Reinforcement Learning Algorithms for Controlled Queueing Systems

Louis-Sébastien Rebuffi

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques et Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Algorithmes d'apprentissage par renforcement pour le contrôle de systèmes de files d'attente

Reinforcement Learning Algorithms for Controlled Queueing Systems

Présentée par :

Louis-Sébastien REBUFFI

Direction de thèse :

Bruno GAUJAL DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse
Jonathan ANSELMI INRIA	Co-encadrant de thèse

Rapporteurs :

GER KOOLE
PROFESSEUR, VRIJE UNIVERSITEIT AMSTERDAM
MATTHIEU JONCKHEERE
DIRECTEUR DE RECHERCHE, CNRS DELEGATION OCCITANIE OUEST

Thèse soutenue publiquement le **11 décembre 2023**, devant le jury composé de :

NADIA BRAUNER, PROFESSEURE DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Présidente
BRUNO GAUJAL, DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse
GER KOOLE, PROFESSEUR, VRIJE UNIVERSITEIT AMSTERDAM	Rapporteur
MATTHIEU JONCKHEERE, DIRECTEUR DE RECHERCHE, CNRS DELEGATION OCCITANIE OUEST	Rapporteur
ALAIN JEAN-MARIE, DIRECTEUR DE RECHERCHE, CENTRE INRIA D'UNIVERSITE COTE D'AZUR	Examinateur

Invités :

JONATHAN ANSELMI
CHARGE DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES

Remerciements

Merci d'abord Jonatha et Bruno d'avoir encadré ma thèse, de m'avoir permis de travailler avec un bon suivi. Nos discussions claires et rigoureuses m'ont efficacement guidé pendant ces trois années, j'ai pu beaucoup apprendre sur les MDPs, les files d'attente, et surtout sur l'univers du cyclisme ! Nos moments de travail réguliers, engagés mais détendus et légers, ont été importants pour mener à bien la thèse dans des conditions idéales.

Je remercie aussi Ger et Matthieu d'avoir accepté de lire et d'apporter leur précieux point de vue sur mon manuscrit, Alain pour ses commentaires détaillés et son suivi intéressé de ma thèse, et Nadia pour sa vision plus ouverte sur mes travaux.

Merci à mes collègues de bureau, d'équipe et du laboratoire d'avoir contribué à l'ambiance de travail détendue et même plaisante qui y régnait, en particulier à Victor et Julien pour les discussions aux cafés, Arnaud, Romain et Sebastian pour votre passion des jeux, et toute l'équipe pour la sympathie et l'accueil dont j'ai pu bénéficier au laboratoire. Vous avez su me permettre de trouver un bon équilibre entre le travail et les moments plus relâchés.

Pour tous les bons moments passés à Grenoble, je remercie mes coéquipiers de volley Baptiste, Charly, Manon, Pierre, Antho, Alex et l'équipe, et pour leur capacité sans pareille à organiser tant de pots et dîners ces dernières années.

Merci aux personnes rencontrées au cours de ma scolarité à diverses occasions, que j'ai souvent pu revoir, Julien, Lucas, Héloïse, Ioannis et Ugo pour les discussions non-mathématiques. Merci à Jad pour sa vivacité, son naturel et le partage de ses goûts artistiques. Merci à Corentin, Manon, Antoine, Théophile, Gabrielle, Gabrielle et Laurane pour avoir lancé ma carrière de volley, même si elle n'a pas encore vraiment décollé.

Merci enfin à ma famille, à mes frères Sylvestre, Jean-Baptiste et Pierre-Alexandre et mes parents Florence et Luigi.

Abstract

Although reinforcement learning has been recently primarily studied in the generic case of Markov decision processes, the queueing systems case stands out in particular. To deal with the potentially extremely large state space *a priori*, learning algorithms must take into account the structure of the systems in order to extract as much information as possible and choose the best control that optimizes the system performance in the long run.

In this thesis, we present algorithms adapted from classical algorithms in the context of queueing systems, and we study their performance to demonstrate a weak dependence on the state space compared to results obtained in the general case.

Résumé

Bien que l'apprentissage par renforcement ait été récemment principalement étudié dans le cas générique des processus de décisions markoviens, le cas des systèmes de files d'attente se dinstigue particulièrement. Pour compenser la taille de l'espace d'état qui peut être extrêmement grande *a priori*, les algorithmes d'apprentissage doivent tenir compte de la structure des systèmes afin d'en extraire le plus d'information et de choisir le meilleur contrôle qui optimisent au mieux les performances du système sur le long terme.

Dans cette thèse, nous présentons des algorithmes construits à partir d'algorithmes classiques, adaptés au contexte des système de file d'attente, et nous étudions les performances de ceux-ci pour montrer une dépendance faible à l'espace d'états comparativement aux résultats obtenus dans le cas général.

Contents

Notation Table

$\mathcal{A}$	Action space
A	Size of the action space $\mathcal{A}$
a	Action in $\mathcal{A}$
$a_{\max}$	Maximal action for $\mathcal{A} = [0, a_{\max}]$
$c(s,a)$	Mean cost in state s under action a
D	Diameter
$\mathbb{E}_s^\pi[\cdot]$	Expectation with initial state s under policy π
G^π	Gain under policy π in a continuous-time MDP (Chapter 3)
g^π	Gain under policy π
g^*	Optimal gain
$h^\pi(s)$	Bias under a policy π in state s
h^*	Bias of an optimal policy, defined up to a constant
$\partial h^\pi(s)$	Variation of the bias $h^\pi(s) - h^\pi(s-1)$
k	Episode index
K_t	Number of episodes until time t
L	Routing matrix of a queueing network
$\mathbb{L}$	Learning algorithm
m	Module index
$\mathcal{M}$	Class of MDPs
M	MDP in $\mathcal{M}$ to be learned
$\mathcal{M}_k$	Confidence set of MDPs at episode k
$N_t(s,a)$	Number of visits of the state-action pair (s,a) until time t
N	Number of queues in the network (Chapter 5)
$P(s'\|s,a)$	Transition probabilities
$p_{s,s'}$	Transition probabilities when a is omitted
$q(s'\|s,a)$	(or equiv. $q_{s,s'}(a)$) Transition rates
$r(s,a)$	Mean reward in state s under action a
$R(s,a)$	Random reward in state s under action a

$r_{\max}$	Maximal reward
$\mathcal{S}$	State space
S	Size of the state space $\mathcal{S}$
S'	$S' = S - 1$, usually the maximal queue size
s, s'	States in $\mathcal{S}$
T	Time horizon
t_k	Starting time of episode k
u_i	Value vector at step i in Extended Value Iteration
U	Uniformization constant
V	Number of visit during episode
δ_r	Maximal variation of the mean rewards
Δ	Bound on the variation of the bias
ρ	Mixing rate
Π	Set of stationary deterministic policies
π	Stationary deterministic policy in Π
π^*	Optimal policy
π^0	Constant policy equal to 0 for each state
λ	Arrival rate in a birth-and-death process
μ	Leaving rate in a birth-and-death process
ν^π	Stationary measure under policy π
τ_{mix}	Mixing time
$\mathbb{1}$	Indicator function
$\hat{}$	Denote an empirical quantity
$\tilde{}$	Denotes an optimistic quantity

1 Introduction

In reinforcement learning, an operator aims to maximize a reward signal received from a system with unknown characteristics. Using a trial-and-error method, unlike in supervised learning, the operator does not access any preliminary training set and must improve their decision over time, while acting on the system to optimize. As foreseen in the seminal paper [Robbins, 1952], decisions and observations thus depend on the history of the former ones and are interdependent.

1.1 Queueing and Learning

Reinforcement learning has become a mainstream tool optimization in various setups. It has been taking off in the field of Markov Decision Processes (MDPs), which are very generic tools to modelize many control problems. A natural follow-up after the general case has been the study of structured MDPs, such as parametric MDPs. A particularly important subclass of MDPs is the class of controlled queueing systems; see, e.g., [Martin L. Puterman, 1994, Chapters 1–3] and [Q.-L. Li et al., 2019], and [Walton and K. Xu, 2021, Section 5] for a review on learning problems in queues. Distinguishing themselves among the MDPs, typical control problems on queues have the following characteristics:

1. *No discount.* Discounting costs or rewards is common practice in the reinforcement learning literature, especially in Q-learning algorithms [Sutton and Barto, 1998]. However, in the context of queueing, optimizing with respect to the average reward is particularly interesting. Moreover, considering the average reward criterion tends to highlight the importance of the structure instead of the initial state.

2. *Large diameter.* Queueing systems are usually investigated under a drift condition that makes the system *stable*, i.e., positive recurrent. This condition implies that some states are hard to reach. In fact, for many queueing control problems, the diameter, which measures the time needed to cross the MDP, is exponential in the size of the state space. Even in the simple case of an M/M/1 queue with a finite buffer, or equivalently a birth-and-death process with a finite state space

and constant birth and death rates, the diameter is exponential in the size of the state space. With the state space possibly much larger in the queueing examples than in the generic MDP toy examples, it is mandatory to find more accurate ways to describe the difficulty of exploring an MDP.

3. *Structured transition matrices.* Queueing models describe how jobs join and leave queues, and this yields bounded state transitions. As a result, MDPs on queues have sparse and structured transition matrices. It allows more explicit computations of typical quantities that appear in the performance bounds, such as the bias.

To illustrate these specificities of the queueing systems, let us introduce a few examples we will study separately in the next chapters. In Chapters 3 and 4, we give the example of a Dynamic Voltage and Frequency Scaling (DVFS) processor dealing with impatient jobs with soft Markovian deadlines: the operator needs to choose the processor speed to find a balance between the energy consumption and the abandonment. In this case, in the Markovian setting, under any speed policy, the resulting Markov chain is a birth-and-death process, whose optimality properties can be theoretically studied. In Chapter 5, we consider the now classical admission control problem, where the operator chooses whether to admit a new job in the queueing network or reject it, inducing a cost in the latter case. In these examples, the number of states may be arbitrarily large, so our saving grace lies in exploiting the structure of the system to learn the best policies efficiently.

1.2 State of the Art

Regarding reinforcement learning in MDPs, as stated previously, the setting is very general. Using discounted rewards for the infinite horizon case enables better control of the rewards in the long term, and it gives access to additional technical tools to prove the convergence of algorithms, which ensures the convergence of Q-learning algorithms, like in [Jin, Allen-Zhu, et al., 2018] in the finite horizon case, or in [Dong et al., 2019] for the discounted case. Stronger assumptions are needed for Q-learning algorithms with average rewards, as in [Wei et al., 2020], where the knowledge of the mixing time of the ergodic MDP is available.

Besides model-free algorithms, in the average reward setting, a fundamental learning algorithm is UCRL2 in [Jaksch et al., 2010], which can be seen as the MDP adaptation of the Upper Confidence Bound (UCB) algorithm from the bandit problems. The performance of this algorithm is measured by the regret, which is upper bounded

by $\tilde{O}(DS\sqrt{AT})$, where S is the size of the state space, A is the size of the action space, T is the time horizon, and D the *diameter*, a measure of the difficulty of exploration in the MDP. Moreover, a lower bound for a class of MDPs with such parameters has been proposed in $\Omega(\sqrt{DSAT})$. This lower bound became the target for matching upper bounds: New algorithms inspired by UCRL2 have improved regret bounds, $\tilde{O}(D\sqrt{SAT})$ in [Azar et al., 2017] and even $\tilde{O}(\sqrt{DSAT})$ according to [Tossou et al., 2019; Zhang and Ji, 2019]. We can already remark that these bounds concern a broad class of MDPs. We can expect these algorithms to have a better performance than what their upper bounds suggest on a restricted class of MDPs, for example queueing systems.

Other papers have been proposing different upgrades of the UCRL2 algorithm. A possibility is to try to use a more accurate definition of the structure of the MDP, taking into account the support of the transitions as in [Fruit, Pirotta, and Lazaric, 2020] or also introducing the *local diameter* as in [Bourel et al., 2020]. Another and more straightforward way to involve the structure of the MDP in the learning problem is to consider parametric MDPs. In [Jin, Yang, et al., 2020], linear models with d parameters achieve a regret upper bound of $\tilde{O}(\sqrt{d^3H^3T})$ in finite horizon. An even broader extension of this concept is the class of linear mixture models: in the discounted case in [Zhou et al., 2021], a regret upper bound of $O(d\sqrt{T}/(1-\gamma)^2)$ is proved, and in the average reward case a bound in $O(d\sqrt{DT})$ is shown in [Wu et al., 2022].

A natural extension of the problem of reinforcement learning in MDPs is to consider the case of Partially Observable MDPs (POMDPs): at each time step, the state is not fully known, and the learner only gets this information partially. For example in a queueing network, they may only access the total number of jobs in the system rather than the precise number of jobs in each queue. Generally, reinforcement learning in POMDPs is intractable [Jin, Kakade, et al., 2020, Propositions 1 and 2], so it is necessary to restrict this learning task to subclasses of POMDPs. Again some of these works have been in the context of finite horizon or with discounted rewards [Even-Dar et al., 2005; Ross et al., 2007; Poupart and N. A. Vlassis, 2008]. In the case of the infinite horizon and undiscounted rewards, the dependence on the diameter still appears in [Azizzadenesheli et al., 2016], so that the same difficulties we noticed for queueing systems as MDPs emerge in the case of POMDPs.

1.3 Contributions

In this thesis, we show that leveraging the knowledge of the structure of the queueing systems, despite their large number of states, yields interesting results to overcome some of the limitations of reinforcement learning in generic MDPs. This will highlight the importance of considering the structure of these MDPs when dealing specifically with queueing systems.

We use the example of the DVFS processor as a basis to study the analysis of UCRL2 in a restricted class of stable birth-and-death processes. In this typical example, the number of states may be arbitrarily large, and we show that the diameter is at least exponential in the number of states. Therefore, the classical upper and lower bounds in $\Omega(\sqrt{DSAT})$ for the general case are unsatisfactory. Since the analysis in [Jaksch et al., 2010] uses a minimax approach, it considers the worst case for MDPs in their class. We present an adapted analysis of a slightly tweaked version of UCRL2 that uses the knowledge of the birth-and-death structure in order to get an upper bound in $\tilde{O}(\sqrt{E_2AT})$, where E_2 depends on the stationary measure of a reference policy. Notably, the bound does not depend on the diameter D nor on S, unlike in the parametric MDPs we presented earlier. The main observation is that to learn the MDP, while the algorithm needs to make sure every state is visited linearly often, the tail end of the state space is barely involved in the average reward, and in consequence in the regret bound.

Another typical example we use to showcase the importance of the structure of queueing systems is the admission control problem. Consider a queueing network where the operator admits or rejects jobs entering the network depending only on the total number of jobs: this process can be modelized as a POMDP. Rather than attempting to directly learn on this POMDP, we present an algorithm that compares it to an asymptotically equivalent MDP we can actually build and learn. This algorithm relies on an equivalence theorem for product-form networks and on the ergodicity of the POMDP to ensure the observations on the POMDP are valuable information to build the equivalent MDP, a birth-and-death process. Using our previous result on such a process, we eventually obtain an upper bound of the regret in $\tilde{O}(ST)$, where the dependency on S comes from the ergodicity requirement. Once again, exploiting the structure of the queueing network is an important step to show this new bound.

1.4 Organization of the Thesis

We describe how the thesis is organized and briefly present the content of each chapter. In Chapter 2, we remind the main MDP definitions and lemmas that we will use in this thesis. We also describe the classical reinforcement learning algorithm UCRL2 and what the main ideas surrounding it are, as it will be the foundation for the algorithms in the following chapters. This chapter barely contains any new result, as its goal is to clarify and introduce the main objects we deal with.

Then in Chapter 3, without using reinforcement learning yet, we will consider the example of a DVFS processor to study the structural properties of the optimal speed policies.

We come back to the properties above in Chapter 4, where they are used to adapt the analysis of the regret of the UCRL2 algorithm to the example of a birth-and-death process MDP. In that case, we show that using this extra knowledge on the MDP lets us derive bounds that showcase the importance of the structure of the queue and its stationary measure under any policy, with a lower dependence on the size of the state space itself.

Finally, in Chapter 5, we study an admission control problem on a more complex queueing network. We show we can use ergodicity properties to fall back to learning a simple birth-and-death process, as in Chapter 4. We present an algorithm inspired by UCRL2 and compute a regret bound where the main dependence on the state space comes from the ergodicity property.

In this thesis, for clarity, results that are not original work will be followed by their reference to mark a difference between original results and what is already known.

2 Markov Decision Processes and Model-Based Reinforcement Learning

We introduce in this chapter the main notations and definitions we will use in the next chapters. More precisely, we present first some MDP definitions before highlighting a few structural properties of birth and death processes. Then we introduce the fundamentals of reinforcement learning on MDPs and expose the classic algorithm UCRL2 with the ideas surrounding it. Indeed the same ideas will be used in Chapters 4 and 5. Note that this chapter is not intended to be an exhaustive presentation on MDPs or on model-based reinforcement learning.

2.1 Markov Decision Processes

In this section, the definitions are mainly based on those from [Martin L. Puterman, 1994, Chapter 8].

2.1.1 Definitions and Notations

Definition 2.1
A discrete-time MDP M is the tuple:

$$M = (\mathcal{S}, \mathcal{A}, P(s'|s,a), \mathcal{R}(s,a)) \,,$$

where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ the finite action space, $P(s'|s,a)$ the transition probability under action a to go from s to s' and $\mathcal{R}(s,a)$ the reward distribution under action a in state s.

We let $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$ where $|\cdot|$ is the set cardinality operator.

The dynamic behaviour of the MDP M is as follows. At each time step, if the MDP is in state s and action a is chosen, an immediate reward $R(s,a)$ with distribution $\mathcal{R}(s,a)$ is obtained independently of all else. Subsequently, the state changes to s' with probability $P(s'|s,a)$ in a Markovian manner.

In the remainder of this thesis, we assume that the reward distributions have bounded support. We will also be interested in measuring performance in expectation. We therefore call $r(s,a) = \mathbb{E}[R(s,a)]$ the expected reward in state s under action a.

We will mainly consider stationary deterministic policies, which are functions $\pi : s \in S \mapsto a \in \mathcal{A}$. Under such a policy, the MDP follows the dynamics of a Markov chain with transition matrix $(P(s'|s,\pi(s)))_{s,s'}$.

Gain

We will consider a unichain MDP M, that is, such that for every policy in $\Pi := \{\pi : \mathcal{S} \to \mathcal{A}\}$ the set of stationary and deterministic policies, there is a single recurrent class for the Markov chains, and possibly some transient states. Call s_t the random variable of the state of the system at time t, with $s_1 = s$ a fixed initial state. In this case, independently of this initial state, the gain or average reward induced by a policy $\pi \in \Pi$ is:

$$g^\pi(M) = g(M,\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_\pi[r(s_t,\pi(s_t))|s_1 = s]. \tag{2.1}$$

Since M has finite state and action spaces, the limit in (2.1) always exists. Notice also that the restriction to stationary and deterministic policies is not a loss of optimality [Martin L. Puterman, 1994, Theorem 8.4.5].

We will denote by $g^* := g^*(M) := \max_{\pi\in\Pi} g(M,\pi)$ the optimal average reward and we denote by π^* an optimal policy.

We recall that for queueing systems, we are specifically interested in the optimization of the average reward rather than the finite horizon reward $\sum_{t=1}^{H} \mathbb{E}_\pi[r(s_t,\pi(s_t))]$ with horizon H, or the discounted reward $\lim_{T\to\infty} \sum_{t=1}^{T} \gamma^{t-1}\mathbb{E}_\pi[r(s_t,\pi(s_t))]$ with discount factor $\gamma \in (0,1)$.

Since the MDP M is unichain, for every policy we can define the stationary measure ν^π. We rewrite the gain:

$$g^\pi = \sum_{s\in\mathcal{S}} r(s,\pi(s))\nu_s^\pi, \tag{2.2}$$

where ν^π is the stationary measure under policy π.

Bias

For a policy π, we define the bias at state s by the following Cesarò sum:

$$h^\pi(s) := \lim_{N\to\infty} \frac{1}{N} \sum_{T=1}^{N} \sum_{t=1}^{T} \mathbb{E}_\pi \left[\left(r\left(s_t, \pi(s_t)\right) - g^\pi \right) \mid s_1 = s \right], \quad \forall\, 0 \le s \le S-1. \tag{2.3}$$

The bias at s can be seen as the initial offset before scoring the gain at every step in the long run. In other words, asymptotically, the cumulative reward after T steps starting from s is roughly equal to $h^\pi(s) + T \cdot g^\pi$.

The bias is also related to the gain and the rewards through the Bellman equation:

$$g^\pi + h^\pi(s) = r(s, \pi(s)) + \sum_{s' \in \mathcal{S}} h^\pi(s') P(s'|s, \pi(s)) \quad \text{for } s \in \mathcal{S},$$

furthermore, in the unichain case, the Bellman equation characterizes the gain g^π and the bias h^π up to an additive constant.

Diameter

We define the diameter of a MDP, as given in [Jaksch et al., 2010], intuitively as the average time needed to reach the furthest state from any other state with a well-chosen policy.

Definition 2.2 (Diameter of a MDP [Jaksch et al., 2010])
Let $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary policy of M with initial state s. Let $T\left(s'|M, \pi, s\right) := \min\{t \ge 0 : s_t = s' | s_1 = s\}$ 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*

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

It should be clear that the diameter of an MDP can be large if there exist states that are hard to reach: we will see that for a stable basic birth and death process, we expect the diameter to scale exponentially in S. The diameter is a commonly used parameter to characterize the difficulty to learn a specific MDP.

Ergodicity Rates

Consider an ergodic MDP M under any policy $\pi \in \Pi$, that is, the Markov chain defined by every policy $\pi \in \Pi$ is ergodic over the state space $\mathcal{S}$. Denote, again by ν^π its stationary measure under policy π. There exists $C > 0$, $\rho(\pi) \in (0, 1)$ such that:

$$\sup_{s_1 \in \mathcal{S}} \|\mathbb{P}^\pi{}_{s_1}(s_t = \cdot) - \nu^\pi\|_{TV} \leq C\rho(\pi)^t \quad \forall t > 0, \tag{2.4}$$

where $\mathbb{P}^\pi{}_{s_1}(s_t = \cdot)$ is the probability distribution of s_t if the MDP starts at s_1 and applies the policy π. While the convergence rate ρ depends on the choice of π *a priori*, we will show in Chapter 5 that we can bound it by a quantity independent of π.

Continuous-time MDP and Uniformization

Common applications are usually given in continuous-time models. We therefore give a quick overview of the continuous-time MDP and the unfirmization method. A continuous-time MDP M' is defined by:

$$M' = (\mathcal{S}, \mathcal{A}, q(s'|s, a), \mathcal{R}'(s, a)) \,,$$

where $\mathcal{S}$ and $\mathcal{A}$ are respectively the state space and the action space, for $s' \neq s$, $q(s'|s, a) \geq 0$ is the transition rate under action a to go to the state s' from s and $\mathcal{R}'(s, a)$ is the reward distribution under action a at s, with reward $R'(s, a)$ and mean reward $r'(s, a)$. The reward is sampled when the action a is chosen. By definition of the rates, we also require that $q(s|s, a) = -\sum_{s' \neq s} q(s'|s, a)$.

We briefly explain the dynamics in M' in the context of stationary policies: for a chosen fixed action a at state s, the MDP transitions to the next state $s' \neq s$ after an exponentially distributed random time $\tau_{s'}$ with parameter $q(s'|s, a)$, whichever time comes first. For such s', the reward scored is $\tau_{s'} R'(s, a)$, and only then the controller chooses the next action to be used in s'. In this setup, $-q(s|s, a)^{-1}$ is the mean time spent at state s before the transition.

Rather than studying these continuous-time models, it is more practical and convenient to focus on their discrete-time equivalent, by uniformizing the MDP, as follows: let $U \geq (-\max_{s \in S, a \in \mathcal{A}} q(s|s, a))$ be the chosen uniformization constant, which is

well defined as S is finite. To build the equivalent discrete-time MDP M_U, we define the transition matrix P such that:

$$\begin{cases} P(s'|s,a) = q(s'|s,a)/U & \text{for } s' \neq s \\ P(s|s,a) = 1 - \sum_{s' \neq s} P(s'|s,a) & \text{otherwise} \end{cases}$$

We also redefine the rewards $R(s,a) = R'(s,a)/U$. In this newly defined process, a time step corresponds to an observation of the original MDP M' at a random time with an exponential distribution of parameter U, while in the original process, we had different exponential parameters for each transition.

This uniformization method preserves the relative performance of the policies so that we can directly study the discrete-time MDPs and use algorithms designed for these MDPs. In the remainder of the thesis, we will mainly use discrete-time MDPs and note them M instead of M_U when no confusion is possible.

2.1.2 The Special Case of Birth and Death Processes: a Comparison of the Bias and the Diameter

In algorithms based on episodes, the algorithm executes a policy π_k for each episode k, so that the cumulative reward for a given episode is roughly, as stated in the definition of the bias: $h^{\pi_k} + T_k \cdot g^{\pi_k}$, with T_k the length of the episode. To control the reward scored by an algorithm, we therefore need to be able to control the bias, possibly for any policy.

With little knowledge of the bias, it has been common to use the diameter D in the learning bounds. However, as we work within classes of MDPs with a strong structure, we can instead expect to control the bias of any policy and use this knowledge to improve the bounds. We first study the scaling of the diameter in a stable birth and death process as given in the following basic example.

We consider a birth and death process with Poisson arrivals with rate λ_s and leaving exponential rate μ_s over the S states $\{0, \dots, S-1\}$. We can represent the dynamics on the following continuous-time Markov chain:

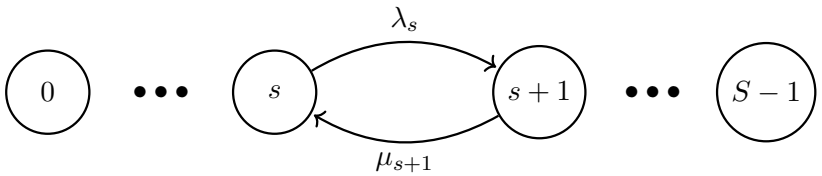


Figure 2.1: Markov chain diagram of the example queue.

Diameter Scaling in Birth and Death Processes

In this example, we assume that Figure 2.1 is a representation of the Markov chain under the best policy to consider for the diameter, and we will choose the rates to be $\lambda_s = \lambda$ and $\mu_s = \mu$ for some constants $\lambda < \mu$.

We can expect here the diameter to be closely related to the worst expected hitting times. These hitting times are intuitively close to the return times to the last states of the queue, which scale exponentially in the size of the state space S. The diameter should therefore also have this exponential scaling.

This discussion is formalized in the following result.

> **Lemma 2.3**
> *For any MDP M and any policy π that induce a birth-and-death process, the diameter D^π under policy π as well as the local diameter defined as $D^\pi(s-1, s) := \mathbb{E}^\pi\left[T\left(s|\pi, s-1\right)\right]$ grow exponentially in S.*

Proof. Under policy π:

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

where $\tau^\pi(s, s') = \mathbb{E}\left[T\left(s'|\pi, s\right)\right]$ is the expected time to go from s to s' under policy π.

Denote by U the choice for the uniformization constant. Starting from $S-1$, we can write the hitting time equations:

$$\tau^\pi(S-1, S-1) = 1 + P^\pi_{S-1,S-2}\tau^\pi(S-2, S-1),$$

and we notice that the left-hand side term actually is the inverse of the stationary measure at $S-1$, so that $\nu^\pi(S-1)^{-1} = \tau^\pi(S-1, S-1)$. We therefore obtain:

$$D^\pi \geq U\frac{\nu^\pi(S-1)^{-1} - 1}{\mu_{S-1}} \geq \nu^\pi(S-1)^{-1} - 1.$$

Now by reversibility, in our specific case, we have that: $\nu^\pi(s) = \nu^\pi(0)\frac{\lambda^s}{\mu^s}$, with $\nu^\pi(0)^{-1} = \frac{1-\lambda^S/\mu^S}{1-\lambda/\mu}$, so that

$$D^\pi \geq \frac{1-\frac{\lambda^S}{\mu^S}}{1-\frac{\lambda}{\mu}}\left(\frac{\mu}{\lambda}\right)^{S-1} - 1.$$

As for the maximal local diameter, $\max_s D^\pi(s-1,s) \geq \max_s \tau^\pi(s-1,s) \geq \tau^\pi(S-2,S-1)$ and the same argument as before applies, so that both the local diameter and the diameter scale exponentially in S, as $\lambda < \mu$.

□

Variation of the Bias over the State Space

We show here how we control the bias in a birth and death process on the state space $\mathcal{S} := \{0,\dots,S-1\}$ with arrival rates λ_s and leaving rates μ_s for a fixed policy π, as in figure 2.1. We will denote by $P := P^\pi$ the transition matrix for this fixed policy.

We first show a bound on hitting times that will be used to control the variations of the bias:

Lemma 2.4

Let π be any policy. Consider the Markov chain with policy π and transitions P starting from any state s. Denote by τ_s the random time to hit 0 from state s. Then:

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

where U is the chosen uniformization constant

Proof. We write the expected hitting time equations. Let $\mathbf{e}$ be the unit vector. We have the system:

$$\mathbb{E}^\pi[\boldsymbol{\tau}] = \mathbf{e} + P\,\mathbb{E}^\pi[\boldsymbol{\tau}], \tag{2.5}$$

with $\boldsymbol{\tau}$ the vector of hitting times, and adding to the system of equation $\tau_0 = 0$.

We will show the result by induction. The system gives for $s = S-1$:

$$\mathbb{E}^\pi[\tau_s] = 1 + \mathbb{E}^\pi[\tau_s]\frac{1-\mu_s}{U} + \mathbb{E}^\pi[\tau_{s-1}]\frac{\mu_s}{U},$$

so that:

$$\mathbb{E}\tau_s = \frac{U}{\mu_s} + \mathbb{E}^\pi\left[\tau_{s-1}\right].$$

Then with an induction, we want to prove the equation for $s < S-1$:

$$\mathbb{E}^\pi\left[\tau_s\right] = \mathbb{E}^\pi\left[\tau_{s-1}\right] + \frac{U}{\mu_s}\sum_{s'=s}^{S-1}\prod_{i=s+1}^{s'}\frac{\lambda_{i-1}}{\mu_i}. \tag{2.6}$$

For $s < S-1$, assume (2.6) is true for $\mathbb{E}^\pi\left[\tau_{s+1}\right]$:

$$\begin{aligned}
\mathbb{E}^\pi\left[\tau_s\right] &= 1 + \mathbb{E}^\pi\left[\tau_{s+1}\right]\frac{\lambda_s}{U} + \mathbb{E}^\pi\left[\tau_s\right]\frac{1-\mu_s-\lambda_s}{U} + \mathbb{E}^\pi\left[\tau_{s-1}\right]\frac{\mu_s}{U} \\
&= 1 + \mathbb{E}^\pi\left[\tau_s\right]\frac{1-\mu_s-\lambda_s}{U} + \mathbb{E}^\pi\left[\tau_{s-1}\right]\frac{\mu_s}{U} + \mathbb{E}\tau_s\frac{\lambda_s}{U} + \sum_{s'=s+1}^{S-1}\prod_{i=s+1}^{s'}\frac{\lambda_{i-1}}{\mu_i} \\
&= \frac{U}{\mu_s} + \mathbb{E}^\pi\left[\tau_{s-1}\right] + \frac{U}{\mu_s}\sum_{s'=s+1}^{S-1}\prod_{i=s+1}^{s'}\frac{\lambda_{i-1}}{\mu_i} \text{ by gathering the } \tau_s \text{ terms} \\
&= \mathbb{E}^\pi\left[\tau_{s-1}\right] + \frac{U}{\mu_s}\sum_{s'=s}^{S-1}\prod_{i=s+1}^{s'}\frac{\lambda_{i-1}}{\mu_i},
\end{aligned}$$

the induction is therefore true, and by definition of $\nu^\pi(0)$ we have:

$$\mathbb{E}^\pi\left[\tau_s\right] \leq \mathbb{E}^\pi\left[\tau_{s-1}\right] + \frac{U}{\mu_s}\nu^\pi(0)^{-1}.$$

□

Proposition 2.5

For any policy π, define for $s \in \{1, \dots, S-1\}$ the variation of the bias

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

Assume that the sequences (λ_s) and (μ_s) are respectively non-increasing and non-decreasing. Then:

$$\partial h^\pi(s) \leq 2\delta_r\nu^\pi(0)^{-1}\sum_{i=1}^{s}\frac{U}{\mu_i},$$

where $\delta_r := \max_{s,a,a'}|r(s,a) - r(s-1,a')|$ is the largest reward variation between neighbouring states.

Proof. To control the difference of probabilities in the bias, we will define a coupling, which will follow the inequality given in [Levin et al., 2008, Proposition 4.7]

$$\left\|P^t(s,\cdot) - P^t(s-1,\cdot)\right\|_{\text{TV}} \leq \inf_{X,Y}\{\mathbb{P}(X \neq Y) : (X_t, Y_t) \text{ is a coupling of } P^t(s,\cdot) \text{ and } P^t(s-1,\cdot)\}. \quad (2.7)$$

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:

$$X_{t+1} = \begin{cases} X_t - 1 & \text{if } 0 \leq U_t \leq \mu_{X_t} \\ X_t & \text{if } \mu_{X_t} \leq U_t \leq 1 - \lambda_{X_t} \\ X_t + 1 & \text{if } 1 - \lambda_{X_t} \leq U_t \leq 1, \end{cases} \quad (2.8)$$

and define Y_{t+1} the same way from Y_t. As (λ_s) and (μ_s) are respectively non-increasing and non-decreasing, $0 \leq X_t - Y_t \leq 1$ for this coupling. Moreover, we have from (2.7), reminding that $\delta_r := \max_{s,a,a'} |r(s,a) - r(s-1,a')|$:

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

As τ_s is the time needed for X_t to hit 0 starting from s, 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:

$$\partial h^\pi(s) \leq 2\delta_r \mathbb{E}^\pi\left[\tau_s\right],$$

and now using Lemma 2.4:

$$\partial h^\pi(s) \leq 2\delta_r \nu^\pi(0)^{-1} \sum_{i=1}^{s} \frac{U}{\mu_i}.$$

□

Remark 2.6

Without the assumption that (λ_s) and (μ_s) are respectively non-increasing and non-decreasing, Proposition 2.5 still holds with $r_{\max}$ rather than δ_r in the inequality, with $r_{\max}$ being the span of the reward in the MDP.

In view of the previous lemma, with the example of figure 2.1 with rates $\lambda_s = \lambda$ and $\mu_s = \mu$ with $\lambda < \mu$ constants, and we get that the variation of the bias is bounded in the following way:

$$\partial h^\pi(s) \leq 2\delta_r s \frac{1 - \frac{\lambda^S}{\mu^S}}{1 - \frac{\lambda}{\mu}} \frac{U}{\mu}.$$

We notice here that the span of the bias, being $\max_s h(s) - \min_s h(s)$, is bounded by a quadratic function, which is much better than the exponential lower bound we had for the diameter. For this reason, in the algorithms, we will try as much as possible to use our knowledge of the span of the bias rather than resorting to the diameter.

2.2 Model-Based Reinforcement Learning in a MDP

2.2.1 Reinforcement Learning Framework

We will consider a unichain discrete-time MDP $M = (\mathcal{S}, \mathcal{A}, p, \mathcal{R})$ in discrete time.

The model-based reinforcement learning problem consists in finding a *learning algorithm* $\mathbb{L}$, or learner, that chooses actions to maximize a cumulative reward over a finite time horizon T. At each time step $t \in \mathbb{N}$, the system is in state $s_t \in \mathcal{S}$ and the learner chooses an action $a_t \in \mathcal{A}$. When executing a_t, the learner receives a random reward $r_t(s_t, a_t)$ with mean $r(s_t, a_t)$ and the system moves, at time step $t + 1$, to the next state s_{t+1} with probability $p(s_{t+1}|s_t, a_t)$. The learner also gets information on the reached state s'. The learning algorithm does not know the MDP M except for the sets $\mathcal{S}$ and $\mathcal{A}$.

In different terms, an agent uses the algorithm $\mathbb{L}$ to act on the system (or environment). This is synthesized in Figure 2.2.

Note already that in Chapter 5, we will face the case of POMDPs, where the state s_{t+1} is not fully observed, and only partial information is retrieved from this next state. We will continue this discussion in the concerned chapter.

In the model-based setup that we will consider, by opposition to the model-free one, the learner relies on a model and its estimates to choose their next action. Model-based algorithms need a large memory space to store the estimated transition probabilities (S^2A). Nevertheless, we will mainly focus on this kind of algorithm as they are efficient for regret minimization.

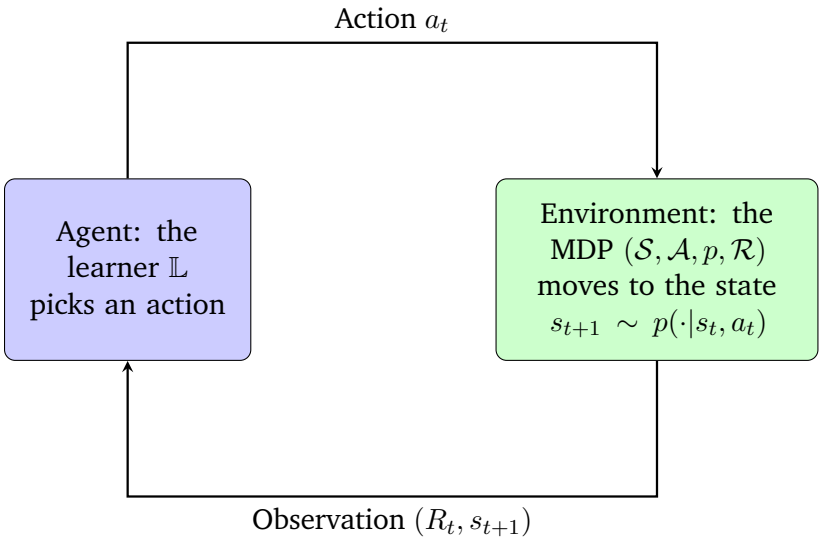


Figure 2.2: Agent and environment interaction

In order to maximize the cumulative reward, the learner cannot afford to dedicate all their time to learning the optimal policy, as many of the observed rewards would be much worse than the best rewards. They also cannot commit to the empirically best policy without regularly getting new information from the environment. Instead, they must find a balance between *exploration* and *exploitation* to ensure optimal performance in the long run. The performance metric we are interested in to match this goal is called the *regret*, which we define in the next subsection.

2.2.2 Definition of the Regret

We remind that $g^* := g^*(M) := \max_{\pi \in \Pi} g(M, \pi)$ is the optimal average reward in the MDP M.

Definition 2.7 (Regret)
The regret *at time T of the learning algorithm $\mathbb{L}$ is*

$$Reg(M, \mathbb{L}, T) := Tg^*(M) - \sum_{t=1}^{T} r_t^{\mathbb{L}}, \tag{2.9}$$

where $r_t^{\mathbb{L}}$ is the mean reward at time t under the algorithm $\mathbb{L}$.

The regret given in 2.9 is sometimes called the *pseudo-regret* as it involves mean rewards rather than random rewards. For bounded rewards R, this difference

is negligible when T is large. The regret is a natural benchmark for evaluating the performance of a learning algorithm. As stated in the previous subsection, in order to minimize the regret, algorithms need to find a balance over time between exploration and exploitation, as opposed to identification algorithms that need to find an optimal policy as fast as possible, disregarding the possible disappointing performance of some policies.

Remark 2.8
The definition of the regret depends a priori on the initial state of the algorithm. In the following, we will consider the case of unichain MDPs, so that we can always work with the regret with a reference initial state, which would only add a constant term to the regret bound with respect to any initial state. For queueing systems, we will usually choose the empty queue as the reference initial state.

In [Jaksch et al., 2010], a *universal* lower bound on $\text{Reg}(M, \mathbb{L}, T)$ has been developed in terms of the *diameter* of the underlying MDP.

Theorem 2.9 (Universal lower bound [Jaksch et al., 2010])
For any learning algorithm $\mathbb{L}$, any natural numbers $S, A \geq 10$, $D \geq 20 \log_A S$, and $T \geq DSA$, there is an MDP M with S states, A actions, and diameter D such that for any initial state $s \in \mathcal{S}$,

$$\mathbb{E}[Reg(M, \mathbb{L}, T)] \geq 0.015\sqrt{DSAT}. \tag{2.10}$$

On this large class of MDP, the regret lower bound depends on the diameter. As seen in the previous section, we expect the regret to be much lower when restricting the class of MDP to birth and death processes for example, when we get to control and know the variation of the bias for any policy. We discuss this point in Chapter 4.

2.2.3 Presentation of UCRL2

We now present UCRL2, a classic reinforcement learning algorithm introduced in [Jaksch et al., 2010] that is a variant of UCRL [Auer and Ortner, 2006]. Let us present an intuition of this algorithm. UCRL2 aims to learn the transition probabilities and mean rewards for each state-action pair (s, a). It is based on *episodes* so that during an episode k, a policy π_k is chosen and used until a stopping criterion is met. To choose its next policy, the algorithm relies on the *Optimism in Face of Uncertainty* (OFU) principle: from the empirical transition probabilities and rewards observed, a

confidence set is built of the optimistic MDPs, so that with high probability, the true MDP M belongs to this set. After each episode, this confidence set is updated, and then the algorithm chooses a policy within this set with the highest gain, where the *optimism* comes from.

We can already compare UCRL2 to the UCB algorithm, which also uses the OFU principle for regret minimization in a bandit setting. In the former case however, a new policy is chosen only after each episode as opposed to after each time step, the idea being that we not only need to learn the reward, but the eventual target is the gain, so that the episode length needs to grow over time to ensure exploration.

Let us now give a more technical presentation of the algorithm. For each episode k, let t_k denote its start time and π_k the policy used during this episode. For each state s and action a, let $V_k(s,a)$ denote the number of visits of (s,a) during episode k and let $N_t(s,a) := \#\{\tau < t : s_\tau = s, a_\tau = a\}$ denote the number of visits of (s,a) until timestep t. Let $\mathcal{M}_k$ be the confidence set of MDPs with transition probabilities $\tilde{p}$ and rewards $\tilde{r}$ that are "close" to the empirical MDP at episode k, $\hat{p}_k$ and $\hat{r}_k$, i.e., $\tilde{p}$ and $\tilde{r}$ satisfy

$$\forall (s,a), \quad |\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq r_{\max} \sqrt{\frac{7 \log(2SAt_k/\delta)}{2 \max\{1, N_{t_k}(s,a)\}}} \tag{2.11}$$

$$\forall (s,a), \quad \|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S \log(2At_k/\delta)}{\max\{1, N_{t_k}(s,a)\}}}, \tag{2.12}$$

where δ is a confidence parameter, technically necessary to compute upper bound on the regret with high probability. It is usually of order $1/T$.

With these quantities, a pseudocode for UCRL2 is given in Algorithm 1. We notice that UCRL2 relies on Extended Value Iteration (EVI), that is a variant of the celebrated Value Iteration (VI) algorithm [Martin L. Puterman, 1994]; for further details about EVI, we point the reader to [Jaksch et al., 2010, Section 3.1]. Let us comment on how UCRL2 works.

Algorithm 1: The UCRL2 algorithm.

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

Set $t := 1$ and observe s_1

for *episodes* $k = 1, 2, \ldots$ **do**

  Set $t_k := t$ and compute the estimates $\hat{r}(s,a)$ and $\hat{p}_k(s'|s,a)$ as in (2.14).

  Use "Extended Value Iteration" to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that

$$g(\tilde{M}_k, \tilde{\pi}_k) \geq \max_{M' \in \mathcal{M}_k, \pi} g(M', \pi) - \frac{1}{\sqrt{t_k}} \tag{2.13}$$

  while $V_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_{t_k}(s_t, \tilde{\pi}_k(s_t))\}$ **do**

    Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t and observe s_{t+1};

    $V_k(s_t, a_t) := V_k(s_t, a_t) + 1$;

    $t := t + 1$;

There are three main steps. First, at the start of each episode, UCRL2 computes the empirical estimates

$$\hat{r}_k(s,a) := \frac{\sum_{t=1}^{t_k-1} r_t \mathbb{1}_{\{s_t=s, a_t=a\}}}{\max\{1, N_{t_k}(s,a)\}}, \qquad \hat{p}_k(s'|s,a) := \frac{\sum_{t=1}^{t_k-1} \mathbb{1}_{\{s_t=s, a_t=a, s_{t+1}=s'\}}}{\max\{1, N_{t_k}(s,a)\}} \tag{2.14}$$

of the reward and probability transitions, respectively,

where $\mathbb{1}_E$ is the indicator function of E. Then, it applies Extended Value Iteration (EVI) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that (2.13) holds true. Finally, it executes policy $\tilde{\pi}_k$ until it finds a state-action pair (s,a) whose count within episode k is greater than the corresponding state-action count before episode k. This method is sometimes referred to as the *doubling trick*, as the number of visits for a state-action pair is at most doubled during an episode. Through this criterion, we get an improved control over the number of episodes, and over the regret induced by each state-action pair.

2.2.4 Classical Lemmas for UCRL2

The two following lemmas are proved in [Jaksch et al., 2010, Appendix C.2 and Appendix C.3] respectively. The first one bounds the number of episodes, using the doubling trick.

Lemma 2.10 ([Jaksch et al., 2010])
Denote by K_t the number of episodes up to time t, and let $t > SA$. It is bounded by:

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

The logarithmic scaling on the number of episodes implies that the $\sqrt{T}$ bound stems from the confidence bounds rather than from the episode changes themselves.

The following lemma is used to simplify regret terms when summing over the episodes for a single state-action pair.

Lemma 2.11 ([Jaksch et al., 2010])
For any fixed state action pair (s, a) and time T, we have:

$$\sum_{k=1} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \leq 3\sqrt{N_{T+1}(s,a)},$$

This lemma is the main ingredient to translate the confidence bounds into a quantity of the same order of $\sqrt{T}$.

We now present a version of the Azuma-Hoeffding inequality that is mainly used to control the regret coming from the episode changes.

Lemma 2.12 (Azuma-Hoeffding inequality [Williams, 1991])
Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leq C$ for all i and some constant $C > 0$. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$:

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

These lemmas are now known tools to prove the regret bound for UCRL2 given in Theorem 2.13. We will also rely on them to prove regret bounds in Chapters 4 and 5.

Theorem 2.13 (Consequence of Theorem 4 in [Jaksch et al., 2010])
Consider a MDP M with diameter D, S states and A actions. For any initial state $s \in \mathcal{S}$ and $T > 1$, the expected regret of UCRL2 *is bounded as follows:*

$$\mathbb{E}\left[Reg(M, \text{UCRL2}, T)\right] \leq 35DS\sqrt{AT \log T}.$$

In comparison with Theorem 2.9, the upper bound of the regret has the same dependence in the time horizon T, up to a logarithmic factor. There is however an extra $\sqrt{DS}$ factor, that has been the main focus for improvement in [Fruit, Pirotta, and Lazaric, 2020; Tossou et al., 2019]. In Chapter 4, we will also improve the upper bound on the specific case of a birth-and-death process.

3 Optimal Control of a Large Queue: the Case of a DVFS Processor

In the previous chapter, we saw how we would uniformize the state space to define a discrete-time MDP from a continuous-time model. In this chapter, we study the specific example of a DVFS processor: an operator changes the speed of the processor, using real-time information on the system, in order to optimize the overall performance of the system for lower energy consumption. In this model, the state space is infinite and the uniformization method is impossible *a priori*. We show how to fall back to this uniformization method on a cunningly truncated state space, and we prove structural properties of the optimal policy that remain valid on the infinite state space.

This chapter is based on the published work [Anselmi, Gaujal, and Rebuffi, 2021].

3.1 Introduction

3.1.1 Related Work

In the deterministic case where job sizes and arrival times are known, a vast literature addressed the problem of designing both off-line and on-line algorithms to compute speed profiles that minimize the energy consumption subject to hard real-time constraints (deadlines) on job execution times; see, e.g., [Yao et al., 1995; Bansal et al., 2007; M. Li et al., 2017] and the references therein. In a *stochastic* environment where only statistical information is available about job sizes and arrival times, it turns out that combining hard deadlines and energy minimization via DVFS-based techniques is much more difficult. In fact, forcing hard deadlines requires to be very conservative, i.e., to consider the worst cases. In spite of these difficulties, this problem has been investigated in [Lorch and Smith, 2001] for a single job and in [Gaujal et al., 2020] for multiple jobs. The former approach constructs the optimal speed profile explicitly in “closed form” while the latter relies on the numerical

solution of a discrete time Markov Decision Process (MDP) [Martin L Puterman, 2014]. The latter approach has several drawbacks: i) it requires a discretization of both time and space, which introduces by itself an approximation on the optimal solution, ii) deadlines and job sizes need to be bounded, and iii) the size of the state space of the underlying MDP is exponential in the size of the maximal deadline. These issues makes this approach unusable in practice.

The approach followed in this chapter circumvents the difficulties described above by replacing the hard real-time constraints, i.e., jobs have hard deadlines that must be satisfied, by *soft* real-time constraints, i.e., jobs may miss their deadlines, at some cost. While the hard deadline of a job must be known at the job arrival, soft deadlines allow for a different information structure: here, only the deadline distribution is known at the job arrival. In this chapter, we further assume that jobs missing their deadlines become obsolete and are dropped. Obsolescence is often found in real-time systems where the information carried by jobs may not be valid any longer after their deadline as it will be replaced by fresher input coming from other jobs. Therefore, obsolete jobs become useless and can get discarded from the queue. Dropping obsolete jobs can also model impatient customers: customers wait for service for some time (deadline) and quit (are dropped) if not served before that time.

3.1.2 Contribution

We investigate the problem above in a Markovian setting where jobs join the system following a Poisson process and both the deadlines and sizes of jobs are exponentially distributed. Under these assumptions, our goal is to minimize the *average cost*, i.e., the average energy spent by the processor per second plus the penalty due to jobs missing their deadlines. We formulate this problem as an MDP in continuous time where the state is the number of jobs in the system and the action is the processor speed.

Our main result, Theorem 3.2, shows the existence of an optimal speed profile that is increasing in the number of jobs in the system and upper bounded by some constant. This constant is defined in (3.4) as the minimizer of a function that comes out from our analysis. Surprisingly, our bound does not depend on the deadlines and arrival rates. In other words, our bound on the optimal speed does not change upon variations of these job characteristics. In addition, it yields a simple approximation for the optimal policy and several numerical tests show that such approximation is accurate in heavy-traffic conditions. Finally, the proposed approximation is used

to control the proportion of jobs that leave the system because they missed their deadline in a simple manner.

Underlying the proof of our main result, there are some technical challenges that we now discuss. The proposed MDP satisfies the regularity assumptions (stability, unichain) needed to establish an optimality equation as described in [X. Guo and Hernandez-Lerma, 2009]. However, this is not enough to show structural properties of the optimal policy. In fact, the common approach to do this is to uniformize the MDP and to investigate the properties of the corresponding discrete time value iteration operator. Unfortunately, this is not possible in our case because the transition rates are unbounded. To uniformize the MDP, a typical approach consists of truncating the state space. Indeed, this is the approach we follow. However, we notice that a naive truncation will not help because the truncation barrier has a strong impact on the structure of the optimal policy in the sense that it would not preserve any monotonicity property that it may have without truncation. This is shown in Figure 3.2. Instead, we use the technique proposed by Blok and Spieksma in [Blok and Spieksma, 2015], which smoothly scales down the upward rates of the truncated system as a function of the size of its state space. This technique has been successfully used in [Hyon and Jean-Marie, 2020; Bhulai et al., 2014] to show structural properties of controlled queueing systems. However, these works focus on discounted costs. Here, we use the same truncation technique but we apply it to the average cost. To the best of our knowledge, this has never been done before. In our specific case, the convergence to the infinite system will be guaranteed by the monotone convergence theorem.

This chapter is organized as follows. In Section 3.2, our model and the corresponding MDP are described in detail. We also present our main result (Theorem 3.2) as well as some hindsight on the construction of the proof. Section 3.3 shows how the MDP is truncated and scaled and shows the proof of the monotonicity as well as the construction of the upper bound on the optimal speed. Section 3.4 focuses on the convergence when the truncation point goes to infinity. Section 3.5 uses Theorem 3.2 to provide an approximation on the optimal policy and estimate the deadline-miss probability. Finally, Section 3.6 draws the conclusions of our work and addresses further research.

3.2 System Description, Problem Statement and Main Result

The system described here is a model for the dynamics of a real-time device composed of a single computing resource (a processor) where incoming jobs need to be executed under a constraint on the amount of time that they spend in the system.

Processor This is a DVFS processor whose speed can continuously vary in the interval $[0, a_{\max}]$. We consider that speed changes are immediate and induce no energy cost. When the processor works at speed a, it processes a units of work per second while its power dissipation is $w(a)$ watts. The classic simple model for the dynamic power dissipation of any CMOS circuit is $w = K\alpha V^2 f$, see, e.g., [Snowdon et al., 2005], where K is a constant, α measures the activity of the logical gates, V is the supply voltage and f is the clock frequency. The clock frequency of the gates is often linearly related to the voltage and therefore DVFS processors adjust both variables together. Within the model above, this means that $w(a)$ is cubic in the speed a. In this chapter, we just require that $w(a)$ is continuous, increasing and strictly convex in the speed a.

Jobs They form a stochastic point process, with Poisson arrivals with rate λ, i.i.d. deadlines exponentially distributed with rate μ and i.i.d. sizes exponentially distributed with rate σ. Without loss of generality, we assume that $\sigma = 1$.

Dynamics At any point in time t, the processor chooses its speed $a(t)$ and executes one of the jobs in its backlog queue. We notice that the choice of the job in execution, named *active* in the following description, is irrelevant here because of the memoryless property of the deadlines and of the sizes. Thus, at any point in time, at most one job can be active. As mentioned above, this induces an instantaneous energy cost of $w(a(t))$. Now, three events can happen in continuous time:

1. A new job may join the queue.
2. The active job is completed before its deadline. In this case, the job leaves the system.
3. One job (active or inactive) reaches its deadline. In this case, this job becomes obsolete, it is removed from the queue and an immediate cost equal to C is paid.

Cost Function If we denote by M_T the number of missed deadlines in the time interval $[1, T]$, the objective of this chapter is to study the speed profile $a(t)$ of the processor that minimizes the long-run average cost given by the missed deadlines plus power consumption. Specifically, this is given by

$$\mathbb{E}\left[\limsup_{T\to\infty}\frac{1}{T}\left(CM_T + \int_1^T w(a(t))\,\mathrm{d}t\right)\right]. \tag{3.1}$$

At this point, we claim that this problem can be modeled by a continuous time Markov decision process with a discrete state space. To see this, let us consider the system at time t under the speed profile $a(\cdot)$ and let $s(t)$ denote the number of jobs present in the system at time t. Each job is characterized by two values: (u_k, v_k), respectively, the sojourn time of job k (how long the job has been present) and its service quantity (how many elementary operations have already been executed on this job by the processor). Also let e_t be the time elapsed since the last arrival. Then the state of the system at time t is the tuple $(e_t, (u_1, v_1), \ldots, (u_{s(t)}, v_{s(t)}))$. From time t, the cost function only depends on the expected future behavior of the system, namely, the future arrivals, future services and future drops of packets due to deadline misses. The memoryless property of the exponential distribution implies that this expected future behavior does not really depend on $(e_t, (u_1, v_1), \ldots, (u_{s(t)}, v_{s(t)}))$, but only on $s(t)$, the current number of jobs. As a by-product, this implies that the choice of the optimal speed at time t only depends on $s(t)$, so that the optimal speed profile can only change when $s(t)$ changes, i.e., when an arrival, a service or a drop occur.

Markov Decision Process We now formulate the problem of interest as an MDP. As mentioned before, the state space is $\mathbb{N}$ and a state represents the number of jobs in the system. The action space is $[0, a_{\max}]$, i.e., the set of available speeds for the DVFS processor. Let $\pi = (\pi_s)_{s\in\mathbb{N}}$ denote a stationary and deterministic speed policy adopted by the processor, i.e., $\pi_s \in [0, a_{\max}]$ is the speed used in state s. It is well known that focusing on stationary and deterministic policies can be done with no loss of optimality in our case [X. Guo and Hernandez-Lerma, 2009, Theorem 5.9]; in other words, we may ignore the broader set of *history dependent randomized* policies and rather focus on the stationary and deterministic policies. For $s, s' \in \{0, \ldots, S'\}$,

we will denote the transition rates $q_{s,s'}(\pi_s) := q(s'|s, \pi_s)$ for the action π_s. These rates are given by:

$$q_{s,s'}(\pi_s) = \begin{cases} \lambda & \text{if } s' = s+1 \text{ (arrival of a new job)} \\ \pi_s + s\mu & \text{if } s \geq 1 \text{ and } s' = s-1 \text{ (completion of the active job} \\ & \qquad \text{or obsolescence of one job)} \\ -\lambda - s\mu - \pi_s & \text{if } s' = s \\ 0 & \text{otherwise.} \end{cases}$$

Under the speed policy $\pi = (\pi_s)_{s\in\mathbb{N}}$, the induced Markov chain, denoted by X^π, is a birth-and-death process that resembles an $M/M/\infty$ queue but with an additional decreasing rate, which comes from the processing speed policy; see Figure 3.1.

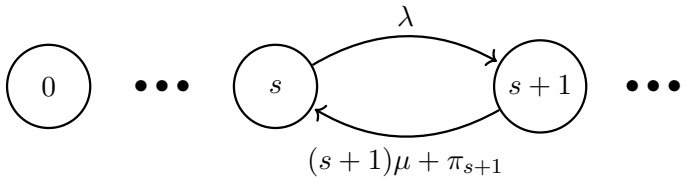


Figure 3.1: Markov chain diagram under policy π.

By ergodicity of the Markov chain X^π under all policies π, the *a priori* random long-run cost in the brackets in equation (3.1) is almost surely equal to the long-run expected cost. Letting $\mathbb{E}^\pi_{s_1}$ denote the expectation given a speed policy π and starting state s_1, the long-run cost becomes:

$$G(s_1, \pi) := \limsup_{T\to\infty} \frac{1}{T} \int_1^T \mathbb{E}^\pi_{s_1} c(X^\pi(t), \pi)\, \mathrm{d}t.$$

In this equation, the immediate cost function $c(\cdot, \cdot)$ is the expected cost incurred by the system at time t. It only depends on the current state and the current speed. Conditionally on the state ($X^\pi(t) = s$), the obsolescence rate is $s\mu$. Thus, the expected cost is:

$$c(s, \pi) := Cs\mu + w(\pi_s).$$

With a slight abuse of notation, we will use both notations $c(s, \pi)$ or $c(s, \pi_s)$ since c only depends on the speed used in state s and not on the whole policy.

As mentioned before, for given speed policy π, X^π is ergodic so that the MDP is unichain (all states are positive recurrent under all policies). This implies that

the cost can be defined independently of the starting state. Moreover, for a given π, there exists a unique stationary measure ν^π for X^π so that we can define the cost independently of the initial state and express it as a function of the stationary measure:

$$\forall s_1, \quad G^\pi := G(s_1, \pi) = \mathbb{E}^{\nu^\pi} c(X^\pi, \pi) = \sum_s \nu_s^\pi c(s, \pi_s). \tag{3.2}$$

Here, $\mathbb{E}^{\nu^\pi}$ is the expectation with respect to the invariant measure of X^π. Stationary policies that minimize (3.2) are optimal speed policies for the model. In particular, they are also optimal over all policies (history dependent and randomized) [X. Guo and Hernandez-Lerma, 2009, Theorem 5.9] . Also, our MDP satisfies all the conditions given in [X. Guo and Hernandez-Lerma, 2009, Theorem 5.9] to assert the existence of an optimal stationary deterministic policy π^* and an optimality equation of the form

$$\begin{aligned} G^* = G^{\pi^*} &= c(s, \pi_s^*) + \sum_{s'} h^*(s') q_{s,s'}(\pi_s^*) \\ &= \min_{a \in [0, a_{\max}]} c(s, a) + \sum_j h^*(s') q_{s,s'}(a), \quad \forall s \in \mathbb{N}, \end{aligned} \tag{3.3}$$

where h^* is a real-valued function on $\mathbb{N}$, referred to as *bias* of the optimal policy.

Main result The goal of this chapter is to investigate structural properties on π^* and G^*. First, let us define B as

$$B := \arg\min_{a \in \mathbb{R}^+} (w(a) + C(\lambda - a)). \tag{3.4}$$

This constant is well defined in $\mathbb{R}^+ \cup \{+\infty\}$ because w is strictly convex. In addition, we have the following remark.

Remark 3.1

If w is super-linear, i.e., $\lim_{a \to \infty} \frac{w(a)}{a} = \infty$, then $w(a) + C(\lambda - a)$ is also super-linear and B is finite. In practice, all models of power dissipation are super-linear in a, e.g., [Chandrakasan et al., 1992]. In the simple case where $w(a) = Ka^3$, we first notice that the constant K can be set to 1 without loss of generality because to compensate its effect one can adjust the missed deadline cost C accordingly. Then, in this case, we obtain $B = \sqrt{\frac{C}{3}}$.

Our main result is the following.

Theorem 3.2
There exists a deterministic optimal policy $\pi^* = (\pi^*_s)_{s\in\mathbb{N}}$ that is increasing in s and upper bounded by B.

Remark 3.3
The optimal speed policy of the processor is always bounded by a finite constant, namely $\min(B, a_{\max})$. By definition, B only depends on w (the power dissipation of the processor) and C the cost of each missed deadline. Thus, we remark that B is independent of the job characteristics *(arrival rate, deadline and size distributions). This is both surprising and helpful in practice. Indeed, if B is finite, one can set a priori the maximal speed of the processor to $a_{\max} := B$. This guarantees that in all cases, no cost reduction would be possible by using a more powerful processor. Further discussion about parameter settings, in particular the link between C and the probability that jobs miss their deadline under the optimal policy, will be discussed in Section 3.5.*

A proof of Theorem 3.2 is developed in Sections 3.3 and 3.4. Before delving into the proof, we devote the remainder of this section to explain the technical difficulties underlying our problem and the general approach that we follow.

The optimality equation (3.3) cannot be uniformized because the rates $q_{s,s}(\pi)$ are unbounded in s. Therefore, the study of structural properties of the optimal policy must be done by constructing a sequence of truncated MDPs whose optimal policies converge to π^* and for which we can prove monotonicity and boundedness. This approach has been proposed for the first time in [Blok and Spieksma, 2015] for MDPs with discounted cost by truncating the state space and scaling the rates of all the events that take the system out of the truncated space. This has been successfully applied in, e.g., [Bhulai et al., 2014; Hyon and Jean-Marie, 2020], to show that threshold type policies yield optimal admission control in one queue. However, all these applications consider discounted costs. To the best of our knowledge, no work has been done for the average cost. In the following, we will show that in our case, the scaling technique of [Blok and Spieksma, 2015] also works for the average cost and is the key ingredient to show Theorem 3.2, which gives new hindsights on the optimal policy. Thus, our result is another evidence of the power of this scaling approach, though our proof is quite different from the approach used in the discounted case. In fact, the common approach is to show that the value iteration operator preserves structural properties of the cost and of the policy (typically convexity properties of the cost and level sets of the policy), so that successive iterations of the operator will also preserve the properties and converge

to the optimal cost/policy. Here, we will directly consider the fixed point optimality equation and prove monotonicity of the policy by induction on the state (see Sections 3.3.1, 3.3.2, 3.3.3).

3.3 Truncated Model

As mentioned before, the original MDP cannot be uniformized because the transition rates $q_{s,s}(\pi)$ grow to infinity when s goes to infinity. To construct a discrete time model for a fixed maximal number of jobs $S' := S - 1$, we truncate the state space following the guidelines from [Blok and Spieksma, 2015] to construct a finite state MDP $\mathcal{M}_{S'}$ with linearly decaying arrival rates. The new state space has S states $\{0, \cdots, S'\}$, with the same continuous action space $[0, a_{\max}]$ and an average cost per second $G^{\pi}_{S'}$ defined as

$$J^{\pi}_{S'} := \limsup_{T\to\infty} \frac{1}{T-1} \int_1^T \mathbb{E}^{\pi}_{s_0} c(X^{\pi}(t), \pi)\, \mathrm{d}t \tag{3.5}$$

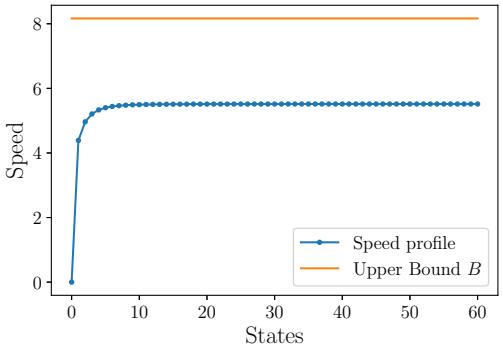
and with transition rates for all $s, s' \leq S'$ given by

$$q^{S'}_{s,s'}(\pi) := \begin{cases} \lambda^{S'}_s & \text{if } s < S' \text{ and } s' = s+1 \\ \pi_s + s\mu & \text{if } s > 0 \text{ and } s' = s-1 \\ -\lambda_s - s\mu - \pi_s & \text{if } s' = s \\ 0 & \text{otherwise,} \end{cases}$$

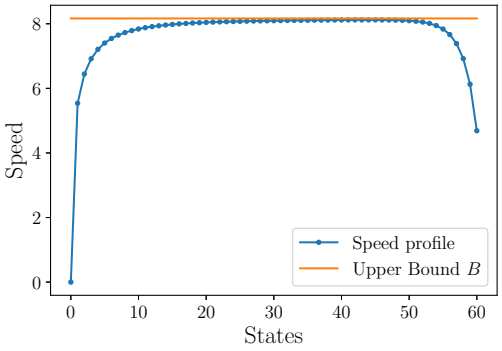
where the decaying arrival rate is $\lambda^{S'}_s := \lambda\left(1 - \frac{s}{S'}\right)$. Using decaying arrival rates will be a key ingredient in this chapter. To illustrate this, let us also consider a naive truncated MDP, $\mathcal{M}'_{S'}$ with fixed arrival rates in each state given by λ. The state space is $\{0, \cdots, S'\}$, with the same continuous action space $[0, a_{\max}]$ and the same cost function but with modified transition rates given by

$$q'^{S'}_{s,s'}(\pi) := \begin{cases} \lambda & \text{if } s < S' \text{ and } s' = s+1 \\ \pi_s + s\mu & \text{if } s > 0 \text{ and } s' = s-1 \\ -\lambda - s\mu - \pi_s & \text{if } s' = s \\ 0 & \text{otherwise,} \end{cases}$$

The two finite MDPs $\mathcal{M}_{S'}$ and $\mathcal{M}'_{S'}$ have been solved numerically using the following parameters: $\lambda = 10, \mu = 0.14, C = 200, w(a) = a^3, S' = 60$.



(a) $\mathcal{M}_{S'}$ with decaying arrival rates



(b) $\mathcal{M}'_{S'}$ with no decay of the arrival rates

Figure 3.2: Optimal policies π^* and $\pi^{*'}$ for the two truncated MDPs, respectively, $\mathcal{M}_{S'}$ (a) and $\mathcal{M}'_{S'}$ (b).

The respective optimal policies π^* and $\pi^{*'}$ are displayed in Figures 3.2(a)-(b). As one can see, the two models behave very differently. At the last state of $\mathcal{M}'_{S'}$, the speed does not need to be as high as the maximal speed, as the arrival rate drops from λ to 0, so that the speed plummets near the last state. In contrast, the optimal policy for $\mathcal{M}_{S'}$ is increasing from 0 to some bound slightly below $\sqrt{\frac{C}{3}}$. Additional numerical tests where we let S' grow (reported in Figure 3.3) further suggest that π^* is increasing in the state s as well as in the level of truncation S'. The bound B also appears to be rather tight when S' is large in the example reported in Figure 3.3.

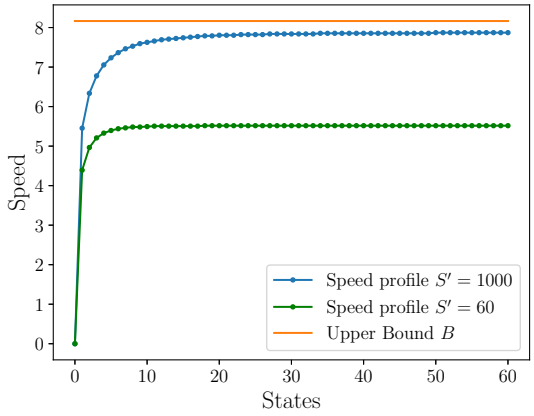


Figure 3.3: Two optimal speed policies in $\mathcal{M}_{S'}$, for $S' = 60$ and $S' = 1000$ and the bound $B = \sqrt{\frac{C}{3}}$.

This makes the study of $\mathcal{M}_{S'}$ promising and, in the remainder, we focus on this MDP with decaying arrival rates.

Since S' is fixed here, we may remove it in the notation for simplicity. As the state space is finite, we can uniformize this MDP to get a discrete time MDP. Choosing

$$U := \lambda + S'\mu + a_{\max} \tag{3.6}$$

as uniformization constant, we get a discrete time MDP $\mathcal{D}^{S'}$ with transition probabilities given by

$$p^{S'}_{s,s'}(\pi) = \begin{cases} \frac{1}{U}\lambda_s & \text{if } s < S' \text{ and } s' = s+1 \\ \frac{1}{U}(\pi_s + s\mu) & \text{if } s > 0 \text{ and } s' = s-1 \\ \frac{1}{U}\bar{U}_{s,\pi_s} & \text{if } s' = s \\ 0 & \text{otherwise,} \end{cases}$$

where the complementary probability to stay in state s is $\bar{U}_{s,\pi_s} := U - \lambda_s - \mu s - \pi_s$. Again, we focus on stationary policies π that minimize the cost $G^{\pi}_{S'}$. The long-run average cost per step for the discrete time MDP is

$$g^{\pi} := \limsup_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}^{\pi}_{s_1} c(X^{\pi}(k), \pi).$$

A classic result for MDPs (see for example [Martin L Puterman, 2014, Section 11.5.3]) says that the discrete and continuous time MDPs are related by the following relations: for any policy π, $Ug^{\pi} = G^{\pi}_{S'}$. This has the following consequences:

- Both MDPs have the same optimal policies.
- Optimal long-run average costs coincide up to the multiplicative uniformization constant: $Ug^* = Ug^{\pi^*} = G^{\pi^*}_{S'} = G^*_{S'}$.

We will first show the counterpart of Theorem 3.2 in the finite case. Let us define

$$B^{S'} := \arg\min_{a\in\mathbb{R}^+} \left(w(a) + \frac{C(\lambda - a)}{1 + \frac{\lambda}{\mu S'}} \right), \tag{3.7}$$

which is well defined in $\mathbb{R}^+ \cup \{\infty\}$ and unique because w is strictly convex.

We have the following properties for the optimal speed policy.

Theorem 3.4 *(i) The optimal speed policy π^* minimizing* (3.5) *is unique.*

(ii) The optimal speed policy is increasing in s: $\forall s \leq S', \pi_s^* < \pi_{s+1}^*$.

(iii) The optimal speed policy is upper-bounded: $\forall s \leq S', \pi_s^* \leq B^{S'}$.

The proof of this result will be the object of Section 3.3.1 (monotonicity, item (ii)), 3.3.2 (upper-bound, item (iii)), 3.3.3 (uniqueness, item (i)). Then, the relation with the original infinite MDP will be shown in Section 3.4.

3.3.1 Proof of Theorem 3.4(ii): Monotonicity of the Optimal Speed

In this subsection, we denote by π^* any optimal speed policy, X^* the associated Markov chain and $G_{S'}^*$ the optimal cost induced by π^*. Thus, the optimal long-run average gain per step is

$$g^* := \lim_{K\to\infty} \frac{1}{K} \mathbb{E}^{\pi^*} \sum_{k=1}^{K} \frac{1}{U} c\left(X^{\pi^*}(k), \pi^*\right). \tag{3.8}$$

When the state space is finite, the *bias* $h^* \in \mathbb{R}^S$ for the optimal speed π^* is defined up to an additive constant by

$$h^*(s) := \mathbb{E}_s^{\pi^*} \sum_{k=1}^{\infty} \left(c\left(X^{\pi^*}(k), \pi^*\right) - Ug^*\right), \quad \forall s \leq S'. \tag{3.9}$$

To fix the value of the bias vector, we set $h^*(0) := 0$.

Since the MDP is finite, unichain, the action space is compact and the costs and transition probabilities are continuous and bounded in the actions, [Martin L Puterman, 2014, Theorem 8.4.7] guarantees the existence of the optimality equations for the optimal cost and for the bias. Specifically, for any state $s \in \{0, \ldots, S'\}$,

$$g^* + h^*(s) = \frac{1}{U} \min_{a\in[0,a_{\max}]} \left\{w(a) + Cs\mu + (\mu s + a)h^*(s-1) + \bar{U}_{s,a}h^*(s) + \lambda_s h^*(s+1)\right\} \tag{3.10}$$

with $h^*(-1) = h^*(S'+1) = 0$ by convention.

For each state s, an optimal action π_s^* is the choice of a speed minimizing the right hand side term. Notice that necessarily, $\pi_0^* = 0$ (the speed of the processor must be 0 when there is no work to do).

Using (3.10), for $s \geq 1$, we can subtract $h^*(s-1)$ from $h^*(s)$ and choose $a = \pi_{s-1}^*$ in (3.10) to get

$$U(h^*(s) - h^*(s-1)) \leq \mu C + \lambda \left(1 - \frac{s}{S'}\right)(h^*(s+1) - h^*(s)) \\ + \left(\bar{U}_{s,\pi_s^*} - \frac{\lambda}{S'}\right)(h^*(s) - h^*(s-1)) + (\mu(s-1) + \pi_{s-1}^*)(h^*(s-1) - h^*(s-2)), \tag{3.11}$$

Doing the same subtraction for $a = \pi_s^*$, we obtain

$$U(h^*(s) - h^*(s-1)) \geq \mu C + \lambda \left(1 - \frac{s}{S'}\right)(h^*(s+1) - h^*(s)) \\ + \left(\bar{U}_{s,\pi_s^*} - \frac{\lambda}{S'}\right)(h^*(s) - h^*(s-1)) + (\mu(s-1) + \pi_s^*)(h^*(s-1) - h^*(s-2)). \tag{3.12}$$

Combining both inequalities together we get the inequality

$$(\pi_s^* - \pi_{s-1}^*)\,(h^*(s) - 2h^*(s-1) + h^*(s-2)) \geq 0.$$

From this, we can deduce the following property of the model.

Proposition 3.5
If h^* satisfies the following notion of discrete convexity:

$$\forall s \geq 2, (h^*(s) - 2h^*(s-1) + h^*(s-2)) > 0, \tag{3.13}$$

then the optimal speed policy π^* is increasing.

We give additional properties of the finite MDP needed to prove the main result.

Lemma 3.6
The asymptotic cost per second is upper-bounded: $Ug^* \leq \frac{C\lambda}{1+\frac{\lambda}{S'\mu}}$.

Proof. The cost Ug^* is the optimal asymptotic cost per unit of time. Therefore, we have that $Ug^* \leq Ug^{\pi^0}$, with Ug^{π^0} being the asymptotic cost per unit of time for policy π^0, when the speed is 0 for each state. Let ν^{π^0} be the asymptotic distribution in that

case. Computations of the stationary measure give $\nu_s^{\pi^0} = \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \binom{S'}{s} \left(\frac{\lambda}{S'\mu}\right)^s$ (Lemma 4.8) and thus we can then compute the associated asymptotic cost as follows:

$$\begin{aligned} g^{\pi^0} &= \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \sum_{s=0}^{S'} \nu_s^{\pi^0} \mu s C = \mu C \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \sum_{s=0}^{S'} s \binom{S'}{s} \left(\frac{\lambda}{S'\mu}\right)^s \\ &= \mu C \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \sum_{s=1}^{S'} S' \binom{S'-1}{s-1} \left(\frac{\lambda}{S'\mu}\right)^s \\ &= \mu C \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \frac{\lambda}{\mu} \sum_{s=0}^{S'-1} \binom{S'-1}{s} \left(\frac{\lambda}{S'\mu}\right)^s \\ &= C\lambda \left(1+\frac{\lambda}{S'\mu}\right)^{-S'} \left(1+\frac{\lambda}{S'\mu}\right)^{S'-1} = \frac{C\lambda}{1+\frac{\lambda}{S'\mu}}, \end{aligned}$$

which concludes the proof. □

We now want to show by backward induction on s that h^* satisfies (3.13). For $0 \le s \le S'-1$, the exact property $\mathcal{P}(s)$ that we will show is

$$h^*(s) - h^*(s-1) < h^*(s+1) - h^*(s) < \frac{C}{1+\frac{\lambda}{S'\mu}}. \tag{3.14}$$

Before investigating the initialization step, we show the following preliminary inequality on the optimal cost.

Lemma 3.7
Let $\overline{\pi} := \mathbb{E}^{\pi^*}[\pi^*_{X^*}]$ be the average speed under the stationary measure ν^{π^*}. Then,

$$Ug^* > w(\overline{\pi}) + C\frac{\lambda - \overline{\pi}}{1+\frac{\lambda}{S'\mu}}.$$

Proof. Consider the optimal Markov chain X^* displayed in Figure 3.4, where for clarity we represented the continuous Markov chain, as its behavior is similar as the uniformized discrete time Markov chain.

Using Jensen's inequality, since w is strictly convex and the stationary measure ν^{π^*} is non trivial, we get the following strict inequality:

$$\begin{aligned} Ug^* = G^*_{S'} = \mathbb{E}^{\pi^*}[w(\pi^*_{X^*})] + C\mu\mathbb{E}^{\pi^*}[X^*] &> w(\mathbb{E}^{\pi^*}[\pi^*_{X^*}]) + C\mu\mathbb{E}^{\pi^*}[X^*] \\ &= w(\overline{\pi}) + C\mu\mathbb{E}^{\pi^*}[X^*]. \end{aligned}$$

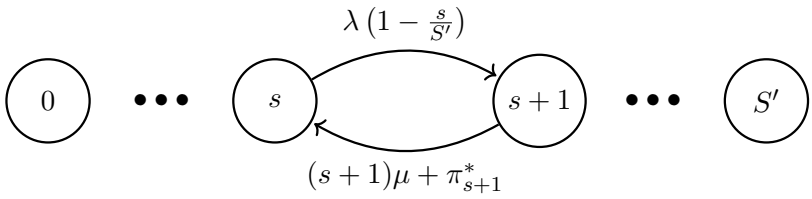


Figure 3.4: The transition rates of the Markov chain X^*.

Now, we calculate $\mathbb{E}^{\pi^*}[X^*]$. The underlying detailed balance equations are

$$\nu^{\pi^*}_{s+1}[\mu(s+1)+\pi^*_{s+1}] = \nu^{\pi^*}_s\left(\lambda - s\frac{\lambda}{S'}\right).$$

The LHS is 0 for $s=-1$, and the RHS is also 0 for $s=S'$. When we sum these equations from 0 to $S'-1$, we can therefore write

$$\sum_{s=0}^{S'} \nu^{\pi^*}_s(\mu s+\pi^*_s) = \sum_{s=0}^{S'} \nu^{\pi^*}_s\left(\lambda - s\frac{\lambda}{S'}\right),$$

which gives $\mu\mathbb{E}^{\pi^*}[X^*]+\overline{\pi} = \lambda - \mathbb{E}^{\pi^*}[X^*]\frac{\lambda}{S'}$, i.e.,

$$\mathbb{E}^{\pi^*}[X^*] = \frac{\lambda-\overline{\pi}}{\mu+\frac{\lambda}{S'}}. \tag{3.15}$$

Thus, we finally have

$$Ug^* > w(\overline{\pi}) + C\frac{\lambda-\overline{\pi}}{1+\frac{\lambda}{S'\mu}}$$

as desired. □

Remark 3.8

The computation of the expectation that gave (3.15) *remains true for any speed policy, and in that case* $\overline{\pi}$ *becomes the average speed of that policy. This formula of the expectation will be needed later to rewrite the cost* $G^*_{S'}$ *in a different way.*

For the initialization of the induction, we need the following lemma.

Lemma 3.9

$h^*(S') - h^*(S'-1) < \frac{C}{1+\frac{\lambda}{S'\mu}}$.

Proof. For any π, the optimality equation (3.10) at $s = S'$ gives

$$U(g^* + h^*(S')) \leq w(a) + CS'\mu + (\mu S' + a)h^*(S' - 1) + (U - \mu S' - a)h^*(S'),$$

which holds if and only if

$$Ug^* + (\mu S' + a)h^*(S') \leq w(a) + CS'\mu + (\mu S' + a)h^*(S' - 1),$$

which holds if and only if

$$h^*(S') - h^*(S' - 1) \leq \frac{w(a) + CS'\mu - Ug}{\mu S' + a}, \tag{3.16}$$

with equality if the chosen π is optimal. We therefore want to show the following inequality for some a:

$$h(a) := \frac{w(a) + CS'\mu - Ug^*}{\mu S' + a} < \frac{C}{1 + \frac{\lambda}{S'\mu}}.$$

To show this, we use the inequality (3.7) on Ug^* and choose $s = \overline{\pi}$. This gives

$$h(\overline{\pi}) < C\frac{S'\mu - \frac{\lambda - \overline{\pi}}{1 + \frac{\lambda}{S'\mu}}}{\mu S' + \overline{\pi}} < C\frac{S'\mu + \overline{\pi}}{(\mu S' + \overline{\pi})(1 + \frac{\lambda}{S'\mu})} < \frac{C}{1 + \frac{\lambda}{S'\mu}}.$$

This concludes the proof by (3.16). □

Therefore, for the initialization step we write the inequality (3.12) for $s = S'$ to obtain

$$\begin{aligned} U\left(h^*(S') - h^*(S' - 1)\right) \geq \mu C + \left(U - \mu S' - a - \frac{\lambda}{S'}\right)\left(h^*(S') - h^*(S' - 1)\right) \\ + (\mu(S' - 1) + a)(h^*(S' - 1) - h^*(S' - 2)), \end{aligned}$$

which implies

$$\left(\mu S' + a + \frac{\lambda}{S'}\right)\left(h^*(S') - h^*(S' - 1)\right) \geq \mu C + (\mu(S' - 1) + a)(h^*(S' - 1) - h^*(S' - 2)),$$

which implies (using Lemma 3.9)

$$(\mu(S' - 1) + a)\left(h^*(S') - h^*(S' - 1)\right) > (\mu(S' - 1) + a)(h^*(S' - 1) - h^*(S' - 2)),$$

which as desired gives $h^*(S') - h^*(S' - 1) > h^*(S' - 1) - h^*(S' - 2)$. This new inequality and Lemma 3.9 imply $\mathcal{P}(S' - 1)$.

Now, let us assume that the property $\mathcal{P}(s)$ is true for some $i \leq S' - 1$. We then have, writing (3.12) with $s = \pi_s^*$ and using the first inequality of $\mathcal{P}(s)$ (3.14),

$$U(h^*(s) - h^*(s-1)) \geq \mu C + \left(U - \mu s - a - \frac{\lambda}{S'}\right)((h^*(s) - h^*(s-1)) \\ + (\mu(s-1) + a)(h^*(s-1) - h^*(s-2))$$

which implies

$$\left(\mu s + a + \frac{\lambda}{S'}\right)(h^*(s) - h^*(s-1)) \geq \mu C + (\mu(s-1) + a)(h^*(s-1) - h^*(s-2))$$

which implies

$$(\mu(s-1) + a)\,(h^*(s) - h^*(s-1)) > (\mu(s-1) + a)(h^*(s-1) - h^*(s-2)) \quad (3.17)$$

The inequality (3.17) comes from the second inequality of $\mathcal{P}(s)$ (3.14), and from its first inequality we finally obtain both inequalities of $\mathcal{P}(s-1)$.

Thus, $h^*(s) - h^*(s-1) > h^*(s-1) - h^*(s-2)$ and the backward induction is complete. Since h^* is strictly convex, with Proposition 3.5 we deduce that π^* is increasing.

3.3.2 Proof of Theorem 3.4(iii): Upper Bound on the Optimal Speed

We call for all $a \geq 0$:

$$u(a) := w(a) + C\frac{\lambda - a}{1 + \frac{\lambda}{S'\mu}}.$$

By definition, we recall that $B^{S'}$ is the unique minimum of u, see (3.7). We also have

$$G^{\pi}_{S'} = \sum_{s \in S} \nu^{\pi}_s u(\pi_s).$$

Proposition 3.10
If π^* is an optimal increasing speed policy, then it is upper-bounded by $B^{S'}$, which means $\pi^*_s \leq B^{S'}$ for all $s \leq S'$.

Proof. We show this statement by contradiction. Let π^* be an optimal increasing speed policy. As π^* is increasing, assume that $\pi^*_{S'} > B^{S'}$. Let ν be the associated

stationary measure. As $\pi^*_{S'} > B^{S'} \geq 0 = \pi^*_0$, we can define $s_0 > 0$ as the smallest state such that $\pi^*_{s_0} > B^{S'}$. Now, we can define the following policy

$$\widetilde{\pi}_s = \begin{cases} \pi^*_s & \text{if } i < s_0 \\ B^{S'} & \text{if } s \geq s_0 \end{cases}.$$

and let $\widetilde{\nu}$ denote its associated stationary measure. By definition, we notice that the minimum of the function u is reached at $B^{S'}$. Moreover, for all $i < s_0$, we get from the local balance equations that $\nu_s > \widetilde{\nu}_s$. Indeed, when $s \geq s_0 - 1$:

$$\frac{\nu_{s+1}}{\nu_s} = \frac{\lambda\left(1 - \frac{s}{S'}\right)}{(s+1)\mu + \pi^*_{s+1}} < \frac{\widetilde{\nu}_{s+1}}{\widetilde{\nu}_s},$$

with equality only when $i < s_0 - 1$. It remains to show that the cost associated to this new speed is lower than the original one. We have

$$\begin{aligned} J^{\pi^*} - J^{\widetilde{\pi}} &= \sum_{s=0}^{S'} \nu_s u(\pi^*_s) - \sum_{s=0}^{S'} \widetilde{\nu}_s u(\widetilde{\pi}_s) \\ &= \sum_{s=0}^{s_0-1} (\nu_s - \widetilde{\nu}_s) u(\pi^*_s) + \sum_{s=s_0}^{S'} \nu_s u(\pi^*_s) - \sum_{s=s_0}^{S'} \widetilde{\nu}_s u(B^{S'}) \\ &> \sum_{s=0}^{s_0-1} (\nu_s - \widetilde{\nu}_s) u(B^{S'}) + \sum_{s=s_0}^{S'} \nu_s u(B^{S'}) - \sum_{s=s_0}^{S'} \widetilde{\nu}_s u(B^{S'}) > 0. \end{aligned}$$

This contradicts the optimality of π^* and concludes the proof. □

3.3.3 Proof of Theorem 3.4(i): Uniqueness of the Optimal Policy

The following lemma gives a different expression for π^*_s using the notion of generalized inverse of the derivative w' of w. First, since w is strictly convex, w' is continuous and well defined everywhere but on a countable subset $Q \subseteq [0, a_{\max}]$. Moreover, w' is strictly increasing and diverges, so that we can correctly define the inverse

$$w'^{-1} : y \mapsto \inf\{a \in [0, a_{\max}] \setminus Q, w'(a) \geq y\},$$

and w'^{-1} is increasing.

Lemma 3.11

If π^* is an optimal speed policy with bias function h^*, we can relate the variation of the bias to the speed at a given state:

$$\pi_s^* = w'^{-1}(\partial h_s^*), \qquad \forall s \geq 1$$

with $\partial h_s^* := h^*(s) - h^*(s-1)$ for $s \in \{1, \dots, S'\}$.

Proof. Let π^* be an optimal speed policy with bias function h^*. For $s \geq 1$, the optimality equation (3.10) can be written as

$$G_{S'}^* = \min_s \{w(a) + Cs\mu - (\mu s + a)\partial h_s^* + \lambda_s \partial h_{s+1}^*\}, \tag{3.18}$$

so that for any speed a:

$$w(\pi_s^*) - \pi_s^* \partial h_s^* \leq w(a) - a\partial h_s^*.$$

This yields

$$\begin{cases} \frac{w(\pi_s^*) - w(a)}{\pi_s^* - a} \leq \partial h_s^* & \text{for } s \leq \pi_s^* \\ \frac{w(\pi_s^*) - w(a)}{\pi_s^* - a} \geq \partial h_s^* & \text{for } s \geq \pi_s^*. \end{cases}$$

Let $\overline{Q} = [0, a_{\max}] \setminus Q$. These inequalities on ∂h_s^* give: $w'(\pi_s^{*-}) \leq \partial h_s^* \leq w'(\pi_s^{*+})$, so that there are two possibilities. Either i) $\partial h_s^* \in w'\left(\overline{Q}\right)$ and $w'(\pi_s^*) = \partial h_s^*$, or ii) $\partial h_s^* \notin w'\left(\overline{Q}\right)$, so that $\pi_s^* = w'^{-1}(\partial h_s^*) \in Q$. In both cases, we have $\pi_s^* = w'^{-1}(\partial h_s^*)$. □

Proposition 3.12

The optimal speed policy is unique, and therefore Blackwell optimal.

Proof. Let $\pi, \widetilde{\pi}$ be two optimal speed policies, $h^*, \widetilde{h}^*$ their respective biases and $\partial h, \widetilde{\partial h}^*$ the respective variations of the biases. We will show by induction that the variation of the speed and biases are equal.

We already have that $\pi(0) = \widetilde{\pi}(0) = 0$. The optimality equation (3.10) for $s = 0$ then gives

$$\partial h_1^* = \widetilde{\partial h}_1^* = \frac{Ug^*}{\lambda} = \frac{G_{S'}^*}{\lambda}.$$

Using Lemma 3.11, we then have $\pi_1 = w'^{-1}(\partial h_1^*) = w'^{-1}(\widetilde{\partial h}_1^*) = \widetilde{\pi}_1$.

Assume that for some $s \geq 1$, $\pi_s = \widetilde{\pi}_s$ and $\partial h_s^* = \widetilde{\partial h}_s^*$. Writing (3.18) for both optimal speeds, and using the assumption of the induction, we have $\partial h_{s+1}^* = \widetilde{\partial h}_{s+1}^*$. Then by using Lemma 3.11 again, $\pi_{s+1} = w'^{-1}(\partial h_{s+1}^*) = w'^{-1}(\widetilde{\partial h}_{s+1}^*) = \widetilde{\pi}_{s+1}$.

The induction is complete, so that the optimal speed policy is unique. □

This completes the last part of the proof of Theorem 3.4.

3.4 Convergence of the Truncated MDPs

To show the convergence of the sequence of the truncated MDPs to the infinite one as S' goes to infinity, we first show monotonicity properties in S'. These guarantee the existence of the limit and allow us to invoke the monotone convergence theorem to show that this limit satisfies the optimality equation of the infinite MDP.

3.4.1 The Optimal Speed is Increasing in the Size of the State Space

The next proposition states that the optimal cost is increasing in the number of states.

> **Proposition 3.13**
> *Let $S' \geq 1$. Then, $G_{S'}^* \leq G_{S'+1}^*$.*

Proof. Let $S' \geq 1$ and let π be the optimal speed policy when the state space is $\{0, \cdots, S'+1\}$. Let $\widetilde{\pi}$ be a speed policy for the MDP with state space $\{0, \cdots, S'\}$, defined as the truncation of π: for $s \leq S'$, $\widetilde{\pi}_s = \pi_s$. Recall that $X^{\widetilde{\pi}}$ is the continuous time Markov chain with speed policy $\widetilde{\pi}$ on the reduced state space $\{0, \cdots, S'\}$, and therefore cannot be compared directly with X^{π}, which is defined on $\{0, \cdots, S'+1\}$.

Thus, let $\widetilde{X}$ be the following discrete time Markov chain, with state space $\{0, \cdots, S'+1\}$ and transition probabilities given by, for $s \leq S'+1$,

$$\widetilde{p}_{s,s'} = \begin{cases} \frac{1}{U^{(S'+1)}} \lambda_s^{(S')} & \text{if } 1 \leq s' = s+1 \leq S' \\ \frac{1}{U^{(S'+1)}} (\pi_s + s\mu) & \text{if } 0 \leq s' = s-1 \leq S'-1 \\ \frac{1}{U^{(S'+1)}} \bar{U}_s^{(S'+1)} & \text{if } 0 \leq s' = s \leq S' \\ \frac{1}{U^{(S'+1)}} (U - (S'+1)\mu - \pi_{S'+1}) & \text{if } s' = s = S'+1 \\ 0 & \text{otherwise,} \end{cases}$$

with $U^{(S'+1)} := 2\lambda + (S'+1)\mu + \pi_{max}$, $\lambda_s^{(S')} := \lambda\left(1 - \frac{s}{S'}\right)$ and $\bar{U}_s^{(S'+1)} := U - \lambda_s^{(S')} - \mu s - \pi_s$.

With a slight abuse of notation, we denote by X^π and $X^{\widetilde{\pi}}$ the uniformized Markov chains with the same uniformization constant $U^{(S'+1)}$, so that we will be able to compare both Markov chains defined with the same time step. Moreover, notice that the Markov chain $\widetilde{X}$ is not irreducible: the use of the last state is only to extend the chain to a larger number of states while keeping the behavior of $\widetilde{X}$ similar to the one of $X^{\widetilde{\pi}}$.

In order to effectively compare $\widetilde{X}$ and X^π, we will now define a coupling $(\widetilde{Y}, Y)$ such that $\widetilde{Y}$ and Y have the same distributions as $\widetilde{X}$ and X^π respectively. For each time step $k \in \mathbb{N}$, let $A(k)$ be a uniformly distributed random variable on $[0, U^{(S'+1)}]$.

$$\widetilde{Y}(k+1) = \begin{cases} \widetilde{Y}(k)+1 & \text{if } A(k) \in [0, \lambda_{\widetilde{Y}(k)}^{(S')}] \text{ and } \widetilde{Y}(k) \leq S' \\ \widetilde{Y}(k) & \text{if } A(k) \in [\lambda_{\widetilde{Y}(k)}^{(S')}, U^{(S'+1)} - \widetilde{Y}(k)\mu - \pi_{\widetilde{Y}(k)}] \text{ and } \widetilde{Y}(k) \leq S' \\ \widetilde{Y}(k) & \text{if } A(k) \in [0, U^{(S'+1)} - \widetilde{Y}(k)\mu - \pi_{\widetilde{Y}(k)}] \text{ and } \widetilde{Y}(k) = S'+1 \\ \widetilde{Y}(k)-1 & \text{if } A(k) \in [U^{(S'+1)} - \widetilde{Y}(k)\mu - \pi_{\widetilde{Y}(k)}, U^{(S'+1)}] \text{ and any } \widetilde{Y}(k), \end{cases}$$

and similarly:

$$Y(k+1) = \begin{cases} Y(k)+1 & \text{if } A(k) \in [0, \lambda_{Y(k)}^{(S'+1)}] \\ Y(k) & \text{if } A(k) \in [\lambda_{Y(k)}^{(S'+1)}, U^{(S'+1)} - Y(k)\mu - \pi_{Y(k)}] \\ Y(k)-1 & \text{if } A(k) \in [U^{(S'+1)} - Y(k)\mu - \pi_{Y(k)}, U^{(S'+1)}]. \end{cases}$$

By construction, if $\widetilde{Y}(0) = Y(0)$, then we show by induction that for all n, $\widetilde{Y}(k) \leq Y(k)$.

We now check for all possible cases. For all $s \leq S'+1$, $\lambda_s^{(S')} < \lambda_s^{(S'+1)}$, therefore if $\widetilde{Y}(k) = Y(k)$, then $\widetilde{Y}(k+1) \leq Y(k+1)$. If $\widetilde{Y}(k) = Y(k)-1$ with $\widetilde{Y}(k+1) = \widetilde{Y}(k)+1$,

then $Y(k+1) \geq Y(k)$ by definition of U, as $U - Y(k)\mu - \pi_{Y(k)} \geq \lambda \geq \lambda^{(S')}_{\widetilde{Y}(k)}$, and therefore $Y(k+1) \geq \widetilde{Y}(k+1)$. In the remaining cases, $\widetilde{Y}(k+1) \leq Y(k+1)$.

Hence, for all k, $\widetilde{Y}(k) \leq Y(k)$, so that $\widetilde{X} \leq_{st} X^{\pi}$; here, $\leq_{st}$ denotes the stochastic order [Shaked and Shanthikumar, 1994a]. As $s \mapsto c(s, \pi_s)$ is increasing, this implies:

$$\mathbb{E}^{\pi} c(\widetilde{X}, \pi) \leq \mathbb{E}^{\pi} c(X^{\pi}, \pi) = G^*_{S'+1},$$

and we have

$$G^{\widetilde{\pi}}_{S'} = \mathbb{E}^{\widetilde{\pi}} c(X^{\widetilde{\pi}}, \widetilde{\pi}) = \mathbb{E}^{\pi} c(\widetilde{X}, \pi).$$

Therefore,

$$G^*_{S'} \leq G^{\widetilde{\pi}}_{S'} \leq G^{\pi}_{S'+1} = G^*_{S'+1}$$

as desired. □

The next proposition states that the optimal speed policy is increasing in the size of the state space.

Proposition 3.14
Let $S' \geq 1$ and let $\pi^{(S')}$ be the unique optimal speed policy for the S'-th MDP. Then, $\pi^{(S')}_s \leq \pi^{(S'+1)}_s$ and $\partial h^{(S')}_s \leq \partial h^{(S'+1)}_s$ for all $s \leq S'$.

Proof. We use the expression of the optimal speed from Lemma 3.11 and the bias function to show by induction on the states $s \geq 1$ that:

$$\partial h^{(S')}_s \leq \partial h^{(S'+1)}_s \text{ and } \pi^{(S')}_s \leq \pi^{(S'+1)}_s, \tag{3.19}$$

where $\partial h^{(S')}_s = h^{(S')}(s) - h^{(S')}(s-1)$ and $h^{(S')}$ is the bias function for the MDP with state space $\{0, \ldots, S'\}$.

We first have that $\pi^{(S')}_0 = \pi^{(S'+1)}_0 = 0$ and that

$$\partial h^{(S')}_1 = \frac{G^*_{S'}}{\lambda} \leq \frac{G^*_{S'+1}}{\lambda} = \partial h^{(S'+1)}_1,$$

where the inequality comes from Proposition 3.13. Let us now assume that for some $s \geq 1$, $\partial h^{(S')}_s \leq \partial h^{(S'+1)}_s$ and $\pi^{(S')}_s \leq \pi^{(S'+1)}_s$. $\partial h^{(S')}_s \leq \partial h^{(S'+1)}_s$, we directly have that $\pi^{(S')}_s \leq \pi^{(S'+1)}_s$.

To show the first inequality, we write the optimality equation (3.18) for S' and $S'+1$ with their respective optimal speed, and subtract one from the other to get:

$$G^*_{S'+1} - G^*_{S'} = -(s\mu + \pi_s^{(S'+1)})\partial h_s^{(S'+1)} + (s\mu + \pi_s^{(S')})\partial h_s^{(S')} + \lambda_s(\partial h_{s+1}^{(S'+1)} - \partial h_{s+1}^{(S')}).$$

Using Proposition 3.13 and the induction assumption, we get

$$\lambda_s(\partial h_{s+1}^{(S'+1)} - \partial h_{s+1}^{(S')}) \geq 0.$$

Then using Lemma 3.11, as $\partial h_{s+1}^{(S')} \leq \partial h_{s+1}^{(S'+1)}$ and w'^{-1} is increasing,

$$\pi_{s+1}^{(S'+1)} = w'^{-1}(\partial h_{s+1}^{(S'+1)}) \geq w'^{-1}(\partial h_{s+1}^{(S')}) = \pi_{s+1}^{(S')}.$$

The induction is therefore complete, and the optimal speed policy is increasing. □

3.4.2 Convergence Results and Proof of Theorem 3.2

For the truncated MDP in discrete time, let g^*, $\pi^{(S')}$ and $h^{(S')}$ be the optimal average cost, the optimal policy and its bias. They satisfy the optimality equation.

$$g^* + h^{(S')}(i) = \frac{1}{U}\left(c(s, \pi_s^{(S')}) + \sum_{s'} h^{(S')}(s') p_{s,s'}^{S'}(\pi_s^{(S')})\right) \quad \forall s \leq S'.$$

This implies that for the truncated model in continuous time, $J_{S'} = Ug^*$, $\pi^{(S')}$ and $h^{(S')}$ also satisfy the optimality equation

$$G_{S'} = c(s, \pi_s^{(S')}) + \sum_{s'} h^{(S')}(s') q_{s,s'}^{S'}(\pi_s^{(S')}) \quad \forall s \leq S'.$$

Furthermore, for all S', $G_{S'} \leq C\lambda$ by Lemma 3.6, $\pi^{(S')} \leq B^{S'} \leq B$ by Theorem 3.4 and since the function w' is increasing, $h^{(S')}(s) \leq sw'(B)$ for all $s \leq S'$ by Lemma 3.11.

Now, by the monotonicity of $G_{S'}$ (Proposition 3.13) and the monotonicity of $\pi^{(S')}$ and $h^{(S')}$ (Proposition 3.14), they all converge to finite non-negative limits when S' goes to infinity, denoted respectively by G_∞, $\pi^{(\infty)}$ and $h^{(\infty)}$.

As for the rates, $q_{s,s'}^{S'}(a)$ also converges monotonically to $q_{s,s'}(a)$ and is continuous in a. Finally, the immediate cost c is continuous in π. The monotone convergence theorem implies that these limits satisfy an optimality equation,

$$G_\infty = c(s, \pi_s^{(\infty)}) + \sum_{s'} h^{(\infty)}(s') q_{s,s'}(\pi_s^{(\infty)}) \quad \forall s \in \mathbb{N}.$$

This shows that these limits are respectively the optimal average cost (G_∞), an optimal policy ($\pi^{(\infty)}$) and its bias ($h^{(\infty)}$) for the original MDP. This completes the proof of Theorem 3.2.

3.5 Cost and Deadline-Miss Probability Approximations

Our main result, Theorem 3.2, may suggest to consider the simple policy π^B defined by $\pi_s^B = B\,\mathbb{1}_{\{s>0\}}$ for all $s \geq 0$, where B is defined in (3.4) and $\mathbb{1}_A$ is the indicator function of the event A. Thus, π^B uses constant speed B whenever the system is busy. In this section, we numerically show that the dynamics induced by π^B are "close" to the ones induced by π^* in the sense that the optimal average cost J^{π^*} is very well approximated by G^{π^B}. Then, we investigate mathematical properties of the Markov chain induced by π^B and we give an upper bound on the stationary probability of missing deadlines as a function of the model parameters. In particular, by varying the cost parameter C, this bound can be used to keep such probabilities below a desired threshold.

In the following, we let ν^{π^B} denote the stationary probability of the Markov chain induced by policy π^B. Using the detailed balance equations, we obtain

$$\nu_s^{\pi^B} = \nu_0^{\pi^B} \frac{\lambda^s}{\prod_{s'=1}^{s}(\mu s' + B)}, \qquad s \geq 1$$

and, using $\sum_{s\geq 0} \nu_s^{\pi^B} = 1$,

$$\nu_0^{\pi^B} = \left(\sum_{s\geq 0} \frac{\lambda^s}{\prod_{s'=1}^{s}(\mu s' + B)} \right)^{-1} = \frac{\mu}{B} \frac{\left(\frac{\lambda}{\mu}\right)^{\frac{B}{\mu}}}{e^{\frac{\lambda}{\mu}} \gamma\left(\frac{B}{\mu}, \frac{\lambda}{\mu}\right)} \tag{3.20}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function.

We will refer to the following proposition, which can be proven by summing the detailed balance equations of the underlying Markov chain as done in the proof of Lemma 3.7; for this reason, we omit the proof.

Proposition 3.15
Let X be distributed as ν^{π^B}. Then, $\mathbb{E}[X] = \frac{1}{\mu}(\lambda - B(1-\nu_0^{\pi^B}))$.

3.5.1 Approximation of the Average Cost

By definition, the average cost induced by π^B, G^{π^B}, is an upper bound on the optimal average cost, G^{π^*}. In particular,

$$G^{\pi^*} \leq G^{\pi^B} = \sum_{s\geq 0} c(s,\pi^B)\,\nu_s^{\pi^B} = C\mu \sum_{s\geq 1} s\,\nu_s^{\pi^B} + \sum_{s\geq 1} w(\pi_s^B)\,\nu_s^{\pi^B} \tag{3.21a}$$

$$= C(\lambda - B(1-\nu_0^{\pi^B})) + w(B)(1-\nu_0^{\pi^B}). \tag{3.21b}$$

where the last equality follows by Lemma 3.15.

3.5.2 Deadline-Miss Probabilities

Let us consider the probability p_M that a job misses its deadline under the stationary regime of policy π^B. This is defined by

$$p_M := \sum_{s\geq 1} \frac{\nu_s^{\pi^B}}{1-\nu_0^{\pi^B}} \frac{\mu s}{\mu s + B}. \tag{3.22}$$

Our objective here is to control p_M by fine-tuning the model parameter C. In other words, we want to design C such that p_M remains below a given threshold. Though the structure of $\nu_0^{\pi^B}$ in (3.20) implies that the exact relation between C and p_M is not trivial, this problem can be clearly solved numerically. Nonetheless, we aim at developing an upper bound on p_M allowing for a simple analytical evaluation.

The following proposition provides a first upper bound on p_M.

Proposition 3.16

$$p_M \leq \frac{1}{1-\nu_0^{\pi^B}} \left(1 - \frac{\mu\nu_0^{\pi^B}}{\mu + B}\right) - \frac{B}{\lambda}. \tag{3.23}$$

Proof. This proposition is proven by the following inequalities

$$p_M \leq \sum_{s\geq 1} \frac{{\nu_s^\pi}^B}{1-{\nu_0^\pi}^B} \frac{\mu(s+1)}{\mu(s+1)+B} = \frac{\mu}{\lambda(1-{\nu_0^\pi}^B)} \sum_{s\geq 1}(s+1)\frac{\lambda^{s+1}}{\prod_{s'=1}^{s+1}(\mu s'+B)}{\nu_0^\pi}^B$$
$$= \frac{1}{1-{\nu_0^\pi}^B}\frac{\mu}{\lambda}\sum_{s\geq 1}(s+1){\nu_{s+1}^\pi}^B = \frac{1}{1-{\nu_0^\pi}^B}\left(1-\frac{B}{\lambda}(1-{\nu_0^\pi}^B)\right) - \frac{\mu}{\mu+B}\frac{{\nu_0^\pi}^B}{1-{\nu_0^\pi}^B} \tag{3.24}$$
$$= \frac{1}{1-{\nu_0^\pi}^B} - \frac{B}{\lambda} - \frac{\mu}{\mu+B}\frac{{\nu_0^\pi}^B}{1-{\nu_0^\pi}^B}.$$

In (3.24), we have used Lemma 3.15. □

By coupling the underlying Markov chain under π^B with an auxiliary $M/M/\infty$ queue with arrival rate λ and service rate $\mu+B$, we notice that the state of the former is stochastically dominated by the latter. Therefore, ${\nu_0^\pi}^B \leq e^{-\frac{\lambda}{\mu+B}}$. Using that the mapping $x \mapsto \frac{1}{1-x}\left(1-\frac{\mu x}{\mu+B}\right)$ is increasing over $[0,1)$ and the previous inequality on ${\nu_0^\pi}^B$ in (3.23), we obtain

$$p_M \leq \overline{p_M} = \frac{1}{1-e^{-\frac{\lambda}{\mu+B}}}\left(1-\frac{\mu e^{-\frac{\lambda}{\mu+B}}}{\mu+B}\right) - \frac{B}{\lambda}. \tag{3.25}$$

We notice that $\overline{p_M} = 1$ if $B = 0$ and that $\overline{p_M} \to 0$ as $B \to \infty$.

Now, this bound $\overline{p_M}$ can also be used to adjust the cost of missed deadlines C so that the proportion of jobs that miss their deadline will stay below some acceptable level α.

Since Equation (3.25) cannot be inverted in close form, a first order Taylor expansion gives $\overline{p_M} \approx \frac{\mu}{\mu+B}$ when B goes to infinity. Using the value $B = \sqrt{C/3}$ for the classic power dissipation $w(a) = a^3$ (see Remark 3.3), we get $C \approx 3\mu^2\left(\frac{1-\overline{p_M}}{\overline{p_M}}\right)^2$. Therefore, $C = 3\mu^2\left(\frac{1-\alpha}{\alpha}\right)^2$ is the cost of missed deadlines that keeps the deadline-miss probability below α.

3.5.3 Accuracy Assessment

By means of numerical calculations, we now evaluate the accuracy of the bound in (3.21). For this purpose, we consider the two following set-ups. In the first one, we focus on one acceptable value for the probability to miss a deadline under policy π^B and we consider that such an acceptable level is $p_M = 0.1$. We let the deadline

rate μ and the arrival rate λ vary in $[0.1, 1]$ and $[0.1, 50]$, respectively. Each value of the couple (λ, μ) induces a unique value for B through equation (3.22) and for C through equation (3.4). Then, for each value of (λ, μ), we compute the percentage relative error

$$\mathcal{E} := \frac{G^{\pi^B} - G^{\pi^*}}{G^{\pi^*}} \times 100 \tag{3.26}$$

where G^{π^B} and G^{π^*} are computed numerically by truncating the state space to some S' large enough so that increasing S' does not change the average cost by more than 10^{-4}; in all cases, $S' \leq 250$.

Figure 3.5 despicts the level sets for the values of $\mathcal{E}$. We observe that $\mathcal{E}$ decreases

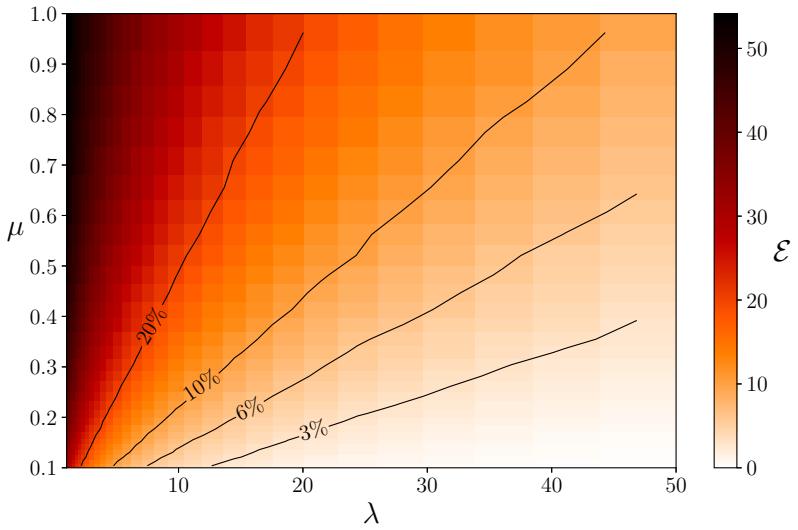


Figure 3.5: Level sets of the percentage relative error $\mathcal{E}$ with a fixed miss probability $p_M = 0.1$. The darker (lighter) the zone, the higher (lower) the error.

as the arrival rate increases and that it is small under a wide set of parameters. In particular, in heavy traffic conditions and with "large" deadlines the percentage relative error can be smaller than 3%. Thus, we conclude that our approximation is accurate within these conditions. On the other hand, in light load conditions and with "short" deadlines, $\mathcal{E}$ can be above 20% and in this case the optimal speed profile is more complex.

In the second set of experiments, we consider a fixed cost C instead of a fixed miss probability. Specifically, we let $C = 300$, which implies that the upper bound on the speed is $B = \sqrt{C/3} = 10$. For each pair (λ, μ), we compute the relative error $\mathcal{E}$ for policy π^B compared with the optimal policy, π^*.

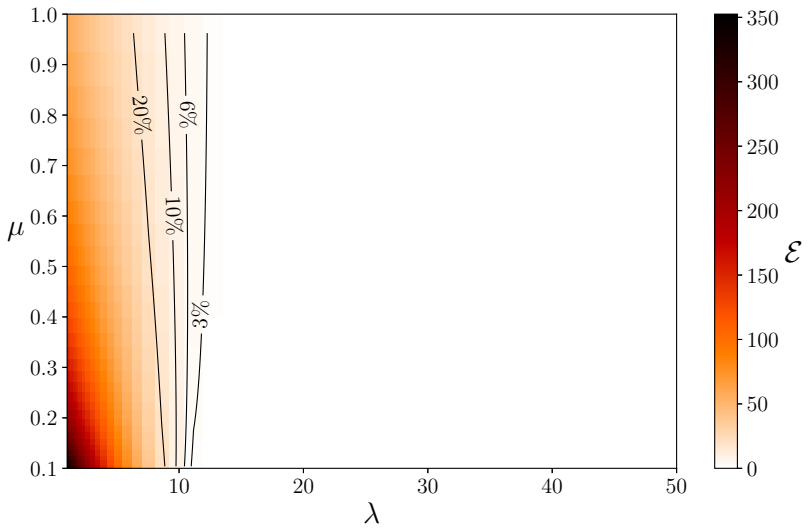


Figure 3.6: Level sets of the percentage relative error $\mathcal{E}$ with a fixed cost per deadline, $C = 300$.

The corresponding results are displayed in Figure 3.6. Unlike for the case with fixed probabilty, the error decreases sharply as the arrival rate λ increases. In particular, the error is very small in heavy traffic ($\lambda \geq B = 10$).

3.6 Conclusion and Perspectives

In a stochastic environment, it is well known that the problem of determining the load-dependent speed profile that a DVFS processor should follow to minimize the mean energy consumption under hard deadlines is difficult. In this chapter, the basic idea is to follow a Lagrangian approach where hard deadlines are replaced by *soft* deadlines, meaning that jobs are allowed to miss their deadline, though each missed deadline comes with a penalty that can be fine-tuned to keep the proportion of jobs missing their deadline as small as desired. The resulting advantage of this approach stands in the ability of formulating the problem above as a Markov decision process for which we can establish constructive structural results (Theorem 3.2). Beyond the existence of monotone optimal policies, we have found that the optimal speed is bounded from above by some constant that does not depend on the deadline and the arrival rates, which may be quite surprising, and that such constant let us define an extremely simple policy whose average cost is close to the optimal one.

There are some open questions that we leave as future research. First, we have assumed that the available processing speeds vary continuously on a compact set.

On the other hand, it may be convenient to consider the case where only a finite number of speeds is available. Second, it is interesting to investigate whether our results are insensitive to the job size and/or deadline distributions. This may be justified by the fact that the proposed queueing system is somewhat similar, when the speeds are bounded, to an $M/M/\infty$ queue. Furthermore, if service times were to follow a phase-type distribution, we could use again an MDP formulation to model the problem but the analysis would require more work as the state space would be much larger. Finally, we wonder whether the optimal policy could be "learned" and, in this respect, how the proposed upper bound could speed up the learning process.

4 Reinforcement Learning in a Birth-and-Death Process: Breaking the Dependence of the State Space

We introduced the example of a DVFS processor in Chapter 3 as a typical birth-and-death process we could study, with an arbitrarily large number of states. In Chapter 2, we saw how we could compute bounds on the bias of any policy in this case, and how the diameter would be exponentially large in a basic example. Considering these particularities, we will discuss in this chapter how we can learn the transitions and rewards in the DVFS example, were they unknown, using an adapted version of UCRL2. This adaptation lets us tackle the diameter problem and use the specific structure of a birth-and-death process to our advantage, despite the *a priori* longer exploration time of the MDP.

4.1 Introduction

In the context of undiscounted reinforcement learning in Markov decision processes (MDPs), it has been shown in the seminal work [Jaksch et al., 2010] that the total regret of any learning algorithm with respect to an optimal policy is lower bounded by $\Omega(\sqrt{DSAT})$, where S is the number of states, A the number of actions, T the time horizon and D the *diameter* of the MDP. Roughly speaking, the diameter is the mean time to move from any state s to any other state s' within an appropriate policy. In the literature, several efforts have been dedicated to approach this lower bound. As a result, learning algorithms have been developed with a total regret of $\tilde{O}(DS\sqrt{AT})$ in [Jaksch et al., 2010], $\tilde{O}(D\sqrt{SAT})$ in [Azar et al., 2017] and even $\tilde{O}(\sqrt{DSAT})$ according to [Tossou et al., 2019; Zhang and Ji, 2019]. These results may give a sense of optimality since the lower bound is attained up to some universal constant. However, lower bounds are based on the minimax approach, which relies on the worst possible MDP with given D, A and S. This means that

when a reinforcement learning algorithm is used on a given MDP, one can expect a much better performance.

One way to alleviate the minimax lower bound is to consider *structured reinforcement learning*, or equivalently MDPs with some specific structure. The exploitation of such structure may yield more efficient learning algorithms or tighter regret analyses of existing learning algorithms. In this context, a first example is to consider *factored* MDPs [Boutilier et al., 2000; Guestrin et al., 2003], i.e., MDPs where the state space can be factored into a number of components; in this case, roughly speaking, $S = K^n$ where n is the number of "factors" and K is the number of states in each factor. The regret of learning algorithms in factored MDPs has been analyzed in [Tian et al., 2020; Rosenberg and Mansour, 2020; Z. Xu and Tewari, 2020; Osband and Roy, 2014] and it is found that the S term of existing upper bounds can be replaced by nK. A similar approach is used in [Gast et al., 2021] to learn the optimal policy in stochastic bandits with a regret that is logarithmic in the number of states. There is also a line of research works that exploit the parametric nature of MDPs. Inspired by parametric bandits, a d-linear additive model was introduced in [Jin, Yang, et al., 2020], where it is shown that an optimistic modification of Least-Squares Value Iteration, see [Osband, Van Roy, et al., 2016], achieves a regret over a finite horizon H of $\tilde{O}(\sqrt{d^3H^3T})$ where d is the ambient dimension of the feature space (the number of unknown parameters). In this case, the regret does not depend on the number of states and actions and the diameter is replaced by the horizon. A discussion about the inapplicability of this approach to our case is postponed to Section 4.3.2.

The regret bounds discussed above and the discussion on the scaling of the diameter in Lemma 2.3 may suggest that the total regret of existing learning algorithm, when applied to queueing systems, is large. However, they often work well in practice and this bring us to consider the following question:

When the underlying MDP has the structure of a queueing system, do the diameter D or the number of states S actually play a role in the regret?

Our Contribution. In this chapter, we examine the previous question with respect to the class of control problems presented in Chapter 3. Specifically, an infinite sequence of jobs joins a service system over time to receive some processing according to the first-come first-served scheduling rule; the system can buffer at most S' jobs and in fact it corresponds to an M/M/1/S' queue. In addition, each job comes with a deadline constraint, and if a job is not completed before its deadline, then it becomes obsolete and is removed from the system. The controller chooses the server

processing speed and the objective is to design a speed policy for the server that minimizes its average energy consumption plus an obsolescence cost per deadline miss. Although this may look quite specific, this problem captures the typical characteristics of a controlled queue: i) the transition matrix has the structure of a birth-and-death process with jump probabilities that are affine functions of the state and ii) the reward is linear in the state and convex in the action. For any MDP in this class, defined in full details in Section 4.2, we show that the diameter is $D = \Omega(S^{S-2})$; see Subsection 4.5.2. Thus, without exploiting the particular structure of this MDP, the existing lower and upper bounds do not justify the reason why standard learning algorithms work efficiently here.

We provide a slight variation of the learning algorithm UCRL2, introduced in [Jaksch et al., 2010], and show in our main result that the resulting regret is upper bounded by $\tilde{O}(\sqrt{E_2 AT})$ where E_2 is a term that depends on the stationary measure of a reference policy defined in Section 4.2.1. Importantly, E_2 does not depend on S. Thus, efficient reinforcement learning can be achieved independently of the number of states by exploiting the stationary structure of the MDP. Let us provide some intuition about our result. First, one may think that any learning algorithm should visit each state a sufficient number of times, which justifies why the diameter of an MDP appears in existing regret analyses. However, this point of view does not take into account the fact that the value of an MDP is the scalar product of the reward and the stationary measure of the optimal policy. If this stationary measure is "highly non-uniform", then some states are rarely visited under the optimal policy and barely contribute to the value. In this case, we claim that the learner may not need to visit the rare states that often to get a good estimation of the value, and thus it may not need to pay for the diameter.

This chapter is based on the published work [Anselmi, Gaujal, and Rebuffi, 2022].

4.2 Controlled Birth-and-Death Processes for Energy Minimization

We place ourselves in the reinforcement learning framework presented in Section 2.2.

We will focus on a specific class of MDPs that has been introduced in Chapter 3, which provides a rather general example of a controlled birth-and-death process with convex costs on the actions and linear rates. We will denote by $\mathcal{M}$ the set of MDPs

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

After uniformization, it is shown in Chapter 3 that this control problem can be modeled as an MDP in discrete time with a "birth-and-death" transition matrix of size S. Specifically, we have an MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ where $\mathcal{S} = \{0, \ldots, S'\}$, with $s \in \mathcal{S}$ representing the number of jobs in the system, and $\mathcal{A} = \{0, \ldots, a_{\max}\}$, with $a \in \mathcal{A}$ representing the processor speed. Then, the transition probabilities under policy π are given by

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

where $U := \lambda_{\max} + S'\mu_{\max} + a_{\max}$ is a uniformization constant, $P_{s,s} = \frac{1}{U}(U - \lambda_s - \mu s - \pi(s))$ and $\lambda_s := \lambda\left(1 - \frac{s}{S'}\right)$ is the *decaying* arrival rate. We have replaced the constant arrival rate λ by a decaying arrival rate λ_i because we want to learn an optimal policy that does not exploit the buffer size S'; see Chapter 3 for further details. For conciseness, Figure 4.1 displays the transition diagram of the Markov chain induced by policy π.

Finally, the reward is a combination of C, the constant cost due to a departing job missing its deadline and $w(a)$, an arbitrary convex function of a, giving the energy cost for using speed a. The immediate cost $c(s, a)$ in state s under action a is a random variable whose value is $w(a)/U + C$ with probability $s\mu/U$ (missed deadline) and $w(a)/U$ otherwise. To keep in line with the use of rewards instead of

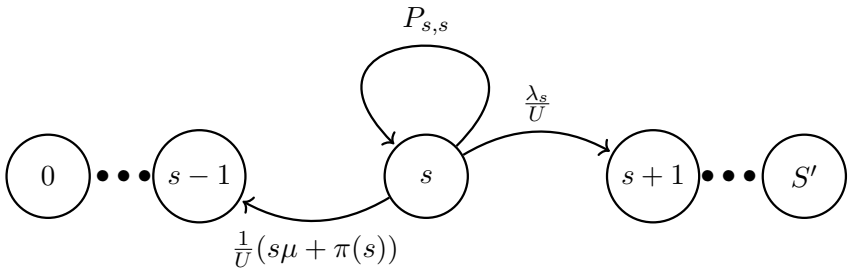


Figure 4.1: Transition diagram of the Markov chain induced by policy π of an MDP in $\mathcal{M}$.

costs, we introduce a bound on the cost, $r_{\max} := C + w(a_{\max})/\mu$ so that the reward in state s under action a is a positive and bounded random variable given by

$$R(s,a) := r_{\max} - c(s,a). \tag{4.1}$$

As defined in Chapter 2 $g^*(M)$ is the optimal average cost and $g(M,\pi)$ is the average cost induced by policy π, where π belongs to the set of deterministic and stationary policies Π. Since the underlying Markov chain induced by any policy is ergodic, we observe that

$$g(M,\pi) = \sum_{s=0}^{S'} \mathbb{E}[r(s,\pi(s))]\nu_s^\pi, \tag{4.2}$$

where ν^π is the stationary measure under policy π. In Chapter 3, it has been shown that the optimal policy is unique and will be denoted by π^*.

4.2.1 Properties of $\mathcal{M}$

In the following, we will use the "reference" (or bounding) policy $\pi^0(s) = 0$ for all $s \in \mathcal{S}$, which thus assigns speed 0 to all states. This policy provides a stochastic bound on all policies in the following sense. Let s_t^π be the state under policy π and let $\leq_{st}$ denote the *stochastic order* [Shaked and Shanthikumar, 1994b]; given two random variables X and Y on $\mathbb{R}_+$, we recall that $X \leq_{st} Y$ if $\mathbb{P}(X \geq s) \leq \mathbb{P}(Y \geq s)$ for all s.

Lemma 4.1
Consider an MDP in $\mathcal{M}$. For all t and policy $\pi \in \Pi$, $s_t^\pi \leq_{st} s_t^{\pi^0}$, provided that $s_1^\pi \leq_{st} s_1^{\pi^0}$.

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

Therefore, $\mathbb{P}(s_t^\pi \geq s) \leq \mathbb{P}(s_t^{\pi^0} \geq s)$ for all s and t, which also implies that the respective stationary measures are comparable, i.e., $\sum_{i=s}^{S'} \nu_i^\pi \leq \sum_{i=s}^{S'} \nu_i^{\pi^0}$.

Let us now consider $h^\pi(s)$, the *bias* at state s of a policy π, defined by

$$h^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \left(r\left(s_t^\pi, \pi(s_t^\pi)\right) - g^\pi(M) \right) \mid s_1^\pi = s \right], \quad \forall\, 0 \leq s \leq S'. \tag{4.3}$$

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

The following result on the variations of the bias holds (see Proposition 2.5):

Lemma 4.2
The local variation of the bias $\partial h^\pi(s)$ is bounded: $|\partial h^\pi(s)| \leq \Delta(s)$ with $\Delta(s) := 2r_{\max} e^{\lambda/\mu}(1+\log s)$ for all $1 \leq s \leq S'$.

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

4.2.2 Applying UCRL2 in $\mathcal{M}$

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

In the remainder, we will apply UCRL2 over an MDP in $\mathcal{M}$ with a change in the confidence bounds to take into account the support of P. Indeed we impose that the confidence set $\mathcal{M}_k$ only contains matrices with the same support as P. Moreover, the confidence bounds of UCRL2 (see Chapter 2) are replaced by:

$$\forall (s,a), \quad |\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq r_{\max} \sqrt{\frac{2\log(2At_k)}{\max\{1, N_{t_k}(s,a)\}}} \tag{4.4}$$

$$\forall (s,a), \quad \|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{8\log(2At_k)}{\max\{1, N_{t_k}(s,a)\}}} \tag{4.5}$$

Removing S in the confidence bounds does help to reduce the regret. However, by using existing analysis, this only removes a factor $\sqrt{S}$ in the regret bound (for example, see [Azar et al., 2017]).

This adapted version of UCRL2 can be rewritten as follows:

Algorithm 2: The adapted UCRL2 algorithm.

Set $t = 1$, $k = 1$;
while $t \leq T$ **do**
  Initialize episode k with $t_k := t$
  Compute the confidence region $\mathcal{M}_k$ as in (4.5);
  Find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ with "Extended Value Iteration" such that

$$g(\tilde{M}_k, \tilde{\pi}_k) \geq \max_{M_k \in \mathcal{M}_k} \max_{\pi} g(M_k, \pi) - \frac{\delta_{\max}}{\sqrt{t_k}}. \tag{4.6}$$

  Exploration: while $V_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_{t_k}(s_t, \tilde{\pi}_k(s_t))\}$, **do**
    1. Choose action $a_t = \tilde{\pi}_k(s_t)$;
    2. Observe s_{t+1};
    3. Update $V_k(s_t, a_t) := V_k(s_t, a_t) + 1$;
    4. Set $t := t + 1$.

Finally, note that Algorithm 2 does not benefit from the parametric nature of the MDPs in $\mathcal{M}$, which is essentially defined by three parameters (λ, μ and C) and the real convex function $w(\cdot)$.

4.3 Regret of the Adapted UCRL2 Algorithm on $\mathcal{M}$

Our objective is to develop an upper bound on the regret of the learning Algorithm 2 when applied to MDPs in our class $\mathcal{M}$. The driving idea is to construct a bound that exploits the structure of the stationary measure of all policies, as they all make some states hard to reach, and to control the number of visits to these states to get a new type of bound.

4.3.1 Main Result

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

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

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

Theorem 4.3
Let $M \in \mathcal{M}$. Define $Q_{\max} := \left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^2 \log\left(\left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^4\right)$. Then, assuming $T \geq \max\{\frac{e^8}{2A}, S^2A^2\}$ and $SA > 4$,

$$\mathbb{E}\left[\mathit{Reg}(M, \text{Algorithm 2}, T)\right] \leq 19\sqrt{E_2 AT \log(2AT)} + 97 r_{\max} D^2 SA \max\{Q_{\max}, T^{1/4}\} \log^2(2AT). \quad (4.7)$$

Here, $E_2 \leq 60 e^{2\lambda/\mu} r_{\max}^2 \left(1 + \frac{\lambda^2}{\mu^2}\right)$, so that the regret satisfies

$$\mathbb{E}\left[\mathit{Reg}(M, \text{Algorithm 2}, T)\right] = O\left(r_{\max} e^{\lambda/\mu} \sqrt{AT\left(1 + \frac{\lambda^2}{\mu^2}\right)\log(AT)}\right).$$

Before giving a sketch of the proof, let us comment on the bound (4.7). Although the first term is of order $\sqrt{T}$ with a multiplicative constant independent of S - as desired - the second term, of order $T^{1/4}$, contains very large terms. However, its interest lies in the novel approach used in the proof that uses the stationary behavior of the algorithm.

4.3.2 Comparison with Other Bounds

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

First, let us compare with the bound given in [Jaksch et al., 2010], which states that with probability $1-\delta$, $\text{Reg}(M,T) \leq 34DS\sqrt{AT\log\left(\frac{T}{\delta}\right)}$ for any $T > 1$. For any $M \in \mathcal{M}$, the diameter grows as S^S (see Subsection 4.5.2), thus this bound is very loose here. More recent works have improved this bound by replacing the term D by the local diameter of the MDP [Bourel et al., 2020]. In Subsection 4.5.2, we show that the local diameter grows again as S^S for any $M \in \mathcal{M}$, and thus these results do not yield significant improvements. Other papers show that the diameter can be replaced by the span of the bias, see [Fruit, Pirotta, Lazaric, and Ortner, 2018; Zhang and Ji, 2019]. This has a big impact because the span of the bias, for any $M \in \mathcal{M}$, is linear in S (instead of S^S for the diameter); see Subsection 4.5.2. However, this is still not as good as the bound given in Theorem 4.3, which is asymptotically independent of S.

Now, let us compare with existing bounds for *parametric* MDPs, as mentioned in the introduction. The d-linear additive model, $d < S$, introduced in [Jin, Yang, et al., 2020] assumes that $P(\cdot|s,a) = \langle\phi(s,a), \theta(\cdot)\rangle$, where $\phi(s,a)$ is a known feature mapping and θ is an unknown measure on $\mathbb{R}^d$. This form of $P(\cdot|s,a)$ implies that the transition kernel is of rank d. Unfortunately, this property does not hold true in birth-and-death processes. In fact, the kernel of any $M \in \mathcal{M}$ has almost *full* rank under all policies. The *linear mixture model* introduced in [Zhou et al., 2021] assumes instead that $P(s'|s,a) = \langle\phi(s'|s,a), \theta\rangle$, $\theta \in \mathbb{R}^d$. This is more adapted to our case, which can be (almost) seen as a linear mixture model of dimension $d = 3$. The bound on the discounted regret of the algorithm proposed in [Zhou et al., 2021] is $\text{Reg}(M,T) \leq d\sqrt{T}/(1-\gamma)^2$ where γ is a discount factor. In contrast to our work, this regret analysis holds for *discounted* problems, where we remark that both the diameter and the span are irrelevant. On the other hand, both are replaced by a term of the form $(1-\gamma)^{-2}$, which implies that the previous bound grows to infinity as $\gamma \uparrow 1$. More recently, a regret bound of $O(d\sqrt{DT})$ has been proven in [Wu et al., 2022] in the undiscounted case, that is the case considered in our work. However, the algorithm presented in that reference highly depends on the diameter and is unsuitable for MDPs with a birth-and-death structure.

Finally, our bound depends on the second moment of the stationary measure of a reference policy, i.e., E_2, which can be bounded independently of the state space size. This is structurally different from the ones existing in the literature. We believe

that this structure holds as well in a class of MDPs much larger than $\mathcal{M}$. In particular, if ν is the stationary measure of some bounding/reference policy, and if the critical quantity $\mathbb{E}_\nu[\Delta \cdot f]$ is small for a well chosen function f, then the regret of a learning algorithm navigating the MDP should also be small. A deeper analysis is left as future work.

4.3.3 Sketch of the Proof

Our proof for Theorem 4.3 is technical and will be provided in Section 4.4. First, in this section, we present the main ideas and its general structure. It initially relies on the regret analysis of UCRL2 developed in [Jaksch et al., 2010], and the differences are highlighted below. First, we consider the mean rewards and split the regret into episodes to separately treat the cases where the true MDP is in the confidence set of optimistic MDPs $\mathcal{M}_k$ or not. Thus, let $R_k := \sum_{s,a} V_k(s,a)(g^* - r(s,a))$ denote the regret in episode k. This split can be written:

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

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

To control R_{out}, we use, as in [Jaksch et al., 2010], the stopping criterion and the confidence bounds. This gives $\mathbb{E}[R_{\text{out}}] \leq r_{\max} S$, so that the regret due to episodes where the confidence regions fail will be negligible next to the main terms. Then, when the true MDP belongs to the confidence region, we use the properties of Extended Value Iteration (EVI) to decompose R_{in} into

$$\underbrace{\sum_{k,s,a} V_k(s,a)(\tilde{r}_k - r(s,a))}_{R_{\text{rewards}}} + \underbrace{\sum_k V_k \left(\tilde{P}_k - I\right) \tilde{h}_k}_{R_{\text{bias}}} + \underbrace{\sum_k V_k \left(\tilde{P}_k - I\right) d_k + 2 r_{\max} \sum_{k,s,a} \frac{V_k(s,a)}{\sqrt{t_k}}}_{R_{\text{EVI}}},$$

where we let V_k be also the vector of the state-action counts, $\tilde{P}_k$ and $\tilde{h}_k$ are respectively the transition matrix and the bias in $\tilde{M}_k$ under policy $\tilde{\pi}_k$, and d_k is the profile difference between the last step of EVI and the bias (see Subsection 4.4.3).

We now show how to handle R_{rewards}, R_{EVI}, R_{bias}. First, we deal with the terms that do not involve the bias. Using the confidence bound on the rewards (see Subsection 4.4.3:

$$R_{\text{rewards}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{4.8}$$

Now, let us consider R_{EVI}. Since d_k becomes arbitrarily small after enough iterations of EVI (see Subsection 4.4.1), for $T \geq \frac{e^8}{2AT}$, we get

$$R_{\text{EVI}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{4.9}$$

The analysis of R_{bias} is different from the one in [Jaksch et al., 2010]: While in [Jaksch et al., 2010] the bias is directly bounded by the diameter, we can use the variations of the bias to control the regret more precisely. Using P_k and h_k, i.e., the transitions and the bias in the true MDP under $\tilde{\pi}_k$, R_{bias} is further decomposed into:

$$\underbrace{\sum_k V_k \left(\tilde{P}_k - P_k\right) h_k}_{R_{\text{trans}}} + \underbrace{\sum_k V_k \left(\tilde{P}_k - P_k\right) \left(\tilde{h}_k - h_k\right)}_{R_{\text{diff}}} + \underbrace{\sum_k V_k \left(P_k - I\right) \tilde{h}_k}_{R_{\text{ep}}}.$$

The term R_{ep} can be treated in a similar manner as in [Jaksch et al., 2010] by bounding the bias terms with the diameter to apply an Azuma-Hoeffding inequality (see Subsection 4.4.3). Here, we obtain

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

Next, we show in Subsection 4.4.3 that R_{diff} does not contribute to the main term of the regret. This is one of the hard point in our proof. First, linear algebra techniques are used to bound $||\tilde{h}_k - h_k||_\infty$ by $D(2r_{\max} D||\tilde{P}_k - P_k||_\infty + ||\tilde{r}_k - r_k||_\infty)$. Each norm is then bounded using Hoeffding inequality. This introduces the special quantity $N_{t_k}(x_k, \pi_k(x_k))$ that yields to the worst confidence bound in episode k. Then, an adaptation of McDiarmid's inequality to Markov chains is used to show that $N_{t_k}(x_k, \pi_k(x_k)) \geq (t_{k+1} - t_k)\nu^{\pi^{\max}}(S')/2$ with high probability, where $\nu^{\pi^{\max}}(S')$ is the stationary measure of state S' under the uniform policy $\pi^{\max}(s) = a_{\max}$. This eventually implies that

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

where $Q_{\max} := \left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^2 \log\left(\left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^4\right)$.

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

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

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

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

We can now aggregate R_{trans}, R_{rewards} and R_{EVI} to compute the main term of the regret (see Subsection 4.4.3). Here, the key ingredient is to bound

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

independently of S. This is the second main difference with [Jaksch et al., 2010]. Instead of exploring the MDP uniformly, we know that the algorithm will mostly visit the first states of the MDP, regardless of the chosen policy. As shown in [Jaksch et al., 2010], for a fixed state s:

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

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

$$\mathbb{E}\left[\sum_{s\geq 0} f(s) N_t(s)\right] \leq t \sum_{s\geq 0} f(s) \nu^{\pi^0}(s). \tag{4.10}$$

Let us choose $f : s \mapsto \frac{\max\{1, s(s-1)\}}{(\Delta(s+1)+r_{\max})^2}$ and let $F := \sum_s f(s)^{-1} \leq 60 e^{2\lambda/\mu} r_{\max}^2$. Define also $E_2 := F\mathbb{E}^{\pi^0}\left[(\Delta + r_{\max})^2 \cdot f\right]$. Then,

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

In Subsection 4.4.3, we further show that $E_2 \leq 60e^{2\lambda/\mu} r_{\max}^2 \left(1 + \frac{\lambda^2}{\mu^2}\right)$. Therefore, for the three main terms, we obtain

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 19\sqrt{E_2 AT \log(2AT)} \tag{4.11}$$

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

4.4 Proof of Theorem 4.3

From now on, the following sections are organized as follows: We first provide some insights on extended value iterations useful in our construction of the regret.Then, the detailed proof of theorem 4.3 is given with bounds on the five terms in our decomposition of the regret. A final section provides technical lemmas about MDPs in $\mathcal{M}$.

4.4.1 Extended Value Iteration

For each episode k, we use the extended value iteration algorithm described in [Jaksch et al., 2010] to compute $\tilde{\pi}_k$ and $\tilde{M} \in \mathcal{M}_k$, an optimistic policy and MDP. The values we iteratively get are defined in the following way:

$$\begin{aligned} u_0^{(k)}(s) &= 0 \\ u_{i+1}^{(k)}(s) &= \max_{a \in \mathcal{A}} \left\{ \tilde{r}(s,a) + \max_{p(\cdot) \in \mathcal{P}(s,a)} \left\{ \sum_{s \in S} p(s') u_i^{(k)}(s') \right\} \right\}, \end{aligned} \tag{4.12}$$

where $\tilde{r}$ is the maximal reward from (4.4) and $\mathcal{P}(s,a)$ is the set of probabilities from (4.5).

Now, from [Jaksch et al., 2010, Theorem 7], we obtain the following lemma on the iterations of extended value iteration.

Lemma 4.4
For episode k, denote by i the last step of extended value iteration, stopped when:

$$\max_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} - \min_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} < \frac{r_{\max}}{\sqrt{t_k}}. \tag{4.13}$$

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

$$\tilde{g}_k := \min_s g(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} g(M', \pi, s') - \frac{r_{\max}}{\sqrt{t_k}}. \tag{4.14}$$

Moreover from [Martin L. Puterman, 1994, Theorem 8.5.6]:

$$\left| u_{i+1}^{(k)}(s) - u_i^{(k)}(s) - \tilde{g}_k \right| \leq \frac{r_{\max}}{\sqrt{t_k}}, \tag{4.15}$$

and when the optimal policy yields an irreducible and aperiodic 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:

$$\tilde{h}_k(s_1) = \mathbb{E}_{s_1}\left[\sum_{t=1}^{\infty}(\tilde{r}(s_t, a_t) - \tilde{g}_k)\right]. \tag{4.16}$$

By choosing iteration i large enough, from [Martin L. Puterman, 1994, Equation 8.2.5], we can also ensure that:

$$\left| u_i^{(k)}(s) - (i-1)\tilde{g}_k - \tilde{h}_k(s) \right| < \frac{r_{\max}}{2\sqrt{t_k}}, \tag{4.17}$$

so that we can define the following difference

$$d_k(s) := \left| u_i^{(k)}(s) - \min_s u_i^{(k)}(s) - \left(\tilde{h}_k(s) - \min_s \tilde{h}_k(s)\right) \right| < \frac{r_{\max}}{\sqrt{t_k}}. \tag{4.18}$$

4.4.2 Regret when M is out of the Confidence Bound

Let us compute $\mathbb{E}[\text{Reg}]$, the expected regret. We will mainly follow the approach in [Jaksch et al., 2010, Section 4], with a few tweaks. We start by splitting the regret into a sum over episodes and states.

We remind that $r(s,a)$ is the overall mean reward and $N_T(s,a)$ the total count of visits. We also define $R_k(s) := \sum_a V_k(s,a)(g^* - r(s,a))$ the regret at episode k induced by state s, with $V_k(s,a)$ the number of visit of (s,a) during episode k and K_T the number of episodes at time T.

Let $R_{\text{in}} := \sum_s \sum_{k=1}^{K_T} R_k(s) \mathbb{1}_{M \in \mathcal{M}_k}$ and $R_{\text{out}} := \sum_s \sum_{k=1}^{K_T} R_k(s) \mathbb{1}_{M \notin \mathcal{M}_k}$. We therefore have the split:

$$\mathbb{E}\left[Reg\right] \leq \mathbb{E}\left[R_{\text{in}}\right] + \mathbb{E}\left[R_{\text{out}}\right]. \tag{4.19}$$

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

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

Taking the expectations:

$$\begin{aligned}
\mathbb{E}\left[R_{\text{out}}\right] &\leq r_{\max} \sum_{t=1}^{T} t \mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \\
&\leq r_{\max} \sum_{t=1}^{T} \frac{tS}{2t^3} \leq r_{\max} \sum_{t=1}^{T} \frac{S}{2t^2} \text{ by Lemma 4.7} \\
&\leq r_{\max} S. \qquad (4.20)
\end{aligned}$$

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

4.4.3 Regret Terms when M is in the Confidence Bound

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

We defined $\tilde{\pi}_k$ the optimistic policy computed at episode k, now define $\tilde{P}_k := (\tilde{p}_k(s'|s, \tilde{\pi}_k(s)))$ the transition matrix of that policy on the optimistic MDP $\tilde{M}_k$. We also let V_k be the row vector of visit counts during episode k. Following the same

steps as in [Jaksch et al., 2010], we get the inequality on the regret of episode k, assuming $M \in \mathcal{M}_k$, using Lemma 4.4:

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

Then with (4.15) and using the definition of the iterated values from EVI, we have for a given state s and $a_s := \tilde{\pi}_k(s)$:

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

so that:

$$R_{\text{in},k} \leq V_k \left(\tilde{P}_k - I \right) u_i + \sum_{s,a} V_k(s,a)(\tilde{r}_k - r(s,a)) + 2 r_{\max} \sum_{s,a} \frac{V_k(s,a)}{\sqrt{t_k}}.$$

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

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

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

$$V_k \left(\tilde{P}_k - I \right) u_i = V_k \left(\tilde{P}_k - I \right) \tilde{h}_k + V_k \left(\tilde{P}_k - I \right) d_k.$$

so that:

$$\begin{aligned} R_{\text{in}} \leq \underbrace{\sum_k \sum_{s,a} V_k(s,a)(\tilde{r}_k - r(s,a))}_{R_{\text{rewards}}} + \underbrace{\sum_k V_k \left(\tilde{P}_k - I \right) \tilde{h}_k}_{R_{\text{bias}}} \\ + \underbrace{\sum_k V_k \left(\tilde{P}_k - I \right) d_k + 2 r_{\max} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{t_k}}}_{R_{\text{EVI}}}. \end{aligned}$$

Then, using reward boounds (4.4), as $M \in \mathcal{M}_k$, and noticing that $N_{t_k} \leq t_k$:

$$R_{\text{rewards}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{4.21}$$

For the term $V_k\left(\tilde{P}_k - I\right) d_k$, which does not appear in the analysis of [Jaksch et al., 2010], we obtain

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

where in the last inequality we used that $\max\{1, N_{t_k}(s,a)\} \leq t_k \leq T$. Thus, for $T \geq \frac{e^8}{2A}$ the regret term coming from the consequences and approximations of EVI satisfies

$$R_{\text{EVI}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{4.22}$$

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

$$\underbrace{\sum_k V_k\left(\tilde{P}_k - P_k\right) h_k}_{R_{\text{trans}}} + \underbrace{\sum_k V_k\left(\tilde{P}_k - P_k\right)\left(\tilde{h}_k - h_k\right)}_{R_{\text{diff}}} + \underbrace{\sum_k V_k\left(P_k - I\right)\tilde{h}_k}_{R_{\text{ep}}}.$$

Overall:

$$\begin{aligned} R_{\text{in}} \leq \underbrace{\sum_k V_k\left(\tilde{P}_k - P_k\right) h_k}_{R_{\text{trans}}} + \underbrace{\sum_k V_k\left(\tilde{P}_k - P_k\right)\left(\tilde{h}_k - h_k\right)}_{R_{\text{diff}}} + \underbrace{\sum_k V_k\left(P_k - I\right)\tilde{h}_k}_{R_{\text{ep}}} \\ \underbrace{+ r_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}}_{R_{\text{EVI}} + R_{\text{rewards}}}. \end{aligned} \tag{4.23}$$

Bound on R_{trans}

Let us deal with the first term R_{trans}. To bound this term, we will use our knowledge of the bias in the true MDP h_k and on the control of the difference of the transition matrices, and for the second term we will control the difference of the biases.

Notice that for a fixed state $0 \leq s \leq S'$:

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

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

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

where in the last inequality, we used the knowledge on the bounded variations of the bias from Lemma 4.2, and that the optimistic MDP has transitions close to the true transitions.

Bound on R_{diff}

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

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

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

Then using Lemma 4.13, and noticing that the bias $\tilde{h}_k$ and the quantity $\|\sum_{t=1}^T \tilde{P}_k^t \tilde{r}_k\|$ is bounded by the same diameter D, using the same argument as in [Jaksch et al., 2010] (Equation (11)):

$$\|\tilde{h}_k - h_k\|_\infty \leq 12 T_{hit} r_{\max} D \sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}}. \tag{4.24}$$

Hence,

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

where in the last inequality we have used (4.24) and that by definition of D, for S large enough

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

and we called

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

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

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

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

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

Using the stopping criterion for episodes:

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

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

$$\mathcal{E} = \left\{\forall k \text{ s.t } I_k > I(T), \ \frac{1}{\max\{1, \nu(x_k, a_k)\}} \leq \frac{2}{\nu^{\pi^{\max}}(S') I_k}\right\}.$$

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

$$\begin{aligned}\Sigma &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \sum_{k, I_k > I(T)} \frac{2}{\nu^{\pi^{\max}}(S')} + \mathbb{1}_{\bar{\mathcal{E}}} \sum_{k, I_k > I(T)} I_k \\ &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \left(K_T - K_{\leq I}\right) \frac{2}{\nu^{\pi^{\max}}(S')} + \mathbb{1}_{\bar{\mathcal{E}}} T.\end{aligned}$$

We use Corollary 4.16 to get $\mathbb{P}(\bar{\mathcal{E}}) \leq \frac{1}{4T}$, so that when taking the expectation:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_{\leq I}\right] I(T) + \mathbb{E}\left[\left(K_T - K_{\leq I}\right)\right] \frac{2}{\nu^{\pi^{\max}}(S')} + \frac{1}{4}.$$

Now using Lemma 2.10, $SA \geq 4$, $I(T) \geq \frac{2}{\nu^{\pi^{\max}}(S')}$ and that $\frac{1}{\log 2} + \frac{1}{4} \leq 2$:

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

We therefore have that:

$$\mathbb{E}\left[R_{\text{diff}}\right] \leq 96 r_{\max} SAD^2 I(T) \log^2\left(2AT\right). \tag{4.25}$$

Bound on the Main Terms: Exploiting the Stochastic Ordering

In Section 4.3.3 we have shown that:

$$R_{\text{trans}} \leq 4\sqrt{2\log\left(2AT\right)} \sum_{s,a} \frac{\Delta(s+1) V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{4.26}$$

To control this term as well as R_{EVI} (4.22) and R_{rewards} (4.21), we need to control the terms in the sum in a way that does not make the parameters D or S appear, as this will be one of the main contributing terms. To do so, we need to sum over the episodes and take the expectation, so that with Lemma 2.11 from Chapter 2, we get:

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

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

Lemma 4.5

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

$$\mathbb{E}\left[\sum_{s' \geq s} f(s') N_t(s')\right] \leq t \sum_{s' \geq s} f(s') \nu^{\pi^0}(s'). \tag{4.27}$$

Proof. Let $s \in \mathcal{S}$. For any state s', define $N_t^{\nu^{\pi^0}, \pi^0}(s')$ the number of visits when the starting state follows the initial distribution ν^{π^0}, and the MDP always executes the policy π^0 at every time step instead of the policy determined by Algorithm 2. Notice already that for any state s':

$$\mathbb{E}\left[N_t^{\nu^{\pi^0}, \pi^0}(s')\right] = t \nu^{\pi^0}(s')$$

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

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

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

$$\begin{cases} (f(x) - f(x-1)) \sum_{s' \geq x} N_t(s') \leq_{st} (f(x) - f(x-1)) \sum_{s' \geq x} N_t^{\nu^{\pi^0}, \pi^0}(s') \\ f(s-1) \sum_{s' \geq s} N_t(s') \leq_{st} f(s-1) \sum_{s' \geq s} N_t^{\nu^{\pi^0}, \pi^0}(s'), \end{cases} \tag{4.28}$$

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

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

which simplifies to:

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

Now summing with the second equation in (4.28), we get the following equation:

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

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

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

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}(\Delta(s+1)+r_{\max})\frac{V_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{A}\sum_{s\geq 0}(\Delta(s+1)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]}, \tag{4.29}$$

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

$$\begin{aligned}
\sum_{s\geq 0}(\Delta(s+1)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]} &= F\sum_{s\geq 0}\frac{1}{Ff(s)}\sqrt{f(s)^2(\Delta(s+1)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq F\sqrt{\sum_{s\geq 0}\frac{f(s)^2(\Delta(s+1)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]}{Ff(s)}} \text{ by concavity} \\
&= \sqrt{F\sum_{s\geq 0}f(s)(\Delta(s+1)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq \sqrt{TF\sum_{s\geq 0}f(s)(\Delta(s+1)+r_{\max})^2\nu^{\pi^0}(s)} \text{ using Lemma 4.5,}
\end{aligned}$$

so that overall, (4.29) becomes:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{(\Delta(s+1)+r_{\max})V_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{ATF}\sqrt{\sum_{s\geq 0}f(s)(\Delta(s+1)+r_{\max})^2\nu^{\pi^0}(s)}. \tag{4.30}$$

This is the term mainly contributing to the regret.

Bound on the Main Terms: Introducing E_2

Now, using Lemma 4.8 which gives the stationary distribution of ν^0, we can compute the expectation under ν^{π^0} of a well-chosen function f:

Lemma 4.6
Choose the increasing function $f : s \mapsto \frac{\max\{1,s(s-1)\}}{(\Delta(s+1)+r_{\max})^2}$. *Then* $F \leq 60e^{2\lambda/\mu}\, r_{\max}^2$ *and* $\sum_{s\geq 0}(\Delta(s+1)+r_{\max})^2 f(s)\nu^{\pi^0}(s) = \mathbb{E}^{\pi^0}\left[(\Delta+r_{\max})^2\cdot f\right] \leq \left(1+\frac{\lambda^2}{\mu^2}\right)$, *so that:*

$$E_2 := F\mathbb{E}^{\pi^0}\left[(\Delta+r_{\max})^2\cdot f\right] \leq 60e^{2\lambda/\mu}\, r_{\max}^2\left(1+\frac{\lambda^2}{\mu^2}\right).$$

Proof. For F, we obtain:

$$F \leq (2e^{\lambda/\mu} r_{\max})^2 3 \left(\sum_{s=0}^{S'} \frac{1 + \log(s+1)}{\max\{1, ss(s-1)\}} \right) \leq 12 e^{2\lambda/\mu} r_{\max}^2 \left(3 + \sum_{s=2}^{S'} \frac{1 + \log(s+1)}{s(s-1)} \right)$$
$$\leq 12 e^{2\lambda/\mu} r_{\max}^2 \left(4 + \sum_{s=2}^{S-2} \frac{\log(1+1/s)}{s} \right) \leq 60 e^{2\lambda/\mu} r_{\max}^2.$$

Using the following computations:

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

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

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

so that finally

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

which concludes the proof. □

Finally (4.30) becomes:

$$\mathbb{E}\left[\sum_{s,a} \sum_{k} \frac{(\Delta(s+1) + r_{\max}) V_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \right] \leq 3\sqrt{E_2 AT}, \tag{4.31}$$

and thus:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 12\sqrt{2E_2 AT \log(2AT)}. \tag{4.32}$$

In particular:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 132 e^{\lambda/\mu} r_{\max} \sqrt{\left(1 + \frac{\lambda^2}{\mu^2}\right) AT \log(2AT)}. \tag{4.33}$$

Bound on R_{ep}

It remains to deal with the following regret term:

$$R_{\text{ep}} = \sum_k V_k \left(P_k - I\right) \tilde{h}_k.$$

We will follow the core of the proof from [Jaksch et al., 2010]. Define $X_t := \left(p(\cdot|s_t, a_t) - e_{s_t}\right) \tilde{h}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}}$, where $k(t)$ is the episode containing step t and e_i the vector with i-th coordinate 1 and 0 for the other coordinates.

$$\begin{aligned} V_k \left(P_k - I\right) \tilde{h}_k &= \sum_{t=t_k}^{t_{k+1}-1} X_t + \tilde{h}_k(s_{t_{k+1}}) - \tilde{h}_k(s_{t_k}) \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + D r_{\max}. \end{aligned}$$

By summing over the episodes we get:

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

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

On the other side, using Lemma 2.10, we get when taking the expectation:

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

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

$$\mathbb{E}\left[R_{\text{ep}}\right] \leq 2 r_{\max} SAD \log(2AT). \tag{4.34}$$

We can now gather the expected regret terms when the true MDP is within the confidence bounds. Using (4.25), (4.32) and (4.34):

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

which gives with (4.19) and (4.20), assuming that $T \geq S^2$:

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

which finally gives:

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

4.5 Technical Lemmas

4.5.1 Probability of the Confidence Bounds

This first lemma is from [Jaksch et al., 2010, Lemma 17] and adapted to our confidence bounds.

Lemma 4.7
For $t > 1$, the probability that the MDP M is not within the set of plausible MDPs $\mathcal{M}_t$ is bounded by:

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

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

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

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

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

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

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

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

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

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

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

and

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

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

4.5.2 Diameter and Span of MDPs in $\mathcal{M}$

For completeness, and to support the discussion in Section 4.3.2, the section details the behavior of the diameter and the span of MDPs in $\mathcal{M}$, as functions of S.

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

Lemma 4.8
Under the stationary policy π^0, the stationary measure $\nu^{\pi^0}(s)$ of the MDP is given by:

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

This lemma is also presented in the proof of Lemma 3.6.

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

This discussion is formalized in the following result, which is a reminder of Lemma 2.3.

Lemma 4.9

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

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

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

where $\tau^\pi(x, y)$ is the expected time to go from x to y under policy π. Starting from S', we can write the hitting time equations:

$$\tau^{\pi^0}(S', S') = 1 + P^{\pi^0}_{S', S'-1}\tau^{\pi^0}(S'-1, S'),$$

and we notice that the left-hand side term actually is the inverse of the stationary measure at S', so that $\nu^{\pi^0}(S')^{-1} = \tau^{\pi^0}(S', S')$. We therefore obtain:

$$D^\pi \geq U\frac{\nu^{\pi^0}(S')^{-1} - 1}{\mu_{S'}} \geq \nu^{\pi^0}(S')^{-1} - 1.$$

Now using Lemma 4.8, we have that, for $s \in \mathcal{S}$:

$$\nu^{\pi^0}(s) = \frac{\binom{S'}{s}\left(\frac{\lambda}{S'\mu}\right)^s}{\left(1 + \frac{\lambda}{S'\mu}\right)^{S'}},$$

and assuming $\frac{\lambda}{S'\mu} \leq e - 1$, by concavity of the log function $\left(1 + \frac{\lambda}{S'\mu}\right)^{S'} \geq \exp\left(\frac{\lambda}{2\mu}\right)$, so that:

$$D^\pi \geq \left(\frac{S'\mu}{\lambda}\right)^{S'} e^{-\frac{\lambda}{2\mu}} - 1.$$

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

□

Let us now consider the bias of the optimal policy in M. From Chapter 3, the bias $h^*(s)$ is decreasing and concave in s, with increments bounded by C. Therefore, its span, defined as span $(h^*) := \max_s h^*(s) - \min_s h^*(s)$, satisfies

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

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

4.6 Generic Lemmas on Ergodic MDPs

4.6.1 Sensitivity of the Bias to the MDP Variations

The three first lemmas of this subsection are used in the proof of Lemma 4.13. This lemma is needed to obtain equation (4.24).

Lemma 4.10
For an MDP with rewards $r \in [0, r_{\max}]$ and transition matrix P, denote by $J_s(\pi, T) := \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, \pi(s_t))\right]$ the expected cumulative rewards until time T starting from state s, under policy π. Let D^π be the diameter under policy π. The following inequality holds: span $(J(\pi, T)) \leq r_{\max} D^\pi$.

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

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

□

Lemma 4.11
Consider two ergodic MDPs M and M'. Let $r, r' \in [0, r_{\max}]$ and P, P' respectively be the rewards and transition matrices of MDPs M and M' under policies π, where both MDPs have the same state and action spaces. Denote by g, g' the average reward obtained under policy π in the MDP M and M'. Then the difference of the gains is upper bounded.

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

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

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

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

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

Notice that, for any state s:

$$\begin{aligned}
\left((P - P') \sum_{t=0}^{T} P^t r \right)(s) &\leq \|p(\cdot|s) - p'(\cdot|s)\| \cdot \text{span}\,(J(T)) \\
&\leq r_{\max} D^{\pi} \|p(\cdot|s) - p'(\cdot|s)\| \text{ by Lemma 4.10} \\
&= b(s).
\end{aligned}$$

In the same manner we show that:

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

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

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

We can also show that:

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

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

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

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

Lemma 4.12

Let P be the stochastic matrix of an ergodic Markov chain with state space $\{1, \dots, S\}$. 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 $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i\ [\tau_S]$, where $\mathbb{E}_i\ [\tau_S]$ is the expected time to reach state S from state i.

Remark that this lemma is true for any state instead of S.

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

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

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

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

e the unit vector and e_S the vector with value 1 for the last state and 0 otherwise. Then $\tilde{A}$ 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 $i \neq S$, $(\mathbb{E}_i\ [\tau_S])_i = A_S^{-1} e$. Notice that A_S is an M-matrix, that is a matrix whose off-diagonal components are non-positive and that can be written $A_S = \kappa I - B$, where B is a matrix with positive components, and κ is larger than the eigenvalues of B . Its inverse therefore has non-negative components, and using the definition of the inifinite matrix norm: $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i\, \tau_S$. □

In the following lemma, we use the same notations as in Lemma 4.11 with a common state space $\{1, \dots S\}$.

Lemma 4.13
Let the biases h, h' be the biases of the two MDPs verify their respective Bellman equations with the renormalization choice $h(S) = h'(S) = 0$. Let $\sup_{s\in\mathcal{S}} \mathbb{E}_s\ [\tau^{\pi}_{s'}]$ be the worst expected hitting time to reach the state s' with policy π, and call $T_{hit} := \inf_{s'\in\mathcal{S}} \sup_{s\in\mathcal{S}} \mathbb{E}_s\ [\tau_{s'}]$. We have the following control of the difference:

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

Remark 4.14
The renormalization choice made in Lemma 4.13 does not matter when we use it to compute the bound on R_{diff} as defined in (4.23)*. We can indeed write the bias term in R_{diff} as $\tilde{h}_k - h_k + c_b e$, where c_b is a real constant and e the unit vector, and then notice that $\left(\tilde{P}_k - P_k\right) e = 0$, so that the renormalization choice does not appear anymore in the final computations.*

Proof. The computations in this proof follow the same idea as in the proof of [Ipsen and Meyer, 1994, 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 5.11, we can write the system of equations $\tilde{A}h = \tilde{r} - \tilde{g}$, with $\tilde{r}$ and $\tilde{g}$ respectively equal to r and g everywhere but on the last state, where their value is replaced by 0.

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

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

The previously defined block decompositions are:

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

For $s < S$, $dh(s) = -e_s^T A_S^{-1}(dA_S h' + d\tilde{r} - d\tilde{g})$ and $dh(S) = 0$. Now by taking the norm and using Lemma 4.10:

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

Notice that $\|dA_S\|_\infty \leq \|dP\|_\infty$, $\|d\tilde{r}\| \leq \|dr\|$ and $\|d\tilde{g}\| = |dg|$. Using Lemma 4.11 and Lemma 5.11, and taking the infimum for the choice of the state of renormalization implies the claimed inequality for the biases. □

4.6.2 A McDiarmid's Inequality

Lemma 4.15
Recall that $\nu^{\pi^{\max}}$ is the stationary measure of the Markov chain under policy $\pi^{\max}$, such that for every state s: $\pi^{\max}(s) = a_{\max}$.
Let k be an episode, and assume that the length of this episode I_k is at least $I(T) = 1 + \max\left\{Q_{\max}, T^{1/4}\right\}$, with $Q_{\max} := \left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^2 \log\left(\left(\frac{10D}{\nu^{\pi^{\max}}(S')}\right)^4\right)$.
Then, with probability at least $1 - \frac{1}{4T}$:

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

We will now prove Lemma 4.15:

Proof. Let k be an episode such that $I_k \geq I(T)$. We will first condition its length to be $I_k = I$, so that $t_{k+1} = t_k + I$. Denote by $\mathring{r}$ the vector of reward equal to 1 if the current state is x_k and 0 otherwise. Denote by $\mathring{g}_{\pi_k}$ the gain associated to the policy π_k for the transitions p and rewards $\mathring{r}$, and similarly define $\mathring{h}_{\pi_k}$ the bias, translated so that $\mathring{h}_{\pi_k}(S') = 0$. Notice in that case, that if we denote by ν^{π_k} the stationary distribution under policy π_k, that $\nu^{\pi^{\max}}(S') \leq \nu^{\pi_k}(S') \leq \nu^{\pi_k}(s)$ for any state s, for large enough $S' \geq \frac{\lambda}{\mu}$. Then:

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

By Azuma-Hoeffding inequality 2.12, following the same proof as in section 4.3.2 of [Jaksch et al., 2010], notice that $X_u = \mathring{h}_{\pi_k}(s_{u+1}) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k}\right\rangle$ form a martingale difference sequence with $|X_u| \leq D$:

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

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

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

On the other hand:

$$\sum_{u=t_k}^{t_k+I-1} \mathring{g}_{\pi_k} = V_k(s_k, a_k)\nu^{\pi_k}(x_k),$$

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

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

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

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

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

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

□

We can show a corollary of Lemma 4.15 that we will use for the regret computations:

Corollary 4.16

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

4.7 Conclusions

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

5 Reinforcement Learning in a Partially Observable Queueing Network: Optimal Admission

We have seen in Chapter 4 how we could improve the analysis of the upper bound of UCRL2 when restricting the class of MDPs to the MDPs that would modelize a birth-and-death process MDP. This restriction allowed us to use queueing properties of the bias, presented in Chapter 2, in order to remove the dependence on the diameter of the MDP, and instead introduce a quantity referring to the stability of the studied queue. We will now use the same properties on another classical example, an admission control problem with only partial information on the current state of the system. We will also need ergodicity properties of the queueing system to make correct use of its structural characteristics.

5.1 Introduction

We now know that in the generic MDP setting, a regret bound in $\tilde{O}(\sqrt{DSAT})$ has been reached [Tossou et al., 2019], and there have been many leads for results in MDPs with structure in the case of average rewards, as we have seen in Chapter 1 with [Fruit, Pirotta, and Lazaric, 2020; Bourel et al., 2020; Wu et al., 2022]. 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 [Jin, Kakade, et al., 2020; Azizzadenesheli et al., 2016; Z. D. Guo et al., 2016]. In this chapter, we address this question in the context of queueing networks where we assume that the learner has only access to the *total* number of jobs in the network, and this makes our problem fall in the family of Partially Observable MDPs (POMDPs).

5.1.1 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 [Papadimitriou and Tsitsiklis, 1987]. Furthermore, it is NP-hard to compute the optimal memoryless policy [N. Vlassis et al., 2012]. In reinforcement learning, where (some of) the model parameters are unknown, the lower bound on the average-case complexity developed in [Jin, Kakade, et al., 2020, 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 [Azizzadenesheli et al., 2016, Section 1] for a detailed discussion. In the attempt to reduce this computational burden, researchers focused on reinforcement learning in *subclasses* of POMDPs [Jin, Kakade, et al., 2020], and we will also follow this approach. The algorithm in [Even-Dar et al., 2005] assumes POMDPs without resets and has sample complexity scaling exponentially with a certain horizon time. The Bayesian algorithms proposed in [Ross et al., 2007; Poupart and N. A. Vlassis, 2008] learn POMDPs but bounds on the mean regret remain unknown for these approaches. A sample-efficient algorithm for episodic finite POMDPs is given in [Jin, Kakade, et al., 2020]. 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 [Azizzadenesheli et al., 2016]. 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, as seen in Chapter 2. Although the additional assumptions 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.

5.1.2 Contribution and Methodology

In this chapter, we will propose a learning algorithm for the optimal job-admission policy in a partially observable queueing network with regret $\mathrm{Reg}(T) \leq \tilde{O}(\sqrt{ST})$. Thus, our main contribution is a learning algorithm with a regret bound that does

not depend on the diameter D, and whose dependence on the state space is very small.

Optimal admission control is one of the most classical control problems in queues. It has been investigated in several works; see, e.g., [Borgs et al., 2014; Xia, 2014] and the references therein. However, these works consider the case where the model parameters are known, i.e., no learning mechanism is used. The novelty of our approach is to leverage i) Norton's equivalence theorem for closed product-form queueing networks [Chandy et al., 1975] and ii) the efficiency of reinforcement learning in MDPs with the structure of birth-and-death processes, as in Chapter 4. More specifically, our result is achieved by using 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. 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.

5.1.3 Organization

The remainder of the chapter is organized as follows. The model of the queueing network, its practical motivation and Norton's equivalent queue are presented in Section 5.2. Section 5.2.1 presents the problem addressed in detail, Section 5.4 is dedicated to the presentation of our learning algorithm (UCRL-M) and Section 5.5 to the analysis of its regret. In the latter, we state our main result in Theorem 5.5. Then, Section 5.6 discusses some technical aspects of our regret bound. Section 5.7 showcases the behaviour of the algorithm on a multi-tier queueing network. From Section 5.8 to Section 5.12, we first prove the main theorem and then the lemmas we get to use. Finally, Section 5.13 draws the conclusions of our work.

This chapter is based on the submitted work [Anselmi, Gaujal, and Rebuffi, 2023].

5.2 Admission Control in a Queueing Network

We consider a Jackson network with N queues (or stations) having service rates $\mu_1^o, \ldots, \mu_N^o$, a 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). To guarantee stability, i.e., positive recurrence of the underlying Markov chain, we require that $\lambda_i^o < \mu_i^o$, for all i, where λ_i^o is the arrival rate at queue i. It is given by the assumed unique solution of the traffic equations $\lambda_i^o = \lambda L_{0,i} + \sum_j \lambda_j^o L_{j,i}$, for all i.

We further assume that the total number of jobs in the network cannot exceed $S' := S - 1$. Under this constraint, the global system can be seen as a *closed* network with S' jobs. This network is identical to the original one except for an auxiliary queue, say queue 0, that represents the outside world with service rate λ. The departures of queue 0 correspond to the arrivals of the initial open network (see Figure 5.1).

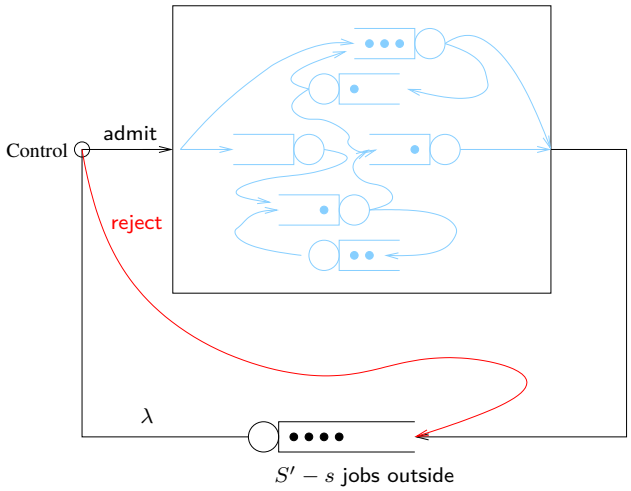


Figure 5.1: Admission control: rejected jobs immediately return to the outside queue.

Jobs that want to enter the network are subjected to admission control. If a job is rejected this can be modeled in the closed network as the job being sent back to the outside queue (see Figure 5.1). The goal of the admission controller is to minimize a 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: $\lim_{T\to\infty} \frac{1}{T-1} \int_{t=1}^{T} c_t$.

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 [Xia, 2014]. In the 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 [Borgs et al., 2014].

5.2.1 Problem Formulation

In this chapter, 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 [Borgs et al., 2014]);
- the controller takes decisions only relying on its set of observations up to time t.

5.2.2 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 [J.A. Rolia and Sevcik, 1995; Jerry Rolia et al., 2009], 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 [*Configuring Concurrency in Knative* 2022] 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 [Wang et al., 2022]. 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".

5.3 MDP Model

This section is dedicated to the construction of an MDP model of the system as well as an artificial aggregated MDP that is 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, non-aggregated. The aggregated system is only used for the performance analysis of the algorithm.

5.3.1 Original MDP

Let us model the problem as an MDP, $M^o = (\mathcal{X}^o, \mathcal{A}^o, P^o, r^o)$, where the super-index o stands for "original" throughout the chapter. We first use uniformization to see the process in discrete time. The uniformization constant U 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.
- 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, $\gamma_{\text{reject}}(1-a)\mathbb{1}_{\text{job-arrival}}$ (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

$$r^o(\mathbf{x}, a) := r_{\max} - \frac{\lambda\gamma_{\text{reject}}(1-a)+\gamma_{\text{hold}}s}{U} = \frac{\lambda\gamma_{\text{reject}}a+\gamma_{\text{hold}}(S'-s)}{U}, \quad (5.1)$$

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:

$$g^o(M^o, \pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^o(\mathbf{x}_t, \pi(\mathbf{x}_t))]. \quad (5.2)$$

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

5.3.2 Aggregated Model

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

Norton Equivalent Queue

In this subsection, let us consider the system defined in Section 5.2 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 [Chandy et al., 1975], also known in the literature as Flow Equivalent Server (FES) method [Krieger, 2008]. Towards this purpose, let v_i denote the average number of visits at queue i relative to queue 0. Set $v_0 = 1$ and let $(v_1, \dots, v_N)$ be the unique solution of $v_i = \sum_{j=0}^{N} L_{j,i} v_j$, for all $i = 1, \dots, N$. Then, when containing S' jobs, the vector of the number of jobs in each queue forms a continuous-time Markov chain with stationary measure [Krieger, 2008]

$$\nu^o(\mathbf{x}) = \frac{1}{\mathcal{G}(S')} \prod_{i=0}^{N} \left(\frac{v_i}{\mu_i}\right)^{x_i}, \tag{5.3}$$

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

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*[1] of the network with s jobs in total (see Figure 5.2 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,

$$\sum_{\mathbf{x}:|\mathbf{x}|=s} \nu^o(\mathbf{x}) = \nu(S' - s, s), \quad \forall s = 0, \dots, S' \tag{5.4}$$

   where $\nu(S'-s, s)$, for all $s = 0, \dots, S'$, 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 $\mu(s)$ [Kameda, 1984].

[1]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 (5.3) of the stationary measure m^o.

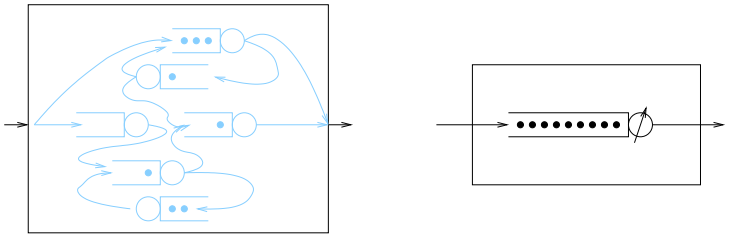


Figure 5.2: Illustration of Norton Equivalence theorem.

> **Lemma 5.1**
> *The throughput function $s \mapsto \mu(s)$ is increasing, concave and bounded by $\mu_{\max} := \sum_{i \leq N} \mu_i^o$.*

The throughput bound $\mu_{\max}$ can be significantly improved [Krieger, 2008] but this will not change the structure of our results.

Aggregated 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 $\mathcal{S} = \{0, \ldots, S'\}$ consists of all possible numbers of jobs in the queueing network. We denote by $S := S' + 1$ the number of states of the aggregated 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 (5.1) 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

$$r(s,a) = \frac{\lambda \gamma_{\text{reject}} a + \gamma_{\text{hold}}(S' - s)}{U} \tag{5.5}$$

- 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 $\mathbf{x} \in \mathcal{X}^o$, then $\pi(\mathbf{x}) := \pi(|\mathbf{x}|)$. We can now define the transition matrix for policy π as the transition matrix in the aggregated MDP M under the stationary measure $\nu^{o,\pi}$:

$$P(s' \mid s, \pi(s)) = \sum_{\mathbf{x},|\mathbf{x}|=s} \sum_{\mathbf{y},|\mathbf{y}|=s'} \frac{\nu^{o,\pi}(\mathbf{x})}{\nu^{\pi}(s)} P^o(\mathbf{y} \mid \mathbf{x}, \pi(\mathbf{x})), \tag{5.6}$$

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.

- Regarding the diameter as defined in Chapter 2, we will only consider the diameter on the aggregated MDP for the computations of the regret bound, as it is needed to control the bias terms of the aggregated MDP (see Section 5.12). We will never need to consider the bias or the diameter of the original MDP.

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

Definition 5.2

The average gain induced by policy π is:

$$g(M, \pi) := \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(s_t, \pi(s_t))]. \tag{5.7}$$

The optimal policy π^* achieves

$$g(M, \pi^*) := g^*(M) := \sup_{\pi \in \Pi} g(M, \pi). \tag{5.8}$$

5.3.3 Comparison Between both MDPs

It should be clear that the original MDP M^o has a greater set of policies than the aggregated 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

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

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

5.3.4 Reinforcement Learning

Here, we consider a learner that can observe the arrivals and departures of jobs in the original MDP and makes admission decisions for each arriving job.

What Does the Learner Know?

- As mentioned earlier, the learner can observe the external events: arrivals and departures of jobs. This implies that at any discrete time-step 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 expected cost in state-action pair (s,a) is unknown as it depends on the unknown parameter λ (see (5.5)). However, the parameters γ_{reject}, γ_{hold} and the uniformization constant U are known, and the learner knows how the cost depends on λ, which will be important for the definition of the confidence regions. We will often use an upper bound on the difference of rewards between two neighboring states $\delta_{\max} := \gamma_{\text{reject}} + \frac{\gamma_{\text{hold}}}{U}$. We will use $\delta_{\max}$ instead of $r_{\max}$ 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. We will also make

the assumption that the learner knows the reward function up to the actual value of the arrival rate λ, that must be learned.

- The learning algorithm knows T, the number of time steps where it can take observations and actions. This is not a strong requirement as one can make the algorithm oblivious to T by using a classical doubling trick on T.

Regret

We recall the definition of the regret, as in Chapter 2:

Definition 5.3 (Regret)
The regret *at time T of the learning algorithm $\mathcal{L}$ is*

$$Reg(M, \mathcal{L}, T) := Tg^*(M) - \sum_{t=1}^{T} r_t. \tag{5.9}$$

Here, $g^*(M)$ is the optimal gain defined in (5.8). The reward r_t is the reward of the state visited at time t under the current policy used by the learning algorithm.

5.4 Learning Algorithm

5.4.1 High-Level Description of the Proposed Algorithm

Our algorithm is *episodic*, *model-based* and *optimistic*. More precisely, the interactions of the learner with the MDP M^o are decomposed into *episodes*. In each episode k, of duration $[t_k, t_{k+1} - 1]$, 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*).

In our case with partial observations, the number of jobs at time t, $(s_t)_{t \leq T}$ is not Markovian, therefore it does not provide enough information to make good estimates on the underlying MDP. Instead, we collect a set $\{s_1, \ldots, s_{\tau_{\text{mix}}}\}$ 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 $s_i, s_{i+\tau_{\text{mix}}}, s_{i+2\tau_{\text{mix}}}, \dots$ 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. Each learning module $\mathcal{L}_i$ behaves similarly as the classical 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 5.4.4. The main technical difficulties in the control of the behavior of the algorithm are:

1. The observations used by 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 sequence of observations $s_i, s_{i+\tau_{\text{mix}}}, s_{i+2\tau_{\text{mix}}}, \dots$ is not really stationary and independent, but only weakly correlated.

5.4.2 Number of Modules: τ_{mix}

Let us first give a more precise definition of the τ_{mix} modules, where the 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 $1 \le t \le \tau_{\text{mix}}$, 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:

$$\max_{\pi\in\Pi} \sup_{\mathbf{x}_1\in\mathcal{S}} \left\| \mathbb{P}^{o,\pi}_{\mathbf{x}_1}(\mathbf{x}_t = \cdot) - \nu^{o,\pi} \right\|_{TV} \le C\rho^t \quad \forall t > 0, \tag{5.10}$$

where $\mathbb{P}^{o,\pi}_{\mathbf{x}_1}(\mathbf{x}_t = \cdot)$ is the distribution of the state at time t under policy π in the original MDP, with initial state $\mathbf{x}_1$. Let us then define

$$\tau_{\text{mix}} := \lceil 5 \log T / \log \rho^{-1} \rceil. \tag{5.11}$$

The reason for this precise choice will appear in the analysis of the regret (see Section 5.5) but the general idea behind this choice comes from Lemma 5.7 given in Section 5.10, 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 5.6.

5.4.3 UCRL-M: Learning with τ_{mix} Modules

Algorithm UCRL-M (Upper Confidence Reinforcement Learning with several Modules) is given in Algorithm 3 below. 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 (5.15) and (5.14). Then, it applies Extended Value Iteration (EVI) (Section 5.9) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ according to (5.12). 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 from the stationary distribution of the current policy. This phase is necessary to guarantee that observations within a module are nearly independent. Afterward, 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 (5.18) 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 (5.17)).

5.4.4 Confidence Region

As mentioned earlier, the 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).

Algorithm 3: The UCRL-M algorithm.

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

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

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

  Initialize episode k with $t_k := t$

  Compute for all (s, a) the modules $m_k(s, a)$ according to (5.13);

  Compute the confidence region $\mathcal{M}_k$ as in (5.17);

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

$$g(\tilde{M}_k, \tilde{\pi}_k) \geq \max_{M_k \in \mathcal{M}_k} \max_{\pi} g(M_k, \pi) - \frac{\delta_{\max}}{\sqrt{t_k}}. \tag{5.12}$$

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

  Exploration: while $V_k^{(m_t)}(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_{t_k}^{(m_t)}(s_t, \tilde{\pi}_k(s_t))\}$, **do**

    1. Choose action $a_t = \tilde{\pi}_k(s_t)$;
    2. Observe s_{t+1};
    3. Update $V_k^{(m_t)}(s_t, a_t) := V_k^{(m_t)}(s_t, a_t) + 1$;
    4. Set $t := t + 1$.

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,

$$m_k(s, a) \in \arg\max_m N_{t_k}^{(m)}(s, a), \tag{5.13}$$

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:

$$\hat{p}_k^{(m)}(s'|s, a) := \frac{\sum_{t=1}^{t_k - 1} \mathbb{1}_{\{s_t = s, a_t = a, s_{t+1} = s', m_t = m\}} \mathbb{1}_{\{t \notin \Phi\}}}{\max\left\{1, N_{t_k}^{(m)}(s, a)\right\}} \tag{5.14}$$

$$\hat{r}_k^{(m)}(s, a) := \hat{p}_k^{(m)}(s + 1|s, a)\gamma_{\text{reject}} + \gamma_{\text{hold}} \frac{S' - s}{U}, \tag{5.15}$$

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:

$$\forall(s,a), \quad \left|\tilde{r}(s,a) - \hat{r}_k^{(m_k(s,a))}(s,a)\right| \leq \delta_{\max}\sqrt{\frac{2\log(2At_k)}{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}}; \tag{5.16}$$

$$\forall(s,a), \quad \|\tilde{p}(\cdot \mid s,a) - \hat{p}_k^{(m_k(s,a))}(\cdot \mid s,a)\|_1 \leq \sqrt{\frac{8\log(2At_k)}{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{5.17}$$

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 (5.13), and we do not build a specific confidence region for each module.

The algorithm finds the best optimistic MDP and policy within this confidence set, 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:

$$\forall(s,m) \quad V_k^{(m)}(s,\tilde{\pi}_k(s)) \leq \max\{1, N_{t_k}^{(m)}(s,\tilde{\pi}_k(s))\}. \tag{5.18}$$

5.4.5 Time Complexity of UCRL-M

Proposition 5.4

The time complexity of UCRL-M *is* $O(KS\tau_{mix} + Kt_{evi} + T)$*, 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 4 and 5 is $O(KS\tau_{\text{mix}})$. The complexity of line 6 is $O(Kt_{\text{evi}})$. The complexity of line 7 is $O(K\tau_{\text{mix}})$. The complexity of line 8 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 [Jaksch et al., 2010] 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.

5.5 Regret of UCRL-M

5.5.1 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 aggregated MDP. Also, $\nu^{\pi^{\max}}(s)$ is the stationary measure in the aggregated MDP under the policy that accepts all jobs, ρ is defined in Section 5.4 and $\mu(i)$ is the service rate in the aggregated 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 $\mu(i_0) > \lambda$. Such a i_0 exists because the unconstrained network is assumed to be stable (see Section 5.2) regardless of S. 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 5.5
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 $\kappa = 228(\gamma_{reject} + \frac{\gamma_{hold}}{U})\frac{U}{\mu(1)}C_1\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^{-3}$. For the choice $\tau_{mix} = 5\frac{\log T}{\log 1/\rho}$, and $A = 2$, assuming $\tau_{mix}S \geq 2$ and $T > \max\{\frac{e^2}{4T}, \tau_{mix}\}$, we have:

$$\mathbb{E}\left[Reg(M, \text{UCRL-M}, T)\right] \leq \kappa \log(2T)\sqrt{T\log^{-1}(1/\rho)} + R_{LO}, \tag{5.19}$$

where $R_{LO} := 138 r_{\max} D^2 \max\left\{Q_{\max}, T^{1/4}\right\}\frac{\log^4(4T)}{\log^2 1/\rho}$ 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 the first order term of 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 5.6.

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.

5.5.2 Outline of the Proof

To compute the expected regret $\mathbb{E}[\text{Reg}]$, we will mainly follow the strategy from [Jaksch et al., 2010, Section 4]. First, we deal with the regret term corresponding to the initialization phase of each episode, which depends on 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 5.7 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: $R_k^{(m)}(s) := \sum_a V_k^{(m)}(s,a)(g^* - r(s,a))$ 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:

$$\mathbb{E}\left[\text{Reg}\right] \leq \mathbb{E}\left[R_{\text{in}}\right] + \mathbb{E}\left[R_{\text{out}}\right] + \mathbb{E}\left[R_{\text{ramp}}\right] \tag{5.20}$$

with K the number of episodes and the regret where the MDP is in the confidence region being $R_{\text{in}} := \sum_m \sum_s \sum_{k=1}^{K} R_k^{(m)}(s) \mathbb{1}_{M \in \mathcal{M}_k}$, and when it is outside $R_{\text{out}} := \sum_m \sum_s \sum_{k=1}^{K} R_k^{(m)}(s) \mathbb{1}_{M \notin \mathcal{M}_k}$ and the regret of the ramping phases $R_{\text{ramp}} = \sum_k \sum_{t=t_k}^{t_k+\tau_{\text{mix}}-1} r(s_t, \tilde{\pi}_k(s,t))$. Each term is then bounded as explained in Section 5.8.

5.6 Controlling the Regret Bound Parameter ρ

The efficiency of UCRL-M is critically based on controlling τ_{mix} and ρ. In particular, Theorem 5.5 says that the regret of UCRL-M depends on $W := \log^{-1/2}(1/\rho)$.

5.6.1 Bounds Using Mixing and Coupling Times

In Section 5.4, the number of modules τ_{mix} is defined as $\tau_{\text{mix}} := 5 \log T / \log \rho^{-1}$. where ρ is such that

$$\max_{\pi} \sup_{\mathbf{x}_0 \in \mathcal{S}} \left\| \mathbb{P}^{o,\pi}_{\mathbf{x}_1}(\mathbf{x}_t = \cdot) - \nu^o(\pi) \right\|_{TV} \leq C\rho^t \quad \forall t > 0. \tag{5.21}$$

Let us first recall classical results from Markov chain theory [Levin et al., 2008] 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 (5.21)). Let us define $d(t) := \sup_{\mathbf{x}_1 \in \mathcal{S}} \|\mathbb{P}_{\mathbf{x}_1}(\mathbf{x}_t = \cdot) - \nu\|_{TV}$. 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:

$$\rho \leq \frac{1}{2^{t_{mix}^{-1}}} \tag{5.22}$$

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 $\tau_{\mathbf{x},\mathbf{y}} := \min\{t : X_t = Y_t\}$. If X_t and Y_t are coupled and start at $X_1 = \mathbf{x}$ and $Y_1 = \mathbf{y}$ respectively. Then, $d(t) \leq \max_{\mathbf{x},\mathbf{y}} \mathbb{P}(\tau_{\mathbf{x},\mathbf{y}} > t)$. By using Markov inequality, this implies that

$$t_{mix} \leq 4 \max_{\mathbf{x},\mathbf{y}} \mathbb{E}[\tau_{\mathbf{x},\mathbf{y}}]. \tag{5.23}$$

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

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 [Dopper et al., 2006, Theorem 5.3], where the following result is proved in the stable case. Using our notation,

$$\max_{\mathbf{x},\mathbf{y}} \mathbb{E}[\tau_{\mathbf{x},\mathbf{y}}] \leq \sum_{i=1}^{N} \frac{U^2}{(\lambda_i^o + \mu_i^o)(\mu_i^o - \lambda_i^o)} S', \tag{5.24}$$

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

According to Equation (5.22) and (5.23), this induces the following bound on the term W in the regret:

$$W \leq \kappa_0 \sqrt{N S'},$$

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

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 [Anselmi and Gaujal, 2014], where the following bound is given (Theorem 2):

$$\max_{\mathbf{x},\mathbf{y}} \mathbb{E}[\tau_{\mathbf{x},\mathbf{y}}] \leq \kappa_2 N^2 S' \sum_{i=1}^{N} \frac{\lambda_i^o}{\mu_i^o - \lambda_i^o}, \tag{5.25}$$

where κ_2 is a constant. Using Equation (5.22) and (5.23), 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.

5.6.2 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 [Azar et al., 2017] 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 5.5.

5.7 Numerical Experiments

5.7.1 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 5.3. 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 [Urgaonkar et al., 2005] and is the default architecture of web applications deployed on Amazon Elastic Compute Cloud (EC2) [*AWS Architecture Center* 2022].

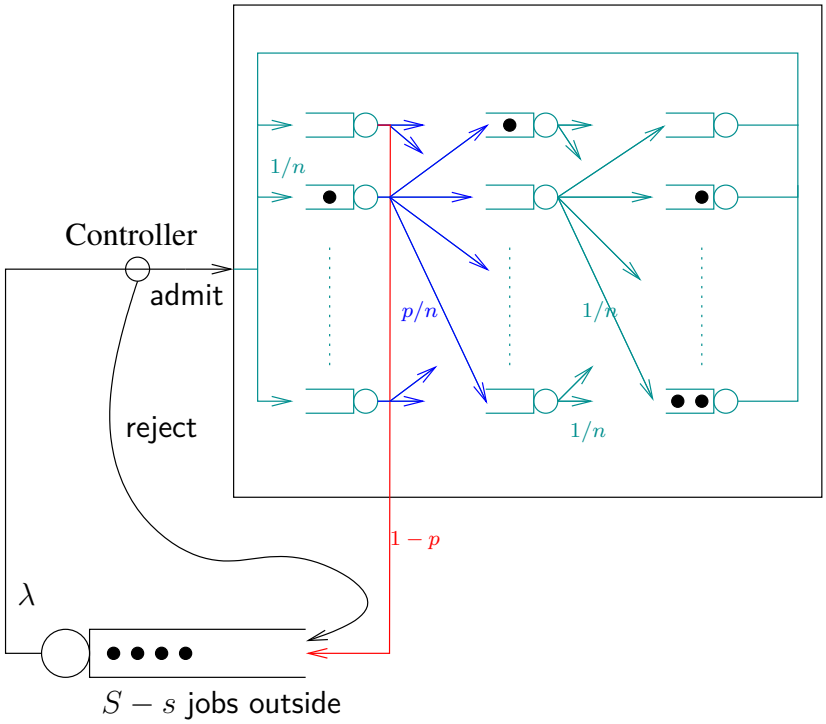


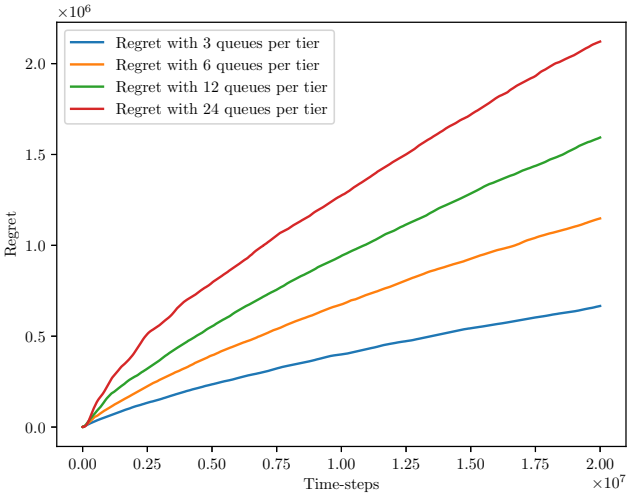
Figure 5.3: A queueing network model with three interconnected tiers. Each tier contains n queues and the total capacity is of S' jobs.

This model may be studied as an example of the generic case described in Section 5.2. Notice that given the routing from Figure 5.3, 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.

5.7.2 Regret of UCRL-M on the Multi-Tier Queueing Network

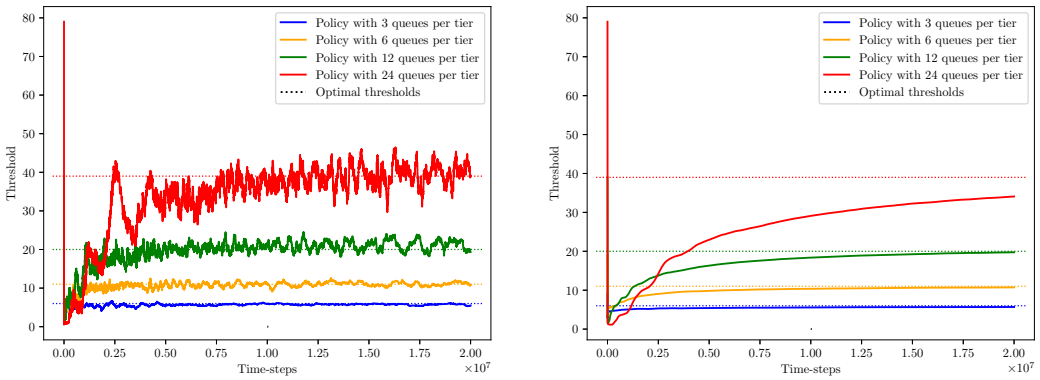
We provide the performance of UCRL-M over the queueing network described above when the number of queues per tier n and the total number of jobs S' vary. In Figure 5.4, we display the average regret over 66 runs of the UCRL-M algorithm when n varies, and with parameters scaling with n to keep the systems 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 5.6 and even in Section 5.2.

In Figure 5.4, we remark that as we let the number of queues n (and the number of jobs S') scale multiplicatively, the regret is increasing in $\log(S')$. Knowing that



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 5.4: Regret of the UCRL-M algorithm on the queueing network for different values



(a) Average threshold over time.

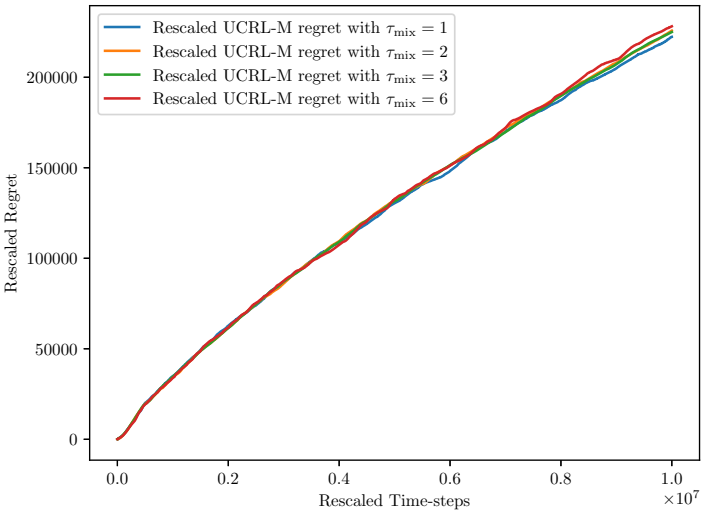
(b) Cesàro sum of the average threshold.

Figure 5.5: Comparison between the average threshold and its Cesàro sum.

the dependency in S' of the regret bound from Theorem 5.5 mainly comes from ρ, this is much slower than the square root bounds given in Section 5.6 (under strong assumptions). This can be interpreted as the bound of equation 5.21 being too large as it considers the mixing from the worst state, while in 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 5.5 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 5.4.1, UCRL-M could be compared to τ_{mix} instances of UCRL2 [Jaksch et al., 2010], 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 5.6 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'	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 5.6: 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 5.5.

5.8 Proof of Theorem 5.5

From this point onwards, we give the detailed proof of Theorem 5.5 and a term-by-term analysis of the regret.

5.8.1 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:

$$R_{\text{ramp}} = \sum_k \sum_{t=t_k}^{t_k+\tau_{\text{mix}}-1} r(s_t, \tilde{\pi}_k(s,t)) \leq K_T \tau_{\text{mix}} r_{\max} \leq r_{\max} SA\tau_{\text{mix}}^2 \log_2\left(\frac{8T}{SA\tau_{\text{mix}}}\right), \tag{5.26}$$

where in the last inequality we used Lemma 2.10. Assuming $\tau_{\text{mix}} SA \geq 4$, and using $\log(2) \geq \frac{1}{2}$, we rewrite it:

$$R_{\text{ramp}} \leq 2 r_{\max} SA\tau_{\text{mix}}^2 \log(2T). \tag{5.27}$$

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

5.8.2 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 of state s during episode k in module m.
- $N_t^{(m)}(s)$ is the number of visits of 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 5.8 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 5.8, we obtain

$$
\mathbb{E}[R_{\text{out}}] \leq r_{\max} \sum_{t=1}^{T} t \mathbb{P}\{M \notin \mathcal{M}(t)\} \leq r_{\max} \sum_{t=1}^{T} \frac{S + 16CSA}{2t^2} \leq r_{\max}(S + 16CSA). \tag{5.28}
$$

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

5.8.3 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 for simplicity

- $R_{\text{in},k}^{(m)} := \sum_s R_k^{(m)}$,
- $\tilde{\pi}_k$ 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 $\tilde{M}_k$,
- $V_k^{(m)} := \left(V_k^{(m)}(s, \tilde{\pi}_k\right)$ the row vector of visit counts,
- 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 5.6 on the accuracy of EVI, we get:

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

In the next few steps, we will focus on rewriting the first sum. With (5.43) 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:

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

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

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

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

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

$$V_k^{(m)} \left(\tilde{P}_k - I \right) u_i = V_k^{(m)} \left(\tilde{P}_k - I \right) \tilde{h}_k + V_k^{(m)} \left(\tilde{P}_k - I \right) d_k.$$

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

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

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

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

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

$$R_{\text{EVI}}^{(m)} \leq \delta_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k^{(m)}(s,a)}{\sqrt{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}}. \tag{5.29}$$

Let us now deal with the term $R_{\text{rewards}}^{(m)}$, as it will be bounded by a similar term as in equation (5.29). Indeed, as $M \in \mathcal{M}_k$, we may use that both the optimistic and true rewards are within the confidence region from equation 5.16, and use that $t_k < T$, so that:

$$R_{\text{rewards}}^{(m)} \leq \delta_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}} \tag{5.30}$$

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

$$\begin{aligned} R_{\text{in}}^{(m)} \leq & \underbrace{\sum_k V_k^{(m)}\left(\tilde{P}_k - P_k\right) h_k}_{R_{\text{trans}}^{(m)}} + \underbrace{\sum_k V_k^{(m)}\left(\tilde{P}_k - P_k\right)\left(\tilde{h}_k - h_k\right)}_{R_{\text{diff}}^{(m)}} + \underbrace{\sum_k V_k^{(m)}(P_k - I)\tilde{h}_k}_{R_{\text{ep}}^{(m)}} \\ & \underbrace{+ \delta_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k^{(m)}(s,a)}{\sqrt{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}}}_{R_{\text{EVI}}^{(m)} + R_{\text{rewards}}^{(m)}}. \end{aligned} \tag{5.31}$$

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:

$$R_{\text{EVI}} + R_{\text{rewards}} = \delta_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{V_k(s,a)}{\sqrt{\max\left\{1, N_{t_k}^{(m_k(s,a))}(s,a)\right\}}}. \tag{5.32}$$

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.

5.8.4 Bound on $R_{\text{trans}}^{(m)}$

To bound $R_{\text{trans}}^{(m)}$, we can follow the computations from Chapter 4. We will use our knowledge of the bias h_k and the control on the transitions in the optimistic MDP to simplify the regret term.

Notice that for a fixed state $0 \leq s \leq S'$:

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

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

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

where Δ is the difference of bias in the last inequality, we used the bound on the variations of the bias from Proposition 5.16, and that the optimistic MDP has transitions close to the true transitions with inequality (5.17). Notice that the final term looks similar to the term coming from EVI and rewards related computations

(5.32). We will deal with these terms together in the next subsection, as they are both mainly contributing to the regret.

5.8.5 Bound on the Main Term

In the previous Section 5.8.4, we have shown that:

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

Summing over the modules m, we get:

$$R_{\text{trans}} \leq 4\sqrt{2\log(2AT)}\sum_{s,a}\frac{\Delta(s+1)V_k(s,a)}{\sqrt{\max\{1, N_{t_k}^{(m_k(s,a))}(s,a)\}}}. \tag{5.33}$$

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 2.11, 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 (5.13), we get:

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

Therefore:

$$R_{\text{trans}} \leq 12\sqrt{2A\tau_{\text{mix}}\log(2AT)}\sum_{s=0}^{S'}\Delta(s+1)\sqrt{\mathbb{E}\left[N_T(s)\right]}. \tag{5.34}$$

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

$$R_{\text{EVI}} + R_{\text{rewards}} \leq 12\delta_{\max}\sqrt{2A\tau_{\text{mix}}\log(2AT)}\sum_{s\geq 0}\sqrt{\mathbb{E}\left[N_T(s)\right]} \tag{5.35}$$

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:

$$12\sqrt{2A\tau_{\text{mix}}T\log\left(2A\tau_{\text{mix}}T\right)}\sum_{s=0}^{S'}\left(\Delta(s+1)+\delta_{\max}\right)\sqrt{\nu^{\pi^{\max}}(s)}. \tag{5.36}$$

Replace in the equation the choice $\tau_{\text{mix}} = 5\log T/\log\rho^{-1}$ and recall that we had, from Proposition 5.16, $\Delta(s) := 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\sum_{i=1}^{s}\frac{U}{\mu(i)} \leq 2\delta_{\max}\nu^{\pi^{\max}}(0)^{-1}\frac{U}{\mu(1)}s$.

Using Lemma 5.15, since

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

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

$$72\delta_{\max}\frac{U}{\mu(1)}C_1\left(1-\sqrt{\frac{\lambda}{\mu(i_0)}}\right)^{-3}\log\left(AT\right)\sqrt{5AT\log^{-1}(\rho^{-1})}. \tag{5.37}$$

5.8.6 Bound on $R_{\text{diff}}^{(m)}$

We now deal with the term involving the difference of bias $R_{\text{diff}}^{(m)}$, defined in equation 5.31. The proof mainly follows the one from Chapter 4, 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 $\tilde{h}_k$ and h_k, so that we will need Lemma 5.12, and we are interested in controlling $\|\tilde{h}_k - h_k\|_\infty$.

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 5.12. 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 5.12, as $\text{span}(h^{\tilde{\pi}_k}) \leq D$ by construction of $\tilde{\pi}_k$ with EVI, following the same argument as in [Jaksch et al., 2010, Equation (11)].

$$\|\tilde{h}_k - h_k\|_\infty \leq 8r_{\max}D^2\sqrt{\frac{2\log(2At_k)}{\max\{1, N_{t_k}^{(m')}(x_k, a_k)\}}}. \tag{5.38}$$

Hence,

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

where in the last inequality we have used (5.38) 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 $Q_{\max} := \left(\frac{10C_2S^2}{\nu^{\pi^{\max}}(S')}\right)^2 \log\left(\left(\frac{10C_2S^2}{\nu^{\pi^{\max}}(S')}\right)^4\right)$ 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 (5.13) 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 Chapter 4. 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:

$$\begin{aligned}
\Sigma &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \sum_{k, I_k > I(T)} \frac{2\tau_{\text{mix}}}{\nu^{\pi^{\max}}(S')} + \mathbb{1}_{\bar{\mathcal{E}}} \sum_{k, I_k > I(T)} \tau_{\text{mix}} I_k \\
&\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \left(K - K_{\leq I}\right) \frac{2\tau_{\text{mix}}}{\nu^{\pi^{\max}}(S')} + \mathbb{1}_{\bar{\mathcal{E}}} \tau_{\text{mix}} T.
\end{aligned}$$

We use Corollary 5.14 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 2.10, $SA \geq 4$, $I(T) \geq \frac{2}{\nu^{\pi^{\max}}(S')}$ and that $\frac{1}{\log 2} + \frac{1}{4} \leq 2$:

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

Therefore, we have that:

$$\mathbb{E}\left[R_{\text{diff}}^{(m)}\right] \leq 64 r_{\max} SAD^2 \tau_{\text{mix}} I(T) \log^2\left(2AT\right). \tag{5.39}$$

5.8.7 Bound on R_{ep}

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

$$R_{\text{ep}}^{(m)} = \sum_k V_k^{(m)} \left(P_k - I\right) \tilde{h}_k.$$

We first want to sum over the modules to get the same kind of term as in [Jaksch et al., 2010], written as a martingale difference sequence, and then take the expectation. Following that proof, we define $X_t := \left(p(\cdot|s_t, a_t) - e_{s_t}\right) \tilde{h}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}}$, where $k(t)$ is

the episode containing step t and e_i the vector with i-th coordinate 1 and 0 for the other coordinates. We obtain

$$\sum_m V_k^{(m)} (P_k - I) \tilde{h}_k \leq V_k (P_k - I) \tilde{h}_k \leq \sum_{t=t_k}^{t_{k+1}-1} X_t + \tilde{h}_k(s_{t_{k+1}}) - \tilde{h}_k(s_{t_k})$$
$$\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + Dr_{\max},$$

and by summing over the episodes we get

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

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 2.10 on the number of episodes, when taking the expectation we obtain

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

As for the computation of (5.27), assuming $\tau_{\text{mix}} SA \geq 4$:

$$\mathbb{E}[R_{\text{ep}}] \leq 2r_{\max} SAD\tau_{\text{mix}} \log(2AT). \tag{5.40}$$

5.8.8 Total Sum

We remind that we showed in subsection 5.8.5 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 it remains now to compute the lower order term of the regret R_{LO}. Using (5.27), (5.28), (5.39) and (5.40), the lower order term of the regret is upper bounded by, omitting the $r_{\max}$ factor:

$$64SAD^2\tau_{\text{mix}} I(T) \log^2(2AT) + 2SA\tau_{\text{mix}}(D + \tau_{\text{mix}}) \log(2AT) + (S + 16CSA),$$

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

$$r_{\max} 69 SAD^2 \tau_{\text{mix}}^2 I(T) \log^2(2AT),$$

which concludes the proof of Theorem 5.5.

5.9 Lemmas on Extended Value Iteration

We remind the fundamental properties of the Extended Value Iteration (EVI) algorithm, first described in [Jaksch et al., 2010], 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 5.8.3. EVI iteratively computes values in the following way:

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

where $\mathcal{P}(s,a)$ is the set of probabilities from (5.17), and the iterations are stopped with respect to the following lemma [Jaksch et al., 2010, Theorem 7].

Lemma 5.6

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:

$$\max_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} - \min_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} < \varepsilon_k. \tag{5.41}$$

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:

$$\tilde{g}_k := \min_s g(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} g(M', \pi, s') - \varepsilon_k. \tag{5.42}$$

Moreover, from [Martin L. Puterman, 1994, Theorem 8.5.6]:

$$\left| u_{i+1}^{(k)}(s) - u_i^{(k)}(s) - \tilde{g}_k \right| \leq \varepsilon_k, \tag{5.43}$$

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:

$$\tilde{h}_k(s_0) = \mathbb{E}_{s_0}\left[\sum_{t=0}^{\infty}(\tilde{r}(s_t, a_t) - \tilde{g}_k)\right]. \tag{5.44}$$

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 [Martin L. Puterman, 1994, Equation 8.2.5], we can ensure that:

$$\left|u_i^{(k)}(s) - (i-1)\tilde{g}_k - \tilde{h}_k(s)\right| < \frac{\varepsilon_k}{2}, \tag{5.45}$$

so that we can define the following difference

$$d_k(s) := \left|u_i^{(k)}(s) - \min_{s'} u_i^{(k)}(s') - \left(\tilde{h}_k(s) - \min_{s'}\tilde{h}_k(s')\right)\right| < \varepsilon_k. \tag{5.46}$$

5.10 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 5.8.2 when we consider the episodes k with $M \notin \mathcal{M}_k$.

Let us first prove the key Lemma 5.7.

Lemma 5.7

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:

$$\max_{\pi\in\Pi}\sup_{\mathbf{x}_0\in\mathcal{S}} \|\mathbb{P}^\pi{}_{\mathbf{x}_0}(\mathbf{x}_t = \cdot) - \nu^\pi\|_{TV} \leq C\rho^t \quad \forall t > 0. \tag{5.47}$$

Let $t, t' > 0$ such that $t' - \tau_{mix} \geq t$, with t' and $t' - \tau_{mix}$ 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:

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

Proof. The proof is essentially the same as in [Bhandari et al., 2018, 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 aggregated 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 distributions $P := \mathbb{P}\{X \in \cdot, Y \in \cdot\}$ and $Q := \mathbb{P}\{X \in \cdot\} \otimes \mathbb{P}\{Y \in \cdot\}$, and we define the total variation information

$$I_{TV}(X, Y) := \sum_{\mathbf{x}} \mathbb{P}\{X = \mathbf{x}\} \left\| \mathbb{P}\{Y = \cdot \mid X = \mathbf{x}\} - \mathbb{P}\{Y = \mathbf{y}\} \right\|_{TV}.$$

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 [Bhandari et al., 2018], we obtain

$$\begin{aligned} \|P - Q\|_{TV} &\leq I_{TV}(X, Y) \leq I_{TV}(\mathbf{x}_t, \mathbf{x}_{t'}) \leq I_{TV}(\mathbf{x}_{t'-\tau_{\text{mix}}}, \mathbf{x}_{t'}) \\ &\leq \sum_{\mathbf{x}} \mathbb{P}\{\mathbf{x}_{t'-\tau_{\text{mix}}} = \mathbf{x}\} \left\| \mathbb{P}\{\mathbf{x}_{t'} = \cdot \mid \mathbf{x}_{t'-\tau_{\text{mix}}} = \mathbf{x}\} - \mathbb{P}\{\mathbf{x}_t = \cdot\} \right\|_{TV} \end{aligned}$$

then using a triangle inequality:

$$\begin{aligned} \left\| \mathbb{P}\{\mathbf{x}_{t'} = \cdot \mid \mathbf{x}_{t'-\tau_{\text{mix}}} = \mathbf{x}\} - \mathbb{P}\{\mathbf{x}_t = \cdot\} \right\|_{TV} \leq & \left\| \mathbb{P}\{\mathbf{x}_{t'} = \cdot\} - \nu^{\tilde{\pi}_k} \right\|_{TV} + \\ & \left\| \mathbb{P}\{\mathbf{x}_{t'} = \cdot \mid \mathbf{x}_{t'-\tau_{\text{mix}}} = \mathbf{x}\} - \nu^{\tilde{\pi}_k} \right\|_{TV}, \end{aligned}$$

we get

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

where in the last inequality we used assumption (5.10) 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 aggregated 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 5.8
For $t > 1$, the probability that the MDP M is not within the set of plausible MDPs $\mathcal{M}(t)$ is bounded by:

$$\mathbb{P}\{M \notin \mathcal{M}(t)\} \leq \frac{S}{2t^3} + \frac{8CSA}{t^3}.$$

Compared to [Jaksch et al., 2010, 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 5.7 to consider independent events again, and to be able to use concentration inequalities.

Let us now prove Lemma 5.8.

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 $\varepsilon_p = \sqrt{\frac{2}{n}\log(16At^4)} \leq \sqrt{\frac{8}{n}\log(2At)}$. Define the events:

$$A_n = \left(\|\hat{p}^{(m)}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log(2At)} \right) \tag{5.48}$$

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:

$$A_n^{\perp} = \left(\|\hat{p}^{\perp}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log(2At)} \right). \tag{5.49}$$

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 5.7:

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

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},

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

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

$$B_n := \left(|\hat{r}^{(m)}(s,a) - r(s,a)| \geq \delta_{\max}\sqrt{\frac{2}{n}\log(2At)} \right). \tag{5.50}$$

By definition of $\hat{r}^{(m)}(s,a) = \gamma_{\text{reject}}\hat{p}^{(m)}(s+1|s,a) + \frac{\gamma_{\text{hold}}}{U}(S'-s)$ 5.15, and using that $\gamma_{\text{reject}} \leq \delta_{\max}$, we can write:

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

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:

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \leq \frac{S}{2t^3} + \frac{8CSA}{t^3}$$

as desired. □

5.11 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 5.8.6 to compare the optimistic MDP and the true MDP.

5.11.1 Lemmas on the Bias Differences

The next four lemmas of this subsection are already proved in Chapter 4, for the sake of completeness, we rewrite them in this subsection. They are used in the proof of Lemma 5.12, to control the difference between the bias of the policy $\tilde{\pi}_k$ in the optimistic MDP and in the true MDP.

Lemma 5.9

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:

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

which proves the lemma. □

Lemma 5.10

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:

$$\begin{aligned}\left((P - P') \sum_{t=0}^{T} P^t r \right)(s) &\leq \|p(\cdot|s) - p'(\cdot|s)\|_1 \cdot \text{span } (J(T)) \\ &\leq r_{\max} D^\pi \|p(\cdot|s) - p'(\cdot|s)\|_1 \text{ by Lemma 5.9} \\ &= b(s).\end{aligned}$$

In the same manner we show that:

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

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

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

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 5.11

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 $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}\left[\tau_{i \to S}\right]$, where $\mathbb{E}\left[\tau_{i \to S}\right]$ is the expected time to reach state S from state i.

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

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

In the following lemma, we use the same notations as in Lemma 5.10 with a common state space $\{0, 1, \dots S\}$.

Lemma 5.12
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 $\sup_{s \in \mathcal{S}} \mathbb{E}\left[\tau^{\pi}_{s \to s'}\right]$ be the worst expected hitting time to reach the state s' with policy π, and call $T_{hit} := \inf_{s' \in \mathcal{S}} \sup_{s \in \mathcal{S}} \mathbb{E}\left[\tau^{\pi}_{s \to s'}\right]$. We have the following control of the difference:

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

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 [Ipsen and Meyer, 1994, 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 5.11, 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 5.9:

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

Notice that $\|dA_S\|_\infty \leq \|dP\|_\infty$ and $|d\tilde{g}| = |dg|$. Using Lemma 5.10 and Lemma 5.11, and taking the infimum for the choice of the state of renormalization implies the claimed inequality for the biases. □

5.11.2 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 also be found in Chapter 4.

Lemma 5.13

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 $Q_{\max} := \left(\frac{10C_2S^2}{\nu^{\pi^{\max}}(S')}\right)^2 \log\left(\left(\frac{10C_2S^2}{\nu^{\pi^{\max}}(S')}\right)^4\right)$,

$C_2 := \frac{(\lambda\gamma_{reject}+\gamma_{hold})C_1}{\mu(1)(1-\lambda/\mu(i_0))}$ *and* C_1 *as in Lemma 5.15. 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 5.13:

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 2.12, following the same proof as in section 4.3.2 of [Jaksch et al., 2010], 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 5.16 proved in Section 5.12, 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 5.13 that we will use for the regret computations:

Corollary 5.14
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 5.13, 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. □

5.12 Properties of the Aggregated MDP

In this section, we prove properties on the aggregated 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 5.8.4 and 5.8.5.

5.12.1 Properties of the Policies in the Aggregated 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 aggregated MDP is stable (as seen in Section 5.2), 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 5.15

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 $s \leq S$:*

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

In order to control the variation of the bias of any policy π, we refer to Lemma 2.4 to first compute hitting times in the MDP under this policy, in order to show a more complete version of Proposition 2.5. The variation of the bias indeed play a major role in the computations of the main term of the regret (see 5.8.4).

Proposition 5.16

For any policy π, define for $s \in \{1, \dots, S'\}$ the variation of the bias

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

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

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

Using the monotonicity of the rates μ from Lemma 5.1, we therefore have :

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

Proof. We continue the proof of Proposition 2.5 where it stopped. We can now use the specific stationary measure we computed previously. We had:

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

and using Lemma 2.4 and Lemma 5.15:

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

□

5.13 Conclusion

In this chapter, 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 5.2, 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 [Jaksch et al., 2010] and that is exponential in the size of the space S in most queueing systems.

While our result strongly relies on Norton's equivalent 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 [Kritzinger et al., 1982] and also used in the context of non-product-form queueing networks for approximate analysis [Krieger, 2008].

6 Conclusions and Future Work

6.1 Conclusions

In this thesis, we mainly focused on highlighting the relevance of the queueing structure in some model-based reinforcement learning problems in the average reward setting. More precisely, we presented the example of a controlled birth-and-death process and showed how an adapted version of the classical algorithm UCRL2 could take into account the structure of the queue. In this case, we found a regret upper bound that is independent of the diameter and the size of the state space. Instead, it depends on the number of relevant states to explore. To achieve this bound, we controlled the bias variations for any policy of the MDP and used these estimations of the bias rather than the diameter. We also showed that the diameter increases exponentially in the size of the state space in simple queueing problems, which is the main motivation to remove it from the regret bounds.

We then investigated the typical example of the admission control problem in a queueing network whose states are not observed that leverages the structure of the problem. Using Norton's equivalence theorem and ergodicity properties, it is possible to simplify the underlying POMDP by learning an asymptotically equivalent MDP with a birth-and-death structure instead. In this case again, the proposed regret upper bound does not depend on the diameter anymore, and it only depends on the size of the state space through the mixing time of the original MDP.

6.2 Future Work

Improvement on the regret bound in the adapted UCRL2 and UCRL-M In these algorithms, we observe that the regret bounds depend on the stationary measure and on the mixing rate for the worst policies. Intuitively, the main contributions to the regret stem from the execution of suboptimal policies that get closer and closer to the optimal one in performance. We could expect that the stationary measures of these policies also get close to the stationary measure of the optimal one, so that only the optimal policy should be involved in the final regret term, rather than

the reference policy we had in Chapter 4. Regarding the mixing rate of policies in ergodic MDPs, we built modules in Chapter 5 according to the worst mixing time, being the time needed to reach stationarity from the furthest state, which therefore depends on S. However, by stability, the average queue size does not depend on S, so that the average mixing time is much lower than S, and we could build modules with this property in mind to improve the algorithm and its regret bound.

Lower bound of the regret for structured MDPs One of the most natural questions about the algorithms in Chapters 4 and 5 is about finding a lower bound of the regret, just like UCRL2 had a lower bound in $\Omega(\sqrt{DSAT})$, evaluate the tightness of our upper bounds. To do so, it would require clearly defining and choosing the class of MDPs we would like our algorithms to be applied to, while still keeping the question relevant. The lower bound in [Jaksch et al., 2010] typically revolved around finding a specific MDP with a given diameter D that would make the exploration as slow as possible, so that we could expect in our case to find a queue within the given class with a well-chosen stationary measure and rewards to hinder exploration.

Using the queueing structure in the parametric case The MDPs considered in [Wu et al., 2022] are linear mixture MDPs where the goal is to learn an unknown feature vector in $\mathbb{R}^d$, with d the dimension of the parameter space. We recall that the regret bound achieved in this case is in $O(D\sqrt{dT})$ by using Hoeffding-type bounds, and in $O(\sqrt{dDT})$ with Bernstein-type bounds. While the latter bound is stronger, it seems to rely on the knowledge of the diameter D to control the variance of the transitions. Sticking to Hoeffding-type bounds, however, would be an interesting first step to deal with linear mixture MDPs with a queueing structure. In this setting, it could be expected to work around the diameter in the analysis as done in Chapter 4. Moreover, it could still be interesting to study this algorithm with the Bernstein-type bounds, keeping in mind that the algorithm UCRL2B [Fruit, Pirotta, and Lazaric, 2020] relies on the empirical variance and not directly on the diameter, so that we could find an improvement on the bound, just like UCRL2B improved the regret bound from being proportional to D to $\sqrt{D}$.

Additionally in the parametric case, we could also explore model-free algorithms, that is, algorithms that do not compute and update estimates of every reward and transition to learn an optimal policy. Such algorithms are interesting for queueing systems where the state space is large, as the space complexity is larger for model-based algorithms in comparison to model-free ones. Following the same ideas from Chapter 5, we could use the ergodicity of the MDP in a Kiefer-Wolfowitz type

of algorithm [Kiefer and Wolfowitz, 1952] where the reward to maximize would depend on the stationary measure, as in the models of Chapters 3 or Chapter 5. The ergodicity could be used to sample approximately a state from the stationary measure and estimate the gradient of the average reward.

Involving other queueing properties to find algorithms The algorithms we presented relied so far on the birth-and-death structure of transitions to control the bias of policies. Just like we used Norton's equivalence theorem in Chapter 5, we could hope to design reinforcement learning algorithms using specific queueing properties, such as reversibility for example. The intuition is that these properties are connected to the stationary measures, and the optimal gain itself depends on the stationary measure of the optimal policy.

List of publications

Anselmi, Jonatha, Bruno Gaujal, and Louis-Sébastien Rebuffi (Dec. 2021). "Optimal Speed Profile of a DVFS Processor under Soft Deadlines". In: *Performance Evaluation* 152. DOI: 10.1016/j.peva.2021.102245. URL: https://hal.archives-ouvertes.fr/hal-03364880. cit. on p. 23

– (Nov. 2022). "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. URL: https://hal.science/hal-03799394. cit. on p. 55

– (2023). *Learning Optimal Admission Control in Partially Observable Queueing Networks*. arXiv: 2308.02391 [cs.LG]. cit. on p. 89

Bibliography

Anselmi, Jonatha and Bruno Gaujal (Jan. 2014). "Efficiency of simulation in monotone hyper-stable queueing networks". In: *Queueing Systems* 76.1, pp. 51–72. DOI: 10.1007/s11134-013-9357-7. URL: https://hal.science/hal-01102977. cit. on p. 106

Auer, Peter and Ronald Ortner (2006). "Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press. URL: https://proceedings.neurips.cc/paper/2006/file/c1b70d965ca504aa751ddb62ad69c63f-Paper.pdf. cit. on p. 18

AWS Architecture Center (2022). https://aws.amazon.com/architecture. Online; accessed: 2023-06-19. cit. on p. 107

Azar, Mohammad Gheshlaghi, Ian Osband, and Rémi Munos (2017). "Minimax Regret Bounds for Reinforcement Learning". In: *International Conference on Machine Learning*, pp. 263–272. cit. on pp. 3, 53, 59, 107

Azizzadenesheli, Kamyar, Alessandro Lazaric, and Animashree Anandkumar (2016). "Reinforcement Learning of POMDPs using Spectral Methods". In: *COLT*. Vol. 49. JMLR Workshop and Conference Proceedings, pp. 193–256. cit. on pp. 3, 87, 88

Bansal, Nikhil, Tracy Kimbrel, and Kirk Pruhs (2007). "Speed Scaling to Manage Energy and Temperature". In: *Journal of the ACM* 54.1. cit. on p. 23

Bhandari, Jalaj, Daniel Russo, and Raghav Singal (June 2018). “A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research, pp. 1691–1692. URL: https://proceedings.mlr.press/v75/bhandari18a.html. cit. on p. 123

Bhulai, S., A. C. Brooms, and F. M. Spieksma (2014). “On structural properties of the value function for an unbounded jump Markov process with an application to a processor sharing retrial queue”. In: *Queueing Systems* 76.4, pp. 425–446. cit. on pp. 25, 30

Blok, H. and F. M. Spieksma (2015). “Countable state Markov decision processes with unbounded jump rates and discounted cost: optimality equation and approximations”. In: *Advances in Applied Probability* 47.4, pp. 1088–1107. cit. on pp. 25, 30, 31

Borgs, Christian, Jennifer Chayes, Sherwin Doroudi, Mor Harchol-Balter, and Kuang Xu (Jan. 2014). “The Optimal Admission Threshold in Observable Queues with State Dependent Pricing”. In: *Probability in the Engineering and Informational Sciences 28*, pp. 101–110. URL: https://www.microsoft.com/en-us/research/publication/optimal-admission-threshold-observable-queues-state-dependent-pricing/. cit. on pp. 89, 90, 91

Bourel, Hippolyte, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi (2020). “Tightening Exploration in Upper Confidence Reinforcement Learning”. In: *Proceedings of the 37th International Conference on Machine Learning*. ICML’20. JMLR.org. cit. on pp. 3, 61, 87

Boutilier, Craig, Richard Dearden, and Moisés Goldszmidt (Aug. 2000). “Stochastic Dynamic Programming with Factored Representations”. In: *Artif. Intell.* 121.1–2, pp. 49–107. DOI: 10.1016/S0004-3702(00)00033-3. URL: https://doi.org/10.1016/S0004-3702(00)00033-3. cit. on p. 54

Chandrakasan, A. P., S. Sheng, and R. W. Brodersen (1992). “Low-power CMOS digital design”. In: *IEEE Journal of Solid-State Circuits* 27.4, pp. 473–484. cit. on p. 29

Chandy, K. M., U. Herzog, and L. Woo (1975). “Parametric Analysis of Queuing Networks”. In: *IBM Journal of Research and Development* 19.1, pp. 36–42. DOI: 10.1147/rd.191.0036. cit. on pp. 89, 94

Configuring Concurrency in Knative (2022). Online; accessed: 2023-01-30. URL: https://knative.dev/docs/serving/autoscaling/concurrency/. cit. on p. 91

Dong, Kefan, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang (2019). “Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP”. In: *CoRR* abs/1901.09311. URL: http://arxiv.org/abs/1901.09311. cit. on p. 2

Dopper, Jantien G., Bruno Gaujal, and Jean-Marc Vincent (June 2006). “Bounds for the Coupling Time in Queueing Networks Perfect Simulation”. In: *MAM, 150th Anniversary of A.A. Markov*. Ed. by A.N. Langville and W.J. Stewart. Charleston, SC. cit. on p. 106

Even-Dar, Eyal, Sham M. Kakade, and Yishay Mansour (2005). “Reinforcement Learning in POMDPs without Resets”. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI’05. Edinburgh, Scotland, pp. 690–695. cit. on pp. 3, 88

Fruit, Ronan, Matteo Pirotta, and Alessandro Lazaric (2020). "Improved Analysis of UCRL2 with Empirical Bernstein Inequality". In: *ArXiv* abs/2007.05456. URL: https://api.semanticscholar.org/CorpusID:220486831. cit. on pp. 3, 22, 87, 138

Fruit, Ronan, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner (Feb. 2018). "Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning". In. : cit. on p. 61

Gast, Nicolas, Bruno Gaujal, and Kimang Khun (June 2021). *Reinforcement Learning for Markovian Bandits: Is Posterior Sampling more Scalable than Optimism?* Tech. rep. hal-03262006. HAL-Inria. cit. on p. 54

Gaujal, Bruno, Alain Girault, and Stéphan Plassart (July 2020). "Dynamic Speed Scaling Minimizing Expected Energy Consumption for Real-Time Tasks". In: *Journal of Scheduling*, pp. 1–25. URL: https://hal.inria.fr/hal-02888573. cit. on p. 23

Guestrin, Carlos, Daphne Koller, Ronald Parr, and Shobha Venkataraman (Oct. 2003). "Efficient Solution Algorithms for Factored MDPs". In: *Journal of Artificial Intelligence Research* 19.1, pp. 399–468. cit. on p. 54

Guo, Xianping and Onesimo Hernandez-Lerma (Jan. 2009). *Continuous-time Markov decision processes. Theory and applications*. Vol. 62. cit. on pp. 25, 27, 29

Guo, Zhaohan Daniel, Shayan Doroudi, and Emma Brunskill (2016). "A PAC RL Algorithm for Episodic POMDPs". In: *AISTATS*. Vol. 51. JMLR Workshop and Conference Proceedings, pp. 510–518. cit. on p. 87

Hyon, Emmanuel and Alain Jean-Marie (2020). "Optimal control of admission in service in a queue with impatience and setup costs". In: *Performance Evaluation* 144. Ed. by Elsevier. cit. on pp. 25, 30

Ipsen, Ilse C. F. and Carl Dean Meyer (1994). "Uniform Stability of Markov Chains". In: *SIAM Journal on Matrix Analysis and Applications* 15, pp. 1061–1074. cit. on pp. 83, 130

Jaksch, T., R. Ortner, and P. Auer (2010). "Near-optimal Regret Bounds for Reinforcement Learning." In: *Journal of Machine Learning Research* 11.4, pp. 1563–1600. cit. on pp. 2, 4, 9, 18, 19, 20, 21, 22, 53, 55, 61, 62, 63, 64, 65, 66, 68, 69, 70, 76, 77, 85, 102, 104, 110, 118, 119, 121, 124, 131, 134, 138

Jin, Chi, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan (2018). "Is Q-Learning Provably Efficient?" In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 4868–4878. cit. on p. 2

Jin, Chi, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu (2020). "Sample-Efficient Reinforcement Learning of Undercomplete POMDPs". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33, pp. 18530–18539. cit. on pp. 3, 87, 88

Jin, Chi, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan (Sept. 2020). "Provably efficient reinforcement learning with linear function approximation". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 2137–2143. URL: https://proceedings.mlr.press/v125/jin20a.html. cit. on pp. 3, 54, 61

Kameda, Hisao (1984). "A Property of Normalization Constants for Closed Queueing Networks". In: *IEEE Transactions on Software Engineering* SE-10.6, pp. 856–857. DOI: 10.1109/TSE.1984.5010314. cit. on p. 94

Kiefer, J. and J. Wolfowitz (1952). "Stochastic Estimation of the Maximum of a Regression Function". en. In: *Ann. Math. Statist.* 23.4, pp. 462–466. URL: http://dml.mathdoc.fr/item/1177729392. cit. on p. 139

Krieger, Udo R. (2008). "Queueing Networks and Markov Chains, 2nd Edition by G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi". In: *IIE Transactions* 40.5, pp. 567–568. DOI: 10.1080/07408170701623187. cit. on pp. 94, 95, 135

Kritzinger, P.S, S van Wyk, and A.E Krzesinski (1982). "A generalisation of Norton's theorem for multiclass queueing networks". In: *Performance Evaluation* 2.2, pp. 98–107. DOI: https://doi.org/10.1016/0166-5316(82)90002-5. URL: https://www.sciencedirect.com/science/article/pii/0166531682900025. cit. on p. 135

Levin, David A., Yuval Peres, and Elizabeth L. Wilmer (2008). *Markov chains and mixing times*. American Mathematical Society. URL: http://scholar.google.com/scholar.bib?q=info:3wf9IU94tyMJ:scholar.google.com/&output=citation&hl=en&as_sdt=2000&ct=citation&cd=0. cit. on pp. 15, 105

Li, Minming, Frances F. Yao, and Hao Yuan (Apr. 2017). "An $O(n^2)$ Algorithm for Computing Optimal Continuous Voltage Schedules". In: *TAMC'17*. Vol. 10185. LNCS. Bern, Switzerland, pp. 389–400. cit. on p. 23

Li, Quan-Lin, Jing-Yu Ma, Rui-Na Fan, and Li Xia (2019). "An Overview for Markov Decision Processes in Queues and Networks". In: *Stochastic Models in Reliability, Network Security and System Safety: Essays Dedicated to Professor Jinhua Cao on the Occasion of His 80th Birthday*. Ed. by Quan-Lin Li, Jinting Wang, and Hai-Bo Yu. Singapore: Springer Singapore, pp. 44–71. DOI: 10.1007/978-981-15-0864-6_3. URL: https://doi.org/10.1007/978-981-15-0864-6_3. cit. on p. 1

Lorch, Jacob R. and Alan Jay Smith (2001). "Improving Dynamic Voltage Scaling Algorithms with PACE". In: *ACM SIGMETRICS 2001 Conference*, pp. 50–61. cit. on p. 23

Osband, Ian and Benjamin Van Roy (Dec. 2014). "Near-Optimal Reinforcement Learning in Factored MDPs". In: *Proc. of the 27th Int. Conf. on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press, pp. 604–612. cit. on p. 54

Osband, Ian, Benjamin Van Roy, and Zheng Wen (2016). "Generalization and Exploration via Randomized Value Functions". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, pp. 2377–2386. cit. on p. 54

Papadimitriou, Christos H. and John N. Tsitsiklis (Aug. 1987). "The Complexity of Markov Decision Processes". In: *Math. Oper. Res.* 12.3, pp. 441–450. cit. on p. 88

Poupart, Pascal and Nikos A. Vlassis (2008). "Model-based Bayesian Reinforcement Learning in Partially Observable Domains". In: *International Symposium on Artificial Intelligence and Mathematics*. cit. on pp. 3, 88

Puterman, Martin L (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons. cit. on pp. 24, 33, 34

– (Apr. 1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. en. 1st ed. Wiley Series in Probability and Statistics. Wiley. cit. on pp. 1, 7, 8, 19, 66, 121, 122

Robbins, Herbert E. (1952). "Some aspects of the sequential design of experiments". In: *Bulletin of the American Mathematical Society* 58, pp. 527–535. URL: https://api.semanticscholar.org/CorpusID:15556973. cit. on p. 1

Rolia, J.A. and K.C. Sevcik (1995). "The Method of Layers". In: *IEEE Transactions on Software Engineering* 21.8, pp. 689–700. DOI: 10.1109/32.403785. cit. on p. 91

Rolia, Jerry, Giuliano Casale, Diwakar Krishnamurthy, Stephen Dawson, and Stephan Kraft (2009). "Predictive Modelling of SAP ERP Applications: Challenges and Solutions". In: VALUETOOLS '09. Pisa, Italy. DOI: 10.4108/ICST.VALUETOOLS2009.7988. URL: https://doi.org/10.4108/ICST.VALUETOOLS2009.7988. cit. on p. 91

Rosenberg, Aviv and Yishay Mansour (Sept. 2020). "Oracle-Efficient Reinforcement Learning in Factored MDPs with Unknown Structure". In: *arXiv:2009.05986 [cs, stat]*. arXiv: 2009.05986 [cs, stat]. cit. on p. 54

Ross, Stephane, Brahim Chaib-draa, and Joelle Pineau (2007). "Bayes-Adaptive POMDPs". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. URL: https://proceedings.neurips.cc/paper/2007/file/3b3dbaf68507998acd6a5a5254ab2d76-Paper.pdf. cit. on pp. 3, 88

Shaked, Moshe and J George Shanthikumar (1994a). *Stochastic orders and their applications*. Academic Pr. cit. on p. 44

– (1994b). *Stochastic orders and their applications*. Academic Pr. cit. on p. 57

Snowdon, David C., Sergio Ruocco, and Gernot Heiser (Sept. 2005). "Power Management and Dynamic Voltage Scaling: Myths and Facts". In: *Proc. of the 2005 Workshop on Power Aware Real-time Computing*. New Jersey, USA. cit. on p. 26

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press. cit. on p. 1

Tian, Yi, Jian Qian, and Suvrit Sra (2020). "Towards Minimax Optimal Reinforcement Learning in Factored Markov Decision Processes". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. cit. on p. 54

Tossou, Aristide, Debabrota Basu, and Christos Dimitrakakis (2019). *Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities*. DOI: 10.48550/ARXIV.1905.12425. URL: https://arxiv.org/abs/1905.12425. cit. on pp. 3, 22, 53, 87

Urgaonkar, Bhuvan, Giovanni Pacifici, Prashant Shenoy, Mike Spreitzer, and Asser Tantawi (June 2005). "An Analytical Model for Multi-Tier Internet Services and Its Applications". In: *SIGMETRICS Perform. Eval. Rev.* 33.1, pp. 291–302. DOI: 10.1145/1071690.1064252. URL: https://doi.org/10.1145/1071690.1064252. cit. on p. 107

Vlassis, Nikos, Michael L. Littman, and David Barber (Nov. 2012). "On the Computational Complexity of Stochastic Controller Optimization in POMDPs". In: *ACM Trans. Comput. Theory* 4.4. DOI: 10.1145/2382559.2382563. URL: https://doi.org/10.1145/2382559.2382563. cit. on p. 88

Walton, Neil and Kuang Xu (2021). *Learning and Information in Stochastic Networks and Queues*. DOI: 10.48550/ARXIV.2105.08769. URL: https://arxiv.org/abs/2105.08769. cit. on p. 1

Wang, Runan, Giuliano Casale, and Antonio Filieri (Nov. 2022). "Estimating Multiclass Service Demand Distributions Using Markovian Arrival Processes". In: *ACM Trans. Model. Comput. Simul.* DOI: 10.1145/3570924. URL: https://doi.org/10.1145/3570924. cit. on p. 92

Wei, Chen-Yu, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain (July 2020). "Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research, pp. 10170–10180. URL: https://proceedings.mlr.press/v119/wei20c.html. cit. on p. 2

Williams, David (1991). *Probability with Martingales*. Cambridge University Press. DOI: 10.1017/CBO9780511813658. cit. on p. 21

Wu, Yue, Dongruo Zhou, and Quanquan Gu (28–30 Mar 2022). "Nearly Minimax Optimal Regret for Learning Infinite-horizon Average-reward MDPs with Linear Function Approximation". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 3883–3913. URL: https://proceedings.mlr.press/v151/wu22a.html. cit. on pp. 3, 61, 87, 138

Xia, Li (June 2014). "Event-Based Optimization of Admission Control in Open Queueing Networks". In: *Discrete Event Dynamic Systems* 24.2, pp. 133–151. DOI: 10.1007/s10626-013-0167-1. URL: https://doi.org/10.1007/s10626-013-0167-1. cit. on pp. 89, 90

Xu, Ziping and Ambuj Tewari (2020). "Reinforcement Learning in Factored MDPs: Oracle-Efficient Algorithms and Tighter Regret Bounds for the Non-Episodic Setting". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. URL: https://proceedings.neurips.cc/paper/2020/hash/d3b1fb02964aa64e257f9f26a31f72cf-Abstract.html. cit. on p. 54

Yao, F., A. Demers, and S. Shenker (1995). "A scheduling model for reduced CPU energy". In: *Proceedings of IEEE Annual Foundations of Computer Science*, pp. 374–382. cit. on p. 23

Zhang, Zihan and Xiangyang Ji (2019). *Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function*. DOI: 10.48550/ARXIV.1906.05110. URL: https://arxiv.org/abs/1906.05110. cit. on pp. 3, 53, 61

Zhou, Dongruo, Jiafan He, and Quanquan Gu (18–24 Jul 2021). "Provably Efficient Reinforcement Learning for Discounted MDPs with Feature Mapping". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 12793–12802. URL: https://proceedings.mlr.press/v139/zhou21a.html. cit. on pp. 3, 61

Abstract

Although reinforcement learning has been recently primarily studied in the generic case of Markov decision processes, the queueing systems case stands out in particular. To deal with the potentially extremely large state space *a priori*, learning algorithms must take into account the structure of the systems in order to extract as much information as possible and choose the best control that optimizes the system performance in the long run.

In this thesis, we present algorithms adapted from classical algorithms in the context of queueing systems, and we study their performance to demonstrate a weak dependence on the state space compared to results obtained in the general case.

Résumé

Bien que l'apprentissage par renforcement ait été récemment principalement étudié dans le cas générique des processus de décisions markoviens, le cas des systèmes de files d'attente se dinstigue particulièrement. Pour compenser la taille de l'espace d'état qui peut être extrêmement grande *a priori*, les algorithmes d'apprentissage doivent tenir compte de la structure des systèmes afin d'en extraire le plus d'information et de choisir le meilleur contrôle qui optimisent au mieux les performances du système sur le long terme.

Dans cette thèse, nous présentons des algorithmes construits à partir d'algorithmes classiques, adaptés au contexte des système de file d'attente, et nous étudions les performances de ceux-ci pour montrer une dépendance faible à l'espace d'états comparativement aux résultats obtenus dans le cas général.