s'|s, a)\) (of the associated \(M_{cp}\) belonging to the maintained confidence sets. Similarly to (Jaksch et al., 2010), we note that \(u^{(i)}_k(s)\) amounts to the total expected \(i\)-step reward of an optimal non-stationary \(i\)-step policy starting in state \(s\) on the MDPRM \(\hat{M}_k\) with the extended action set. The RM-restricted diameter of state \(s\) of this extended MDPRM is at most \(D_s\), as by assumption \(k\) is a good episode and hence \(\hat{M}_k\) contains the true MDPRM \(\hat{M}\), and therefore, the actions of the true MDPRM are contained in the continuous action set of \(\hat{M}_k\). Now, if there were states \(s_1, s_2 \in \mathcal{B}_s \times \mathcal{O}\) with \(u^{(i)}_k(s_1) - u^{(i)}_k(s_2) > D_s\), then an improved value for \(u^{(i)}_k(s_1)\) 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 \(u^{(i)}_k(s_1) \geq u^{(i)}_k(s_2) - D_s\), 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
\[
y(1 - y) = x(1 - x) + (1 - 2x)(y - x) - (y - x)^2
= x(1 - x) + (1 - x - y)(y - x)
\leq x(1 - x) + |1 - x - y| \left(\sqrt{2y(1-y)}\zeta + \frac{1}{3}\zeta\right)
\leq x(1 - x) + \sqrt{2y(1-y)}\zeta + \frac{1}{3}\zeta.
\]

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
\[
y(1 - y) \leq x(1 - x) + \frac{1}{3}\zeta + \sqrt{2\zeta \left(x(1 - x) + \frac{1}{3}\zeta\right)} + 2\zeta
\leq x(1 - x) + \sqrt{2\zeta x(1 - x)} + 3.15\zeta
= \left(\sqrt{x(1 - x)} + \sqrt{\frac{1}{3}\zeta}\right)^2 + 2.65\zeta,\quad (14)
\]
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:
\[ \sqrt{y(1-y)} \leq \sqrt{x(1-x)} + \sqrt{\frac{1}{2}\zeta} + \sqrt{2.65\zeta} \leq \sqrt{x(1-x)} + 2.4\sqrt{\zeta}. \]

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
\[ (i) \quad \sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq (\sqrt{2} + 1)\sqrt{Z_n}. \]
\[ (ii) \quad \sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq 2\log(Z_n) + 1. \]

Lemma C.5 (Jaksch et al., 2010, Proposition 18)). The number $m(T)$ of episodes up to time $T \geq OA$ satisfies
\[ m(T) \leq OA \log_2 \left( \frac{8T}{\delta^2} \right). \]

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 $\sigma$-sub-Gaussian noise if
\[ \forall t, \forall \lambda \in \mathbb{R}, \quad \log \mathbb{E} \left[ \exp \left( \lambda (Y_t - \mathbb{E}[Y_t|\mathcal{F}_{t-1}]) \right) \mid \mathcal{F}_{t-1} \right] \leq \frac{\lambda^2 \sigma^2}{2}, \]
where $\mathcal{F}_{t-1}$ denotes the $\sigma$-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 $\mu$, such that $Y_n - \mu$ is $\sigma$-sub-Gaussian. Let $\bar{\mu}_n = \frac{1}{n} \sum_{s=1}^{n} Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0, 1)$, it holds
\[ \mathbb{P} \left( \exists n \in \mathbb{N}, \quad |\bar{\mu}_n - \mu| \geq \sigma \sqrt{\left( 1 + \frac{1}{n} \right) \frac{2 \ln \left( \frac{\sqrt{n} + 1}{\delta} \right)}{n}} \right) \leq \delta. \]

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

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 $\hat{P}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i = x\}}$ for all $x \in \mathcal{X}$. Then, for all $\delta \in (0, 1)$,
\[ \mathbb{P} \left( \exists n \in \mathbb{N} : \|P - \hat{P}_n\|_1 \geq \sqrt{\frac{2}{n} \left( 1 + \frac{1}{n} \right) \log \left( \frac{\sqrt{n + 1} |\mathcal{X}| - 2}{\delta} \right)} \right) \leq \delta. \]
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 $\mathcal{F} = (\mathcal{F}_t)$, be its natural filtration. Assume for all $t \in \mathbb{N}$, $|Z_t| \leq b$ and $\mathbb{E}[Z_t^2 | \mathcal{F}_{t-1}] \leq v$ 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 $\delta \in (0, 1)$,

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

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

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 $\mu$, we have, for all $\delta \in (0, 1)$,

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

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

**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 MDP 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).

![Diagram](image-url)

*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).
To build the RM, we let $\Delta$ be a value greater than the cross-product diameter of $\mathcal{P}$. Then, we build the following RM transitions function $r$ and reward function $r$. For all actions $a \in \mathcal{A}$ and transitions to either $o_{A}$ or $o_{B}$ according to:

$$p(o_{A}|o_{i}, a) = \frac{1}{2} + \varepsilon(a, i) \quad \text{and} \quad p(o_{B}|o_{i}, a) = \frac{1}{2} - \varepsilon(a, i),$$

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:

$$p(o|o_{A}, a) = 1 - \delta, \quad p(o_{A}|o_{A}, a) = \delta, \quad o \in \{o_{A}, o_{B}\}.$$  

Let us choose $\delta = \frac{6Q}{D_{\text{cp}}}$. 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 $D_{\text{cp}}$, regardless of the value of $\Delta$. Also, for the diameter of the labeled MDP, $D$, we will have $D = \frac{4}{3}$.

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 $a \in \mathcal{A}$,

$$L(o_{A}, a, o_{A}) = \sigma_{A}, \quad L(o_{A}, a, o_{B}) = \sigma_{A\cap B}, \quad L(o_{B}, a, o_{B}) = \sigma_{B}, \quad L(o_{A}, a, o_{0}) = \sigma_{A\cap B}, \quad L(o_{B}, a, o_{0}) = \sigma_{A\cap B}.$$  

To build the RM, we let $N = \lceil (Q - 1)/2 \rceil$ and $N' = \lceil (Q - 1)/2 \rceil$. 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 $r$ and reward function $r$, for all $i \in 1, N$ and all $j \in 1, N'$.
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 $\sigma_A$ 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 $\sigma_B$ 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 $\delta$ 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 $\frac{2T}{kD}$ decisions and the rewards are roughly in $[0, \frac{DQ}{8}]$, or $3DQ = 2Dcp$, so we should expect the regret to be $\Omega(Dcp \sqrt{kT/Dcp}) = \Omega(\sqrt{kTDcpOA})$.

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

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

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

By definition, both have $k$ elements. Then, let $M_0$ be the MDPRM with $\varepsilon(o, a) = 0$ for all pairs in $\mathcal{L}^M$. Then let $M_j$ be the MDPRM with $\varepsilon(o, a) = \Delta$ for the $j$-th observation-action pair in the set $\mathcal{L}^M$. Similarly to (Lattimore & 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 $T/Dcp - 1$, or $T$ if the state $(q_0, s_0)$ is not visited enough:

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

Also, let $T_j$ be the number of visits to the $j$-th triplet of $\mathcal{L}$ until $T_{stop}$ and $T_{tot} = \sum_{j = 1}^{k} T_j$. 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 $\pi$ and $M_j$ and let $\mathbb{E}_j[i]$ 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 $Dcp(M_j)$ cross-product diameter of the MDPRM $M_j$:

$$Dcp(M_j) \leq DN + DN \sum_{i = 1}^{\infty} \frac{1}{2^i} + DN' \leq \frac{3}{2} QD = Dcp,$$
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 \( \Delta \) 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 \( D_{\text{cp}} E_0[T_0]/T \in [c_1, c_2] \). By construction, we have

\[
\frac{D_{\text{cp}} E_0[T_0]}{T} \leq \frac{E[T_0]}{OA} \leq \frac{T}{DN/\text{OA}} \leq c_2
\]

Similarly,

\[
\frac{D_{\text{cp}} E[T_0]}{T} \geq \frac{E_0[T_0]}{OA} \geq \frac{T}{D\text{N}/\text{OA}} \geq c_1.
\]

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

\[
E[\mathcal{R}_j(T)] \geq c_3 \Delta D_{\text{cp}} E[T_0 - T_j]
\]

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

\[
E[\mathcal{R}_j(T)] = \sum_{i=1}^{T} \mathcal{R}_j[T_i] - \sum_{i=1}^{T} \mathcal{R}_j[T_i],
\]

where \( \mathcal{R}_j \) 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:

\[
E[\mathcal{R}_j(T)] \geq \sum_{i=1}^{T} \mathcal{R}_j(T_i) \left( \frac{1}{2} + \Delta \right) \frac{DN}{4} - E[T_0 - T_j] \left( \frac{DN}{8} - \mathcal{R}_j[T_j] \left( \frac{1}{2} + \Delta \right) \right) \frac{DN}{4} - \mathcal{R}_j[T_j] \Delta \frac{DN}{4},
\]

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

**The Final Lower Bound.** Let \( D(P, Q) \) 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 \( D(P_0, P_j) = E_0[T_j] d(1/2, 1/2 + \Delta) \), 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 \( \Delta \) will satisfy \( \Delta \leq 1/4 \), then using the entropy inequalities from (Lattimore & Szepesvári, 2020, Equation 14.16), we have:

\[
D(P_0, P_j) \leq 4\Delta^2 E_0[T_j].
\]

Then following the same steps as in (Lattimore & Szepesvári, 2020, Chapter 38.7) and using Pinsker’s inequality, and using the fact that \( 0 \leq T_0 - T_j \leq T_0 - \frac{T}{D_{\text{cp}}} \), we have

\[
E_j[T_0 - T_j] \geq E[T_0 - T_j] - \frac{T}{D_{\text{cp}}} \sqrt{D(P_0, P_j)} \geq E_0[T_0 - T_j] - \frac{T\Delta}{D_{\text{cp}}} \sqrt{2E_0[T_j]}.
\]

Summing over \( j \) and applying Cauchy-Schwarz give us

\[
\frac{1}{k} \sum_{j=1}^{k} E_j[T_0 - T_j] \geq \frac{1}{k} \sum_{j=1}^{k} E_0[T_0 - T_j] - \frac{T\Delta}{D_{\text{cp}}} \sqrt{\frac{2E_0[T]_j}{2k}}
\]

\[
\geq \frac{c_1 T(k - 1)}{D_{\text{cp}}} - \frac{T\Delta}{D_{\text{cp}}} \sqrt{\frac{2c_3 T}{D_{\text{cp}}}.
\]
Now choosing $\Delta = \frac{c_1(k-1)}{2} \sqrt{\frac{D_{cp}}{2cT_k}}$ yields

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

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

$$E[R_j(T)] \geq c_3D_{cp}\Delta E_j[T_{tot} - T_j] \geq \frac{c_1^2c_3 T(k-1)^2}{4k} \sqrt{\frac{D_{cp}}{2cT_k}} = c_0 \sqrt{D_{cp}OAT},$$

with $c_0 > 0$ being a universal constant.

E. Details of Diameter Computations

**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 $q \in Q$ coincides with the diameter of the underlying MDP. Hence, $D_{o_0,q} = \frac{1}{\delta}$ for all $q \in Q$. Now, observe that the diameter for $\delta \in (0, \frac{1}{2})$ and $L \geq 2$ then we immediately have $D = \frac{1}{\delta}$, and the construction of MDPRM ensures that $D_{cp} > \frac{1}{\delta^2}$. Thus when $\delta \to 0$ we can immediately conclude that $D \to 1$ and $D_{cp} \to \infty$.

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 $D_{cp} \leq DN$ 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 matters 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](image)

<table>
<thead>
<tr>
<th>$O$</th>
<th>$\sqrt{O} K_{c,M}$</th>
<th>$\sqrt{c_{cp}^2}$</th>
<th>$D_{cp} \sqrt{\sum_{q,o,a} K_{(q,o),a}}$</th>
<th>$D_{cp} QO \sqrt{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>93.8</td>
<td>54.0</td>
<td>133.6</td>
<td>334.3</td>
</tr>
<tr>
<td>12</td>
<td>319.3</td>
<td>130.28</td>
<td>443.1</td>
<td>1551.1</td>
</tr>
<tr>
<td>20</td>
<td>726.0</td>
<td>229.6</td>
<td>1009.4</td>
<td>4542.5</td>
</tr>
<tr>
<td>40</td>
<td>2130.5</td>
<td>476.4</td>
<td>2978.0</td>
<td>18893.9</td>
</tr>
<tr>
<td>70</td>
<td>5005.4</td>
<td>846.1</td>
<td>7013.4</td>
<td>58783.2</td>
</tr>
<tr>
<td>100</td>
<td>8995.8</td>
<td>1219.8</td>
<td>12044.3</td>
<td>120745.6</td>
</tr>
</tbody>
</table>

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).
Table 2. Various quantities related to the regret bounds for the Multitask MDP RM 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).

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\sqrt{O \alpha c_M}$</th>
<th>$\sqrt{c_M'}$</th>
<th>$D_{cp} \sqrt{\sum_{q_0, a} K_{(q_0, a)}}$</th>
<th>$D_{cp} Q O \sqrt{\delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>468.0</td>
<td>272.4</td>
<td>3032.0</td>
<td>20339.3</td>
</tr>
<tr>
<td>5</td>
<td>468.1</td>
<td>272.4</td>
<td>3360.6</td>
<td>23100.5</td>
</tr>
<tr>
<td>6</td>
<td>468.2</td>
<td>272.5</td>
<td>3699.7</td>
<td>26029.4</td>
</tr>
<tr>
<td>8</td>
<td>504.2</td>
<td>293.2</td>
<td>4407.0</td>
<td>32384.4</td>
</tr>
<tr>
<td>10</td>
<td>550.1</td>
<td>319.5</td>
<td>5151.9</td>
<td>39404.6</td>
</tr>
<tr>
<td>12</td>
<td>600.2</td>
<td>348.3</td>
<td>5932.8</td>
<td>47090.0</td>
</tr>
</tbody>
</table>

Figure 8. The Multitask MDP RM.

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