Learning Optimal Admission Control in Partially Observable Queueing Networks

Jonatha Anselmi, Bruno Gaujal, Louis-Sébastien Rebuffi

Learning Optimal Admission Control in Partially Observable Queueing Networks

Jonatha Anselmi[1*], Bruno Gaujal[1†] and Louis-Sébastien Rebuffi[1†]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, Grenoble, 38000, France.

*Corresponding author(s). E-mail(s): jonatha.anselmi@inria.fr;
Contributing authors: bruno.gaujal@inria.fr;
louis-sebastien.rebuffi@univ-grenoble-alpes.fr;
†These authors contributed equally to this work.

Abstract

We develop an efficient reinforcement learning algorithm that learns the optimal admission control policy in a partially observable queueing network. Specifically, only the arrival and departure times from the network are observable, optimality refers to the average holding/rejection cost in infinite horizon, and efficiency is with respect to regret performance.

While reinforcement learning in partially-observable Markov Decision Processes (MDP) is prohibitively expensive in general, we show that the regret at time $\boldsymbol{T}$ induced by our algorithm is $\tilde{\boldsymbol{O}}(\sqrt{\boldsymbol{T}\,\mathbf{log}(\mathbf{1}/\boldsymbol{\rho})})$ where $\boldsymbol{\rho} \in (\mathbf{0}, \mathbf{1})$ is connected to the mixing time of the underlying MDP. In contrast with existing regret bounds, ours does not depend on the *diameter* ($\boldsymbol{D}$) of the underlying MDP, which in most queueing systems is at least exponential in $\boldsymbol{S}$, i.e., the maximal number of jobs in the network. Instead, the role of the diameter is played by the $\mathbf{log}(\mathbf{1}/\boldsymbol{\rho})$ term, which may depend on $\boldsymbol{S}$ but we find that such dependence is "minimal". In the case of acyclic or *hyperstable* queueing networks, we prove that $\mathbf{log}(\mathbf{1}/\boldsymbol{\rho}) = \boldsymbol{O}(\boldsymbol{S})$, which overall provides a regret bound of the order of $\tilde{\boldsymbol{O}}(\sqrt{\boldsymbol{TS}})$. In the general case, numerical simulations support the claim that the term $\mathbf{log}(\mathbf{1}/\boldsymbol{\rho})$ remains extremely small compared to the diameter.

The novelty of our approach is to leverage Norton's theorem for queueing networks and an efficient reinforcement learning algorithm for MDPs with the structure of birth-and-death processes.

Keywords: Product-form queueing networks, Norton's theorem, admission control, reinforcement learning, regret

1 Introduction

Research on reinforcement learning in Markov Decision Processes (MDP) has been flourishing since Walkins' work on Q-learning [1], the celebrated model-free learning algorithm. Since then, several extensions of Q-learning [2], Bayesian [3, 4] or model-based [5] approaches have been proposed to improve learning efficiency [6, 7] up to the point where the *regret* of the most recent algorithms is "asymptotically optimal" in the sense that it matches a universal lower bound. The regret of the best learning algorithm with respect to an optimal policy is $\tilde{O}(\sqrt{DSAT})$[1], where S is the number of states, A the number of actions, T the time horizon and D the *diameter* of the MDP; we point the reader to [8] for a recent overview of this quest for optimal regret.

Since this problem has reached a satisfactory solution, the following natural question arises: Can one learn *efficiently* the optimal policy of an MDP not only when the rewards and the transition kernel are unknown *but also when the state is partially observable*? Recently, this question has been investigated under certain assumptions on the structure of model parameters [9–11]. In this paper, we address this question in the context of queueing networks assuming that the learner only has access to the *total* number of jobs in the network. This places our problem in the family of Partially Observable MDPs (POMDPs).

1.1 Reinforcement learning in POMDPs

It is well-known that POMDPs are prohibitively expensive to solve. If the parameters are known, the problem of computing an optimal policy is PSPACE-complete even in finite horizon [12]. Furthermore, it is NP-hard to compute the optimal memoryless policy [13]. In reinforcement learning, where (some of) the model parameters are unknown, the lower bound on the average-case complexity developed in [9, Propositions 1 and 2] confirms with no surprise that reinforcement learning in POMDPs remains intractable. Matter of fact, the design of effective exploration–exploitation strategies in POMDPs is still relatively unexplored; see [10, Section 1] for a detailed discussion. In the attempt to reduce this computational burden, researchers focused on reinforcement learning in *subclasses* of POMDPs [9], and we will also follow this approach. The algorithm in [14] assumes POMDPs without resets and has sample complexity scaling exponentially with a certain horizon time. The Bayesian algorithms proposed in [15, 16] learn POMDPs but bounds on the mean regret remain unknown for these approaches. A sample-efficient algorithm for episodic finite POMDPs is given in [9]. Here, it is assumed that the number of observations is larger than the number of latent states.

The works above have focused on reinforcement learning over a finite or discounted horizon. In contrast, we will be interested in the (undiscounted) infinite horizon case, which is technically more challenging. In infinite horizon, a POMDP algorithm based on spectral methods is proposed in [10]. For this algorithm, the authors find an order-optimal regret bound with respect to the optimal memoryless policy. However, it exhibits a linear dependence on the diameter D of the underlying MDP. This dependence makes this type of bounds not interesting in the context of queueing systems as the diameter is usually exponential in the number of states [17]. Although the additional assumptions

[1] The $\tilde{O}$ notation is a variant of big-O that ignores logarithmic factors.

on the structure of the model mitigate, to some extent, the intrinsic complexity of POMDPs, learning algorithms with regret $\tilde{O}(\sqrt{DSAT})$ have remained elusive for all but trivial cases to the best of our knowledge.

1.2 Contribution and methodology

In this paper, we propose a learning algorithm for the optimal job-admission policy in a partially observable queueing network with regret $\tilde{O}(\sqrt{T\log(1/\rho)})$, where $\rho \in (0,1)$ is a parameter that is connected to the mixing time of the underlying MDP; see Section 5.2. The proposed algorithm, UCRL-M, builds on URCL2 [5] and its main novelty relies on the use of "modules"; this will be explained in Section 5. Thus, our main contribution is a learning algorithm with a regret bound that does not depend on the diameter D; we recall that D grows exponentially in the size of the state space, denoted by S, even in the simple case of an M/M/1 queue with a finite buffer [17]. In general, ρ depends on S but we claim that such dependence is minimal. In the case of acyclic or *hyperstable* networks, we show that $\log(1/\rho) = O(S)$, which overall provides a regret bound of the order of $\tilde{O}(\sqrt{TS})$, while in the general case, this is corroborated by numerical simulations.

Optimal admission control is one of the most classical control problems in queues. It has been investigated in several works; see, e.g., [18, 19] and the references therein. However, these works consider the case where the model parameters are known, i.e., no learning mechanism is employed. The novelty of our approach is to leverage i) Norton's theorem for closed product-form queueing networks [20] and ii) the efficiency of reinforcement learning in MDPs with the structure of birth-and-death processes [17]. More specifically, our result uses Norton's theorem to replace the whole network by a single load-dependent queue in its stationary regime and relies on the mixing time τ_{mix} of the network to apply this equivalence every τ_{mix} time-steps. We notice that Norton's theorem is only used for the performance analysis of the algorithm. The key observation is that Norton's theorem helps us to somewhat cast the original partially-observable MDP to a standard (fully-observable) MDP. In other words, the resulting asymptotically equivalent POMDP becomes an MDP with the structure of a birth and death process. This structure is then exploited to construct tight bounds on the regret of our algorithm by controlling the bias of the current policy as well as its stationary measure.

1.3 Organization

The remainder of the paper is organized as follows. The model of the queueing network, its practical motivation and Norton's equivalent queue are presented in Section 2. Section 3 presents the problem addressed in the paper in detail and Section 4 presents how reinforcement learning works in the considered context. Section 5 is dedicated to the presentation of our learning algorithm (UCRL-M) and Section 6 to the analysis of its regret. In the latter, we state our main result in Theorem 6.1. Then, Section 7 discusses some technical aspects of our regret bound. Section 8 showcases the behavior of our algorithm on a multi-tier queueing network, and, finally, Section 9 draws the conclusions of our work.

2 Admission control in a queueing network

We consider an (open) Jackson network with N queues (or stations) having service rates $\mu_1^o, \ldots, \mu_N^o$, routing probability matrix $L = [L_{i,j} : 0 \leq i, j \leq N]$ and exogenous arrivals occurring with rate λ. Here, $L_{0,j}$ (resp. $L_{i,0}$) represents the probability that a job joins queue j from outside (resp. leaves the network after service at queue i). We assume that $I + L + L^2 + \cdots$ is convergent, which implies that all jobs eventually leave the network. Let also λ_i^o be the arrival rate at queue i, which is given by the unique solution of the traffic equations $\lambda_i^o = \lambda L_{0,i} + \sum_j \lambda_j^o L_{j,i}$, for all i. For all i, we assume that $\lambda_i^o < \mu_i^o$, which would ensure stability (positive recurrence) even if the total number of jobs in the network were unlimited. We further assume that the total number of jobs in the network cannot exceed S.

Before joining the network, jobs go through an admission controller. The purpose of the admission controller is to either admit or reject jobs in order to minimize some cost function. For each job, the immediate cost c_t is decomposed into a per-rejection cost γ_{reject} and a per-time-unit holding cost γ_{hold}. This cost function is the long-run average cost per time unit: $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} c_t$.

When the controller can observe the state of the network and knows the parameters of the system (λ, μ^o, L), this classical problem has been solved in [19]. In the degenerate case where the network is a single M/M/1 queue, there exists an explicit formula (involving the Lambert W function) for the optimal admission policy [18].

2.1 Problem formulation

In this paper, we consider an admission controller that can only observe arrivals to and departures from the network. More precisely, the network topology, internal service rates and routing probabilities are not known, and the movements of the jobs inside the network are not observable. Our objective is to design a learning algorithm that learns the optimal admission policy with a *small* regret in the sense that its dependence on the network *complexity* is minimal.

For the cost to be minimized, we assume that:

- the controller may choose to reject jobs arriving in the network, at the price of a fixed γ_{reject} for each rejected job;
- for every time unit in the system, each job induces a holding cost γ_{hold} (this is the classical cost function for admission control, see [18]);
- the controller makes decisions only relying on its set of observations up to time t.

2.2 Motivating applications

Our main motivation is the control of computer and software systems. These systems are composed of multiple interconnected *containers*, where a container can be a cluster of servers or a modular software system, and admission control mechanisms are commonly employed to optimize performance. In the literature, containers are usually modeled via product-form queueing networks (for tractability) or layered queueing networks [21, 22], which justifies our modeling approach. In serverless computing, for instance, users of the serverless platform can control the overall number of simultaneous requests

that can be processed in a cluster of servers (each with its own queue) at any given time. In Knative, a Kubernetes-based platform to deploy and manage modern serverless workloads that is used among others by Google Cloud Run, admission thresholds are set via the `container-concurrency-target-default` global key [23] and the upper limit on the number of jobs that can be active running at the same time, i.e., S, can be controlled via the `max-scale-limit` global key. In Kubernetes, an open-source system for the management of containerized applications, admission controllers are configured via the `-enable-admission-plugins` and `-admission-control-config-file` flags and can be leveraged in case the pod (or application) is requesting too many resources.

Because of the complex relationships among containers, which can also be nested in multiple layers, i) a detailed knowledge of the current *state* is expensive to obtain at any point in time and ii) the internal container structure is also subject to estimation errors and may vary over time [24]. This leads us to our learning model, which is meant to capture both of these aspects: we do not know the network topology, routing probabilities and service rates as well as the current "state".

3 Markov Decision Process Formulation

In this section, we construct an MDP model for the system under investigation (referred to as "original MDP") as well as an artificial MDP (referred to as "aggregate MDP") that will be equivalent to the original MDP under its stationary regime. Note however that the learning algorithm constructed in the following only interacts with the original system. The aggregate system is only used for the performance analysis of the algorithm.

Before introducing the original MDP, we notice that under the constraint that at most S jobs can circulate in the network, the open Jackson network above is equivalent to a *closed* network with S jobs. This closed network, illustrated in Figure 1, is identical to the open one except for an auxiliary queue, say queue 0, that represents the outside world and has service rate λ. The departures of queue 0 correspond to the arrivals of the initial open network (see Figure 1). In this setting, if a job is rejected, it can be represented in the closed network as the job being returned to the external queue.

We are now in a position to define both the original and aggregate MDPs for the closed queueing network.

3.1 Original MDP

Let us model the problem as an MDP, $M^o = (\mathcal{X}^o, \mathcal{A}^o, P^o, r^o)$, where the superscript o stands for "original" throughout the paper. We first use uniformization to see the process in discrete time. The uniformization constant U (defined later, see Eq. (9)) is lower bounded by the sum of the rates: $U \geq \lambda + \sum_{i=1}^{N} \mu_i^o$. Thus, the time steps, which will be indexed by t, follow a Poisson process with rate U, and events (arrivals, services, routings and control actions) can only occur at these times. In the following $1/U$ will be seen as one time unit.

- The state space $\mathcal{X}^o$ is the set of all tuples $(x_1, \ldots, x_N)$ given by the number of jobs x_i in each queue i.
- The action space is $\mathcal{A}^o := \{0, 1\}$ where 0 stands for rejection and 1 for admission.

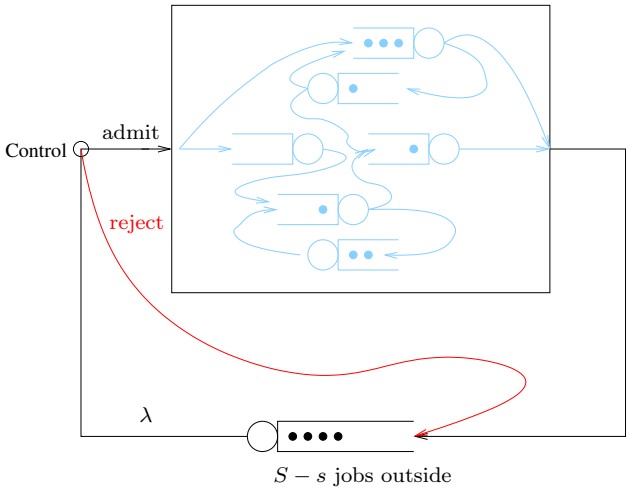


Figure 1: Admission control: rejected jobs immediately return to the outside queue.

- The transition matrix P^o is simply constructed by using the routing matrix L, the arrival rate λ and the service rates $(\mu_i^o)_i$.
- The mean rewards r^o are constructed from the cost function. The immediate cost for each state-action pair $(\mathbf{x}, a)$, is Bernoulli distributed. It is decomposed into:
  - a deterministic part, $\frac{1}{U}(\gamma_{\text{hold}} \sum_{i=1}^N x_i)$ (each present job incurs a cost γ_{hold} per time unit),
  - and a stochastic part, $\gamma_{\text{reject}}(1-a)\mathbf{1}_{\text{job-arrival}}$ (if a job arrives and the action is reject).

  To be consistent with the learning literature, where rewards are used instead of costs, we first define $r_{\max} := \frac{\lambda\gamma_{\text{reject}} + \gamma_{\text{hold}} S}{U}$ and for each state-action pair $(\mathbf{x}, a)$, the reward are Bernoulli distributed with expected value

$$r^o(\mathbf{x}, a) := r_{\max} - \frac{\lambda\gamma_{\text{reject}}(1-a) + \gamma_{\text{hold}} s}{U} = \frac{\lambda\gamma_{\text{reject}} a + \gamma_{\text{hold}}(S-s)}{U}, \quad (1)$$

where $s := \sum_{i=1}^N x_i$.

Let $\Pi^o := \{\pi : \mathcal{X}^o \to \mathcal{A}^o\}$ denote the set of stationary and deterministic policies. A stationary *policy* π is a deterministic function from $\mathcal{X}^o$ to $\mathcal{A}^o$.

Then, the MDP evolves under π in the standard Markovian way. At each time-step t, the system is in state $\mathbf{x}_t$, the controller chooses the action $a_t = \pi(\mathbf{x}_t)$ and receives a random reward whose expected value is $r^o(\mathbf{x}_t, a_t)$, and the system moves to state $\mathbf{x}'$ at time $t+1$ with probability $P^o(\mathbf{x}' \mid \mathbf{x}_t, a_t)$. The objective function is to minimize the long run average cost.

The average reward induced by policy π is:

$$g^o(M^o, \pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^o(\mathbf{x}_t, \pi(\mathbf{x}_t))]. \quad (2)$$

An optimal policy π^* for the original MDP achieves the best average reward $g^o(M^o, \pi^*) = \sup_{\pi \in \Pi^o} g^o(M^o, \pi)$.

3.2 Aggregate model

Let us define an aggregate MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ where the network is replaced by a single queue.

3.2.1 Norton equivalent queue

In this subsection, let us consider the system defined in Section 2 without control (all jobs are admitted).

The stationary measure of the network can be connected to the stationary measure of a birth-and-death process via Norton's theorem of queueing networks [20], also known in the literature as Flow Equivalent Server (FES) method [25]. When containing S jobs, the vector of the number of jobs in each queue forms a continuous-time Markov chain with stationary measure [25]

$$\nu^o(\mathbf{x}) = \frac{1}{G(S)} \prod_{i=0}^{N} \left(\frac{\lambda_i^o}{\mu_i} \right)^{x_i}, \tag{3}$$

for all $\mathbf{x} \in \{\mathbf{x} \in \mathbb{N}^N : |\mathbf{x}| \leq S\}$, where $G(S)$ is a normalization constant, $|\cdot|$ denotes the L_1 norm and λ_i^o was defined in Section 2.

The construction of our equivalent queue works as follows:

1. Given the closed Jackson network above, consider the (closed) network where queue 0 is short circuited (this means set $\mu_0^o = \infty$) and let $\mu(s)$ denote the *throughput*[2] of the network with s jobs in total (see Figure 2 for an illustration).
2. Consider the original network where all queues except 0 are all replaced by a *single* queue that operates with rate $\mu(s)$ if it contains s jobs.
3. Then,

$$\sum_{\mathbf{x}:|\mathbf{x}|=s} \nu^o(\mathbf{x}) = \nu(S-s, s), \quad \forall s = 0, \ldots, S \tag{4}$$

where $\nu(S-s, s)$, for all $s = 0, \ldots, S$, is the stationary measure of the reduced network with two queues.

We remark that $\nu(S-s, s)$ is indeed the stationary measure of a birth-and-death process with birth rate $\lambda \mathbb{1}_{s<S}$ and death rate $\mu(s)$, a fact that will be key in the regret analysis of our learning algorithm. In particular, we will use the following lemma, which provides some known properties about the throughput function $\mu(s)$ [26].

Lemma 3.1. *The throughput function $s \mapsto \mu(s)$ is increasing, concave and bounded by $\mu_{\max} := \sum_{i \leq N} \mu_i^o$.*

[2]The throughput of a closed Jackson queueing network with s jobs is the rate at which jobs flow at a reference queue (queue 0 in our case) and is defined by $\mu(s) := \frac{G(s-1)}{G(s)}$ where $G(s)$ is the normalizing constant appearing in the product-form expression (3) of the stationary measure m^o.

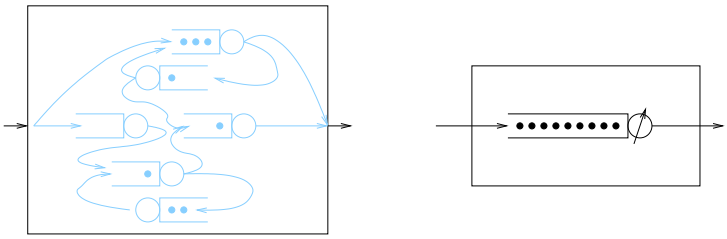


Figure 2: Illustration of Norton Equivalence theorem.

The throughput bound $\mu_{\max}$ can be significantly improved [25] but this will not change the structure of our results.

3.2.2 Aggregate MDP

Notice that, in the original MDP M^o, the rewards do not depend on the state but only on the number of jobs in the network, therefore, the Norton equivalent queue can also be used to construct an equivalent MDP.

Define the simplified equivalent MDP $M = (\mathcal{S}, \mathcal{A}, r, P)$.

- The state space $\mathcal{S} = \{0, \dots, S\}$ consists of all possible numbers of jobs in the queueing network. We denote by $S' := S + 1$ the number of states of the aggregate MDP.
- The actions are the same as for the original MDP: $\mathcal{A} = \mathcal{A}^o = \{0, 1\}$ (reject or accept).
- The original reward in (1) does not depend on the precise position of the jobs in the network but only on their number. Therefore for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we can define the expected reward as

$$r(s, a) = \frac{\lambda \gamma_{\text{reject}} a + \gamma_{\text{hold}}(S - s)}{U}. \tag{5}$$

  On a side note that anticipates on the future, we will use an upper bound on the difference of expected rewards between two neighboring states $\delta_{\max} := \gamma_{\text{reject}} + \frac{\gamma_{\text{hold}}}{U}$. We will use $\delta_{\max}$ instead of $r_{\max}$ (defined in Section 3.1) in the following derivation of the regret, as $\delta_{\max}$ does not depend on S and this will help us gain a factor S in the regret bound.
- The transition probabilities (P^π) are defined as follows.

  Let π be a policy (a function from $\mathcal{S} \to \mathcal{A}$) on M. By convention, π will also be seen as a policy in the original MDP M^o using the natural extension, i.e., if $\mathbf{x} \in \mathcal{S}^o$, then $\pi(\mathbf{x}) := \pi(|\mathbf{x}|)$. We can now define the transition matrix for policy π as the transition matrix in the aggregate MDP M under the stationary measure $\nu^{o,\pi}$:

$$P(s' \mid s, \pi(s)) = \sum_{\mathbf{x}, |\mathbf{x}| = s} \sum_{\mathbf{y}, |\mathbf{y}| = s'} \frac{\nu^{o,\pi}(\mathbf{x})}{\nu^{\pi}(s)} P^o(\mathbf{y} \mid \mathbf{x}, \pi(\mathbf{x})), \tag{6}$$

where $\nu^\pi(s) = \sum_{\mathbf{x},|\mathbf{x}|=s} \nu^{o,\pi}(\mathbf{x})$ is the equivalent stationary measure. Under this construction, these probabilities are those for the Norton equivalent queue. Also, notice that the equivalent stationary measure $\nu^\pi(s)$ is also the stationary measure of the Norton equivalent queue with transition matrix P^π under policy π.

Let $\Pi := \{\pi : \mathcal{S} \to \mathcal{A}\}$ denote the set of stationary and deterministic policies.

Definition 3.2. The average gain induced by policy π is:

$$g(M,\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(s_t, \pi(s_t))]. \tag{7}$$

The optimal policy π^* achieves

$$g(M,\pi^*) := g^*(M) := \sup_{\pi\in\Pi} g(M,\pi). \tag{8}$$

3.3 Comparison between both MDPs

It should be clear that the original MDP M^o has a greater set of policies than the aggregate MDP M because it has more states. Therefore, $g^o(M^o,\pi^*) \geq g^*(M)$. However, if we only consider the set of policies in the original MDP M^o that take the same action (reject or accept) in all the states with the same total number of jobs, then optimal gains coincide. More precisely, let Π^o_{sum} be the subset of policies in M^o such that for all $\pi \in \Pi^o_{sum}, \pi(\mathbf{x}) = \pi(\mathbf{y})$ if $|\mathbf{x}| = |\mathbf{y}|$. Then, the stationary measure on M^o under any policy π in Π^o_{sum}, and the stationary measure under π on M satisfy $\sum_{\mathbf{x},|\mathbf{x}|=s} \nu^{o,\pi}(\mathbf{x}) = \nu^\pi(s)$. Therefore, we get for all π in Π^o_{sum},

$$\begin{aligned} g^o(M^o,\pi) &= \sum_{\mathbf{x}} \nu^{o,\pi}(\mathbf{x}) r^o(\mathbf{x}, \pi(\mathbf{x})) \\ &= \sum_{s} \sum_{\mathbf{x},|\mathbf{x}|=s} \nu^{o,\pi}(\mathbf{x}) r(s, \pi(s)) \\ &= \sum_{s} \nu^{\pi}(s) r(s, \pi(s)) = g(M,\pi). \end{aligned}$$

Now taking the maximum over all policies in Π^o_{sum} yields

$$\max_{\pi\in\Pi^o_{sum}} g^o(M^o,\pi) = g^*(M).$$

When the full state is not observable, the best one can aim for is to learn $\max_{\pi\in\Pi^o_{sum}} g^o(M^o,\pi)$. Therefore, in the following we will consider the regret with respect to this oblivious optimal gain $g^*(M)$.

3.4 Bias and diameter

In the following, we will heavily use the *bias* [27] of a policy on M (although it has no intuitive meaning relative to the original MDP) as well the *diameter* of the aggregate MDP.

Definition 3.3 (Diameter of an MDP)**.** Let $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary policy of any MDP 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

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

Again, let us point out that we will only consider the diameter on the aggregate MDP for the computations, as it is needed to control the bias terms of the aggregate MDP (see Appendix F). We will never need to consider the bias or the diameter of the original MDP.

4 Reinforcement Learning

This section presents how reinforcement learning works in this context. The basic idea is that the learner can observe the external behavior of the network but has no information about its internal behavior. Despite this lack of information, it decides which jobs to admit to eventually be as efficient as if it knew everything about the system.

4.1 What does the learner know?

- The learner has some static information about the system. It knows an upper-bound on the number of servers ($N \leq N_{\max}$) inside the network and on their service rates ($\forall i, \mu_i^o < \mu_i^{\max}$). This means that the partial uniformization constant $U_1 = \sum_{i=1}^{N_{\max}} \mu_i^{\max}$ is known by the learner.
- The expected cost in state-action pair (s, a) is unknown as it depends on the unknown parameter λ (see (5)). However, the cost parameters γ_{reject} and γ_{hold} are known.
- On the dynamic side, the learner can observe the external events, i.e., the arrivals of jobs and the global departure process (the learner does not see the individual departures from the different queues). This implies that at any time t, the total number of jobs in the system, s_t, is known to the learner and will be seen as the partially observed *state*.
- The learning algorithm knows T, i.e., the number of time steps where it can take observations and actions. Notice that this is not a strong requirement as one can make the algorithm oblivious to T by using a classical doubling trick on T [5].

All these assumptions are either classical in reinforcement learning (as for example the bounds on the parameters) or natural in this context.

4.2 When does the learner acts?

The queueing system evolves in continuous time. Its external events are the arrival and departure processes. We assume that the learning algorithm is equipped with an independent Poisson clock that ticks at rate U_1. The algorithm will take decisions at every tick of its own clock as well as at each arrival and departure.

Notice that since we assume that the network is a stable Jackson network, then, under its stationary state, the arrival process and the departure process are both Poisson with rate λ and independent (this is a classical property of stable Jackson networks, see for example Corollary 5.7.3 in [28]). Therefore, the total point process indicating when the learner acts forms a Poisson process with a unknown rate

$$U = 2\lambda + \sum_{i=1}^{N_{\max}} \mu_i^{\max}. \tag{9}$$

The rate U is used as the uniformization constant to see the continuous time MDP as a discrete time one as in Section 3.

4.3 How do we measure the learner performance?

We will use the classical definition on the regret to assess the learning performance.
Definition 4.1 (Regret)**.** The *regret* at time T of the learning algorithm $\mathcal{L}$ is

$$\mathrm{Reg}(M, \mathcal{L}, T) := Tg^*(M) - \sum_{t=1}^{T} r_t^{\mathcal{L}}. \tag{10}$$

Here, $g^*(M)$ is the oblivious optimal gain defined in (8). The reward $r_t^{\mathcal{L}}$ is the reward of the state visited at time t by the learning algorithm.

5 Learning Algorithm with Modules: UCRL-M

5.1 High-level description of the proposed algorithm

Our algorithm is *episodic*, *model-based* and *optimistic*. As we now discuss, it can seen as a tweaked version of the UCRL2 algorithm introduced in [5].

The interactions of the learner with the MDP M^o are decomposed into *episodes*. In each episode k, spanning over the interval $[t_k, t_{k+1} - 1]$, one admission policy π_k is used to control the network and the learner observes the system (arrivals and departures) while collecting rewards under π_k. At the end of the episode, the estimation of the true transition probabilities and rewards (the *model*), $\hat{p}_k$ and $\hat{r}_k$ respectively, as well as the *confidence region* $\mathcal{M}_k$ are updated using the samples collected during the episode. This gives $\hat{p}_{k+1}$, $\hat{r}_{k+1}$ and $\mathcal{M}_{k+1}$. The next policy π_{k+1} is the best policy for the best MDP inside the confidence region $\mathcal{M}_{k+1}$ (*optimism*); with 'best MDP', we mean the MDP that is returned by Extended Value Iteration (EVI) as discussed in Appendix B.

In our case with partial observations, the number of jobs at time t, $(s_t)_{t \le T}$ is not Markovian, therefore it does not provide enough information to make good estimates on the underlying MDP. Instead, we collect a set $\{s_1, \ldots, s_{\tau_{\mathrm{mix}}}\}$ of observations and try to learn using this extended information. If τ_{mix} is well chosen, i.e., larger than the mixing time of the MDP, then each subsequence $s_i, s_{i+\tau_{\mathrm{mix}}}, s_{i+2\tau_{\mathrm{mix}}}, \ldots$ forms an "almost" independent sequence and therefore can be used for statistical estimations.

Our learning algorithm is based on the following idea. It can be seen as a collection of τ_{mix} learning algorithms $\mathcal{L}_1, \dots, \mathcal{L}_{\tau_{\text{mix}}}$, using respectively the subsequence $(s_{i+k\tau_{\text{mix}}})_{k\in\mathbb{N}}$ of observations, which are called *modules* in the following. The behavior of the modules is illustrated in Figure 3. Each learning module $\mathcal{L}_i$ behaves similarly as the classical

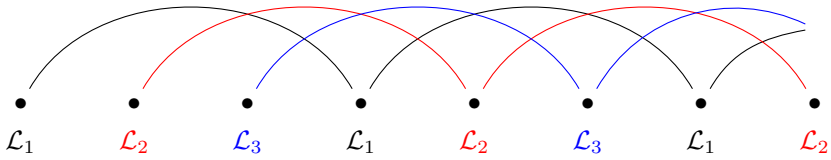


Figure 3: Illustration of the interleaving of the modules over time with $\tau_{\text{mix}} = 3$.

optimistic algorithm described above. There are no interactions between modules except for the number of visits that contributes to the construction of the global confidence region, as detailed in Section 5.3. The main technical difficulties in the control of the behavior of the algorithm are:

1. The learning modules $\mathcal{L}_1, \dots, \mathcal{L}_{\tau_{\text{mix}}}$ are not independent of each other, so one must be careful in assessing the interplay between the modules.
2. For each learning module $\mathcal{L}_i$, its own sequence of observations $s_{i+\tau_{\text{mix}}}, s_{i+2\tau_{\text{mix}}}, s_{i+3\tau_{\text{mix}}}, \dots$ is not really stationary and independent, but only weakly correlated.

Therefore, roughly speaking, our algorithm resembles τ_{mix} instances of UCRL2 [5] and the main contribution is the introduction of modules.

5.2 Number of modules: τ_{mix}

Let us first give a more precise definition of the modules, where their number τ_{mix} is yet to be chosen carefully. At the beginning of the algorithm, each time-step t is attributed a module m_t, so that these modules form a partition of the time-steps. For $0 \le t \le \tau_{\text{mix}} - 1$, the module m_t is defined in the following way: first $t \in m_t$, then we wait τ_{mix} steps to add the next time-step to that module, so that $t, t+\tau_{\text{mix}}, t+2\tau_{\text{mix}}, \dots \in m_t$, until time-step T is reached. More formally one can identify, $m_t = t \mod \tau_{\text{mix}}$.

The number of modules τ_{mix} is chosen using the following construction. Let us consider the original MDP under any policy π, with stationary measure $\nu^{o,\pi}$. There exists $C > 0$, $\rho \in (0,1)$ such that:

$$\max_{\pi\in\Pi} \sup_{x_0\in\mathcal{S}} \left\| \mathbb{P}_{x_0}^{o,\pi}(x_t = \cdot) - \nu^{o,\pi} \right\|_{TV} \le C\rho^t \quad \forall t > 0, \tag{11}$$

where $\mathbb{P}_{x_0}^{o,\pi}(x_t = \cdot)$ is the distribution of the state at time t under policy π in the original MDP, with starting state x_0. Let us then define

$$\tau_{\text{mix}} := \lceil 5 \log T / \log \rho^{-1} \rceil. \tag{12}$$

The reason for this precise choice will appear in the analysis of the regret (see Section 6) but the general idea behind this choice comes from Lemma D.1 given in appendix, that basically says that after τ_{mix} steps, the correlation between the state at time t and the state at time $t+\tau_{\text{mix}}$, under any policy, is smaller than $C'\rho^{\tau_{\text{mix}}}$, where C' is a constant.

The fact that the number of modules used by the algorithm depends on ρ can be seen as a weakness of our approach because it means that the learner needs to know *a priori* a bound on the mixing time of the unknown MDP. This point will be addressed in Section 7.

5.3 Confidence region

As mentioned earlier, our learning algorithm relies on the "Optimism in face of uncertainty" principle. Here, we provide the explicit construction of a confidence region $\mathcal{M}_k$ based on the observations, which depends on the visit counts. For each state-action pair (s,a) and each module m, let $N_{t_k}^{(m)}(s,a)$ be the cumulative number of visits to (s,a) at all times $t = m \mod \tau_{\text{mix}}$ smaller than t_k, and excluding the visits during the ramping phases Φ (see the UCRL-M algorithm).

We also define the *most frequent module* for each state-action pair (s,a): Let $m_k(s,a)$ be a module with the highest visit count until episode k,

$$m_k(s,a) \in \arg\max_m N_{t_k}^{(m)}(s,a), \tag{13}$$

so that for this module, the empirical observations are the most accurate, and we can relate the number of observations for this module to the total number of visits $N_{t_k}(s,a)$ of the pair (s,a) with the inequality: $N_{t_k}^{(m_k(s,a))}(s,a) \geq \frac{1}{\tau_{\text{mix}}} N_{t_k}(s,a)$.

To define the confidence region $\mathcal{M}_k$, first define $\hat{r}_k^{(m)}$ and $\hat{p}_k^{(m)}$ the empirical reward and transition estimates in module m:

$$\hat{p}_k^{(m)}(s'|s,a) := \frac{\sum_{t=1}^{t_k-1} \mathbb{1}_{\{s_t=s,a_t=a,s_{t+1}=s',m_t=m\}} \mathbb{1}_{\{t\notin\Phi\}}}{\max\left\{1, N_{t_k}^{(m)}(s,a)\right\}} \tag{14}$$

$$\hat{r}_k^{(m)}(s,a) := \hat{p}_k^{(m)}(s+1|s,a)\gamma_{\text{reject}} + \gamma_{\text{hold}}\frac{S-s}{U}, \tag{15}$$

where Φ is the set of the time steps in the ramping phases defined in the algorithm. $\mathcal{M}_k$ is the confidence set of MDPs whose rewards $\tilde{r}$ and transitions $\tilde{p}$ satisfy:

$$\forall (s,a), \quad \left|\tilde{r}(s,a) - \hat{r}_k^{(m_k(s,a))}(s,a)\right| \leq \delta_{\max}\sqrt{\frac{2\log(2At_k)}{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}}; \tag{16}$$

$$\forall (s,a), \quad \|\tilde{p}(\cdot \mid s,a) - \hat{p}_k^{(m_k(s,a))}(\cdot \mid s,a)\|_1 \leq \sqrt{\frac{8\log(2At_k)}{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{17}$$

Notice that for each state-action pair (s,a), we only need the empirical reward and transition estimates for the module $m_k(s,a)$: this means that the confidence region

$\mathcal{M}_k$ is built from the comparison between modules from (13), and we do not build a specific confidence region for each module.

The algorithm, formally defined below, finds the best optimistic MDP and policy within this confidence set using Extended Value Iteration, and executes the policy on the true MDP until the stopping criterion is met, that is when for any module m the number of visits $V_k^{(m)}(s,a)$ in the current episode of a state-action pair (s,a) reaches the number of visits of this pair and module until time t_k. More formally, if at episode k we choose the policy $\tilde{\pi}_k$, then the stopping criterion gives the following guarantee:

$$\forall (s,m) \quad V_k^{(m)}(s,\tilde{\pi}_k(s)) \leq \max\{1, N_{t_k}^{(m)}(s,\tilde{\pi}_k(s))\}. \tag{18}$$

5.4 UCRL-M: learning with $\tau_{\mathbf{mix}}$ modules

We are now in a position to give our learning algorithm, UCRL-M (Upper Confidence Reinforcement Learning with several Modules), in Algorithm 1. First, the algorithm initializes the different modules. Here, for each episode k and module m, it computes the empirical estimates of the reward and probability transition as in (15) and (14). Then, it applies Extended Value Iteration (EVI) (Appendix B) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ according to (19). Finally, to explore the MDP at episode k, it first iterates on the MDP over τ_{mix} time-steps and discards these samples (ramping phase) to start the observations closer to the stationary distribution of the current policy. This phase is necessary to guarantee that observations within a module are nearly independent. Afterwards, UCRL-M explores the true MDP with the optimistic policy $\tilde{\pi}_k$ and updates the empirical estimates with its observations.

The episode ends when the stopping criterion (18) is met. The next optimistic policy for the episode $k+1$ is found with respect to the observations inducing the confidence region $\mathcal{M}_k$ that is built using all modules (see (17)).

Algorithm 1: The UCRL-M algorithm.

Input: $\mathcal{S}$ and $\mathcal{A}$.

1 Set $t = 0$, $k = 0$;

2 **while** $t \leq T$ **do**

3 **Initialize** episode k with $t_k := t$;

4 **Compute** for all (s, a) the modules $m_k(s, a)$ according to (13);

5 **Compute** the confidence region $\mathcal{M}_k$ as in (17);

6 **Find** a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ with "Extended Value Iteration" such that

$$g(\tilde{M}_k, \tilde{\pi}_k) \geq \max_{M_k \in \mathcal{M}_k} \max_{\pi} g(M_k, \pi) - \frac{\delta_{\max}}{\sqrt{t_k}}. \tag{19}$$

7 **Ramping phase (Φ):** Iterate the MDP with policy $\tilde{\pi}_k$ for τ_{mix} time-steps, discard the observations and set $t := t + \tau_{\text{mix}}$.

8 **Exploration: while** *criterion* (18) *is true* **do**

1. Use policy $\tilde{\pi}_k$; // choose action $a_t = \tilde{\pi}_k(s_t)$ and observe state s_{t+1};
2. Set $t := t + 1$;

9 **end**

10 $k := k + 1$;

11 **end**

5.5 Time complexity of UCRL-M

Proposition 1. The time complexity of UCRL-M is $O(KS\tau_{\text{mix}} + Kt_{\text{evi}} + T)$, where K is the number of episodes and t_{evi} the time complexity of extended value iteration. Furthermore, $\mathbb{E}(K) = O(\log T)$.

Proof. The time complexity of lines 5 and 6 is $O(KS\tau_{\text{mix}})$. The complexity of line 7 is $O(Kt_{\text{evi}})$. The complexity of line 8 is $O(K\tau_{\text{mix}})$. The complexity of line 9 is $O(T - K\tau_{\text{mix}})$, the number of useful observations. As for the expected number of episodes, $\mathbb{E}K = O(\log T)$ because of the doubling trick used to end the episodes (see [5] for example). □

Note that the total number of useful samples (excluding the steps made during the ramping phases) is $T - K\tau_{\text{mix}}$, and each module uses $\frac{T - K\tau_{\text{mix}}}{\tau_{\text{mix}}}$ samples. As for the time complexity of EVI, each iteration of EVI is $O(S^3)$ and the number of iterations depends on the starting point and is more difficult to estimate. In total, the time complexity does not really depend on τ_{mix} or t_{evi} that only appears at the beginning of each episode, and the number of episodes is small w.r.t. T.

6 Regret of UCRL-M

6.1 Main result

Let us recall that S is the global bound on the number of jobs, $S' = S + 1$ is the number of states, γ_{reject} is the rejection cost, γ_{hold} is the unit-time holding cost and D is the diameter of the aggregate MDP. Also, $\nu^{\pi^{\max}}(s)$ is the stationary measure in the aggregate MDP under the policy that accepts all jobs, $\rho \in (0,1)$ is such that (11) holds true and $\mu(i)$ is the service rate in the aggregate MDP when i jobs are in the system.

Define the constant $C_1 := \prod_{i=1}^{i_0-1} \frac{\mu(i_0)}{\mu(i)} \geq 1$, where i_0 is chosen such that $\mu(i_0) > \lambda$. We notice that i_0 exists because the open (unconstrained) network defined in Section 2 is assumed to be stable (see Section 2) even with $S = +\infty$. Hence, the flow equivalent queue is also stable regardless of S. Define also $C_2 := \frac{(\lambda\gamma_{\text{reject}} + \gamma_{\text{hold}})C_1}{\mu(1)(1-\lambda/\mu(i_0))}$.

Theorem 6.1. *Let $M \in \mathcal{M}$. Define $Q_{\max} := \left(\frac{10C_2S'^2}{\nu^{\pi^{\max}}(S)}\right)^2 \log\left(\left(\frac{10C_2S'^2}{\nu^{\pi^{\max}}(S)}\right)^4\right)$. Define also the constant $\kappa = 228(\gamma_{\text{reject}} + \frac{\gamma_{\text{hold}}}{U})\frac{U}{\mu(1)}C_1\left(1 - \sqrt{\frac{\lambda}{\mu(i_0)}}\right)^{-3}$. For the choice $\tau_{\text{mix}} = 5\frac{\log T}{\log 1/\rho}$ and $A = 2$, and assuming $\tau_{\text{mix}}S \geq 2$ and $T > \max\left\{\frac{e^2}{4T}, \tau_{\text{mix}}\right\}$, it holds that*

$$\mathbb{E}\left[Reg(M, UCRL\text{-}M, T)\right] \leq \kappa \log(2T)\sqrt{T\log^{-1}(1/\rho)} + R_{\text{LO}}, \tag{20}$$

where $R_{\text{LO}} := 138 r_{\max} D^2 \max\left\{Q_{\max}, T^{1/4}\right\} \frac{\log^4(4T)}{\log^2 1/\rho}$ is a lower order term of the regret.

Before diving into the proof, which involves many technical points, let us comment on our result. In contrast with most bounds from the literature, the most remarkable point is that both the diameter and the size of the state space do not appear in our bound. These are both replaced by $\log^{-1/2}(1/\rho)$.

Although we do not know any explicit bounds on ρ for all possible networks, it is quite reasonable to predict that $\log^{-1/2}(1/\rho)$ can be of order $\sqrt{S}$. In fact, this can be shown for acyclic networks as well as for hyper-stable networks as it will be shown in Section 7.

This implies that the regret of UCRL-M is $\tilde{O}(\sqrt{ST})$, which is a major improvement over the best bound for general MDPs, namely $\tilde{O}(\sqrt{DSAT})$. This further confirms the fact that exploiting the structure of the learned system actually leads to more efficient algorithms as well as tighter analysis of their performance.

6.2 Outline of the proof

To compute the expected regret $\mathbb{E}[\text{Reg}]$, we will mainly follow the strategy from [5, Section 4]. First, we deal with the regret term corresponding to the initialization phase of each episode, which depends in the number of episodes. Then, for each episode k, we consider the case where the true MDP M does not belong to the confidence region $\mathcal{M}_k$, and use concentration inequalities along with the independence Lemma D.1 to show that this regret term will remain low. Then, we consider the case where the true MDP belongs to the confidence region, and for each episode, we split the regret into

relevant comparisons. Here, we expose terms depending on the difference of rewards and transitions between the true and optimistic MDPs, terms depending on the difference of biases, a term depending on the number of episodes and a term coming from the the computation of the optimistic policy and MDP with EVI.

To achieve the first split, we need to define: $R_k^{(m)}(s) := \sum_a V_k^{(m)}(s,a)(\rho^* - r(s,a))$ the regret at episode k induced by state s in module m, with $V_k^{(m_t)}(s,a)$ the number of visit of (s,a) during episode k in module m. We split the regret into terms where the true MDP belongs to the confidence region, terms where it does not, and the terms from initializing the episodes:

$$\mathbb{E}\left[\mathrm{Reg}\right] \leq \mathbb{E}\left[R_{\mathrm{in}}\right] + \mathbb{E}\left[R_{\mathrm{out}}\right] + \mathbb{E}\left[R_{\mathrm{ramp}}\right] \tag{21}$$

with K the number of episodes and the regret where the MDP is in the confidence region being $R_{\mathrm{in}} := \sum_m \sum_s \sum_{k=1}^K R_k^{(m)}(s)\mathbb{1}_{M\in\mathcal{M}_k}$, and when it is outside $R_{\mathrm{out}} := \sum_m \sum_s \sum_{k=1}^K R_k^{(m)}(s)\mathbb{1}_{M\notin\mathcal{M}_k}$ and the regret of the ramping phases $R_{\mathrm{ramp}} = \sum_k \sum_{t=t_k}^{t_k+\tau_{\mathrm{mix}}-1} r(s_t, \tilde{\pi}_k(s,t))$. Each term is then bounded as explained in Appendix A.

7 Controlling the regret bound parameter ρ

The efficiency of UCRL-M is critically based on controlling τ_{mix} and ρ. In particular, Theorem 6.1 says that the regret of UCRL-M depends on $W := \log^{-1/2}(1/\rho)$.

7.1 Bounds using mixing and coupling times

In Section 5, the number of modules τ_{mix} is defined as $\tau_{\mathrm{mix}} := 5\log T/\log\rho^{-1}$. where ρ is such that

$$\max_{\pi} \sup_{x_0\in\mathcal{S}} \left\| \mathbb{P}_{x_0}^{o,\pi}(x_t = \cdot) - \nu^o(\pi) \right\|_{TV} \leq C\rho^t \quad \forall t > 0. \tag{22}$$

Let us first recall classical results from Markov chain theory [29] relating ρ with the mixing and coupling time of a Markov chain. Let us consider any Markov chain with transition matrix P and stationary distribution ν (in our case, consider the Markov chain under the policy that attains the maximum in (22)). Let us define $d(t) := \sup_{x_0\in\mathcal{S}} \|\mathbb{P}_{x_0}(x_t = \cdot) - \nu\|_{TV}$. Then, the *mixing time* of the chain is defined as $t_{mix} := \min\{t : d(t) \leq 1/4\}$.

A classical bound on ρ is then obtained by using the mixing time:

$$\rho \leq \frac{1}{2^{t_{mix}^{-1}}} \tag{23}$$

This implies that $W \leq \sqrt{t_{mix}\log(2)}$.

Another bound on ρ can be obtained by using the coupling time. The coupling time is $\tau_{x,y} := \min\{t : X_t = Y_t\}$. If X_t and Y_t are coupled and start at $X_0 = x$ and $Y_0 = y$ respectively. Then, $d(t) \leq \max_{x,y} \mathbb{P}(\tau_{x,y} > t)$. By using Markov inequality, this implies that

$$t_{mix} \leq 4 \max_{x,y} \mathbb{E}[\tau_{x,y}]. \tag{24}$$

Therefore, a bound on the expected coupling time translates into a bound on ρ.

7.1.1 Acyclic networks

In our model, if the queueing network is acyclic, then the coupling time is controllable because whenever a queue couples it stays coupled forever.

More precisely, since the total number of states in the network increases with the admission threshold, the threshold policy under which the coupling time is the largest is when all jobs are admitted. Under this policy, by monotonicity, the coupling time is upper bounded by the coupling in an open network where all the N queues have buffers bounded by S. In this case, the coupling time has been studied in [30, Theorem 5.3], where the following result is proved in the stable case. Using our notation,

$$\max_{x,y} \mathbb{E}[\tau_{x,y}] \leq \sum_{i=1}^{N} \frac{U^2}{(\lambda_i^o + \mu_i^o)(\mu_i^o - \lambda_i^o)} S, \tag{25}$$

where U is the uniformization constant and $(\lambda_i)_{i \leq N}$ is the solution of the traffic equations.

According to Equations (23) and (24), this induces the following bound on the term W in the regret:

$$W \leq \kappa_0 \sqrt{NS},$$

where κ_0 is a constant: $\kappa_0 = \max_i \sum_{i=1}^{N} \frac{U}{\lambda_i^o + \mu_i^o}$.

7.1.2 Hyperstable networks

This is another type of networks for which an explicit bound on the coupling time exists. A network is called *hyperstable* if for each queue i, $\sum_j L_{ji} \mu_j^o + L_{0i} \lambda < \mu_i^o$.

As in the acyclic case, the threshold policy under which the coupling time is the largest is when all jobs are admitted. Under this policy, as for the acyclic case, the coupling time is upper bounded by the coupling in an open network where all the N queues have buffers bounded by S.

Coupling times of hyperstable networks with finite buffer queues have been studied in [31], where the following bound is given (Theorem 2):

$$\max_{x,y} \mathbb{E}[\tau_{x,y}] \leq \kappa_2 N^2 S \sum_{i=1}^{N} \frac{\lambda_i^o}{\mu_i^o - \lambda_i^o}, \tag{26}$$

where κ_2 is a constant. Using Equations (23) and (24), this induces a similar bound on the term W in the regret:

$$W \leq \kappa_3 N \sqrt{S},$$

where κ_3 is yet another constant.

7.2 Making the algorithm oblivious to ρ

By construction, the current version of UCRL-M uses explicitly $\tau_{\text{mix}} = 5 \log T / \log \rho^{-1}$ modules. This can be a problem as it implies an *a priori* knowledge of ρ, and of the mixing time (or at least an upper bound) of the network being learned.

These types of assumptions are sometimes made in the reinforcement learning literature. For example, the UCBVI algorithm [32] requires the knowledge of the diameter of the MDP being learned.

Here, we can patch UCRL-M to make it oblivious to ρ by making sure that $\tau_{\text{mix}} \geq 5 \log T / \log \rho^{-1}$ for any large enough T. For example, one can chose $\tau_{\text{mix}} := \log^2(T)$, as it is asymptotically larger than the previous one. This patch adds a multiplicative $\log(T)$ term in the asymptotic bound of the regret given in Theorem 6.1.

8 Numerical experiments

8.1 A multi-tier queueing network

To assess the performance of UCRL-M, we rely on a standard multi-tier queueing network as displayed in Figure 4. The topology of this network is composed of three tiers. Namely, tiers 1, 2 and 3 represent the web, application and database stages of a typical web-application request. Each tier is composed of multiple servers, each with its own queue. After accessing the web tier, a request may either return back to the issuing user with probability $1 - p$ or flow through the application and database tiers. This multi-tier structure is common in empirical studies of computer systems [33] and is the default architecture of web applications deployed on Amazon Elastic Compute Cloud (EC2) [34].

This model may be studied as an example of the generic case described in Section 2. Notice that given the routing from Figure 4, the stability condition is met if $\frac{\lambda}{1-p} < \mu$, where $\mu = \mu_1^o = \ldots = \mu_{3n}^o$ is the service rate of the queues in the network.

8.2 Regret of UCRL-M on the multi-tier queueing network

We provide the performance of UCRL-M over the queueing network described above when the number of queues per tier n and the total number of jobs S vary. In Figure 5, we display the average regret over 66 runs of the UCRL-M algorithm when n varies, and with parameters scaling with n to keep the systems proportionally comparable. More precisely, the scaling in S and μ is such that as the number of queues increases, the waiting time in each tier remains roughly identical for a job in each tier, and the scaling in the holding cost is also consistent with the increase of the number of jobs in the system. Notice that for our choice of parameters, the network is not stable, so that we use the UCRL-M algorithm under more general conditions than those assumed in Section 7 and even in Section 2.

In Figure 5, we remark that as we let the number of queues n (and the number of jobs S) scales multiplicatively, the regret increases as $\log(S)$. Knowing that the dependency in S of the regret bound from Theorem 6.1 mainly comes from ρ, this is much slower than the square root bounds given in Section 7 (under strong assumptions). This can be interpreted as the bound of Equation (22) being too large as it considers

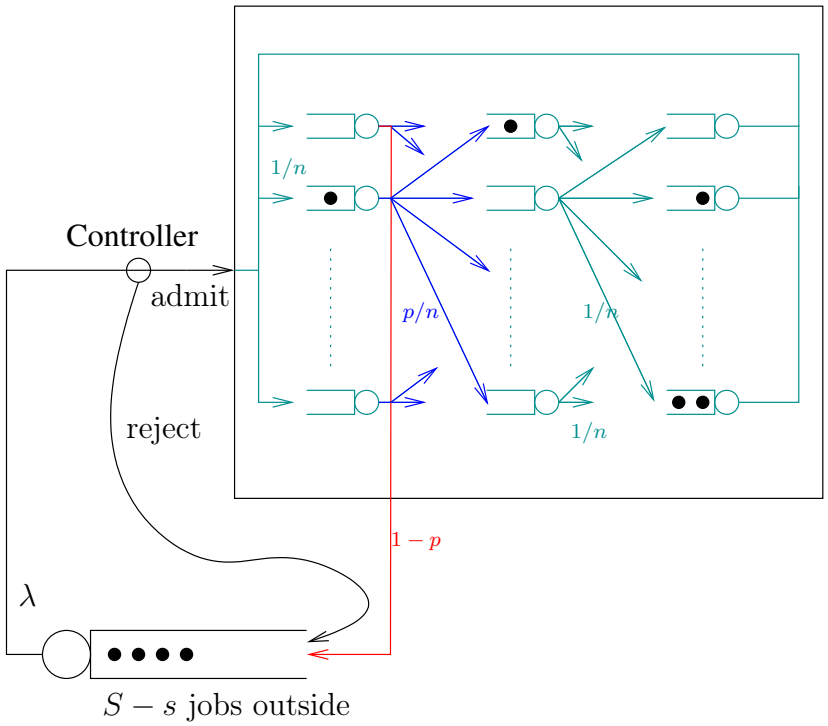
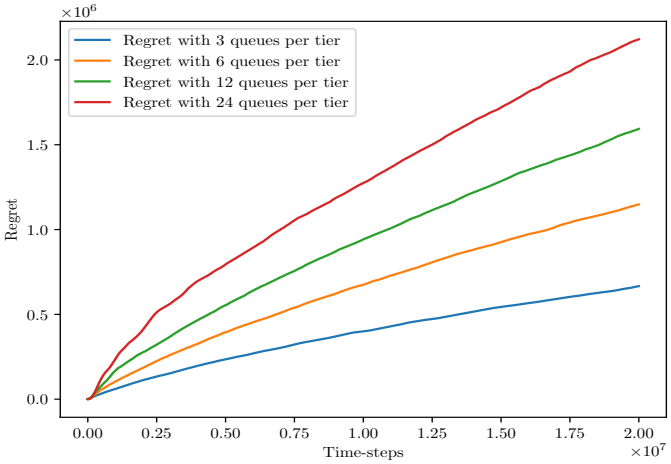


Figure 4: A queueing network model with three interconnected tiers. Each tier contains n queues and the total capacity is of S jobs.

the mixing from the worst state, while on average it is more likely for the algorithm to mix from states that are visited the most, which are already close to stationary states.

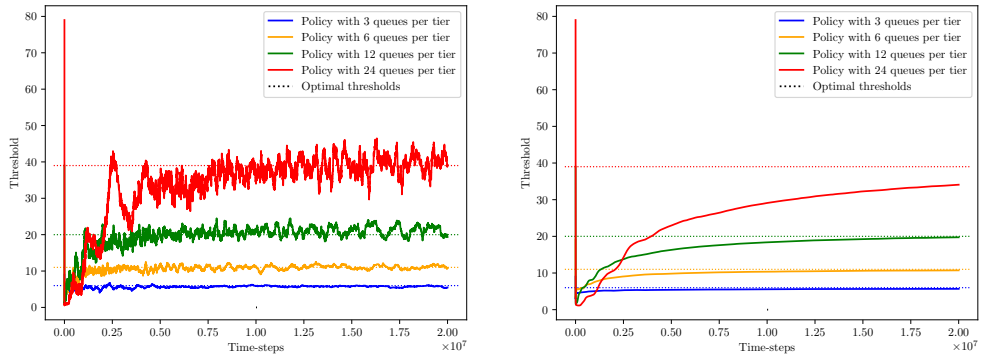
We see in Figure 6 that the chosen policy does not converge to the optimal threshold, as the algorithm needs to ensure exploration phases. Its Cesàro sum however does converge to the optimal threshold, for each value of n. It suggests that the optimal threshold is scaling linearly with n, and that the convergence is slower as S increases.

In the previous experiments, the number of modules is arbitrarily fixed to $\tau_{\text{mix}} = 3$. Now, we perform another experiment to observe the dependency of the regret in the choice of τ_{mix} for this queueing system. The intuition is the following: as explained in the high-level description of the algorithm in Subsection 5.1, UCRL-M could be compared to τ_{mix} instances of UCRL2 [5], where all modules but the best one is discarded at each episode. This best module runs on roughly $\frac{T}{\tau_{\text{mix}}}$ time-steps, and its regret can be compared to $\frac{1}{\tau_{\text{mix}}}$ times the expected regret of UCRL-M. With this intuition in mind, we plot in Figure 7 the regret of UCRL-M, where we rescaled both the regret and the time-steps by a factor $\frac{1}{\tau_{\text{mix}}}$.



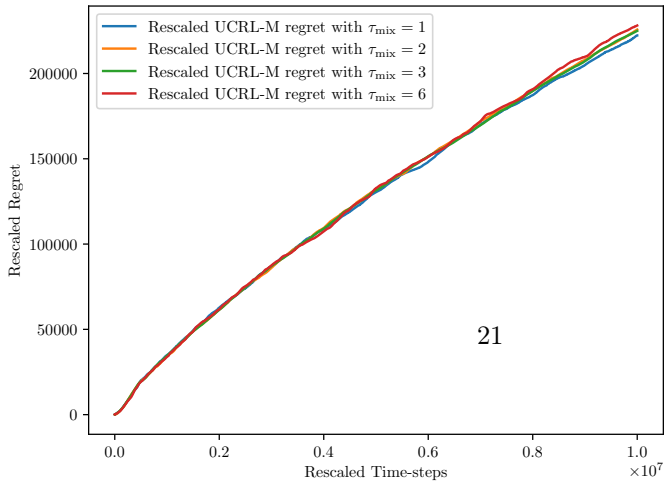
Parameter	Value
S	$\frac{10}{3}n$
λ	0.99
μ	$\frac{1}{n}$
γ_{reject}	10
γ_{hold}	$\frac{4}{3n}$
U	$\lambda + 3n\mu$
p	0.2
τ_{mix}	3

Figure 5: Regret of the UCRL-M algorithm on the queueing network for different values of n and scaling parameters.



(a) Average threshold over time. (b) Cesàro sum of the average threshold.

Figure 6: Comparison between the average threshold and its Cesàro sum.



Parameter	Value
S	20
n	6
λ	0.99
μ	$\frac{1}{n}$
γ_{reject}	10
γ_{hold}	$\frac{4}{3n}$
U	$\lambda + 3n\mu$
p	0.2

Figure 7: Rescaled regret of the UCRL-M algorithm on the queueing network for different values of τ_{mix}.

Within the considered queueing model, we notice that the modules do not seem to bring any practical upside because the regret is almost perfectly linear in the number of modules. In this particular example, the observations behave as if they were independent even if the algorithm only uses a single module. Intuitively, the system remains close to stationarity despite the policy changes, which could explain the limited effect of the modules. However, they remain necessary to guarantee the correctness of the confidence sets and to get the theoretical bound on the regret given in Theorem 6.1.

9 Conclusion

In the context of queueing networks, we have shown that efficient learning in POMDPs is possible. Provided that the learner's objective is to learn the optimal admission control policy, which is a problem appearing in a number of applications as discussed in Section 2, we have proposed UCRL-M, an optimistic algorithm whose regret is independent of the diameter D, i.e., a quantity that appears in most of the existing regret analyses [5] and that is exponential in the size of the space S in most queueing systems.

While our result strongly relies on Norton's equivalence theorem, which only applies exactly to product-form queueing networks, our main perspective is that this type of results under partial observations may be found in several other models from queueing theory. In fact, Norton's theorem has been generalized to multiclass networks [35] and also used in the context of non-product-form queueing networks for approximate analysis [25, 36].

References

[1] Walkins, C.J.: Learning from delayed rewards. PhD thesis, Cambridge University (1989)

[2] Jin, C., Yang, Z., Wang, Z., Jordan, M.I.: Provably efficient reinforcement learning with linear function approximation. In: Abernethy, J., Agarwal, S. (eds.) Proceedings of Thirty Third Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 125, pp. 2137–2143 (2020). https://proceedings.mlr.press/v125/jin20a.html

[3] Ouyang, Y., Gagrani, M., Nayyar, A., Jain, R.: Learning unknown Markov decision processes: A Thompson sampling approach. arXiv preprint arXiv:1709.04570 (2017)

[4] Ouyang, Y., Gagrani, M., Nayyar, A., Jain, R.: Learning unknown markov decision processes: A thompson sampling approach. In: 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA (2017)

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

[6] Fruit, R., Pirotta, M., Lazaric, A., Ortner, R.: Efficient bias-span-constrained exploration-exploitation in reinforcement learning (2018)

[7] Tossou, A., Basu, D., Dimitrakakis, C.: Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities (2019). https://doi.org/10.48550/ARXIV.1905.12425 . https://arxiv.org/abs/1905.12425

[8] Wei, C.-Y., Jahromi, M.J., Luo, H., Sharma, H., Jain, R.: Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10170–10180 (2020). https://proceedings.mlr.press/v119/wei20c.html

[9] Jin, C., Kakade, S., Krishnamurthy, A., Liu, Q.: Sample-efficient reinforcement learning of undercomplete pomdps. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 18530–18539 (2020)

[10] Azizzadenesheli, K., Lazaric, A., Anandkumar, A.: Reinforcement learning of POMDPs using spectral methods. In: COLT. JMLR Workshop and Conference Proceedings, vol. 49, pp. 193–256 (2016)

[11] Guo, Z.D., Doroudi, S., Brunskill, E.: A PAC RL algorithm for episodic pomdps. In: AISTATS. JMLR Workshop and Conference Proceedings, vol. 51, pp. 510–518 (2016)

[12] Papadimitriou, C.H., Tsitsiklis, J.N.: The complexity of markov decision processes. Math. Oper. Res. **12**(3), 441–450 (1987)

[13] Vlassis, N., Littman, M.L., Barber, D.: On the computational complexity of stochastic controller optimization in pomdps. ACM Trans. Comput. Theory **4**(4) (2012) https://doi.org/10.1145/2382559.2382563

[14] Even-Dar, E., Kakade, S.M., Mansour, Y.: Reinforcement learning in POMDPs without resets. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. IJCAI'05, pp. 690–695, San Francisco, CA, USA (2005)

[15] Ross, S., Chaib-draa, B., Pineau, J.: Bayes-adaptive POMDPs. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20 (2007). https://proceedings.neurips.cc/paper/2007/file/3b3dbaf68507998acd6a5a5254ab2d76-Paper.pdf

[16] Poupart, P., Vlassis, N.A.: Model-based bayesian reinforcement learning in partially observable domains. In: International Symposium on Artificial Intelligence and Mathematics (2008)

[17] Anselmi, J., Gaujal, B., Rebuffi, L.-S.: Reinforcement Learning in a Birth and

Death Process: Breaking the Dependence on the State Space. In: NeurIPS 2022 - 36th Conference on Neural Information Processing Systems, New Orleans, United States (2022). https://hal.science/hal-03799394

[18] Borgs, C., Chayes, J., Doroudi, S., Harchol-Balter, M., Xu, K.: The optimal admission threshold in observable queues with state dependent pricing. In: Probability in the Engineering and Informational Sciences 28, pp. 101–110 (2014). https://www.microsoft.com/en-us/research/publication/optimal-admission-threshold-observable-queues-state-dependent-pricing/

[19] Xia, L.: Event-based optimization of admission control in open queueing networks. Discrete Event Dynamic Systems **24**(2), 133–151 (2014) https://doi.org/10.1007/s10626-013-0167-1

[20] Chandy, K.M., Herzog, U., Woo, L.: Parametric analysis of queuing networks. IBM Journal of Research and Development **19**(1), 36–42 (1975) https://doi.org/10.1147/rd.191.0036

[21] Rolia, J.A., Sevcik, K.C.: The method of layers. IEEE Transactions on Software Engineering **21**(8), 689–700 (1995) https://doi.org/10.1109/32.403785

[22] Rolia, J., Casale, G., Krishnamurthy, D., Dawson, S., Kraft, S.: Predictive modelling of SAP ERP applications: Challenges and solutions. VALUETOOLS '09, Brussels, BEL (2009). https://doi.org/10.4108/ICST.VALUETOOLS2009.7988 . https://doi.org/10.4108/ICST.VALUETOOLS2009.7988

[23] Configuring Concurrency in Knative. https://knative.dev/docs/serving/autoscaling/concurrency/. Online; accessed: 2023-01-30 (2022)

[24] Wang, R., Casale, G., Filieri, A.: Estimating multiclass service demand distributions using markovian arrival processes. ACM Trans. Model. Comput. Simul. (2022) https://doi.org/10.1145/3570924

[25] Krieger, U.R.: Queueing networks and markov chains, 2nd edition by g. bolch, s. greiner, h. de meer, and k.s. trivedi. IIE Transactions **40**(5), 567–568 (2008) https://doi.org/10.1080/07408170701623187

[26] Kameda, H.: A property of normalization constants for closed queueing networks. IEEE Transactions on Software Engineering **SE-10**(6), 856–857 (1984) https://doi.org/10.1109/TSE.1984.5010314

[27] Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming, (2014)

[28] Ross, S.: Stochastic Processes. Wiley series in probability and mathematical statistics. Wiley, ??? (1983)

[29] Levin, D.A., Peres, Y., Wilmer, E.L.: Markov Chains and Mixing Times, (2008)

[30] Dopper, J.G., Gaujal, B., Vincent, J.-M.: Bounds for the coupling time in queueing networks perfect simulation. In: Langville, A.N., Stewart, W.J. (eds.) MAM, 150th Anniversary of A.A. Markov, Charleston, SC (2006)

[31] Anselmi, J., Gaujal, B.: Efficiency of simulation in monotone hyper-stable queueing networks. Queueing Systems **76**(1), 51–72 (2014) https://doi.org/10.1007/s11134-013-9357-7

[32] Azar, M.G., Osband, I., Munos, R.: Minimax regret bounds for reinforcement learning. In: International Conference on Machine Learning, pp. 263–272 (2017)

[33] Urgaonkar, B., Pacifici, G., Shenoy, P., Spreitzer, M., Tantawi, A.: An analytical model for multi-tier internet services and its applications. SIGMETRICS Perform. Eval. Rev. **33**(1), 291–302 (2005) https://doi.org/10.1145/1071690.1064252

[34] AWS Architecture Center. https://aws.amazon.com/architecture. Online; accessed: 2023-06-19 (2022)

[35] Kritzinger, P.S., van Wyk, S., Krzesinski, A.E.: A generalisation of norton's theorem for multiclass queueing networks. Performance Evaluation **2**(2), 98–107 (1982) https://doi.org/10.1016/0166-5316(82)90002-5

[36] Anselmi, J., Casale, G., Cremonesi, P.: Approximate solution of multiclass queuing networks with region constraints. In: 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, pp. 225–230 (2007). https://doi.org/10.1109/MASCOTS.2007.10

[37] Bhandari, J., Russo, D., Singal, R.: A finite time analysis of temporal difference learning with linear function approximation. In: Bubeck, S., Perchet, V., Rigollet, P. (eds.) Proceedings of the 31st Conference On Learning Theory. Proceedings of Machine Learning Research, vol. 75, pp. 1691–1692 (2018). https://proceedings.mlr.press/v75/bhandari18a.html

[38] Ipsen, I.C.F., Meyer, C.D.: Uniform stability of markov chains. SIAM Journal on Matrix Analysis and Applications **15**, 1061–1074 (1994)

Appendix A Proof of Theorem 6.1

The proof of Theorem 6.1 is quite cumbersome and is decompsed into many lemmas. The scheme in Figure A1 gives an overall picture of the interactions between them.

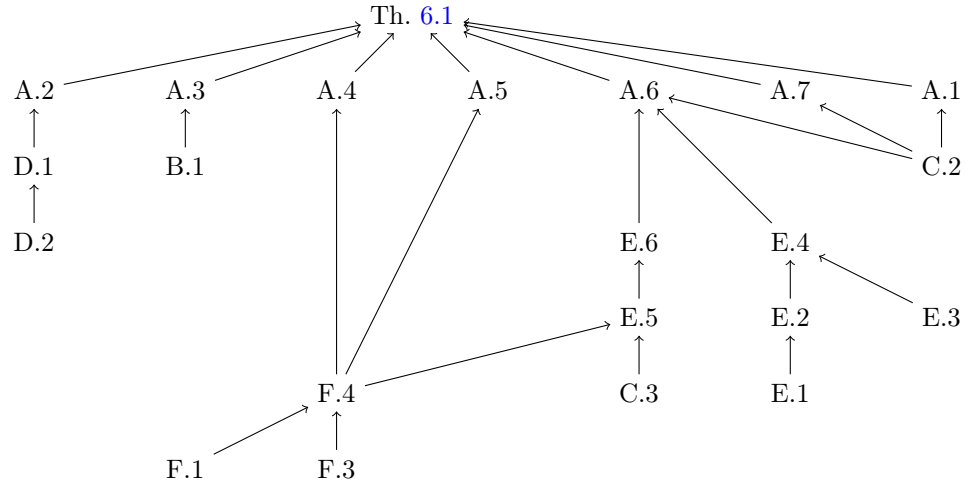


Figure A1: Dependencies in the proof of Theorem 6.1

A.1 Terms for the ramping phases

We first briefly deal with the terms coming from the ramping phases Φ of the episodes, R_{ramp}. We have:

$$R_{\text{ramp}} = \sum_k \sum_{t=t_k}^{t_k+\tau_{\text{mix}}-1} r(s_t, \tilde{\pi}_k(s,t)) \leq K\tau_{\text{mix}} r_{\max} \leq r_{\max} SA\tau_{\text{mix}}^2 \log_2\left(\frac{8T}{SA\tau_{\text{mix}}}\right), \tag{A1}$$

where in the last inequality we used Lemma C.2. Assuming $\tau_{\text{mix}} SA \geq 4$, and using $\log(2) \geq \frac{1}{2}$, we rewrite it:

$$R_{\text{ramp}} \leq 2 r_{\max} SA\tau_{\text{mix}}^2 \log(2T). \tag{A2}$$

This term is therefore among the lower-order terms of the regret.

A.2 Terms in the confidence bound

We start with the terms coming from the case where the MDP is out of the confidence regions $\mathcal{M}_k$. For each episode k, we define:

- $V_k^{(m)}(s)$ the number of visits to state s during episode k in module m.
- $N_t^{(m)}(s)$ is the number of visits to state s until time-step t excluded, in module m.

- $\mathcal{M}(t)$ 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:

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

We now need Lemma D.2 to control the probability that the MDP fails to be within the confidence bounds $\mathbb{P}\{M \notin \mathcal{M}(t)\}$. Taking the expectations and using Lemma D.2, we obtain

$$\mathbb{E}[R_{\text{out}}] \leq r_{\max} \sum_{t=1}^{T} t \mathbb{P}\{M \notin \mathcal{M}(t)\} \leq r_{\max} \sum_{t=1}^{T} \frac{S' + 16CS'A}{2t^2} \leq r_{\max}(S' + 16CS'A). \tag{A3}$$

This term is constant in T and therefore it does not significantly contribute to the regret.

A.3 Split of confidence bound

We assume that $M \in \mathcal{M}_k$ and to simplify the notations, we will omit the use of the indicator functions $\mathbb{1}_{M \in \mathcal{M}_k}$. For each episode k and module m, let us define

- $R_{\text{in},k}^{(m)} := \sum_s R_k^{(m)}$,
- $\tilde{\pi}_k$ the optimistic policy,
- $\tilde{P}_k := (\tilde{p}_k(s'|s, \tilde{\pi}_k(s)))$ the transition matrix of policy $\tilde{\pi}_k$ on the optimistic MDP $\tilde{M}_k$,
- $\mathbf{v}_k^{(m)} := (V_k(s, \tilde{\pi}_k)$ the row vector of visit counts,
- $\mathbf{h}_k$ the bias vector of the Markov chain in the true MDP M with policy $\tilde{\pi}_k$.

Now, we split the regret term $R_{\text{in}}^{(m)}$ into subterms that have different meaning. Assuming $M \in \mathcal{M}_k$ and using Lemma B.1 on the accuracy of EVI, we get:

$$\begin{aligned}
R_{\text{in},k}^{(m)} &= \sum_{s,a} V_k^{(m)}(s,a)(g^* - r(s,a)) \\
&\leq \sum_{s,a} V_k^{(m)}(s,a)(\tilde{g}_k - r(s,a)) + \varepsilon_k \sum_{s,a} V_k^{(m)}(s,a)
\end{aligned}$$

$$= \sum_{s,a} V_k^{(m)}(s,a)(\tilde{g}_k - \tilde{r}_k(s,a)) + \sum_{s,a} V_k^{(m)}(s,a)(\tilde{r}_k(s,a) - r(s,a) + \varepsilon_k).$$

In the next few steps, we will focus on rewriting the first sum. With (B18) and using the definition of the iterated values from EVI, we have for a given state s and $a_s := \tilde{\pi}_k(s)$:

$$\left| (\tilde{g}_k - \tilde{r}_k(s,a_s)) - \left(\sum_{s'} \tilde{p}_k(s'|s,a_s) u_i^{(k)}(s') - u_i^{(k)}(s) \right) \right| \leq \varepsilon_k,$$

so that:

$$R_{\text{in},k}^{(m)} \leq \mathbf{v}_k^{(m)} \left(\tilde{\mathbf{P}}_k - \mathbf{I} \right) \mathbf{u}_i + \sum_{s,a} V_k^{(m)}(s,a)(\tilde{r}_k(s,a) - r(s,a)) + \varepsilon_k \sum_{s,a} V_k^{(m)}(s,a).$$

Again, with $\tilde{\mathbf{h}}_k$ being the bias of the average optimal policy for the optimist MDP, define:

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

Then for any s: $|d_k(s)| \leq \varepsilon_k$.

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:

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

so that, using the definition of ε_k, we have that overall:

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

We can already further simplify the term related to EVI. Notice that:

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

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

$$R_{\text{EVI}}^{(m)} \leq \delta_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k^{(m)}(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{A4}$$

Let us now deal with the term $R_{\text{rewards}}^{(m)}$, as it will be bounded by a similar term as in equation (A4). Indeed, as $M \in \mathcal{M}_k$, we may use that both the optimistic and true rewards are within the confidence region from equation 16, and use that $t_k < T$, so that:

$$R_{\text{rewards}}^{(m)} \leq \delta_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}} \tag{A5}$$

On the other hand, we can also split more precisely the term that depends on the bias. Define P_k as the transition matrix of the optimistic policy $\tilde{\pi}_k$ in the true MDP M. We get

$$\begin{aligned} R_{\text{in}}^{(m)} \leq & \underbrace{\sum_k \mathbf{v}_k^{(m)} \Big(\tilde{\mathbf{P}}_k - \mathbf{P}_k\Big)\mathbf{h}_k}_{R_{\text{trans}}^{(m)}} + \underbrace{\sum_k \mathbf{v}_k^{(m)} \Big(\tilde{\mathbf{P}}_k - \mathbf{P}_k\Big)\Big(\tilde{\mathbf{h}}_k - \mathbf{h}_k\Big)}_{R_{\text{diff}}^{(m)}} + \underbrace{\sum_k \mathbf{v}_k^{(m)} (\mathbf{P}_k - \mathbf{I})\tilde{\mathbf{h}}_k}_{R_{\text{ep}}^{(m)}} \\ & \underbrace{+ \delta_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k^{(m)}(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}}_{R_{\text{EVI}}^{(m)} + R_{\text{rewards}}^{(m)}}. \end{aligned} \tag{A6}$$

Now that we split the regret into several terms, we still need to sum over the modules and analyze for each term its contribution to the regret. For instance, we can sum over the modules the terms depending on EVI and the reward differences to get:

$$R_{\text{EVI}} + R_{\text{rewards}} = \delta_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{A7}$$

This term is related to the choice of the confidence bounds, and it will contribute to the main term of the regret. Regarding the other terms, $R_{\text{trans}}^{(m)}$ will also use the confidence bounds on the transition as well as our knowledge of the bias in the true MDP. $R_{\text{diff}}^{(m)}$ will be a lower order term in the regret, using the confidence bounds for both the comparisons between the transitions and the biases. Finally, $R_{\text{ep}}^{(m)}$ will be related to the count of episodes, so that it will also be a lower order term. The discussion for each of these terms will be spread over the next subsections.

A.4 Bound on $R_{\text{trans}}^{(m)}$

To bound $R_{\text{trans}}^{(m)}$, we can follow the computations from [17]. We will use our knowledge of the bias $\mathbf{h}_k$ and the control on the transitions in the optimistic MDP to simplify the regret term.

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

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

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

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

where Δ is the difference of bias in the last inequality, we used the bound on the variations of the bias from Proposition F.4, and that the optimistic MDP has transitions close to the true transitions with inequality (17). Notice that the final term looks similar to the term coming from EVI and rewards related computations (A7). We will deal with these terms together in the next subsection, as they are both mainly contributing to the regret.

A.5 Bound on the main term

In the previous Section A.4, we have shown that:

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

Summing over the modules m, we get:

$$R_{\text{trans}} \leq 4\sqrt{2\log\left(2AT\right)} \sum_{s,a} \frac{\Delta(s+1) V_k(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{A8}$$

We now wish to control this term, R_{EVI} and R_{rewards} using our knowledge of the bias, rather than bounding it directly with the diameter D. We first sum over the episodes and take the expectation, so that with Lemma C.1, and using that $N_{t_k}^{(m_k(s,a))}(s,a) \geq$

$\frac{1}{\tau_{\text{mix}}} N_{t_k}(s,a)$ we had from Equation (13), we get:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{\sqrt{\tau_{\text{mix}}}V_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] \leq 3\mathbb{E}\left[\sum_{s,a}\sqrt{\tau_{\text{mix}}N_T(s,a)}\right]$$
$$\leq 3\sum_{s}\sqrt{\tau_{\text{mix}}\mathbb{E}\left[N_T(s)\right]A}, \qquad \text{by Jensen's inequality.}$$

Therefore:

$$R_{\text{trans}} \leq 12\sqrt{2A\tau_{\text{mix}}\log(2AT)}\sum_{s=0}^{S}\Delta(s+1)\sqrt{\mathbb{E}\left[N_T(s)\right]}. \tag{A9}$$

This is one of the terms mainly contributing to the regret, the other one being, doing similar computations:

$$R_{\text{EVI}} + R_{\text{rewards}} \leq 12\delta_{\max}\sqrt{2A\tau_{\text{mix}}\log(2AT)}\sum_{s\geq 0}\sqrt{\mathbb{E}\left[N_T(s)\right]} \tag{A10}$$

Now, let $N_T^{\pi^{\max}}$ be the number of visits when the starting state is sampled randomly from the initial distribution $\nu^{\pi^{\max}}$ and the policy $\pi^{\max}$ is always chosen. By stochastic ordering, as $N_T(s) \leq_{st} N_T^{\pi^{\max}}$, we have $\mathbb{E}\left[N_T(s)\right] \leq \mathbb{E}\left[N_T^{\pi^{\max}}\right] = T\nu^{\pi^{\max}}(s)$. We can therefore rewrite the main contributing term to the regret as:

$$12\sqrt{2A\tau_{\text{mix}}T\log(2A\tau_{\text{mix}}T)}\sum_{s=0}^{S}\left(\Delta(s+1)+\delta_{\max}\right)\sqrt{\nu^{\pi^{\max}}(s)}. \tag{A11}$$

Replace in the equation the choice $\tau_{\text{mix}} = 5\log T/\log\rho^{-1}$ and recall that we had, from Proposition F.4, $\Delta(s) := 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\sum_{i=1}^{s}\frac{U}{\mu(i)} \leq 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\frac{U}{\mu(1)}s$. Using Lemma F.1, since

$$\sum_{s=0}^{S}\left(\Delta(s+1)+\delta_{\max}\right)\sqrt{\nu^{\pi^{\max}}(s)} \leq 3\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\frac{U}{\mu(1)}\sum_{s=0}^{S}(s+1)\sqrt{\nu^{\pi^{\max}}(s)}$$
$$\leq 3\delta_{\max}\nu^{\pi^{\max}}(0)^{-1/2}\frac{U}{\mu(1)}\sqrt{C_1}\sum_{s\geq 0}s\left(\frac{\lambda}{\mu(i_0)}\right)^{s/2}$$
$$\leq 3\delta_{\max}\nu^{\pi^{\max}}(0)^{-1/2}\frac{U}{\mu(1)}\sqrt{C_1}\frac{1}{\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^2}$$
$$\leq 3\delta_{\max}\frac{U}{\mu(1)}C_1\frac{1}{\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^3},$$

then, assuming $\tau_{\text{mix}} \leq T$, the main term is upper bounded by:

$$72\delta_{\max}\frac{U}{\mu(1)}C_1\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^{-3}\log(AT)\sqrt{5AT\log^{-1}(\rho^{-1})}. \tag{A12}$$

A.6 Bound on $R_{\text{diff}}^{(m)}$

We now deal with the term involving the difference of bias $R_{\text{diff}}^{(m)}$, defined in equation A6. The proof mainly follows the one from [17], with a final tweak to relate the visits from a module to the total number of visits. Notice that we cannot directly use the confidence regions to control the difference between $\tilde{\mathbf{h}}_k$ and $\mathbf{h}_k$, so that we will need Lemma E.4, and we are interested in controlling $\|\tilde{\mathbf{h}}_k - \mathbf{h}_k\|_\infty$.

Fix the module m and the episode k, with policy $\tilde{\pi}_k$. Choose a state minimizing $N_{t_k}^{(m_k(s,\tilde{\pi}_k(s)))}(s,\tilde{\pi}_k(s))$, and call this state x_k, $a_k := \tilde{\pi}_k(x_k)$ and $m' := m_k(x_k,a_k)$: for this state, the confidence bounds are at their worst, and $\sqrt{\frac{\log(2At_k)}{\max\{1,N_{t_k}^{(m')}(x_k,a_k)\}}}$ is maximal for episode k. This means that controlling the number of visits of the worst state lets us control the number of visits for any state. As the true MDP is within the confidence bounds, with a triangle inequality we get:

$$\|\tilde{P}_k - P_k\|_\infty \leq 4\sqrt{\frac{2\log(2At_k)}{\max\{1,N_{t_k}^{(m')}(x_k,a_k)\}}}.$$

We now want to use Lemma E.4. In our case, notice that in the true MDP we have $D \geq T_{hit}^{\tilde{\pi}_k} \geq 1$ for S large enough. Remark also that $D^{\tilde{\pi}_k}$ can be replaced by D in the last inequality of the proof of E.4, as $\text{span}(h^{\tilde{\pi}_k}) \leq D$ by construction of $\tilde{\pi}_k$ with EVI, following the same argument as in [5, Equation (11)].

$$\|\tilde{\mathbf{h}}_k - \mathbf{h}_k\|_\infty \leq 8r_{\max}D^2\sqrt{\frac{2\log(2At_k)}{\max\{1,N_{t_k}^{(m')}(x_k,a_k)\}}}. \tag{A13}$$

Hence,

$$\begin{aligned}
R_{\text{diff}}^{(m)} &\leq \sum_s\sum_{s'} V_k^{(m)}(s,\tilde{\pi}_k(s))\cdot(\tilde{p}_k(s'|s,\tilde{\pi}_k(s)) - p(s'|s,\tilde{\pi}_k(s)))\cdot(\tilde{h}_k(s') - h_k(s')) \\
&\leq \sum_s V_k^{(m)}(s,\tilde{\pi}_k(s))\cdot\|\tilde{p}_k(\cdot|s,\tilde{\pi}_k(s)) - p(\cdot|s,\tilde{\pi}_k(s))\|_1\,\|\tilde{\mathbf{h}}_k - \mathbf{h}_k\|_\infty \\
&\leq 32D^2 r_{\max}\log(2AT)\,\Sigma^{(m)},
\end{aligned}$$

where in the last inequality we have used (A13) and defined

$$\Sigma^{(m)} := \sum_{s,a}\sum_k\sum_{t=t_k}^{t_{k+1}-1}\frac{\mathbb{1}_{\{s_t,a_t=s,a\}}\mathbb{1}_{\{t\in m\}}}{\sqrt{\max\{1,N_{t_k}^{(m_k(s,a))}(s,a)\}}\sqrt{\max\{1,N_{t_k}^{(m')}(x_k,a_k)\}}}.$$

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

$$\Sigma \leq \sum_m \sum_{s,a} \sum_k \sum_{t=t_k}^{t_{k+1}-1} \frac{\mathbb{1}_{\{s_t, a_t = s,a\}} \mathbb{1}_{\{t \in m\}}}{\max\{1, N_{t_k}^{(m')}(x_k, a_k)\}} = \sum_k \frac{I_k}{\max\{1, N_{t_k}^{(m')}(x_k, a_k)\}}.$$

Now, define $Q_{\max} := \left(\frac{10 C_2 S'^2}{\nu^{\pi^{\max}}(S)}\right)^2 \log\left(\left(\frac{10 C_2 S'^2}{\nu^{\pi^{\max}}(S)}\right)^4\right)$ where we defined the constant $C_2 = \frac{(\lambda \gamma_{\text{reject}} + \gamma_{\text{hold}}) C_1}{\mu(1)(1 - \lambda/\mu(i_0))}$, 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:

$$\Sigma \leq K_{\leq I} I(T) + \sum_{k, I_k > I(T)} \frac{I_k}{\max\{1, N_{t_k}^{(m')}(x_k, a_k)\}}.$$

Using the stopping criterion for episodes, and that we have chosen the module m' in equation (13) to have the inequality $V_k^{(m')}(x_k, a_k) \geq \frac{1}{\tau_{\text{mix}}} V_k(x_k, a_k)$:

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

Now we can end the computations as in [17]. Denote by $\mathcal{E}$ the event:

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

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

$$\begin{aligned}
\Sigma &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \sum_{k, I_k > I(T)} \frac{2\tau_{\text{mix}}}{\nu^{\pi^{\max}}(S)} + \mathbb{1}_{\bar{\mathcal{E}}} \sum_{k, I_k > I(T)} \tau_{\text{mix}} I_k \\
&\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} (K - K_{\leq I}) \frac{2\tau_{\text{mix}}}{\nu^{\pi^{\max}}(S)} + \mathbb{1}_{\bar{\mathcal{E}}} \tau_{\text{mix}} T.
\end{aligned}$$

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

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_{\leq I}\right] I(T) + \mathbb{E}\left[(K - K_{\leq I})\right] \frac{2\tau_{\text{mix}}}{\nu^{\pi^{\max}}(S)} + \frac{\tau_{\text{mix}}}{4}.$$

Now using Lemma C.2, $S'A \geq 4$, $I(T) \geq \frac{2}{\nu^{\pi^{\max}}(S)}$ and that $\frac{1}{\log 2} + \frac{1}{4} \leq 2$:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K\right] I(T) \tau_{\text{mix}} + \frac{\tau_{\text{mix}}}{4} \leq 2 S' A \tau_{\text{mix}} \log(2AT) I(T).$$

Therefore, we have that:

$$\mathbb{E}\left[R_{\text{diff}}^{(m)}\right] \leq 64 r_{\max} S' A D^2 \tau_{\text{mix}} I(T) \log^2(2AT). \tag{A14}$$

A.7 Bound on R_{ep}

The last regret term we have to bound is related to the count of episodes.

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

We first want to sum over the modules to get the same kind of term as in [5], written as a martingale difference sequence, and then take the expectation. Following that proof, we define $X_t := (p(\cdot|s_t, a_t) - \mathbf{e}_{s_t}) \tilde{\mathbf{h}}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}}$, 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. We obtain

$$\begin{aligned}\sum_c \mathbf{v}_k^{(m)} (\mathbf{P}_k - \mathbf{I}) \tilde{\mathbf{h}}_{\mathbf{k}} \leq \mathbf{v}_k (\mathbf{P}_k - \mathbf{I}) \tilde{\mathbf{h}}_{\mathbf{k}} &\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + \tilde{\mathbf{h}}_k(s_{t_{k+1}}) - \tilde{\mathbf{h}}_k(s_{t_k}) \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + D r_{\max},\end{aligned}$$

and by summing over the episodes we get

$$\sum_m R_{\text{ep}}^{(m)} \leq \sum_{t=1}^{T} X_t + K D r_{\max}.$$

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

On the other hand, using Lemma C.2 on the number of episodes, when taking the expectation we obtain

$$\mathbb{E}\left[\sum_m R_{\text{ep}}^{(m)}\right] \leq S' A \tau_{\text{mix}} \log_2\left(\frac{8T}{S' A \tau_{\text{mix}}}\right) \cdot D r_{\max}.$$

As for the computation of (A2), assuming $\tau_{\text{mix}} S' A \geq 4$:

$$\mathbb{E}[R_{\text{ep}}] \leq 2 r_{\max} S' A D \tau_{\text{mix}} \log(2AT). \tag{A15}$$

A.8 Total sum

We remind that we showed in subsection A.5 that the main term of the regret is:

$$72\delta_{\max}\frac{U}{\mu(1)}C_1\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^{-3}\log(AT)\sqrt{5AT\log^{-1}(\rho^{-1})},$$

and now it remains to compute the lower order term of the regret R_{LO}. Using (A2), (A3), (A14) and (A15), the lower order term of the regret is upper bounded by, omitting the $r_{\max}$ factor:

$$64S'AD^2\tau_{\mathrm{mix}}I(T)\log^2(2AT)+2S'A\tau_{\mathrm{mix}}(D+\tau_{\mathrm{mix}})\log(2AT)+(S'+16CS'A),$$

and for T large enough so that $1+16C\leq\log^2(T)$ the upper bound is :

$$r_{\max}69S'AD^2\tau_{\mathrm{mix}}^2I(T)\log^2(2AT),$$

which concludes the proof of Theorem 6.1.

Appendix B Lemmas on Extended Value Iteration

We remind the fundamental properties of the Extended Value Iteration (EVI) algorithm, first described in [5], which is used to find the optimistic MDP $\tilde{M}_k$ and the policy $\tilde{\pi}_k$ for each episode k given a confidence region $\mathcal{M}_k$. These properties are useful notably in the first splits of the regret terms in Section A.3. EVI iteratively computes values in the following way:

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

where $\mathcal{P}(s,a)$ is the set of probabilities from (17), and the iterations are stopped with respect to the following lemma [5, Theorem 7].

Lemma B.1. *For episode k and accuracy $\varepsilon_k:=\frac{\delta_{\max}}{\sqrt{t_k}}$, denote by i the last step of extended value iteration, stopped when:*

$$\max_s\{u_{i+1}^{(k)}(s)-u_i^{(k)}(s)\}-\min_s\{u_{i+1}^{(k)}(s)-u_i^{(k)}(s)\}<\varepsilon_k. \tag{B16}$$

The optimistic MDP $\tilde{M}_k$ and the optimistic policy $\tilde{\pi}_k$ at the last step of EVI are so that the gain is ε_k- close to the optimal gain:

$$\tilde{g}_k:=\min_s g(\tilde{M}_k,\tilde{\pi}_k,s)\geq\max_{M'\in\mathcal{M}_k,\pi,s'}g(M',\pi,s')-\varepsilon_k. \tag{B17}$$

Moreover, from [27, Theorem 8.5.6]:

$$\left|u_{i+1}^{(k)}(s) - u_i^{(k)}(s) - \tilde{g}_k\right| \leq \varepsilon_k, \tag{B18}$$

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

$$\tilde{h}_k(s_0) = \mathbb{E}_{s_0}\left[\sum_{t=0}^{\infty}(\tilde{r}(s_t, a_t) - \tilde{g}_k)\right]. \tag{B19}$$

Rather than using the last value of EVI in the computations of the regret, we rely on the bias to show that the last value and the optimistic bias are nearly equal, up to a translation. By choosing iteration i large enough, from [27, Equation 8.2.5], we can ensure that:

$$\left|u_i^{(k)}(s) - (i-1)\tilde{g}_k - \tilde{h}_k(s)\right| < \frac{\varepsilon_k}{2}, \tag{B20}$$

so that we can define the following difference

$$d_k(s) := \left|u_i^{(k)}(s) - \min_s u_i^{(k)}(s) - \left(\tilde{h}_k(s) - \min_s \tilde{h}_k(s)\right)\right| < \varepsilon_k. \tag{B21}$$

Appendix C Classical lemmas for the regret computation in UCRL-like algorithms

We introduce classic lemmas from [5] that are needed for the regret computations. The first lemma, proven in [5, Appendix C.3], is used to simplify the main regret terms (A7) and (A8).

Lemma C.1. *For any fixed state action pair (s, a), and time T, we have:*

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

The next lemma, proven in [5, Appendix C.2] is useful to bound the term from in A.1 and in equation (A14).

Lemma C.2. *Denote by K_t the number of episodes up to time t, and let $t > SA\tau_{\text{mix}}$. It is bounded by:*

$$K_t \leq S'A\tau_{\text{mix}} \log_2\left(\frac{8t}{S'A\tau_{\text{mix}}}\right).$$

The next lemma is needed in the proof of Lemma E.5. While it includes the diameter, this will only impact a lower-order term of the regret.

Lemma C.3 (Azuma-Hoeffding inequality)**.** *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leq RD$ for all i and some $R > 0$. Then, for all $\varepsilon > 0$ and $n \in \mathbb{N}$:*

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

Appendix D Probability of not being in the confidence region

We compute the probability that the true MDP M fails to be in the confidence set. This lemma controls the corresponding regret terms in Section A.2 when we consider the episodes k with $M \notin \mathcal{M}_k$.

Let us first prove the key Lemma D.1.

Lemma D.1. *Let us consider the original MDP under any policy π, with stationary measure ν^π. There exists $C > 0$, $\rho \in (0,1)$ such that:*

$$\max_{\pi \in \Pi} \sup_{x_0 \in \mathcal{S}} \|\mathbb{P}^\pi{}_{x_0}(x_t = \cdot) - \nu^\pi\|_{TV} \leq C\rho^t \quad \forall t > 0. \tag{D22}$$

Let $t, t' > 0$ such that $t' - \tau_{\text{mix}} \geq t$, with t' and $t' - \tau_{\text{mix}}$ belonging to the same episode. Let X be a function of the state of the original MDP until time t and Y function of the state of the original MDP from time t'. Let $\hat{Y}$ be a random variable following the same distribution as Y independently from X. Let f be a real-valued, bounded function. Then:

$$\left|\mathbb{E}\left[f(X,Y)\right] - \mathbb{E}\left[f(X,\hat{Y})\right]\right| \leq 4C\|f\|_\infty \rho^{\tau_{\text{mix}}}.$$

Proof. The proof is essentially the same as in [37, Lemma 9], but as states are sampled from the original MDP and not a single Markov chain, we cannot just assume that the starting distribution at time 0 is a stationary distribution. Instead, we have to make sure that it is the case for each start of the episodes, hence the initial phase where τ_{mix} samples of the aggregate MDP are discarded, so that the original MDP is close to its stationary distribution. Due to this ramping time, we can make sure that t' and $t' - \tau_{\text{mix}}$ belong to the same episodes and can therefore be related to the same stationary distribution.

Let $t, t' > 0$ such that $t' - \tau_{\text{mix}} \geq t$, with t' and $t' - \tau_{\text{mix}}$ belonging to the same episode k. Now, X is a function of the state of the original MDP until time t, so there are t observed transitions but there might be many more that are hidden. In turn, Y is a function of the state of the original MDP from time t'. Let $\hat{Y}$ be a random variable following the same distribution as Y and independent of X. Note that there are at least τ_{mix} observed or hidden transitions between t and t' on the original MDP.

We also define the distribution $P := \mathbb{P}\{X \in \cdot, Y \in \cdot\}$ and the distribution $Q := \mathbb{P}\{X \in \cdot\} \otimes \mathbb{P}\{Y \in \cdot\}$, and we define the total variation information $I_{TV}(X,Y) := \sum_x \mathbb{P}\{X = x\} \|\mathbb{P}\{Y = \cdot \mid X = x\} - \mathbb{P}\{Y = y\}\|_{TV}$. To simplify, assume that $\|f\|_\infty \leq \frac{1}{2}$. By definition of the total variation distance, we first have that:

$$\left|\mathbb{E}\left[f(X,Y)\right] - \mathbb{E}\left[f(X,\hat{Y})\right]\right| \leq \|P - Q\|_{TV},$$

Then, using the properties of the total variation information related to a Markov chain described in [37], we obtain

$$\|P - Q\|_{TV} \leq I_{TV}(X,Y) \leq I_{TV}(x_t, x_{t'}) \leq I_{TV}(x_{t'-\tau_{\text{mix}}}, x_{t'})$$

$$\leq \sum_x \mathbb{P}\left\{x_{t'-\tau_{\text{mix}}} = x\right\} \left\|\mathbb{P}\left\{x_{t'} = \cdot \mid x_{t'-\tau_{\text{mix}}} = x\right\} - \mathbb{P}\left\{x_t = \cdot\right\}\right\|_{TV}$$

then using a triangle inequality:

$$\left\|\mathbb{P}\left\{x_{t'} = \cdot \mid x_{t'-\tau_{\text{mix}}} = x\right\} - \mathbb{P}\left\{x_t = \cdot\right\}\right\|_{TV} \leq \left\|\mathbb{P}\left\{x_{t'} = \cdot\right\} - \nu^{\tilde{\pi}_k}\right\|_{TV} + \left\|\mathbb{P}\left\{x_{t'} = \cdot \mid x_{t'-\tau_{\text{mix}}} = x\right\} - \nu^{\tilde{\pi}_k}\right\|_{TV},$$

we get

$$\|P - Q\|_{TV} \leq 2C\rho^{\tau_{\text{mix}}},$$

where in the last inequality we used assumption (11) twice, as t' and $t' - \tau_{\text{mix}}$ belong to the same episode, and therefore can be related to the same stationary measure $\nu^{\tilde{\pi}_k}$. To clarify, the exponent τ_{mix} in the inequality is loose, as τ_{mix} is the number of time-steps in the aggregate MDP, so there are at least as many time steps in the original MDP, and the mixing is confirmed. □

We can now give the lemma that actually shows that M is likely to be in the confidence set of MDPs.

Lemma D.2. *For $t > 1$, the probability that the MDP M is not within the set of plausible MDPs $\mathcal{M}(t)$ is bounded by:*

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \leq \frac{S'}{2t^3} + \frac{8CS'A}{t^3}.$$

Compared to [5, Lemma 17], we notice that the first term comes from the choice of the confidence bound adapted to the birth and death structure of the MDP, but the second one comes from the imperfect independence of the observations. To prove this inequality, we will need Lemma D.1 to consider independent events again, and to be able to use concentration inequalities.

Let us now prove Lemma D.2.

Proof. Fix a state-action pair (s, a), m any module and n the number of visits of this pair within the module before time t. We will first consider the confidence around the empirical transitions, and then the confidence around the rewards. Let $\varepsilon_p = \sqrt{\frac{2}{n}\log\left(16At^4\right)} \leq \sqrt{\frac{8}{n}\log\left(2At\right)}$. Define the events:

$$A_n = \left(\|\hat{p}^{(m)}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log\left(2At\right)}\right) \tag{D23}$$

Here, we aim to control these events but the difficulty is that the observations from the state-action pairs are not independent. On the other hand, we notice that the observations within a fixed module are nearly independent, which is why we needed to introduce these modules in the first place.

Define $\hat{p}^{\perp}(\cdot|s,a)$ the empirical transition probabilities from n independent observations of the state-action pair (s,a). Define events that are copies of A_n but with independent observations:

$$A_n^{\perp} = \left(\|\hat{p}^{\perp}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log(2At)} \right). \tag{D24}$$

Similarly, define $A_n^{\perp,k}$ events such that the first $n-k$ observations are the same as the ones for A_n and the next k observations are independent, so that for example $A_n^{\perp,0} = A_n$ and $A_n^{\perp,n-1} = A_n^{\perp}$. Then, applying $n-1$ times Lemma D.1:

$$\left|\mathbb{P}\left\{A_n\right\} - \mathbb{P}\left\{A_n^{\perp}\right\}\right| \leq \sum_{k=1}^{n-1} \left|\mathbb{P}\left\{A_n^{\perp,k-1}\right\} - \mathbb{P}\left\{A_n^{\perp,k}\right\}\right| \leq 4Cn\rho^{\tau_{\text{mix}}} \leq 4CT^{1-5}.$$

We can therefore work on the events with independent observations. Knowing that from each pair, there are at most 3 transitions, a Weissman's inequality gives:

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

and we get

$$\mathbb{P}\left\{A_n^{\perp}\right\} \leq \frac{3}{8At^4},$$

and within our choice of τ_{mix},

$$\mathbb{P}\left\{A_n\right\} \leq \frac{3}{8At^4} + \frac{4C}{t^4}.$$

We deal with the rewards in a similar manner. Define the events:

$$B_n := \left(|\hat{r}^{(m)}(s,a) - r(s,a)| \geq \delta_{\max}\sqrt{\frac{2}{n}\log(2At)} \right). \tag{D25}$$

By definition of $\hat{r}^{(m)}(s,a) = \gamma_{\text{reject}}\hat{p}^{(m)}(s+1|s,a) + \frac{\gamma_{\text{hold}}}{U}(S-s)$ 15, and using that $\gamma_{\text{reject}} \leq \delta_{\max}$, we can write:

$$\mathbb{P}\left\{B_n\right\} \leq \mathbb{P}\left\{ |\hat{p}^{(m)}(s+1|s,a) - p(s+1|s,a)| \geq \sqrt{\frac{2}{n}\log(2At)} \right\}.$$

Once again, we consider $\hat{p}^{\perp}(s+1|s,a)$ the empirical transition probabilities from independent observations of (s,a) to $s+1$, and we look to control the probability of the events $B_n^{\perp}$. With the independence, we may now use the following Hoeffding

inequality on the Bernoulli random variable of parameter $p(s+1|s,a)$:

$$\mathbb{P}\left\{|\hat{p}^{(m)}(s+1|s,a) - p(s+1|s,a)| \geq \varepsilon_r\right\} \leq 2\exp\left(-2n\varepsilon_r^2\right),$$

where $\varepsilon_r = \sqrt{\frac{1}{2n}\log\left(16At^4\right)} \leq \sqrt{\frac{2}{n}\log\left(2At\right)}$. We therefore get:

$$\mathbb{P}\left\{B_n^{\perp}\right\} \leq \frac{1}{8At^4},$$

and with the previous choice of τ_{mix},

$$\mathbb{P}\left\{B_n\right\} \leq \frac{1}{8At^4} + \frac{4C}{t^4}.$$

Overall:

$$\mathbb{P}\left\{A_n \cup B_n\right\} \leq \frac{1}{2At^4} + \frac{8C}{t^4}.$$

Now, with a union bound for all values of $n = \max\{1, N_t^{(m)}(s,a)\} \in \left\{0, 1, \cdots, \left\lceil \frac{t-1}{\tau_{\text{mix}}} \right\rceil\right\}$ and all τ_{mix} possible modules, and also summing over all state-action pairs:

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \leq \frac{S'}{2t^3} + \frac{8CS'A}{t^3}$$

as desired. □

Appendix E Lemmas specific to our regret computations

In this section, we prove generic properties on the difference of biases between two MDPs. This control on the difference is needed in subsection A.6 to compare the optimistic MDP and the true MDP.

E.1 Lemmas on the bias differences

The next three lemmas of this subsection are already proved in [17], for the sake of completeness, we rewrite them in this appendix. They are used in the proof of Lemma E.4, to control the difference between the bias of the policy $\tilde{\pi}_k$ in the optimistic MDP and in the true MDP.

Lemma E.1. *For an 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 $s, s' \in \mathcal{S}$ be recurrent states under policy π. Call $\tau_{s\to s'}$ the random time needed to reach state s' from state s. 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 E.2. *Consider two unichain MDPs M and M'. Let $r = r' \in [0, r_{\max}]$ and P, P' be the rewards and transition matrix of MDP M, M' under policy π, π' respectively, where both MDPs have the same state and action spaces. Denote by g, g' the average reward obtained under policy π, π' in the MDP M, M' respectively. Then the difference of the gains is upper bounded.*

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

Proof. Define for any state s the following correction term $b(s) := r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\|_1$. 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 recurrent state s for policy π:

$$\begin{aligned} \left((P - P')\sum_{t=0}^{T} P^t r\right)(s) &\leq \|p(\cdot|s) - p'(\cdot|s)\|_1 \cdot \text{span}\ (J(T)) \\ &\leq r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\|_1 \text{ by Lemma E.1} \end{aligned}$$

$$= b(s).$$

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:

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

As $r = r'$, with a multiplication by $\frac{1}{T+1}$ and by taking the Cesáro limit :

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

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

Lemma E.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 $S \times S$ is invertible and $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}\,\tau_{i \to S}$, where $\mathbb{E}\,\tau_{i \to S}$ is the expected time to reach state S from state i.

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

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

In the following lemma, we use the same notations as in Lemma E.2 with a common state space $\{0, 1, \dots S\}$.

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

$$\|h - h'\|_\infty \leq 2T^\pi_{hit} D^{\pi'} r_{\max} \|P - P'\|_\infty.$$

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 [38, 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 E.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, as $r = r'$:

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

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

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

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

E.2 Visits of the furthest state

We also need the next lemmas to bound R_{diff} by controlling the number of visits of the state with the fewest visits. If we can guarantee that each state receives enough visits, then we will have a good approximation of the biases and transition probabilities. The proof can be found in [17].

Lemma E.5. *Let $\nu^{\pi^{\max}}$ be the stationary measure of the Markov chain under policy $\pi^{\max}$, such that for every state s: $\pi^{\max}(s) = 1$, so that every job is admitted in the network until maximal capacity S is reached.*

Let k be an episode and assume that the length of this episode I_k is at least $I(T) = 1 + \max\left\{Q_{\max}, T^{1/4}\right\}$, with $Q_{\max} := \left(\frac{10C_2S'^2}{\nu^{\pi^{\max}}(S)}\right)^2 \log\left(\left(\frac{10C_2S'^2}{\nu^{\pi^{\max}}(S)}\right)^4\right)$, $C_2 := \frac{(\lambda\gamma_{\text{reject}}+\gamma_{\text{hold}})C_1}{\mu(1)(1-\lambda/\mu(i_0))}$ and C_1 as in Lemma F.1. Then, with probability at least $1 - \frac{1}{4T}$:

$$V_k(x_k, a_k) \geq \nu^{\pi^{\max}}(S) I_k - 5C_2S'^2\sqrt{I_k \log I_k}.$$

We will now prove Lemma E.5:

Proof. Let k be an episode such that $I_k \geq I(T)$, and first consider it is of fixed length I. Let $x_k \in \mathcal{S}$ be a recurrent state, $a_k = \tilde{\pi}_k(s_k)$. Denote by ν_k the stationary distribution under policy $\tilde{\pi}_k$. Notice that $\nu^{\pi^{\max}}(S) \leq \nu_k(x_k)$ for S large enough.

Define a new Markov reward process: consider again the original state space $\mathcal{S}'$ and the transitions p' with policy $\tilde{\pi}_k$, but the rewards $\mathring{r}$, where $\mathring{r}(s') = 1$ for states s' such that $|s'| = x_k$ and 0 otherwise. Denote by $\mathring{g}_{\tilde{\pi}_k}$ the gain associated to the policy $\tilde{\pi}_k$ and similarly define $\mathring{\mathbf{h}}_{\tilde{\pi}_k}$ the bias, translated so that $\mathring{h}_{\tilde{\pi}_k}(S) = 0$. Then:

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

By Azuma-Hoeffding inequality C.3, following the same proof as in section 4.3.2 of [5], notice that $X_u = \mathring{h}_{\tilde{\pi}_k}(s'_{u+1}) - \left\langle p'\left(\cdot \mid s'_u, \tilde{\pi}_k(s'_u)\right), \mathring{h}_{\tilde{\pi}_k}\right\rangle$ form a martingale difference sequence with the bound $|X_u| \leq \operatorname{span} \mathring{h}_{\tilde{\pi}_k}$:

$$\mathbb{P}\left\{\sum_{u=t_k}^{t_{k+1}-1} X_u \geq C_2S'^2\sqrt{10I \log I}\right\} \leq \frac{1}{I^5}.$$

With Proposition F.4 proved in Appendix F, we have $\operatorname{span} \mathring{h}_{\tilde{\pi}_k} \leq C_2S'^2$ with $C_2 = \frac{(\lambda\gamma_{\text{reject}}+\gamma_{\text{hold}})C_1}{\mu(1)(1-\lambda/\mu(i_0))}$, so that with probability at least $1 - \frac{1}{I^2}$:

$$V_k(x_k, a_k) \geq \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\tilde{\pi}_k} - 5C_2S'^2\sqrt{I \log I}.$$

On the other hand:

$$\sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\tilde{\pi}_k} = V_k(s_k, a_k)\nu_k(x_k),$$

so that, using that $\nu_k(x_k) \geq \nu^{\pi^{\max}}(S)$, with probability at least $1 - \frac{1}{I^5}$:

$$V_k(x_k, a_k) \geq \nu^{\pi^{\max}}(S)I - 5C_2S'^2\sqrt{I \log I}.$$

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

$$\begin{aligned}\mathbb{P}\left\{V_k(x_k, a_k) < \nu^{\pi^{\max}}(S)I_k - 5C_2S'^2\sqrt{I_k \log I_k}\right\} &\leq \sum_{I=I(T)+1}^{T} \frac{1}{I^5} \leq \sum_{I=T^{1/4}+1}^{T} \frac{1}{I^5} \\ &\leq \frac{1}{4T},\end{aligned}$$

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

$$V_k(x_k, a_k) \geq \nu^{\pi^{\max}}(S)I_k - 5C_2S'^2\sqrt{I_k \log I_k}.$$

□

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

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

$$V_k(x_k, a_k) \geq \frac{\nu^{\pi^{\max}}(S)}{2} I_k.$$

Proof. With Lemma E.5, it is enough to show that $5C_2S'^2\sqrt{I_k \log I_k} \leq \frac{\nu^{\pi^{\max}}(S)}{2} I_k$, *i.e.* that $\sqrt{\frac{I_k}{\log I_k}} \geq \frac{10C_2S'^2}{\nu^{\pi^{\max}}(S)} =: B$. 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. □

Appendix F Properties of the aggregate MDP

In this section, we prove properties on the aggregate MDP that are needed to control the average number of visits of the states of the MDP under any policy. We also prove a bound on the bias of the true MDP under any policy, which is eventually needed to control the main term in subsections A.4 and A.5.

F.1 Properties of the policies in the aggregate MDP

We may only consider policies that are threshold policies, as we are mainly interested in the average reward scored by these policies, so that we consider that the policies chosen by EVI are threshold policies. We remind that the aggregate MDP is stable (as seen in Section 2), so that there exists a i_0 large enough for which $i \geq i_0$, $\mu(i) \geq \mu(i_0) > \lambda$.

With the following lemma, we compute the stationary measures ν^π and give a comparison between any ν^π with the stationary measure $\nu^{\pi^{\max}}$ of the maximal policy $\pi^{\max}$, that admits every job into the queue, by relating these Markov chains to the $M/M/1/S$ queue with rates λ and $\mu(i_0)$.

Lemma F.1. *Denote by $\bar{s}$ the last recurrent state of the MDP for policy π, so that $\pi(s) = 0$ for $s \geq \bar{s}$. Define the constant $C_1 := \prod_{i=1}^{i_0-1} \frac{\mu(i_0)}{\mu(i)} \geq 1$, independent of S.*

We have the following inequalities

- *On the stationary measure of the maximal policy:*

$$\nu^{\pi^{\max}}(0)^{-1} := \sum_{s'=0}^{S} \prod_{i=1}^{s'} \frac{\lambda}{\mu(i)} \leq \frac{C_1}{1 - \frac{\lambda}{\mu(i_0)}},$$

- *On the stationary measure of any policy:*

$$\nu^{\pi}(0)^{-1} := \sum_{s'=0}^{\bar{s}} \prod_{i=1}^{s'} \frac{\lambda}{\mu(i)} \leq \nu^{\pi^{\max}}(0)^{-1} \leq \frac{C_1}{1 - \frac{\lambda}{\mu(i_0)}},$$

- *Also we can compute for $s \leq S$:*

$$\nu^{\pi^{\max}}(s) := \nu^{\pi^{\max}}(0) \prod_{i=1}^{s} \frac{\lambda}{\mu(i)} = \nu^{\pi^{\max}}(0) C_1 \left(\frac{\lambda}{\mu(i_0)}\right)^s.$$

We now remind a definition of the bias for any policy π and control its variations, as they play a major role in the computations of the main term of the regret (see A.4).

Definition F.2 (Bias)**.** Let π be a policy, P the transition matrix and ν^π the stationary measure of the Markov chain under policy π. The bias $\mathbf{h}^\pi$ of this policy is defined as:

$$h^\pi(s) = \sum_{t=1}^{\infty} (P^t(s, \cdot) - \nu^\pi)\mathbf{r}. \tag{F26}$$

In order to control the variation of the bias of any policy, we will relate the bias to the expected hitting time to hit the state 0 from state s, so that we first need to compute the hitting times:

Lemma F.3. *Let π be any policy, $(X_t)_t$ be the Markov chain with policy π and transitions P starting from any state s. Denote by τ_s the random time needed for X_t to hit 0. Then:*

$$\mathbb{E}\tau_s \leq \nu^\pi(0)^{-1} \sum_{i=1}^{s} \frac{U}{\mu(i)}$$

Proof. We write the expected hitting time equations, and use induction. Let τ_i be the hitting time to 0 starting from state i, and $\mathbf{e}$ be the unit vector. We have the system:

$$\mathbb{E}\boldsymbol{\tau} = \mathbf{e} + P\,\mathbb{E}\boldsymbol{\tau}, \tag{F27}$$

with the extra equation $\tau_0 = 0$. The system gives for $s \geq \bar{s}$:

$$\mathbb{E}\tau_s = 1 + \mathbb{E}\tau_s \frac{1-\mu(s)}{U} + \mathbb{E}\tau_{s-1}\frac{\mu(s)}{U},$$

so that:

$$\mathbb{E}\tau_s = \frac{U}{\mu(s)} + \mathbb{E}\tau_{s-1}$$

Then, by induction, we want to prove the equation for $s < \bar{s}$:

$$\mathbb{E}\tau_s = \mathbb{E}\tau_{s-1} + \frac{U}{\mu(s)} \sum_{s'=s}^{\bar{s}} \prod_{i=s+1}^{s'} \frac{\lambda}{\mu(i)} \tag{F28}$$

For $s < \bar{s}$, assume (F28) is true for $\mathbb{E}\tau_{s+1}$:

$$\begin{aligned}
\mathbb{E}\tau_s &= 1 + \mathbb{E}\tau_{s+1}\frac{\lambda}{U} + \mathbb{E}\tau_s \frac{1-\mu(s)-\lambda}{U} + \mathbb{E}\tau_{s-1}\frac{\mu(s)}{U} \\
&= 1 + \mathbb{E}\tau_s \frac{1-\mu(s)-\lambda}{U} + \mathbb{E}\tau_{s-1}\frac{\mu(s)}{U} + \mathbb{E}\tau_s \frac{\lambda}{U} + \sum_{s'=s+1}^{\bar{s}} \prod_{i=s+1}^{s'} \frac{\lambda}{\mu(i)} \\
&= \frac{U}{\mu(s)} + \mathbb{E}\tau_{s-1} + \frac{U}{\mu(s)} \sum_{s'=s+1}^{\bar{s}} \prod_{i=s+1}^{s'} \frac{\lambda}{\mu(i)} \text{ by gathering the } \tau_s \text{ terms} \\
&= \mathbb{E}\tau_{s-1} + \frac{U}{\mu(s)} \sum_{s'=s}^{\bar{s}} \prod_{i=s+1}^{s'} \frac{\lambda}{\mu(i)},
\end{aligned}$$

the induction is therefore true, and we have: $\mathbb{E}\tau_s \leq \mathbb{E}\tau_{s-1} + \frac{U}{\mu(s)}\nu^{\pi}(0)^{-1}$. □

Proposition F.4. *For any policy π, define for $s \in \{1, \dots, S\}$ the variation of the bias*

$$\Delta^{\pi}(s) := h^{\pi}(s) - h^{\pi}(s-1) = \sum_{t=1}^{\infty} \left(P^t(s,\cdot) - P^t(s-1,\cdot)\right)\mathbf{r}.$$

Remind that $\delta_{\max} := \max_{s,a,a'} |r(s,a) - r(s-1,a')| = \frac{\lambda\gamma_{\text{reject}} + \gamma_{\text{hold}}}{U}$:

$$\Delta^{\pi}(s) \leq \Delta(s) := 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1} \sum_{i=1}^{s} \frac{U}{\mu(i)}.$$

Using the monotonicity of the rates μ from Lemma 3.1, we therefore have :

$$\Delta(s) \leq 2(\lambda\gamma_{\text{reject}} + \gamma_{\text{hold}})\nu^{\pi^{\max}}(0)^{-1} s \frac{1}{\mu(1)}$$

Proof. We will use an optimal coupling, that is, a coupling such that the following infimum is reached, as defined in [29].

$$\left\|P^t(s,\cdot)-P^t(s-1,\cdot)\right\|_{\mathrm{TV}} = \inf\{\mathbb{P}(X_t \neq Y_t) : (X_t, Y_t) \text{ couples } P^t(s,\cdot) \text{ and } P^t(s-1,\cdot)\}. \tag{F29}$$

More precisely, let X and Y be Markov chains with transition matrix P and starting states $X_1 = s$, $Y_1 = s-1$, coupled in the following way: For each time-step $t \geq 2$, let $U_t \sim \mathcal{U}([0,1])$ be a sequence of independent random variables sampled uniformly on $[0,1]$. We have:

$$X_{t+1} = \begin{cases} X_t - 1 & \text{if } U_t \leq \mu(X_t) \\ X_t & \text{if } \mu(X_t) \leq U_t \leq 1-\lambda \\ X_t + 1 & \text{if } 1-\lambda \leq U_t \end{cases} \tag{F30}$$

and define Y_{t+1} the same way from Y_t. This coupling is optimal, but in particular we have:

$$(P^t(s,\cdot) - P^t(s-1,\cdot))\mathbf{r} \leq 2\mathbb{P}(X_t \neq Y_t)\delta_{\max}.$$

We remind that τ_s is the random time needed for the Markov chain X_t to hit 0. The coupling time is lower than τ_s:

$$\mathbb{P}(X_t \neq Y_t) \leq \mathbb{P}\left(\tau_{s\to 0} > t\right),$$

so that summing over t gives:

$$\Delta^\pi(s) \leq 2\delta_{\max}\mathbb{E}\tau_s,$$

and using Lemma F.3 and Lemma F.1:

$$\Delta^\pi(s) \leq 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\sum_{i=1}^{s}\frac{U}{\mu(i)}.$$

□