Logarithmic regret in communicating MDPs: Leveraging known dynamics with bandits

Hassan Saber HASSAN.SABER@INRIA.FR
Fabien Pesquerel FABIEN.PESQUEREL@INRIA.FR
Odalric-Ambrym Maillard ODALRIC.MAILLLARD@INRIA.FR
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL, F-59000, Lille, France

Mohammad Sadegh Talebi M.SHAHI@DI.KU.DK
Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

We study regret minimization in an average-reward and communicating Markov Decision Process (MDP) with known dynamics, but unknown reward function. Although learning in such MDPs is a priori easier than in fully unknown ones, they are still largely challenging as they include as special cases large classes of problems such as combinatorial semi-bandits. Leveraging the knowledge on transition function in regret minimization, in a statistically efficient way, appears largely unexplored. As it is conjectured that achieving exact optimality in generic MDPs is NP-hard, even with known transitions, we focus on a computationally efficient relaxation, at the cost of achieving order-optimal logarithmic regret instead of exact optimality. We contribute to filling this gap by introducing a novel algorithm based on the popular Indexed Minimum Empirical Divergence strategy for bandits. A key component of the proposed algorithm is a carefully designed stopping criterion leveraging the recurrent classes induced by stationary policies. We derive a non-asymptotic, problem-dependent, and logarithmic regret bound for this algorithm, which relies on a novel regret decomposition leveraging the structure. We further provide an efficient implementation and experiments illustrating its promising empirical performance.

Keywords: Average-reward Markov decision process, regret minimization, logarithmic regret, Markov chain, recurrent classes

1. Introduction

In Reinforcement learning (RL), a learning agent (henceforth, learner) interacts with an environment that is often modeled using a Markov Decision Process (MDP), and her goal is to optimize a notion of reward (Puterman, 1994; Sutton and Barto, 2018). The learner does not fully know the underlying MDP, and tries to learn a near-optimal behavior quickly from the experience collected via interaction. In the average-reward setting, the learner's performance is often measured in terms of regret, which compares her cumulative reward to that of an optimal policy (Jaksch et al., 2010); equivalently, the learner's goal is to minimize regret. A standard assumption in most settings in RL is that the environment's dynamics is unknown, while the reward function may be known. This assumption is justified since state dynamics are not controlled by the learner, but is also in line with the argument that the main challenge in RL stems from unknown dynamics rather than unknown rewards.

Consequently, the vast majority of existing regret minimizing algorithms have some key ingredient, in design or analysis, to tackle unknown transition probabilities. In model-based algorithms (e.g., Jaksch et al. (2010); Filippi et al. (2010); Burnetas and Katehakis (1997)), this is featured in the form of confidence sets around the empirical transition distributions.

In contrast, in some applications of RL, the learner has some prior knowledge on the transition function; for example, she may know the associated support sets, some transition probabilities, or even the entire transition function up to some small deviation error. This could arise, for example, when the learner has access to an accurate estimate of the transition function via data collected while performing another task on the same environment (but with a different reward function). For instance, in the context of personalized recommendation, where the rewards are given by a user based on her internal evaluation of the recommendations, and where the task (hence transitions) is fixed across users, it is natural to assume that based on previous interactions, the transitions of the system are perfectly known, but the rewards associated to the current user are unknown. Note that although rewards are provided by a user, this does not mean they are known, as evaluation at a point in time can be subjective and noisy. Another scenario could arise in learning tasks where the dynamics are governed by some physical phenomena that are perfectly known to the learner. In such scenarios, the following question arises naturally: *What is the most statistically efficient way to perform exploration when the dynamics are known?*

While any form of prior knowledge on the transition function do not appear directly advantageous to model-free algorithms, which is in line with their design principle, model-based algorithms can benefit directly from it. In the case of perfectly known dynamics, most off-the-shelf algorithms can simply remove the relevant confidence sets, which would lead to improved exploration, and hence, smaller regret bounds.[1] Despite such straightforward modifications of model-based algorithms, it still remains open as to what the best way is to incorporate such prior knowledge into algorithm design in a non-trivial manner, and whether it could lead to instance-dependent (and logarithmic) regret bounds. To our best knowledge, existing literature on learning in MDPs, albeit rich, fails to provide algorithmic ideas to leverage such prior knowledge in a statistically efficient way, and the potential gains thereof in terms of regret or sample complexity remain largely unexplored.

Contributions We focus on regret minimization in communicating MDPs with known dynamics but unknown reward functions, and introduce a class of strategies called *rarely-switching algorithms*, which provide a principled way to leverage the connectivity structure in the MDP through viewing the problem as a multi-policy Multi-Armed Bandit (MAB), thanks to the prior knowledge on the dynamics. The novel design of these strategies considers recurrent classes induced by stationary policies as well as a carefully designed stopping criterion based on the said classes. For these strategies, we present a generic regret bound, which relies on a novel regret decomposition leveraging the structure, which could be of independent interest for learning in MDPs in general. Then, we instantiate a specific rarely-switching algorithm called `IMED-KD`, which uses the popular Indexed Minimum Empirical Divergence (IMED) strategy for MABs (Honda and Takemura, 2015). IMED

1. For example, it is straightforward to show that UCRL2 (Jaksch et al., 2010), when equipped with the knowledge on dynamics, attains a regret bound of $\widetilde{\mathcal{O}}(\sqrt{(SA+D)T})$ with high probability, in any communicating MDP with S states, A actions, and diameter D, where $\widetilde{\mathcal{O}}(\cdot)$ hides $\log(T)$ terms and universal constants. In contrast, without prior knowledge, UCRL2 achieves a regret of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$.

offers an interesting alternative to optimistic strategies such as UCB or KL-UCB, and to Bayesian strategies such as Thompson sampling. Owing to its form directly inspired by the constraints of the optimization problem appearing in asymptotic regret lower bounds, it has been shown to yield optimal regret performance, like KL-UCB or Thompson Sampling. We stress that a key departure from existing IMED-style algorithms for MDPs (e.g., IMED-RL (Pesquerel and Maillard, 2022)) is to exploit the intrinsic structure of the problem via use of a rarely-switching algorithm. Under some standard assumption on the reward function and MDP regularity, as well as a mild assumption on the involved hitting times (Assumption 5), we derive a non-asymptotic, problem-dependent, and logarithmic regret bound for `IMED-KD`, whose proof relies on the generic properties of rarely-switching algorithms as well as proof machinery of IMED-style indices adapted to MDPs. We further provide an efficient implementation and experiments illustrating its promising empirical performance. To the best of our knowledge, `IMED-KD` is the first algorithm specifically designed to leverage the structure in MDPs with known dynamics.

Related work There is a rich literature on regret minimization in average-reward MDPs. Early papers like (Burnetas and Katehakis, 1997; Graves and Lai, 1997) mostly presented regret bounds for ergodic MDPs and with an asymptotic flavour, whereas more recent literature, e.g., (Jaksch et al., 2010; Filippi et al., 2010; Talebi and Maillard, 2018; Fruit et al., 2018; Zhang and Ji, 2019; Wei et al., 2020; Bourel et al., 2020; Pesquerel and Maillard, 2022), reported non-asymptotic regret guarantees and, often, for the bigger of class of (weakly) communicating MDPs. The majority of recent literature on learning in MDPs, following Jaksch et al. (2010), report worst-case regret bounds growing as $\widetilde{\mathcal{O}}(\sqrt{T})$ after T steps. In contrast, comparatively there exists little work that present logarithmic and instance-dependent regret bounds for average-reward MDPs. The most notable exceptions include (Jaksch et al., 2010), which reports a logarithmic regret bound for UCRL2 (albeit with a large mixing-time related additive term), and more recent papers (Gopalan and Mannor, 2015; Pesquerel and Maillard, 2022), which only consider ergodic MDPs. We also mention the logarithmic regret bounds derived in (Ortner, 2009; Tranos and Proutiere, 2021) for the much simpler setting of MDPs with deterministic transitions.

We also mention that some studies consider regret minimization in MDPs in the *episodic* setting, with a fixed and known horizon; see, e.g., Osband et al. (2013); Azar et al. (2017); Simchowitz and Jamieson (2019), where the latter work presents a problem-dependent, logarithmic regret bound. However, the proof machinery used in episodic RL often fails to work in average-reward RL due to relying on the fixed episode length and resetting of the state. Finally, it is worth remarking that an MDP with known transitions but unknown rewards may be viewed as a MAB instance with highly structured actions (one action corresponding to a policy), in a way which is reminiscent of combinatorial MABs (Chen et al., 2013; Combes et al., 2015). Despite such resemblance, the problem is more challenging as the learner is traversing an MDP without a resetting device. As a result, algorithmic ideas for combinatorial MABs or those with generic structure (Combes et al., 2017; Saber et al., 2020) do not directly carry over to MDPs with known dynamics.

Notations For an integer $n \in \mathbb{N} \cup \{0\}$, we denote $[n] = \{0, \dots, n\}$. For a Boolean event A, $\mathbb{I}\{A\} \in \{0, 1\}$ denotes the indicator function of A. For a sequence $(h_t)_{t\in\mathbb{N}}$, and $t_1 < t_2 \in \mathbb{N}$, $h_{t_1:t_2} := (h_t)_{t\in\{t_1,\dots,t_2\}}$ denotes the sub-sequence of elements indexed in between t_1 and t_2. Last, for a set A, $\mathcal{P}(A)$ denotes the set of probability distributions over A.

2. Problem formulation

We consider an average-reward Markov Decision Process $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$, where $\mathcal{S}$ is the set of states with cardinality S, and $\mathcal{A} = (\mathcal{A}_s)_{s\in\mathcal{S}}$, where $\mathcal{A}_s$ specifies the set of actions available in $s \in \mathcal{S}$. For convenience, we introduce the set of pairs $\mathcal{C} = \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$. Further, $\mathbf{p} : \mathcal{C} \to \mathcal{P}(\mathcal{S})$ denotes the transition function, and $\mathbf{r} : \mathcal{C} \to \mathcal{P}(\mathbb{R})$ the reward function. We denote the corresponding mean reward function by $\mathbf{m} : \mathcal{C} \to \mathbb{R}$.

Policies Each stationary policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ acting on $\mathbf{M}$ induces a Markov chain on $\mathcal{C}$, with corresponding transition probability $\mathbf{p}_\pi : \mathcal{C}^2 \to \mathcal{P}(\mathcal{C})$, defined by $\mathbf{p}_\pi(s,a)(s',a') = \mathbf{p}(s'|s,a)\pi(a'|s')$. We denote by $\overline{\mathbf{p}}_\pi : \mathcal{C}^2 \to \mathcal{P}(\mathcal{C})$ the Cesaro-average of $\mathbf{p}_\pi$; formally, $\overline{\mathbf{p}}_\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{p}_\pi^{t-1}$, where $\overline{\mathbf{p}}_\pi(c_1, c)$ captures the frequency of reaching the pair $c \in \mathcal{C}$ under policy π starting in pair $c_1 \in \mathcal{C}$. This enables to introduce the gain of policy π, when starting from state-action pair $c_1 = (s_1, a_1)$, defined by $\mathbf{g}_{c_1,\pi} := (\overline{\mathbf{p}}_\pi \mathbf{m})(c_1) = \sum_{c\in\mathcal{C}} \overline{\mathbf{p}}_\pi(c_1, c) \cdot \mathbf{m}(c)$, where we recall that $\mathbf{m}(c)$ is the mean reward of pair c. Given a *finite* set of stationary policies Π, we define $\mathbf{g}_c^\star = \max_{\pi\in\Pi} \mathbf{g}_{c,\pi}$ the optimal gain starting from c, and $\Pi_c^\star = \{\pi \in \Pi : \mathbf{g}_{c,\pi} = \mathbf{g}_c^\star\}$ the set of policies achieving the optimal gain.

Cycles The set of *(positive) recurrent* state-action pairs (i.e., pairs with finite return times) under π is defined as $\mathcal{C}_\pi^+ = \{c \in \mathcal{C} : \overline{\mathbf{p}}_\pi(c)(c) > 0\}$. Further, the relation $\sim_\pi$ such that $c \sim_\pi c' \Leftrightarrow \overline{\mathbf{p}}_\pi(c)(c') \cdot \overline{\mathbf{p}}_\pi(c')(c) > 0$, is an equivalence relation on $\mathcal{C}_\pi^+$. Denoting $[c]_\pi$ the class of $c \in \mathcal{C}_\pi^+$ for relation $\sim_\pi$, the asymptotic *cycles* under policy π are defined as $\mathcal{X}_\pi = \mathcal{C}_\pi^+ / \sim_\pi = \left\{ [c]_\pi : \ c \in \mathcal{C}_\pi^+ \right\}$. Distinct elements of $\mathcal{X}_\pi$ correspond to disjoint cycles. A policy π with $|\mathcal{X}_\pi| = 1$ is called a *unichain* policy.

Remark 1 *A remarkable property is that for a* unichain policy π *and recurrent* $c' \in \mathcal{C}_\pi^+$, $\overline{\mathbf{p}}_\pi(c)(c')$ *is independent of the starting pair* c *and equals* $1/\boldsymbol{\tau}_\pi(c',c')$, *where* $\boldsymbol{\tau}_\pi(c',c')$ *is the expected hitting time of* c' *when starting from* c' *and following policy* π*; see (Puterman, 1994). As a consequence,* $\mathbf{g}_{c,\pi}$ *also does not depend on* c.

We consider the two following assumptions on MDP regularity and the reward function:

Assumption 1 (MDP) $\mathbf{M}$ *is communicating, that is,* $\forall c, c', \exists \pi, t \in \mathbb{N} : \mathbf{p}_\pi^t(c)(c') > 0$. *Also,* Π *is proper, that is, the Cesaro-average* $\overline{\mathbf{p}}_\pi$ *of* $\mathbf{p}_\pi$ *exists for each* $\pi \in \Pi$. *There is a unique gain-optimal policy* $\pi^\star \in \bigcap_{c\in\mathcal{C}} \Pi_c^\star$ *that is* unichain *(i.e., it has a unique asymptotic cycle).*

Assumption 2 (Reward function) *For each* $c \in \mathcal{C}$, *the reward distribution* $\mathbf{r}(c)$ *is supported on* $[0,1]$ *(in particular, it is* $1/2$*-sub-Gaussian), with bounded mean* $\mathbf{m}(c) \in [0,1)$.

In particular, under Assumption 2, $\forall c \in \mathcal{C}, \pi \in \Pi$, the gain is bounded: $\mathbf{g}_{c,\pi} \in [0,1)$. To gain insight into the motivations behind this assumption, we refer to discussion in Appendix F.

Local monotony Finally, a key property of *unichain* MDPs is that there always exists a modification of a sub-optimal policy in a single state having (stricly) larger gain (Puterman, 1994). We generalize this useful monotony property to *larger neighborhoods* as follows:

Assumption 3 (Policy-improving neighborhood) $\forall c \in \mathcal{C}, \pi \notin \Pi_c^\star, \exists \pi' \in \Pi, \ \boldsymbol{h}(\pi,\pi') \leqslant k$, *such that* $\mathbf{g}_{c,\pi'} > \mathbf{g}_{c,\pi}$, *where* $\boldsymbol{h}$ *denotes the Hamming distance between two policies.*

Here, k is a given constant. Note that as k increases from 1 to S, Assumption 3 interpolates between (at least) all unichain MDPs, when $k = 1$, and all discrete MDPs, when $k = S$.

Remark 2 *In a communicating MDP, for* Π *consisting of all stationary policies, the set of optimal policies does not depend on the starting pair and is simply denoted by* $\Pi^\star$*. Moreover, an optimal policy is also unichain in this case. The gain of a unichain policy* π *does not depend on the starting state-action pair, and is simply denoted by* $\mathbf{g}_\pi$ *(and* $\mathbf{g}^\star$ *for an optimal policy); hence, we denote* $\Pi^\star_c = \Pi^\star$ *and* $\mathbf{g}^\star_c = \mathbf{g}^\star$ *for all* c*. Note, however, that for sub-optimal policies* $\pi \in \Pi$ *that are not unichain,* $\mathbf{g}_{c,\pi}$ *may still depend on the initial state-action* $c \in \mathcal{C}$*.*

The online learning problem The learner interacts with MDP $\mathbf{M}$ for T time steps, starting in an initial state-action pair $(s_1, a_1) \in \mathcal{C}$ chosen by Nature. At each time $t \geq 2$, she is in state $s_t \in \mathcal{S}$ and chooses an action $a_t \in \mathcal{A}_{s_t}$ according to a stationary policy $\pi_t \in \Pi$, that is $a_t \sim \pi_t(s_t)$. The stationary policy π_t is selected based on the learner's observations so far. Then, (i) she receives a reward $r_t \in [0, 1]$, where $r_t \sim \mathbf{r}(s_t, a_t)$; and (ii) Nature decides a next state $s_{t+1} \in \mathcal{S}$, where $s_{t+1} \sim \mathbf{p}(\cdot|s_t, a_t)$. The sequence of chosen policies is denoted by $(\pi_t)_{t \geqslant 1}$, and simply by (π) when for all time step $t \geqslant 1$, $\pi_t = \pi$. Further, we denote by $c_t = (s_t, a_t)$ the state-action pair at time step t. We assume that the learner *does not know* the reward function $\mathbf{r}$, but *knows* the transition function $\mathbf{p}$, and can thus compute $\overline{\mathbf{p}}_\pi$ for each $\pi \in \Pi$. Her performance is measured through the notion of (expected) regret, as defined next. Let $V_{\mathbf{M}}(\mathbb{A}, T)$ denote the cumulative reward of an algorithm $\mathbb{A}$ following a policy sequence $(\pi_t)_{t \leqslant T}$ up to time T:

$$V_{\mathbf{M}}(\mathbb{A}, T) = \mathbb{E}_{(\pi_t)}\left[\sum_{t=1}^{T} r_t\right].$$

For a policy sequence (π), it is simply denoted by $V_{\mathbf{M}}((\pi), T)$. The (expected) regret with respect to playing a gain-optimal policy sequence $(\pi^\star)$, up to time T, is defined as:

$$\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T) = V_{\mathbf{M}}((\pi^\star), T) - V_{\mathbf{M}}(\mathbb{A}, T). \tag{1}$$

Remark 3 (Pseudo-regret) *For each given* T*, the quantity* $V^\star_{\mathbf{M}}(T) = \max_{\pi \in \Pi} V_{\mathbf{M}}((\pi), T)$ *and the set* $\operatorname{Argmax}_{\pi \in \Pi} V_{\mathbf{M}}((\pi), T)$ *differ a priori from the cumulative reward of gain-optimal policies* $(V_{\mathbf{M}}((\pi^\star), T))_{\pi^\star \in \Pi^\star}$ *and the set* $\Pi^\star$*, respectively. However, it is easily checked that* $\lim_{T \to \infty} V^\star_{\mathbf{M}}(T)/T = \lim_{T \to \infty} V_{\mathbf{M}}((\pi^\star), T)/T = \mathbf{g}^\star$*, for all gain-optimal stationary policy* $\pi^\star \in \Pi^\star$*. That is, the asymptotic maximal average reward coincides both with the asymptotic average reward of gain-optimal policies and the optimal gain. Since the set of considered stationary policies* Π *is finite, this further implies that* $\Pi^\star_T \subset \Pi^\star$ *when horizon* T *is large enough, which also implies* $V^\star_{\mathbf{M}}(T) - V_{\mathbf{M}}((\pi^\star), T) \overset{T \to \infty}{=} O(1)$*.*

3. Rarely-switching Algorithms

Rarely-switching learners We choose to restrict the learner to follow a *rarely-switching* strategy, which forces the learner to keep playing the same policy until some criterion — to be introduced momentarily — is met. The T time steps are divided into episodes of

random durations, where episode $k \in \{1, 2, \dots\}$ starts at random time $\tau_{k-1}+1$ and ends at random time τ_k (with $\tau_0 = 0$). We gather in the sequence $\mathcal{T} = (\tau_k)_{k\in\mathbb{N}}$ the last time step before starting each new episode. Hence, for $\tau \in \mathcal{T}$, the learner starts at time step $\tau+1$ a new episode (to which we refer as "episode τ"), and after pulling state-action pair $c_{\tau+1} = (s_{\tau+1}, a_{\tau+1})$, she follows the same policy $\pi = \pi_{\tau+1}$ until the event `Event` is triggered (and the episode ends). `Event` is a generic function of the current policy π and the history $h_{\tau+1:t}$ of all observations and decisions made from the beginning of the episode until the current time. We resume the generic structure of rarely-switching learners in Algorithm 1.

Algorithm 1 Rarely-switching learner

input: $(\mathbf{p}_\pi)_{\pi\in\Pi}$, (s_1, π_1) and `Event` function
Start a new episode $\tau \leftarrow 0$, $\pi \leftarrow \pi_1$
Pull action $a_1 \sim \pi(s_1)$
for time step $t \geqslant 1$ **do**
  Receive reward r_t, update history $h_t = (s_t, \pi, a_t, r_t)$
  if $\neg$`Event`$(\pi, h_{\tau+1:t})$ **then**
    Keep the same policy $\pi \leftarrow \pi_{\tau+1}$
    Pull action $a_{t+1} \sim \pi(s_{t+1})$
  else
    Start a new episode $\tau \leftarrow t$
    Compute a new policy $\pi_{\tau+1}$ and update $\pi \leftarrow \pi_{\tau+1}$.
    Pull action $a_{t+1} \sim \pi(s_{t+1})$
  end if
end for

Counters For a rarely-switching learner, let $N^{\mathrm{ini}}_{c,\pi}(0:T) = \sum_{\tau\in\mathcal{T}\cap[T]} \mathbb{I}\{\pi_{\tau+1} = \pi, c_{\tau+1} = c\}$ denote the number of times when an episode starts in pair c and follows policy π until time T. This quantity should not be confused with the (possibly much larger) number of visits $N_{c,\pi}(T) = \sum_{t\in[T]} \mathbb{I}\{\pi_t = \pi, c_t = c\}$ of pair c by policy π until time T. In view of the introduction of `Event`, it is also convenient to introduce $N_c(h) = \sum_{(s,\pi,a,r)\in h} \mathbb{I}\{(s,a) = c\}$ that counts the number of visits of pair c on the piece of history h.

Owing to the fact that the criterion used to stop an episode is *independent* of the rewards accumulated during the episode, and using properties of the expectation, we can show the following decomposition lemma, somewhat reminiscent of bandit analyses.

Assumption 4 (Whole number of episodes) *We assume that $T \in \mathcal{T}$, that is horizon time T coincides with the last time step of an episode. We abusively conserve the notation $\mathbb{E}_{(\pi_t)}[Z]$ instead of $\mathbb{E}_{(\pi_t)}[Z|T \in \mathcal{T}]$ to compute the expectation of any random variable Z.*

Lemma 4 (Cumulative reward and regret decomposition) *Under Assumption 4, the cumulative reward of a rarely-switching algorithm $\mathbb{A}$ satisfies*

$$V_{\mathbf{M}}(\mathbb{A}, T) = \sum_{c\in\mathcal{C}} \sum_{\pi\in\Pi} \mathbb{E}_{(\pi_t)}\big[N^{ini}_{c,\pi}(0:T)\big] \cdot \mathbb{E}[\ell_{c,\pi}] \cdot G_{c,\pi}\,,$$

where $\ell_{c,\pi} = \min\{t > 0 : \texttt{Event}(\pi, h_{1:t}), c_1 = c\}$ denotes the (random) length of the episode, and where $G_{c,\pi} = \mathbb{E}_{(\pi_t)}\Big[\frac{1}{\ell_{c,\pi}} \sum_{t=1}^{\ell_{c,\pi}} r_t \Big| (s_1, a_1) = c\Big]$ denotes the expected average reward of an episode starting in pair c and following policy π. When `Event` further ensures an episode

running policy π always stops in a same reference pair $c_\pi \in \mathcal{C}$, then writing $G^\star = G_{c_{\pi^\star},\pi^\star}$, it holds

$$\begin{aligned}\mathcal{R}_{\mathbf{M}}(\mathbb{A},T) &= \sum_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbb{E}_{(\pi_t)}\big[N^{ini}_{c,\pi}(0{:}T)\big]\cdot\mathbb{E}[\ell_{c,\pi}]\cdot(G^\star - G_{c,\pi}) \\ &\quad + \sum_{c\neq c_{\pi^\star}} \mathbb{E}_{(\pi_t)}\big[N^{ini}_{c,\pi^\star}(1{:}T)\big]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star - G_{c,\pi^\star})\,. \qquad (2)\end{aligned}$$

Note that $N^{\mathrm{ini}}_{c,\pi^\star}(1{:}T)$ excludes the first episode. Furthermore, we stress that $G^\star$ is defined using the stopping time induced by $\pi^\star$, and $G_{c,\pi}$ by the one induced by π.

The proof of Lemma 4 is provided in Appendix A. To give some intuition, Lemma 4 decomposes the cumulative reward of a rarely-switching learner according to *each configuration* when the policy being played is π and the initial pair in this episode is c. Thus, it makes appear the number of times such a configuration happens, $\mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi}(0{:}T)\big]$, as well as the reward accumulated in that episode. A similar decomposition can be written for the optimal policy, and using a reference state ensures that $R_{\mathbf{M}}(\pi^\star,T) \simeq TG^\star$, up to the contribution of the first episode in which the episode may not start from $c_{\pi^\star}$. Combining the two cumulative reward decompositions yields the convenient form in Equation (2).

Further, the product form term $\mathbb{E}[\ell_{c,\pi}]\cdot G_{c,\pi}$ reveals that Lemma 4 offers a decoupling between the expected number $\mathbb{E}[\ell_{c,\pi}]$ of steps of an episode starting in c with policy π, and its average reward $G_{c,\pi}$ received during that episode. It is worth mentioning that the decoupling between the gain and the length of an episode holds by virtue of the Markov property and since we consider a decomposition in expectation.

Remark 5 (Simplifications) *Note that $\mathbb{E}_{(\pi_t)}\big[N^{ini}_{c,\pi}(0{:}T)\big] = 0$ for policies π not explored by a rarely-switching algorithm. Typically, a learning algorithm will progressively focus on a few policies, and hence, the sum over all stationary policies π should effectively involve much fewer terms than A^S (i.e., the number of all stationary deterministic policies). Interestingly, in the case of bandits, there is a unique state, and hence, Equation (2) simplifies to the classical regret decomposition, in which case the second term disappears:*

$$\sum_{c\neq c_{\pi^\star}} \mathbb{E}_{(\pi_t)}\big[N^{ini}_{c,\pi^\star}(1{:}T)\big]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star - G_{c,\pi^\star}) = 0\,.$$

Gain One may wonder about the link between $G_{c,\pi}$ and the gain $\mathbf{g}_{c,\pi}$: $G_{c,\pi}$ can be seen as a proxy for the gain $\mathbf{g}_{c,\pi}$ of the policy, since $\mathbf{g}_{c,\pi} = \lim_{T\to\infty} \mathbb{E}_{(\pi_t)}\Big[\frac{1}{T}\sum_{t=1}^{T} r_t \Big| c_1 = c, \pi_1 = \pi\Big]$, that is, as $\ell_{c,\pi}\to\infty$, then $G_{c,\pi}$ indeed approaches $\mathbf{g}_{c,\pi}$. This interpretation is however valid only when $\ell_{c,\pi}$ is sufficiently large. Luckily, thanks to the regenerating properties of the chain, if we start and stop an episode in the same *recurrent* pair c_π, hence "completing a loop", then the average of the rewards received during that episode must, in expectation, equal that of infinitely many such loops. More formally:

Lemma 6 (Regeneration property) *For any unichain policy π, any recurrent reference pair $c_\pi \in \mathcal{C}^+_\pi$, and any function* `Event` *ensuring that an episode always stops in c_π when we play π, then $G_{c_\pi,\pi} = \mathbf{g}_{c_\pi,\pi}$, that is the expected average reward received during an episode starting and ending at pair c_π is equal to the gain of the policy.*

A proof of Lemma 6 is provided in Appendix A.1. This motivates us to introduce for each π a reference pair $c_\pi \in \operatorname{Argmax}_{c\in\mathcal{C}} \overline{\mathbf{p}}_\pi(c)(c)$ (which belongs to $\mathcal{C}_\pi^+$ by construction), and define $\texttt{Event}(\pi, h_{\tau+1:t})$ to ensure that $(s_t, a_t) = c_\pi$. Indeed, this choice of c_π also minimizes $\boldsymbol{\tau}_\pi(c,c)$ over c, hence tends to reduce $\mathbb{E}[\ell_{c_\pi,\pi}]$. This construction of events further yields the following useful control on the regret:

Proposition 7 (Rarely-switching learners with reference pair) *Under Assumption 1, if the rarely-switching learner $\mathbb{A}$ specifies for each π to stop the episode starting with π in the same reference pair $c_\pi \in \mathcal{C}_\pi^+$, then the following bound holds almost surely:*

$$\sum_{c\neq c_{\pi^\star}} N_{c,\pi^\star}^{ini}(1\!:\!T) \leqslant \sum_{c\in\mathcal{C}} \sum_{\pi\neq\pi^\star} N_{c,\pi}^{ini}(0\!:\!T)\,.$$

Moreover, the cumulative regret of any such rarely-switching algorithm $\mathbb{A}$ with respect to the unique optimal policy $\pi^\star$, up to the end T of any episode, is upper-bounded by

$$\mathcal{R}_{\mathbf{M}}(\mathbb{A},T) \;\leqslant\; \mathbb{E}_{(\pi_t)}\left[\sum_{c\in\mathcal{C},\pi\neq\pi^\star} N_{c,\pi}^{ini}(0\!:\!T)\right]\cdot\left(\max_{(c,\pi)\neq(c_{\pi^\star},\pi^\star)} \mathbb{E}[\ell_{c,\pi}](G^\star - G_{c,\pi}) + \mathbf{B}_\star\right),$$

where $\mathbf{B}_\star := \max_{c\neq c_{\pi^\star}} \mathbb{E}_{(\pi_t)}[\ell_{c,\pi^\star}](G^\star - G_{c,\pi^\star})$ is a problem-dependent quantity.

Remark 8 *It holds $\mathbf{B}_\star \leqslant \max_{(c,\pi)\neq(c_{\pi^\star},\pi^\star)} \mathbb{E}[\ell_{c,\pi}](G^\star - G_{c,\pi})$. Further, $\mathbf{B}_\star = 0$ for bandits.*

Estimation and covering time Before we specify the algorithm, let us remind that since the transitions are known, only the mean rewards need to be estimated. Since $\mathbf{g}_{c,\pi} = \sum_{c'\in\mathcal{C}_\pi^+} \overline{\mathbf{p}}_\pi(c)(c')\mathbf{m}(c')$, where $\mathbf{m}$ is unknown, it is natural to collect observations of pairs $c' \in \mathcal{C}_\pi^+$ to estimate the corresponding $\mathbf{m}(c')$, and hence the gain $\mathbf{g}_\pi$. A natural way to ensure the estimation error reduces in each episode is to stop an episode when all pairs in $\mathcal{C}_\pi^+$ have been visited at least once: Formally, $\min_{c'\in\mathcal{C}_\pi^+} N_{c'}(h_{\tau+1:t}) > 0$, that is after *covering* the set $\mathcal{C}_\pi^+$. In order to control the resulting episode length, unfortunately, there is in general no simple control of the cover time by a policy π of its recurrent pairs. The policy could be diffusive or lazy (see Appendix B), yielding an arbitrarily large cover time. Formally, given $C \subset \mathcal{C}$ and $c \in \mathcal{C}$, we denote by $\pi_c^H(C)$ a policy that minimizes over policies π the expected time $\boldsymbol{\tau}_{c,\pi}^H(C)$ to reach *any element* of C starting from c and following π. In a similar manner, we let $\overline{\pi}_c(C)$ denote a policy minimizing over π the expected time $\overline{\boldsymbol{\tau}}_{c,\pi}(C)$ to cover *all elements* of C starting from c and following π. Letting $D_{\mathbf{M}}$ denote the diameter of $\mathbf{M}$,[2] it holds: $\min_\pi \boldsymbol{\tau}_{c,\pi}^H(C) \leqslant D_{\mathbf{M}}$ and $\min_\pi \overline{\boldsymbol{\tau}}_{c,\pi}(C) \leqslant |C| D_{\mathbf{M}}$ for all c and C. In contrast, $\overline{\boldsymbol{\tau}}_{c,\pi}(\mathcal{C}_\pi^+)$ could be *arbitrarily large*, even for a gain-optimal policy π.

2. The diameter of a finite MDP $\mathbf{M}$ is defined as $D_{\mathbf{M}} = \max_{s\neq s'} \min_\pi \mathbb{E}[T^\pi(s,s')]$, where $T^\pi(s,s')$ denotes the number of steps it takes to reach s' starting from s and following policy π (Jaksch et al., 2010).

Frequently recurrent pairs and restricted gain This motivates us to discard states with *too small return frequency*. To formalize this, we introduce a notion of gain, which we call η-*restricted gain*, defined using a parameter $\eta \in \mathbb{R}^+$. Formally, for a constant $\eta \in \mathbb{R}^+$, define the set of frequently recurrent pairs of a stationary policy π:

$$\text{(Frequently recurrent pairs)} \qquad \mathcal{C}^+_{c,\pi}(\eta) := \{c' \in \mathcal{C} : \overline{\mathbf{p}}_\pi(c)(c') > \eta\}\,,$$

which leads to defining the corresponding η-restricted gain function:

$$(\eta\text{-restricted gain}) \qquad \mathbf{g}_{c,\pi}(\eta) := \sum_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c)(c') \cdot \mathbf{m}(c') \Big/ \sum_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c)(c')\,.$$

We further naturally introduce $\mathbf{g}^\star_c(\eta) = \max_{\pi \in \Pi} \mathbf{g}_{c,\pi}(\eta)$ and $\Pi^\star_c(\eta) = \operatorname{Argmax}_{\pi \in \Pi} \mathbf{g}_{c,\pi}(\eta)$. Note that for $\eta = 0$, we recover the usual definitions (e.g., $\mathbf{g}_{c,\pi}(0) = \mathbf{g}_{c,\pi}$). More generally:

Lemma 9 (Restricted-gain approximation)

$$\forall \pi, c, \eta, \quad \mathbf{g}_{c,\pi} - \mathbf{g}_{c,\pi}(\eta) \;\leqslant\; \eta \mathbf{m}_{\max} |\mathcal{C}^+_{c,\pi} \setminus \mathcal{C}^+_{c,\pi}(\eta)|\,,$$

where $\mathbf{m}_{\max} = \max_{c \in \mathcal{C}} \mathbf{m}(c)$ *is the maximal state-action pair mean.*

This lemma is proven in Appendix C. In particular, for a given ε, choosing $\eta \leqslant \frac{\varepsilon}{\mathbf{m}_{\max} |\mathcal{C}^+_{c,\pi} \setminus \mathcal{C}^+_{c,\pi}(\eta)|}$ (for instance, $\eta = \varepsilon/(\mathbf{m}_{\max} S)$) ensures that the gain is still well-approximated by the η-restricted gain up to the desired precision ε. Hence, we can restrict to cover $C = \mathcal{C}^+_{c,\pi}(\eta)$ instead of $\mathcal{C}^+_{c,\pi}$ and define $\texttt{Event}(\pi, h_{\tau+1:t})$ accordingly. Unfortunately, $\overline{\boldsymbol{\tau}}_{c,\pi}(\mathcal{C}^+_\pi(\eta))$ can still be arbitrary in general. This motivates us to introduce:

Definition 10 (Laziness) *A chain induced by* π *is* (B, η)*-lazy if* $\max_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \overline{\boldsymbol{\tau}}_{c',\pi}(\mathcal{C}^+_{c,\pi}(\eta)) > B$.

We assume (the laziness constant B may be unknown to the learner, or computed offline):

Assumption 5 (No-laziness) $\mathbf{M}$ *has no* (B, η)*-lazy chain, where* $\eta \in [0, 1]$ *is given.*

Structure of policies We conclude this section by showing that choosing this specific form of event further enables us to revisit the decomposition of regret to better exploit structure of the policies. Indeed, while $\mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi}(0\!:\!T)\big] = 0$ for policies π not explored by a rarely-switching learner, there is more: policies are structured, in the sense that visiting one state-action pair (s, a) is not only informative about the actual policy π playing a in state s, but all such ones as well. Using Proposition 7 and the form of stopping event introduced in Lemma 12 (in the next section), we derive the following result, showing, remarkably, that the **sum over all policies can be removed in favor of a maximum**.

Theorem 11 (Rarely-switching learners exploiting recurrence structure) *Let* $\mathbb{A}$ *be a rarely-switching algorithm using stopping event* $\texttt{Event}(\pi, h_{\tau+1:t}) = \{\min_{c' \in C} N_{c'}(h_{\tau+1:t}) > 0 \text{ and } (s_t, a_t) = c_\pi\}$ *where* $C = \mathcal{C}^+_{c,\pi}(\eta)$ *is parameterized by* η. *Then,*

$$\sum_{c \in \mathcal{C},\ \pi \neq \pi^\star} N^{ini}_{c,\pi}(0\!:\!T) \leqslant |\mathcal{C}| \max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \mathbf{N}^\eta_{c,\pi}(T)\,,$$

where we introduced $\mathbf{N}^{\eta}_{c,\pi}(t) = \min_{c' \in \mathcal{C}^{+}_{c,\pi}(\eta)} N_{c'}(h_{1:t})$. *In particular, using Remark 8,*

$$\frac{\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T)}{\mathbb{E}_{(\pi_t)}[\max_{c \in \mathcal{C}, \pi \neq \pi^\star} \mathbf{N}^{\eta}_{c,\pi}(T)]} \leqslant |\mathcal{C}| \Bigg(\underbrace{\max_{(c,\pi) \neq (c_{\pi^\star}, \pi^\star)} \mathbb{E}[\ell_{c,\pi}]}_{=:\mathbf{L}} \underbrace{(G^\star - G_{c,\pi})}_{\in [-1,1]} + \mathbf{B}_\star \Bigg) \leqslant 2|\mathcal{C}|\mathbf{L}\,. \tag{3}$$

4. The IMED-KD strategy

In this section, we present `IMED-KD` (Indexed Minimum Empirical Divergence for MDPs with Known Dynamics), which is a rarely-switching algorithm that uses an IMED-type index together with the knowledge of $\mathbf{p}$ to attain a logarithmic regret in communicating MDPs. The IMED strategy (Honda and Takemura, 2015) has been proven asymptotically optimal in stochastic MABs and is computationally appealing when compared with the optimistic KL-UCB or the Bayesian Thompson sampling (TS) strategy that require, at each step, solving an optimization problem or sampling from a posterior, respectively. Although posterior sampling can be made efficient for some parametric distributions such as Gaussians, current extensions of TS to MDPs require introducing a forced optimism mechanism (Agrawal and Jia, 2017), which makes it less appealing both from theory and computational perspectives.

High-level description At a high level, the algorithm computes at the beginning of each episode τ an empirical best candidate policy $\widehat{\pi}^\star_\tau$, as well as a best informative policy $\widehat{\pi}^I_\tau$. The algorithm considers the stopping event targeting $C = \mathcal{C}^{+}_{c_\tau,\pi}(\eta)$ and final pair $c_0 = c_\pi$ for the policy $\pi = \widehat{\pi}^I_\tau$. It runs the episode using $\widehat{\pi}^H_\tau$ until hitting C, followed by policy π (so if $c_\tau \in C$, this reduces to running π). We now detail the computation of $\widehat{\pi}^\star_\tau, \widehat{\pi}^I_\tau$, and $\widehat{\pi}^H_\tau$.

a. Empirical best policy $\widehat{\pi}^\star_\tau$ is computed via classical value (or policy) iteration algorithms in the MDP $\widehat{\mathbf{M}}_\tau = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \widehat{\mathbf{r}}_\tau)$ where for each $c \in \mathcal{C}$, we introduce $\widehat{\mathbf{r}}_\tau(c) = \mathcal{N}(\widehat{\mathbf{m}}_\tau(c), \sigma^2)$ with $\widehat{\mathbf{m}}_\tau(c) = \frac{1}{N_c(h_\tau)} \sum_{t'=1}^{\tau} r_{t'} \mathbb{I}\{c_{t'} = c\}$ being the classical empirical estimate of the mean $\mathbf{m}(c)$ computed on observations received until time τ.

b. Informative policy To compute $\widehat{\pi}^I_\tau$, we first form $\widehat{\mathbf{g}}^\star_{c,\tau}(\eta) = \widehat{\mathbf{g}}_{c,\widehat{\pi}^\star_\tau,\tau}(\eta)$, where for each policy π, we introduced its *η-restricted gain estimate* defined by

$$\widehat{\mathbf{g}}_{c,\pi,\tau}(\eta) = \frac{\sum_{c' \in \mathcal{C}^{+}_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c)(c') \widehat{\mathbf{m}}_\tau(c')}{\sum_{c' \in \mathcal{C}^{+}_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c)(c')}.$$

We further introduce for each policy the notation $\mathbf{N}_\pi(\tau) = \mathbf{N}^{\eta}_{c_\tau,\pi}(\tau)$ and the IMED-type index, inspired by (Honda and Takemura, 2015) for MABs,

$$I_\tau(\pi) = \mathbf{N}_\pi(\tau) \mathtt{d}\big(\widehat{\mathbf{g}}_{c_\tau,\pi,\tau}(\eta) \big| \widehat{\mathbf{g}}^\star_{c_\tau,\tau}(\eta)\big) + \log(\mathbf{N}_\pi(\tau)),$$

where $\mathtt{d}(x|y) = \frac{(x-y)^2}{2\sigma^2} = 2(x-y)^2$ denotes the Kullback-Leibler divergence between Gaussian distributions with respective means x and y, and identical standard deviation $\sigma = 1/2$. This is justified since under Assumption 2, all gains fall in $[0,1]$, and hence can be considered $1/2$-sub-Gaussians. Finally, we let $\widehat{\pi}^I_\tau$ (also written $\widetilde{\pi}_{\tau+1}$) be a policy minimizing I_τ over a subset of policies $\Pi_\tau \subset \Pi$ containing $\widehat{\pi}^\star_\tau$. Following Assumption 3, we introduce $\mathcal{V}_{\widehat{\pi}^\star_\tau}(k) = \{\pi : \mathtt{h}(\widehat{\pi}^\star_\tau, \pi) \leqslant k\}$, and define Π_τ such that $\mathcal{V}_{\widehat{\pi}^\star_\tau} \subset \Pi_\tau$. We discuss choices of Π_τ in Section 6.

c. Exploratory policy To compute the fast hitting policy $\widehat{\pi}^H_\tau = \pi^H_c(C)$ that tries to reach $C = \mathcal{C}^+_{c_\tau,\widehat{\pi}^I_\tau}(\eta)$ as fast as possible starting from $c = c_\tau$, we introduce a specific MDP $\mathbf{M}^H_\tau = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r}^H_\tau)$ with modified reward function $\mathbf{r}^H_\tau(c) = \begin{cases} 1 & \text{if } c \in C \\ 0 & \text{else} \end{cases}$. We compute an optimal policy for this MDP, under the average reward criterion, using value iteration. This policy is used to reach the set C and ensures the hitting time is always finite.

Strategy Finally, we define `IMED-KD` to be the rarely-switching algorithm with update rule (line 11 of Algorithm 1) given by choosing at each new episode $\tau \in \mathcal{T}$ the policy

$$\pi_{\tau+1} = \pi^H_{c_\tau}(\mathcal{C}^+_{c_\tau,\widehat{\pi}^I_\tau}(\eta)) \text{ followed by } \widehat{\pi}^I_\tau\,,$$

with stopping event $\texttt{Event}(\pi_{\tau+1}, h_{\tau+1:t}) = \left\{ \min_{c' \in \mathcal{C}^+_{c_\tau,\widehat{\pi}^I_\tau}(\eta)} N_{c'}(h_{\tau+1:t}) > 0 \,\text{and}\, (s_t, a_t) = c_{\widehat{\pi}^I_\tau} \right\}$.

We provide the following control on the length of episodes run with `IMED-KD`, whose proof is given in Appendix C (together with a complementary control for generic learners).

Lemma 12 (Bound on episode lengths) *Assuming* $\mathbf{M}$ *has diameter* $D_\mathbf{M}$ *and no* (B,η)*-lazy chain, the expected length of an episode of* `IMED-KD` *started at* τ *satisfies*

$$\mathbb{E}[\ell_{c,\pi} | h_{1:\tau}, c_\tau = c] \leqslant D_\mathbf{M} + 2B.$$

5. Regret performances

In this section, we provide performance bounds of the `IMED-KD` strategy, starting with a non-asymptotic control on the number of visits of sub-optimal policies. We stress that the existing lower performance bounds from, e.g., (Burnetas and Katehakis, 1997) are explicit only for ergodic MDPs, and presumably NP-hard to compute in general. Hence, we allow for deviating from this and derive an upper-bound involving a different problem-dependent term. Closing the gap is an interesting challenge (both computationally and theoretically).

Theorem 13 (Performance bound of `IMED-KD`) *For an MDP* $\mathbf{M}$ *with diameter* $D_\mathbf{M}$ *and satisfying Assumptions 1–5, the* `IMED-KD` *strategy ensures, provided that* $\eta < \frac{\varepsilon_\mathbf{M}(0)}{2\mathbf{m}_{\max} S}$ *where* $\varepsilon_\mathbf{M}(\eta) = \min_{\substack{c \in \mathcal{C} \\ \pi \notin \Pi^\star}} \left\{ \max_{\pi' \in \mathcal{V}_\pi} \mathbf{g}_{c,\pi'}(\eta) - \mathbf{g}_{c,\pi}(\eta) \right\}$*, the following*

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \mathbf{N}^\eta_{c,\pi}(T) \right] \leqslant \max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \frac{(1+\alpha_\mathbf{M}(\varepsilon))\log(T)}{\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta) \big| \mathbf{g}^\star_c(\eta)\big)} + K_T(\varepsilon,\eta)(D_\mathbf{M} + 2B)\,,$$

for all accuracy $0 < \varepsilon < \dfrac{\varepsilon_\mathbf{M}(\eta)}{2}$*, where* $\lim_{\varepsilon \to 0} \alpha_\mathbf{M}(\varepsilon) = 0$ *and*

$$K_T(\varepsilon,\eta) \leqslant \frac{5\,|\mathcal{C}|\, e^{2\varepsilon^2}}{2\varepsilon^2} + |\mathcal{C}| \left(1 + c^{-1}_{\varepsilon_\mathbf{M}(\eta)} + 2C_{\varepsilon_\mathbf{M}(\eta)} \sqrt{\log(c_{\varepsilon_\mathbf{M}(\eta)} T)}\right)\,.$$

with C_ε *and* c_ε *being constants independent of* $\mathbf{M}$ *and* T*.*

We combine this result together with Equation (3) and the fact that $G_{c,\pi} \leqslant 1$ to obtain

$$\mathcal{R}_\mathbf{M}(\mathbb{A}, T) \leqslant \left[\max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \frac{(1+\alpha_\mathbf{M}(\varepsilon))\log(T)}{\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta) \big| \mathbf{g}^\star_c(\eta)\big)} + K_T(\varepsilon,\eta)(D_\mathbf{M} + 2B) \right] \cdot 2(D_\mathbf{M} + 2B)\,|\mathcal{C}|\,. \quad (4)$$

Remark 14 *A natural question is how to ensure η is small enough since $\varepsilon_{\mathbf{M}}(0)$ is a priori unknown. One possible way to accommodate this is to consider near-optimality instead, with given precision $\tilde{\varepsilon}$, and simply choose $\eta = \tilde{\varepsilon}/2S$. In practice, we may choose η adaptively (i.e., $\eta = \eta_t$); we discuss in Appendix H some simple adaptive choices of η, and demonstrate that they lead to promising empirical performance, though not directly covered by Theorem 13.*

The regret bound in Equation (4) grows logarithmically with T, where the leading constant is determined by a notion of gap with respect to η-restricted gains. In this respect, the bound bears some similarity with logarithmic regret bounds for MABs. This is consistent with the design principle behind the rarely-switching algorithms wherein the MDP was viewed as a multi-policy MAB. In contrast, the bound in Equation (4) is inversely proportional to the square of the gap terms, which stems from the technical difficulties arising in the regret analysis in the average-reward setting. Jaksch et al. (2010) report a logarithmic regret bound for UCRL2 that depends on a similar notion of gap term. However, their bound involves an additive term, which depends on mixing time quantities and has an implicit dependence on $\log(T)$. It thus could grow very large. In empirical evaluation of UCRL2, it is often witnessed that the logarithmic regime in the regret actually kicks in after very long burn-in phase. While Equation (4) offers a bound with an optimal dependence on T, it is not clear whether the gap in terms of policy gains — appearing in both Equation (4) and (Jaksch et al., 2010) — is the best one could get. Indeed, we recall that regret *lower* bounds for (non-ergodic) average-reward MDPs are open and deriving them even for the case of known dynamics is a very interesting, yet challenging, topic of future research.

6. Choice of policies

In this section, we discuss the construction of the set Π_τ. Hereafter, we consider that a set of policies to be small if its size does not exceed 10^6, somewhat arbitrarily.

First, there are cases in which Π is small. This situation may typically happen in real-world applications when a learner must choose between a limited set of policies prescribed by experts. A typical example is that of agriculture in which policies are intervention plans carefully built by agronomists, with a few parameters, despite considering a complex system.

Then, even when Π is large, there are cases when $\Pi^\star$ is known to belong to a small set of policies. For instance in (Puterman, 1994)[Theorem 8.11.3], the author detail the case of an inventory problem when an optimal policy can be searched in a restricted set of $\binom{A+S-1}{S}$ many *non-decreasing* policies instead of all possible A^S ones. For an MDP with $S = 150$ states and $A = 4$ actions, there are over 10^{90} deterministic policies but only $585276 \simeq 10^6$ non-decreasing policies. Likewise, in goal-state MDPs, one can restrict to policies aiming at reaching (and staying) in a single state as fast as possible (they can be computed knowing the transitions of the MDP), yielding only S many policies to consider.

Finally, generic structural properties of the MDP can be used, such as restricting to stationary and unichain policies since an optimal policy satisfies both conditions. Also, when the MDP is known to be unichain, it then satisfies Assumption 3 with $k = 1$, which suggests to simply choose $\Pi_\tau = \mathcal{V}_{\widehat{\pi}^\star_\tau}(1)$. More generally, one can set $\Pi_\tau = \mathcal{V}_{\widehat{\pi}^\star_\tau}(k)$ provided that $|\mathcal{V}_\pi(k)| = \binom{S}{k}A^k$ is small. When k is unknown, one may choose $\Pi_\tau = \mathcal{V}_{\widehat{\pi}^\star_\tau}(\tilde{k}) \cup \Gamma$ where $\tilde{k}$ satisfies $\binom{S}{\tilde{k}}A^{\tilde{k}} \leqslant 10^6$ and Γ is a small set of policies uniformly randomly chosen in $\Pi \setminus \mathcal{V}_{\widehat{\pi}^\star_\tau}(\tilde{k})$. This indeed ensures that Π_τ contains an improving policy over $\widehat{\pi}^\star_\tau$ with positive

probability, which may be interesting for the practitioner. In Appendix G, we detail an alternative way of exploiting policies having more than one recurrent class.

7. Numerical experiments

In this section, we discuss the practical implementation of the presented `IMED-KD` algorithm, and present some numerical experiments[3]. We consider three environments: *RiverSwim* (Fig. 1), which is difficult to navigate; *nasty* (Fig. 3), where two high reward cycles are separated by a bottleneck action; and 4-*rooms* (Fig. 4), which is a sparse reward environment with close-to-deterministic transitions.

Practical comparison In those environments, we illustrate the performance of `IMED-KD` against the strategies `UCRL3` (Bourel et al., 2020), `PSRL` (Osband et al., 2013) and Q-learning (run with discount $\gamma = 0.99$ and optimistic initialization). `PSRL` and `UCRL3` use a confidence parameter to control the quality of the MDP approximation, which is set to 0.05 in the experiments. Further, we adapt both strategies to receive exact knowledge of the transition. The η parameter of `IMED-KD` plays a similar role, and we therefore use $\eta = 0.05/|\mathcal{S}|$ to ensure a fair comparison. `IMED-KD` uses value iteration as a routine, which is faster than the extended value iteration used in `UCRL3`. Q-learning takes an exploration parameter, ε, or exploration scheme when ε is slowly decreased with time. We report regret curves averaged over 2048 independent runs along with quantiles 0.1 and 0.9.

RiverSwim In each of the L states, there are two actions: `RIGHT` and `LEFT`. In Fig. 1, the `LEFT` action is represented with a dashed line and the `RIGHT` with solid line. Rewards are located at the extremities of the MDP, with a small reward in left initial state s_1 and large reward in the rightmost state s_L. Starting from state s_1, this MDP has proven challenging because of the large amount of non-rewarding exploration necessary to find the optimal policy. We consider the 6-state and 25-state instances, which allows us to compare how algorithms behave depending on the amount of necessary exploration; see Fig. 2. Q-learning is struggling despite its optimistic initialization, while `IMED-KD` is on par with `PSRL` on both experiments. The regret of `UCRL3` scales differently with L than the one of `IMED-KD` and `PSRL`, although it remains controlled.

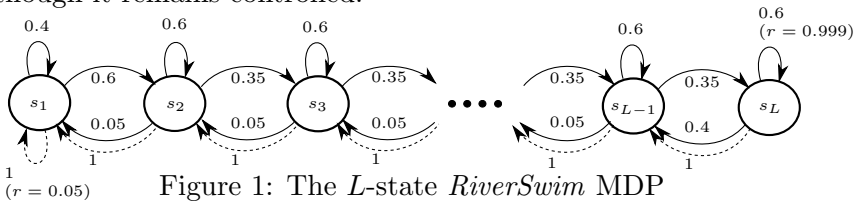

Figure 1: The L-state *RiverSwim* MDP

Nasty In this setting, there are two promising cycles separated by a small chain of one bottleneck state with no associated reward, which may induce an "oscillation" of a learner between the two cycles, paying the cost of the travel along the chain each time it changes cycle (policy). Q-learning exhibits a bad performance, suffering from a large, linear-shaped regret. `UCRL3` attains an even worse regret than Q-learning. In contrast, `IMED-KD` and `PSRL` are highly competitive and perform similarly.

n-rooms 4-rooms is a grid-like environment with 20 states and 4 cardinal actions where transitions are close to deterministic with a 0.8 chance of going in the intended direction. A

3. Source code is available via https://github.com/fabienpesquerel/Logarithmic-regret-in-communicating-MDPs-Leveraging-known-dynamics-with-bandits.git.

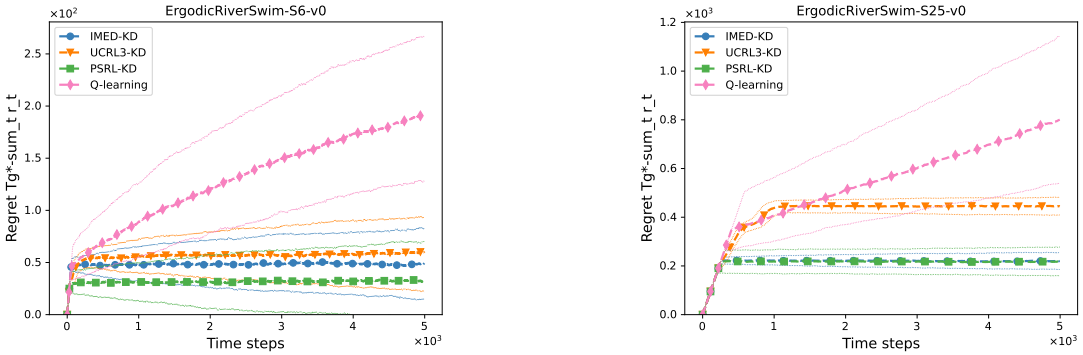

Figure 2: Regret on RiverSwim MDPs: 6-state (left) and 25-state (right)

reward of 0.99 is located in the goal state (highlighted in yellow), while it is zero elsewhere. Upon reaching the goal, the learner is positioned again in the initial red-state. As shown in Fig. 4, `IMED-KD` significantly outperforms the others in this environment —also in 2-rooms as shown in Appendix H. Even for horizons as large as 10^5, we cannot observe a bend in the Q-learning regret curve while it occurs around time step 6×10^4 for `UCRL3` (see Appendix H).

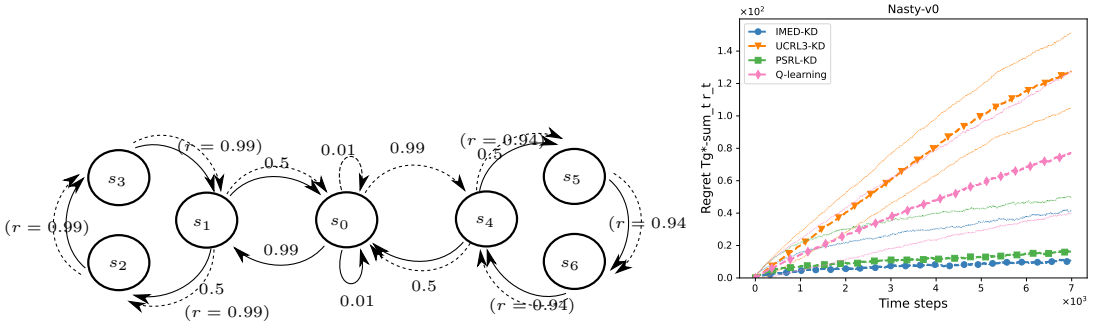

Figure 3: The *Nasty* environment (left) and regret curves (right)

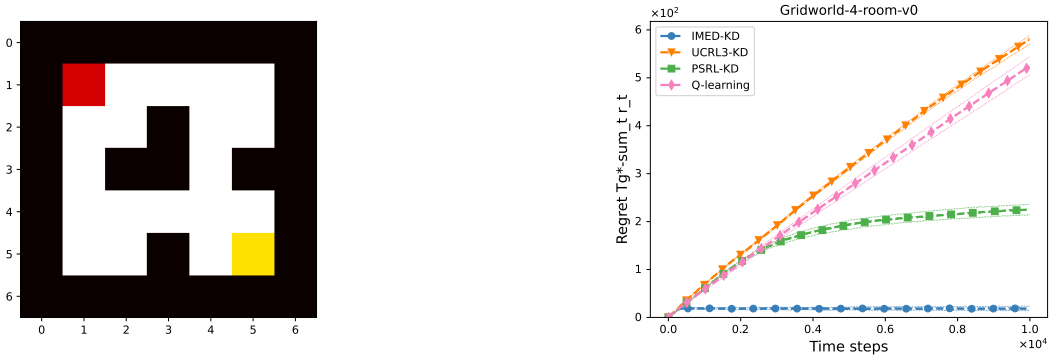

Figure 4: The *4-rooms* environment (left) and regret curves (right)

8. Conclusion

We studied regret minimization in communicating MDPs with known dynamics but unknown reward functions, and introduced a class of rarely-switching algorithms, whose design allows for leveraging the connectivity structure induced by the (known) transition function via considering the recurrent classes of the stationary policies. We presented `IMED-KD`, a rarely-switching algorithm that relies on an IMED-style index function. It admits an efficient implementation and significantly outperforms existing algorithms empirically. Under mild assumptions, we derived a finite-time, problem-dependent, and logarithmic regret bound for `IMED-KD`. Regret lower bounds for this setting (and communicating MDPs in general) are open, to our best knowledge, and deriving them is an interesting, yet challenging, direction for future work. Other interesting future directions include deriving adaptive rules to tune the parameter η (used to control the gains) and to relax the laziness assumption, even though some restrictive assumption seems required to ensure computational efficiency.

Acknowledgements The authors acknowledge the funding of the French National Research Agency, the French Ministry of Higher Education and Research, Inria, the MEL and the I-Site ULNE regarding project R-PILOTE 19-004-APPRENF. MST acknowledges the funding of the Independent Research Fund Denmark (DFF), under grant number 1026-00397B. The authors thank the anonymous reviewers for their careful reading of the paper and their suggestions for improvements.

References

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.

M. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *ICML*, pages 263–272, 2017.

H. Bourel, O.-A. Maillard, and M. S. Talebi. Tightening exploration in upper confidence reinforcement learning. In *ICML*, pages 1056–1066, 2020.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Math. Oper. Res.*, 22(1):222–255, 1997.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann. Stat.*, 41(3):1516–1541, 2013.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *ICML*, pages 151–159, 2013.

R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In *NIPS*, pages 2116–2124, 2015.

R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *NIPS*, pages 1763–1771, 2017.

S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Allerton*, pages 115–122, 2010.

R. Fruit, M. Pirotta, and A. Lazaric. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *NeurIPS*, pages 2998–3008, 2018.

A. Gopalan and S. Mannor. Thompson sampling for learning parameterized Markov decision processes. In *COLT*, pages 861–898, 2015.

W. K. Grassmann, M. I. Taksar, and D. P. Heyman. Regenerative analysis and steady state distributions for Markov chains. *Oper. Res.*, 33:1107–1116, 1985.

T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM J. Control. Optim.*, 35(3):715–743, 1997.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

O.-A. Maillard. Boundary crossing probabilities for general exponential families. *Math. Methods Stat.*, 27(1):1–31, 2018.

R. Ortner. Online regret bounds for Markov decision processes with deterministic transitions. In *ALT*, pages 123–137, 2009.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

F. Pesquerel and O.-A. Maillard. IMED-RL: Regret optimal learning of ergodic Markov decision processes. In *NeurIPS*, pages 26363–26374, 2022.

M. L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

H. Saber, P. Ménard, and O.-A. Maillard. Optimal strategies for graph-structured bandits. *arXiv preprint arXiv:2007.03224*, 2020.

M. Simchowitz and K. G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *NeurIPS*, pages 1153–1162, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

M. S. Talebi and O.-A. Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *ALT*, pages 770–805, 2018.

D. Tranos and A. Proutiere. Regret analysis in deterministic reinforcement learning. In *CDC*, pages 2246–2251, 2021.

C.-Y. Wei, M. Jafarnia Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *ICML*, pages 10170–10180, 2020.

Z. Zhang and X. Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *NeurIPS*, pages 2823–2832, 2019.

Table of Notation

$\mathbf{M}$ average-reward Markov Decision Process

$\mathcal{C}$ set of state-action pairs

$\mathbf{p}$ transition distribution function on $\mathcal{C}$

$\mathbf{r}$ reward distribution function on $\mathcal{C}$

$\mathbf{m}$ mean reward function on $\mathcal{C}$

Π finite set of stationary policies

T horizon time

s_t state visited at time step t

a_t action played at time step t

c_t state-action pair (s_t, a_t)

$\mathbb{A}$ algorithm following a policy sequence $(\pi_t)_{1 \leqslant t \leqslant T} \subset \Pi$

$V_{\mathbf{M}}(\mathbb{A}, T)$ cumulative reward of an algorithm $\mathbb{A}$

$\mathbf{M}_\pi$ Markov chain induced by stationary policy π acting on $\mathbf{M}$

$\mathbf{p}_\pi$ transition probability on $\mathcal{C}^2$ of Markov chain $\mathbf{M}_\pi$

$\overline{\mathbf{p}}_\pi$ Cesaro-average of $\mathbf{p}_\pi$

$\mathbf{g}_{c,\pi}$ gain of stationary policy π starting from state-action pair c

$\mathbf{g}_c^\star$ maximal gain of stationary policies starting from state-action pair c

$\mathbf{g}^\star$ maximal gain of stationary policies

$\Pi_c^\star$ set of stationary policies achieving maximal gain $\mathbf{g}_c^\star$ when starting from state-action pair c

$\Pi^\star$ set of stationary policies achieving maximal gain $\mathbf{g}^\star$

$\pi^\star$ stationary policy achieving maximal gain $\mathbf{g}^\star$

$\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T)$ expected regret with respect to playing a gain-maximal stationary policy (up to time T)

$\mathcal{C}_\pi^+$ set of recurrent state-action pairs (with finite expected return times) when following stationary policy π

c_π reference recurrent state-action pair in $\mathcal{C}_\pi^+$ with minimal expected return time when following stationary policy π

$\mathcal{X}_\pi$ set of disjoint recurrent cycles when following stationary policy π

$\boldsymbol{\tau}_\pi(c, c')$ expected hitting time of state-action pair c' when starting from state-action pair c and following stationary policy π

τ_k (random) last time step of episode k when following a rarely-switching algorithm

$\mathcal{T}$ sequence (τ_k) of last time steps of episodes when following a rarely-switching algorithm

$N^{\mathbf{ini}}_{c,\pi}(0{:}T)$ number of times when an episode starts in state-action pair c and follows stationary policy π

$N_{c,\pi}(T)$ number of visits of state-action pair c when following stationary policy π (possibly much larger than $N^{\mathrm{ini}}_{c,\pi}(0{:}T)$)

h history of all the observations and decisions

$h_{\tau+1:t}$ history of all observations and decisions made from the beginning of the episode starting in time step $\tau + 1$ until time step t

`Event` (random) event following which an episode ends

$\ell_{c,\pi}$ random length of an episode starting in state-action pair c when following stationary policy π

$\boldsymbol{\ell}_{c,\pi}$ expected length of an episode starting in state-action pair c when following stationary policy π

$G_{c,\pi}$ expected average reward of an episode starting in state-action pair c when following statonary policy π

Appendix A. Regret decomposition for rarely-switching learners

A.1. Regret decomposition generic events

Proof of Lemma 4:

Part 1 Let us consider the (random) sequence increasing of episodes $(\tau_i)_{i=0\dots|\mathcal{T}|} \subset \mathcal{T}$. Then the duration of episode $\tau_i \in \mathcal{T}$ is $\tau_{i+1} - \tau_i$. We then define respectively the total duration and the average gain of policy $\pi \in \Pi$ started in state-action pair $c \in \mathcal{C}$ as

$$L_{c,\pi}(T) = \sum_{i=0}^{|\mathcal{T}|} \mathbb{I}\{c_{\tau_i+1} = c, \pi_{\tau_i+1} = \pi\}(\tau_{i+1} - \tau_i)\,, \tag{5}$$

and

$$\widehat{G}_{c,\pi}(T) = \frac{1}{L_{c,\pi}(T)} \sum_{i=0}^{|\mathcal{T}|} \mathbb{I}\{c_{\tau_i+1} = c, \pi_{\tau_i+1} = \pi\} \sum_{t=\tau_i+1}^{\tau_{i+1}} r_t\,.$$

Then, the cumulative reward rewrites,

$$\mathbb{E}_{(\pi_t)}\left[\sum_{t=1}^{T} r_t\right] = \sum_{c\in\mathcal{C}} \sum_{\pi\in\Pi} \mathbb{E}_{(\pi_t)}\left[L_{c,\pi}(T) \cdot \widehat{G}_{c,\pi}(T)\right].$$

The Markov property implies $\mathbb{E}_{(\pi_t)}\left[\widehat{G}_{c,\pi}(T)\,\middle|(c_\tau),(\pi_\tau),(\tau_i)\right] = G_{c,\pi}$, and from previous equality we have

$$\begin{aligned}
\mathbb{E}_{(\pi_t)}\left[\sum_{t=1}^{T} r_t\right] &= \sum_{c\in\mathcal{C}} \sum_{\pi\in\Pi} \mathbb{E}_{(\pi_t)}\left[L_{c,\pi}(T)\ \mathbb{E}_{(\pi_t)}\left[\widehat{G}_{c,\pi}(T)\,\middle|(c_\tau),(\pi_\tau),(\tau_i)\right]\right] \\
&= \sum_{c\in\mathcal{C}} \sum_{\pi\in\Pi} \mathbb{E}_{(\pi_t)}[L_{c,\pi}(T)]\, G_{c,\pi}\,.
\end{aligned} \tag{6}$$

Similarly, the Markov property implies $\mathbb{E}_{(\pi_t)}\left[\tau_{i+1} - \tau_i\,\middle|(c_\tau),(\pi_\tau)\right] = \ell_{c_{\tau_i+1},\pi_{\tau_i+1}}$. This implies for all $c \in \mathcal{C}$ and for all $\pi \in \Pi$,

$$\begin{aligned}
\mathbb{E}_{(\pi_t)}[L_{c,\pi}(T)] &= \mathbb{E}_{(\pi_t)}\left[\mathbb{E}_{(\pi_t)}\left[L_{c,\pi}(T)\,\middle|(c_\tau),(\pi_\tau)\right]\right] \\
&= \mathbb{E}_{(\pi_t)}\left[\sum_{i\geqslant 0} \mathbb{I}\{c_{\tau_i+1} = c, \pi_{\tau_i+1} = \pi\}\ell_{c,\pi}\right] \\
&= \mathbb{E}_{(\pi_t)}\left[\sum_{i\geqslant 0} \mathbb{I}\{c_{\tau_i+1} = c, \pi_{\tau_i+1} = \pi\}\right] \mathbb{E}[\ell_{c,\pi}] \\
&= \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \mathbb{E}[\ell_{c,\pi}]\,,
\end{aligned} \tag{7}$$

where we recall that $N^{\mathrm{ini}}_{c,\pi}(0:T)$ is the (random) number of episodes with starting state-action pair c and followed stationary policy π.

We conclude the proof of the cumulative reward decomposition by combining Equations (6) and (7).

Part 2 We now turn to the decomposition of the regret. To this end, we apply the same decomposition to an optimal policy $\pi^\star$. Then since the event `Event` stops in the same reference state c_π for a policy π by assumption, then except possibly for the first episode, all next episodes under $\pi^\star$ start in the same reference state. This enables us to use the following Markov regeneration property from Lemma 6.

Hence, we obtain that

$$\mathbb{E}_{(\pi^\star)}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1 \neq c^\star\}\ell_{c_1,\pi^\star} \cdot G_{c_1,\pi^\star}] + \mathbb{E}_{(\pi^\star)}\left[N^{\mathrm{ini}}_{c^\star,\pi^\star}(1:T)\right] \cdot \mathbb{E}[\ell_{c^\star,\pi^\star}] \cdot G^\star\,, \quad (8)$$

$$T = \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1 \neq c^\star\}\ell_{c_1,\pi^\star}] + \mathbb{E}_{(\pi^\star)}\left[N^{\mathrm{ini}}_{c^\star,\pi^\star}(1:T)\right] \cdot \mathbb{E}[\ell_{c^\star,\pi^\star}]\,, \quad (9)$$

where $G^\star = G_{c^\star,\pi^\star}$ and c_1 is generated from $\pi^\star$. In particular, both Equations (8) and (9) imply

$$\mathbb{E}_{(\pi^\star)}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1 \neq c^\star\}\ell_{c_1,\pi^\star} \cdot (G_{c_1,\pi^\star} - G^\star)] + T\,G^\star\,. \quad (10)$$

Note that, taking limits as $T \to \infty$, we recover that indeed $G^\star = \mathbf{g}^\star$.

Thus, from the cumulative reward decomposition and previous Equation (10), and using that, on the other hand $T = \sum_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi}]$, we obtain

$$\mathcal{R}_{\mathbf{M}}(\mathbb{A},T) = \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1 \neq c^\star\}\ell_{c_1,\pi^\star} \cdot (G_{c_1,\pi^\star} - G^\star)] + \sum_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi}] \cdot (G^\star - G_{c,\pi})\,. \quad (11)$$

At this point, focusing on the second term, we remark that

$$\begin{aligned}\sum_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi}] \cdot (G^\star - G_{c,\pi}) \;&=\; \sum_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi}] \cdot (G^\star - G_{c,\pi}) \\ &\quad + \sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi^\star}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi^\star}] \cdot (G^\star - G_{c,\pi^\star})\,,\end{aligned} \quad (12)$$

from which we deduce that

$$\begin{aligned}\mathcal{R}_{\mathbf{M}}(\mathbb{A},T) \;&=\; \sum_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi}] \cdot (G^\star - G_{c,\pi}) \\ &\quad + \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1 \neq c^\star\}\ell_{c_1,\pi^\star} \cdot (G_{c_1,\pi^\star} - G^\star)] \\ &\quad + \sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}\left[N^{\mathrm{ini}}_{c,\pi^\star}(0:T)\right] \cdot \mathbb{E}[\ell_{c,\pi^\star}] \cdot (G^\star - G_{c,\pi^\star})\,.\end{aligned}$$

To better control the last term we first isolate the first episode from the next ones. Indeed the first episode may be a bit special, compared to next episodes, as we do not necessary start from a reference state for the current policy. Using the definition of $N^{\mathrm{ini}}_{c,\pi^\star}(0\!:\!T)$, we isolate the first episode and write

$$\begin{aligned} N^{\mathrm{ini}}_{c,\pi^\star}(0\!:\!T) &= \mathbb{I}\{\pi_1=\pi^\star, c_1=c\} + \sum_{\tau\in\mathcal{T}\cap[T],\tau>0} \mathbb{I}\{\pi_{\tau+1}=\pi^\star, c_{\tau+1}=c\} \\ &= \mathbb{I}\{\pi_1=\pi^\star, c_1=c\} + N^{\mathrm{ini}}_{c,\pi^\star}(1\!:\!T)\,. \end{aligned}$$

This leads to a first interesting reduction. Indeed, we then realize that

$$\begin{aligned} \mathcal{R}_{\mathbf{M}}(\mathbb{A},T) &= \sum_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi}(0\!:\!T)\big]\cdot\mathbb{E}[\ell_{c,\pi}]\cdot(G^\star - G_{c,\pi}) \\ &\quad + \mathbb{E}_{(\pi^\star)}[\mathbb{I}\{c_1\neq c^\star\}\ell_{c_1,\pi^\star}\cdot(G_{c_1,\pi^\star}-G^\star)] \\ &\quad + \sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}[\mathbb{I}\{\pi_1=\pi^\star,c_1=c\}]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star-G_{c,\pi^\star}) \\ &\quad + \sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi^\star}(1\!:\!T)\big]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star-G_{c,\pi^\star})\ , \end{aligned}$$

in which the second and third term telescope, owing to the fact that since c_1 is fully determined under $\pi^\star$, then

$$\sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}[\mathbb{I}\{\pi_1=\pi^\star,c_1=c\}]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star-G_{c,\pi^\star}) = \mathbb{E}_{(\pi_t)}[\mathbb{I}\{\pi_1=\pi^\star,c_1\neq c^\star\}\cdot\ell_{c_1,\pi^\star}\cdot(G^\star-G_{c_1,\pi^\star})]\ .$$

Hence, we obtain

$$\begin{aligned} \mathcal{R}_{\mathbf{M}}(\mathbb{A},T) &= \sum_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi}(0\!:\!T)\big]\cdot\mathbb{E}[\ell_{c,\pi}]\cdot(G^\star - G_{c,\pi}) \\ &\quad + \sum_{c\neq c^\star} \mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi^\star}(1\!:\!T)\big]\cdot\mathbb{E}[\ell_{c,\pi^\star}]\cdot(G^\star-G_{c,\pi^\star})\ , \end{aligned} \tag{13}$$

thus completing the proof. □

Proof of Lemma 6:

We denote by (π, c_π) the constant policy (π) started at reference pair c_π. (π, c_π) can be seen as a rarely-switching policy such that $\pi_{\tau+1}=\pi$ at each new episode τ. Since the episode starts and stops in c_π, its length is a multiple of the recurrent time of c_π when playing π. Note that the multiple can be larger than 1 has we may require visiting c_π several times during an episode. Also, by definition of reference pair c_π, for all pair $c\neq c_\pi$, $L_{c,\pi}(T)=0$ and $L_{c_\pi,\pi}(T)=T$ (see Equation (5)).

Hence, from Equation (6) we have

$$\mathbb{E}_{(\pi,c_\pi)}\left[\sum_{t=1}^{T} r_t\right] = T \cdot G_{c_\pi,\pi}\,.$$

This implies in particular

$$\mathbf{g}_{c_\pi,\pi} \coloneqq \lim_{T\to\infty} \mathbb{E}_{(\pi)}\left[\frac{1}{T}\sum_{t=1}^{T} r_t \;\middle|\; c_1 = c_\pi\right] = \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}_{(\pi,c_\pi)}\left[\sum_{t=1}^{T} r_t\right] = G_{c_\pi,\pi}\,.$$

□

A.2. Regret decomposition for specific stopping events

Proof of Proposition 7:

We prove an upper bound on the regret by making appear the gaps $G_{c^\star,\pi^\star} - G_{c,\pi}$, for $c \in \mathcal{C}$, $\pi \in \Pi$, using Lemma 4. We recall that $c^\star$ is the unique state-action pair of reference of the unique optimal stationary strategy $\pi^\star$ (Assumption 1).

The key property we use is that when the episode always stops in the same reference pair for a given policy, since $\pi^\star$ is unichain, then except for the first episode, an episode under $\pi^\star$ that does not start in state-action pair of reference $c^\star$ implies that the stationary policy from the previous episode differs from $\pi^\star$. This shows that

$$\sum_{c\neq c_{\pi^\star}} N^{\mathrm{ini}}_{c,\pi^\star}(1\!:\!T) \leqslant \sum_{c\in\mathcal{C}} \sum_{\pi'\neq\pi^\star} N^{\mathrm{ini}}_{c,\pi'}(0\!:\!T)\,. \tag{14}$$

By combining Equations (2) and (14), we prove the following upper bound on the regret:

$$\begin{aligned}\mathcal{R}_{\mathbf{M}}(\mathbb{A},T) \;\leqslant\; & \mathbb{E}_{(\pi_t)}\left[\sum_{c\in\mathcal{C},\ \pi\neq\pi^\star} N^{\mathrm{ini}}_{c,\pi}(0\!:\!T)\right] \\ & \times\left(\max_{(c,\pi)\neq(c_{\pi^\star},\pi^\star)} \mathbb{E}[\ell_{c,\pi}](G^\star - G_{c,\pi}) + \max_{c\neq c_{\pi^\star}} \mathbb{E}[\ell_{c,\pi^\star}](G^\star - G_{c,\pi^\star})\right).\end{aligned}$$

□

A.3. Regret decomposition for stopping events using recurrent sets

Proof of Theorem 11:

In order to prove the upper bound on the regret we prove an upper on the total number of episodes under a sub-optimal stationary strategy, that is $\sum_{c\in\mathcal{C},\ \pi\neq\pi^\star} \mathbb{E}_{(\pi_t)}\big[N^{\mathrm{ini}}_{c,\pi}(0\!:\!T)\big]$.

Let us consider the number of pulls $\mathbf{N}^{\eta}_{c,\pi}(t) = \min_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} N_{c'}(h_{1:t})$ associated with stationary policy $\pi \in \Pi$ started in state-action pair $c \in \mathcal{C}$. When considering rarely-switching algorithms, due to the definition stopping event, all state-action pairs $c' \in \mathcal{C}^+_{c,\pi}(\eta)$ are visited at each episode started in state-action pair c under policy π. This implies

$$\forall c' \in \mathcal{C}, \quad \sum_{c \in \mathcal{C},\ \pi \neq \pi^\star} \mathbb{I}\{c' \in \mathcal{C}^+_{c,\pi}(\eta)\} N^{\mathrm{ini}}_{c,\pi}(0{:}T) \leqslant N_{c'}(h_{1:T})\,. \tag{15}$$

By considering the definition of the associated numbers of pulls and introducing the argmin set $\underline{\mathcal{C}}^+_{c,\pi}(\eta, T) = \operatorname*{Argmin}_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} N_{c'}(h_{1:T}) \subset \mathcal{C}^+_{c,\pi}(\eta)$ in previous Equation (15), it holds

$$\forall c' \in \mathcal{C}, \sum_{c \in \mathcal{C}, \pi \neq \pi^\star} \mathbb{I}\{c' \in \underline{\mathcal{C}}^+_{c,\pi}(\eta, T)\} N^{\mathrm{ini}}_{c,\pi}(0{:}T) \leqslant \sum_{c \in \mathcal{C}, \pi \neq \pi^\star} \mathbb{I}\{c' \in \mathcal{C}^+_{c,\pi}(\eta)\} N^{\mathrm{ini}}_{c,\pi}(0{:}T) \leqslant N_{c'}(h_{1:T})\,, \tag{16}$$

where $\mathbf{N}^{\eta}_{c,\pi}(T) = N_{c'}(h_{1:T})$ by construction when $c' \in \underline{\mathcal{C}}^+_{c,\pi}(\eta, T)$. This implies

$$\forall c' \in \mathcal{C}, \quad \sum_{c \in \mathcal{C},\ \pi \neq \pi^\star} \mathbb{I}\{c' \in \underline{\mathcal{C}}^+_{c,\pi}(\eta, T)\} N^{\mathrm{ini}}_{c,\pi}(0{:}T) \leqslant \max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \mathbf{N}^{\eta}_{c,\pi}(T)\,. \tag{17}$$

then, summing over $c' \in \mathcal{C}$, and using that $|\underline{\mathcal{C}}^+_{c,\pi}(\eta, T)| \geqslant 1$, it comes

$$\sum_{c \in \mathcal{C},\ \pi \neq \pi^\star} N^{\mathrm{ini}}_{c,\pi}(0{:}T) = \sum_{c' \in \mathcal{C}} \sum_{c \in \mathcal{C},\ \pi \neq \pi^\star} \frac{\mathbb{I}\{c' \in \underline{\mathcal{C}}^+_{c,\pi}(\eta, T)\}}{|\underline{\mathcal{C}}^+_{c,\pi}(\eta, T)|} N^{\mathrm{ini}}_{c,\pi}(0{:}T) \leqslant |\mathcal{C}| \max_{\substack{c \in \mathcal{C} \\ \pi \neq \pi^\star}} \mathbf{N}^{\eta}_{c,\pi}(T)\,. \tag{18}$$

□

Appendix B. Cover times and episode lengths

In this section, we provide a few illustrative examples that highlight the challenges of having a long enough episode on the one hand, while ensuring the cover time of recurrent pairs is controlled.

Long enough episode To give intuition about what it means to have a sufficiently large episode length, consider the example of Fig. 5 depicting an MDP (here deterministic for illustration purpose) with two actions (solid/dashed transitions). In this case, starting from state s_0 following solid actions yields a cycle of length 4 and dashed actions a cycle of length 13. An episode smaller than the length of the cycle will not result in a good approximation of $\mathbf{g}_{c,\pi}$. On the other hand, the average gain is equal to the expected average reward on the cycle starting and ending in s_0 (due to the Markov property hence the regenerative property at state s_0). So, a good estimation of it can be obtained by completing exactly at least one full cycle and using the corresponding average reward to update the estimate.

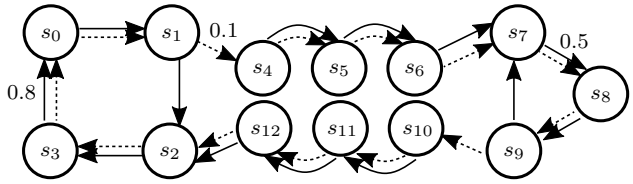

Figure 5: A deterministic MDP with two actions (solid/dashed line) and sparse rewards

Covering time of diffusive policies While in deterministic systems, the covering time of $\mathcal{C}_\pi^+$ by policy π would be of order the cardinal of $\mathcal{C}_\pi^+$, there is some difficulty when considering a stochastic system: Indeed, take the case of a diffusion process as illustrated in Fig. 6. In this case, starting from s_0 with a policy π always playing $\mathtt{up}$, $\mathcal{C}_\pi^+$ contains all pairs $(s, \mathtt{up})$ with $s \in \mathcal{S}$. However, reaching the states s_k or s_{-k} takes time exponential in their distance k to s_0, which is undesirable. At the same time, the contribution of these states to the gain is much smaller than that of s_0 as their return frequency is also much smaller than that of s_0.

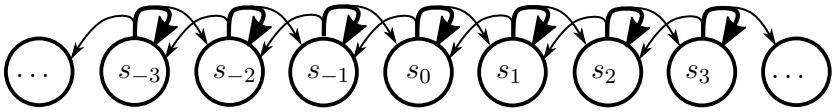

Figure 6: Diffusion process: Action $\mathtt{up}$ (solid line) has high probability $1-2\varepsilon$ to self-loop (thick arrow), and ε probability (thin arrows) to go left or right, causing an arbitrary large covering time

Covering time of lazy chains Now, let us consider the case of an MDP $\mathbf{M}$ and a deterministic policy π that induces the following chain for some small $\varepsilon > 0$:

$$\mathbf{p}_\pi(s'|s) = \begin{cases} 1-\varepsilon & \text{if } s' = s \\ \frac{\varepsilon}{S-1} & \text{else} \end{cases}.$$

Note that such situation can happen for an optimal, or near-optimal policy, that cycles on (a subset of) states in a lazy way. In such a chain, all states s are recurrent and asymptotically visited at same frequency, yielding $\overline{\mathbf{p}}_\pi((s,\pi(s))) = \frac{1}{S}$. Hence, $|\mathcal{C}_{c,\pi}^+(\eta)| = S$ for $\eta \in [0, 1/S)$. On the other hand, we observe that it takes about $1/\varepsilon$ steps to move from one state to another one, due to high probability of self-loop $1-\varepsilon$ that makes the chain lazy. Further, it is easily shown that the expected time to cover the set, starting from any s is $O(S\ln(S)/\varepsilon)$. Hence for each recurrent c', $\overline{\boldsymbol{\tau}}_{c',\pi}(\mathcal{C}_{c,\pi}^+(\eta)) = O(S\ln(S)/\varepsilon)$, which can be made arbitrarily large independently of the values of $\overline{\mathbf{p}}_\pi$. This shows that there is in general no direct control of the covering time of a set as a function of the asymptotic visiting probabilities.

Appendix C. Technical lemmas

Lemma 15 (Restricted-gain approximation)

$$\forall \pi, c, \eta, \quad \mathbf{g}_{c,\pi} - \mathbf{g}_{c,\pi}(\eta) \;\leqslant\; \eta \mathbf{m}_{\max} |\mathcal{C}_{c,\pi}^+ \setminus \mathcal{C}_{c,\pi}^+(\eta)| \,.$$

Proof :

Indeed, it holds that

$$\mathbf{g}_{c,\pi} - \mathbf{g}_{c,\pi}(\eta) = \sum_{c' \notin \mathcal{C}^+_{c,\pi}(\eta)} \underbrace{\overline{\mathbf{p}}_\pi(c,c')}_{\leqslant \eta} \cdot \underbrace{\mathbf{m}(c')}_{\leqslant \mathbf{m}_{\max}} - \underbrace{\sum_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c,c') \cdot \mathbf{m}(c') \frac{\sum_{c' \notin \mathcal{C}^+_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c,c')}{\sum_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \overline{\mathbf{p}}_\pi(c,c')}}_{\geqslant 0}.$$

□

We first provide a useful control of the episode lenghts for any rarely-switching algorithm using a specific stopping event. In general, without further assumption on the structure of the MDP or laziness of its chains, we can prove the following.

Lemma 16 (Episode length) *Assume that for some subset $C \subset \mathcal{C}$ and target $c_0 \in C$, the stopping event is of the form $\texttt{Event}(\pi, h_{\tau+1:t}) = \{\min_{c' \in C} N_{c'}(h_{\tau+1:t}) > 0 \text{ and } (s_t, a_t) = c_0\}$. Then, the following holds*

$$\boldsymbol{\ell}_{c,\pi}(\eta) \leqslant \begin{cases} \boldsymbol{\tau}_\pi(c, c_0) & \text{if } C = \emptyset \\ \sum_{c' \in C} \boldsymbol{\tau}_\pi(c, c') + \mathbb{I}\{c' \neq c_0\} \boldsymbol{\tau}_\pi(c', c_0) & \text{else.} \end{cases} \tag{19}$$

where $\boldsymbol{\tau}_\pi(c,c')$ denotes the expected first (random) passage time from state-action pair $c \in \mathcal{C}$ to state-action pair $c' \in \mathcal{C}$ in the Markov Process $M_\pi = (\mathcal{C}, \mathbf{p}_\pi)$.

Proof of Lemma 16:

Let us consider $\tau_\pi(c,c')$ the first (random) passage time from state-action pair $c \in \mathcal{C}$ to state-action pair $c' \in \mathcal{C}$ in the Markov process $M_\pi = (\mathcal{C}, \mathbf{p}_\pi)$, whose expectation is $\boldsymbol{\tau}_\pi(c,c')$. For a set $C \subset \mathcal{C}$, the first (random) cover time of C, denoted by $\tau_\pi(c, C)$, corresponds to the first time when all elements of C have been visited at least once. Ordering the elements of C from the (random) first element $c_{(1)}$ visited in C to the (random) last one $c_{(|C|)}$, we have $\tau_\pi(c, c_{(1)}) < \cdots < \tau_\pi(c, c_{(|C|)})$, where $\tau_\pi(c, C)$ coincides with $\tau_\pi(c, c_{(|C|)})$, that is we have $\tau_\pi(c, C) = \max_{c' \in C} \tau_\pi(c, c')$. Note that this quantity is very different from and should not be confused with $\max_{c' \in \mathcal{C}} \boldsymbol{\tau}_\pi(c, c')$, which can be much smaller than the expected covering time $\mathbb{E}[\tau_\pi(c, C)]$. Now, considering the set C to be non-empty, we thus introduce the (random) state-action pair

$$c_{\eta,\pi} = \operatorname*{argmax}_{c' \in C} \tau_\pi(c, c')$$

such that for all state-action pair $c' \in C$, $\tau_\pi(c,c') \leqslant \tau_\pi(c, c_{\eta,\pi})$. By construction of the stopping event `Event`, if $c_{\eta,\pi} = c_0$ then the episode stops immediately, otherwise one has to wait to reach c_0 from $c_{\eta,\pi}$, that is $\tau_\pi(c_{\eta,\pi}, c_0)$ many steps. Hence, under any rarely-switching

algorithm using such event, the expected duration of an episode started in state-action pair c under the policy π is given by the following expression

$$\begin{aligned}\boldsymbol{\ell}_{c,\pi}(\eta) &= \mathbb{E}_{(\pi)}[\tau_\pi(c, c_{\eta,\pi}) + \mathbb{I}\{c_{\eta,\pi} \neq c_0\}\tau_\pi(c_{\eta,\pi}, c_0)] \\ &\leqslant \mathbb{E}_{(\pi)}\left[\max_{c' \in C} \tau_\pi\big(c, c'\big) + \mathbb{I}\{c' \neq c_0\}\tau_\pi\big(c', c_0\big)\right]\end{aligned}$$

Upper-bounding the maximum over c' by a sum over the possible c', and using that $\mathbb{E}_{(\pi)}[\tau_\pi(c, c')] = \boldsymbol{\tau}_\pi(c, c')$ this implies

$$\boldsymbol{\ell}_{c,\pi}(\eta) \leqslant \sum_{c' \in C} \Big(\boldsymbol{\tau}_\pi(c, c') + \mathbb{I}\{c' \neq c_0\}\boldsymbol{\tau}_\pi(c', c_0)\Big)\,. \tag{20}$$

Now, if the set C is empty then the episode stops when c_0 is reached, that is $\boldsymbol{\ell}_{c,\pi}(\eta) \leqslant \mathbb{E}_{(\pi)}[\tau_\pi(c, c_0)] = \boldsymbol{\tau}_\pi(c, c_0)$. □

Now, we provide the control of the length of episodes for `IMED-KD` below, under the assumption that there are no (B, η)-lazy chains.

Proof of Lemma 12:

`IMED-KD` runs a policy that first reaches $\mathcal{C}^+_{c,\widehat{\pi}^I_\tau}$ as fast as possible, but then simply run the policy $\widehat{\pi}^I_\tau$. Hence, it takes at most $D_{\mathbf{M}}$ expected steps to reach $\mathcal{C}^+_{c,\widehat{\pi}^I_\tau}$ but then B many steps to cover the set $\mathcal{C}^+_{c,\widehat{\pi}^I_\tau}$ under Assumption 5, and at most B more steps to reach the reference pair $c_{\widehat{\pi}^I_\tau}$. This proves that

$$\mathbb{E}_{(\pi_t)}[\ell_{c,\pi} | h_{1:\tau}, c_\tau = c] \leqslant D_{\mathbf{M}} + 2B\,.$$

□

Appendix D. Finite time analysis of `IMED-KD`

At a high level, the key interesting step of the proof is to realize that the considered algorithm implies empirical lower and empirical upper bounds on the numbers of pulls (see Lemmas 18 and 19). Then, based on concentration lemmas (see Section F as well as the discussion below Theorem 24), the algorithm-based empirical lower bounds ensure the reliability of the estimators of interest (Lemma 21). Interestingly, this makes use of arguments based on recent concentration of measure that enable to control the concentration without adding some log log bonus —such a bonus was required for example in the initial analysis of the KL-UCB strategy from (Cappé et al., 2013). Then, combining the reliability of these estimators with the obtained algorithm-based empirical upper bounds, we obtain upper bounds on the average numbers of pulls (Theorem 13). Interestingly, most of the proof is agnostic to the length of an episode (that is handled separately). We only use the property that the algorithm guarantees in each episode an increase by at least one of the number of pulls of each η-recurrent pair. The proof is concise to fit mostly in the next few pages.

D.1. Notations

For an MDP $\mathbf{M}$, we denote by $\mathcal{V}_\pi = \{\pi' \in \Pi : \mathtt{h}(\pi, \pi') \leqslant k\}$ the neighbourhood of policy $\pi \in \Pi$ at radius k in Hamming distance $\mathtt{h}$. For constant $\eta \geqslant 0$, we let

$$\varepsilon_{\mathbf{M}}(\eta) = \min_{\substack{c \in \mathcal{C} \\ \pi \notin \Pi^\star}} \left\{ \max_{\pi' \in \mathcal{V}_\pi} \mathbf{g}_{c,\pi'}(\eta) - \mathbf{g}_{c,\pi}(\eta) \right\} . \tag{21}$$

According to policy-improvement Assumption 3, $\varepsilon_{\mathbf{M}}(0) > 0$. Then, from Lemma 9, provided that $\eta < \varepsilon_{\mathbf{M}}(0)/(2\mathbf{m}_{\max} S)$, we also have $\varepsilon_{\mathbf{M}}(\eta) > 0$. Furthermore, it holds that

$$\varepsilon_{\mathbf{M}}(\eta) \leqslant \min_{\substack{c \in \mathcal{C} \\ \pi \notin \Pi^\star}} \{\mathbf{g}_{c,\pi^\star}(\eta) - \mathbf{g}_{c,\pi}(\eta)\} .$$

Note that this value of η also ensures that $\Pi^\star := \operatorname*{argmax}_{\pi \in \Pi} \max_{c \in \mathcal{C}} \mathbf{g}_{c,\pi} = \operatorname*{argmax}_{\pi \in \Pi} \max_{c \in \mathcal{C}} \mathbf{g}_{c,\pi}(\eta)$. Indeed, $\varepsilon_{\mathbf{M}}(0) \leqslant \min_{\substack{c \in \mathcal{C} \\ \pi \notin \Pi^\star}} \left\{ \max_{\pi' \in \Pi} \mathbf{g}_{c,\pi'} - \mathbf{g}_{c,\pi} \right\}$, which ensures that η-restriction does not modify the best policy. Then, there exists a function $\alpha_{\mathbf{M}}(\cdot)$ with $\lim_{\varepsilon \to 0} \alpha_{\mathbf{M}}(\varepsilon) = 0$ such that for all $0 \leqslant \varepsilon < \varepsilon_{\mathbf{M}}(\eta)/2$, for all state-action pair $c \in \mathcal{C}$, for all sub-optimal stationary policy $\pi \notin \Pi^\star$,

$$\mathtt{d}(\mathbf{g}_{c,\pi}(\eta) + \varepsilon | \mathbf{g}^\star_c(\eta) - \varepsilon) \leqslant (1 + \alpha_{\mathbf{M}}(\varepsilon))^{-1} \, \mathtt{d}(\mathbf{g}_{c,\pi}(\eta) | \mathbf{g}^\star_c(\eta)) . \tag{22}$$

In the sequel, for notational convenience and avoid cluttering the notations. We denote

$$\begin{aligned} N_\pi(\tau) &:= \mathbf{N}^\eta_{c_{\tau+1},\pi}(\tau) \\ \widehat{g}_\pi(\tau) &:= \widehat{\mathbf{g}}_{c_{\tau+1},\pi,\tau}(\eta) \\ \widehat{g}^\star(\tau) &:= \widehat{\mathbf{g}}^\star_{c_{\tau+1},\tau}(\eta) \end{aligned}$$

Last, we recall that $\mathcal{T}$ denotes the set of starting times, namely the set of time steps that start a new episode.

D.2. Algorithm-based empirical bounds

The `IMED-KD` algorithm implies inequalities between the indexes that can be rewritten as inequalities on the numbers of pulls. While lower bounds involving $\log(t)$ (or $\log(\tau)$) may be expected in view of the asymptotic regret bounds, we show lower bounds on the numbers of pulls involving instead $\log\big(N_{\tilde{\pi}_{\tau+1}}(\tau)\big) = \log\big(\mathbf{N}^\eta_{c_{\tau+1},\tilde{\pi}_{\tau+1}}(\tau)\big)$, the logarithm of the number of pulls of the current index policy. We also provide upper bounds on $N_{\tilde{\pi}_{\tau+1}}(\tau)$ involving $\log(\tau)$. We believe that establishing these empirical lower and upper bounds is a key element of our proof technique, which is of independent interest.

Remark 17 *According to the `IMED-KD` algorithm, $\widehat{\pi}^\star_\tau \in \mathcal{V}_{\widehat{\pi}^\star_\tau} \subset \Pi_\tau$.*

Lemma 18 (Empirical lower bounds) *Under `IMED-KD`, for all starting time $\tau \in \mathcal{T}$, for all stationary policy $\pi \in \Pi_\tau$,*

$$\log\big(N_{\tilde{\pi}_{\tau+1}}(\tau)\big) \leqslant N_\pi(\tau) \, \mathtt{d}(\widehat{g}_\pi(\tau) | \widehat{g}^\star(\tau)) + \log(N_\pi(\tau)) , \tag{23}$$

and for the empirical best policy $\widehat{\pi}^\star_\tau$,

$$N_{\tilde{\pi}_{\tau+1}}(\tau) \leqslant N_{\widehat{\pi}^\star_\tau}(\tau) . \tag{24}$$

Proof :

For all stationary policy $\pi \in \Pi$, we have $I_\pi(\tau) = N_\pi(\tau)\mathtt{d}(\widehat{g}_\pi(\tau)|\widehat{g}^\star(\tau)) + \log(N_\pi(\tau))$ by definition. Hence, by non-negativity of the first term, it comes

$$\log(N_\pi(\tau)) \leqslant I_\pi(\tau)\,.$$

This implies, since the policy $\tilde{\pi}_{\tau+1}$ with minimum index is chosen,

$$\log\big(N_{\tilde{\pi}_{\tau+1}}(\tau)\big) \leqslant I_{\tilde{\pi}_{\tau+1}}(\tau) = \min_{\pi\in\Pi} I_\pi(\tau) \leqslant I_{\widehat{\pi}^\star_\tau}(\tau) = \log\big(N_{\widehat{\pi}^\star_\tau}(\tau)\big)\,.$$

Taking $\exp(\cdot) = \log^{-1}(\cdot)$ on both side concludes the proof. □

Lemma 19 (Empirical upper bounds) *Under* `IMED-KD`*, for all starting time $\tau \in \mathcal{T}$, for all stationary policy $\pi \in \Pi_\tau$,*

$$N_{\tilde{\pi}_{\tau+1}}(\tau)\; \mathtt{d}\big(\widehat{g}_{\tilde{\pi}_{\tau+1}}(\tau)|\widehat{g}^\star(\tau)\big) \leqslant \log(\tau)\,. \tag{25}$$

Proof :

By construction, since policy $\tilde{\pi}_{\tau+1}$ has minimum index, we have

$$I_{\tilde{\pi}_{\tau+1}}(\tau) \leqslant I_{\widehat{\pi}^\star_t}(\tau)\,.$$

To conclude, it remains to note that on one hand,

$$N_{\tilde{\pi}_{\tau+1}}(\tau)\mathtt{d}\big(\widehat{g}_{\tilde{\pi}_{\tau+1}}(\tau)|\widehat{g}^\star(\tau)\big) \leqslant I_{\tilde{\pi}_{\tau+1}}(\tau)\,,$$

and on the other hand,

$$I_{\widehat{\pi}^\star_\tau}(\tau) = \log(N_{\widehat{\pi}^\star_\tau}(\tau)) \leqslant \log(\tau)\,.$$

□

D.3. Non-reliable current best stationary policy

For accuracy $\varepsilon > 0$, stationary policy $\pi \in \Pi$, and state-action pair $c \in \mathcal{C}$, let $\mathcal{M}^\star_{c,\pi}(\varepsilon)$ be the set of starting times $\tau \in \mathcal{T}$ such that $c_{\tau+1} = c$ and $\tilde{\pi}_{\tau+1} = \pi$ and where some of the current best stationary policy $\widehat{\pi}^\star_\tau$ has not too optimistic gain and does not belong to $\Pi^\star$:

$$\mathcal{M}^\star_{c,\pi}(\varepsilon) \stackrel{\text{def}}{=} \left\{ \tau \in \mathcal{T} : \begin{array}{ll} (1) & c_{\tau+1} = c \\ (2) & \tilde{\pi}_{\tau+1} = \pi \\ (3) & \widehat{g}_{\widehat{\pi}^\star_\tau}(\tau) < \mathbf{g}_{c,\widehat{\pi}^\star_\tau}(\eta) + \varepsilon \\ (4) & \widehat{\pi}^\star_\tau \notin \Pi^\star \end{array} \right\}. \tag{26}$$

For all couple of stationary policies $(\pi,\pi')\in\Pi^2$, initial state-action pair $c\in\mathcal{C}$, and for all accuracy $\varepsilon>0$, let us further introduce $\mathcal{K}^-_{c,\pi,\pi'}(\varepsilon)$ as the set of starting times where couple of stationary policy (π,π') shows ε-$\mathtt{d}$-log deviation, that is

$$\mathcal{K}^-_{c,\pi,\pi'}(\varepsilon)\overset{\text{def}}{=}\left\{t\in\mathcal{T}:\begin{array}{ll}(1) & c_{\tau+1}=c\\ (2) & \tilde{\pi}_{\tau+1}=\pi'\\ (3) & \widehat{g}_\pi(\tau)\leqslant \mathbf{g}_{c,\pi}(\eta)-\varepsilon\\ (4) & \log(N_{\pi'}(\tau))\leqslant N_\pi(\tau)\,\mathtt{d}(\widehat{g}_\pi(\tau)|\mathbf{g}_{c,\pi}(\eta)-\varepsilon)+\log(N_\pi(\tau))\end{array}\right\}. \tag{27}$$

The two sets are related thanks to the following result.

Lemma 20 (Relation between subsets of times) *Under* `IMED-KD`*, for all accuracy* $0<\varepsilon<\varepsilon_{\mathbf{M}}(\eta)/2$*, for all stationary policy* $\pi\in\Pi$*, and all starting state-action pair* $c\in\mathcal{C}$*,*

$$\mathcal{M}^\star_{c,\pi}(\varepsilon)\subset\bigcap_{\pi^+\in\mathcal{V}^\star}\mathcal{K}^-_{c,\pi^+,\pi}(\varepsilon_{\mathbf{M}}(\eta)/2)\,, \tag{28}$$

where we introduced the set $\mathcal{V}^\star=\bigcup_{\widehat{\pi}^\star\notin\Pi^\star}\operatorname*{Argmax}_{\pi'\in\mathcal{V}_{\widehat{\pi}^\star}}\mathbf{g}_{c,\pi'}(\eta)$.

Proof :

Let us consider $\tau\in\mathcal{M}^\star_{c,\pi}(\varepsilon)$. Since $\widehat{\pi}^\star_\tau\notin\Pi^\star$ is not a best stationary policy, then according to policy-improvement Assumption 3 and for a value of η ensuring that $\varepsilon_{\mathbf{M}}(\eta)>0$, and for any $\pi^+\in\operatorname*{argmax}_{\pi'\in\mathcal{V}_{\widehat{\pi}^\star_\tau}}\mathbf{g}_{c_{\tau+1},\pi'}(\eta)$, we have

$$\mathbf{g}_{c_{\tau+1},\pi^+}(\eta)>\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)\,. \tag{29}$$

Then, since $\widehat{\pi}^\star_\tau\in\operatorname*{argmax}_{\pi\in\Pi}\widehat{g}(\tau)$ and $\mathcal{V}_{\widehat{\pi}^\star_\tau}\subset\Pi_\tau\subset\Pi$, we have on the other hand

$$\widehat{g}_{\widehat{\pi}^\star_\tau}(\tau)=\widehat{g}^\star(\tau)\geqslant\widehat{g}_{\pi^+}(\tau)\,, \tag{30}$$

where $\pi^+\in\operatorname*{argmax}_{\pi'\in\mathcal{V}_{\widehat{\pi}^\star_\tau}}\mathbf{g}_{c_{\tau+1},\pi'}(\eta)\subset\mathcal{V}_{\widehat{\pi}^\star_\tau}\subset\Pi$ by the design of `IMED-KD`. Since $\tau\in\mathcal{M}^\star_{c,\pi}(\varepsilon)$, we have by construction

$$\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)+\varepsilon\geqslant\widehat{g}_{\widehat{\pi}^\star_\tau}(\tau)\,. \tag{31}$$

By combining Equations (30) and (31), it comes

$$\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)+\varepsilon\geqslant\widehat{g}^\star(\tau)\geqslant\widehat{g}_{\pi^+}(\tau)\,. \tag{32}$$

Since $\varepsilon_{\mathbf{M}}(\eta)\leqslant\mathbf{g}_{c_{\tau+1},\pi^+}(\eta)-\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)$, Equation (29) implies $\mathbf{g}_{c_{\tau+1},\pi^+}(\eta)>\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)+\varepsilon_{\mathbf{M}}(\eta)$. Then, since $\varepsilon\leqslant\varepsilon_{\mathbf{M}}(\eta)/2$, Equation (32) implies

$$\mathbf{g}_{c_{\tau+1},\pi^+}(\eta)-\varepsilon_{\mathbf{M}}(\eta)/2>\mathbf{g}_{c_{\tau+1},\pi^+}(\eta)+\varepsilon\geqslant\widehat{g}^\star(\tau)\geqslant\widehat{g}_{\pi^+}(\tau)\,. \tag{33}$$

At this point, since $\pi^+\in\mathcal{V}_{\widehat{\pi}^\star_\tau}\subset\Pi_\tau$, empirical lower bounds in Equation (23) imply

$$\log\big(N_{\tilde{\pi}_{\tau+1}}(\tau)\big)\leqslant N_{\pi^+}(\tau)\,\mathtt{d}(\widehat{g}_{\pi^+}(\tau)|\widehat{g}^\star(\tau))+\log(N_{\pi^+}(\tau))\,. \tag{34}$$

The classical monotonic properties of $\mathtt{d}(\cdot|\cdot)$ and Equation (33) imply

$$\begin{cases} \widehat{g}_{\pi^+}(\tau) \leqslant \widehat{g}^\star(\tau) < \mathbf{g}_{c_{\tau+1},\pi^+}(\eta) - \varepsilon_{\mathbf{M}}(\eta)/2 \\ \mathtt{d}(\widehat{g}_{\pi^+}(\tau)|\widehat{g}^\star(\tau)) \leqslant \mathtt{d}\big(\widehat{g}_{\pi^+}(\tau)\big|\mathbf{g}_{c_{\tau+1},\pi^+}(\eta) - \varepsilon_{\mathbf{M}}(\eta)/2\big)\,. \end{cases} \tag{35}$$

Combining Equations (33) and (35), we finally get

$$\begin{cases} \widehat{g}_{\pi^+}(\tau) < \mathbf{g}_{c_{\tau+1},\pi^+}(\eta) - \varepsilon_{\mathbf{M}}(\eta)/2 \\ \log\big(N_{\tilde{\pi}_{\tau+1}}(t)\big) \leqslant N_{\pi^+}(\tau)\, \mathtt{d}\big(\widehat{g}_{\pi^+}(\tau)\big|\mathbf{g}_{c_{\tau+1},\pi^+}(\eta) - \varepsilon_{\mathbf{M}}(\eta)/2\big) + \log(N_{\pi^+}(\tau))\,, \end{cases} \tag{36}$$

which means $\tau \in \mathcal{K}^-_{c,\pi^+,\tilde{\pi}_{\tau+1}}(\varepsilon_{\mathbf{M}}(\eta)/2)$, hence concluding the proof. □

D.4. Reliable current gains and current best stationary policy

In this subsection, we characterize subsets of starting times where both the gain of current played stationary policy and the optimal gain are well-estimated.

Let us consider for an accuracy $0 < \varepsilon < \varepsilon_{\mathbf{M}}$, a sub-optimal stationary policy $\pi \in \Pi$ and a starting state-action pair $c \in \mathcal{C}$, the following set of starting times

$$\mathcal{U}_{c,\pi}(\varepsilon) \;=\; \left\{ \tau \in \mathcal{T} : \begin{array}{ll} (1) & c_{\tau+1} = c \\ (2) & \tilde{\pi}_{\tau+1} = \pi \notin \Pi^\star \\ (3) & (3a) \text{ or } (3b) \text{ or } (3c) \text{ or } (3d) \text{ where} \\ (3a) & \widehat{g}_{\pi}(\tau) \geqslant \mathbf{g}_{c,\pi}(\eta) + \varepsilon \\ (3b) & \widehat{g}_{\widehat{\pi}^\star_\tau}(\tau) \geqslant \mathbf{g}_{c,\widehat{\pi}^\star_\tau}(\eta) + \varepsilon \text{ and } N_{\pi}(\tau) \leqslant N_{\widehat{\pi}^\star_\tau}(\tau) \\ (3c) & \widehat{g}_{\widehat{\pi}^\star_\tau}(\tau) \leqslant \mathbf{g}_{c,\widehat{\pi}^\star_\tau}(\eta) - \varepsilon \text{ and } N_{\pi}(\tau) \leqslant N_{\widehat{\pi}^\star_\tau}(\tau) \\ (3d) & \widehat{g}_{\widehat{\pi}^\star_\tau}(\tau) < \mathbf{g}_{c,\widehat{\pi}^\star_\tau}(\eta) + \varepsilon \text{ and } \widehat{\pi}^\star_\tau \notin \Pi^\star \end{array} \right\},$$

where we recall that whenever $\tilde{\pi}_{\tau+1} = \pi$, then $N_{\tilde{\pi}_{\tau+1}}(\tau) \leqslant N_{\widehat{\pi}^\star_\tau}(\tau)$, by Equation (24). By construction of this set we have the following result.

Lemma 21 (Reliable current means) *Under* `IMED-KD`*, for all accuracy* $0 < \varepsilon < \varepsilon_{\mathbf{M}}(\eta)/2$*, for all stationary policy* $\pi \in \Pi$ *and starting state-action pair* $c \in \mathcal{C}$*, for all starting time* $\tau \notin \mathcal{U}_{c,\pi}(\varepsilon)$ *such that* $c_{\tau+1} = c$ *and* $\tilde{\pi}_{\tau+1} = \pi \notin \Pi^\star$*,*

$$\begin{cases} \widehat{\pi}^\star_\tau \in \Pi^\star_{c_{\tau+1}} \\ \widehat{g}^\star(\tau) \geqslant \mathbf{g}^\star_{c_{\tau+1}}(\eta) - \varepsilon \\ \widehat{g}_{\pi}(\tau) \leqslant \mathbf{g}_{c_{\tau+1},\pi}(\eta) + \varepsilon\,. \end{cases}$$

Proof :

For $0 < \varepsilon < \varepsilon_{\mathbf{M}}(\eta)/2$, for stationary policy $\pi \in \Pi$, let us consider a starting time $\tau \notin \mathcal{U}_{c,\pi}(\varepsilon)$, such that $\tilde{\pi}_{\tau+1} = \pi \notin \Pi^\star$.

Since $\tilde{\pi}_{\tau+1}=\pi\notin\Pi^\star$ and $\tau\notin\mathcal{U}_{c_{\tau+1},\tilde{\pi}_{\tau+1}}(\varepsilon)$, then $\widehat{g}_{\tilde{\pi}_{\tau+1}}(\tau)<\mathbf{g}_{c_{\tau+1},\tilde{\pi}_{\tau+1}}(\eta)+\varepsilon$, which rewrites $\widehat{g}_{\pi}(\tau)<\mathbf{g}_{c_{\tau+1},\pi}(\eta)+\varepsilon$ (since $\tilde{\pi}_{\tau+1}=\pi$).

Likewise, since $\tilde{\pi}_{\tau+1}=\pi\notin\Pi^\star_{c_{\tau+1}}$ and $\tau\notin\mathcal{U}_{c_{\tau+1},\tilde{\pi}_{\tau+1}}(\varepsilon)$, and $N_\pi(\tau)\leqslant N_{\widehat{\pi}^\star_\tau}(\tau)$, then

$$\widehat{g}^\star(\tau)=\widehat{g}_{\widehat{\pi}^\star_\tau}(\tau)>\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)-\varepsilon\,. \tag{37}$$

Likewise, we must have

$$\widehat{g}^\star(\tau)=\widehat{g}_{\widehat{\pi}^\star_\tau}(\tau)<\mathbf{g}_{c_{\tau+1},\widehat{\pi}^\star_\tau}(\eta)+\varepsilon\,. \tag{38}$$

Since $\tilde{\pi}_{\tau+1}=\pi\notin\Pi^\star$ and $\tau\notin\mathcal{U}_{c_{\tau+1},\tilde{\pi}_{\tau+1}}(\varepsilon)$, then this must in turn imply

$$\widehat{\pi}^\star_\tau\in\Pi^\star\,. \tag{39}$$

By combining this with Equations (37) and (38), we get

$$\widehat{g}^\star(\tau)>\mathbf{g}^\star_{c_{\tau+1}}(\eta)-\varepsilon\,. \tag{40}$$

□

Size of the set $\mathcal{U}_{c,\pi}(\varepsilon)$ We now want to control the expected size of the set $\mathcal{U}_{c,\pi}(\varepsilon)$. To this end, we first note that Lemma 20 enables to replace the equality in the definition of $\mathcal{U}_{c,\pi}(\varepsilon)$ with an inclusion, and (3d) with

$$\begin{aligned}\forall\pi^+\in\mathcal{V}^\star,\qquad &\widehat{g}_{\pi^+}(\tau)\leqslant\mathbf{g}_{c,\pi^+}(\eta)-\tilde{\varepsilon}\\ &\log(N_\pi(\tau))\leqslant N_{\pi^+}(\tau)\,\mathtt{d}\big(\widehat{g}_{\pi^+}(\tau)\big|\mathbf{g}_{c,\pi^+}(\eta)-\tilde{\varepsilon}\big)+\log(N_{\pi^+}(\tau))\end{aligned} \tag{41}$$

where $\tilde{\varepsilon}=\varepsilon_{\mathbf{M}}(\eta)/2\geqslant\varepsilon$. Further, we realize that we can decompose the set as follows

$$\mathcal{U}_{c,\pi}(\varepsilon)\subset\mathcal{E}_{c,\pi}(\varepsilon)\cup\overline{\mathcal{E}}_{c,\pi}(\varepsilon)\cup\mathcal{K}_{c,\pi}(\tilde{\varepsilon})\,,$$

where we introduced the following convenient events

$$\begin{aligned}\mathcal{E}_{c,\pi}(\varepsilon)&=\{\tau:(1),(2),\widehat{g}_\pi(\tau)\geqslant\mathbf{g}_{c,\pi}(\eta)+\varepsilon\}\\ \overline{\mathcal{E}}_{c,\pi}(\varepsilon)&=\{\tau:(1),(2),\exists\pi':|\widehat{g}_{\pi'}(\tau)-\mathbf{g}_{c,\pi'}(\eta)|\geqslant\varepsilon\text{ and }N_\pi(\tau)\leqslant N_{\pi'}(\tau)\}\\ \mathcal{K}_{c,\pi}(\tilde{\varepsilon})&=\{\tau:(1),(2),(41)\}\end{aligned}$$

Interestingly, we note that if $\tau\in\mathcal{K}_{c,\pi}(\tilde{\varepsilon})\setminus\overline{\mathcal{E}}_{c,\pi}(\varepsilon)$ and since $\tilde{\varepsilon}\geqslant\varepsilon$, then for each $\pi^+\in\mathcal{V}^\star$, on top of Equation (41) we must also have $N_\pi(\tau)>N_{\pi'}(\tau)$, which motivates to introduce

$$\overline{\mathcal{K}}_{c,\pi}(\tilde{\varepsilon})=\{\tau:(1),(2),(41)\text{ and }\forall\pi^+\in\mathcal{V}^\star,N_\pi(\tau)>N_{\pi^+}(\tau)\}\,.$$

Using this decomposition, we get the following control

$$\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}|\mathcal{U}_{c,\pi}(\varepsilon)|\leqslant\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}|\mathcal{E}_{c,\pi}(\varepsilon)|+\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\big|\overline{\mathcal{E}}_{c,\pi}(\varepsilon)\big|+\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\big|\overline{\mathcal{K}}_{c,\pi}(\tilde{\varepsilon})\big|\,. \tag{42}$$

We can now resort to concentration arguments in order to control the size of these sets under rarely-switching algorithms, which yields the following upper bounds. We defer the proof to Appendix E.

Lemma 22 (Bounded subsets of times) *Under any rarely-switching algorithm, for all accuracy $\varepsilon>0$, for all stationary policy $\pi\in\Pi$ and starting state-action pair $c\in\mathcal{C}$,*

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\left|\mathcal{E}_{c,\pi}(\varepsilon)\right|\right],\ \mathbb{E}_{(\pi_t)}\left[\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\left|\overline{\mathcal{E}}_{c,\pi}(\varepsilon)\right|\right]\leqslant 2\left|\mathcal{C}\right|b_\varepsilon\,,$$

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\left|\overline{\mathcal{K}}_{c,\pi}(\varepsilon)\right|\right]\leqslant\left|\mathcal{C}\right|\left(b_\varepsilon+1+c_\varepsilon^{-1}+2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\right),$$

where $b_\varepsilon = 2\sigma^2 e^{\varepsilon^2/2\sigma^2}/\varepsilon^2$ with $\sigma^2=1/4$, considering concentration for σ-sub-Gaussian distributions, and $c_\varepsilon, C_\varepsilon>0$ are the constants involved in the concentration Theorem 24.

In particular, using this lemma, it holds

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\left|\mathcal{U}_{c,\pi}(\varepsilon)\right|\right]\leqslant 5\left|\mathcal{C}\right|b_\varepsilon+\left|\mathcal{C}\right|\left(c_\varepsilon^{-1}+2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\right). \tag{43}$$

D.5. Upper bounds on the numbers of pulls of sub-optimal policies

In this subsection, we now combine the different results of the previous subsections to prove Theorem 13.

Proof of Theorem 13:

For all accuracy $0<\varepsilon<\varepsilon_{\mathbf{M}}(\eta)/2$, for all stationary policy $\pi\in\Pi$, for all starting time $\tau\notin\mathcal{U}_{c,\pi}(\varepsilon)$ such that $\tilde{\pi}_{\tau+1}=\pi\notin\Pi^\star$, we derive the following steps. From empirical upper bounds (25), we have

$$N_\pi(\tau)\,\mathtt{d}(\widehat{g}_\pi(\tau)|\widehat{g}^\star(\tau))\leqslant\log(\tau)\,. \tag{44}$$

From Lemma 21, we have $\widehat{g}_\pi(\tau)\leqslant\mathbf{g}_{c_{\tau+1},\pi}(\eta)+\varepsilon<\mathbf{g}^\star_{c_{\tau+1}}(\eta)-\varepsilon\leqslant\widehat{g}^\star(\tau)$. From classical monotonic properties of $\mathtt{d}(\cdot|\cdot)$ and Equation (22), we have $\mathtt{d}(\widehat{g}_\pi(\tau)|\widehat{g}^\star(\tau))\geqslant\mathtt{d}\Big(\mathbf{g}_{c_{\tau+1},\pi}(\eta)+\varepsilon\Big|\mathbf{g}^\star_{c_{\tau+1}}(\eta)-\varepsilon\Big)\geqslant(1+\alpha_{\mathbf{M}}(\varepsilon))^{-1}\,\mathtt{d}\Big(\mathbf{g}_{c_{\tau+1}}(\eta)\Big|\mathbf{g}^\star_{c_{\tau+1}}(\eta)\Big)$. In view of Equation (44), and recalling that $N_\pi(\tau)=\mathbf{N}^\eta_{c_{\tau+1},\pi}(\tau)$, this implies

$$\forall\tau\notin\mathcal{U}_{c,\pi}(\varepsilon)\text{ such that }\tilde{\pi}_{\tau+1}=\pi\notin\Pi^\star,\quad\mathbf{N}^\eta_{c_{\tau+1},\pi}(\tau)\leqslant\frac{(1+\alpha_{\mathbf{M}}(\varepsilon))\log(\tau)}{\mathtt{d}\big(\mathbf{g}_{c_{\tau+1},\pi}(\eta)\big|\mathbf{g}^\star_{c_{\tau+1}}(\eta)\big)}\,. \tag{45}$$

For state-action pair $c\in\mathcal{C}$ and for stationary policy $\pi\notin\Pi^\star$, we denote by

$$\tau_{c,\pi}=\max\left\{\tau\in\mathcal{T}:\ c_{\tau+1}=c,\quad\tilde{\pi}_{\tau+1}=\pi\quad\text{and}\quad\tau\notin\mathcal{U}_{c,\pi}(\varepsilon)\right\} \tag{46}$$

the last starting time that does not belong to $\mathcal{U}_{c,\pi}(\varepsilon)$ such that we play stationary policy π.

Now, using Equation (46) and that by definition, when $\tau\in\mathcal{U}_{c,\pi}(\varepsilon)$, it must be that $c_{\tau+1}=c$, $\tilde{\pi}_{\tau+1}=\pi$, we obtain

$$\begin{aligned}
\mathbf{N}^{\eta}_{c,\pi}(T) &= \min_{c'} N_{c'}(h_{1:T}) \\
&= \min_{c'} \sum_{k\in\mathbb{N}} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}\{c_t = c'\} \\
&= \min_{c'} \sum_{k\in\mathbb{N}} \mathbb{I}\{c_{\tau+1} = c,\ \tilde{\pi}_{\tau+1} = \pi, \tau_k \notin \mathcal{U}_{c,\pi}(\varepsilon)\} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}\{c_t = c'\} \\
&\quad + \sum_{k\in\mathbb{N}} \mathbb{I}\{c_{\tau+1} \neq c \text{ or } \tilde{\pi}_{\tau+1} \neq \pi \text{ or } \tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}\{c_t = c'\} \\
&\leqslant \min_{c'} N_{c'}(h_{1:\tau_{c,\pi}}) + \sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}\{c_t = c'\} \\
&\leqslant \min_{c'} N_{c'}(h_{1:\tau_{c,\pi}}) + \sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\}(\tau_{k+1} - \tau_k) \\
&\overset{\mathcal{L}}{=} \mathbf{N}^{\eta}_{c,\pi}(\tau_{c,\pi}) + \sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\}\ell_{c,\pi}
\end{aligned}$$

where the last equality holds in law, using that $\tau_{k+1} - \tau_k$ is equal in law to $\ell_{c,\pi}$.

Now, from Lemma 12, we can control the expected value of $\ell_{c,\pi}$ conditionally on the past history before each episode, by the deterministic quantity

$$\mathbb{E}_{(\pi_t)}[\ell_{c,\pi}|h_{1:\tau}, c_\tau] \leqslant \mathbf{L} \overset{\text{def}}{=} \max\{(|\mathcal{C}|+2)D_{\mathbf{M}}, D_{\mathbf{M}} + 2B\}\,.$$

Since the law of the stopping time $\ell_{c,\pi}$ is independent of other variables before the start of an episode, we deduce that

$$\begin{aligned}
\mathbb{E}_{(\pi_t)}\left[\sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\}\ell_{c,\pi}\right] &= \mathbb{E}_{(\pi_t)}\left[\sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\}\mathbb{E}_{(\pi_t)}[\ell_{c,\pi}|h_{1:\tau_k} c_{\tau_k} = c]\right] \\
&\leqslant \mathbb{E}_{(\pi_t)}\left[\sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\}\right]\mathbf{L}\,.
\end{aligned}$$

Further remarking that $\sum_{k\in\mathbb{N}} \mathbb{I}\{\tau_k \in \mathcal{U}_{c,\pi}(\varepsilon)\} = |\mathcal{U}_{c,\pi}(\varepsilon)|$, we deduce that

$$\mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbf{N}^{\eta}_{c,\pi}(T)] \leqslant \mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbf{N}^{\eta}_{c,\pi}(\tau_{c,\pi})] + \mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} |\mathcal{U}_{c,\pi}(\varepsilon)|]\mathbf{L}\,.$$

Combined with the inequality in Equation (45), and using that $\tau_{c,\pi} \leqslant T$, we obtain

$$\begin{aligned}
\mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \mathbf{N}^{\eta}_{c,\pi}(T)] &\leqslant \mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \frac{(1+\alpha_{\mathbf{M}}(\varepsilon))\log(\tau_{c,\pi})}{\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta)\big|\mathbf{g}^\star_c(\eta)\big)}] + \mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} |\mathcal{U}_{c,\pi}(\varepsilon)|]\mathbf{L} \\
&\leqslant \max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \frac{(1+\alpha_{\mathbf{M}}(\varepsilon))\log(T)}{\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta)\big|\mathbf{g}^\star_c(\eta)\big)} + \mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} |\mathcal{U}_{c,\pi}(\varepsilon)|]\mathbf{L}\,.
\end{aligned}$$

We conclude the proof using Equation (43) to control $\mathbb{E}_{(\pi_t)}[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} |\mathcal{U}_{c,\pi}(\varepsilon)|]$. □

Appendix E. Bounded subsets of time (Proof of Lemma 22)

We regroup in this section, for completeness, the proofs of the remaining lemmas used in the analysis of `IMED-KD` in Section D.

Proof of Lemma 22, part 1:

We detail the proof to bound $\mathbb{E}_{(\pi_t)}\big[\big|\overline{\mathcal{E}}_{c,\pi}(\varepsilon)\big|\big]$. The control of $\mathbb{E}_{(\pi_t)}[|\mathcal{E}_{c,\pi}(\varepsilon)|]$ is similar.

We first write

$$\big|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\big| = \sum_{\tau\in\mathcal{T}} \mathbb{I}\{c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi', \exists\pi:\ N_{\pi'}(\tau)\leqslant N_\pi(\tau),\ |\mathbf{g}_{c,\pi}(\eta)-\widehat{g}_\pi(\tau)|\geqslant\varepsilon\}\,. \tag{47}$$

Considering the stopped stopping times $\tau_n = \inf\{\tau\in\mathcal{T}: c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi' \text{ and } N_{\pi'}(\tau)\geqslant n\}$ for $n\geqslant 0$, we will rewrite the sum of indicators and use Lemma 23. We note that the set

$$\big\{\tau\in\mathcal{T}: c_{\tau+1}=c, \tilde{\pi}_{\tau+1}=\pi' \text{ and } n\leqslant N_{\pi'}(\tau)<n+1\big\}$$

is either empty or equal to $\{\tau_n\}$. This is true for all rarely-switching algorithm (Algorithm 1), by construction of the stopping event that ensures $N_{\pi'}(\tau)$ increases by one in the corresponding episode.

$$\begin{aligned}
\big|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\big| &\leqslant \sum_{n=0}^{T-1}\sum_{\tau\in\mathcal{T}} \mathbb{I}\{c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi',\ n\leqslant N_{\pi'}(\tau)<n+1\} \qquad (48)\\
&\qquad \times\mathbb{I}\{\exists\pi: n\leqslant N_\pi(\tau),\ |\mathbf{g}_{c,\pi}(\eta)-\widehat{g}_\pi(\tau)|\geqslant\varepsilon\}\\
&\leqslant \sum_{n=0}^{T-1}\sum_{\tau\in\mathcal{T}} \mathbb{I}\{\tau=\tau_n,\ c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi',\ n\leqslant N_{\pi'}(\tau)<n+1\} \qquad (49)\\
&\qquad \times\mathbb{I}\{\exists\pi: n\leqslant N_\pi(\tau_n),\ |\mathbf{g}_{c,\pi}(\eta)-\widehat{g}_\pi(\tau_n)|\geqslant\varepsilon\}\\
&\leqslant \sum_{n=0}^{T-1} \mathbb{I}\{\exists\pi: n\leqslant N_\pi(\tau_n),\ |\mathbf{g}_{c,\pi}(\eta)-\widehat{g}_\pi(\tau_n)|\geqslant\varepsilon\}\,,
\end{aligned}$$

where in the last line we use that $N_{\pi'}(\tau)$ does increase by one in episode τ. At this point, we make use of the fact that $\mathbf{g}_{c,\pi}(\eta) = \sum_{c'\in\mathcal{C}^+_{c,\pi}(\eta)} \tilde{\mathbf{p}}_\pi(c)(c')\mathbf{m}(c')$ and $\widehat{g}_\pi(\tau) = \sum_{c'\in\mathcal{C}^+_{c,\pi}(\eta)} \tilde{\mathbf{p}}_\pi(c)(c')\widehat{\mathbf{m}}_\tau(c')$, with $\tilde{\mathbf{p}}_\pi(c)(c') = \frac{\overline{\mathbf{p}}_\pi(c)(c')}{\sum_{c'\in\mathcal{C}^+_{c,\pi}(\eta)}\overline{\mathbf{p}}_\pi(c)(c')}$ so that

$$\mathbf{g}_{c,\pi}(\eta)-\widehat{g}_\pi(\tau) = \sum_{c'\in\mathcal{C}^+_{c,\pi}(\eta)} \tilde{\mathbf{p}}_\pi(c)(c')\big(\mathbf{m}(c')-\widehat{\mathbf{m}}_\tau(c')\big)\,.$$

In particular, $|\mathbf{g}_{c,\pi}(\eta) - \widehat{g}_\pi(\tau)| \geqslant \varepsilon$ implies that $\exists c' \in \mathcal{C}^+_{c,\pi}(\eta), |\mathbf{m}(c') - \widehat{\mathbf{m}}_\tau(c')| \geqslant \varepsilon$, otherwise one would have $|\mathbf{g}_{c,\pi}(\eta) - \widehat{g}_\pi(\tau)| < \Big(\sum_{c' \in \mathcal{C}^+_{c,\pi}(\eta)} \tilde{\mathbf{p}}_\pi(c)(c') \Big) \varepsilon = \varepsilon$. Hence, this shows that

$$\begin{aligned}
\left|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\right| &\leqslant \sum_{n=0}^{T-1} \mathbb{I}\{\exists \pi, c' \in \mathcal{C}^+_{c,\pi}(\eta),\ n \leqslant N_{c'}(\tau_n) :\ |\mathbf{m}(c') - \widehat{\mathbf{m}}_{\tau_n}(c')| \geqslant \varepsilon\} \\
&\leqslant \sum_{n=0}^{T-1} \mathbb{I}\{\exists c' \in \bigcup_\pi \mathcal{C}^+_{c,\pi}(\eta),\ n \leqslant N_{c'}(\tau_n) :\ |\mathbf{m}(c') - \widehat{\mathbf{m}}_{\tau_n}(c')| \geqslant \varepsilon\}\,. \\
&\leqslant \sum_{n=0}^{T-1} \mathbb{I}\{\exists c' \in \mathcal{C},\ n \leqslant N_{c'}(\tau_n) :\ |\mathbf{m}(c') - \widehat{\mathbf{m}}_{\tau_n}(c')| \geqslant \varepsilon\}\,.
\end{aligned}$$

The last inequality implies

$$\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\right| \leqslant \sum_{c \in \mathcal{C}} \sum_{n=0}^{T-1} \mathbb{I}\{n \leqslant N_c(\tau_n),\ |\mathbf{m}(c) - \widehat{\mathbf{m}}_{\tau_n}(c)| \geqslant \varepsilon\}\,. \tag{50}$$

Taking the expectation of Equation (50), it comes

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\right|\right] \leqslant \sum_{c \in \mathcal{C}} \sum_{n \geqslant 0} \mathbb{P}\left(\bigcup_{\substack{t \geqslant 1 \\ N_c(t) \geqslant n}} |\widehat{\mathbf{m}}_t(c) - \mathbf{m}(c)| \geqslant \varepsilon \right). \tag{51}$$

From Lemma 23, previous Equation (51) implies

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\right|\right] \leqslant \sum_{c \in \mathcal{C}} \sum_{n \geqslant 0} 2 \exp\big(- n\, \mathtt{d}(\mathbf{m}(c) - \varepsilon | \mathbf{m}(c)) \big). \tag{52}$$

From Pinsker's inequality, previous Equation (52) implies

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{E}}_{c,\pi'}(\varepsilon)\right|\right] \leqslant \sum_{c \in \mathcal{C}} \sum_{n \geqslant 0} 2 \exp\big(-n\varepsilon^2/2\sigma^2\big) = \frac{2\,|\mathcal{C}|}{1 - e^{-\varepsilon^2/2\sigma^2}}\,, \tag{53}$$

where $\sigma^2 = 1/4$, assuming 1/2-sub-Gaussian reward distributions. Finally we note that

$$\frac{1}{1 - e^{-\varepsilon^2/2\sigma^2}} = \frac{e^{\varepsilon^2/2\sigma^2}}{e^{\varepsilon^2/2\sigma^2} - 1} \leqslant \frac{2\sigma^2 e^{\varepsilon^2/2\sigma^2}}{\varepsilon^2} = b_\varepsilon\,.$$

□

Proof of Lemma 22, part 2:

We now prove the upper bound on $\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c\in\mathcal{C}\\ \pi\in\Pi}}\left|\overline{\mathcal{K}}_{c,\pi}(\varepsilon)\right|\right]$.

By definition of the set, we have

$$\begin{aligned}\left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right| \;=\; & \sum_{\tau\in\mathcal{T}} \mathbb{I}\{c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi',\ \forall \pi^+\in\mathcal{V}^\star, 0<N_{\pi^+}(\tau)<N_{\pi'}(\tau)\}\\ & \times\mathbb{I}\{\ \widehat{g}_{\pi^+}(\tau)\leqslant \mathbf{g}_{c,\pi^+}(\eta)-\varepsilon\}\\ & \times\mathbb{I}\{\log\left(N_{\pi'}(\tau)\right)\leqslant N_{\pi^+}(\tau)\,\mathtt{d}(\widehat{g}_{\pi^+}(\tau)|\mathbf{g}_{c,\pi^+}(\eta)-\varepsilon)+\log\left(N_{\pi^+}(\tau)\right)\}.\end{aligned} \tag{54}$$

Considering again the stopped stopping times

$$\tau_n=\inf\left\{\tau\in\mathcal{T}\colon c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi' \text{ and } N_{\pi'}(\tau)\geqslant n\right\}$$

for $n\geqslant 0$, we will rewrite the previous sum and use boundary crossing probabilities for one-dimensional exponential family distributions. We recall that the set

$$\left\{\tau\in\mathcal{T}\colon c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi' \text{ and } n\leqslant N_{\pi'}(\tau)<n+1\right\}$$

is either empty or equal to $\{\tau_n\}$.

$$\begin{aligned}&\left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right|\\ \leqslant\;& \sum_{\tau\in\mathcal{T}} \mathbb{I}\{c_{\tau+1}=c,\ \tilde{\pi}_{\tau+1}=\pi',\ \forall\pi\in\mathcal{V}^\star, 0<N_{\pi}(\tau)<N_{\pi'}(\tau),\ \widehat{g}_{\pi}(\tau)\leqslant \mathbf{g}_{c,\pi}(\eta)-\varepsilon\}\\ &\quad\times\mathbb{I}\{\log\left(N_{\pi'}(\tau)\right)\leqslant N_{\pi}(\tau)\,\mathtt{d}(\widehat{g}_{\pi}(\tau)|\mathbf{g}_{c,\pi}(\eta)-\varepsilon)+\log\left(N_{\pi}(\tau)\right)\}\\ \leqslant\;& \sum_{n=0}^{T-1}\sum_{\tau\in\mathcal{T}} \mathbb{I}\{c_{\tau+1}=c,\ \tau=\tau_n,\ \tilde{\pi}_{\tau+1}=\pi',\ n\leqslant N_{\pi'}(\tau)<n+1\} \qquad (55)\\ &\quad\times\mathbb{I}\{\forall\pi\in\mathcal{V}^\star, 0<N_{\pi}(\tau),\ \widehat{g}_{\pi}(\tau)\leqslant \mathbf{g}_{c,\pi}(\eta)-\varepsilon\}\\ &\quad\times\mathbb{I}\{\forall\pi\in\mathcal{V}^\star, \log(n)\leqslant N_{\pi}(\tau)\,\mathtt{d}(\widehat{g}_{\pi}(\tau)|\mathbf{g}_{c,\pi}(\eta)-\varepsilon)+\log\left(N_{\pi}(\tau)\right)\}\\ \leqslant\;& \sum_{n=0}^{T-1} \mathbb{I}\{\forall\pi\in\mathcal{V}^\star, 0<N_{\pi}(\tau_n),\ \widehat{g}_{\pi}(\tau_n)\leqslant \mathbf{g}_{c,\pi}(\eta)-\varepsilon\}\\ &\quad\times\mathbb{I}\{\forall\pi\in\mathcal{V}^\star, \log(n)\leqslant N_{\pi}(\tau_n)\,\mathtt{d}(\widehat{g}_{\pi}(\tau_n)|\mathbf{g}_{c,\pi}(\eta)-\varepsilon)+\log\left(N_{\pi}(\tau_n)\right)\} \qquad (56)\end{aligned}$$

Let us consider for stationary policy $\pi\in\Pi$ and starting time $\tau\in\mathcal{T}$,

$$c_\tau^\pi\in\operatorname*{argmin}_{c'\in\mathcal{C}^+_{c,\pi}(\eta)}\widehat{\mathbf{m}}_\tau(c')-\mathbf{m}(c')\,.$$

Then, the following inequality and implication holds (see Lemma 25):

$$\widehat{\mathbf{m}}_\tau(c_\tau^\pi)-\mathbf{m}(c_\tau^\pi)\leqslant\widehat{g}_\tau(\tau)-\mathbf{g}_{c,\pi}(\eta)=\sum_{c'\in\mathcal{C}^+_{c,\pi}(\eta)}\tilde{\mathbf{p}}_\pi(c)(c')\big(\widehat{\mathbf{m}}_\tau(c')-\mathbf{m}(c')\big)\,.$$

$$\left(\widehat{g}_\pi(\tau) \leqslant \mathbf{g}_{c,\pi}(\eta) - \varepsilon \right)$$
$$\implies \quad \left(\widehat{\mathbf{m}}_\tau(c_\tau^\pi) \leqslant \mathbf{m}(c_\tau^\pi) - \varepsilon \text{ and } \mathtt{d}(\widehat{g}_\pi(\tau) | \mathbf{g}_{c,\pi}(\eta) - \varepsilon) \leqslant \mathtt{d}(\widehat{\mathbf{m}}_\tau(c_\tau^\pi) | \mathbf{m}(c_\tau^\pi) - \varepsilon) \right) \quad (57)$$

Note also that since $c_\tau^\pi \in \mathcal{C}_{c,\pi}^+(\eta)$ and by construction $N_\pi(\tau) = \min_{c' \in \mathcal{C}_{c,\pi}^+(\eta)} N_{c'}(\tau)$, then $N_\pi(\tau_n) > 0$ implies $N_{c_{\tau_n}^\pi}(\tau_n) > 0$. In particular, Equations (55) and (57) imply

$$\begin{aligned}
\left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right| &\leqslant \min_{\pi \in \mathcal{V}^\star} \sum_{n=0}^{T-1} \mathbb{I}\{0 < N_{c_{\tau_n}^\pi}(\tau_n),\ \widehat{\mathbf{m}}_{\tau_n}(c_{\tau_n}^\pi) \leqslant \mathbf{m}(c_{\tau_n}^\pi) - \varepsilon\} \\
&\quad \times \mathbb{I}\{\log(n) \leqslant N_{c_{\tau_n}^\pi}(\tau_n)\, \mathtt{d}(\widehat{\mathbf{m}}_{\tau_n}(c_{\tau_n}^\pi) | \mathbf{m}(c_{\tau_n}^\pi) - \varepsilon) + \log\left(N_{c_{\tau_n}^\pi}(\tau_n)\right)\} \\
&\leqslant \min_{\pi \in \mathcal{V}^\star} \sum_{c' \in \mathcal{C}_{c_{\tau+1},\pi}^+(\eta)} \sum_{n=0}^{T-1} \mathbb{I}\{0 < N_{c'}(\tau_n),\ \widehat{\mathbf{m}}_{\tau_n}(c') \leqslant \mathbf{m}(c') - \varepsilon\} \\
&\quad \times \mathbb{I}\{\log(n) \leqslant N_{c'}(\tau_n)\, \mathtt{d}(\widehat{\mathbf{m}}_{\tau_n}(c') | \mathbf{m}(c') - \varepsilon) + \log\left(N_{c'}(\tau_n)\right)\}.
\end{aligned}$$

This last inequality implies

$$\begin{aligned}
\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right| &\leqslant \sum_{c \in \mathcal{C}} \sum_{n=0}^{T-1} \mathbb{I}\{1 \leqslant N_c(\tau_n),\ \widehat{\mathbf{m}}_{\tau_n}(c) \leqslant \mathbf{m}(c) - \varepsilon\} \\
&\quad \times \mathbb{I}\{\log(n) \leqslant N_c(\tau_n)\, \mathtt{d}(\widehat{\mathbf{m}}_{\tau_n}(c) | \mathbf{m}(c) - \varepsilon) + \log\left(N_c(\tau_n)\right)\} \quad (58)
\end{aligned}$$

Taking the expectation of Equation (58), it comes

$$\begin{aligned}
\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right|\right] &\leqslant \sum_{c \in \mathcal{C}} \sum_{n=0}^{T-1} \mathbb{P}\left(\bigcup_{\substack{t \geqslant 1 \\ N_c(t) \geqslant n}} \widehat{\mathbf{m}}_t(c) < \mathbf{m}(c) - \varepsilon \right) \\
&\quad + \sum_{c \in \mathcal{C}} \sum_{n=2}^{T-1} \mathbb{P}\left(\bigcup_{\substack{t \geqslant 1 \\ \widehat{\mathbf{m}}_t(c) < \mathbf{m}(c) - \varepsilon \\ 1 \leqslant N_c(t) \leqslant n}} N_c(t) \mathtt{d}(\widehat{\mathbf{m}}_t(c) | \mathbf{m}(c) - \varepsilon) \geqslant \log(n / N_c(t)) \right).
\end{aligned}$$

Invoking Lemma 23 and Theorem 24, Equation (59) implies

$$\mathbb{E}_{(\pi_t)}\left[\max_{\substack{c \in \mathcal{C} \\ \pi' \in \Pi}} \left|\overline{\mathcal{K}}_{c,\pi'}(\varepsilon)\right|\right] \leqslant |\mathcal{C}| \left(\frac{2\sigma^2 e^{\varepsilon^2/2\sigma^2}}{\varepsilon^2} + 1 + c_\varepsilon^{-1} + 2C_\varepsilon \sqrt{\log(c_\varepsilon T)} \right). \quad (59)$$

Indeed, we have

$$\begin{aligned}
\sum_{n=2}^{T-1} \frac{C_\varepsilon}{n\sqrt{\log(c_\varepsilon n)}} &\leqslant 1 + c_\varepsilon^{-1} + C_\varepsilon \sum_{n \geqslant 1 + c_\varepsilon^{-1}}^{T-1} \frac{c_\varepsilon}{c_\varepsilon n \sqrt{\log(c_\varepsilon n)}} \\
&\leqslant 1 + c_\varepsilon^{-1} + C_\varepsilon \int_{c_\varepsilon^{-1}}^{T} \frac{c_\varepsilon\, dx}{c_\varepsilon x \sqrt{\log(c_\varepsilon x)}} \\
&= 1 + c_\varepsilon^{-1} + 2C_\varepsilon \sqrt{\log(c_\varepsilon T)}.
\end{aligned}$$

□

Appendix F. Concentration inequalities

In this section, we state two concentration results used in the proof. First, a classical maximal inequality for sub-Gaussian distributions. Then, a boundary crossing probability result suitable for the analysis of an IMED strategy, adapted here to sub-Gaussian distributions.

Lemma 23 (Maximal concentration inequality) *Under Assumption 2, for any $c \in \mathcal{C}$, for $x < \mathbf{m}(c)$, and integer $n \geqslant 0$, we have*

$$\mathbb{P}\left(\bigcup_{\substack{t \geqslant 1 \\ N_c(t) \geqslant n}} \widehat{\mathbf{m}}_t(c) < x \right) \leqslant \exp\big(- n \, \mathtt{d}(x|\mathbf{m}(c)) \big).$$

Proof :

Indeed, by a Chernoff method, introducing some $\lambda > 0$, and $\varphi(\lambda) = \sigma^2\lambda^2/2$ where $\sigma = 1/2$, we get, provided that $\lambda\varepsilon \geqslant \varphi(\lambda)$,

$$\begin{aligned}
&\mathbb{P}(\exists t, N_c(t) \geqslant n, \widehat{\mathbf{m}}_t(c) - \mathbf{m}(c) > \varepsilon) \\
&= \mathbb{P}\left(\exists t, N_c(t) \geqslant n, \exp\left(\lambda \sum_{j=1}^{N_c(t)} Z_j \right) > \exp(\lambda N_c(t)\varepsilon) \right) \\
&\leqslant \mathbb{P}\left(\exists N \geqslant n, \exp\left(\lambda \sum_{j=1}^{N} Z_j - N\varphi(\lambda) \right) > \exp(N(\lambda\varepsilon - \varphi(\lambda)) \right) \\
&\leqslant \mathbb{P}\left(\exists N \geqslant n, \exp\left(\lambda \sum_{j=1}^{N} Z_j - N\varphi(\lambda) \right) > \exp(n(\lambda\varepsilon - \varphi(\lambda)) \right) \\
&= \mathbb{E}_{(\pi_t)}\left[\max_{N \geqslant n} \exp\left(\lambda \sum_{j=1}^{N} Z_j - N\varphi(\lambda) \right) \right] \exp\left(-n(\lambda\varepsilon - \varphi(\lambda)) \right),
\end{aligned}$$

where $Z_j = r_{(j)} - \mathbf{m}(c)$ with $r_{(j)} \sim \mathbf{r}(c)$ being the j-th sample collected from $\mathbf{r}(c)$, and where N denotes a random stopping time and $W_n = \exp(\lambda \sum_{j=1}^{n} Z_j - n\varphi(\lambda))$ is a non-negative supermartingale bounded by 1. By Doob's maximal inequality the expectation term is thus upper bounded by 1. Optimizing over λ, we get

$$\begin{aligned}
\mathbb{P}(\exists t, N_c(t) \geqslant n, \widehat{\mathbf{m}}_t(c) > \mathbf{m}(c) + \varepsilon) &\leqslant \exp(-n\varepsilon^2/(2\sigma^2)) \\
&= \exp\left(-n\mathtt{d}(\mathbf{m}(c) + \varepsilon | \mathbf{m}(c)) \right).
\end{aligned}$$

□

For reward distributions belonging to generic exponential families, concentration result is obtained in (Maillard, 2018, Corollary 2). Specializing this to Gaussian distributions with variance of $1/2$, the Kullback-Leibler divergence between two distributions reduces to $\mathtt{d}$. When assuming regular canonical one-dimensional exponential reward distributions $\mathbf{r}(\cdot)$, we can define the reward distributions as a function their means and abusively write $\mathbf{r}(\cdot) = (\mathbf{r}_c) = \big(\mathbf{r}(\mathbf{m}(c))\big)$. Using this, we obtain more precisely

Theorem 24 (Boundary crossing probabilities) *For all pair $c \in \mathcal{C}$, for all $\varepsilon > 0$, for all $n \geqslant 1$, we have for one-dimensional exponential distributions*

$$\mathbb{P}\left(\bigcup_{\substack{t \geqslant 1 \\ \widehat{\mathbf{m}}_t(c) < \mathbf{m}(c) - \varepsilon \\ 1 \leqslant N_c(t) \leqslant n}} N_c(t) \mathit{KL}\Big(\mathbf{r}(\widehat{\mathbf{m}}_t(c)), \mathbf{r}(\mathbf{m}(c))\Big) \geqslant \log\left(n/N_c(t)\right) \right) \leqslant \frac{C_\varepsilon}{n\sqrt{\log(c_\varepsilon n)}}\,,$$

where $\mathit{KL}\big(\mathbf{r}(\widehat{\mathbf{m}}_t(c)), \mathbf{r}(\mathbf{m}(c))\big) = \mathtt{d}(\widehat{\mathbf{m}}_t(c) | \mathbf{m}(c) - \varepsilon)$ when assuming Gaussian distributions with standard deviation $\sigma = 1/2$, and where $c_\varepsilon, C_\varepsilon > 0$ are explained in (Maillard, 2018, Corollary 2).

It is then not difficult, scrutinizing the proof, to show that the same bound still holds now for sub-Gaussian distributions. Importantly, in this case $\mathtt{d}$ is no longer the natural metric, but by Pinsker's inequality, $\mathtt{d}$ always controls it, although now in a possibly loose way. More precisely, in the case of Gaussian reward distributions, we have $\mathtt{KL}(\mathbf{r}_c, \mathbf{r}'_c) = (\mathbf{m}'(c) - \mathbf{m}(c))^2 / 2\sigma^2 = \mathtt{d}(\mathbf{m}(c) | \mathbf{m}'(c))$ while for the case of reward distributions supported on $[0,1]$ that we consider (that are $1/2$-sub-Gaussian), by Pinsker's inequality combined with properties of total variation norm, it holds

$$\mathtt{KL}(\mathbf{r}_c, \mathbf{r}'_c) \geqslant \|\mathbf{r}_c - \mathbf{r}'_c\|_{\mathrm{TV}}^2 / 2 \geqslant 2(\mathbf{m}'(c) - \mathbf{m}(c))^2 = \mathtt{d}\big(\mathbf{m}(c) \big| \mathbf{m}'(c)\big)\,.$$

That is, concentration provides a high probability upper bound only on $\mathtt{d}(\mathbf{m}(c) | \mathbf{m}'(c))$ but (of course) not on the more demanding $\mathtt{KL}(\mathbf{r}_c, \mathbf{r}'_c)$.

Discussion. The previous restriction is not problematic, as in the analysis we use the concentration tools of Lemma 23 and Theorem 24 after we deduce from inequalities involving the gains (and hence, the state-action means). This is what is done especially in Appendix E and Equation (57) by considering this straightforward lemma.

Lemma 25 (Mean to Gain) *For a finite set $\mathcal{C}$, consider aggregates $g = \sum_{c \in \mathcal{C}} p_c \cdot \mathbf{m}(c)$ and $g' = \sum_{c \in \mathcal{C}} p_c \cdot \mathbf{m}'(c)$, where $p \in \mathcal{P}(\mathcal{C})$, and where $\mathbf{m}(c), \mathbf{m}'(c) \in \mathbb{R}$ for all $\mathcal{C}$. Let $\overline{c} \in \mathrm{Argmax}_{c \in \mathcal{C}}\big(\mathbf{m}'(c) - \mathbf{m}(c)\big)$ such that gap $\mathbf{m}'(\overline{c}) - \mathbf{m}(\overline{c})$ is maximal. Then, for a given accuracy $\varepsilon > 0$, $g' - g \geqslant \varepsilon$ implies:*

$$(i) \quad \mathbf{m}'(\overline{c}) - \mathbf{m}(\overline{c}) \geqslant g' - g \geqslant \varepsilon, \qquad (ii) \quad \big(\mathbf{m}'(\overline{c}) - \mathbf{m}(\overline{c})\big)^2 / \sigma^2 \geqslant \big(g' - g\big)^2 / \sigma^2\,, \quad \forall \sigma > 0.$$

Reminding the reader that $\mathtt{d}(x|y) = \frac{(x-y)^2}{2\sigma^2} = 2(x-y)^2$, then (ii) above rewrites $\mathtt{d}(\mathbf{m}(\overline{c}) | \mathbf{m}'(\overline{c})) \geqslant \mathtt{d}(g | g')$ as desired for Equation (57) to hold. Hence, we only need a high-probability control on $\mathtt{d}(\mathbf{m}(\overline{c}) | \mathbf{m}'(\overline{c}))$ to conclude and not on $\mathtt{KL}(\mathbf{r}_{\overline{c}}, \mathbf{r}'_{\overline{c}})$. Now, one may prefer to use

a more refined bound. In order to use a more refined (pseudo-)metric $\mathtt{d}'$ such as involving $\mathtt{d}' = \mathtt{KL}$ instead of $\mathtt{d}$, one would need reverse inequalities of the form $\mathtt{d}(g', g) = (g' - g)^2 / 2\sigma^2 \geqslant \kappa_{g,g'} \mathtt{d}'(g, g')$ for some $\kappa_{g,g'} \leqslant 1$. Whenever these are available (such as in one-dimensional exponential families), one would derive from (ii) $\mathtt{d}(\mathbf{m}'(\bar{c})|\mathbf{m}(\bar{c})) \geqslant \mathtt{d}(g'|g)$ the bound $\mathtt{d}(\mathbf{m}'(\bar{c})|\mathbf{m}(\bar{c})) \geqslant \kappa_{g,g'} \mathtt{d}'(g, g')$. This in turn makes appear factors of the form $1/\kappa_{g,g'}$ in the corresponding regret bounds, unless we slightly reshape the IMED-type indexes to handle such factors.

Furthermore, we note it is possible to resort to another variant of Pinsker's inequality specific to regular canonical one-dimensional exponential distributions, in lieu of the one applied to distributions with bounded support. Such distributions include Gaussians with known variance, Bernoulli or Poisson distributions as special cases; we refer to, e.g., (Cappé et al., 2013) for further examples, as well as the proof of the following result.

Lemma 26 (A variant of Pinsker's inequality) *When assuming regular canonical one-dimensional exponential reward distributions* $\mathbf{r}(\cdot)$*, for* $m < m'$*, it holds that*

$$KL\big(\mathbf{r}(m), \mathbf{r}(m')\big) \geqslant \frac{(m' - m)^2}{2\sigma^2},$$

where $\sigma^2 = \max\Big\{\mathbb{V}_{X\sim\mathbf{r}(m'')}(X) \colon m'' \in [m\,, m']\Big\}$.

We refer to Lemma 3 in Appendix A.2.A of (Cappé et al., 2013) for more insights. Thus, using this result we can assume regular canonical one-dimensional exponential reward distributions and ensure the same theoretical guarantees by replacing $\sigma^2 = 1/4$ with $\max_{m\in[m^-,m^+]} \mathbb{V}_{X\sim\mathbf{r}(m)}(X)$ in metric $\mathtt{d}(\cdot|\cdot)$, where m^- and m^+ are such that $m \subset [m^-\,, m^+]$. This enables to replace Assumption 2, assuming bounded support, with the following one:

Assumption 6 (An alternative to Assumption 2) *We assume regular canonical one-dimensional exponential family reward distributions* $\mathbf{r}(\cdot)$ *with bounded mean* $\mathbf{m}(\cdot) \in [0, 1)$.

Appendix G. Computing the set of neighborhood policies

In this section, we explain how we compute $\Pi_\tau(1)$ which is the 1-neighborhood of the empirical optimal policy augmented by some randomly chosen policies. Computing the indexes of policies in $\Pi_\tau(1)$ is more complicated than it seems since some polices might be multi-chain, i.e., they have multiple recurrent classes. In such cases, neither the gain nor the index is uniquely defined and we must decompose the policy on all its recurrent classes in order to compute one gain and one index per class.

The k-neighborhood of a policy can be computed recursively from the 1-neighborhood. To compute the 1-neighborhood, we iterate through all states and actions to create $S \times (A-1)$ new policies. Let us denote by π_{sa} the policy that corresponds to the modification of $\widehat{\pi}_\star$ where action a (different from $\widehat{\pi}_\star(s)$) is chosen in state s. We compute the associated Markov chain thanks to our knowledge of transitions. If the policy is unichain, i.e., the associated Markov chain has only one recurrent class, then we compute the (unique) gain of the policy and add it to the 1-neighborhood pool. On the other hand, if the policy is multi-chain — namely the associated Markov chain has $p > 1$ recurrent classes —, then

there are up to p gains associated to this policy, one for each recurrent class. In this case, we compute all the recurrent classes, all the associated gains and we register p different policies in $\Pi_\tau(1)$, one for each recurrent classes. This way, we will compute p different indexes for the policy π_{sa}, associated to each possible gain. In a sense, we derive from a multi-chain policy p unichain policies.

To find the different stationary distributions, we use the Grassmann–Taksar–Heyman (GTH) algorithm of Grassmann et al. (1985). The set of neighborhood policies is re-computed only when the empirical policy changes but we still randomly sample distributions to add to the set Π_τ as described in the main part of the paper. Also, if a randomly selected policy is multi-chain, we apply the same decomposition procedure we described.

Appendix H. Experiments

In this section, we report additional experimental results. One is concerned about confirming the effectiveness of `IMED-KD` in grid-like environment. To this end, we run an experiment in the 2-rooms environment. Another one is conducted to investigate the efficacy of an adaptively chosen parameter η. To this effect, we present an environment in which `IMED-KD` is not highly competitive and we show that its performances can drastically improve if we chose to use an adaptive η.

n-rooms We recall in Fig. 7 the results that we got in 4-rooms environment. We confirm the impressive performances of `IMED-KD` in grid-like environment by running an experiment in the 2-rooms environments depicted in Fig. 10 along with the regret curves. In this environment with 9×11 states, there are 4 actions, similar to what we described in the main experimental section 7. It can be observed that `IMED-KD` is again, and by far, the best of tested algorithm in this environment. Even `PSRL` that was somewhat competitive in 2-rooms suffers a large linear regret for a very long time. In fact, we were unable to find a reasonable horizon under which regret curves start bending. Those two results emphasize the effectiveness of `IMED-KD` in grid-like environments.

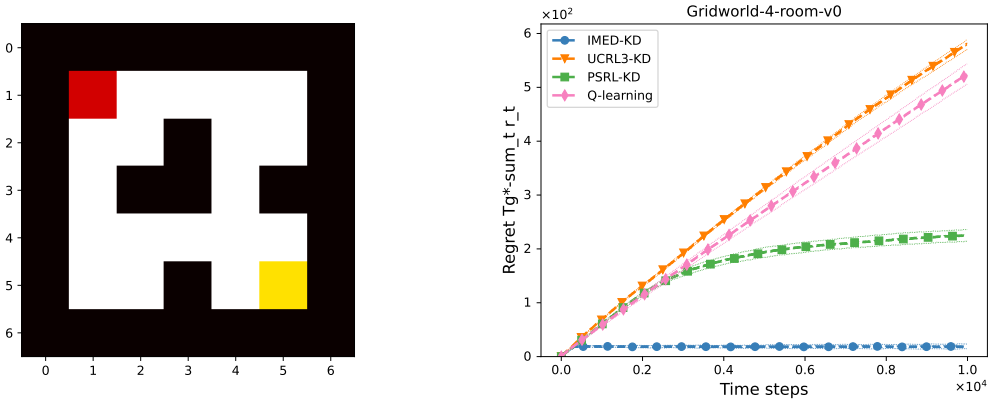

Figure 7: The *4-rooms* environment (left) and regret curves (right)

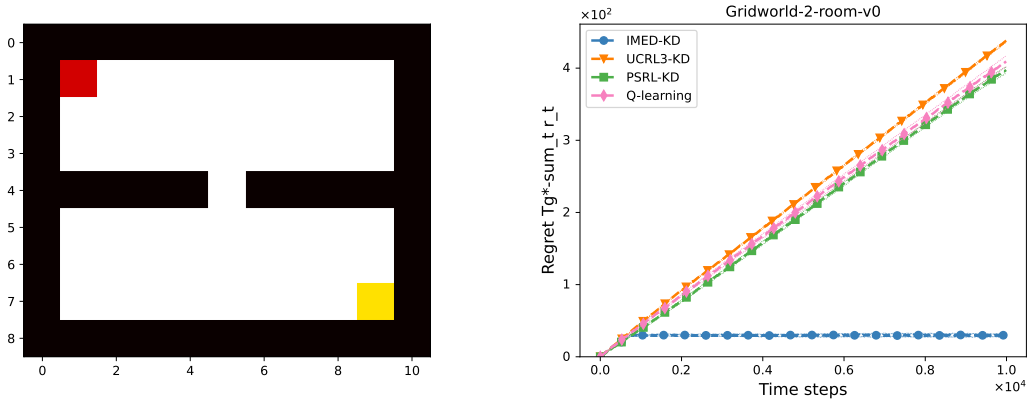

Figure 8: The *2-rooms* environment (left) and regret curves (right)

Finally, we run an experiment in the 4-rooms environment with a horizon of 10^5 (but only using 1024 runs). As depicted in Fig. 9, the regret under `IMED-KD` is substantially smaller than the rest. Furthermore, while Q-learning suffers from a linear regret, it can be observed that the regret of `UCRL3` enters the sublinear phase around time step 6×10^4.

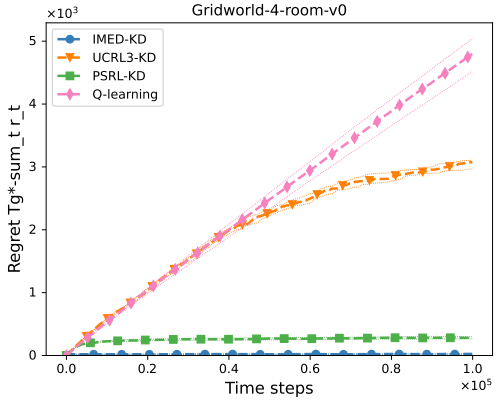

Figure 9: Regret curves in the *4-rooms* environment with horizon of 10^5

Adaptive η When there is a large gap between the gain of an optimal policy and the second largest gain achievable on an MDP, it could be that the chosen value of η impair our `IMED-KD` from distinguishing between the two and maybe even worse, reverse the order of the best two gains. Mathematically, it could be that the ordering of the policies according to g_π is different from the one computed from the modified gains g_π^η where only state-action pairs that are visited with a frequency larger than η are taken into account in the computation of the modified gain. Because rewards are bounded, the error is controlled by η and therefore the order of policies with different gains is preserved if η is smaller than half of the smallest difference between gains.

Therefore, we know that if η is small enough, one can preserve the ordering of policies, thus making IMED-KD safe. Since we do not know the gaps in advance, there could be no constant choice of η that will do well in every environment. One promising solution seems to make η a decreasing function of the time T, where $\eta(T)$ tends to 0 as T grows large. This ensures that after a horizon T_0, $\eta(T_0)$ is small enough to preserve the ordering of gains. We provide below a sketch of proof.

Regret under adaptive parameter (sketch of proof) Note that provided that $\eta < \frac{\varepsilon_{\mathbf{M}}(0)}{2\mathbf{m}_{\max}S}$, if some $\pi' \in \mathcal{V}_\pi$ satisfies $\mathbf{g}_{c,\pi'}(\eta) > \mathbf{g}_{c,\pi}(\eta)$, then it also satisfies $\mathbf{g}_{c,\pi'} > \mathbf{g}_{c,\pi}$, and thus a policy improvement can be obtained correctly in a neighborhood of any policy. Unfortunately, since $\varepsilon_{\mathbf{M}}(0)$ is unknown, this motivates us to consider some η_t decreasing with t and to introduce $T_0 = \min\{t : \eta_t < \frac{\varepsilon_{\mathbf{M}}(0)}{2\mathbf{m}_{\max}S}\}$. For $\eta_t \to 0$, $T_0 < \infty$ and for all $T \geqslant T_0$, then the regret incurred in the subsequent time steps is controlled by $\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T - T_0)$, hence $\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T) \leqslant T_0 + \mathcal{R}_{\mathbf{M}}(\mathbb{A}, T - T_0)$. Upper bounding $K(\varepsilon, \eta)$ by $K(\varepsilon, 0)$ and replacing $\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta)\big|\mathbf{g}^\star_c(\eta)\big)$ with its worst approximation $\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta_{T_0})\big|\mathbf{g}^\star_c(\eta_{T_0})\big)$, we can thus obtain the following bound:

$$\mathcal{R}_{\mathbf{M}}(\mathbb{A}, T) \leqslant T_0 + \left[\max_{\substack{c\in\mathcal{C}\\ \pi\neq\pi^\star}} \frac{(1+\alpha_{\mathbf{M}}(\varepsilon))\log(T-T_0)}{\mathtt{d}\big(\mathbf{g}_{c,\pi}(\eta_{T_0})\big|\mathbf{g}^\star_c(\eta_{T_0})\big)} + K_T(\varepsilon,0)(D_{\mathbf{M}}+2B)\right] \cdot 2(D_{\mathbf{M}}+2B)\,|\mathcal{C}|\ ,$$

under the only assumption that η_t decreases towards 0 with t.

The benefit of rarely-switching algorithm (Algorithm 1) In this last series of experiments, we illustrate the potential benefit of rarely-switching algorithm design (in Algorithm 1) in itself, by experimentally comparing classical strategies to variants of them using Algorithm 1. Indeed, IMED-KD is derived by instantiating Algorithm 1 with the IMED approach. There the IMED algorithm is used in line 11 of the Algorithm 1, where IMED indexes are computed for each of the policies in the considered policy space and the IMED selection rule is applied to compute the policy to play in the next episode. Therefore, one could in principle consider a UCB or TS approach as well (although these are not analyzed).

While we do not extend our theoretical analysis to other bandit algorithms, in this section we report numerical experiments illustrating the performance of the rarely-switching versions of TS and UCB. For these experiments, we assume Gaussian-like reward distributions. Because the rewards are bounded in $[0, 1]$, we can assume that the variance is upper bounded by $\frac{1}{2}$. More precisely, TS-RS consists in using the following TS selection rule in line 11 of Algorithm 1. For all policy a, we sample a reward $x_a(t)$ from posterior distribution $\mathcal{N}(\widehat{\mu}_a(t), \frac{1}{2\sqrt{N_a(t)}})$ and select the policy with the largest sample, $\pi_{t+1} \in \operatorname{argmax} x_a(t)$. Whereas UCB-RS consists in using the following UCB selection rule in line 11 of Algorithm 1. For all policy a, we compute a reward upper bound $u_a(t)$ from UCB index for $\frac{1}{2}$ sub-Gaussian distributions, $u_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{\log t}{2N_a(t)}}$ and select the policy with the largest upper bound, $\pi_{t+1} \in \operatorname{argmax} u_a(t)$.

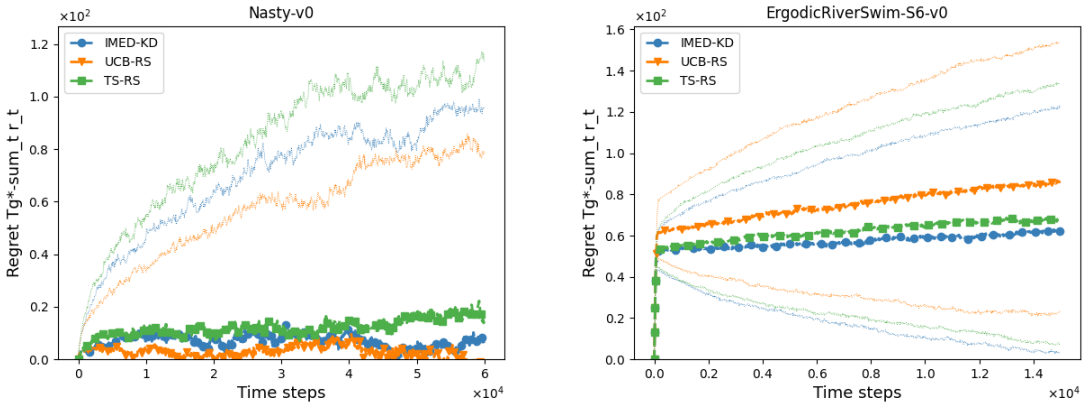

Figure 10: `IMED-KD` against Algorithm 1 combined with TS and UCB in Nasty (left) and 6-state RiverSwim (right)

We compare those strategies in the 6-state RiverSwim environment (see Fig. 10 (right)). It can be seen that Algorithm 1 can indeed be used, at least experimentally, with other bandit sampling strategies with good empirical performances. Still, we observe that our original design suffer the smallest regret. The experiment was run on 5120 independent runs and the horizon was fixed to 20000. We also compare the designs on the Nasty environment in Fig. 10 (left), where it is shown that `IMED-KD` still is better that the other algorithms, albeit by a margin so small that its can be considered equivalent of other designs in this experimental setting. Whatever the experiments, Algorithm 1 seems like a good enough design to be used with other bandit strategies, and our preferred and studied strategy `IMED-KD` proves the best empirically.