IMED-RL: Regret optimal learning of ergodic Markov decision processes

Anonymous Author(s) email
Affiliation
Address

Abstract

We consider reinforcement learning in a discrete, undiscounted, infinite-horizon Markov decision problem (MDP) under the average-reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor transitions of the MDP. In light of their success at regret minimization in multi-armed bandits, popular bandit strategies, such as the optimistic `UCB`, `KL-UCB` or the Bayesian Thompson sampling strategy, have been extended to the MDP setup. Despite some key successes, existing strategies for solving this problem either fail to be provably asymptotically optimal, or suffer from prohibitive burn-in phase and computational complexity when implemented in practice. In this work, we shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (`IMED`) bandit algorithm. Traditional asymptotic problem dependent lower bounds on the regret are known under the assumption that the MDP is *ergodic*. Under this assumption, we introduce `IMED-RL` and prove that its regret upper bound asymptotically matches the regret lower bound. We discuss both the case when the supports of transitions are unknown, and the more informative but a priori harder-to-exploit-optimally case when they are known. Rewards are assumed light-tailed, semi-bounded from above. Last, we provide numerical illustrations on classical tabular MDPs, *ergodic* and *communicative* only, showing the competitiveness of `IMED-RL` in finite-time against state-of-the-art algorithms. `IMED-RL` also benefits from a light complexity.

Keywords: Reinforcement Learning, Markov Decision Process, average-reward, bandit, lower bound

1. Introduction

We study Reinforcement Learning (RL) with an unknown finite Markov Decision Problem (MDP) under the average-reward criterion in which a learning algorithm interacts sequentially with the dynamical system, without any reset, in a single and infinite sequence of observations, actions, and rewards while trying to maximize its total accumulated rewards over time. Formally, we consider a finite MDP $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$, where $\mathcal{S}$ is the finite set of states, $\mathcal{A} = (\mathcal{A}_s)_{s\in\mathcal{S}}$ specifies the set of actions available in each state and we introduce the set of pairs $\mathcal{X}_\mathbf{M} = \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$ for convenience. Further[1], $\mathbf{p} : \mathcal{X}_\mathbf{M} \to \mathcal{P}(\mathcal{S})$ is the transition distribution function and $\mathbf{r} : \mathcal{X}_\mathbf{M} \to \mathcal{P}(\mathbb{R})$ the reward distribution function, with corresponding mean reward function denoted by $\mathbf{m} : \mathcal{X}_\mathbf{M} \to \mathbb{R}$. An agent interacts with the MDP at discrete time steps $t \in \mathbb{N}^*$ and yields a random sequence $(s_t, a_t, r_t)_t$ of states, actions, and rewards in the following way. At each time step t, the agent observes the current state s_t and decides the action a_t to take based on s_t and possibly past information, *i.e.* previous elements of the sequence. After playing a_t, it observes a reward $r_t \sim \mathbf{r}(s_t, a_t)$, the current state of the MDP changes to $s_{t+1} \sim \mathbf{p}(\cdot|s_t, a_t)$ and the agent proceeds sequentially. In the *average-reward setting*, one is interested in maximizing the limit, $\frac{1}{T}\sum_{t=1}^{T} r_t$, when $T \to \infty$, providing it exists. This setting is a popular framework for studying sequential decision making problems; it can be traced back to seminal papers such as those of Graves and Lai (1997), and Burnetas and Katehakis (1997). This theoretical framework allows to study the *exploration-exploitation* trade-off that arises from the sequential optimization problem a learner is trying to solve while being uncertain about the very problem it is optimizing.

In this paper, one is interested in developing a sampling strategy that is *optimal* amongst strategies that aim at maximizing the average-reward, *i.e.* balancing exploration and exploitation in an optimal way. To assert optimality, we define the notion of *regret* and state a *regret lower bound* with the purpose of defining a theoretically sound notion of optimality that is *problem dependent*. While *regret* defines the discrepancy to optimality of a learning strategy, a *problem dependent regret lower bound* will formally assess the minimal regret that any learning algorithm must incur on a given MDP problem by computing a minimal rate of exploration. Because this minimal rate of exploration depends on the problem,

1. Given a set E, $\mathcal{P}(E)$ denotes the set of probability distributions on E.

©2022 Anonymous Author(s).
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/.

it is said to be problem dependent, as opposed to worst case regret study that can exist in the MDP literature (*e.g.* Jaksch et al. (2010)). Regret lower bounds currently exist in the literature when the MDP $\mathbf{M}$ is assumed to be *ergodic*[2]. Hence we hereafter make this assumption, in order to be able to compare the regret of our algorithm to an optimal bound. Similarly, to ensure fast enough convergence of the empirical estimate of the reward to the true mean, an assumption controlling the rate of convergence to the mean is necessary.

Assumption 1 (Light-tail rewards)**.** *For all $x \in \mathcal{X}_{\mathbf{M}}$, the moment generating function of the reward exists in a neighborhood of* 0*:* $\exists \lambda_x > 0, \forall \lambda \in \mathbb{R}$ *such that* $|\lambda| < \lambda_x, \mathbb{E}_{R\sim \mathbf{r}(x)}[\exp(\lambda R)] < \infty$.

Policy Regret and ergodicity are defined using properties of the set of stationary deterministic policies $\Pi(\mathbf{M})$ on $\mathbf{M}$. On $\mathbf{M}$, each stationary deterministic policy $\pi : \mathcal{S} \to \mathcal{A}_s$ defines a Markov reward process, *i.e.* a Markov chain on $\mathcal{S}$ with kernel $\mathbf{p}_\pi : s \in \mathcal{S} \mapsto \mathbf{p}\left(\cdot|s, \pi(s)\right) \in \mathcal{P}\left(\mathcal{S}\right)$ together with rewards $\mathbf{r}_\pi : s \in \mathcal{S} \mapsto \mathbf{r}\left(s, \pi(s)\right) \in \mathcal{P}\left(\mathbb{R}\right)$ and associated mean rewards $\mathbf{m}_\pi : s \in \mathcal{S} \mapsto \mathbf{m}\left(s, \pi(s)\right) \in \mathbb{R}$. The t-steps transition kernel of π on $\mathbf{M}$ is denoted $\mathbf{p}_\pi^t$. We denote $\overline{\mathbf{p}}_\pi = \lim\limits_{T\to\infty} \frac{1}{T}\sum\limits_{t=1}^{T} \mathbf{p}_\pi^{t-1} : \mathcal{S} \to \mathcal{P}(\mathcal{S})$ the Cesaro-average of $\mathbf{p}_\pi$. A learning agent is executing a sequence of policies $\pi_t \in \Pi(\mathbf{M})$, $t \geqslant 1$, where π_t depends on past information $\left(s_{t'}, a_{t'}, r_{t'}\right)_{t'<t}$. With a slight abuse of notation, a sequence of identical decision rules, $\pi_t = \pi$ for all t, is also denoted π.

Gain The cumulative reward (value) at time T, starting from an initial state s_1, of policy $\pi = (\pi_t)_t$ is formally given by

$$V_{s_1}(\mathbf{M}, \pi, T) = \mathbb{E}_{\pi,\mathbf{M},s_1}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{\pi,\mathbf{M},s_1}\left[\sum_{t=1}^{T} \mathbf{m}(s_t, a_t)\right] = \sum_{t=1}^{T}\Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}}\mathbf{m}_{\pi_{t'}}\Big)(s_1)\,. \tag{1}$$

For $\pi \in \Pi(\mathbf{M})$, the average-reward $\frac{1}{T}V_{s_1}(\mathbf{M}, \pi, T)$ tends to $\left(\overline{\mathbf{p}}_\pi \mathbf{m}_\pi\right)(s_1)$ as $T \to \infty$. The gain of policy $\pi \in \Pi(\mathbf{M})$, when starting from state s_1 is defined by $\mathbf{g}_\pi(s_1) = (\overline{\mathbf{p}}_\pi \mathbf{m}_\pi)(s_1)$ and the optimal gain $\mathbf{g}^{\mathbf{M}}(s_1)$ is defined as the maximum of the gains achievable by a deterministic policy, $\mathbf{g}^{\mathbf{M}}(s_1) = \max_{\pi\in\Pi(\mathbf{M})} \mathbf{g}_\pi(s_1)$. The set of policies achieving maximal gain on $\mathbf{M}$ starting from state s is $\mathcal{O}_s(\mathbf{M}) = \{\pi \in \Pi\left(\mathbf{M}\right) : \mathbf{g}_\pi(s) = \mathbf{g}^{\mathbf{M}}(s)\}$.

Definition 1 (Regret)**.** *The regret at time T of a learning policy $\pi = (\pi_t)_t$ starting at state s on an MDP $\mathbf{M}$ is defined with respect to any $\pi^\star \in \mathcal{O}_s\left(\mathbf{M}\right)$, as*

$$\mathcal{R}_{\pi,s}\left(\mathbf{M}, T; \pi^\star\right) = V_s(\mathbf{M}, \pi^\star, T) - V_s(\mathbf{M}, \pi, T)\,. \tag{2}$$

In this paper, we aim to find a learning algorithm with *asymptotic* minimal regret. The Lemma 2 will prove that for all optimal policies, $\pi^\star$, regrets are the same up to a bounded term that therefore does not count in asymptotic analysis. Some authors such as Bourel et al. (2020) define the regret as $T\mathbf{g}^{\mathbf{M}}(s) - V_s(\mathbf{M}, \pi, T)$ which is equal to the one we defined up to a bounded term (again by Lemma 2). No stationary policy can be optimal at all time and the important fact is that all those notions of regret induce the same asymptotic lower bound. In the considered setting, the learning agent interacts with the MDP without any reset. The minimal assumption would be to allow the agent to come back with positive probability from any initial mistake in finite time, so that the agent is not stuck in a sub-optimal area of the system. This is assuming that the MDP is *communicating*, that is $\forall s, s', \exists \pi, t \in \mathbb{N} : \mathbf{p}_\pi^t(s'|s) > 0$. However, in the literature, lower bounds on the regret are stated for MDPs satisfying a stronger assumption, *ergodicity*. Since one is interested in crafting an algorithm matching a lower bound, we consider this stronger assumption.

Assumption 2 (Ergodic MDP)**.** *The MDP $\mathbf{M}$ is ergodic, that is* $\forall s, s', \forall \pi, \exists t \in \mathbb{N} : \mathbf{p}_\pi^t(s'|s) > 0$.

Intuitively, this means that for all policies and all couples of states, there exists a finite trajectory of positive probability between the states. Interestingly, the ergodic property can be assumed on the MDP or on the set of policies in which we seek an optimal one. For instance, in any communicating MDP, all ε-soft policies[3] are ergodic; more in the experiment Section 5 and Appendix E.

2. We prefer the term *ergodic* over the more accurate one, *irreducible* as it is a standard abuse of terminology in the MDP community. Mathematically, an MDP is ergodic if irreducible, aperiodic and positive recurrent. In this paper, irreducibility only is used.
3. A policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}_s)$ is ε-soft if $\pi(a|s] \geqslant \varepsilon/|\mathcal{A}_s|$ for all s and a.

Related work Had the MDP only one state, it would be a bandit problem. Lower bound on the bandit regret and algorithms matching this lower bound, sometimes up to a constant factor, are well studied in the bandit literature. Therefore, bandit sampling strategies with known theoretical guarantees have inspired RL algorithms. The `KL-UCB` algorithm (Burnetas and Katehakis (1996); Maillard et al. (2011)), has inspired the strategy of the seminal paper of Burnetas and Katehakis (1997), as well the more recent `KL-UCRL` strategy (Filippi et al. (2010); Talebi and Maillard (2018)). Inspired by the infamous `UCB` algorithm (Agrawal (1995); Auer et al. (2002)), a number of strategies implementing the optimism principle have emerged such as `UCRL` (Auer and Ortner (2006)), `UCRL2` (Jaksch et al. (2010)) and `UCRL3` (Bourel et al. (2020) (and beyond, for the related episodic setup). The strategy `PSRL` (Osband et al. (2013)) is inspired by Thompson sampling (Thompson (1933)).

Outline and contribution In this work, we build on the `IMED` strategy (Honda and Takemura (2015)), a bandit algorithm that benefits from practical and optimal guarantees but has never been used by the RL community. We fill this gap by proposing the `IMED-RL` algorithm which we prove to be asymptotically optimal for the average-reward criterion. We revisit the notion of skeleton (equation 11) introduced in the seminal work of Burnetas and Katehakis (1997), with a subtle but key modification that prevents a prohibitive burn-in phase (see Appendix G for further details). Further, this novel notion of skeleton enables `IMED-RL` to remove any tracking or hyper-parameter and mimic a *stochastic-policy-iteration-like* algorithm. [4] Further, this skeleton scales naturally with the studied MDP as it does not explicitly refer to absolute quantities such as the time. We prove that our proposed `IMED-RL` is asymptotically optimal and show its numerical competitivity.

Building on `IMED`, we make an additional assumption on the reward that is less restrictive than the common bounded reward hypothesis made in the RL community.

Assumption 3 (Semi-bounded rewards)**.** *For all $x \in \mathcal{X}_{\mathbf{M}}$, $r(x)$ belongs to a subset $\mathcal{F}_x \subset \mathcal{P}(\mathbb{R})$ known to the learner.*[5] *There exists a known quantity $m_{\max}(x) \in \mathbb{R}$ such that for all $x \in \mathcal{X}$, the support $\mathtt{Supp}(\mathbf{r}(x))$ of the reward distribution is semi-bounded from above, $\mathtt{Supp}(\mathbf{r}(x)) \subset]-\infty, m_{max}(x)]$, and its mean satisfies $\mathbf{m}(x) < m_{\max}(x)$.*

2. Regret lower bound

In this section, we recall the regret lower bound for ergodic MDPs and provide a few insights about it.

Characterizing optimal policies Relying on classical results that can be found in the books of Puterman (1994) and Hernández-Lerma and Lasserre (1996), we give a useful characterization of optimal policies that is used to derive a regret lower bound. Under the ergodic Assumption 2 of MDP $\mathbf{M}$, for all policy $\pi \in \Pi(\mathbf{M})$, the gain is independent from the initial state, *i.e.* $\mathbf{g}_\pi(s) = \mathbf{g}_\pi(s')$ for all states s and s', and we denote it $\mathbf{g}_\pi$. Similarly, the set of optimal policies $\mathcal{O}(\mathbf{M})$ is state-independent since $\mathcal{O}_s(\mathbf{M}) = \mathcal{O}_{s'}(\mathbf{M})$. Any policy π satisfy the following fixed point property

$$\textit{(Poisson equation)} \qquad \mathbf{g}_\pi + \mathbf{b}_\pi(s) = \mathbf{m}_\pi(s) + (\mathbf{p}_\pi \mathbf{b}_\pi)(s)\,, \tag{3}$$

where $\mathbf{b}_\pi : \mathcal{S} \to \mathbb{R}$ is called the bias function and is defined up to an additive constant by

$$\mathbf{b}_\pi(s) = \Big(\sum_{t=1}^{\infty} (\mathbf{p}_\pi^{t-1} - \overline{\mathbf{p}}_\pi)\mathbf{m}_\pi \Big)(s).$$

We highlight that bias plays a role similar to the value function in the discounted reward setting in which the gain is always zero and equation 3 reduces to the Bellman equation, giving a direction in which extend our results to this other RL setting. In the average-reward setting, the bias function can also be interpreted as an "intrinsic desirability". For instance, in a tabular environment where all rewards are 0 except for one 1 in one action a in one state s such that $\mathbf{p}(s|s,a) = 1$, the bias function of an optimal policy can be interpreted as the opposite of the average time to get to that desirable state. Under the ergodic Assumption 2 of MDP $\mathbf{M}$, all optimal policies satisfy a Poisson equation while some are also being characterized by the optimal Poisson equation (see Hernández-Lerma and Lasserre (1996)),

$$\mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s) = \max_{a \in \mathcal{A}_s} \left\{ \mathbf{m}(s,a) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s,a)\mathbf{b}^{\mathbf{M}}(s') \right\} . \tag{4}$$

4. The skeleton in Burnetas and Katehakis (1997) is sometimes empty at some states, when t is too small, this causes the strategy to work well only after t is large enough to ensure that the skeleton contains at least one action in each state.

5. *e.g.* Bernoulli, multinomial with unknown support, beta, truncated Gaussians, a mixture of those, *etc*.

This characterization is used to compute the optimal gain and a bias function associated to an optimal policy. Interestingly, for any communicating and a fortiori ergodic MDP, the span $\mathbb{S}(\mathbf{b}_\pi) = \max_{s\in\mathcal{S}} \mathbf{b}_\pi(s) - \min_{s\in\mathcal{S}} \mathbf{b}_\pi(s)$ of the bias function of any policy is bounded, which allows to decompose the regret in the useful following way.

Lemma 2 (Regret decomposition)**.** *Under the ergodic Assumption 2, for all optimal policy $\star \in \mathcal{O}(\mathbf{M})$, the regret of any policy $\pi = (\pi_t)_t$ can be decomposed as*

$$\mathcal{R}_{\pi,s_1}(\mathbf{M}, T; \star) = \sum_{x\in\mathcal{X}_\mathbf{M}} \mathbb{E}_{\pi,s_1}[N_x(T)]\,\Delta_x(\mathbf{M}) + \underbrace{\left(\left[\prod_{t=1}^{T} \mathbf{p}_{\pi_t} - \mathbf{p}_\star^t\right] b_\star\right)(s_1)}_{\leqslant \mathbb{S}(\mathbf{b}_\star)}, \tag{5}$$

where $N_{s,a}(T) = \sum_{t=1}^{T} \mathbb{1}\{s_t = s, a_t = a\}$ counts the number of time the state-action pair (s,a) has been sampled and $\Delta_{s,a}(\mathbf{M})$ is the sub-optimality gap of the state-action pair (s,a) in $\mathbf{M}$,

$$\Delta_{s,a}(\mathbf{M}) = \mathbf{m}(s,a) + \mathbf{p}_a\mathbf{b}_\star(s) - \mathbf{m}_\star(s) - \mathbf{p}_\star\mathbf{b}_\star(s) = \mathbf{m}(s,a) + \mathbf{p}_a\mathbf{b}_\star(s) - \mathbf{g}_\star - \mathbf{b}_\star(s) \tag{6}$$

with $\mathbf{p}_a = \mathbf{p}(\cdot|s,a)$ by a slight abuse of notation. Action $a \in \mathcal{A}_s$ is optimal if and only if $\Delta_{s,a}(\mathbf{M}) = 0$, otherwise, it is said sub-optimal.

This result can be found in Puterman (1994) and is rederived in Appendix C. For the asymptotic lower bound, we will be interested in controlling the first term of Equation 5. Since this quantity does not depend on $\star$ and s_1 does not have much importance under the ergodic assumption, we will simplify the notation by denoting the regret $\mathcal{R}_\pi(\mathbf{M}, T)$.

Lower bound To assess the minimal sampling complexity of a sub-optimal state action pair, one must compute how far a sub-optimal state-action pair is from being optimal from an information point-of-view. A sub-optimal state-action pair $(s,a) \in \mathcal{X}_\mathbf{M}$ is said to be *critical* if it can be made optimal by changing reward $\mathbf{r}(s,a)$ and transition $\mathbf{p}(\cdot|s,a)$ while respecting the assumptions on the rewards and transitions. Formally, let $\varphi_\mathbf{M}(\nu \otimes q) = \mathbb{E}_{R\sim\nu}[R] + q\mathbf{b}^\mathbf{M}$ denotes the potential function of $\nu \otimes q$ in $\mathbf{M}$, where $\nu \otimes q$ is the product measure of ν and q. A pair $(s,a) \in \mathcal{X}_\mathbf{M}$ is *critical* if it is sub-optimal and there exists $\nu \in \mathcal{F}_{s,a}$ and $q \in \mathcal{P}(\mathcal{S})$ such that

$$\varphi_\mathbf{M}(\nu \otimes q) > \gamma_s(\mathbf{M}) \quad \text{where } \gamma_s(\mathbf{M}) \stackrel{\text{def}}{=} \mathbf{g}^\mathbf{M} + \mathbf{b}^\mathbf{M}(s). \tag{7}$$

Note that $\gamma_s(\mathbf{M}) = \max_{a\in\mathcal{A}_s} \varphi_\mathbf{M}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a))$ by the optimal Poisson equation (4).

Definition 3 (Sub-optimality cost)**.** *The **sub-optimality cost** of a sub-optimal state-action pair $(s,a) \in \mathcal{X}_\mathbf{M}$ is defined as $\underline{\mathbf{K}}_{s,a}(\mathbf{M}) \stackrel{\text{def}}{=} \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$ where*

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma) = \inf_{\substack{\nu\in\mathcal{F}_{s,a}\\ q\in\mathcal{P}(\mathcal{S})}} \left\{\mathrm{KL}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a), \nu \otimes q\right) \;:\; \varphi_\mathbf{M}(\nu \otimes q) > \gamma\right\}. \tag{8}$$

A lower bound on the regret may now be stated for a certain class of learner, the set of uniformly consistent learning algorithm, *i.e.* those policies $\pi = (\pi_t)_t$ such that $\mathbb{E}_{\pi,\mathbf{M}}(N_{s,a}(T)) = o(T^\alpha)$ for all sub-optimal state-action pair (s,a) and $0 < \alpha < 1$ (see Agrawal et al. (1989); Burnetas and Katehakis (1997)).

Theorem 4 (Regret lower bound Burnetas and Katehakis (1997))**.** *Let $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ be an MDP satisfying Assumption 1, 2, 3. For all uniformly consistent learning algorithm π,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\pi,\mathbf{M}}[N_{s,a}(T)]}{\log T} \geqslant \frac{1}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})} \tag{9}$$

with the convention that $1/\infty = 0$. The regret lower bound is

$$\liminf_{T\to\infty} \frac{\mathcal{R}_\pi(\mathbf{M}, T)}{\log T} \geqslant \sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})} \tag{10}$$

where $\mathcal{C}(\mathbf{M}) = \left\{(s,a) : 0 < \underline{\mathbf{K}}_{s,a}(\mathbf{M}) < \infty\right\}$ is called the set of critical state-action pairs. Those are the state-action pairs (s,a) that could be confused for an optimal one if we were to change their associated rewards and transitions distributions at the displacement cost of $\underline{\mathbf{K}}_{s,a}(\mathbf{M})$.

3. The IMED-RL Algorithm

In this section we introduce and detail the `IMED-RL` algorithm, whose regret matches this fundamental lower bound and extends the `IMED` strategy from Honda and Takemura (2015) to ergodic MDPs. In particular, when the MDP has only one state, the problem is a Bandit one and `IMED-RL` is `IMED`.

Empirical quantities `IMED-RL` is a *model-based* algorithm that keeps empirical estimates of the transitions $\mathbf{p}$ and rewards $\mathbf{r}$ as opposed to *model-free* algorithm such as Q-learning. We denote by $\hat{\mathbf{r}}_t(s,a) = \hat{\mathbf{r}}(s,a;N_{s,a}(t))$ and $\hat{\mathbf{p}}_t(\cdot|s,a) = \hat{\mathbf{p}}(\cdot|s,a;N_{s,a}(t))$ the empirical reward distributions and transition vectors after t time steps, *i.e.* using $N_{s,a}(t)$ samples from the distribution $\mathbf{r}(s,a)$. Initially, $\hat{\mathbf{p}}(\cdot|s,a;0)$ is the uniform probability over the state space and $\hat{\mathbf{p}}(\cdot|s,a;k) = (1-1/k)\hat{\mathbf{p}}(\cdot|s,a;k-1) + (1/k)\mathbf{s}$, where $\mathbf{s}$ is a vector of zeros except for a one at index s, the k^{th} samples drawn from $\mathbf{p}(\cdot|s,a)$.[6] This defines at each time step t an empirical MDP $\widehat{\mathbf{M}}_t = (\mathcal{S},\mathcal{A},\hat{\mathbf{p}}_t,\hat{\mathbf{r}}_t)$. On this empirical MDP, for each state, some actions have been sampled more than others and their empirical quantities are therefore better estimated. We call *skeleton* at time t the subset of state-action pairs that can be considered sampled enough at time t; it is defined by restricting $\mathcal{A}_s$ to $\mathcal{A}_s(t)$ for all state $s \in \mathcal{S}$, with

$$\mathcal{A}_s(t) = \left\{ a \in \mathcal{A}_s \ : \ N_{s,a}(t) \geqslant \log^2\left(\max_{a' \in \mathcal{A}_s} N_{s,a'}(t)\right)\right\}. \tag{11}$$

Since $x > \log^2 x$, $\mathcal{A}_s(t) \neq \emptyset$, hence $\mathcal{A}(t) = (\mathcal{A}_s(t))_s$ contains at least one deterministic policy. We note that the MDP $\mathbf{M}(\mathcal{A}(t)) \stackrel{\text{def}}{=} (\mathcal{S},\mathcal{A}(t),\mathbf{p},\mathbf{r})$ defined by restricting the set of actions to $\mathcal{A}(t) \subseteq \mathcal{A}$ is an ergodic MDP. The restricted empirical MDP $\widehat{\mathbf{M}}_t(\mathcal{A}(t)) = (\mathcal{S},\mathcal{A}(t),\hat{\mathbf{p}}_t,\hat{\mathbf{r}}_t)$ also is ergodic thanks to the ergodic initialization of the estimate $\hat{\mathbf{p}}$. Inspired by `IMED`, we define the `IMED-RL` index.

Definition 5 (`IMED-RL` index)**.** *For all state-action pairs $(s,a) \in \mathcal{X}_{\mathbf{M}}$, define $\mathbf{K}_{s,a}(t) \stackrel{\text{def}}{=} \underline{\mathbf{K}}_{s,a}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t)), \hat{\gamma}_s(t)\right)$ with empirical threshold $\hat{\gamma}_s(t) \stackrel{\text{def}}{=} \max_{a \in \mathcal{A}_s} \varphi_{\hat{\mathbf{M}}_t(\mathcal{A}(t))}(\hat{\mathbf{r}}(s,a) \otimes \hat{\mathbf{p}}(\cdot|s,a))$. Then, the `IMED-RL` index of (s,a) at time t, $\mathbf{H}_{s,a}(t)$, is defined as*

$$\mathbf{H}_{s,a}(t) \quad = \quad N_{s,a}(t)\underline{\mathbf{K}}_{s,a}(t) + \log N_{s,a}(t)\,, \tag{12}$$

Note that $\hat{\gamma}_s(t) \neq \gamma_s(\hat{\mathbf{M}}_t(\mathcal{A}(t)))$ as the maximum is taken over all $a \in \mathcal{A}_s$ an not just $a \in \mathcal{A}_s(t)$.

Known support of transitions Were the support of transition known, the infimum in sub-optimality cost $\underline{\mathbf{K}}_{s,a}$ defined by equation 8 would be redefined as one over the set $\{q \in \mathcal{P}(\mathcal{S}) : \mathtt{Supp}(q) = \mathtt{Supp}(\mathbf{p}(\cdot|s,a))\}$, modifying both the lower bound and `IMED-RL` index.

`IMED-RL` algorithm The `IMED-RL` algorithm consists in playing at each time step t, an action a_t of minimal `IMED-RL` index at the current state s_t. The intuition behind the `IMED-RL` index is similar stems from a Bayesian point-of-view of the lower bound. At a given time t, the frequency of play $\frac{N_{s,a}(t)}{N_s(t)}$ of action $a \in \mathcal{A}_s$ in state $s \in \mathcal{S}$, should be larger than or equal to its "posterior probability" of being the optimal action in that state, $\exp\left(-N_{s,a}(t)\underline{\mathbf{K}}_{s,a}(t)\right)$, that is to say $\frac{N_{s,a}(t)}{N_s(t)} \geqslant \exp\left(-N_{s,a}(t)\underline{\mathbf{K}}_{s,a}(t)\right)$. Taking the logarithm and rearranging the terms, this condition rewrites $\mathbf{H}_{s,a}(t) \geqslant \log N_s(t)$ at each time step t. An action that is the closest to violate this condition or that violates this condition the most is the one of minimal `IMED-RL` index, thus `IMED-RL` plays an action in $\arg\min_a \mathbf{H}_{s,a}(t)$.

Algorithm 1 `IMED-RL`: **I**ndexed **M**inimum **E**mpirical **D**ivergence for **R**einforcement **L**earning

---

Require: State-Action space $\mathcal{X}_{\mathbf{M}}$ of MDP $\mathbf{M}$, Assumptions 1, 2, 3
Require: Initial state s_1
  for $t \geqslant 1$ **do**
    Sample $a_t \in \arg\min_{a \in \mathcal{A}_{s_t}} \mathbf{H}_{s_t,a}(t)$
  end for

---

6. This estimation of $\mathbf{p}$ is sometimes called a Laplace estimator or an additive smoothing.

Intuitions of the IMED-RL algorithm root to the control theory of MDPs and optimal bandit theory; IMED-RL intertwines the two and the regret proof exactly follows from the following intuitions.

Control In control theory, we assume that both the expected rewards and transitions probabilities of an MDP $\mathbf{M}$ are known. *Policy iteration* (see Puterman (1994), Bertsekas and Shreve (1978)) is an algorithm that computes a sequence $(\pi_n)_n$ of deterministic policies that are increasingly strictly better until an optimal policy is reached. In the average-reward setting and under the ergodic assumption, a policy π is strictly better than another policy π' if $\mathbf{g}_\pi(\mathbf{M}) = \mathbf{g}^{\mathbf{M}(\pi)} > \mathbf{g}_{\pi'}(\mathbf{M}) = \mathbf{g}^{\mathbf{M}(\pi')}$. Here, we emphasized that a policy can be seen as special case of a skeleton. The policy iteration algorithm computes the sequence of policies, recursively in the following way. Initially, an arbitrary deterministic policy π_0 is chosen. At step $n+1 \in \mathbb{N}^*$, it computes $\mathbf{m}_{\pi_n}$ and $\mathbf{b}_{\pi_n}$ then swipes through the states $s \in \mathcal{S}$ in an arbitrary order until it reaches one state s such that there exists $a \in \mathcal{A}(s)$ with $\mathbf{m}(s,a) + \mathbf{p}(\cdot|s,a)\mathbf{b}_{\pi_n} > \mathbf{m}_{\pi_n}(s) + \mathbf{p}_\pi(s)\mathbf{b}_{\pi_n}$. If such an s does not exist, then it returns π_n as an optimal policy. Otherwise, π_{n+1} is defined as $\pi_{n+1}(s') = \pi_n(s')$ for all $s \neq s'$ and $\pi_{n+1}(s) \in \arg\max\{\mathbf{m}(s,a) + \mathbf{p}(\cdot|s,a)\mathbf{b}_{\pi_n}\}$. Such a step is called a policy improvement step. Policy iteration is guaranteed to finish in a finite number as the cardinal of $\Pi(\mathbf{M})$ is finite. At each step $n \in \mathbb{N}^*$, $\varphi_{\mathbf{M}(\pi_n)}$ is a local function that takes into account the whole dynamic of the MDP and allows to compute, *via* an *argmax*, an optimal choice of improvement (or optimal action) based on local information; $\varphi_{\mathbf{M}(\pi_n)}(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a)) = \mathbf{m}(s,a) + \mathbf{p}(\cdot|s,a)\mathbf{b}_{\pi_n}$. IMED-RL uses $\varphi_{\widehat{\mathbf{M}}(\mathcal{A}(t))}$ and improves the skeleton similarly to policy iteration as it can be seen in the analysis 4.

Bandit control A degenerate case of MDP would be one where there is only one state s with $\varphi_\mathbf{M}(\mathbf{r}(s,a)) = \mathbf{m}(s,a)$ by choosing the bias function to be zero.[7] Playing optimally consists in playing an action with largest expected reward at each time step t, $a_t \in \arg\max_{a \in \mathcal{A}_s} \mathbf{m}(s,a)$.

Bandit Learning occurs when rewards are unknown; this is the bandit problem. In that case, a lower bound on the regret similar to 4 exists. Under some assumptions on the reward distributions, optimal algorithms whose regret upper bounds asymptotically match the lower bound can be derived. IMED (Honda and Takemura (2015)) and KL-UCB (Maillard et al. (2011); Cappé et al. (2013)) are two such examples that use indexes, *i.e.* computes a number $I_{s,a}(t)$ at each time step and play $a_t \in \arg\min I_{s,a}(t)$ (for the unique state s). Such indexes are crafted to correctly handle the *exploration-exploitation* trade-off.

RL in Ergodic MDPs The delayed rewards caused by the dynamic of the system is the main source of difficulty arising from having more than one state. IMED-RL combines control and bandit theory in the following way. At each time step t, a restricted MDP $\widehat{\mathbf{M}}_t(\mathcal{A}(t))$ is built from the empirical one $\widehat{\mathbf{M}}_t$. If the condition to belong to the skeleton is selective enough, then the potentials on the restricted empirical MDP $\widehat{\mathbf{M}}_t(\mathcal{A}(t))$ may become close to those of the restricted true MDP $\mathbf{M}(\mathcal{A}(t))$, that is $\|\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \varphi_{\mathbf{M}(\mathcal{A}(t))}\|_\infty$ is small. We want to make policy improvements by finding, at each state s an action $a' \in \arg\max \varphi_{\mathbf{M}(\mathcal{A}(t))}(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a))$, play it enough that it belongs to the skeleton which will modify φ and repeat until $\varphi_{\mathbf{M}(\mathcal{A}(t))} = \varphi_\mathbf{M}$. Using φ, the global dynamic is reduced to a local function so that at each state, the agent is presented a bandit problem. This bandit problem is well estimated if $\|\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \varphi_{\mathbf{M}(\mathcal{A}(t))}\|_\infty$ is small. As opposed to the control setting, the learning agent cannot choose the state in which to make the policy improvement step and it may be possible that no policy improvement step is possible at state s_t. However, thanks to the ergodic Assumption 2 the agent is guaranteed to visit such a state in finite time, if it exists. There is a trade-off between the adptativity of the skeleton, *i.e.* how quickly one can add an improving action to define a new φ, and concentration of statistical quantities defined on the MDP restricted by the skeleton.

4. Regret of IMED-RL

In this section we state the main theoretical result of this paper, which consists in the IMED-RL regret upper bound. We then sketch a few key ingredients of the proof.

7. recall that the bias function is defined up to an additive constant.

Theorem 6 (Regret upper bound for Ergodic MDPs)**.** *Let $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ be an MDP satisfying assumptions 1, 2, 3. Let $0 < \varepsilon \leqslant \frac{1}{3} \min_{\pi \in \Pi(\mathbf{M})} \min_{(s,a) \in \mathcal{X}_{\mathbf{M}}} \{|\Delta_{s,a}(\mathbf{M}(\pi))| : |\Delta_{s,a}(\mathbf{M}(\pi))| > 0\}$. The regret of `IMED-RL` is upper bounded,*

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M}) - \varepsilon \Gamma_s(\mathbf{M})} \right) \log T + O(1), \tag{13}$$

where $\Gamma_s(\mathbf{M})$ is constant that depends on the MDP $\mathbf{M}$ and state s; it is made explicit in the proof detailed in Appendix D. A Taylor expansion allows to write the regret upper bound as

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})} \right) \log T + O\left((\log T)^{10/11} \right). \tag{14}$$

Were the semi-bounded reward assumption changed to a bounded reward one with known upper and lower bound, the $O\left((\log T)^{10/11}\right)$ could be made a $O(1)$ as explained in Appendix E.

This Theorem proves the optimality of `IMED-RL` since the upper bound on the regret matches the lower bound of Theorem 4. Such a bound was asymptotically matched by the algorithm proposed by Burnetas and Katehakis (1997) and we recall that this algorithm and its problems are discussed in Appendix G. On the other hand, the current state-of-the-art algorithms `UCRL3` and `PSRL`, while having some theoretical guarantees, have not been proved to match the regret lower bound. On the practical side, Q-learning is often used without much theoretical guarantee because of its usually strong practical performances. In the experiments, we will compare `IMED-RL` to those three algorithms.

Sketch of proof While a full proof is given in Appendix D, we sketch here the main proof ideas that follow directly from the intuitions behind the `IMED-RL` conception. The regret is decomposed into two terms, the **bandit** term when the local bandit problems defined by $\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}$ is well estimated, and the **skeleton improvement** term that controls the probability that the local bandit problem is not well estimated. This second term is managed by controlling the number of policy improvement steps and using concentration properties of empirical quantities defined on the skeleton.

The main Theorem 6 follows from the following proposition that is proved in Appendix D.

Proposition 7. *For all state-action pair $x \in \mathcal{X}_{\mathbf{M}}$, for all $\varepsilon > 0$,*

$$N_x(t) \leqslant B_x(T) + S(T), \tag{15}$$

where we introduced the following terms

$$B_x(T) = \sum_{t=1}^{T} \mathbb{1}\left\{ x_t = x, \mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \right\} \qquad \textit{Bandit term} \tag{16}$$

$$S(T) = \sum_{t=1}^{T} \mathbb{1}\left\{ \overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon} \right\} \qquad \textit{Skeleton improvement} \tag{17}$$

Furthermore, for $0 < \varepsilon \leqslant \frac{1}{3} \min_{\pi \in \Pi(\mathbf{M})} \min_{(s,a) \in \mathcal{X}_{\mathbf{M}}} \{|\Delta_{s,a}(\mathbf{M}(\pi))| : |\Delta_{s,a}(\mathbf{M}(\pi))| > 0\}$, $\mathbb{E}(S(T)) = O(1)$, and for a critical state-action pair x,

$$\mathbb{E}(B_x(T)) \leqslant \frac{\Delta_x(\mathbf{M})}{\underline{\mathbf{K}}_x(\mathbf{M}) - \varepsilon \Gamma_s(\mathbf{M})} \log T + O(1)$$

while for a non-critical state-action pair, $\mathbb{E}(B_x(T)) = O(1)$.

5. Numerical experiments

In this section, we discuss numerical issues regarding `IMED-RL` and extend the discussion in Appendix F. Experiments in this section take less than an hour to run on a standard machine with 8 CPUs (*Intel(R) Core(TM) i7-10610U CPU @ 1.80GHz*) and 32 GB of RAM.

Computing `IMED-RL` index At each time step, we run the value iteration algorithm on $\widehat{\mathbf{M}}_t(\mathcal{A}(t))$ to compute the optimal bias and the associated potential function $\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}$. This task is standard. Once done, one must compute, for all $a \in \mathcal{A}_{s_t}$, the value of the optimization problem $\underline{\mathbf{K}}_{s_t,a}(t)$ which belongs to the category of convex optimization problem with linear constraint. Such problems have been studied under the name of *partially-finite convex optimization*, *e.g.* in Borwein and Lewis (1991). It is possible to compute $\underline{\mathbf{K}}_{s_t,a}(t)$ by considering the Legendre-Fenchel dual and one does not need to compute the optimal distribution to know the value of the optimization problem.

Proposition 8 (Index computation, Honda and Takemura (2015) Theorem 2)**.** *Let (s,a) be in $\mathcal{X}_{\mathbf{M}}$, $M = m_{max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\mathbf{M}}(s)$, and $\gamma > \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a))$, then*

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \begin{cases} \max\limits_{0 \leqslant x \leqslant \frac{M}{M-\gamma}} \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left[\log\left(M - \left(R + \mathbf{b}^{\mathbf{M}}(S) - \gamma\right)x\right)\right] & \text{if } M > \gamma \\ +\infty & \text{otherwise} \end{cases}. \tag{18}$$

If $\gamma \leqslant \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a))$, then $\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = 0$.

In particular, this Proposition 8 sometimes allows to write $\mathbf{K}_{s,a}(t)$ almost in close form, *e.g.* when $\mathcal{F}_{s,a}$ defined in Asumption 3 is a set of multinomials with unknown support (and only the upper bound m_{max} is known). In Appendix F, we discuss this numerical computation further. In particular, the complexity of computing the next action to take with `IMED-RL` is the one of performing value iteration on a restricted MDP followed by $\mathcal{A}_{s_t}$ convex optimization problems. After some time, the skeleton may only contain one (optimal) policy so that the value iteration step is reduced to a policy evaluation.

Environments In different environments, we illustrate in Figure 2 and Figure 3 the performance of `IMED-RL` 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). As stated during the introduction, any finite communicating MDP can be turned into an ergodic one, since on such MDPs, any stochastic policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}_s)$ with full support $\text{Supp}(\pi(s)) = \mathcal{A}_s$ is ergodic. Hence by mixing its transition $\mathbf{p}$ with that obtained from playing a uniform policy, formally $\mathbf{p}_\varepsilon(\cdot|s,a) = (1-\varepsilon)\mathbf{p}(\cdot|s,a) + \varepsilon \sum_{a' \in \mathcal{A}_s} \mathbf{p}(\cdot|s,a')/|\mathcal{A}_s|$, for an arbitrarily small $\varepsilon > 0$ one obtain an ergodic MDP. In the experiments, we consider ergodic versions of the classical n-state river-swim environment, 2-room and 4-room with $\varepsilon = 10^{-3}$, and classical communicating versions ($\varepsilon = 0$).

n-states *RiverSwim* environment As illustrated by Figure 2, the performances of `IMED-RL` are particularly good and the regret of `IMED-RL` is below the regrets of all its competitors, even when the MDP is communicating only. This numerical performance grounds numerically the previous theoretical analysis. While using `IMED-RL` in communicating MDPs is not endorsed by our theoretically analysis, it is interesting to see how much this hypothesis amounts in the numerical performances of `IMED-RL`. We therefore ran an experiment on another classical environment, 2-rooms.

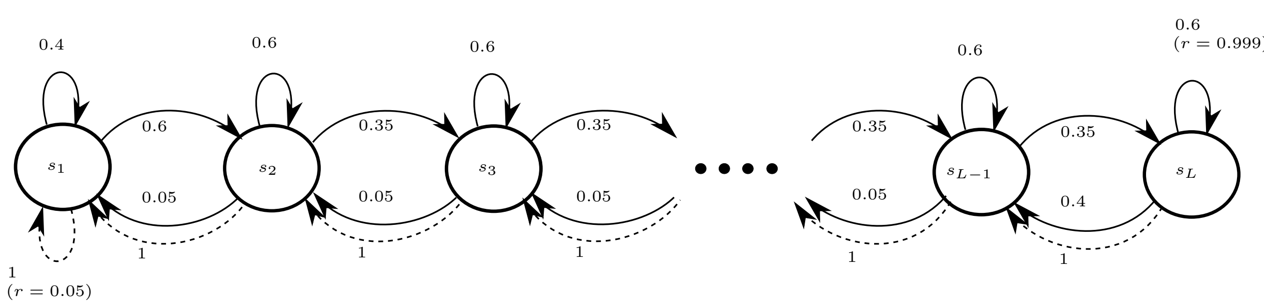

Figure 1: The ergodic n-state *RiverSwim* MDP. In each of the n states, there are two actions `RIGHT` and `LEFT`. The left action is represented with a dashed line and the `RIGHT` with plain line. Rewards are located at the extremities of the MDP.

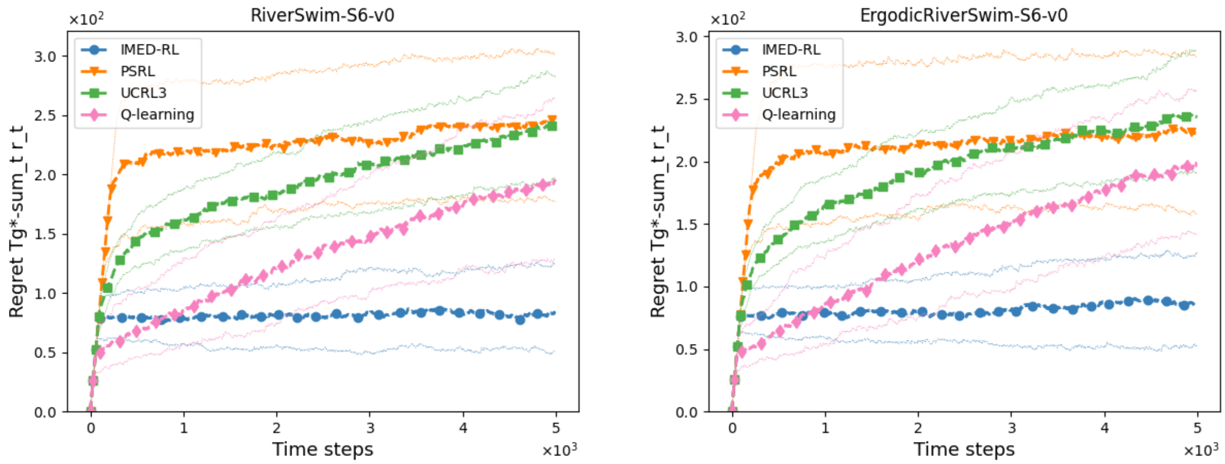

Figure 2: Average (256 independent runs) regret and quantiles (0.25 and 0.75) curves of algorithms on a standard communicating 6-states RiverSwim (left) and an ergodic 6-states RiverSwim (right).

n-rooms environment As illustrated by Figure 3, the performances of `IMED-RL` are particularly good, even surprisingly good, in this communicating only environment. Those experiments are a clue that the `IMED-RL` strategy may still be reasonable, although not necessarily optimal in some communicating MDPs.

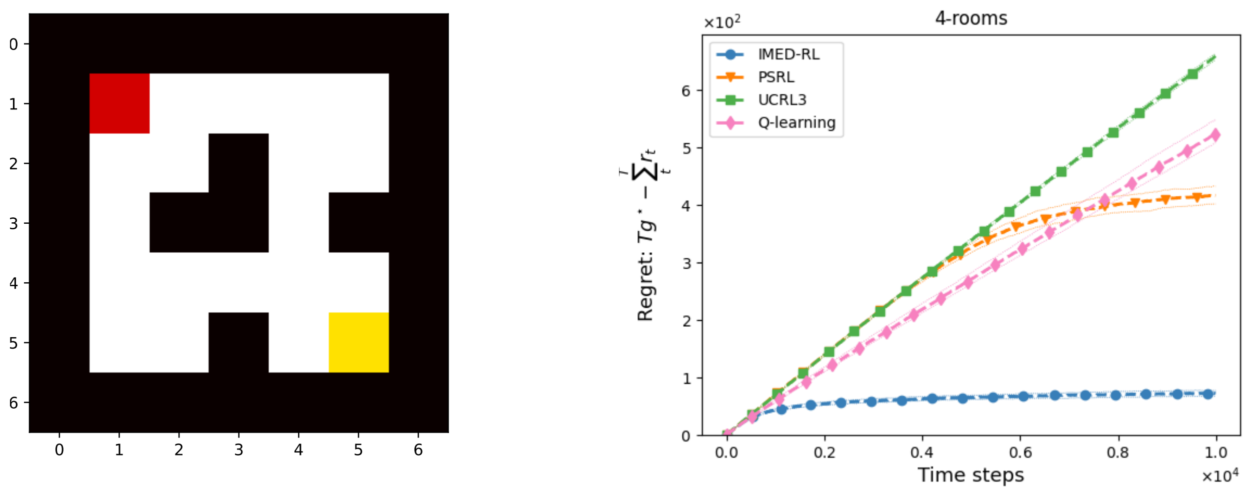

Figure 3: Average regret and quantiles (0.25 and 0.75) curves of algorithms (right) corresponding to learning on a 4-room (left) grid-world environment, with 20 states: the starting state is shown in red, and the rewarding state is shown in yellow. From the yellow state, all actions bring the learner to the red state. Other transitions are noisy as in a *frozen-lake* environment.

6. Conclusion

In this paper, we introduced `IMED-RL`, a numerically efficient algorithm to solve the average-reward criterion problem under the ergodic assumption for which we derive an upper bound on the regret matching the known regret lower bound. Further, its surprisingly good numerical performances in communicating only MDPs open the path to future work in MDPs that are communicating only.

References

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3), 1989.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47 (2-3):235–256, 2002.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *Proceedings of the 20th conference on advances in Neural Information Processing Systems*, NIPS '06, pages 49–56, Vancouver, British Columbia, Canada, dec 2006. MIT Press. ISBN 0-262-19568-2.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control (The Discrete Time Case)*. Academic Press, New York, 1978.

J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problem. *SIAM Journal on Computation and Optimization*, 29(2):325–338, 1991.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pages 1056–1066. PMLR, 2020.

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

Apostolos N. Burnetas and Michaël N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Xi-Ren Cao. *Reinforcement Learning*, pages 289–340. Springer US, Boston, MA, 2007. ISBN 978-0-387-69082-7. doi: 10.1007/978-0-387-69082-7_6. URL https://doi.org/10.1007/978-0-387-69082-7_6.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

Amir Dembo and Ofer Zeitouni. Large deviations techniques and applications. *Elearn*, 1998.

S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, US, 2010.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

Onésimo Hernández-Lerma and Jean-Bernard Lasserre. *Discrete-Time Markov Control Processes*. Springer New York, 1996. doi: 10.1007/978-1-4612-0729-0. URL https://hal.laas.fr/hal-02095866.

J. Honda and A. Takemura. Finite-time regret bound of a bandit algorithm for the semi-bounded support model. arXiv:1202.2277, 2012.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Machine Learning*, 16:3721–3756, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.

O-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Budapest, Hungary, 2011.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805, 2018.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Appendix A. Table of contents

In the following appendices, we recall and introduce notations (Appendix B) useful for understanding IMED-RL and the proof of the regret upper-bound given in Appendix D. Before that, we introduce some technical results in Appendix C that will help for the main proof and increase its readability. In Appendix E, we discuss the ergodic Assumption 2 made in the paper, followed in Appendix F in which we experiments more with IMED-RL and especially, beyond the ergodic assumption, towards the communication one. Finally, we detail in small note why our modification of the skeleton's notion introduced by Burnetas and Katehakis (1997) was key to the empirical success of our proposed IMED-RL.

Appendix B. Notations

Notations of exact quantities

$\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r}\}$, a MDP

$\mathcal{S}$, state space,

$\mathcal{A} = (\mathcal{A}_s)_{s\in\mathcal{S}}$, action space

$\mathcal{X}_{\mathbf{M}} = \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$, state-action space

$\mathbf{p} : \mathcal{X}_{\mathbf{M}} \to \mathcal{P}(\mathcal{S})$, transition distribution

$\mathbf{r} : \mathcal{X}_{\mathbf{M}} \to \mathcal{P}(\mathbb{R})$, reward distribution

$\mathbf{m}(s,a) = \mathbb{E}_{r\sim\mathbf{r}(s,a)}[r]$, mean reward function

$\Pi(\mathbf{M}) = \{\pi : s \in \mathcal{S} \mapsto \pi(s) \in \mathcal{A}_s\}$, stationary deterministic policy space

$\forall \pi \in \Pi(\mathbf{M})$, $\mathbf{p}_\pi : s \in \mathcal{S} \mapsto \mathbf{p}(\cdot|s,\pi(s)) = \mathbf{p}(s,\pi(s)) \in \mathcal{P}(\mathcal{S})$, $\mathbf{r}_\pi : s \in \mathcal{S} \mapsto \mathbf{r}(s,\pi(s)) \in \mathcal{P}(\mathbb{R})$, $\mathbf{m}_\pi : s \in \mathcal{S} \mapsto \mathbf{m}(s,\pi(s)) \in \mathbb{R}$, MDP related quantities defined by a policy π

$\mathbf{M}(\mathcal{D}) = (\mathcal{S}, \mathcal{D}, \mathbf{p}, \mathbf{r})$, MDP $\mathbf{M}$ with action space restricted to $\mathcal{D}$

$\mathbf{b}_\pi : s \in \mathcal{S} \mapsto \mathbf{b}_\pi(s) \in \mathbb{R}$, bias function as defined in Poisson Equation 3

$\mathbf{g}_\star = \mathbf{g}^{\mathbf{M}}$, optimal gain on MDP $\mathbf{M}$

$\mathbf{b}^{\mathbf{M}}$, optimal bias function as defined in the optimal Poisson Equation 4

$\Delta_{s,a}(\mathbf{M}) = \mathbf{m}(s,a) + \mathbf{p}(s,a)\mathbf{b}^{\mathbf{M}} - \mathbf{g}^{\mathbf{M}} - \mathbf{b}^{\mathbf{M}}(s)$, sub-optimality gap

$\gamma_s(\mathbf{M}) = \mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s)$, optimality threshold

$\varphi_{\mathbf{M}}(\nu \otimes q) = \mathbb{E}_{R\sim\nu}[R] + q\mathbf{b}^{\mathbf{M}}$, potential function

$\mathcal{O}_s(\mathbf{M}) = \{a \in \mathcal{A}_s : \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a) = \gamma_s(\mathbf{M}))\}$, optimal action in state s

$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \inf_{\substack{\nu\in\mathcal{F}_{s,a} \\ q\in\mathcal{P}(\mathcal{S})}} \{\mathrm{KL}(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a), \nu \otimes q) \ : \ \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma\}$, sub-optimality cost for threshold γ

$\underline{\mathbf{K}}_{s,a}(\mathbf{M}) = \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$, sub-optimality cost

$\mathcal{C}(\mathbf{M}) = \left\{(s,a) : 0 < \underline{\mathbf{K}}_{s,a}(\mathbf{M}) < \infty\right\}$, set of critical state-action pairs

$\mathcal{C}_{sa,\mathbf{M}}(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})) = \{\nu \otimes q \in \mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}) : \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma_s(\mathbf{M})\}$, set of critical distributions

$\mathcal{C}_{sa,\mathbf{M}}(\mathcal{F}_{s,a} \otimes \mathcal{P}(\mathcal{S}), \delta) = \{\nu \otimes q \in \mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}) : \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma_s(\mathbf{M}) - \varepsilon\}$

Notations of empirical quantities

$N_{s,a}(T) = \sum_{t=1}^{T} \mathbb{1}\{s_t = s, a_t = a\}$, counting random variable for state-action pair (s,a)

$N_s(T) = \sum_{t=1}^{T} \mathbb{1}\{s_t = s\}$, counting random variable for state s

$\mathcal{A}_s(t) = \left\{a \in \mathcal{A}_s : N_{s,a}(t) \geqslant \log^2\left(\max_{a' \in \mathcal{A}_s} N_{s,a'}(t)\right)\right\}$, skeleton at state s at time t (this is a random variable)

$\hat{\gamma}_s(t) \overset{\text{def}}{=} \max_{a \in \mathcal{A}_s} \varphi_{\hat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\hat{\mathbf{r}}(s,a) \otimes \hat{\mathbf{p}}(s,a)\right)$, empirical optimality threshold

$\mathbf{K}_{s,a}(t) \overset{\text{def}}{=} \underline{\mathbf{K}}_{s,a}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t)), \hat{\gamma}_s(t)\right)$, empirical sub-optimality cost

$\mathbf{H}_{s,a}(t) = N_{s,a}(t)\underline{\mathbf{K}}_{s,a}(t) + \log N_{s,a}(t)$, `IMED-RL` index

$\mathcal{R}_{s_1}(\mathbf{M}, \pi, T) = \sum_{x \in \mathcal{X}_{\mathbf{M}}} \mathbb{E}_{\pi, s_1}[N_x(T)]\, \Delta_x(\mathbf{M}) + \left(\left[\prod_{t=1}^{T} \mathbf{p}_{\pi_t} - \mathbf{p}_\star^t\right] b^{\mathbf{M}}\right)(s_1)$, regret of a policy π (with $\star \in \mathcal{O}(\mathbf{M})$, expression proved in Lemma 2)

Useful notations in the proofs

$\Pi = |\Pi(\mathbf{M})|$, cardinal of policy space

$\sigma_t^{\Pi+1} = t+1$, $\sigma_t^{\nu} = \sigma_t^{\nu+1} - \lfloor \frac{t}{\Pi+1} \rfloor$ for all $\nu \in \{1, \cdots, \Pi\}$ and $\sigma_t^0 = 1$, boarders of a $\mathbf{M}$-adapted sub-division (see Definition 11)

$I_t^{\nu} = \left\{k : \sigma_t^{\nu} \leqslant k < \sigma_t^{\nu+1}\right\}$, sub-interval ν

$N_s^{\nu}(t) = \sum_{k \in I_t^{\nu}} \mathbb{1}\{s_t = s\}$, number of visits state at s during sub-interval I_t^{ν}

$0 < \kappa < \frac{1}{\Pi+1}$

$0 < \beta < \beta_{\mathbf{M}}$, where $\beta_{\mathbf{M}}$ is defined in Proposition 10

$V_t^{\nu} = \bigcap_{s \in \mathcal{S}} \{N_s^{\nu}(t) \geqslant \kappa\beta t\}$, an event controlling the number of visit in sub-interval ν

$V_t = \bigcap_{\nu=0}^{\Pi} V_t^{\nu}$, an event controlling the number of visits in each sub-interval

$S_t^{\nu}(\delta) = \bigcap_{k \in I_t^{\nu}} \bigcap_{x \in \mathcal{X}_{\mathbf{M}(\mathcal{A}(k))}} \left\{\|\hat{\mathbf{p}}^k(x) - \mathbf{p}(x)\|_\infty \leqslant \delta, |\hat{\mathbf{m}}^k(x) - \mathbf{m}(x)| \leqslant \delta\right\}$, an event controlling the precision on the empirically restricted MDP during interval ν

$S_t(\delta) = \bigcap_{\nu=1}^{\Pi} S_t^{\nu}$, an event controlling the precision in each sub-interval but the first

$I(s, \mathbf{M}, \gamma) = \{a \in \mathcal{A}_s : \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)) > \gamma\}$

$x^*_{sa,\mathbf{M}} = \arg \max_{0 \leqslant x \leqslant \frac{M}{M - \gamma_s(\mathbf{M})}} \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left[\log\left(B - \left(R + \mathbf{b}^{\mathbf{M}}(S) - \gamma_s(\mathbf{M})\right)x\right)\right]$,
where $B = m_{max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\mathbf{M}}(s)$

$\lambda_{sa,\mathbf{M},\varepsilon} = \sup\left\{\lambda \in \mathbb{R} : \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left(\frac{B - R - \mathbf{b}^{\mathbf{M}}(S)}{B - \mu_{s,a} + \varepsilon}\right)^{\lambda} \geqslant 1\right\}$, where $B = m_{max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\mathbf{M}}(s)$ and $\mu_{s,a} = \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a))$

$$\Lambda^*_{sa,\mathbf{M}}(x) = \sup_\lambda \left\{ \lambda(\mu_{s,a} - x) - \log \mathbb{E}_{\substack{R\sim\mathbf{r}(s,a)\\ S\sim\mathbf{p}(\cdot|s,a)}} \left[\exp\left(\lambda(R + \mathbf{b}^{\mathbf{M}}(S))\right)\right] \right\},$$

where $\mu_{s,a} = \varphi_{\mathbf{M}}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)\right)$

$$\tilde{\Lambda}^*_{sa,\mathbf{M}}(x) = \sup_\lambda \left\{ \lambda x - \log \mathbb{E}_{\substack{R\sim\mathbf{r}(s,a)\\ S\sim\mathbf{p}(\cdot|s,a)}} \left[\left(1 - \left(R + \mathbf{b}^{\mathbf{M}}(S) - \gamma_s\left(\mathbf{M}\right)\right) x^*_{s,a}\right)^\lambda \right] \right\}$$

$$\underline{\mathcal{K}}^{\varepsilon}_{s,a}(t) \stackrel{\text{def}}{=} \inf_{\substack{\nu\in\mathcal{F}_{s,a}\\ q\in\mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\hat{\mathbf{r}}_t(s,a) \otimes \hat{\mathbf{p}}_t(s,a), \nu \otimes q\right) \;:\; \varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - \varepsilon \right\}$$

$$\underline{\mathcal{K}}^{\varepsilon,n}_{s,a} \stackrel{\text{def}}{=} \inf_{\substack{\nu\in\mathcal{F}_{s,a}\\ q\in\mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\hat{\mathbf{r}}(s,a,n) \otimes \hat{\mathbf{p}}_t(s,a,n), \nu \otimes q\right) \;:\; \varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - \varepsilon \right\}$$

Appendix C. Technical results

In this section we state a few lemmas and define a few objects that are useful for the regret analysis of IMED-RL. More precisely,

Lemma 2 is about expressing the regret with respect to the number of pulls $N_x(t)$ of sub-optimal state-action pairs $x \in \mathcal{X}_{\mathbf{M}}$ and show that $\Delta_x(\mathbf{M})$ indeed is roughly the cost an agent suffer each time it plays it. Thanks to this Lemma, controlling the regret is equivalent to controlling $N_x(T)$ for each sub-optimal state-action pair $x \in \mathcal{X}_{\mathbf{M}}$, which we do in Appendix D.

Definition 9 introduces notations used for expressing concentration results of events studied in the main proof (see Appendix D). Those are inspired from Honda and Takemura (2015) and lies on the assumptions made in this paper.

Lemma 10 states that under the ergodic Assumption 2, any learning agent is guaranteed to visit every states sufficiently often with a large enough probability. This lemma is proved in Proposition 2 of Burnetas and Katehakis (1997).

Definition 11 introduces the notion of $\mathbf{M}$-adapted sub-division that split interval $[0, t]$ into $\Pi + 1$ sub-intervals. This notion follows from the developed intuition about Policy Improvement and we will prove that with high probability, a policy improvement occurs (if possible) in-between each sub-interval.

Lemma 12 uses the previous lemma to derive a stronger result by proving that a linear number of visits in each state and in each linear sub-interval of a $\mathbf{M}$-adapted subdivision can be obtained with high probability. It is useful to derive improvement of the skeleton between the beginning and end of a sub-interval.

Lemma 13 states that, by definition of the skeleton, by Lemma 12, and by the reward Assumptions 1, 3, empirical quantities defined on the MDP restricted by the skeleton are well approximated.

Lemma 14 expresses how well the gain $\mathbf{g}_\pi$ and bias $\mathbf{b}_\pi$ of every policy $\pi \in \Pi(\mathbf{M})$ can be estimated given a precision on the mean rewards and transitions of the original MDP.

Lemma 15 expresses how well the optimal gain $\mathbf{g}^{\mathbf{M}}$ and optimal bias $\mathbf{b}^{\mathbf{M}}$ defined by the optimal Poisson Equation 4 can be estimated given a precision on the mean rewards and transitions of the original MDP.

Lemma 16 expresses how well the optimal gain $\mathbf{g}^{\mathbf{M}(\mathcal{A}(t))}$ and optimal bias $\mathbf{b}^{\mathbf{M}(\mathcal{A}(t))}$ defined by the optimal Poisson Equation 4 can be estimated given a precision on the mean rewards and transitions of the MDP $\mathbf{M}\left(\mathcal{A}(t)\right)$ prescribed by a good event.

Lemma 2 (Regret decomposition)**.** *Under the ergodic Assumption 2, for all optimal policy $\star \in \mathcal{O}(\mathbf{M})$, the regret of any policy $\pi = (\pi_t)_t$ can be decomposed as*

$$\mathcal{R}_{s_1}(\mathbf{M}, \pi, T) = \sum_{x \in \mathcal{X}_{\mathbf{M}}} \mathbb{E}_{\pi, s_1}\left[N_x(T)\right] \Delta_x(\mathbf{M}) + \underbrace{\left(\left[\prod_{t=1}^{T} \mathbf{p}_{\pi_t} - \mathbf{p}_\star^t\right] b_\star\right)(s_1)}_{\leqslant \mathbb{S}(\mathbf{b}_\star)} . \tag{2}$$

Proof of Lemma 2. It holds by the Poisson equation that $\mathbf{m}_\pi = \mathbf{g}_\pi + (I - \mathbf{p}_\pi)\mathbf{b}_\pi$. Hence, the cumulative reward of a strategy playing policy π_t at time t until time T and starting from state s_1 is given by

$$\begin{aligned} V_{s_1}(\mathbf{M}, \pi, T) &= \sum_{t=1}^{T} \Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} \mathbf{m}_{\pi_t}\Big)(s_1) \\ &= \sum_{t=1}^{T} \Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} (\mathbf{g}_{\pi_t} + (I - \mathbf{p}_{\pi_t})\mathbf{b}_{\pi_t})\Big)(s_1) . \end{aligned}$$

Under the ergodic Assumption 2, for all optimal policy $\star$, $\mathbf{g}_\star(s_1)$ takes the same value $g_\star$ for all state s_1. In this case, $(\mathbf{p}_\pi \mathbf{g}_\star)(s_1) = \mathbf{g}_\star$ for all π, s_1. Using this property it comes

$$\begin{aligned}
\mathcal{R}_{s_1}(\mathbf{M}, \pi, T) &= V_{s_1}(\mathbf{M}, \star, T) - V_{s_1}(\mathbf{M}, \pi, T) \\
&= \sum_{t=1}^{T} \Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} \underbrace{[(g_\star - \mathbf{g}_{\pi_t}) + (I - \mathbf{p}_{\pi_t})(\mathbf{b}_\star - \mathbf{b}_{\pi_t})]}_{\boldsymbol{\Delta}\pi_t} \Big)(s_1) \\
&\quad + \sum_{t=1}^{T} \Big([\mathbf{p}_\star^{t-1} - \mathbf{p}_\star^{t} - \prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} - \prod_{t'=1}^{t} \mathbf{p}_{\pi_{t'}}] \mathbf{b}_\star \Big)(s_1) \\
&= \Big(\sum_{t=1}^{T} \Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} \Big) \boldsymbol{\Delta}_{\pi_t} \Big)(s_1) + \Big(\Big[\prod_{t'=1}^{T} \mathbf{p}_{\pi_{t'}} - \mathbf{p}_\star^{T} \Big] \mathbf{b}_\star \Big)(s_1)
\end{aligned}$$

At this point, we note that

$$\begin{aligned}
\boldsymbol{\Delta}_\pi(s) &= g_\star - \mathbf{g}_\pi(s) + (I - \mathbf{p}_\pi)(\mathbf{b}_\star - \mathbf{b}_\pi)(s) = \mathbf{m}_\star(s) - \mathbf{m}_\pi(s) + ([\mathbf{p}_\star - \mathbf{p}_\pi]\mathbf{b}_\star)(s) \\
&= \sum_{a \in \mathcal{A}_s} [\mathbf{m}_\star(s) - \mathbf{m}(s,a) + \big((\mathbf{p}_\star - \mathbf{p}_a)\mathbf{b}_\star\big)(s)] \pi(a|s) = \mathbb{E}_\pi[\Delta(s,a)] \,.
\end{aligned}$$

To conclude, we note that

$$\begin{aligned}
\sum_{t=1}^{T} [\mathbf{p}_{\pi_1} \mathbf{p}_{\pi_2} \dots \mathbf{p}_{\pi_{t-1}} \boldsymbol{\Delta}_{\pi_t}](s_1) &= \sum_{s,a} \sum_{t=1}^{T} \mathbb{E}_{\pi_1, \dots \pi_t}[\Delta(s,a) \mathbb{1}\{S_t = s, A_t = a\}] \\
&= \sum_{x} \Delta_x(\mathbf{M}) \mathbb{E}[N_x(T)] \,.
\end{aligned}$$

■

Definition 9. *Let $\mathbf{M}$ be an MDP satisfying Assumption 2 and whose reward distribution $\mathbf{r}$ satisfy Assumptions 1 and 3, then the following quantities[8] are well defined,*

$$\begin{aligned}
x^*_{sa,\mathbf{M}} &= \arg \max_{0 \leqslant x \leqslant \frac{M}{M - \gamma_s(\mathbf{M})}} \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left[\log \left(B - \left(R + \mathbf{b}^{\mathbf{M}}(S) - \gamma_s(\mathbf{M}) \right) x \right) \right], \\
\lambda_{sa,\mathbf{M},\varepsilon} &= \sup \left\{ \lambda \in \mathbb{R} : \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left(\frac{B - R - \mathbf{b}^{\mathbf{M}}(S)}{B - \mu_{s,a} + \varepsilon} \right)^{\lambda} \geqslant 1 \right\}, \\
\Lambda^*_{sa,\mathbf{M}}(x) &= \sup_{\lambda} \left\{ \lambda(\mu_{s,a} - x) - \log \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left[\exp\left(\lambda (R + \mathbf{b}^{\mathbf{M}}(S)) \right) \right] \right\}, \\
\tilde{\Lambda}^*_{sa,\mathbf{M}}(x) &= \sup_{\lambda} \left\{ \lambda x - \log \mathbb{E}_{\substack{R \sim \mathbf{r}(s,a) \\ S \sim \mathbf{p}(\cdot|s,a)}} \left[\left(1 - \left(R + \mathbf{b}^{\mathbf{M}}(S) - \gamma_s(\mathbf{M}) \right) x^*_{s,a} \right)^{\lambda} \right] \right\},
\end{aligned}$$

where $B = m_{max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\mathbf{M}}(s)$ and $\mu_{s,a} = \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a))$. By Sections 6 and 7 of Honda and Takemura (2015), $x^*_{sa,\mathbf{M}}$ exists uniquely when $\varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)) \leqslant \gamma_s(\mathbf{M})$ and $\lambda_{sa,\varepsilon} > 1$ for $\varepsilon > 0$. In Section D.1, we drop the explicit mention to $\mathbf{M}$ as we are referring to the original MDP and because it makes the equations easier to read.

Lemma 10 (Proposition 2 of Burnetas and Katehakis (1997))**.** *Let $\mathbf{M}$ be an MDP satisfying the ergodic Assumption 2. There exists $B_{\mathbf{M}} > 0$ and $\beta_{\mathbf{M}}$ such that for all all $\beta > 0$, for all $s \in \mathcal{S}$, $t > |\mathcal{S}|$ and policy $\pi = (\pi_k)_{0 \leqslant k \leqslant t}$,*

$$\mathbb{P}_{\pi,\mathbf{M},s_0}\left(N_s(t) \leqslant \beta t\right) \leqslant B_{\mathbf{M}} \exp\left(-\left(\beta_{\mathbf{M}} - \beta\right) t\right) . \tag{19}$$

In particular, for $\beta < \beta_{\mathbf{M}}$, the probability that any given state has been visited less than a linear amount of time is exponentially small. Such a $\beta_{\mathbf{M}}$ that satisfies this proposition controls the rate at which all states are visited.

8. see Equations 4, 5, and 6 of Honda and Takemura (2015)

Definition 11 ($\mathbf{M}$-adapted sub-division)**.** *Let $\mathbf{M}$ be an MDP and denote by $\Pi = |\Pi(\mathbf{M})|$ the number of deterministic policies $\Pi(\mathbf{M})$ on $\mathbf{M}$. Let $t > \Pi + 2$, and $I_t = \{1, \cdots, t\}$ the discrete time steps from 1 to t. For all $\nu \in \{0, \cdots, \Pi\}$, let*

$$I_t^\nu = \left\{k : \sigma_t^\nu \leqslant k < \sigma_t^{\nu+1}\right\}, \tag{20}$$

with $\sigma_t^{\Pi+1} = t+1$, $\sigma_t^\nu = \sigma_t^{\nu+1} - \lfloor \frac{t}{\Pi+1} \rfloor$ for all $\nu \in \{1, \cdots, \Pi\}$ and let $\sigma_t^0 = 1$. The sub-division $\bigcup_\nu I_t^\nu$ of I_t induced by $(\sigma_t^\nu)_\nu$ is called an $\mathbf{M}$-adapted sub-division at time t.

It follows immediately from the definition that

$$\sigma_t^\nu = (t+1) - (\Pi + 1 - \nu)\left\lfloor \frac{t}{\Pi+1} \right\rfloor \qquad \forall \nu \in \{1, \cdots, \Pi+1\}, \tag{21}$$

$$\sigma_t^0 = 1, \tag{22}$$

and

$$|I_t^\nu| = \left\lfloor \frac{t}{\Pi+1} \right\rfloor \qquad \forall \nu \in \{1, \cdots, \Pi+1\}, \tag{23}$$

$$|I_t^0| = t - \Pi \left\lfloor \frac{t}{\Pi+1} \right\rfloor \geqslant \left\lfloor \frac{t}{\Pi+1} \right\rfloor. \tag{24}$$

Lemma 12 (Linear visit in each interval of a $\mathbf{M}$-adapted sub-division)**.** *Let $\mathbf{M}$ be an MDP and denote by $\Pi = |\Pi(\mathbf{M})|$ the number of deterministic policies $\Pi(\mathbf{M})$ on $\mathbf{M}$. Let $t > \Pi + 2$, and $(I_t^\nu, \sigma_t^\nu)_\nu$ be an $\mathbf{M}$-adapted sub-division of $\mathbf{M}$ at time t, i.e. a sub-division of $I_t = \{1, \cdots, t\}$. Let $\pi = (\pi_k)_k$ be a policy.*

Let $N_s^\nu(t) = \sum_{k \in I_t^\nu} \mathbb{1}\{s_t = s\}$ be the number of time the agent visit state s during the sub-interval I_t^ν. Let κ be such that $0 < \kappa < \frac{1}{\Pi+1}$ and let $0 < \beta < \beta_{\mathbf{M}}$ with $\beta_{\mathbf{M}}$ as in Lemma 10.

Let $V_t^\nu = \bigcap_{s \in \mathcal{S}} \{N_s^\nu(t) \geqslant \kappa\beta t\}$ the event that all states are visited more than $\kappa\beta t$ times during interval ν. Finally, denote $V_t = \bigcap_{\nu=0}^{\Pi} V_t^\nu$ the event that all states are visited more than $\kappa\beta t$ times during each sub-interval of the sub-division.

Then,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\overline{V_t}\right\}\right] \leqslant (\Pi + 2) + (\Pi + 1)|\mathcal{S}| \frac{B_{\mathbf{M}}}{1 - \exp\left(-\left(\beta_{\mathbf{M}} - \beta\right)\right)}. \tag{25}$$

Proof. By applying Proposition 10 and using a union bound on all states and sub-intervals, the Lemma 12 follows. ∎

Lemma 13 (Uniform concentration on the skeleton)**.** *Let $\mathbf{M}$ be an MDP and denote by $\Pi = |\Pi(\mathbf{M})|$ the number of deterministic policies $\Pi(\mathbf{M})$ on $\mathbf{M}$. Let $t > \Pi + 2$, and $(I_t^\nu, \sigma_t^\nu)_\nu$ be an $\mathbf{M}$-adapted sub-division of $\mathbf{M}$ at time t, i.e. a sub-division of $I_t = \{1, \cdots, t\}$. Let κ, β, V_t^ν and V_t be as in Lemma 12.*

Let $\delta > 0$ be a positive number representing a precision on the skeleton, let

$$S_t^\nu(\delta) = \bigcap_{k \in I_t^\nu} \bigcap_{x \in \mathcal{X}_{\mathbf{M}(\mathcal{A}(k))}} \left\{\|\hat{\mathbf{p}}_k(x) - \mathbf{p}(x)\|_\infty \leqslant \delta, |\hat{\mathbf{m}}_k(x) - \mathbf{m}(x)| \leqslant \delta\right\}, \tag{26}$$

be the event of uniform δ-good approximation on the skeleton for sub-interval ν and let

$$\begin{aligned} S_t(\delta) &= \bigcap_{\nu=1}^{\Pi} S_t^\nu \\ &= \bigcap_{k \geqslant \sigma_t^1} \bigcap_{x \in \mathcal{X}_{\mathbf{M}(\mathcal{A}(k))}} \left\{\|\hat{\mathbf{p}}_k(x) - \mathbf{p}(x)\|_\infty \leqslant \delta, |\hat{\mathbf{m}}_k(x) - \mathbf{m}(x)| \leqslant \delta\right\} \end{aligned} \tag{27}$$

be the event of uniform δ-good approximation on the skeleton for all time steps after the first sub-interval.

Then, for all policy $\pi = (\pi_k)_k$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{V_t \cap \overline{S_t\left(\delta\right)}\right\}\right] \leqslant S_{\mathbf{M}}\left(\delta\right) \tag{28}$$

where $S_{\mathbf{M}}\left(\delta\right)$ is decomposed as

$$S_{\mathbf{M}}\left(\delta\right) = \sum_{(s,a)\in\mathcal{X}_{\mathbf{M}}} S_{sa,\mathbf{M}}(\delta), \tag{29}$$

with

$$S_{sa,\mathbf{M}}(\delta) = \sum_{t=1}^{T} \frac{2t}{\left(1-\exp\left(-\Lambda^*_{s,a}\left(\delta\right)\right)\right)} \exp\left(-\Lambda^*_{s,a}\left(\delta\right)\log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right)\right).$$

For all state-action pair (s,a), $S_{sa,\mathbf{M}}$ is expressed as the limit of a convergent series.

Proof. First, we remark that $V_t \cap \overline{S_t\left(\delta\right)} = \bigcup_{\nu=1}^{\Pi} V_t \cap \overline{S^{\nu}_t\left(\delta\right)}$ so that, by a simple union bound, we only need to control the probability of the event $V_t \cap \overline{S^{\nu}_t\left(\delta\right)}$. We then remark that, since for all state s and for all action $a \in \mathcal{A}_s(k)$, $N_{s,a}(k) \geqslant \log^2 \max_{a'\in\mathcal{A}_s} N_{s,a'}(k)$ and $\max N_{s,a'}(k) \geqslant N_s(k)/|\mathcal{A}_s|$, we have $N_{s,a}(k) \geqslant \log^2\left(N_s(k)/|\mathcal{A}_s|\right)$. Combining with V_t, for all $k \in I^{\nu}_t$, $(s,a) \in \mathcal{X}_{\mathbf{M}(\mathcal{A}(k))}$, the number of samples of (s,a) is lower bounded by

$$N_{s,a}(k) \geqslant \log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right).$$

The event $V_t \cap \overline{S^{\nu}_t\left(\delta\right)}$ therefore satisfies the following inclusion,

$$V_t \cap \overline{S^{\nu}_t\left(\delta\right)} \subseteq \bigcup_{(s,a)\in\mathcal{X}_{\mathbf{M}}} \bigcup_{k\in I^{\nu}_t} \left\{ \begin{array}{c} \max\left(\|\hat{\mathbf{p}}_k(s,a) - \mathbf{p}(s,a)\|_\infty, |\hat{\mathbf{m}}_k(s,a) - \mathbf{m}(s,a)|\right) > \delta \\ N_{s,a}(k) \geqslant \log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right) \end{array} \right\}.$$

Again, by a union bound on state-action pairs, we are interested in controlling the event,

$$\bigcup_{k\in I^{\nu}_t} \left\{ \begin{array}{c} \max\left(\|\hat{\mathbf{p}}(s,a,N_{s,a}(k)) - \mathbf{p}(s,a)\|_\infty, |\hat{\mathbf{m}}(s,a,N_{s,a}(k)) - \mathbf{m}(s,a)|\right) > \delta \\ N_{s,a}(k) \geqslant \log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right) \end{array} \right\},$$

that is to say, the probability of

$$\bigcup_{k\in I^{\nu}_t} \bigcup_{n=\log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right)}^{\sigma^{\nu+1}_t} \left\{\max\left(\|\hat{\mathbf{p}}(s,a,n) - \mathbf{p}(s,a)\|_\infty, |\hat{\mathbf{m}}(s,a,n) - \mathbf{m}(s,a)|\right) > \delta\right\}. \tag{30}$$

Using the light-tail Assumption 1 and the fact that $\mathbf{p}(s,a)$ is a multinomial (hence light-tailed too), we get from Equations (2.2.12) and (2.2.13) of Dembo and Zeitouni (1998),

$$\mathbb{P}\left(\max\left(\|\hat{\mathbf{p}}(s,a,n) - \mathbf{p}(s,a)\|_\infty, |\hat{\mathbf{m}}(s,a,n) - \mathbf{m}(s,a)|\right) > \delta\right) \leqslant 2\exp\left(-n\Lambda^*_{s,a}\left(\delta\right)\right) \tag{31}$$

from which we deduce that the probability of Equation 30 is upper bounded by

$$\frac{2t}{(\Pi+1)\left(1-\exp\left(-\Lambda^*_{s,a}\left(\delta\right)\right)\right)} \exp\left(-\Lambda^*_{s,a}\left(\delta\right)\log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right)\right), \tag{32}$$

which is the term of a convergent series (in t). Denoting $S_{sa,\mathbf{M}}$ the limit of the series

$$S_{sa,\mathbf{M}}(\delta) = \sum_{t=1}^{T} \frac{2t}{\left(1-\exp\left(-\Lambda^*_{s,a}\left(\delta\right)\right)\right)} \exp\left(-\Lambda^*_{s,a}\left(\delta\right)\log^2\left(\frac{\nu\kappa\beta t}{|\mathcal{A}_s|}\right)\right), \tag{33}$$

and then combining all the union bound, we deduce that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{V_t \cap \overline{S_t\left(\delta\right)}\right\}\right] \leqslant \sum_{(s,a)\in\mathcal{X}_{\mathbf{M}}} S_{sa,\mathbf{M}}(\delta). \tag{34}$$

■

Lemma 14 (Sensibility of the Poisson equation, Lemma 7 (i) Burnetas and Katehakis (1997))**.** *Let $\varepsilon > 0$ a real positive number. Let $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ be an MDP and $\mathbf{M}_\delta = (\mathcal{S}, \mathcal{A}, \mathbf{p}_\delta, \mathbf{r}_\delta)$ be another MDP such that for all $x \in \mathcal{X}_{\mathbf{M}} = \mathcal{X}_{\mathbf{M}_\delta}$, $\max\left(\|\mathbf{p}(x) - \mathbf{p}_\delta(x)\|, |\mathbf{m}(x) - \mathbf{m}_\delta(x)|\right) \leqslant \delta$. There exits $\delta_{\mathbf{M}}(\varepsilon)$ such that for all $\delta < \delta_{\mathbf{M}}(\varepsilon)$, for all policy $\pi \in \Pi(\mathbf{M})$,*

$$|\mathbf{g}_\pi^{\mathbf{M}} - \mathbf{g}_\pi^{\mathbf{M}_\delta}| \leqslant \frac{\varepsilon}{2}, \tag{35}$$

$$\|\mathbf{b}_\pi^{\mathbf{M}} - \mathbf{b}_\pi^{\mathbf{M}_\delta}\| \leqslant \frac{\varepsilon}{2}. \tag{36}$$

The proof of this fact is given by Burnetas and Katehakis (1997) (Lemma 7 (i)) and a more modern proof to this result is given in Section 1, Chapter 4 of the book Cao (2007) under the name of perturbation analysis.

Lemma 15 (Sensibility of the optimal Poisson equation, Lemma 8 (ii) Burnetas and Katehakis (1997))**.** *Let $\varepsilon > 0$ a real positive number. Let $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ be an MDP and $\mathbf{M}_\delta = (\mathcal{S}, \mathcal{A}, \mathbf{p}_\delta, \mathbf{r}_\delta)$ be another MDP such that for all $x \in \mathcal{X}_{\mathbf{M}} = \mathcal{X}_{\mathbf{M}_\delta}$, $\max\left(\|\mathbf{p}(x) - \mathbf{p}_\delta(x)\|, |\mathbf{m}(x) - \mathbf{m}_\delta(x)|\right) \leqslant \delta$. There exits $\delta_{\mathbf{M}}(\varepsilon)$ such that for all $\delta < \delta_{\mathbf{M}}(\varepsilon)$,*

$$|\mathbf{g}^{\mathbf{M}} - \mathbf{g}^{\mathbf{M}_\delta}| \leqslant \frac{\varepsilon}{2}, \tag{37}$$

$$\|\mathbf{b}^{\mathbf{M}} - \mathbf{b}^{\mathbf{M}_\delta}\| \leqslant \frac{\varepsilon}{2}. \tag{38}$$

It follows immediately that forall $\nu \otimes q \in \mathcal{P}(\mathbb{R}) \otimes \mathcal{P}(\mathcal{S}))$ (with ν having an expected value), $|\varphi_{\mathbf{M}}(\nu \otimes q) - \varphi_{\mathbf{M}_\delta}(\nu \otimes q)| \leqslant \varepsilon$.

Furthermore, for ε such that $0 < \varepsilon < \varepsilon_{\mathbf{M}}$, $\mathcal{O}(\mathbf{M}_\delta) \subseteq \mathcal{O}(\mathbf{M})$, where

$$\varepsilon_{\mathbf{M}} = \frac{1}{3} \min\left\{|\mathbf{g}_\pi^{\mathbf{M}} - \mathbf{g}_{\pi'}^{\mathbf{M}}| : \pi, \pi' \in \Pi(\mathbf{M}), \mathbf{g}_\pi^{\mathbf{M}} \neq \mathbf{g}_{\pi'}^{\mathbf{M}}\right\}.$$

Proof. The first part is proved in Lemma 8 (ii) of Burnetas and Katehakis (1997) and we prove the last claim for the sake of introducing $\varepsilon_{\mathbf{M}}$.

Let $\varepsilon_{\mathbf{M}} = \frac{1}{3} \min\left\{|\mathbf{g}_\pi^{\mathbf{M}} - \mathbf{g}_{\pi'}^{\mathbf{M}}| : \pi, \pi' \in \Pi(\mathbf{M}), \mathbf{g}_\pi^{\mathbf{M}} \neq \mathbf{g}_{\pi'}^{\mathbf{M}}\right\}$ and ε be such that $0 < \varepsilon < \varepsilon_{\mathbf{M}}$. Let $\delta < \delta_{\mathbf{M}}(\varepsilon)$ where $\delta_{\mathbf{M}}$ is defined in Lemma 14. Let $\star \in \mathcal{O}(\mathbf{M}_\delta)$, then $\star$ also is optimal in MDP $\mathbf{M}$. Indeed, for all $\pi' \in \Pi(\mathbf{M})$,

$$\begin{aligned} \mathbf{g}_\star^{\mathbf{M}} &\geqslant \mathbf{g}_\star^{\mathbf{M}_\delta} - \frac{\varepsilon}{2} \\ &\geqslant \mathbf{g}_{\pi'}^{\mathbf{M}_\delta} - \frac{\varepsilon}{2} \\ &\geqslant \mathbf{g}_{\pi'}^{\mathbf{M}} - \varepsilon \\ &> \mathbf{g}_{\pi'}^{\mathbf{M}} - 2\varepsilon \end{aligned}$$

which implies that $\mathbf{g}_\star^{\mathbf{M}} \geqslant \mathbf{g}_{\pi'}^{\mathbf{M}}$ by definition of $\varepsilon < \varepsilon_{\mathbf{M}}$ which separates policies by at least $3\varepsilon_{\mathbf{M}}$. Therefore, $\mathcal{O}(\mathbf{M}_\delta) \subseteq \mathcal{O}(\mathbf{M})$. ■

Lemma 16. *Let σ_t^ν, and $S_t(\delta)$ be defined as in Lemma 13, let $\delta < \delta_{\mathbf{M}}(\varepsilon)$ with $\delta_{\mathbf{M}}$ and $\varepsilon < \varepsilon_{\mathbf{M}}$ defined in Lemma 15, then, under $S_t(\delta)$, for all $k \geqslant \sigma_t^1$,*

$$\mathcal{O}\left(\widehat{\mathbf{M}}_k(\mathcal{A}(k))\right) \subseteq \mathbf{M}(\mathcal{A}(k)), \tag{39}$$

$$\|\mathbf{b}^{\widehat{\mathbf{M}}_k(\mathcal{A}(k))} - \mathbf{b}^{\mathbf{M}(\mathcal{A}(k))}\| \leqslant \frac{\varepsilon}{2}. \tag{40}$$

Proof. This lemma is a direct consequence of Lemma 15 and definition of $S_t(\delta)$ in Lemma 13. ■

Appendix D. Proof of Theorem (REGRET)

In this appendix, we prove our main Theorem 6 that asses the optimality of the `IMED-RL` algorithm.

Theorem 6 (Regret upper bound for Ergodic MDPs)**.** *Let* $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ *be an MDP satisfying Assumptions 1, 2, 3. Let* $0 < \varepsilon \leqslant \frac{1}{3} \min_{\pi \in \Pi(\mathbf{M})(\mathcal{X}_\mathbf{M})} \min_{(s,a) \in \mathcal{X}_\mathbf{M}} \{|\Delta_{s,a}(\mathbf{M}_\pi)| : |\Delta_{s,a}(\mathbf{M}_\pi)| > 0\}$. *The regret of* `IMED-RL` *is upper bounded as*

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M}) - \varepsilon \Gamma_s(\mathbf{M})} \right) \log T + O(1), \tag{13}$$

where $\Gamma_s(\mathbf{M})$ *is constant that depends on the MDP* $\mathbf{M}$ *and state* s*; it is made explicit in the proof below. A Taylor expansion allows to write the regret upper bound as*

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})} \right) \log T + O\left((\log T)^{10/11} \right). \tag{14}$$

Were the semi-bounded reward assumption changed to a bounded reward one with known upper and lower bound, and the $O\left((\log T)^{10/11}\right)$ *could be made a* $O(1)$ *as explained in Appendix E.*

Outline The proof combines the concentration results obtained by Honda and Takemura (2015) for the family of rewards we study and the skeleton improvement idea from Burnetas and Katehakis (1997). Because we define a new notion of skeleton and that `IMED-RL` does not require forced exploration, we specifically derive the Lemma 25 that is at the heart of the proof that $\varphi_{\mathbf{M}(\mathcal{A}(t))}$ converges fast enough to $\varphi_\mathbf{M}$, thus allowing to optimally leverage the `IMED` algorithm in the MDP setting.

Proposition 17. *For all state-action pair* $x \in \mathcal{X}_\mathbf{M}$*, for all* $\varepsilon > 0$*,*

$$\begin{aligned} N_x(t) \leqslant \quad & \textstyle\sum_{t=1}^T \mathbb{1}\left\{ x_t = x, \mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon \right\} \\ & + \textstyle\sum_{t=1}^T \mathbb{1}\left\{ \overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon} \right\} \end{aligned} \tag{41}$$

where $x_t = (s_t, a_t)$*.*

Proof. The proof is immediate by decomposing the event $\{x_t = x\}$ on

$$\left\{ \mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon \right\}.$$

$$\begin{aligned} \mathbb{1}\{x_t = x\} &= \mathbb{1}\left\{ x_t = x, \mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon \right\} \\ &\quad + \mathbb{1}\left\{ x_t = x, \overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon} \right\} \\ &\leqslant \mathbb{1}\left\{ x_t = x, \mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon \right\} \\ &\quad + \mathbb{1}\left\{ \overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t))\right) \subseteq \mathcal{O}(\mathbf{M}), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^\mathbf{M}\|_\infty \leqslant \varepsilon} \right\}. \end{aligned}$$

■

D.1 Bandit term

Let us recall the definition of the quantity $B_x(T)$ we called the "bandit term",

$$B_x(T) = \sum_{t=1}^{T} \mathbb{1}\left\{x_t = x, \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon\right\}. \tag{42}$$

In order to control this quantity, we first make the following useful observation.

Lemma 18. *Whenever the inequality $\|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon$ holds true, it implies that for all ν, q,*

$$\left|\varphi_{\mathbf{M}}\left(\nu \otimes q\right) - \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right)\right| \leqslant \varepsilon.$$

Proof. Let us assume that $\|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon$. Recall that $\varphi_{\mathbf{M}}\left(\nu \otimes q\right) = \mathbb{E}_{R\sim\nu}[R] + q\mathbf{b}^{\mathbf{M}}$. Then, for all ν, q

$$\begin{aligned}
\left|\varphi_{\mathbf{M}}\left(\nu \otimes q\right) - \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right)\right| &= \left|\mathbb{E}_{R\sim\nu}[R] + q\mathbf{b}^{\mathbf{M}} - \mathbb{E}_{R\sim\nu}[R] + q\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\right| \\
&= \left|q\left(\mathbf{b}^{\mathbf{M}} - \mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\right)\right| \\
&\leqslant \sum_s q(s)\|\mathbf{b}^{\mathbf{M}} - \mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\|_\infty \\
&\leqslant \varepsilon\,.
\end{aligned}$$

■

In order to control (42), we further split the considered event depending on whether the threshold appearing in the complexity term is (subsection D.1.2) or not (subsection D.1.1) underestimated.

D.1.1 THE THRESHOLD IS NOT UNDERESTIMATED

In this subsection, we assume that the threshold is not underestimated, that is to say,

$$\hat{\gamma}_s(t) \geqslant \gamma_s\left(\mathbf{M}\right) - 2\varepsilon.$$

Non-critical state-action pair First, we study the case where the state-action pair (s, a) is not critical, that is to say $\mathcal{C}_{sa,\mathbf{M}}\left(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})\right) \stackrel{\text{def}}{=} \{\nu \otimes q \in \mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}) : \varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right)\} = \emptyset$.

Proposition 19. *For all non-critical state-action pair,*

$$\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) \geqslant \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\} = 0\,. \tag{43}$$

Proof. Let $\mathcal{C}_{sa,\mathbf{M}}\left(\mathcal{F}_{s,a} \otimes \mathcal{P}(\mathcal{S}), \delta\right) \stackrel{\text{def}}{=} \{\nu \otimes q \in \mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}) : \varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - \varepsilon\}$. There exists $\delta_{s,a}$ small enough such that $\mathcal{C}_{s,\mathbf{M}}\left(\mathcal{F}_{s,a} \otimes \mathcal{P}(\mathcal{S}), \delta_{s,a}\right) = \emptyset$. Let ε be strictly smaller than $\delta_{s,a}/3$ (which is the case for ε defined as in the statement of Theorem 6),

$$\begin{aligned}
\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right) &\leqslant \varphi_{\mathbf{M}}\left(\nu \otimes q\right) + \varepsilon \\
&\leqslant \mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s) - \delta_{s,a} + \varepsilon \\
&< \mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s) - 2\varepsilon \\
&= \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \\
&\leqslant \hat{\gamma}_s(t)
\end{aligned}$$

Therefore, for all distributions in $\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})$, $\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(\nu \otimes q) < \hat{\gamma}_s(t)$ meaning that the empirical set of critical distributions is empty, $\mathcal{C}_{s,\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(\mathcal{F}_{s,a} \otimes \mathcal{P}(\mathcal{S})) = \emptyset$. Remembering that we define $\inf \emptyset = +\infty$ by convention, it implies that $\underline{\mathbf{K}}_{s,a}(t) = +\infty$ and therefore, $\mathbf{H}_{s,a}(t) = +\infty$. A state-action pair with infinite index can never be sampled since at least one action at that state, the current empirical best one, has a finite, therefore strictly smaller, index. ■

Proposition 20. *Following immediately from Proposition 19, for all sub-optimal state-action pair $(s, a) \in \mathcal{X}_{\mathbf{M}}$ that are not critical, it holds*

$$\sum_{t=1}^{T} \mathbb{1} \left\{ \begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) \geqslant \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array} \right\} = 0. \tag{44}$$

Critical state-action pair We now study the case were (s, a) is critical, *i.e.* it can be made optimal under the distributions assumptions, formally $\mathcal{C}_{sa,\mathbf{M}}(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})) \neq \emptyset$.

Proposition 21 (Number of pulls of critical state-action pair in the bandit term)**.** *For all sub-optimal state-action pair $(s, a) \in \mathcal{X}_{\mathbf{M}}$ that are critical, it holds*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1} \left\{ \begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) \geqslant \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array} \right\}\right] &\leqslant \frac{\log T}{\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \varepsilon \Gamma_{s,a}\left(\mathbf{M}\right)} \\ &+ \frac{1}{1 - \exp\left(-\tilde{\Lambda}_{s,a}\left(\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \frac{\varepsilon}{2}\Gamma_{s,a}\left(\mathbf{M}\right)\right)\right)},\end{aligned} \tag{45}$$

where $\Gamma_{s,a}(\mathbf{M}) = \frac{\mathbf{m}_{max}(s,a) + \mathbf{b}^{\mathbf{M}}(s)}{\mathbf{m}_{max}(s,a) + \mathbf{b}^{\mathbf{M}}(s) - \gamma_s(\mathbf{M})}$.

Proof. For all $\nu \otimes q \in \mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})$,

$$\begin{aligned}\varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(\nu \otimes q) > \hat{\gamma}_s(t) &\implies \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(\nu \otimes q) > \gamma_s(\mathbf{M}) - 2\varepsilon \\ &\implies \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma_s(\mathbf{M}) - \varepsilon\end{aligned}$$

therefore, $\mathcal{C}_{sa,\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})) \subseteq \mathcal{C}_{sa,\mathbf{M}}(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}), \varepsilon)$. Because the infimum over a larger set is smaller than the infimum over a smaller (for the inclusion order),

$$\underline{\mathcal{K}}^{\varepsilon}_{s,a}(t) \leqslant \underline{\mathbf{K}}_{s,a}(t), \tag{46}$$

where

$$\underline{\mathcal{K}}^{\varepsilon}_{s,a}(t) \overset{\text{def}}{=} \inf_{\substack{\nu \in \mathcal{F}_{s,a} \\ q \in \mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\hat{\mathbf{r}}_t(s,a) \otimes \hat{\mathbf{p}}_t(s,a), \nu \otimes q\right) \,:\, \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma_s(\mathbf{M}) - \varepsilon \right\}. \tag{47}$$

We recall that $\mathbf{r}_t(s,a) \otimes \hat{\mathbf{p}}_t(s,a) = \mathbf{r}_t(s,a,N_{s,a}(t)) \otimes \hat{\mathbf{p}}_t(s,a,N_{s,a}(t))$ and we denote

$$\underline{\mathcal{K}}^{\varepsilon,n}_{s,a} \overset{\text{def}}{=} \inf_{\substack{\nu \in \mathcal{F}_{s,a} \\ q \in \mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\hat{\mathbf{r}}(s,a,n) \otimes \hat{\mathbf{p}}_t(s,a,n), \nu \otimes q\right) \,:\, \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma_s(\mathbf{M}) - \varepsilon \right\} \tag{48}$$

the random variable associated to n samples of state-action pair (s, a). Because $(s_t, a_t) = (s, a)$, under the the studied event,

$$\begin{aligned}N_{s,a}(t)\underline{\mathcal{K}}^{\varepsilon}_{s,a}(t) &\leqslant N_{s,a}(t)\underline{\mathbf{K}}_{s,a}(t) \\ &\leqslant \mathbf{H}_{s,a}(t) \\ &\leqslant \max_{a'} \log N_{s,a'}(t) \\ &\leqslant \log t \\ &\leqslant \log T\end{aligned}$$

Therefore, for a critical state action pair,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) \geqslant \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\} &\leqslant \sum_{n=1}^{T}\sum_{t=1}^{T}\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ N_{s,a}(t) = n \\ n\underline{\mathcal{K}}_{s,a}^{\varepsilon,n} \leqslant \log T \end{array}\right\} \\ &\leqslant \sum_{n=1}^{T}\mathbb{1}\left\{n\underline{\mathcal{K}}_{s,a}^{\varepsilon,n} \leqslant \log T\right\}\sum_{t=1}^{T}\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ N_{s,a}(t) = n \end{array}\right\} \\ &\leqslant \sum_{n=1}^{T}\mathbb{1}\left\{n\underline{\mathcal{K}}_{s,a}^{\varepsilon,n} \leqslant \log T\right\}\end{aligned}$$

It remains to control the expected value of the term

$$\sum_{n=1}^{T}\mathbb{1}\left\{n\underline{\mathcal{K}}_{s,a}^{\varepsilon,n} \leqslant \log T\right\}\ , \tag{49}$$

that is to say, control

$$\inf_{\substack{\nu \in \mathcal{F}_{s,a} \\ q \in \mathcal{P}(\mathcal{S})}}\left\{\mathrm{KL}\left(\hat{\mathbf{r}}(s,a,n) \otimes \hat{\mathbf{p}}_t(s,a,n), \nu \otimes q\right)\ :\ \varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - \varepsilon\right\}\ ,$$

a quantity that has attracted a lot of attention from the theoretical Bandit community. In particular, under the Assumption 3 (semi-bounded) and Assumption 1 (light-tail), one can apply Lemma 7 of Honda and Takemura (2015) (as in their Theorem 3) to deduce that,

$$\mathbb{E}\left[\sum_{n=1}^{T}\mathbb{1}\left\{n\underline{\mathcal{K}}_{s,a}^{\varepsilon,n}\right\}\right] \leqslant \frac{\log T}{\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \varepsilon\Gamma_{s,a}\left(\mathbf{M}\right)} + \frac{1}{1 - \exp\left(-\tilde{\Lambda}_{s,a}\left(\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \frac{\varepsilon}{2}\Gamma_{s,a}\left(\mathbf{M}\right)\right)\right)} \tag{50}$$

where $\Gamma_{s,a}\left(\mathbf{M}\right) = \frac{\mathbf{m}_{max}(s,a) + \mathbf{b}^{\mathbf{M}}(s)}{\mathbf{m}_{max}(s,a) + \mathbf{b}^{\mathbf{M}}(s) - \gamma_s(\mathbf{M})}$. ■

D.1.2 THE THRESHOLD IS UNDERESTIMATED

In this subsection, we now turn to the case when the threshold is underestimated, that is to say,

$$\hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon.$$

In particular, it means that the gain is underestimated since $\gamma_s\left(\mathbf{M}\right) = \mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s)$ and that for the studied bandit term, the bias is ε-well estimated and the empirical set of optimal state-action pairs is included in the true one. Because all empirically optimal actions belong to the skeleton, those are bound to have been sampled enough. Further using the ergodic Assumption 2, we get the concentration we need to bound the expected value of

$$\sum_{t=1}^{T}\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\}.$$

Proposition 22. *For all sub-optimal state-action pair,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\}\right] \leqslant \min_{\star \in \mathcal{O}_s(\mathbf{M})}\zeta_{s\star}\left(\varepsilon\right)\ , \tag{51}$$

where

$$\zeta_{s\star}\left(\varepsilon\right) = \frac{6e}{\left(1 - \frac{1}{\lambda_{s\star,\varepsilon}}\right)\left(1 - \exp\left(-\left(1 - \frac{1}{\lambda_{s\star,\varepsilon}}\right)\Lambda_{s\star}^{*}(\varepsilon)\right)\right)^3}.$$

Proof. Let $\star \in \mathcal{O}_s(\mathbf{M})$ be an optimal action in state s, and denote $\mathbf{H}_s^\star(t) = \min \mathbf{H}_{s,a}(t)$ the value of the minimal index. We note that $\mathbf{H}_s^\star(t) \geqslant \log N_{x_t}(t) = \log N_{s,a}(t)$ since $x_t = (s,a)$ under the current studied event. In particular, $\mathbf{H}_{s,\star}(t) \geqslant \log N_{s,a}(t)$.

For all $\nu \otimes q \in \mathcal{F}_{s\star} \otimes \mathcal{P}(\mathcal{S})$,

$$\begin{aligned}\varphi_{\mathbf{M}}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - \varepsilon &\implies \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right) + \varepsilon > \gamma_s\left(\mathbf{M}\right) - \varepsilon \\ &\implies \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right) > \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \\ &\implies \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\nu \otimes q\right) > \hat{\gamma}_s(t).\end{aligned}$$

This implies that $\mathcal{C}_{sa,\mathbf{M}}\left(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S}), \varepsilon\right) \subseteq \mathcal{C}_{sa,\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\mathcal{F}_{sa} \otimes \mathcal{P}(\mathcal{S})\right)$ and, using notation $\underline{\mathcal{K}}_{s\star}$ introduced Equation 47, proves that $\underline{\mathbf{K}}_{s\star}(t) \leqslant \underline{\mathcal{K}}_{s\star}^\varepsilon(t)$, *i.e.*, combining with a previous inequality,

$$\log N_{s,a}(t) \leqslant \mathbf{H}_{s,\star}(t) \leqslant N_{s,\star}(t)\underline{\mathcal{K}}_{s\star}^\varepsilon(t) + \log N_{s,\star}(t). \tag{52}$$

Furthermore, because $\hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon$, we have that

$$\begin{aligned}\varphi_{\mathbf{M}}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) &\leqslant \varphi_{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) + \varepsilon \\ &\leqslant \hat{\gamma}_s(t) + \varepsilon \\ &< \gamma_s\left(\mathbf{M}\right) - 2\varepsilon + \varepsilon \\ &= \varphi_{\mathbf{M}}\left(\mathbf{r}(s,\star) \otimes \mathbf{p}(s,\star)\right) - \varepsilon\,.\end{aligned}$$

This implies that

$$\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\}$$

is smaller than or equal to

$$\mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \log N_{s,a}(t) \leqslant N_{s,\star}(t)\underline{\mathcal{K}}_{s\star}^\varepsilon(t) + \log N_{s,\star}(t) \\ \varphi_{\mathbf{M}}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) \leqslant \varphi_{\mathbf{M}}\left(\mathbf{r}(s,\star) \otimes \mathbf{p}(s,\star)\right) - \varepsilon \end{array}\right\}. \tag{53}$$

Recalling that, by definition, $\varphi_{\mathbf{M}}\left(\mathbf{r}(s,\star) \otimes \mathbf{p}(s,\star)\right)$ is an expected value, this quantity is controlled by Lemma 14 of Honda and Takemura (2015) with

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \log N_{s,a}(t) \leqslant N_{s,\star}(t)\underline{\mathcal{K}}_{s\star}^\varepsilon(t) + \log N_{s,\star}(t) \\ \varphi_{\mathbf{M}}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) \leqslant \varphi_{\mathbf{M}}\left(\mathbf{r}(s,\star) \otimes \mathbf{p}(s,\star)\right) - \varepsilon \end{array}\right\}\right) \leqslant \zeta_{s\star}\left(\varepsilon\right) \tag{54}$$

with

$$\zeta_{s\star}\left(\varepsilon\right) = \frac{6e}{\left(1 - \frac{1}{\lambda_{s\star,\varepsilon}}\right)\left(1 - \exp\left(-\left(1 - \frac{1}{\lambda_{s\star,\varepsilon}}\right)\Lambda_{s\star}^*(\varepsilon)\right)\right)^3}. \tag{55}$$

Finally, taking the minimum over all optimal arm $\star$, we get the result of the proposition,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\begin{array}{c} s_t = s, a_t = a \\ \mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right) \\ \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon \\ \hat{\gamma}_s(t) < \gamma_s\left(\mathbf{M}\right) - 2\varepsilon \end{array}\right\}\right] \leqslant \min_{\star \in \mathcal{O}_s(\mathbf{M})} \zeta_{s\star}\left(\varepsilon\right).$$

■

D.1.3 BANDIT TERM: UPPER BOUND

Proposition 23 (Upper bound for bandit term)**.** *Wrapping up, Propositions 21 and 22, if (s, a) is a critical state-action pair,*

$$\begin{aligned}\mathbb{E}\left[B_{s,a}(T)\right] \leqslant\, & \frac{\log T}{\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \varepsilon \Gamma_{s,a}\left(\mathbf{M}\right)} \\ & + \frac{1}{1 - \exp\left(-\tilde{\Lambda}_{s,a}\left(\underline{\mathbf{K}}_{s,a}\left(\mathbf{M}\right) - \frac{\varepsilon}{2}\Gamma_{s,a}\left(\mathbf{M}\right)\right)\right)} \\ & + \min_{\star \in \mathcal{O}_s(\mathbf{M})} \zeta_{s\star}\left(\varepsilon\right),\end{aligned} \tag{56}$$

and by Propositions 20 and 22, if (s, a) is not a critical state-action pair,

$$\mathbb{E}\left[B_{s,a}(T)\right] \leqslant \min_{\star \in \mathcal{O}_s(\mathbf{M})} \zeta_{s\star}\left(\varepsilon\right). \tag{57}$$

D.2 Skeleton improvement term

In this part of the main proof, we aim at controlling the expected value of the sum

$$S(T) = \sum_{t=1}^{T} \mathbb{1}\left\{\overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon}\right\} \tag{58}$$

and prove that $\mathbb{E}\left(S(T)\right) = O(1)$. For readability, we denote

$$W_t(\varepsilon) \stackrel{\text{def}}{=} \left\{\overline{\mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right), \|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\|_\infty \leqslant \varepsilon}\right\}$$

in the rest of proof.

Proposition 24. *Let V_t and $S_t\left(\delta\right)$ as in Lemma 12, then*

$$\begin{aligned}\mathbb{E}\left[S(T)\right] \leqslant \mathbb{E}&\left[\sum_{t=1}^{T} \mathbb{1}\left\{W_t(\varepsilon), V_t, S_t\left(\delta\right)\right\}\right] \\ &+ (\Pi+2) + (\Pi+1)\left|\mathcal{S}\right| \frac{B_{\mathbf{M}}}{1-\exp\left(-\left(\beta_{\mathbf{M}}-\beta\right)\right)} + S_{\mathbf{M}}\left(\delta\right)\end{aligned}.$$

Proof. Recall that we denote (σ_t^ν, I_t^ν) the $\mathbf{M}$-adapted sub-division of $I_t = \{1, \cdots, t\}$ used in V_t and $S_t(\delta)$ as in Definition 11, $0 < \beta < \beta_{\mathbf{M}}$ as in Proposition 10, and V_t and $S_t\left(\delta\right)$ are the events defined in Lemma 13. We first decompose $\{W_t(\varepsilon)\}$ on $V_t \cap S_t\left(\delta\right)$ and $\overline{V_t \cap S_t\left(\delta\right)}$ by the law of total probability, and deduce the inequality

$$\begin{aligned}\mathbb{1}\left\{W_t(\varepsilon)\right\} \leqslant\ & \mathbb{1}\left\{W_t(\varepsilon), V_t, S_t\left(\delta\right)\right\} \\ &+ \mathbb{1}\left\{\overline{V_t}\right\} + \mathbb{1}\left\{V_t \cap \overline{S_t(\delta)}\right\}\end{aligned}. \tag{59}$$

By Lemma 12,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\overline{V_t}\right\}\right] \leqslant (\Pi+2) + (\Pi+1)\left|\mathcal{S}\right| \frac{B_{\mathbf{M}}}{1-\exp\left(-\left(\beta_{\mathbf{M}}-\beta\right)\right)}, \tag{25}$$

and by Lemma 13,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{V_t \cap \overline{S_t\left(\delta\right)}\right\}\right] \leqslant S_{\mathbf{M}}\left(\delta\right), \tag{28}$$

therefore,

$$\begin{aligned}\mathbb{E}\left[S(T)\right] \leqslant \mathbb{E}&\left[\sum_{t=1}^{T} \mathbb{1}\left\{W_t(\varepsilon), V_t, S_t\left(\delta\right)\right\}\right] \\ &+ (\Pi+2) + (\Pi+1)\left|\mathcal{S}\right| \frac{B_{\mathbf{M}}}{1-\exp\left(-\left(\beta_{\mathbf{M}}-\beta\right)\right)} + S_{\mathbf{M}}\left(\delta\right)\end{aligned}.$$

■

Outline - Intuition The intuition for controlling the remaining term,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{W_t(\varepsilon), V_t, S_t\left(\delta\right)\right\}\right],$$

is the following. There are $\Pi + 1$ sub-intervals and Π policies. At the end of the first sub-interval, all states have been visited linear amount of time and from that point, the gain and bias are well estimated on the skeleton, *i.e.*, the bandit problem defined by $\varphi_{\mathbf{M}}(\mathcal{A}(k))$ for all $k \geqslant \sigma_t^1$. Because we play a bandit strategy, sub-optimal actions for the current bandit problem are guaranteed to be played a logarithmic amount of time while at least one improving action will be played a linear amount of time until it belongs to the skeleton, defining a new well-estimated bandit problem. Because the condition to belong to the skeleton is in $\log^2$ which is greatly sub-linear, an improving action will quickly makes its way to the skeleton. Each interval lasting a linear amount of time, the probability that a skeleton improvement, if one is possible, occurred between the beginning and end of the sub-interval is high. Since there are Π policies and Π sub-intervals after the first one, by the pigeonhole principle, the probability that the skeleton contains an optimal policy by the end of the last sub-interval is large. The two main propositions for controlling this term are Propositions 26 and 29.

Lemma 25 (Skeleton coherence)**.** *Under the event $V_t \cap S_t(\delta)$ with $\delta < \delta_{\mathbf{M}}(\varepsilon_{\mathbf{M}})$, for all $k \geqslant \sigma_t^1$,*

$$\mathbf{g}^{\mathbf{M}(\mathcal{A}(k+1))} \geqslant \mathbf{g}^{\mathbf{M}(\mathcal{A}(k))}. \tag{60}$$

Proof. It is sufficient to show that $\mathcal{A}(k+1) \cap \mathcal{O}\left(\widehat{\mathbf{M}}_k(\mathcal{A}(k))\right) \neq \emptyset$, *i.e.* to prove that no empirically optimal policy at time k disappear from the skeleton after the action taken at that time. We distinguish between two cases.

- If the sampled action $a_k \notin \mathcal{A}_{s_k}(k)$, *i.e.* the chosen action does not belong to the current skeleton, then the skeleton at s_k can only grow in size, as $a_k \notin \arg\max_a \max_a N_{s_k a}(k)$,
$$\mathcal{A}_{s_k}(k+1) \subseteq \mathcal{A}_{s_k}(k) \cup \{a_k\},$$
and therefore, in that case, $\mathbf{g}^{\mathbf{M}(\mathcal{A}(k+1))} \geqslant \mathbf{g}^{\mathbf{M}(\mathcal{A}(k))}$.
- If the sampled action $a_k \in \mathcal{A}_{s_k}(k)$, *i.e.* the chosen action belongs to the current skeleton, then the size of the skeleton may decrease if $a_k \notin \arg\max_a \max_a N_{s_k a}(k)$. We distinguish again between two cases.
  - If the sampled action $a_k \in \arg\max_a \varphi_{\widehat{\mathbf{M}}_k(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s_k, a) \otimes \mathbf{p}^k(s_k, a)\right)$, *i.e.* the chosen action is empirically optimal and belongs to $\mathcal{O}\left(\widehat{\mathbf{M}}_k(\mathcal{A}(k))\right)$, then this action will belong to the skeleton at time $k+1$ (whether $a_k \notin \arg\max_a \max_a N_{s_k a}(k)$ or not). By Lemma 16, $\mathcal{O}\left(\widehat{\mathbf{M}}_k(\mathcal{A}(k))\right) \subseteq \mathcal{O}(\mathbf{M}(\mathcal{A}(k)))$ and therefore, the true gain on the skeleton will remain the same.
  - If the sampled action $a_k \notin \arg\max_a \varphi_{\widehat{\mathbf{M}}_k(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s_k, a) \otimes \mathbf{p}^k(s_k, a)\right)$, then we show that it cannot belong to $\arg\max_a \max_a N_{s_k a}(k)$ and thus that the skeleton remains the same between times k and $k+1$. We show this fact by contradiction. If a' is an action that is not empirically optimal and belongs to $\arg\max_a \max_a N_{s_k a}(k)$, then $\mathbf{H}_{s_k a'}(k) > \log \max_a N_{s_k a}(k)$. On the other hand, for all empirically optimal action $\star$, $\mathbf{H}_{s_k \star}(k) = \log N_{s_k \star} \leqslant \log \max_a N_{s_k a}(k)$. Therefore, $\mathbf{H}_{s_k a'}(k) > \mathbf{H}_{s_k \star}(k)$ and action a' cannot be sampled.

  Therefore, in that case, $\mathbf{g}^{\mathbf{M}(\mathcal{A}(k+1))} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(k))}$.

We proved that under the event $V_t \cap S_t(\delta)$ with $\delta < \delta_{\mathbf{M}}(\varepsilon_{\mathbf{M}})$, for all $k \geqslant \sigma_t^1$,

$$\mathbf{g}^{\mathbf{M}(\mathcal{A}(k+1))} \geqslant \mathbf{g}^{\mathbf{M}(\mathcal{A}(k))}.$$

where strict improvement can only occur when the sampled action a_k does not belong to the skeleton. If the action a_k belongs to the current skeleton, the gain on the skeleton can only remains the same. ■

An immediate consequence of Lemma 25 is that under the event $V_t \cap S_t(\delta)$, $\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} \geqslant \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^{\nu}))}$. The aim is to prove that the inequality is strict unless the optimal gain has already been reached, *i.e.* an optimal policy belongs to the skeleton. Another consequence is that if the skeleton contains an optimal policy at some time k, then this policy will remain in the skeleton for all subsequent step (under the event $V_t \cap S_t(\delta)$).

For all $\nu \in \{1, \cdots, \Pi - 1\}$, let

$$G_t^{\nu} = \left\{\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} > \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^{\nu}))}\right\} \bigcup \left\{\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^{\nu}))} = \mathbf{g}^{\mathbf{M}}\right\}, \tag{61}$$

and let

$$G_t = \bigcap_{\nu=1}^{\Pi-1} G_t^\nu, \tag{62}$$

where G_t is an event in which there is a still skeleton improvement between each interval until optimality is reached.

Proposition 26.

$$\sum_{t=1}^{T} \mathbb{1}\left\{W_t\left(\varepsilon\right), V_t, S_t\left(\delta\right), G_t\right\} = 0. \tag{63}$$

Proof. One one hand, there are Π policies, therefore the pigeonholes principle implies that under G_t, by the end of the last sub-interval, $\mathcal{O}\left(\mathbf{M}\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right)$. Furthermore, under $V_t \cap S_t\left(\delta\right)$, by Lemma 16, $\mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\left(\mathcal{A}(t)\right)\right)$. Thus, $\mathcal{O}\left(\widehat{\mathbf{M}}_t\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right)$. Furthermore, under the event $S_t\left(\delta\right)$, by Lemma 16, $\|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}(\mathcal{A}(t))}\| \leqslant \frac{\varepsilon}{2}$ and since $\mathcal{O}\left(\mathbf{M}\left(\mathcal{A}(t)\right)\right) \subseteq \mathcal{O}\left(\mathbf{M}\right)$, it follows that $\mathbf{b}^{\mathbf{M}(\mathcal{A}(t))} = \mathbf{b}^{\mathbf{M}}$ by the optimal Poisson equation and therefore, $\|\mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))} - \mathbf{b}^{\mathbf{M}}\| \leqslant \frac{\varepsilon}{2}$. Therefore, event $\left\{V_t, S_t\left(\delta\right), G_t\right\}$ implies $\overline{W_t(\varepsilon)}$.
Thus, $\left\{W_t\left(\varepsilon\right), V_t, S_t\left(\delta\right), G_t\right\} = \emptyset$, the indicator of such an event is always 0 and the sum equally. ■

The last proposition lets us to study, for all $\nu \in \{1, \cdots, \Pi - 1\}$,

$$\sum_{t=1}^{T} \mathbb{1}\left\{W_t(\varepsilon), V_t, S_t\left(\delta\right), \overline{G_t^\nu}\right\} \leqslant \sum_{t=1}^{T} \mathbb{1}\left\{V_t, S_t\left(\delta\right), \overline{G_t^\nu}\right\},$$

because $\overline{G_t} = \bigcup_{\nu=1}^{\Pi-1} \overline{G_t^\nu}$ (union bound).

Lemma 27. *The equality,*

$$\left\{V_t, S_t\left(\delta\right), \overline{G_t^\nu}\right\} = \left\{V_t, S_t\left(\delta\right), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} = \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)} < \mathbf{g}^{\mathbf{M}}\right\}, \tag{64}$$

is true.

Proof. Under $V_t \cap S_t\left(\delta\right)$, Lemma 25 implies that, $\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} \geqslant \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)}$. $\overline{G_t^\nu}$ implies that $\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} < \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)}$ or $\mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} = \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)} < \mathbf{g}^{\mathbf{M}}$. Therefore the result. ■

In the last part of the proof, we aim to control the probability that, under good event $V_t \cap S_t\left(\delta\right)$, no improvement occurs during sub-interval ν. Denote by $I\left(s, \mathbf{M}, \gamma\right)$ the set of improving actions over a threshold γ,

$$I\left(s, \mathbf{M}, \gamma\right) = \left\{a \in \mathcal{A}_s : \varphi_{\mathbf{M}}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)\right) > \gamma\right\}, \tag{65}$$

and $I^+\left(s, \mathbf{M}, \gamma\right)$ the set of maximally improving actions over a threshold (it may be empty),

$$I^+\left(s, \mathbf{M}, \gamma\right) = \arg\max\left\{\varphi_{\mathbf{M}}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)\right) : a \in I\left(s, \mathbf{M}, \gamma\right)\right\}. \tag{66}$$

Lemma 28. *On the event,*

$$\left\{V_t, S_t\left(\delta\right), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} = \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)} < \mathbf{g}^{\mathbf{M}}\right\}$$

we have that for all $k \in I_t^\nu \cup \left\{\sigma_t^{\nu+1}\right\}$

$$\mathbf{g}^{\mathbf{M}(\mathcal{A}(k))} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}, \tag{67}$$

$$\mathbf{b}^{\mathbf{M}(\mathcal{A}(k))} = \mathbf{b}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}, \tag{68}$$

$$\varphi_{\mathbf{M}(\mathcal{A}(k))} = \varphi_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}, \tag{69}$$

$$\tag{70}$$

and for all $s \in \mathcal{S}$,

$$I\left(s, \mathbf{M}\left(\mathcal{A}(k)\right), \gamma_{\mathbf{M}(\mathcal{A}(k))}\right) = I\left(s, \mathbf{M}\left(\mathcal{A}(\sigma_t^\nu)\right), \gamma_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}\right). \tag{71}$$

Proof. The first three equations result from the, definition of the event, the unicity of the optimal Poisson equation (see Equation 5.2.18, Chapter 5, Hernández-Lerma and Lasserre (1996)) and the definition of $\varphi_{\mathbf{M}}$.[9]

The set of improving actions cannot change during $I_t^\nu \cup \left\{\sigma_t^{\nu+1}\right\}$ because if it were, it would mean that an improving action now belong to the skeleton. This would lead to a strict increase of the gain during I_t^ν, which is in contradiction with the studied event. ■

Proposition 29 (Expected time before policy improvement)**.** *Under the event $V_t \cap S_t(\delta)$, the expected time during interval ν without skeleton improvement is bounded, i.e.,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{V_t, S_t\left(\delta\right), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu+1}\right)\right)} = \mathbf{g}^{\mathbf{M}\left(\mathcal{A}\left(\sigma_t^{\nu}\right)\right)} < \mathbf{g}^{\mathbf{M}}\right\}\right] \leqslant \Pi A\left(\tilde{\zeta}\left(\varepsilon\right) + \tau(\mathbf{M})\right), \tag{72}$$

where $\tilde{\zeta}\left(\varepsilon\right)$ is defined by Equation 75, $\tau(\mathbf{M})$ is defined in Equation 81 and $A = \max_s |\mathcal{A}_s|$.

Outline of the proof We aim to control the number of pulls within the improving set during interval ν. The threshold to belong to the skeleton will be at most $\log^2 \sigma_t^{\nu+1}$ and I_ν^t is of linear length. Furthermore, the applied bandit strategy on the well estimated problem given by $\varphi_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}$ will sample roughly a linear number of times an optimal action for that problem, *i.e.* an action in the improving set. This will imply that such an action belong to the skeleton by the end of the interval, contradicting the non-increasing gain assumed by the studied event. We will have to control the probability that no optimal action is sampled more than $\log^2\left(\sigma_t^{\nu+1}\right)$ during interval ν, which is very unlikely.

Proof. First, we remark that there exits a state $s \in \mathcal{S}$ such that $I\left(s, \mathbf{M}\left(\mathcal{A}(\sigma_t^\nu)\right), \gamma_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}\right)$ is not empty because the gain $\mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))} < \mathbf{g}^{\mathbf{M}}$ and an improving action is bound to exist by the policy improvement theorem. Let s be such a state.

We denote by $P_t^\nu(s) = \sum_{k \in I_t^k} \mathbb{1}\left\{s_k = s, a_k \in \mathcal{A}_s(k)\right\}$ the number of times the pulled action belongs to the skeleton at state s, while an improving action outside of it exists. We have,

$$P_t^\nu(s) = \sum_{k \in I_t^k} \mathbb{1}\left\{s_k = s, a_k \in \mathcal{A}_s(k), \hat{\gamma}_s(k) \geqslant \gamma_s\left(k\right) - 2\varepsilon\right\} \tag{73}$$

$$+ \sum_{k \in I_t^k} \mathbb{1}\left\{s_k = s, a_k \in \mathcal{A}_s(k), \hat{\gamma}_s(k) < \gamma_s\left(k\right) - 2\varepsilon\right\} \tag{74}$$

The sum corresponding to Equation 73 is equal to 0 because under $V_t \cap S_t\left(\delta\right)$, for all action $a \in \mathcal{A}_s(k)$,

$$\begin{aligned}
\varphi_{\widehat{\mathbf{M}}_k(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s,a) \otimes \hat{\mathbf{p}}^k(s,a)\right) &\leqslant \varphi_{\mathbf{M}(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s,a) \otimes \hat{\mathbf{p}}^k(s,a)\right) + \frac{\varepsilon}{2} \\
&\leqslant \varphi_{\mathbf{M}(\mathcal{A}(k))}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)\right) + \varepsilon \\
&< \gamma_s\left(k\right) - 3\varepsilon + \varepsilon \\
&= \gamma_s\left(k\right) - 2\varepsilon \\
&\leqslant \hat{\gamma}_s(k).
\end{aligned}$$

By definition of s, no empirically optimal action $\star$ in state s belong to the skeleton and $\hat{\gamma}_s(k)$ is realized by an action $\star \notin \mathcal{A}_s(k)$. Let $\star$ be such an action, its `IMED-RL` index is $\log N_{s,\star}(k) < \log \max_{a'} N_{s,a'}(k)$. Further, for all action $a \in \mathcal{A}_s(k)$, its `IMED-RL` index is strictly larger than $\log N_{s,a}(k) \geqslant \log \max_{a'} N_{s,a'}(k)$. Therefore, a_k cannot belong to $\mathcal{A}_s(k)$ if $\hat{\gamma}_s(k) \geqslant \gamma_s\left(k\right) - 2\varepsilon$ under the favorable event $V_t \cap S_t\left(\delta\right)$. Thus,

$$P_t^\nu(s) = \sum_{k \in I_t^k} \mathbb{1}\left\{s_k = s, a_k \in \mathcal{A}_s(k), \hat{\gamma}_s(k) < \gamma_s\left(k\right) - 2\varepsilon\right\}.$$

9. While the bias is defined up to a constant, all choices are made by comparing $\varphi_{\mathbf{M}(\mathcal{A}(t))}$ on empirical distributions, which cancels out the global constant. All equality results are stated modulo this global additive constant.

For all $a \in I^+(s, \mathbf{M}, \gamma_s(k))$, $\varphi_{\mathbf{M}(\mathcal{A}(k))}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)) = \gamma_s(k)$ and is an optimal choice of action for the bandit problem defined by $\varphi_{\mathbf{M}(\mathcal{A}(k))} = \varphi_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}$. While this fact is not used, we still note that any action $a \in I(s, \mathbf{M}, \gamma_s(k))$ that is sampled enough will make the gain increase.

For all $a \in I^+(s, \mathbf{M}, \gamma_s(k))$, $\varphi_{\mathbf{M}(\mathcal{A}(k))}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)) = \gamma_s(k)$, and therefore

$$\begin{aligned}
\varphi_{\mathbf{M}(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s,a) \otimes \hat{\mathbf{p}}^k(s,a)\right) &\leqslant \varphi_{\widehat{\mathbf{M}}_k(\mathcal{A}(k))}\left(\hat{\mathbf{r}}^k(s,a) \otimes \hat{\mathbf{p}}^k(s,a)\right) + \varepsilon \\
&\leqslant \hat{\gamma}_s(k) + \varepsilon \\
&< \gamma_s(k) - \varepsilon \\
&= \varphi_{\mathbf{M}(\mathcal{A}(k))}(\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)) - \varepsilon
\end{aligned}$$

By a union bound, for all $\star \in I^+(s, \gamma_s(\sigma_t^\nu))$,

$$P_t^\nu(s) \leqslant \sum_{a \notin I^+(s,\gamma_s(\sigma_t^\nu))} \sum_{k \in I_\nu^t} \mathbb{1}\left\{\begin{array}{l} s_k = s, a_k = a \\ \log N_{s,a}(t) \leqslant N_{s,\star}(k)\underline{\mathcal{K}}^\varepsilon_{s\star}(k) + \log N_{s,\star}(k) \\ \varphi_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) \leqslant \gamma_s\left(\mathbf{M}\left(\mathcal{A}(\sigma_t^\nu)\right)\right) - \varepsilon \end{array}\right\},$$

a quantity that is similar to the one controlled in Equation 53. In particular, this quantity is similarly controlled by Lemma 14 of Honda and Takemura (2015) with

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\begin{array}{l} s_k = s, a_k = a \\ \log N_{s,a}(t) \leqslant N_{s,\star}(k)\underline{\mathcal{K}}^\varepsilon_{s\star}(k) + \log N_{s,\star}(k) \\ \varphi_{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))}\left(\hat{\mathbf{r}}(s,\star) \otimes \hat{\mathbf{p}}(s,\star)\right) \leqslant \gamma_s\left(\mathbf{M}\left(\mathcal{A}(\sigma_t^\nu)\right)\right) - \varepsilon \end{array}\right\}\right] \leqslant \tilde{\zeta}(\varepsilon) \tag{75}$$

where

$$\tilde{\zeta}(\varepsilon) = \max_{\pi \in \Pi(\mathbf{M})} \max_{s \in \mathcal{S}} \min_{\star \in I^+(s, \mathbf{M}_\pi)} \frac{6e}{\left(1 - \frac{1}{\lambda_{s\star,\mathbf{M}_\pi,\varepsilon}}\right)\left(1 - \exp\left(-\left(1 - \frac{1}{\lambda_{s\star,\mathbf{M}_\pi,\varepsilon}}\right)\Lambda^*_{s\star,\mathbf{M}_\pi}(\varepsilon)\right)\right)^3}.$$

This proves that $\mathbb{E}\left[P_t^\nu(s)\right] \leqslant A\tilde{\zeta}(\varepsilon)$ where $A = \max_s |\mathcal{A}_s|$, *i.e.*, $P_t^\nu(s)$ is a positive random variable with finite expected value. This implies that, $\mathbb{P}\left(P_t^\nu(s) \geqslant \alpha t\right) = o(1/t)$ for all $\alpha > 0$.

To end the proof, we decompose,

$$\mathbb{1}\left\{V_t, S_t(\delta), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}(\sigma_t^{\nu+1})\right)} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))} < \mathbf{g}^{\mathbf{M}}\right\} \tag{76}$$

as the sum

$$\mathbb{1}\left\{V_t, S_t(\delta), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}(\sigma_t^{\nu+1})\right)} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))} < \mathbf{g}^{\mathbf{M}}, P_t^\nu(s) < \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\} \tag{77}$$

$$+\, \mathbb{1}\left\{V_t, S_t(\delta), \mathbf{g}^{\mathbf{M}\left(\mathcal{A}(\sigma_t^{\nu+1})\right)} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^\nu))} < \mathbf{g}^{\mathbf{M}}, P_t^\nu(s) \geqslant \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\}. \tag{78}$$

We control the term 78. This term is upper bounded by $\mathbb{1}\left\{P_t^\nu(s) \geqslant \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\}$, and because $\mathbb{P}\left(P_t^\nu(s) \geqslant \frac{\kappa\beta t}{|\mathcal{A}_s|}\right) \geqslant \mathbb{P}\left(P_t^\nu(s) \geqslant \left\lfloor \frac{\kappa\beta t}{|\mathcal{A}_s|} \right\rfloor\right)$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{P_t^\nu(s) \geqslant \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\}\right] \leqslant \left\lceil \frac{|\mathcal{A}_s|}{\kappa\beta} \right\rceil \tilde{\zeta}(\varepsilon) \leqslant \left\lceil \frac{A}{\kappa\beta} \right\rceil \tilde{\zeta}(\varepsilon), \tag{79}$$

by using again the fact that $\mathbb{E}\left(P_t^\nu(s)\right) = \sum_{n \in \mathbb{N}} \mathbb{P}\left(P_t^\nu(s) \geqslant n\right)$.

Finally, we control the term 77. The event $\left\{P_t^\nu(s) < \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\}$ implies that the number of times the sampled action belong to the skeleton during I_t^ν is only a fraction of that sub-interval. The remaining sample must therefore not belong to the skeleton. Sub-optimal actions outside the skeleton at time σ_t^ν will be sampled at most $\log^2 \sigma_t^{\nu+1}$ times and there are at most A such actions. Therefore, there are at most $\frac{\kappa\beta t}{|\mathcal{A}_s|} + A\log^2(\nu\beta t)$ samples that are not in the improving set $I(s, \gamma_s(\sigma_t^\nu))$. The improving set is at most of size A and therefore, at least one action in that set is sampled more than

$$\frac{1}{A}\left(\beta t - \left(\frac{\kappa\beta t}{|\mathcal{A}_s|} + A\log^2(\nu\beta t)\right)\right)$$

times. This quantity is linear and for t larger than a constant that depends on κ, β, A, $|\mathcal{A}_s|$ and ν, we have that,

$$\frac{1}{A}\left(\beta t - \left(\frac{\kappa\beta t}{|\mathcal{A}_s|} + A\log^2(\nu\beta t)\right)\right) \geqslant \log^2(\sigma_t^\nu). \tag{80}$$

We denote by $\tau(\mathbf{M})$, the maximum on $s \in \mathcal{S}$ and ν of these constants. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{V_t, S_t(\delta), \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^{\nu+1}))} = \mathbf{g}^{\mathbf{M}(\mathcal{A}(\sigma_t^{\nu}))} < \mathbf{g}^{\mathbf{M}}, P_t^\nu(s) < \frac{\kappa\beta t}{|\mathcal{A}_s|}\right\}\right] \leqslant \tau(\mathbf{M}). \tag{81}$$

■

D.2.1 SKELETON IMPROVEMENT: UPPER BOUND

Proposition 30. *By combining Propositions 24 and 29, then*

$$\begin{aligned}\mathbb{E}[S(T)] \leqslant \Pi A\left(\tilde{\zeta}(\varepsilon) + \tau(\mathbf{M})\right)& \\ + (\Pi+2) + (\Pi+1)|\mathcal{S}|\frac{B_{\mathbf{M}}}{1-\exp(-(\beta_{\mathbf{M}}-\beta))} + S_{\mathbf{M}}(\delta)&.\end{aligned} \tag{82}$$

D.3 Regret upper bound

Finally, one can express the full regret upper bound on the regret of IMED-RL by combining the decomposition of Proposition D.1, the result of Proposition 23 and the one of Proposition 30.

Theorem 31 (Regret upper bound for Ergodic MDPs)**.** *Let* $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ *be an MDP satisfying Assumptions 1, 2, 3. Let* $0 < \varepsilon \leqslant \frac{1}{3} \min_{\pi\in\Pi(\mathbf{M})(\mathbf{M})} \min_{(s,a)\in\mathbf{M}} \min\{|\Delta_{s,a}(\mathbf{M}_\pi)| : |\Delta_{s,a}(\mathbf{M}_\pi)| > 0\}$*. The regret of IMED-RL is upper bounded as*

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M}) - \varepsilon\Gamma_s(\mathbf{M})}\right)\log T + O(1), \tag{13}$$

where $\Gamma_s(\mathbf{M})$ *is defined in Proposition 23 and the O(1) is made explicit by summing terms from Proposition 23 and Proposition 30. A Taylor expansion allows to write the regret upper bound as*

$$\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T) \leqslant \left(\sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})}\right)\log T + O\left((\log T)^{10/11}\right). \tag{14}$$

The Taylor expansion is a direct application of Equation 7, Corollary 4 of Honda and Takemura (2015).

Appendix E. Assumptions

In this section, we discussion a variant of the considered setup when the support of transitions is considered known, and then a possible relaxation of the ergodic assumption.

Known support of transitions As quickly explained in the main article, when the support of transition is known, the infimum in sub-optimality cost $\underline{\mathbf{K}}_{s,a}$ defined by equation 8 is redefined as one over the set $\{q \in \mathcal{P}(\mathcal{S}) : \mathtt{Supp}(q) = \mathtt{Supp}(\mathbf{p}(\cdot|s,a))\}$, modifying the lower bound. Without the knowledge of the support,

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \inf_{\substack{\nu \in \mathcal{F}_{s,a} \\ q \in \mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a), \nu \otimes q\right) \,:\, \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma \right\},$$

and with the knowledge of the support,

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \inf_{\substack{\nu \in \mathcal{F}_{s,a} \\ q \in \mathcal{P}(\mathcal{S}) : \mathtt{Supp}(q) = \mathtt{Supp}(\mathbf{p}(\cdot|s,a))}} \left\{ \mathrm{KL}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a), \nu \otimes q\right) \,:\, \varphi_{\mathbf{M}}(\nu \otimes q) > \gamma \right\}.$$

Hence, two similar but different lower bounds can be derived depending on whether or not, one assumes to know the support $\mathtt{Supp}(\mathbf{p})$ of the transitions. In both cases, it can be written

$$\liminf_{T \to \infty} \frac{\mathcal{R}_\pi(\mathbf{M},T)}{\log T} \geqslant \sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})},$$

where $\mathcal{C}(\mathbf{M}) = \left\{(s,a) : 0 < \underline{\mathbf{K}}_{s,a}(\mathbf{M}) < \infty\right\}$, the set of critical state-action pairs, depends on the made hypothesis. Since the lower bound obtained with the knowledge of the support is smaller than without this knowledge, it is a priori not trivial that an algorithm originally designed for the case when support is unknown can indeed exploit this knowledge. Fortunately, due to form of the IMED-RL, it is enough to use the same restriction on the set $\{q \in \mathcal{P}(\mathcal{S}) : \mathtt{Supp}(q) = \mathtt{Supp}(\mathbf{p}(\cdot|s,a))\}$ in the definition of the index to leverage this knowledge. The resulting algorithm slightly differs from IMED-RL and it can be checked easily that the regret analysis for this modified version can be done similarly to IMED-RL, and leads to algorithm matching the asymptotic lower bound knowing the support of transitions. Please refer to Subsection F.1 to see how the hereafter term denoted B is modified by the knowledge of the support.

Bounded support Similarly, one can modify the Assumption 3 to be one with a bounded reward assumption. In this case, Theorem 5 of Honda and Takemura (2015) shows that the Taylor expansion made in the regret upper bound has the aforementioned form in Theorem 6.

Known reward If the reward distributions or mean rewards are assumed to be known, then the infimum in sub-optimality cost $\underline{\mathbf{K}}_{s,a}$ defined by equation 8 can be redefined as

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \inf_{q \in \mathcal{P}(\mathcal{S})} \left\{ \mathrm{KL}\left(\mathbf{p}(\cdot|s,a), q\right) \,:\, \varphi_{\mathbf{M}}(\mathbf{r}(s,a) \otimes q) > \gamma \right\},$$

and yet another lower bound and corresponding IMED-RL algorithm can be derived from it.

These three paragraphs show that IMED-RL is quite versatile regarding assumptions that can be made on the MDP.

Communicating MDPs and ε-soft policies The ergodic assumption can be limiting in practice, since most common MDPs are not ergodic but only communicating. Interestingly, in a communicating MDP, every stochastic policy $\pi : s \in \mathcal{S} \mapsto \pi(\cdot|s) \in \mathcal{P}(\mathcal{A}_s)$, with full-support (that is $\mathtt{Supp}(\pi(\cdot|s)) = \mathcal{A}_s$ for each $s \in \mathcal{S}$) is ergodic. In particular, the uniform policy is ergodic. Also, ε-soft policies, that satisfy $\pi(a|s) \geqslant \varepsilon$ for all s, a, are ergodic. When restricting to the class of ε-soft policies in a communicating MDPs, it seems that modifying IMED-RL to be also ε-soft should lead to a strategy competitive with an optimal ε-soft policy. For $\varepsilon < 1/|\mathcal{A}_s|$, the modification is to sample the chosen action with probability $1 - (|\mathcal{A}_s| - 1)\varepsilon$ and any other action with probability ε. Now, a precise analysis of this modification is postponed to further work, and going beyond this case to handle the full-blown communicating assumption seem to require other ideas, especially since the lower bound for non-ergodic MDPs is expected to be much different from that of ergodic MDPs.

Appendix F. Numerical Experiments

In this section, we first discuss a few implementation details of the `IMED-RL` index, then present additional numerical experiments as well as some extensions.

F.1 Solving the optimization problem $\mathbf{K}_{s,a}$

Although the `IMED` index involves only a single optimization problem (unlike `KL-UCB` that requires two), computing the Kullback-Leibler projection is not obvious in general. The same remark holds for `IMED-RL`. Luckily, inspired from the work of Honda and Takemura (2015), the `IMED-RL` index can be computed easily when restricting to some families of distributions.

In particular, when the set $\mathcal{F}_{s,a}$ of reward distributions is a set of multinomial distributions over a finite set with largest element $< m_{max}$, then the computation of $\mathbf{K}_{s,a}$ can be done easily, owing to the rewriting Theorem 3 from Honda and Takemura (2012).

Lemma 32. *Let $\mathbf{M}$ be an MDP satisfying Assumptions 2, $B = \mathbf{m}_{\max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\mathbf{M}}(s)$ with $\mathbf{m}_{\max}(s,a)$ as in Assumption 3 and rewards satisfying Assumption 1. Then,*

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \max_{0 \leqslant x \leqslant \frac{B}{B-\gamma}} \sum_{\substack{s' \in \mathcal{S} \\ r \in Supp(\mathbf{r}(s,a))}} \mathbf{p}(s'|s,a)\mathbb{P}_{\mathbf{r}(s,a)}(r) \log\left(B - \left(r + \mathbf{b}^{\mathbf{M}}(s') - \gamma\right)x\right)$$

which is a finite convex optimization problem in $x \in \mathbb{R}$.

For the computation of $\underline{\mathbf{K}}_{s,a}(t)$, we use the empirical support of $\mathbf{r}(s,a)$ computed with the gathered samples. Theoretical guarantees comes from the finitness of the support of the original distribution but most importantly, from the fact that is upper bounded by known constant. If $\mathbb{E}_{R\sim\mathbf{r}(s,a), S\sim\mathbf{p}(s,a)}\left(\frac{B-\gamma}{B-R-\mathbf{b}^{\mathbf{M}}(S)}\right) \leqslant 1$, then this optimization problem even has a closed form formula and the maximum of the right-hand-side is obtained for $x = \frac{1}{B-\gamma}$.

In a run of `IMED-RL`, we compute the solution of the empirical problem the same way,

$$\underline{\mathbf{K}}_{s,a}(t) = \max_{0 \leqslant x \leqslant \frac{B_t}{B_t - \hat{\gamma}_s(t)}} \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathtt{Supp}(\hat{\mathbf{r}}^t(s,a))}} \hat{\mathbf{p}}^t(s'|s,a)\mathbb{P}_{\hat{\mathbf{r}}^t(s,a)}(r) \log\left(B_t - \left(r + \mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}(s') - \hat{\gamma}_s(t)\right)x\right)$$

with $B_t = \mathbf{m}_{\max}(s,a) + \max_{s' \in \mathcal{S}} \mathbf{b}^{\widehat{\mathbf{M}}_t(\mathcal{A}(t))}$.

In the general case, the problem given in Proposition 8 is still convex and can be numerically solved as long as one can correctly approximate an expected value, *i.e.* an integral.

Known support of transition We follow the discussion started in Appendix E. If the support of the transition is known, then the cost is computed as

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\gamma) = \max_{0 \leqslant x \leqslant \frac{B_{s,a}}{B_{s,a}-\gamma}} \sum_{\substack{s' \in \mathtt{Supp}(\mathbf{p}(\cdot|s,a)) \\ r \in \mathtt{Supp}(\mathbf{r}(s,a))}} \mathbf{p}(s'|s,a)\mathbb{P}_{\mathbf{r}(s,a)}(r) \log\left(B_{s,a} - \left(r + \mathbf{b}^{\mathbf{M}}(s') - \gamma\right)x\right)$$

where $B_{s,a} = \mathbf{m}_{\max}(s,a) + \max_{s' \in \mathtt{Supp}(\mathbf{p}(\cdot|s,a))} \mathbf{b}^{\mathbf{M}}(s)$ replacing the initial B in Lemma 32, thus making `IMED-RL` takes the knowledge of the support of the transition into account. Of course, this problem is still a finite convex optimization problem in $x \in \mathbb{R}$.

Lazy updates Numerically, `IMED-RL` benefits from this fast computation and the fact that it employs a Value Iteration in lieu of an Extended Value Iteration for instance used in `UCRL3`. On the other hand, `IMED-RL` updates its policy at each time step, unlike `UCRL3` that proceeds into episodes. On our numerical experiments, the overall running time of `IMED-RL` is only about 5 times that of `UCRL3`, despite updating its policy at each time step. Interestingly, it may be possible to further reduce the numerical complexity of `IMED-RL` by performing *lazy* computation of the indexes after some time. Indeed, by design, with high probability, the potential function $\varphi_{\widehat{\mathbf{M}}(\mathcal{A}(t))}$ is not destined to change nor to be much different from the true $\varphi_{\mathbf{M}}$ once an optimal policy belongs to $\mathcal{A}(t)$. As the number of samples increase, the magnitude of the updates decreases and $\varphi_{\widehat{\mathbf{M}}(\mathcal{A}(t))}$ roughly remains the same, thus allowing the practitioner to perform value iteration every once in a while, when at least one estimate shifted by more than a fraction of the minimal sub-optimality gap for instance. Of course this modification requires to update the regret analysis accordingly.

F.2 Additional experiments

In this section, we detail a few more experiments. In all experiments, we used environments with maximal reward 0.99 and bound $m_{max} = 1$ given to the learner. The code of the experiments will be made fully available in the camera-ready version on a git repository, for reproducibility reasons. Experiments are conducted using 256 replications (independent run), with horizon specified in case.

River-swim We consider one experimentation with a river-swim with 25 states. River-swim environments are sometimes considered hard instances for strategies such as `PSRL`, as the reward signal is sparse. We observe in Figure 4 that indeed `PSRL` struggles in such an environment. The three other strategies work well, with some advantage for `IMED-RL` in the long run.

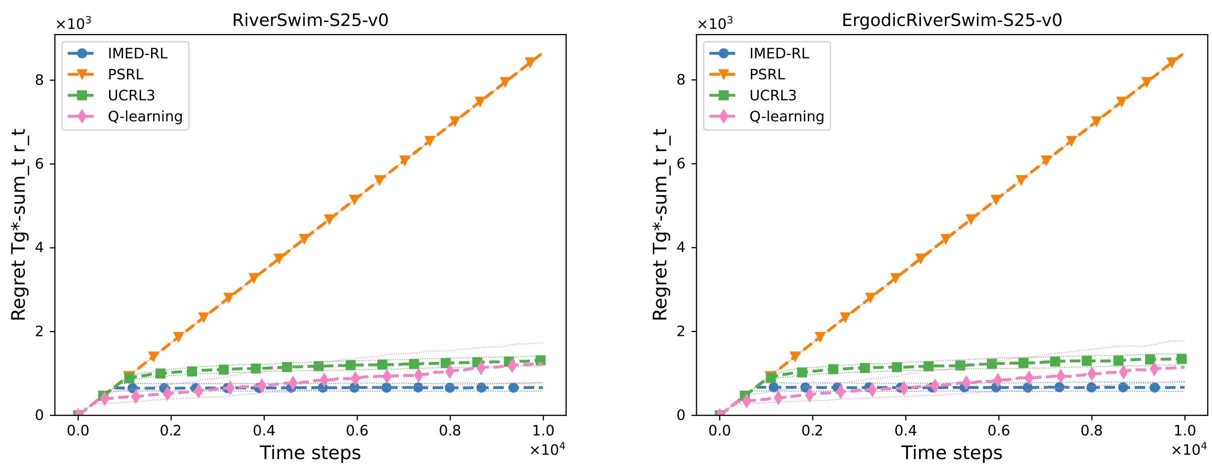

Figure 4: Average regret and quantiles (0.25 and 0.75) curves of algorithms on a standard communicating 25-states RiverSwim (left), and in its ergodic version (right).

For the sake of completeness, we present the average runtime for completing a trajectory of the tested algorithms on both ergodic and non-ergodic RiverSwim.

Table 1: Average runtime (second) on 25-states RiverSwim

	`IMED-RL`	`PSRL`	`UCRL3`	Q-learning
non-ergodic	5.56	0.15	0.42	0.02
ergodic	1.45	0.04	0.23	0.02

Apart from Q-learning, all algorithms seem to benefit a numerical boost from the ergodicity of the environment.

Two-room grid-world The two-room environment we consider in this experiment consists of a 9×11 grid and 4 actions, and is actually a larger state-action space than the four-room MDP considered in the main text, Section 5. Also, it contains a bottleneck state (state in-between the rooms), which is sometimes considered as a hard instance.

Note that since the considered grid-worlds are slippery (frozen-lake style, with 0.1 probability of visiting executing nearby actions), this also means that from the bottleneck state, it is actually possible to enter the bottom room not only with action down, but also left and right. Hence, this MDP does not contain a bottleneck state-action pair. In such environments, although not being ergodic, we expect the `IMED-RL` strategy to work reasonably well, which is confirmed by the experiment in Figure 5.

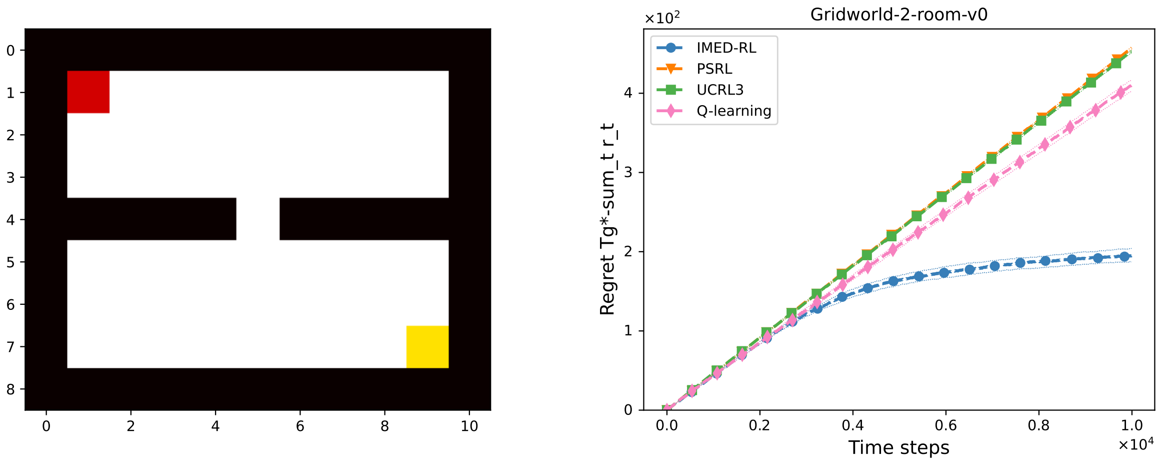

Figure 5: A two room environment with size 9×11 and 4 actions (left), and average regret and quantiles (0.25 and 0.75) curves of algorithms (right).

Another grid-world We further provide below complementary experiments in Figure 6 and Figure 7 with other randomly generated frozen-lake grid-worlds with a unique goal state. The learner jumps to a random initial state each time the goal is reached. The frozen lake part is implemented as slippery actions, where for instance choosing action up has some small probability to move the learner also left or right, or action left has some probability to move the learner up or down, as long as there is no wall (note also that they are coded as toric environments). Although these environments are not ergodic but only communicating. we can observe the striking performance of `IMED-RL` against the state-of-the-art `UCRL3` or related `PSRL` and Q-learning strategies. Note that these other strategies eventually learn as well, but for larger time horizon. In Figure 7, we did not report `UCRL3` and `PSRL` as their computation time were prohibitive compare to `IMED-RL` and Q-learning in this setup.

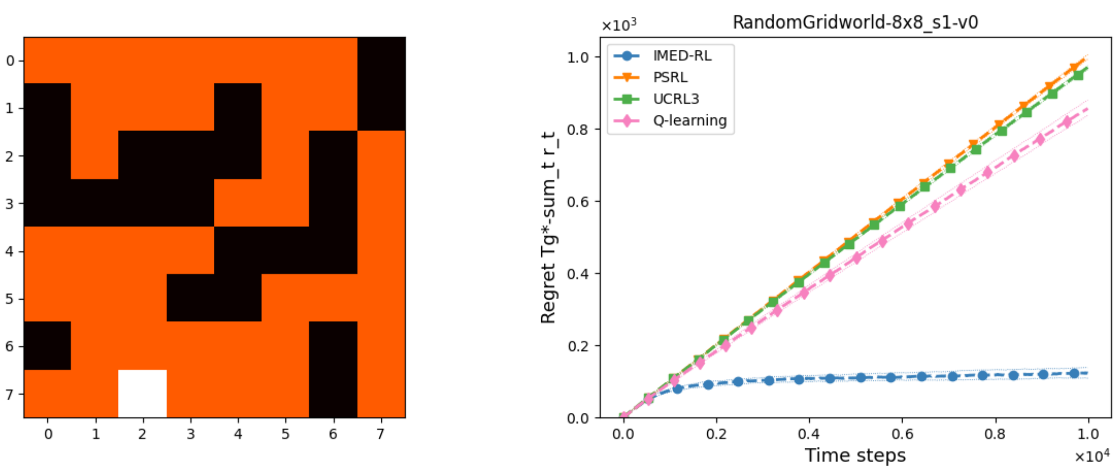

Figure 6: Average regret and quantiles (0.25 and 0.75) curves of algorithms (right) in a randomly generated grid-world (8x8 grid, 4 actions) with reward 0.99 in white state (right).

We present again the average runtime for completing a trajectory of the tested algorithms on such a grid-world environment.

Table 2: Average runtime (second) on 8×8 grid-world

IMED-RL	PSRL	UCRL3	Q-learning
1.82	0.75	6.36	0.03

We can see that the average runtime performances of IMED-RL and UCRL3 were exchanged. Generally, our experiments tends to show that the performances of IMED-RL are quite good on grid-worlds, both from a regret minimization viewpoint and a numerical complexity viewpoint.

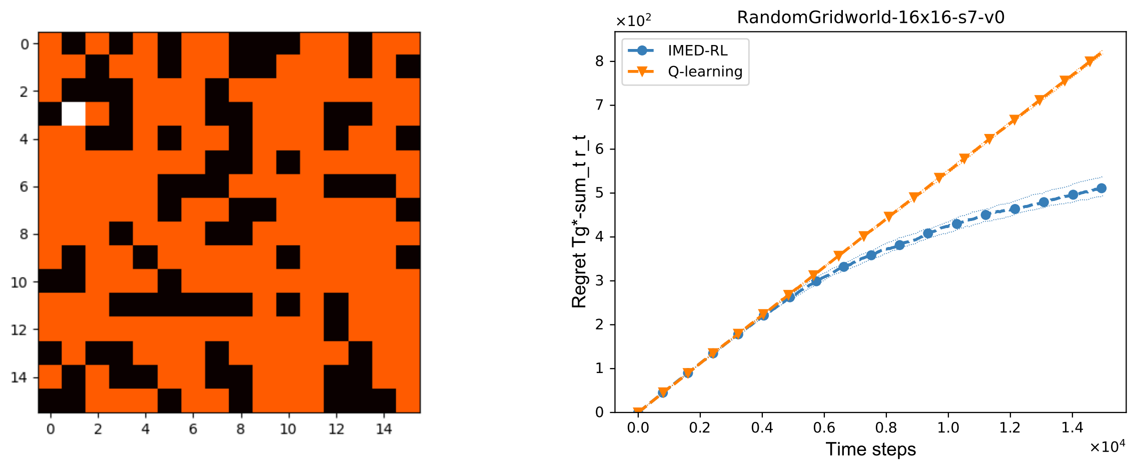

Figure 7: Average regret and quantiles (0.25 and 0.75) curves of algorithms (right) in a randomly generated grid-world (16x16 grid, 4 actions) with reward 0.99 in white state (right).

A reward-rich environment While the previous examples were considering environments with a sparse reward signal, it is interesting to test the behavior of the algorithm in other types of environments. In the following experiment, we consider a reward-rich environment, where about 80% of state-action pairs generate a reward of at least 0.4 (and the maximal reward is 0.99). Such environments are known to favor the PSRL strategy as well as optimistically initialized strategies, that benefit from a reduced burn-in phase thanks to their prior. In Figure 8, we observe that PSRL indeed performs better on this environment than on reward-scarce environments, see Figure 4), and performs on par with Q-learning but IMED-RL still eventually outperforms all other three strategies in this environment (Q-learning algorithm is initialized with $\gamma = 0.99$ and initial value $1/(1-\gamma)$ in each state). When the MDP is modified to have minimal pass $\mathbf{p}(s'|s,a) \geqslant 0.01$ for each s, a, s', the performance of IMED-RL improves and becomes more stable (as well as that of other strategies), as seen in Figure 8-right.

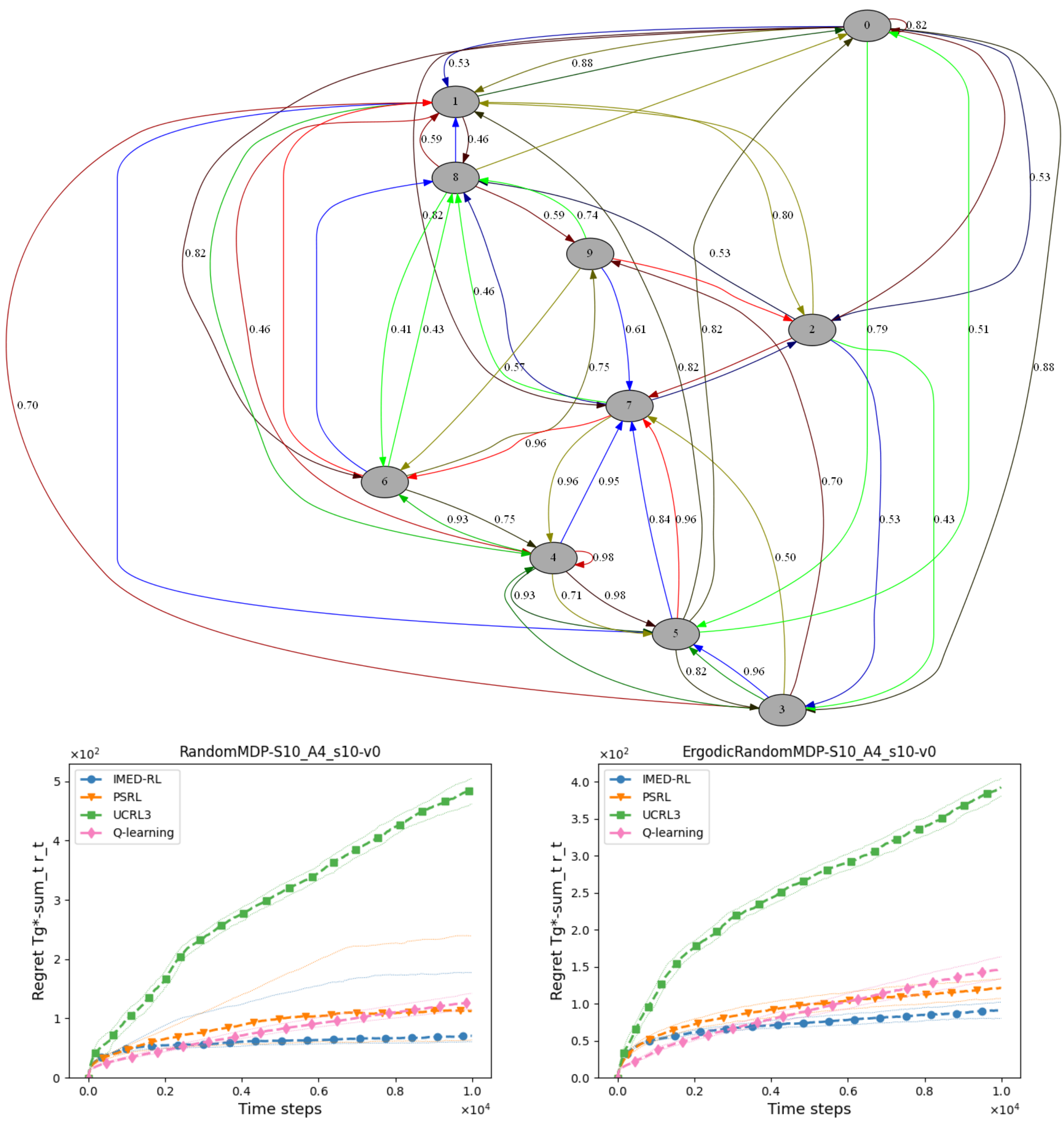

Figure 8: Average regret and quantiles (0.25 and 0.75) curves of algorithms (in log-scale) in a reward-rich environment (10 states, 4 actions) where 80% of state-action pairs give reward of at least 0.4. Right: regret in the ergodic version of the MDP.

A nasty case In order to better understand the limitation of the `IMED-RL` algorithm, we tried (but did not succeed) to craft an environment that would make the `IMED-RL` algorithm fail. The analysis reveals that we should consider a non-ergodic MDP for this purpose. Importantly, the index for pair (s, a) is based on building a modified MDP with unmodified reward and transitions for pairs different than (s, a), which is a feature coming from the ergodic property. However, in a non-ergodic MDP, an optimal policy and a policy playing a in state s may have different recurrent classes, say class $\star$ and $\star_a$. It is not difficult to show that when all paths from a state in $\star$ to a state in $\star_a$ must contain (s, a), that is (s, a) is a bottleneck pair, then changing the MDP only in pair (s, a) to build a "confusing instance" isn't sound anymore, hence the construction of the `IMED-RL` index is no longer justified in such cases. Inspired from this intuition, we build in Figure 9 a specific nasty MDP with such a bottleneck state-action pair, separating two cycles with close value. We further remark that this structure, two promising cycles at two ends of a chain with less rewards in between, may induce an "oscillation" of a learning agent between the two cycles, paying the cost of the travel along the chain each time it "decides" to change cycle. We observe that the quantile tube of `IMED-RL` is larger than before an indeed indicates more struggles but not enough that the `IMED-RL` fails the task. Still, we remark a small advantage of `PSRL` over `IMED-RL` in this environment. Note that the environment is reward rich with rewards close to 1, which also favors `PSRL` and Q-learning with optimistic initialization.

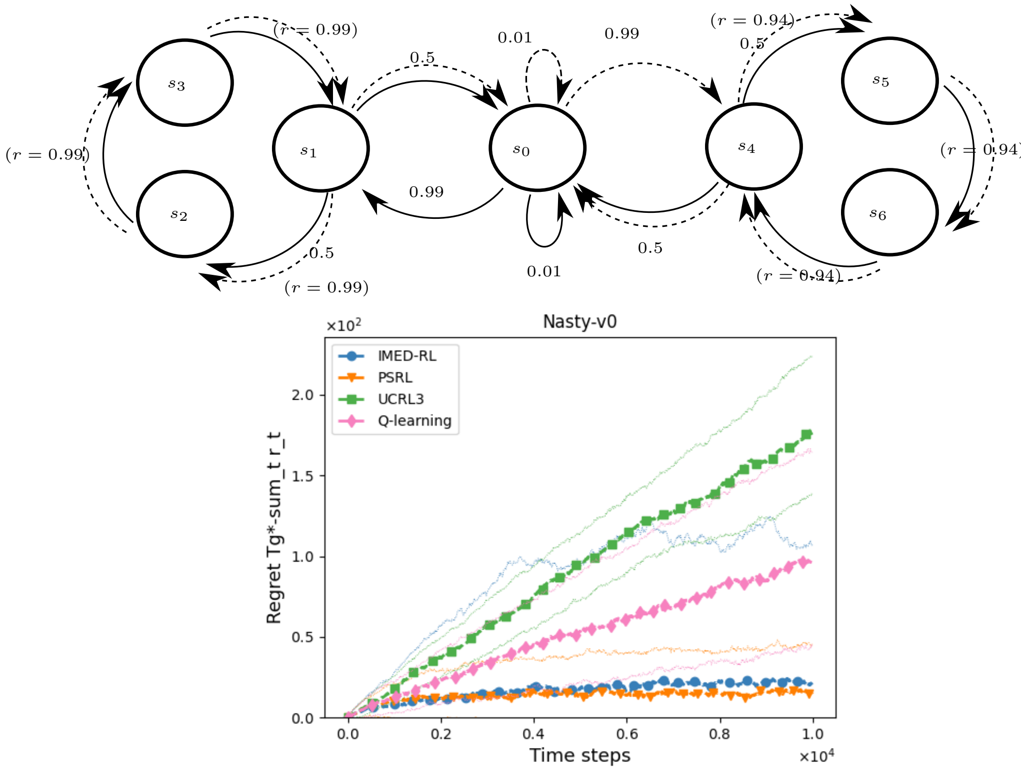

Figure 9: Average regret and quantiles (0.25 and 0.75) curves of algorithms (right) in a nasty environment with two cycles separated by a bottleneck action.

Conclusion The numerical experiments presented in the main paper and this section show that the `IMED-RL` algorithm seems to deliver the promise of the premise. On ergodic environment, its regret is empirically very low and, in the long run, always the smallest in tested environment (and horizon). As explained in this section, it is pleasing to see that its numerical guarantees seems to go beyond the ergodic assumption with particularly good performances in grid-world. In order to better understand the role of the made assumptions, we specifically tried to design a "nasty case" built to make `IMED-RL` fail. The fact that `IMED-RL` still has very good performances in this environment may indicates that theoretical guarantees can be obtained beyond the ergodic assumption and paves the way for future work.

Appendix G. A note on the seminal paper of Burnetas and Katehakis (1997)

In this last appendix, we discuss the subtle but key modification that we made to the notion of *skeleton* introduced in the seminal paper of Burnetas and Katehakis (1997) and defined, for each state s and time t, by,

$$\mathcal{A}_s^{BK}(t) = \left\{ a \in \mathcal{A}_s : N_{s,a}(t) \geqslant \log^2 \left(N_s(t)\right) \right\}. \tag{83}$$

In contrast, the skeleton used in `IMED-RL` is defined replacing the sum $N_s(t) = \sum_{a' \in \mathcal{A}_s} N_{s,a'}(t)$ with a maximum as follows

$$\mathcal{A}_s(t) = \left\{ a \in \mathcal{A}_s : N_{s,a}(t) \geqslant \log^2 \max_{a' \in \mathcal{A}_s} \left(N_{s,a'}(t)\right) \right\}.$$

Correctness The restricted MDP defined by `IMED-RL`, $\mathbf{M}_{\mathcal{A}(t)}$, is well defined in the sense that at least one action is available in each state, that is for all t, for all s, $\mathcal{A}_s(t) \neq \emptyset$. On the other hand, especially at the beginning, $\mathcal{A}_s^{BK}(t)$ could very well be empty. Without saying anything about their algorithm, we just highlight that[10] it must explore all actions in each state at least once before proceeding with non-trivial allocation (this is because the index is $-\infty$ when an arm has not been pulled). Suppose that there are 4 actions in a state s. After the first 3 visits in s, whatever the current time t, $N_s(t) = 3$, $N_{s,a}(t) \leqslant 1$ (as it is 0 for the only unsampled action or 1 for the three others). Because $\log^2 3 \simeq 1.2 > 1$, the skeleton at state s is hence empty and therefore, no action belong to the skeleton. Note that this situation does not happen when using the definition of skeleton used by `IMED-RL`, since $x \geqslant \ln^2(x)$ for all $x \geqslant 0.5$ and at least one action must be sampled ($N_{s,a}(t) \geqslant 1$). In this case, the behaviour of the algorithm presented in the paper of Burnetas and Katehakis (1997) is undefined as it is not specified how to compute the bias and gain on the lacking restricted MDP. Now, in and MDP with a larger number of actions, say 100, the same argument shows that between the 3^{rd} and 100^{th} visit of state s, the skeleton at s is empty and the behaviour undefined. This means that if there are only 20 states in the MDP, the behaviour of the algorithm is undefined for at least about $|\mathcal{S}| \times (A - \log^2(A)) \simeq 2000$ steps (and possibly much more, since one would need all states to be visited about $A - \log^2(A)$ time and it is unlikely that all states are visited equally often).

Incoherence The skeleton as defined in (83) is "incoherent" in the sense that actions may be removed from it for no "justified" reason. In the worst case, all actions may be removed in one step. Assume a state s with 2 actions, one having been sampled 3 times and the other 2 times, *i.e.* $N_s(t) = 5$. Because $2 < \log^2 5 \simeq 2.6 < 3$, one action belongs to the skeleton and the other does not. Assume that the action that have been sampled 2 times is now sampled at time $k > t$. Then both actions have been sampled 3 times, $N_s(k) = 6$ and the skeleton at s is now empty since $\log^2 6 \simeq 3.2 > 3$. While this kind of behaviour disappear for large number of samples, it is not desirable in finite time and introduces incoherence that makes the algorithm undefined and the learning less efficient if we were to resolve undefined behaviour by random choices.

Forced exploration Because of their definition of skeleton, forced exploration is necessary in the analysis of Burnetas and Katehakis (1997) meaning that their algorithm is not purely based on a computed index. While forced exploration and tracking is not inherently an unwanted feature, we think that it should be avoided when possible, hence leaning towards our `IMED-RL` skeleton.

Measuring accuracy The skeleton is used to build a restricted MDP on which the gain and bias can be controlled. This control is due to the fact that, on the skeleton, state-action pairs have been sampled enough. In each state, we are interested in actions with the largest number of pulls amongst all available actions. The most sampled action in each state should therefore obviously belong to the skeleton. Furthermore, it seems natural that the skeleton at a state does not change if the maximal precision in that state, given by the action that has been sampled the most in that state, does not change. This is mainly the rationale behind our subtle but key modification of the notion of skeleton.

10. as it is the case for all learning algorithm without prior information