Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space

Jonatha Anselmi, Bruno Gaujal, Louis-Sébastien Rebuffi

Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space

Jonatha Anselmi
jonatha.anselmi@inria.fr
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.

Bruno Gaujal
bruno.gaujal@inria.fr
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.

Louis Sébastien Rebuffi
louis-sebastien.rebuffi@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.

Abstract

In this paper, we revisit the regret of undiscounted reinforcement learning in MDPs with a birth and death structure. Specifically, we consider a controlled queue with impatient jobs and the main objective is to optimize a trade-off between energy consumption and user-perceived performance. Within this setting, the *diameter* D of the MDP is $\Omega(S^S)$, where S is the number of states. Therefore, the existing lower and upper bounds on the regret at time T, of order $O(\sqrt{DSAT})$ for MDPs with S states and A actions, may suggest that reinforcement learning is inefficient here. In our main result however, we exploit the structure of our MDPs to show that the regret of a slightly-tweaked version of the classical learning algorithm UCRL2 is in fact upper bounded by $\tilde{\mathcal{O}}(\sqrt{E_2AT})$ where E_2 is related to the weighted second moment of the stationary measure of a reference policy. Importantly, E_2 is bounded independently of S. Thus, our bound is asymptotically independent of the number of states and of the diameter. This result is based on a careful study of the number of visits performed by the learning algorithm to the states of the MDP, which is highly non-uniform.

1 Introduction

In the context of undiscounted reinforcement learning in Markov decision processes (MDPs), it has been shown in the seminal work [11] that the total regret of any learning algorithm with respect to an optimal policy is lower bounded by $\Omega(\sqrt{DSAT})$, where S is the number of states, A the number of actions, T the time horizon and D the *diameter* of the MDP. Roughly speaking, the diameter is the mean time to move from any state s to any other state s' within an appropriate policy. In the literature, several efforts have been dedicated to approach this lower bound. As a result, learning algorithms have been developed with a total regret of $\tilde{\mathcal{O}}(DS\sqrt{AT})$ in [11], $\tilde{\mathcal{O}}(D\sqrt{SAT})$ in [3] and even $\tilde{\mathcal{O}}(\sqrt{DSAT})$ according to [21, 25]. These results may give a sense of optimality since the lower bound is attained up to some universal constant. However, lower bounds are based on the minimax approach, which relies on the worst possible MDP with given D, A and S. This means that when a reinforcement learning algorithm is used on a given MDP, one can expect a much better performance.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

One way to alleviate the minimax lower bound is to consider *structured reinforcement learning*, or equivalently MDPs with some specific structure. The exploitation of such structure may yield more efficient learning algorithms or tighter regret analyses of existing learning algorithms. In this context, a first example is to consider *factored* MDPs [6, 9], i.e., MDPs where the state space can be factored into a number of components; in this case, roughly speaking, $S = K^n$ where n is the number of "factors" and K is the number of states in each factor. The regret of learning algorithms in factored MDPs has been analyzed in [20, 17, 24, 14] and it is found that the S term of existing upper bounds can be replaced by nK. A similar approach is used in [8] to learn the optimal policy in stochastic bandits with a regret that is logarithmic in the number of states. There is also a line of research works that exploit the parametric nature of MDPs. Inspired by parametric bandits, a d-linear additive model was introduced in [12], where it is shown that an optimistic modification of Least-Squares Value Iteration, see [15], achieves a regret over a finite horizon H of $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ where d is the ambient dimension of the feature space (the number of unknown parameters). In this case, the regret does not depend on the number of states and actions and the diameter is replaced by the horizon. A discussion about the inapplicability of this approach to our case is postponed to Section 4.2.

Learning in Queueing Systems. The control of queueing systems is undoubtedly one of the main application areas of MDPs; see, e.g., [16, Chapters 1–3] and [13]. Within the rich literature of structured reinforcement learning however, few papers are dedicated to reinforcement learning in queueing systems, see [22, Section 5], and this motivates us to examine the total regret in this context. Typical control problems on queues have the following distinguishing characteristics:

1. *No discount*. Discounting costs or rewards is common practice in the reinforcement learning literature, especially in Q-learning algorithms [19]. However, in queues one is typically interested in optimizing with respect to the average cost.
2. *Large diameter.* Queueing systems are usually investigated under a drift condition that makes the system *stable*, i.e., positive recurrent. This condition implies that some states are hard to reach. In fact, for many queueing control problems, the diameter D is exponential in the size of the state space. Even in the simple case of an M/M/1 queue with a finite buffer, or equivalently a birth–death process with a finite state space and constant birth and death rates, the diameter is exponential in the size of the state space.
3. *Structured transition matrices.* Queueing models describe how jobs join and leave queues, and this yields bounded state transitions. As a result, MDPs on queues have sparse and structured transition matrices.

The regret bounds discussed above and item 2 may suggest that the total regret of existing learning algorithm, when applied to queueing systems, is large. However, they often work well in practice and this bring us to consider the following question: *When the underlying MDP has the structure of a queueing system, do the diameter D or the number of states S actually play a role in the regret?*

Our Contribution. In this paper, we examine the previous question with respect to the class of control problems presented in [1]. Specifically, an infinite sequence of jobs joins a service system over time to receive some processing according to the first-come first-served scheduling rule; the system can buffer at most $S - 1$ jobs and in fact it corresponds to an M/M/1/S-1 queue. In addition, each job comes with a deadline constraint, and if a job is not completed before its deadline, then it becomes obsolete and is removed from the system. The controller chooses the server processing speed and the objective is to design a speed policy for the server that minimizes its average energy consumption plus an obsolescence cost per deadline miss. Although this may look quite specific, this problem captures the typical characteristics of a controlled queue: i) the transition matrix has the structure of a birth and death process with jump probabilities that are affine functions of the state and ii) the reward is linear in the state and convex in the action. For any MDP in this class, defined in full details in Section 3, we show that the diameter is $D = \Omega(S^{S-2})$; see Appendix B.3. Thus, without exploiting the particular structure of this MDP, the existing lower and upper bounds do not justify the reason why standard learning algorithms work efficiently here.

We provide a slight variation of the learning algorithm UCRL2, introduced in [11], and show in our main result that the resulting regret is upper bounded by $\tilde{\mathcal{O}}(\sqrt{E_2AT})$ where E_2 is a term that depends on the stationary measure of a reference policy defined in Section 3.1. Importantly, E_2 does not depend on S. Thus, efficient reinforcement learning can be achieved independently of the number

of states by exploiting the stationary structure of the MDP. Let us provide some intuition about our result. First, one may think that any learning algorithm should visit each state a sufficient number of times, which justifies why the diameter of an MDP appears in existing regret analyses. However, this point of view does not take into account the fact that the value of an MDP is the scalar product of the reward and the stationary measure of the optimal policy. If this stationary measure is "highly non-uniform", then some states are rarely visited under the optimal policy and barely contribute to the value. In this case, we claim that the learner may not need to visit the rare states that often to get a good estimation of the value, and thus it may not need to pay for the diameter.

2 Reinforcement Learning Framework

We consider a unichain Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r)$ in discrete time where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ the finite action space, P the transition probabilities and r the expected reward function [16]. Let also $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$ where $|\cdot|$ is the set cardinality operator. The model-based reinforcement learning problem consists in finding a *learning algorithm*, or learner, that chooses actions to maximize a cumulative reward over a finite time horizon T. At each time step $t \in \mathbb{N}$, the system is in state $s_t \in \mathcal{S}$ and the learner chooses an action $a_t \in \mathcal{A}$. When executing a_t, the learner receives a random reward $r_t(s_t, a_t)$ with mean $r(s_t, a_t)$ and the system moves, at time step $t+1$, to state s' with probability $P(s'|s_t, a_t)$. The learning algorithm does not know the MDP M except for the sets $\mathcal{S}$ and $\mathcal{A}$.

For simplicity, in the following we consider *weakly communicating* MDPs. Since we will be interested in the long-run average cost, this will let us remove the dependence on the initial state for several quantities of interest.

2.1 Undiscounted Regret

Given an MDP M, let $\Pi := \{\pi : \mathcal{S} \to \mathcal{A}\}$ denote the set of stationary and deterministic policies and let

$$\rho(M, \pi) := \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(s_t, \pi(s_t))] \tag{1}$$

denote the average reward induced by policy π. Since M has finite state and action spaces, we notice that i) The limit in (1) always exists, ii) It does not depend on the initial state s_0 when M is unichain [16] and iii) The restriction to stationary and deterministic policies is not a loss of optimality [16, Theorem 8.4.5].

Let also $\rho^* := \rho^*(M) := \max_{\pi \in \Pi} \rho(M, \pi)$ be the optimal average reward.

Definition 2.1 (Regret)**.** The *regret* at time T of the learning algorithm $\mathbb{L}$ is

$$\mathrm{Reg}(M, \mathbb{L}, T) := T\rho^*(M) - \sum_{t=1}^{T} r_t. \tag{2}$$

The regret 2 is a natural benchmark for evaluating the performance of a learning algorithm. In [11], a *universal* lower bound on $\mathrm{Reg}(M, \mathbb{L}, T)$ has been developed in terms of the *diameter* of the underlying MDP.

Definition 2.2 (Diameter of an MDP)**.** Let $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary policy of M with initial state s. Let $T(s'|M, \pi, s) := \min\{t \geq 0 : s_t = s' | s_0 = s\}$ be the random variable for the first time step in which s' is reached from s under π. Then, we say that the *diameter* of M is

$$D(M) := \max_{s \neq s'} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}\left[T(s'|M, \pi, s)\right].$$

It should be clear that the diameter of an MDP can be large if there exist states that are hard to reach. Within the set of structured MDPs considered in this paper, this will be the case and we will show that $D = \Omega(S^{S-2})$. The following result shows that all learning algorithms have a regret that eventually increase with $\sqrt{D}$.

Theorem 2.3 (Universal lower bound [11])**.** *For any learning algorithm $\mathbb{L}$, any natural numbers $S, A \geq 10$, $D \geq 20 \log_A S$, and $T \geq DSA$, there is an MDP M with S states, A actions, and diameter D such that for any initial state $s \in \mathcal{S}$,*

$$\mathbb{E}[\mathrm{Reg}(M, \mathbb{L}, T)] \geq 0.015\sqrt{DSAT}. \tag{3}$$

In view of this result, the diameter of an MDP and its state space appear to be critical parameters for evaluating the performance of a learning algorithm.

2.2 The UCRL2 Algorithm

We now focus on UCRL2, a classical reinforcement learning algorithm introduced in [11] that is a variant of UCRL [2]. While more efficient algorithms have been proposed for the general case (see for example [3, 21]), we will show that UCRL2 already achieves a very low regret, namely $\tilde{\mathcal{O}}(\sqrt{AT})$, independent of S so using more refined algorithms can only bring marginal gains.

UCRL2 is based on *episodes*. For each episode k, let t_k denote its start time. For each state s and action a, let $\nu_k(s,a)$ denote the number of visits of (s,a) during episode k and let $N_t(s,a) := \#\{\tau < t : s_\tau = s, a_\tau = a\}$ denote the number of visits of (s,a) until timestep t. Let $\mathcal{M}_k$ be the confidence set of MDPs with transition probabilities $\tilde{p}$ and rewards $\tilde{r}$ that are "close" to the empirical MDP at episode k, $\hat{p}_k$ and $\hat{r}_k$, i.e., $\tilde{p}$ and $\tilde{r}$ satisfy

$$\forall (s,a), \quad |\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq r_{\max} \sqrt{\frac{7 \log(2SAt_k/\delta)}{2 \max\{1, N_{t_k}(s,a)\}}} \tag{4}$$

$$\forall (s,a), \quad \|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S \log(2At_k/\delta)}{\max\{1, N_{t_k}(s,a)\}}} \tag{5}$$

With these quantities, a pseudocode for UCRL2 is given in Algorithm 1. We notice that UCRL2 relies on Extended Value Iteration (EVI), that is a variant of the celebrated Value Iteration (VI) algorithm [16]; for further details about EVI, we point the reader to [11, Section 3.1]. Let us comment on how UCRL2 works. There are three main steps. First, at the start of each episode, UCRL2

Algorithm 1: The UCRL2 algorithm.

Input: A confidence parameter $\delta \in (0,1)$, $\mathcal{S}$ and $\mathcal{A}$.
Output: .

Set $t := 1$ and observe s_1
for *episodes* $k = 1, 2, \ldots$ **do**
  Compute the estimates $\hat{r}(s,a)$ and $\hat{p}_k(s'|s,a)$ as in (7).
  Use "Extended Value Iteration" to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\rho(\tilde{M}_k, \tilde{\pi}_k) \geq \max_{M' \in \mathcal{M}_k, \pi} \rho(M', \pi) - \frac{1}{\sqrt{t_k}} \tag{6}$$

  while $\nu_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_{t_k}(s_t, \tilde{\pi}_k(s_t))\}$ **do**
    Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t and observe s_{t+1};
    $\nu_k(s_t, a_t) := \nu_k(s_t, a_t) + 1$;
    $t := t + 1$;
  end
end

computes the empirical estimates

$$\hat{r}_k(s,a) := \frac{\sum_{t=1}^{t_k - 1} r_t \mathbb{1}_{\{s_t = s, a_t = a\}}}{\max\{1, N_{t_k}(s,a)\}}, \qquad \hat{p}_k(s'|s,a) := \frac{\sum_{t=1}^{t_k - 1} \mathbb{1}_{\{s_t = s, a_t = a, s_{t+1} = s'\}}}{\max\{1, N_{t_k}(s,a)\}} \tag{7}$$

of the reward and probability transitions, respectively, where $\mathbb{1}_E$ is the indicator function of E. Then, it applies Extended Value Iteration (EVI) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that (6) holds true. Finally, it executes policy $\tilde{\pi}_k$ until it finds a state-action pair (s,a) whose count within episode k is greater than the corresponding state-action count prior to episode k.

3 Controlled Birth and Death Processes for Energy Minimization

Now, we focus on a specific class of MDPs that has been introduced in [1], which provides a rather general example of a controlled birth and death process with convex costs on the actions and linear rates. We will denote by $\mathcal{M}$ the set of MDPs with the structure described below. The MDPs in $\mathcal{M}$ have been proposed to represent a Dynamic Voltage and Frequency Scaling (DVFS) processor executing jobs with soft obsolescence deadlines. Here, jobs arrive according to a Poisson process with rate $\lambda \in [0, \lambda_{\max}]$ in a buffer of size $S-1$. If the buffer is full and a job arrives, then the job is rejected. Each job has a deadline and a size, i.e., amount of work, which are exponentially distributed random variables with rates $\mu \in [0, \mu_{\max}]$ and one, respectively. Job deadlines and sizes are all independent random variables. If a job misses its deadline, which is a real time constraint activated at the moment of its arrival, it is removed from the queue without being served and a cost C is paid. The processor serves jobs under any work-conserving scheduling discipline, e.g., first-come first-served, with a processing speed that belongs to the finite set $\{0, \ldots, A_{\max}\}$. The objective is to design a speed policy that minimizes the sum of the long term power dissipation and the cost induced by jobs missing their deadlines. When the processor works at speed $a \in \{0, \ldots, A_{\max}\}$, it processes a units of work per second while its power dissipation is $w(a)$.

After uniformization, it is shown in [1] that this control problem can be modeled as an MDP in discrete time with a "birth-and-death" transition matrix of size S. Specifically, we have an MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ where $\mathcal{S} = \{0, \ldots, S-1\}$, with $s \in \mathcal{S}$ representing the number of jobs in the system, and $\mathcal{A} = \{0, \ldots, A_{\max}\}$, with $a \in \mathcal{A}$ representing the processor speed. Then, the transition probabilities under policy π are given by

$$P_{i,j}(\pi) = \begin{cases} \frac{1}{U}\lambda_i & \text{if } i < S-1 \text{ and } j = i+1 \\ \frac{1}{U}(\pi(i) + i\mu) & \text{if } i > 0 \text{ and } j = i-1 \\ P_{ii} & \text{if } j = i \\ 0 & \text{otherwise,} \end{cases}$$

where $U := \lambda_{\max} + (S-1)\mu_{\max} + A_{\max}$ is a uniformization constant, $P_{ii} = \frac{1}{U}(U - \lambda_i - \mu i - \pi(i))$ and $\lambda_i := \lambda\left(1 - \frac{i}{S-1}\right)$ is the *decaying* arrival rate. We have replaced the constant arrival rate λ by a decaying arrival rate λ_i because we want to learn an optimal policy that does not exploit the buffer size $S-1$; see [1] for further details. For conciseness, Figure 1 displays the transition diagram of the Markov chain induced by policy π.

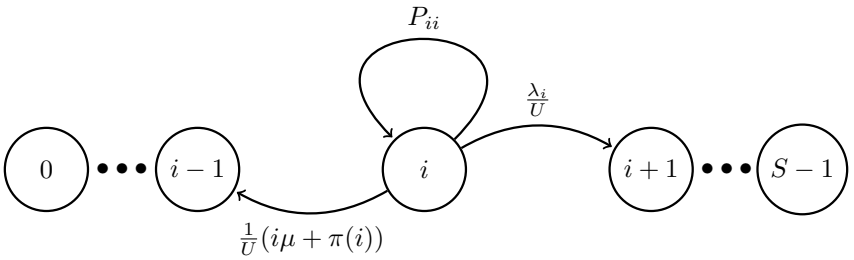

Figure 1: Transition diagram of the Markov chain induced by policy π of an MDP in $\mathcal{M}$.

Finally, the reward is a combination of C, the constant cost due to a departing job missing its deadline and $w(a)$, an arbitrary convex function of a, giving the energy cost for using speed a. The immediate cost $c(s, a)$ in state s under action a is a random variable whose value is $w(a) + C$ with probability $i\mu/U$ (missed deadline) and $w(a)$ otherwise. To keep in line with the use of rewards instead of costs, we introduce a bound on the cost, $r_{\max} := C + w(A_{\max})$ so that the reward in state s under action a is a positive and bounded random variable given by

$$r(s, a) := r_{\max} - c(s, a). \tag{8}$$

As in Section 2.1, $\rho^*(M)$ is the optimal average cost and $\rho(M, \pi)$ is the average cost induced by policy π, where π belongs to the set of deterministic and stationary policies Π. Since the underlying

Markov chain induced by any policy is ergodic, we observe that

$$\rho(M,\pi) = \sum_{s=0}^{S-1} \mathbb{E}[r(s,\pi(s))] m_s^\pi, \tag{9}$$

where m^π is the stationary measure under policy π. In [1], it has been shown that the optimal policy is unique and will be denoted by π^*.

3.1 Properties of $\mathcal{M}$

In the following, we will use the "reference" (or bounding) policy $\pi^0(s) = 0$ for all $s \in \mathcal{S}$, which thus assigns speed 0 to all states. This policy provides a stochastic bound on all policies in the following sense. Let s_t^π be the state under policy π and let $\leq_{st}$ denote the *stochastic order* [18]; given two random variables X and Y on $\mathbb{R}_+$, we recall that $X \leq_{st} Y$ if $\mathbb{P}(X \geq s) \leq \mathbb{P}(Y \geq s)$ for all s.

Lemma 3.1. *Consider an MDP in $\mathcal{M}$. For all t and policy $\pi \in \Pi$, $s_t^\pi \leq_{st} s_t^{\pi^0}$, provided that $s_0^\pi \leq_{st} s_0^{\pi^0}$.*

Proof. (sketch) The proof follows by a simple coupling argument between the two policies. Roughly speaking, each time the Markov chain under π decreases from s to $s-1$ because of the speed $\pi(s)$, it stays in state s under policy π^0. □

Therefore, $\mathbb{P}(s_t^\pi \geq s) \leq \mathbb{P}(s_t^{\pi^0} \geq s)$ for all s and t, which also implies that the respective stationary measures are comparable, i.e., $\sum_{i=s}^{S-1} m_i^\pi \leq \sum_{i=s}^{S-1} m_i^{\pi^0}$.

Let us now consider $H(s)$, the *bias* at state s of the optimal policy π^*, defined by

$$H(s) := \mathbb{E}_{\pi^*}\left[\sum_{t=1}^{\infty} \left(r\left(s_t^{\pi^*}, \pi^*(s_t^{\pi^*})\right) - \rho^*(M)\right) \mid s_0^{\pi^*} = s\right], \quad \forall 0 \leq s \leq S-1, \tag{10}$$

Let also $\partial H(s) := H(s) - H(s-1)$ be the local variation of the bias.

The following result was shown in [1, Lemma 3.8].

Lemma 3.2. *The local variation of the bias, $\partial H(s)$, is negative, decreasing in s, and bounded: $-\partial H(s) \leq \Delta(s)$ with $0 < \Delta(s) \leq C$ for all $1 \leq s \leq S-1$.*

Both m^{π^0} and Δ will play a major role in our analysis of the regret.

3.2 Applying UCRL2 in $\mathcal{M}$

We assume that the bounds $\lambda_{\max}$ and $\mu_{\max}$ are fixed so that $r_{\max}$ is known to the learner. This is a classical assumption, often replaced by assuming that rewards live in $[0,1]$.

In the remainder, we will apply UCRL2 over an MDP in $\mathcal{M}$ with a change in the confidence bounds to take into account the support of P. The confidence bounds in (4) (resp. (5)) are replaced by $r_{\max}\sqrt{\frac{2\log(2At_k)}{\max\{1,N_{t_k}(s,a)\}}}$ (resp. $\sqrt{\frac{8\log(2At_k)}{\max\{1,N_{t_k}(s,a)\}}}$). We also impose that the confidence set $\mathcal{M}_k$ only contains matrices with the same support as P. Removing S in the confidence bounds does help to reduce the regret. However, by using existing analysis, this only removes a factor $\sqrt{S}$ in the regret bound (for example, see [4]).

Finally, note that UCRL2 does not benefit from the parametric nature of the MDPs in $\mathcal{M}$, which is essentially defined by three parameters (λ, μ and C) and the real convex function $w(\cdot)$.

4 Regret of UCRL2 on $\mathcal{M}$

Our objective is to develop an upper bound on the regret of the learning algorithm UCRL2 when applied to MDPs in our class $\mathcal{M}$. The driving idea is to construct a bound that exploits the structure of the stationary measure of all policies, as they all make some states hard to reach, and to control the number of visits to these states to get a new type of bound.

4.1 Main Result

The following theorem gives an upper bound on the regret that does not depend on the classical parameters such as the size of the state space nor on global quantities such as the diameter of the MDP nor the span of the bias of some policy. Instead, the regret bound below mainly depends on the weighted second moment of the stationary measure of the reference policy π^0, which is bounded independently of the size of the state space.

We consider the policy $\pi^{\max}$ such that $\pi^{\max}(s) = A_{\max}$ for all s and $m^{\max}$ its stationary measure.

Let us also recall that m^{π^0} is the stationary measure of the Markov chain under policy $\pi^0(s) = 0$ for all s and that $\Delta : \mathcal{S} \to \mathbb{R}^+$ is a function bounding the local variations of the optimal bias. Let $E_2 := F\, \mathbb{E}_{m^{\pi^0}}\left[(\Delta + r_{\max})^2 \cdot f\right]$ with $f : s \mapsto \frac{\max\{1, s(s-1)\}}{(\Delta(s)+r_{\max})^2}$ and $F := \sum_{s\in\mathcal{S}} f(s)^{-1}$. Here, E_2 is closely related to the second moment of the measure m^{π^0} weighted by the bias variations and the maximal reward.

Theorem 4.1. *Let $M \in \mathcal{M}$. Define $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$.*

$$\mathbb{E}\left[\mathrm{Reg}(M, \mathrm{UCRL2}, T)\right] \leq 19\sqrt{E_2 AT \log(2AT)} + 97 r_{\max} D^2 SA \max\{Q_{\max}, T^{1/4}\} \log^2(2AT). \quad (11)$$

Here, $E_2 \leq 12 r_{\max}^2 \left(1 + \frac{\lambda^2}{\mu^2}\right)$, so that the regret satisfies

$$\mathbb{E}\left[\mathrm{Reg}(M, \mathrm{UCRL2}, T)\right] = \mathcal{O}\left(r_{\max}\sqrt{AT\left(1 + \frac{\lambda^2}{\mu^2}\right)\log(AT)}\right).$$

Before giving a sketch of the proof, let us comment on the bound (11). Although the first term is of order $\sqrt{T}$ with a multiplicative constant independent of S - as desired - the second term, of order $T^{1/4}$, contains very large terms. Its interest however, lies in the novel approach used in the proof that uses the stationary behavior of the algorithm.

4.2 Comparison with Other Bounds

Let us compare our upper bound with the ones existing in the literature, as we claim that ours is of a different nature.

First, let us compare with the bound given in [11], which states that with probability $1 - \delta$, $\mathrm{Reg}(M, T) \leq 34 DS\sqrt{AT \log\left(\frac{T}{\delta}\right)}$ for any $T > 1$. For any $M \in \mathcal{M}$, the diameter grows as S^S (see Appendix B.3), thus this bound is very loose here. More recent works have improved this bound by replacing the term D by the local diameter of the MDP [5]. In Appendix B.3, we show that the local diameter grows again as S^S for any $M \in \mathcal{M}$, and thus these results do not yield significant improvements. Other papers show that the diameter can be replaced by the span of the bias, see [7, 25]. This has a big impact because the span of the bias, for any $M \in \mathcal{M}$, is linear in S (instead of S^S for the diameter); see Appendix B.3. However, this is still not as good as the bound given in Theorem 4.1, which is independent of S.

Now, let us compare with existing bounds for *parametric* MDPs, as mentioned in the introduction. The d-linear additive model, $d < S$, introduced in [12] assumes that $P(\cdot|s,a) = \langle \phi(s,a), \theta(\cdot)\rangle$, where $\phi(s,a)$ is a known feature mapping and θ is an unknown measure on $\mathbb{R}^d$. This form of $P(\cdot|s,a)$ implies that the transition kernel is of rank d. Unfortunately, this property does not hold true in birth and death processes. In fact, the kernel of any $M \in \mathcal{M}$ has almost *full* rank under all policies. The *linear mixture model* introduced in [26] assumes instead that $P(s'|s,a) = \langle \phi(s'|s,a), \theta\rangle$, $\theta \in \mathbb{R}^d$. This is more adapted to our case, which can be (almost) seen as a linear mixture model of dimension $d = 3$. The bound on the discounted regret of the algorithm proposed in [26] is $\mathrm{Reg}(M, T) \leq d\sqrt{T}/(1-\gamma)^2$ where γ is a discount factor. In contrast to our work, this regret analysis holds for *discounted* problems, where we remark that both the diameter and the span are irrelevant. On the other hand, both are replaced by a term of the form $(1-\gamma)^{-2}$, which implies

that the previous bound grows to infinity as $\gamma \uparrow 1$. More recently, a regret bound of $O(d\sqrt{DT})$ has been proven in [23] in the undiscounted case, that is the case considered in our work. However, the algorithm presented in that reference highly depends on the diameter and is unsuitable for MDPs with a birth and death structure.

Finally, our bound depends on the second moment of the stationary measure of a reference policy, i.e., E_2, which can be bounded independently of the state space size. This is structurally different from the ones existing in the literature. We believe that this structure holds as well in a class of MDPs much larger than $\mathcal{M}$. In particular, if m is the stationary measure of some bounding/reference policy, and if the critical quantity $\mathbb{E}_m[\Delta \cdot f]$ is small for a well chosen function f, then the regret of a learning algorithm navigating the MDP should also be small. A deeper analysis is left as future work.

4.3 Sketch of the Proof

Our proof for Theorem 4.1 is technical and is provided in the supplementary material. In this section, we present the main ideas and its general structure. It initially relies on the regret analysis of UCRL2 developed in [11], and the differences are highlighted below. First, we consider the mean rewards and split the regret into episodes to separately treat the cases where the true MDP is in the confidence set of optimistic MDPs $\mathcal{M}_k$ or not. Thus, let $R_k := \sum_{s,a} \nu_k(s,a)(\rho^* - \overline{r}(s,a))$ denote the regret in episode k. This split can be written:

$$\mathbb{E}[\mathrm{Reg}(M,T)] \leq \mathbb{E}[R_{\mathrm{in}}] + \mathbb{E}[R_{\mathrm{out}}],$$

where $R_{\mathrm{in}} := \sum_k R_k \mathbb{1}_{M \in \mathcal{M}_k}$ and $R_{\mathrm{out}} := \sum_k R_k \mathbb{1}_{M \notin \mathcal{M}_k}$.

To control R_{out}, we use, as in [11], the stopping criterion and the confidence bounds. This gives $\mathbb{E}[R_{\mathrm{out}}] \leq r_{\max} S$, so that the regret due to episodes where the confidence regions fail will be negligible next to the main terms. Then, when the true MDP belongs to the confidence region, we use the properties of Extended Value Iteration (EVI) to decompose R_{in} into

$$\underbrace{\sum_{k,s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a))}_{R_{\mathrm{rewards}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\mathrm{bias}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k + 2 r_{\max} \sum_{k,s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}}_{R_{\mathrm{EVI}}},$$

where $\mathbf{v}_k$ is the vector of the state-action counts ν_k's, $\tilde{\mathbf{P}}_k$ and $\tilde{\mathbf{h}}_k$ are respectively the transition matrix and the bias in $\tilde{M}_k$ under policy $\tilde{\pi}_k$, and $\mathbf{d}_k$ is the profile difference between the last step of EVI and the bias (see Appendix A.3).

We now show how to handle R_{rewards}, R_{EVI}, R_{bias}. First, we deal with the terms that do not involve the bias. Using the confidence bound on the rewards (see Appendix A.3.1:

$$R_{\mathrm{rewards}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{12}$$

Now, let us consider R_{EVI}. Since $\mathbf{d}_k$ becomes arbitrarily small after enough iterations of EVI (see Appendix A.1), for $T \geq \frac{e^8}{2AT}$, we get

$$R_{\mathrm{EVI}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{13}$$

The analysis of R_{bias} is different from the one in [11]: While in [11] the bias is directly bounded by the diameter, we can use the variations of the bias to control the regret more precisely. Using $\mathbf{P}_k$, i.e., the transitions in the true MDP under $\tilde{\pi}_k$, R_{bias} is further decomposed into:

$$\underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \mathbf{h}^*}_{R_{\mathrm{trans}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \left(\tilde{\mathbf{h}}_k - \mathbf{h}^*\right)}_{R_{\mathrm{diff}}} + \underbrace{\sum_k \mathbf{v}_k \left(\mathbf{P}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\mathrm{ep}}}.$$

The term R_{ep} can be treated in a similar manner as in [11] by bounding the bias terms with the diameter to apply an Azuma-Hoeffding inequality (see Appendix A.3.5). Here, we obtain

$$\mathbb{E}[R_{\mathrm{ep}}] \leq SAD\, r_{\max} \log_2\left(\frac{8T}{SA}\right).$$

Next, we show in A.3.2 that R_{diff} does not contribute to the main term of the regret. This is one of the hard point in our proof. First, linear algebra techniques are used to bound $||\tilde{\mathbf{h}}_k - \mathbf{h}^*||_\infty$ by $D(2r_{\max}D||\tilde{\mathbf{P}}_k - \mathbf{P}^*||_\infty + ||\tilde{\mathbf{r}}_k - \mathbf{r}^*||_\infty)$. Each norm is then bounded using Hoeffding inequality. This introduces the special quantity $N_{t_k}(x_k, \pi_k(x_k))$ that yields to the worst confidence bound in episode k. Then, an adaptation of McDiarmid's inequality to Markov chains is used to show that $N_{t_k}(x_k, \pi_k(x_k)) \geq (t_{k+1} - t_k)m^{\max}(S-1)/2$ with high probability, where $m^{\max}(S-1)$ is the stationary measure of state $S-1$ under the uniform policy $\pi^{\max}(s) = A_{\max}$. This eventually implies that

$$\mathbb{E}\left[R_{\text{diff}}\right] \leq 96 r_{\max} D^2 SA \max\{Q_{\max}, T^{1/4}\} \log^2(2AT),$$

where $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$.

Then, to deal with the main term R_{trans}, we exploit the optimal bias. The unit vector being in the kernel of $\tilde{\mathbf{P}}_k - \mathbf{P}_k$, we can rewrite:

$$R_{\text{trans}} = \sum_k \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot \left(h^*(s') - h^*(s)\right)$$

and, thus, using the confidence bound and the bounded variations of the bias,

$$R_{\text{trans}} \leq 4\sqrt{2\log\left(2AT\right)} \sum_k \sum_{s,a} \frac{\Delta(s)\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}.$$

We can now aggregate R_{trans}, R_{rewards} and R_{EVI} to compute the main term of the regret (see Appendix A.3.4). Here, the key ingredient is to bound

$$\sum_{k,s,a} \frac{(\Delta(s) + r_{\max})\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}$$

independently of S. This is the second main difference with [11]. Instead of exploring the MDP uniformly, we know that the algorithm will mostly visit the first states of the MDP, regardless of the chosen policy. As shown in [11], for a fixed state s:

$$\mathbb{E}\left[\sum_a \sum_k \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}\right] \leq 3\sqrt{\mathbb{E}\left[N_T(s)\right] A}.$$

Now, instead of summing over the states, we can use properties of stochastic ordering to compare the mean number of visits of a state with the probability measure m^{π^0}; here, we strongly rely on the birth and death structure of the MDPs in $\mathcal{M}$. For any non-negative non-decreasing function $f : \mathcal{S} \to \mathbb{R}^+$, we obtain

$$\mathbb{E}\left[\sum_{s\geq 0} f(s) N_t(s)\right] \leq t \sum_{s \geq 0} f(s) m^{\pi^0}(s). \tag{14}$$

Let us choose $f : s \mapsto \frac{\max\{1, s(s-1)\}}{(\Delta(s) + r_{\max})^2}$ and let $F := \sum_s f(s)^{-1} \leq 3(C + r_{\max})^2$. Let also $E_2 := F\mathbb{E}_{m^{\pi^0}}\left[(\Delta + r_{\max})^2 \cdot f\right]$. Then,

$$\mathbb{E}\left[\sum_k \sum_{s,a} \frac{(\Delta(s) + r_{\max})\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{E_2 AT}.$$

In Appendix A.3.4, we further show that $E_2 \leq 3(C + r_{\max})^2 \left(1 + \frac{\lambda^2}{\mu^2}\right)$. Therefore, for the three main terms, we obtain

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 19\sqrt{E_2 AT \log\left(2AT\right)} \tag{15}$$

and we conclude our proof by combining all of these terms.

5 Conclusions

For learning in a class of birth and death processes, we have shown that exploiting the stationary measure in the analysis of classical learning algorithms yields a $K\sqrt{T}$ regret, where K only depends on the stationary measure of the system under a well chosen policy. Thus, the dependence on the size of the state space as well as on the diameter of the MDP or its span disappears. We believe that this type of results can be generalized to other cases such as optimal routing, admission control and allocation problems in queuing systems, as the stationary distribution under all policies is uneven between the states.

References

[1] Jonatha Anselmi, Bruno Gaujal, and Louis Sébastien Rebuffi. Optimal Speed Profile of a DVFS Processor under Soft Deadlines. *Performance Evaluation*, 152, December 2021.

[2] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[3] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[5] Hippolyte Bourel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[6] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artif. Intell.*, 121(1–2):49–107, aug 2000.

[7] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. 02 2018.

[8] Nicolas Gast, Bruno Gaujal, and Kimang Khun. Reinforcement learning for Markovian bandits: Is posterior sampling more scalable than optimism? Technical Report hal-03262006, HAL-Inria, June 2021.

[9] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19(1):399–468, October 2003.

[10] Ilse C. F. Ipsen and Carl Dean Meyer. Uniform stability of markov chains. *SIAM Journal on Matrix Analysis and Applications*, 15:1061–1074, 1994.

[11] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4):1563–1600, 2010.

[12] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.

[13] Quan-Lin Li, Jing-Yu Ma, Rui-Na Fan, and Li Xia. *An Overview for Markov Decision Processes in Queues and Networks*, pages 44–71. Springer Singapore, Singapore, 2019.

[14] Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Proc. of the 27th Int. Conf. on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 604–612, Montreal, Canada, December 2014. MIT Press.

[15] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2377–2386. JMLR.org, 2016.

[16] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition, April 1994.

[17] Aviv Rosenberg and Yishay Mansour. Oracle-Efficient Reinforcement Learning in Factored MDPs with Unknown Structure. *arXiv:2009.05986 [cs, stat]*, September 2020.

[18] Moshe Shaked and J George Shanthikumar. *Stochastic orders and their applications*. Academic Pr, 1994.

[19] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[20] Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[21] Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities, 2019.

[22] Neil Walton and Kuang Xu. Learning and information in stochastic networks and queues, 2021.

[23] Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR, 28–30 Mar 2022.

[24] Ziping Xu and Ambuj Tewari. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[25] Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function, 2019.

[26] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12793–12802. PMLR, 18–24 Jul 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The abstract and the introduction do provide a concise statement of the original result proved in this paper.
   (b) Did you describe the limitations of your work? [Yes] The limitations of our work are explicit as we described in 3 the subclass of MDPs on which the theorem applies.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work is mainly of mathematical nature and we hope that it has a very little direct societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We confirm that our paper complies with the ethical guidelines of Neurips.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions and the context of the theorem are provided in 2 and 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] A sketch of the proof is given in 4.3 and the complete proof is given in the supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] There is no experiment in this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] There is no experiment in this paper.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] There is no experiment in this paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] There is no experiment in this paper.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We did not work with existing assets or new assets.
   (b) Did you mention the license of the assets? [N/A] We did not work with existing assets or new assets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We did not work with existing assets or new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not work with existing assets or new assets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not work with existing assets or new assets.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not work with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not work with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not work with human subjects.

The appendix is organized as follows: We first provide some insights on extended value iterations useful in our construction of the regret.Then, the detailed proof of theorem 4.1 is given with bounds on the five terms in our decomposition of the regret. A final appendix provides technical lemmas about MDPs in $\mathcal{M}$.

A Proof of Theorem 4.1

A.1 Extended value iteration

For each episode k, we use the extended value iteration algorithm described in [11] to compute $\tilde{\pi}_k$ and $\tilde{M} \in \mathcal{M}_k$, an optimistic policy and MDP. The values we iteratively get are defined in the following way:

$$\begin{aligned} u_0^{(k)}(s) &= 0 \\ u_{i+1}^{(k)}(s) &= \max_{a \in \mathcal{A}} \left\{ \tilde{r}(s,a) + \max_{p(\cdot) \in \mathcal{P}(s,a)} \left\{ \sum_{s \in S} p(s') u_i^{(k)}(s') \right\} \right\}, \end{aligned} \tag{16}$$

where $\tilde{r}$ is the maximal reward from (4) and $\mathcal{P}(s,a)$ is the set of probabilities from (5).

Now, from [11, Theorem 7], we obtain the following lemma on the iterations of extended value iteration.

Lemma A.1. *For episode k, denote by i the last step of extended value iteration, stopped when:*

$$\max_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} - \min_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} < \frac{r_{\max}}{\sqrt{t_k}}. \tag{17}$$

The optimistic MDP $\tilde{M}_k$ and the optimistic policy $\tilde{\pi}_k$ that we choose are so that the gain is $\frac{1}{\sqrt{t_k}}-$ close to the optimal gain:

$$\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{r_{\max}}{\sqrt{t_k}}. \tag{18}$$

Moreover from [16, Theorem 8.5.6]:

$$\left| u_{i+1}^{(k)}(s) - u_i^{(k)}(s) - \tilde{\rho}_k \right| \leq \frac{r_{\max}}{\sqrt{t_k}}, \tag{19}$$

and when the optimal policy yields an irreducible and aperiodic Markov chain, we have that $\tilde{\rho}_k = \rho(\tilde{M}_k, \tilde{\pi}_k, s)$ for any s, so that we can define the bias:

$$\tilde{h}_k(s_0) = \mathbb{E}_{s_0} \left[\sum_{t=0}^{\infty} (\tilde{r}(s_t, a_t) - \tilde{\rho}_k) \right]. \tag{20}$$

By choosing iteration i large enough, from [16, Equation 8.2.5], we can also ensure that:

$$\left| u_i^{(k)}(s) - (i-1)\tilde{\rho}_k - \tilde{h}_k(s) \right| < \frac{r_{\max}}{2\sqrt{t_k}}, \tag{21}$$

so that we can define the following difference

$$d_k(s) := \left| u_i^{(k)}(s) - \min_s u_i^{(k)}(s) - \left(\tilde{h}_k(s) - \min_s \tilde{h}_k(s) \right) \right| < \frac{r_{\max}}{\sqrt{t_k}}. \tag{22}$$

A.2 Regret when M is out of the confidence bound

Let us compute $\mathbb{E}[\mathrm{Reg}]$, the expected regret. We will mainly follow the approach in [11, Section 4], with a few tweaks. We start by splitting the regret into a sum over episodes and states.

We remind that $\overline{r}(s,a)$ is the overall mean reward and $N_T(s,a)$ the total count of visits. We also define $R_k(s) := \sum_a \nu_k(s,a)(\rho^* - \overline{r}(s,a))$ the regret at episode k induced by state s, with $\nu_k(s,a)$ the number of visit of (s,a) during episode k.

Let $R_{\text{in}} := \sum_s \sum_{k=1}^m R_k(s) \mathbb{1}_{M \in \mathcal{M}_k}$ and $R_{\text{out}} := \sum_s \sum_{k=1}^m R_k(s) \mathbb{1}_{M \notin \mathcal{M}_k}$. We therefore have the split, for $T \geq 2$, as $\log T \geq \frac{1}{4}$:

$$\mathbb{E}\left[Reg\right] \leq \mathbb{E}\left[R_{\text{in}}\right] + \mathbb{E}\left[R_{\text{out}}\right] + . \tag{23}$$

Now, let $\nu_k(s) = \sum_a \nu_k(s,a)$ and denote by $\mathcal{M}(t)$ the set of MDPs $\mathcal{M}_k$ such that $t_k \leq t < t_{k+1}$. For the terms out of the confidence sets, we have:

$$\begin{aligned}
R_{\text{out}} &\leq \sum_s \sum_{k=1}^m \nu_k(s) \mathbb{1}_{M \notin \mathcal{M}_k} \\
&\leq \sum_s \sum_{k=1}^m N_{t_k}(s) \mathbb{1}_{M \notin \mathcal{M}_k} \text{ using the stopping criterion} \\
&= \sum_{t=1}^T \sum_s \sum_{k=1}^m \mathbb{1}_{t_k = t} N_t(s) \mathbb{1}_{M \notin \mathcal{M}(t)} \leq \sum_{t=1}^T \sum_s N_t(s) \mathbb{1}_{M \notin \mathcal{M}(t)} \\
&= \sum_{t=1}^T \mathbb{1}_{M \notin \mathcal{M}(t)} \sum_s N_t(s) \leq \sum_{t=1}^T t \mathbb{1}_{M \notin \mathcal{M}(t)}.
\end{aligned}$$

Taking the expectations:

$$\begin{aligned}
\mathbb{E}\left[R_{\text{out}}\right] &\leq r_{\max} \sum_{t=1}^T t \mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \\
&\leq r_{\max} \sum_{t=1}^T \frac{tS}{2t^3} \leq r_{\max} \sum_{t=1}^T \frac{S}{2t^2} \text{ by Lemma B.1} \\
&\leq r_{\max} S.
\end{aligned} \tag{24}$$

Thus, we have dealt with the cases where the MDP M did not belong to any confidence set, for some episodes. We now need to deal with the rest.

A.3 Regret terms when M is in the confidence bound

We now assume that $M \in \mathcal{M}_k$ and deal with the terms in the confidence bound, so that we can omit the repetitions of the indicator functions. For each episode k, let $R_{\text{in},k} := \sum_s R_k$.

We defined $\tilde{\pi}_k$ the optimistic policy computed at episode k, now define $\tilde{P}_k := (\tilde{p}_k(s'|s, \tilde{\pi}_k(s)))$ the transition matrix of that policy on the optimistic MDP $\tilde{M}_k$. Define also $\mathbf{v}_k := (\nu_k(s, \tilde{\pi}_k)$ the row vector of visit counts during episode k. Following the same steps as in [11], we get the inequality on the regret of episode k, assuming $M \in \mathcal{M}_k$, using Lemma A.1:

$$\begin{aligned}
R_{\text{in},k} &= \sum_{s,a} \nu_k(s,a)(\rho^* - \overline{r}(s,a)) \\
&\leq \sum_{s,a} \nu_k(s,a)(\tilde{\rho}_k - \overline{r}(s,a)) + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}} \\
&= \sum_{s,a} \nu_k(s,a)(\tilde{\rho}_k - \tilde{r}_k(s,a)) + \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a)) + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}.
\end{aligned}$$

Then with (19) and using the definition of the iterated values from EVI, we have for a given state s and $a_s := \tilde{\pi}_k(s)$:

$$\left| (\tilde{\rho}_k - \tilde{r}_k(s,a_s)) - \left(\sum_{s'} \tilde{p}_k(s'|s,a_s) u_i^{(k)}(s') - u_i^{(k)}(s) \right) \right| \leq \frac{r_{\max}}{\sqrt{t_k}},$$

so that:

$$R_{\text{in},k} \leq \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I} \right) \mathbf{u}_i + \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a)) + 2 r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}.$$

Remember that for any state s: $|d_k(s)| \leq \frac{r_{\max}}{\sqrt{t_k}}$, where $\tilde{\mathbf{h}}_k$ is the bias of the average optimal policy for the optimist MDP, and:

$$d_k(s) := \left(u_i^{(k)}(s) - \min_x u_i^{(k)}(x)\right) - \left(\tilde{\mathbf{h}}_k(s) - \min_x \tilde{\mathbf{h}}_k(x)\right).$$

Notice that the unit vector is in the kernel of $\left(\tilde{\mathbf{P}}_k - \mathbf{I}\right)$. Therefore, in the first term, we can replace $\mathbf{u}_i$ by any translation of it. We get:

$$\mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{u}_i = \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k + \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k.$$

so that:

$$R_{\text{in}} \leq \underbrace{\sum_k \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a))}_{R_{\text{rewards}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{bias}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k + 2r_{\max} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}}_{R_{\text{EVI}}}.$$

Then, using the assumption on empirical rewards (4), as $M \in \mathcal{M}_k$, and noticing that $N_{t_k} \leq t_k$:

$$R_{\text{rewards}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{25}$$

For the term $\mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k$, which does not appear in the analysis of [11], we obtain

$$\begin{aligned}
\mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k &\leq \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - \mathbb{1}_s\right\|_1 \cdot \sup_{s'} |d_k(s')| \\
&\leq 2r_{\max} \sum_s \frac{\nu_k\left(s, \tilde{\pi}_k(s)\right)}{\sqrt{t_k}} \leq 2r_{\max} \sum_{s,a} \frac{\nu_k\left(s,a\right)}{\sqrt{t_k}} \\
&\leq 2r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}},
\end{aligned}$$

where in the last inequality we used that $\max\{1, N_{t_k}(s,a)\} \leq t_k \leq T$. Thus, for $T \geq \frac{e^2}{2AT}$ the regret term coming from the consequences and approximations of EVI satisfies

$$R_{\text{EVI}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{26}$$

Now, by defining P_k the transition matrix of the optimistic policy $\tilde{\pi}_k$ in the true MDP M, we have the following decomposition of the middle term:

$$\underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \mathbf{h}^*}_{R_{\text{trans}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \left(\tilde{\mathbf{h}}_k - \mathbf{h}^*\right)}_{R_{\text{diff}}} + \underbrace{\sum_k \mathbf{v}_k \left(\mathbf{P}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{ep}}}$$

Overall:

$$R_{\text{in}} \leq \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \mathbf{h}^*}_{R_{\text{trans}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \left(\tilde{\mathbf{h}}_k - \mathbf{h}^*\right)}_{R_{\text{diff}}} + \underbrace{\sum_k \mathbf{v}_k \left(\mathbf{P}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{ep}}} + \underbrace{r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}}_{R_{\text{EVI}} + R_{\text{rewards}}}.$$

A.3.1 Bound on R_{trans}

Let us deal with the first term R_{trans}. To bound this term, we will use our knowledge of the optimal bias $\mathbf{h}^*$ and the control of the difference of the transition matrices, and for the second term we will control the difference of the biases.

Notice that for a fixed state $1 \leq s \leq N-1$:

$$\sum_{s'} p\left(s'|s, \tilde{\pi}_k(s)\right) h^*(s') = \sum_{s'} p\left(s'|s, \tilde{\pi}_k(s)\right) \left(h^*(s') - h^*(s)\right) + h^*(s).$$

The same is true for $\tilde{p}_k$, and knowing the MDP is a birth and death process:

$$\begin{aligned}
R_{\text{trans}} &= \sum_k \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot h^*(s') \\
&= \sum_k \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot \left(h^*(s') - h^*(s)\right) \\
&\leq \sum_k \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - p\left(\cdot|s, \tilde{\pi}_k(s)\right)\right\|_1 \sup_s \partial h^*(s) \\
&\leq 4\sqrt{2\log\left(2AT\right)} \sum_k \sum_{s,a} \frac{\Delta(s)\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}},
\end{aligned}$$

where in the last inequality, we used the knowledge on the bounded variations of the optimal bias from Lemma 3.2, and that the optimistic MDP has transitions close to the true transitions.

A.3.2 Bound on R_{diff}

We now deal with the term involving the difference of bias, R_{diff}. For each episode k with policy π_k, denote by x_k the state such that the confidence bounds are at their worst and denote by $a_k := \pi_k(x_k)$ the corresponding action used at this state, so that $N_{t_k}(x_k, a_k)$ is minimal. We therefore have that $\sqrt{\frac{\log(2At_k)}{\max\{1, N_{t_k}(x_k,a_k)\}}}$ is maximal for episode k. The true MDP being within the confidence bounds, with a triangle inequality:

$$\|P_k - P^*\|_\infty \leq 4\sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}},$$

and

$$\|r_k - r^*\|_\infty \leq 2r_{\max}\sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}}.$$

Then using Lemma C.4, and noticing that to bound the biases $\tilde{\mathbf{h}}_k$, $\mathbf{h}^*$ and the quantity $\|\sum_{t=1}^T \tilde{P}_k^t \tilde{r}_k\|$ is bounded by the same diameter D, using the same argument as in [11] (Equation (11)), and noticing that $D \geq 1$:

$$\|\tilde{\mathbf{h}}_k - \mathbf{h}^*\|_\infty \leq 12 T_{hit} r_{\max} D \sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}}. \tag{27}$$

Hence,

$$\begin{aligned}
R_{\text{diff}} &\leq \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot \left(\tilde{h}_k(s') - h^*(s')\right) \\
&\leq \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - p\left(\cdot|s, \tilde{\pi}_k(s)\right)\right\|_1 \left\|\tilde{\mathbf{h}}_k - \mathbf{h}^*\right\|_\infty \\
&\leq 48 D^2 r_{\max} \log\left(2AT\right) \Sigma,
\end{aligned}$$

where in the last inequality we have used (27) and that by definition of D

$$T_{hit} := \inf_{s' \in \mathcal{S}} \sup_{s \in \mathcal{S}} \mathbb{E}_s\, \tau_{s'}^{\pi^*} \leq \mathbb{E}_{S-1}\, \tau_0^{\pi^0} \leq D,$$

and we called

$$\Sigma := \sum_{s,a}\sum_{k}\sum_{t=t_k}^{t_{k+1}-1} \frac{\mathbb{1}_{\{s_t,a_t=s,a\}}}{\sqrt{\max\{1,N_{t_k}(s,a)\}}\sqrt{\max\{1,N_{t_k}(x_k,a_k\}}}.$$

By the choice of x_k, $N_{t_k}(x_k,a_k) \leq N_{t_k}(s,a)$ for any state-action pair (s,a), so that we can rewrite, with $I_k := t_{k+1} - t_k$ the length of episode k:

$$\Sigma \leq \sum_{s,a}\sum_{k}\sum_{t=t_k}^{t_{k+1}-1} \frac{\mathbb{1}_{\{s_t,a_t=s,a\}}}{\max\{1,N_{t_k}(s,a)\}} = \sum_{k} \frac{I_k}{\max\{1,N_{t_k}(x_k,a_k)\}}.$$

Now define $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$, and $I(T) := \max\left\{Q_{\max}, T^{1/4}\right\}$. We split the sum depending on whether the episodes are shorter than $I(T)$ or not, and call $K_{\leq I}$ the number of such episodes. This yields:

$$\Sigma \leq K_{\leq I} I(T) + \sum_{k, I_k > I(T)} \frac{I_k}{\max\{1,N_{t_k}(x_k,a_k)\}}.$$

Using the stopping criterion for episodes:

$$\Sigma \leq K_{\leq I} I(T) + \sum_{k, I_k > I(T)} \frac{I_k}{\max\{1,\nu_k(x_k,a_k)\}}.$$

Now denote by $\mathcal{E}$ the event:

$$\mathcal{E} = \left\{\forall k \text{ s.t } I_k > I(T),\ \frac{1}{\max\{1,\nu(x_k,a_k)\}} \leq \frac{2}{m^{\max}(S-1)I_k}\right\}.$$

By splitting the sum, using the above event, we get:

$$\begin{aligned}\Sigma &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \sum_{k, I_k > I(T)} \frac{2}{m^{\max}(S-1)} + \mathbb{1}_{\bar{\mathcal{E}}} \sum_{k, I_k > I(T)} I_k \\ &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \left(K_T - K_{\leq I}\right) \frac{2}{m^{\max}(S-1)} + \mathbb{1}_{\bar{\mathcal{E}}} T.\end{aligned}$$

We use Corollary C.6 to get $\mathbb{P}\left(\bar{\mathcal{E}}\right) \leq \frac{1}{4T}$, so that when taking the expectation:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_{\leq I}\right] I(T) + \mathbb{E}\left[\left(K_T - K_{\leq I}\right)\right] \frac{2}{m^{\max}(S-1)} + \frac{1}{4}$$

Now using Lemma B.3, $SA \geq 4$, $I(T) \geq \frac{2}{m^{\max}(S-1)}$ and that $\frac{1}{\log 2} + \frac{1}{4} \leq 2$:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_T\right] I(T) + \frac{1}{4} \leq 2SA \log(2AT) I(T).$$

We therefore have that:

$$\mathbb{E}\left[R_{\text{diff}}\right] \leq 96 r_{\max} SAD^2 I(T) \log^2\left(2AT\right). \tag{28}$$

A.3.3 Bound on the main terms: Exploiting the stochastic ordering

In Section 4.3 we have shown that:

$$R_{\text{trans}} \leq 4\sqrt{2\log\left(2AT\right)} \sum_{s,a} \frac{\Delta(s)\nu_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}. \tag{29}$$

To control this term as well as R_{EVI} (26) and R_{rewards} (25), we need to control the terms in the sum in a way that does not make the parameters D or S appear, as this will be one of the main contributing

terms. To do so, we need to sum over the episodes and take the expectation, so that with Lemma B.4, we get:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\right] \leq 3\mathbb{E}\left[\sum_{s,a}\sqrt{N_T(s,a)}\right]$$
$$\leq 3\sum_{s}\sqrt{\mathbb{E}\left[N_T(s)\right]A} \text{ by Jensen's inequality.}$$

We will use the following lemma to carry on the computations:

Lemma A.2. *Let m^{π^0} be the stationary measure of the Markov chain under policy π^0, such that for every state s: $\pi^0(s)=0$. Let $f:\mathcal{S}\to\mathbb{R}^+$ be a non-negative non-decreasing function on the state space. Then for any state $s\in\mathcal{S}$,*

$$\mathbb{E}\left[\sum_{s'\geq s} f(s')N_t(s')\right] \leq t\sum_{s'\geq s} f(s')m^{\pi^0}(s') \tag{30}$$

Proof. Let $s\in\mathcal{S}$. For any state s', define $N_t^{m^{\pi^0},\pi^0}(s')$ the number of visits when the starting state follows the initial distribution m^{π^0}, and the MDP always executes the policy π^0 at every timestep instead of the policy determined by the algorithm UCRL2. Notice already that for any state s':

$$\mathbb{E}\left[N_t^{m^{\pi^0},\pi^0}(s')\right] = tm^{\pi^0}(s')$$

On the other hand, for $x\in\mathcal{S}$, we have the stochastic ordering:

$$\sum_{s'\geq x} N_t(s') \leq_{st} \sum_{s'\geq x} N_t^{m^{\pi^0},\pi^0}(s'),$$

so that for any non-negative non-decreasing function f, with the convention $f(-1)=0$:

$$\begin{cases} (f(x)-f(x-1))\sum_{s'\geq x} N_t(s') \leq_{st} (f(x)-f(x-1))\sum_{s'\geq x} N_t^{m^{\pi^0},\pi^0}(s') \\ f(s-1)\sum_{s'\geq s} N_t(s') \leq_{st} f(s-1)\sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s'), \end{cases} \tag{31}$$

and then summing the equation above for $s\leq x\leq S-1$ and switching the sums yields:

$$\sum_{s'\geq s} N_t(s')\sum_{x=s}^{s'}[f(x)-f(x-1)] \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')\sum_{x=s}^{s'}[f(x)-f(x-1)],$$

which simplifies to:

$$\sum_{s'\geq s} N_t(s')[f(s')-f(s-1)] \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')[f(s')-f(s-1)].$$

Now summing with the second equation in (31), we get the following equation:

$$\sum_{s'\geq s} N_t(s')f(s') \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')f(s').$$

Taking the expectation in this last inequality finishes the proof. □

Now, we can conclude our bound on R_{trans}. Since

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}(\Delta(s)+r_{\max})\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\right] \leq 3\sqrt{A}\sum_{s\geq 0}(\Delta(s)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]}, \tag{32}$$

let f be a non-negative non-decreasing function over the state space, such that $F := \sum_{s\geq 0} f(s)^{-1}$ exists. Then by concavity:

$$\begin{aligned}
\sum_{s\geq 0}(\Delta(s)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]} &= F\sum_{s\geq 0}\frac{1}{Ff(s)}\sqrt{f(s)^2(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq F\sqrt{\sum_{s\geq 0}\frac{f(s)^2(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]}{Ff(s)}} \text{ by concavity} \\
&= \sqrt{F\sum_{s\geq 0} f(s)(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq \sqrt{TF\sum_{s\geq 0} f(s)(\Delta(s)+r_{\max})^2 m^{\pi^0}(s)} \text{ using Lemma A.2,}
\end{aligned}$$

so that overall, (32) becomes:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{(\Delta(s)+r_{\max})\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\right] \leq 3\sqrt{ATF}\sqrt{\sum_{s\geq 0} f(s)(\Delta(s)+r_{\max})^2 m^{\pi^0}(s)}. \tag{33}$$

This is the term mainly contributing to the regret.

A.3.4 Bound on the main terms: Introducing E_2

Now, using Lemma B.5 which gives the stationary distribution of m^0, we can compute the expectation under m^0 of a well-chosen function f:

Lemma A.3. *Let us choose the increasing function $f : s \mapsto \frac{\max\{1,s(s-1)\}}{(\Delta(s)+r_{\max})^2}$. Then $F \leq 3(C+r_{\max})^2$ and $\sum_{s\geq 0}(\Delta(s)+r_{\max})^2 f(s) m^{\pi^0}(s) = \mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq \left(1+\frac{\lambda^2}{\mu^2}\right)$, so that:*

$$E_2 := F\mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq 3(C+r_{\max})^2\left(1+\frac{\lambda^2}{\mu^2}\right).$$

Proof. For F, we obtain:

$$F \leq (C+r_{\max})^2\left(2+\sum_{s=2}^{S-1}\frac{1}{s(s-1)}\right) = (C+r_{\max})^2\left(2+\sum_{s=2}^{S-1}\left(\frac{1}{s-1}-\frac{1}{s}\right)\right) \leq 3(C+r_{\max})^2$$

Using the following computations:

$$\begin{aligned}
\sum_{s=2}^{S-1} s(s-1)\binom{S-1}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^s &= (S-2)(S-1)\sum_{s=2}^{S}\binom{S-3}{s-2}\left(\frac{\lambda}{(S-1)\mu}\right)^s \\
&= (S-2)(S-1)\left(\frac{\lambda}{(S-1)\mu}\right)^2\sum_{s=0}^{S-3}\binom{S-3}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^s \\
&= (S-2)(S-1)\left(\frac{\lambda}{(S-1)\mu}\right)^2\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-3} \\
&\leq \left(\frac{\lambda}{\mu}\right)^2\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-3},
\end{aligned}$$

and using that $1+\frac{\lambda}{\mu} \leq \left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1}$, we get:

$$\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1}\mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq \left(1+\frac{\lambda^2}{\mu^2}\right)\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1},$$

so that finally

$$\mathbb{E}_{m^{\pi^0}}\left[(\Delta + r_{\max})^2 \cdot f\right] \leq \left(1 + \frac{\lambda^2}{\mu^2}\right),$$

which concludes the proof. □

Finally (33) becomes:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{(\Delta(s) + r_{\max})\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}\right] \leq 3\sqrt{E_2 AT}, \tag{34}$$

and thus:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 12\sqrt{2E_2 AT\log\left(2AT\right)}. \tag{35}$$

In particular:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 30(C + r_{\max})\sqrt{\left(1 + \frac{\lambda^2}{\mu^2}\right) AT\log\left(2AT\right)}. \tag{36}$$

A.3.5 Bound on R_{ep}

It remains to deal with the following regret term:

$$R_{\text{ep}} = \sum_k \mathbf{v}_k\left(\mathbf{P}_k - \mathbf{I}\right)\tilde{\mathbf{h}}_k.$$

We will follow the core of the proof from [11]. Define $X_t := \left(p(\cdot|s_t, a_t) - \mathbf{e}_{s_t}\right)\tilde{\mathbf{h}}_{k(t)}\mathbb{1}_{M\in\mathcal{M}_{k(t)}}$, where $k(t)$ is the episode containing step t and $\mathbf{e}_i$ the vector with i-th coordinate 1 and 0 for the other coordinates.

$$\begin{aligned}\mathbf{v}_k\left(\mathbf{P}_k - \mathbf{I}\right)\tilde{\mathbf{h}}_{\mathbf{k}} &= \sum_{t=t_k}^{t_{k+1}-1} X_t + \tilde{\mathbf{h}}_k(s_{t_{k+1}}) - \tilde{\mathbf{h}}_k(s_{t_k}) \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + Dr_{\max}.\end{aligned}$$

By summing over the episodes we get:

$$R_{\text{ep}} \leq \sum_{t=1}^{T} X_t + K_T D r_{\max}.$$

Notice that $\mathbb{E}\left[X_t|s_1, a_1, \ldots, s_t, a_t\right] = 0$, so that when taking the expectations, only the term in the number of episodes remains.

On the other side, using Lemma B.3, we get when taking the expectation:

$$\mathbb{E}\left[R_{\text{ep}}\right] \leq SA\log_2\left(\frac{8T}{SA}\right)\cdot Dr_{\max}.$$

Assuming $SA \geq 4$, and using $\log(2) \geq \frac{1}{2}$:

$$\mathbb{E}\left[R_{\text{ep}}\right] \leq 2r_{\max}SAD\log(2AT). \tag{37}$$

We can now gather the expected regret terms when the true MDP is within the confidence bounds. Using (28), (35) and (37):

$$\mathbb{E}\left[R_{\text{in}}\right] \leq 96r_{\max}SAD^2 I(T)\log^2\left(2AT\right) + 12\sqrt{2E_2 AT\log\left(2AT\right)} + 2r_{\max}SAD\log(2AT),$$

which gives with (23) and (24), assuming that $T \geq S^2$:

$$\mathbb{E}\left[Reg\right] \leq 97 r_{\max} SAD^2 I(T) \log^2\left(2AT\right) + 12\sqrt{2E_2 AT \log\left(2AT\right)}.$$

which finally gives:

$$\mathbb{E}\left[Reg\right] \leq 97 r_{\max} SAD^2 I(T) \log^2\left(2AT\right) + 19\sqrt{E_2 AT \log\left(2AT\right)}.$$

B Technical Lemmas

B.1 Probability of the confidence bounds

This first lemma is from [11, Lemma 17] and adapted to our confidence bounds.

Lemma B.1. *For $t > 1$, the probability that the MDP M is not within the sate of plausible MDPs $\mathcal{M}_t$ is bounded by:*

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} < \frac{S}{2t^3}.$$

Proof. Fix a state action pair (s, a), and n the number of visits of this pair before time t. Recall that $\hat{p}$ and $\hat{r}$ are the empirical transition probabilities and rewards from the n observations. Knowing that from each pair, there are at most 3 transitions, a Weissman's inequality gives for any $\varepsilon_p > 0$:

$$\mathbb{P}\left\{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \varepsilon_p\right\} \leq 6\exp\left(-\frac{n\varepsilon_p^2}{2}\right).$$

So that for the choice of $\varepsilon_p = \sqrt{\frac{2}{n}\log\left(16At^4\right)} \leq \sqrt{\frac{8}{n}\log\left(2At\right)}$, we get:

$$\mathbb{P}\left\{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log\left(2At\right)}\right\} \leq \frac{3}{8At^4}.$$

We can do similar computations for the confidence on rewards, with a Hoeffding inequality:

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq \varepsilon_r\right\} \leq 2\exp\left(-\frac{2n\varepsilon_r^2}{r_{\max}^2}\right),$$

and choosing $\varepsilon_r = r_{\max}\sqrt{\frac{1}{2n}\log\left(16At^4\right)} \leq r_{\max}\sqrt{\frac{2}{n}\log\left(2At\right)}$, so that:

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq r_{\max}\sqrt{\frac{2}{n}\log\left(2At\right)}\right\} \leq \frac{1}{8At^4}.$$

Now with a union bound for all values of $n \in \{0, 1, \cdots, t-1\}$, we get:

$$\mathbb{P}\left\{\|\hat{p}(\cdot) - p(\cdot)\|_1 \geq \sqrt{\frac{8\log\left(2At\right)}{\max\{1, N_t(s,a)\}}}\right\} \leq \frac{3}{8At^3},$$

and

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq r_{\max}\sqrt{\frac{2\log\left(2At\right)}{\max\{1, N_t(s,a)\}}}\right\} \leq \frac{1}{8At^3},$$

and finally, when summing over all state-action pairs, $\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} < \frac{S}{2t^3}$. □

B.2 Number of visits for an MDP in $\mathcal{M}$

This lemma is needed in the proof of Lemma C.5.

Lemma B.2 (Azuma-Hoeffding inequality)**.** *Let $X_1, X_2, \dots$ be a martingale difference sequence with $|X_i| \leq RD$ for all i and some $R > 0$. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$:*

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq \varepsilon\right\} \leq \exp\left(-\frac{\varepsilon^2}{2nDR}\right).$$

The two following lemmas are proved in [11, Appendix C.2 and Appendix C.3] respectively. Bounding the number of episodes is notably useful to obtain equation (28).

Lemma B.3. *Denote by K_t the number of episodes up to time t, and let $t > SA$. It is bounded by:*

$$K_t \leq SA \log_2\left(\frac{8t}{SA}\right).$$

The following lemma is used to simplify regret terms, notably (29).

Lemma B.4. *For any fixed state action pair (s, a) and time T, we have:*

$$\sum_{t=1}^{T} \frac{\mathbb{1}_{\{s_t, a_t = s, a\}}}{\sqrt{\max\{1, N_t(s,a)\}}} \leq 3\sqrt{N_{T+1}(s,a)},$$

B.3 Diameter and Span of MDPs in $\mathcal{M}$

For completeness, and to support the discussion in Section 4.2, the section details the behavior of the diameter and the span of MDPs in $\mathcal{M}$, as functions of S.

Under policy π^0, it is possible to get en explicit expression for the stationary distribution of the states.

Lemma B.5. *Under the stationary policy π^0, the stationary measure $m^{\pi^0}(s)$ of the MDP is given by:*

$$m^{\pi^0}(s) = \frac{\binom{S-1}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^s}{\left(1 + \frac{\lambda}{(S-1)\mu}\right)^{(S-1)}}.$$

This lemma is shown in the proof of [1, Lemma 3.3].

First, it should be clear that under any policy π, the diameter of the MDP under π is extremely large because the probability to move from state s to state $s+1$ is smaller and smaller as s grows. Actually, this is also true for the local diameter, more precisely the expected time to go from s to $s+1$ grows extremely fast with s.

This discussion is formalized in the following result.

Lemma B.6. *For any $M \in \mathcal{M}$ and any policy π, the diameter D^π as well as the local diameter $D^\pi(s-1, s)$ grow as S^{S-2}.*

Proof. Under policy π, the following sequence of inequalities follows from the stochastic comparison with π^0 and monotonicity under π^0.

$$D^\pi \geq \tau^\pi(0, S-1) \geq \tau^{\pi^0}(0, S-1) \geq \tau^{\pi^0}(S-2, S-1),$$

where $\tau^\pi(x, y)$ is the expected time to go from x to y under policy π.

Now, starting from $S-2$, the Markov chain moves to $S-1$ with probability $p := \lambda/(U(S-1))$ and the time to reach $S-1$ is equal to 1 or moves to $S-2$ or $S-3$ with probability $1-p$. Therefore,

$\tau^{\pi^0}(S-2, S-1)$ is bounded by $1-p$ times the return time to $S-1$, bounded in turn by the inverse of the stationary measure of state $S-2$ in the chain truncated at $S-2$. Using Lemma B.5,

$$\tau^{\pi^0}(S-2, S-1) \geq (1-p)\left(\frac{(S-2)\mu}{\lambda}\right)^{(S-2)}\left(1+\frac{\lambda}{(S-2)\mu}\right)^{(S-2)} \tag{38}$$

$$= \exp\left(\frac{\lambda}{\mu}-2\right)\left(\frac{\mu}{\lambda}\right)^{S-2} S^{S-2}(1+o(1/S)). \tag{39}$$

As for the maximal local diameter, $\max_s D^\pi(s-1,s) \geq \max_s \tau^{\pi^0}(s-1,s) \geq \tau^{\pi^0}(S-2,S-1)$ and the same argument as before applies.

□

Let us now consider the bias of the optimal policy in M. Using Lemma 3.2, the bias $h^*(s)$ is decreasing and concave in s, with bounded increments. Therefore, its span, defined as span $(h^*) := \max_s h^*(s) - \min_s h^*(s)$, satisfies

$$(h^*(0)-h^*(1))S \leq \text{span } (h^*) \leq (h^*(S-2)-h^*(S-1))S \leq C(S-1).$$

This implies that the span of the bias behaves as a linear function of S for all M.

C Generic lemmas on ergodic MDPs

C.1 From bias variations to probability transition variations

The three first lemmas of this subsection are used in the proof of Lemma C.4. This lemma is needed to obtain equation (27).

Lemma C.1. *For a MDP with rewards $r \in [0, r_{\max}]$ and transition matrix P, denote by $J_s(\pi, T) := \mathbb{E}\left[\sum_{t=0}^T r(s_t, \pi(s_t))\right]$ the expected cumulative rewards until time T starting from state s, under policy π. Let D_π be the diameter under policy π. The following inequality holds: span $(J(\pi, T)) \leq r_{\max} D_\pi$.*

Proof. Let $s, s' \in \mathcal{S}$. Call $\tau_{s\to s'}$ the random time needed to reach state s' from state s under policy π. Then:

$$\begin{aligned} J_s(\pi,T) &= \mathbb{E}\left[\sum_{t=0}^{T} r(s_t)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\tau_{s\to s'}-1} r(s_t)\right] + \mathbb{E}\left[\sum_{t=\tau_{s\to s'}}^{T} r(s_t)\right] \\ &\leq r_{\max}\mathbb{E}\left[\tau_{s\to s'}\right] + J_{s'}(\pi,T) \\ &\leq r_{\max} D_\pi + J_{s'}(\pi,T), \end{aligned}$$

which proves the lemma. □

Lemma C.2. *Consider two ergodic MDPs M and M'. For $i \in 1, 2$, let $r_i \in [0, r_{\max}]$ and P_i be the rewards and transition matrix of MDP M_i under policy π_i, where both MDPs have the same state and action spaces. Denote by g_i the average reward obtained under policy π_i in the MDP M_i. Then the difference of the gains is upper bounded.*

$$|g - g'| \leq \|r - r'\|_\infty + r_{\max} D_\pi \|P - P'\|_\infty.$$

Proof. Define for any state s the following correction term $b(s) := r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\|$. Let us show by induction that for $T \geq 0$,

$$\sum_{t=0}^{T-1} P^t r \leq \sum_{t=0}^{T-1} P'^t (r+b).$$

This is true for $T = 0$. Assume that the inequality is true for some $T \geq 0$, then

$$\begin{aligned}
\sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t (r + b) &= -b + P \sum_{t=0}^{T-1} P^t r - P' \sum_{t=0}^{T-1} P'^t (r + b) \\
&= -b + P' \left(\sum_{t=0}^{T-1} P^t r - \sum_{t=0}^{T-1} P'^t (r + b) \right) + (P - P') \sum_{t=0}^{T} P^t r \\
&\leq -b + (P - P') \sum_{t=0}^{T} P^t r \text{ by induction hypothesis}
\end{aligned}$$

Notice that, for any state s:

$$\begin{aligned}
\left((P - P') \sum_{t=0}^{T} P^t r \right)(s) &\leq \|p(\cdot|s) - p'(\cdot|s)\| \cdot \text{span } (J(T)) \\
&\leq r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\| \text{ by Lemma C.1} \\
&= b(s)
\end{aligned}$$

In the same manner we show that:

$$\sum_{t=0}^{T} P^t r \geq \sum_{t=0}^{T} P'^t (r - b).$$

Hence, as P' has non-negative coefficients, denoting by e the unit vector:

$$\left\| \sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t r \right\|_\infty \leq \|b\|_\infty \left\| \sum_{t=0}^{T} P'^t \cdot e \right\|_\infty = \|b\|_\infty (T + 1).$$

We can also show that:

$$\left\| \sum_{t=0}^{T} P'^t r - \sum_{t=0}^{T} P'^t r' \right\|_\infty = \left\| \sum_{t=0}^{T} P'^t (r - r') \right\|_\infty \leq \|r - r'\|_\infty (T + 1)$$

And therefore with a multiplication by $\frac{1}{T+1}$ and by taking the Cesáro limit in $\left\| \sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t r' \right\|_\infty$, and with a triangle inequality:

$$|g - g'| \leq \|r - r'\|_\infty + \|b\|_\infty,$$

where $\|b\|_\infty = r_{\max} D_\pi \|P - P'\|_\infty$. □

Lemma C.3. *Let P be the stochastic matrix of an ergodic Markov chain with state space $1, \ldots, S$. The matrix $A := I - P$ has a block decomposition*

$$A = \begin{pmatrix} A_S & b \\ c & d \end{pmatrix};$$

then A_S, of size $(S - 1) \times (S - 1)$ is invertible and $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i \tau_S$, where $\mathbb{E}_i \tau_S$ is the expected time to reach state S from state i.

Remark that this lemma is true for any state in $\mathcal{S}$.

Proof. $(\mathbb{E}_i \tau_S)_i$ is the unique vector solution to the system:

$$\begin{cases} v(S) = 0 \\ \forall i \neq S,\ v(i) = 1 + \sum_{j \in \mathcal{S}} P(i, j) v(j) \end{cases}$$

We can rewrite this system of equations as: $\tilde{A} v = e - e_S$, where $\tilde{A}$ is the matrix

$$\tilde{A} := \begin{pmatrix} A_S & b \\ 0 & 1 \end{pmatrix},$$

e the unit vector and e_S the vector with value 1 for the last state and 0 otherwise. Then $\widetilde{A}$ and A_S are invertible and we write:

$$\tilde{A}^{-1} = \begin{pmatrix} A_S^{-1} & -A_S^{-1}b \\ 0 & 1 \end{pmatrix}.$$

Thus, by computing $\tilde{A}^{-1}(e - e_S)$, for $i \neq S$, $(\mathbb{E}_i \tau_S)_i = A_S^{-1}e$. By definition of the infinite norm and using that A_S is an M-matrix and that its inverse has non-negative components, $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i \tau_S$. □

In the following lemma, we use the same notations as in Lemma C.2 with a common state space $\{1, \dots S\}$.

Lemma C.4. *Let the biases h, h' be the biases of the two MDPs verify their respective Bellman equations with the renormalization choice $h(S) = h'(S) = 0$. Let $\sup_{s \in \mathcal{S}} \mathbb{E}_s \tau_{s'}^\pi$ be the worst expected hitting time to reach the state s' with policy π, and call $T_{hit} := \inf_{s' \in \mathcal{S}} \sup_{s \in \mathcal{S}} \mathbb{E}_s \tau_{s'}$. We have the following control of the difference:*

$$\|h - h'\|_\infty \leq 2T_{hit}(D' r_{\max} \|P - P'\|_\infty + \|r - r'\|_\infty)$$

Notice that although the biases are unique up to a constant additive term, the renormalization choice does not matter as the unit vector is in the kernel of $(P - P')$.

Proof. The computations in this proof follow the same idea as in the proof of [10, Theorem 4.2]. The biases verify the following Bellman equations $r - ge = (I - P)h$, and also the arbitrary renormalization equations, thanks to the previous remark: $h(S) = 0$. Using the same notations as in the proof of Lemma C.3, we can write the system of equations $\tilde{A}h = \tilde{r} - \tilde{g}$, with $\tilde{r}$ and $\tilde{g}$ respectively equal to r and g everywhere but on the last state, where their value is replaced by 0.

We therefore have that $h = \tilde{A}^{-1}(\tilde{r} - \tilde{g})$, and with identical computations, $h' = \tilde{A'}^{-1}(\tilde{r'} - \tilde{g'})$. By denoting $dX := X - X'$ for any vector or matrix X, we get:

$$dh = -\tilde{A}^{-1}(d\tilde{r} - d\tilde{g} + d\tilde{A}h').$$

The previously defined block decompositions are:

$$\tilde{A}^{-1} = \begin{pmatrix} A_S^{-1} & -A_S^{-1}b \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad d\tilde{A} = \begin{pmatrix} A_S - A'_S & b - b' \\ 0 & 0 \end{pmatrix}.$$

For $s < S$, $dh(s) = -e_s^T A_S^{-1}(dA_S h' + d\tilde{r} - d\tilde{g})$ and $dh(S) = 0$. Now by taking the norm and using C.1:

$$\|dh\|_\infty \leq \|A_S^{-1}\|_\infty (r_{\max} D' \|dA_S\|_\infty + \|d\tilde{r}\| + |d\tilde{g}|).$$

Notice that $\|dA_S\|_\infty \leq \|dP\|_\infty$, $\|d\tilde{r}\| \leq \|dr\|$ and $\|d\tilde{g}\| = |dg|$. Using Lemma C.2 and Lemma C.3, and taking the infimum for the choice of the state of renormalization implies the claimed inequality for the biases. □

C.2 A McDiarmid's inequality

Lemma C.5. *Recall that $m^{\max}$ is the stationary measure of the Markov chain under policy $\pi^{\max}$, such that for every state s: $\pi^{\max}(s) = A_{\max}$.*

Let k be an episode, and assume that the length of this episode I_k is at least $I(T) = 1 + \max\{Q_{\max}, T^{1/4}\}$, with $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$. Then, with probability at least $1 - \frac{1}{4T}$:

$$\nu_k(x_k, a_k) \geq m^{\max}(S-1)I_k - 5D\sqrt{I_k \log I_k}.$$

We will now prove Lemma C.5:

Proof. Let k be an episode such that $I_k \geq I(T)$, and first consider it is of fixed length I. Denote by $\mathring{r}$ the vector of reward equal to 1 if the current state is x_k and 0 otherwise. Denote by $\mathring{g}_{\pi_k}$ the gain associated to the policy π_k for the transitions p and rewards $\mathring{r}$, and similarly define $\mathring{h}_{\pi_k}$ the bias, translated so that $\mathring{h}_{\pi_k}(S-1) = 0$. Notice in that case, that if we denote by m_k the stationary distribution under policy π_k, that $m^{\max}(S-1) \leq m_k(s)$ for any state s, assuming that $S \geq \frac{\lambda}{\mu} + 1$. Then:

$$\begin{aligned}
\nu_k(x_k, a_k) &= \sum_{u=t_k}^{t_{k+1}-1} \mathring{r}(s_u) \\
&= \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} + \mathring{h}_{\pi_k}(s_u) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle \text{ using a Bellman's equation} \\
&= \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} + \mathring{h}_{\pi_k}(s_u) - \mathring{h}_{\pi_k}(s_{u+1}) + \mathring{h}_{\pi_k}(s_{u+1}) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle.
\end{aligned}$$

By Azuma-Hoeffding inequality B.2, following the same proof as in section 4.3.2 of [11], notice that $X_u = \mathring{h}_{\pi_k}(s_{u+1}) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle$ form a martingale difference sequence with $|X_u| \leq D$:

$$\mathbb{P}\left\{ \sum_{u=t_k}^{t_{k+1}-1} X_u \geq D\sqrt{10I \log I} \right\} \leq \frac{1}{I^5}.$$

Using that $\left| \mathring{h}_{\pi_k}(s_{t_k}) - \mathring{h}_{\pi_k}(s_{t_{k+1}}) \right| \leq D$, with probability at least $1 - \frac{1}{I^2}$:

$$\nu_k(x_k, a_k) \geq \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} - 5D\sqrt{I \log I}.$$

On the other hand:

$$\sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} = \nu_k(s_k, a_k) m_k(a_k),$$

so that, using that $m_k(a_k) \geq m^{\max}(S-1)$, with probability at least $1 - \frac{1}{I^5}$:

$$\nu_k(x_k, a_k) \geq m^{\max}(S-1)I - 5D\sqrt{I \log I}.$$

We now use a union bound over the possible values of the episode lengths I_k, between $I(T)+1$ and T:

$$\begin{aligned}
\mathbb{P}\left\{ \nu_k(x_k, a_k) < m^{\max}(S-1)I_k - 5D\sqrt{I_k \log I_k} \right\} &\leq \sum_{I=I(T)+1}^{T} \frac{1}{I^5} \leq \sum_{I=T^{1/4}+1}^{T} \frac{1}{I^5} \\
&\leq \frac{1}{4T},
\end{aligned}$$

so that we now have that with probability at least $1 - \frac{1}{4T}$:

$$\nu_k(x_k, a_k) \geq m^{\max}(S-1)I_k - 5D\sqrt{I_k \log I_k}.$$

□

We can show a corollary of Lemma C.5 that we will use for the regret computations:

Corollary C.6. *For an episode k such that its length I_k is greater than $I(T)$,with probability at least $1 - \frac{1}{4T}$:*

$$\nu_k(x_k, a_k) \geq \frac{m^{\max}(S-1)}{2} I_k.$$

Proof. With Lemma C.5, it is enough to show that $5D\sqrt{I_k \log I_k} \leq \frac{m^{\max}(S-1)}{2} I_k$, *i.e.* that $\sqrt{\frac{I_k}{\log I_k}} \geq \frac{10D}{m^{\max}(S-1)} =: B$. By monotonicity, as $I_k \geq Q_{\max} = B^2 \log B^4$ we can show instead that $B^2 \log B^4 \geq B^2 \log \left(B^2 \log B^4\right)$.

This last inequality is true, using that $\log x \geq \log(2 \log x)$ for $x > 1$. This proves the corollary. $\square$